PERSONALISED DIETARY ASSESSMENT: OPTIMISING THE SIZE OF QUESTIONNAIRES AND DYNAMIC QUESTION ORDERING IN FOOD FREQUENCY QUESTIONNAIRES

Nina Reščič

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Dr. Mitja Luštrek, Jožef Stefan Institute, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
Co-Supervisor: Prof. Dr. Barbara Koroušić Seljak, Jožef Stefan Institute, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

Evaluation Board:
Prof. Dr. Matjaž Gams, Chair, Jožef Stefan Institute, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
Prof. Dr. Zoran Bosnić, Member, Faculty of Computer Science, University of Ljubljana, Ljubljana, Slovenia
Assoc. Prof. Dr. Riste Stojanov, Member, Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, North Macedonia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL



Nina Reščič

PERSONALISED DIETARY ASSESSMENT: OPTIMISING THE SIZE OF QUESTIONNAIRES AND DYNAMIC QUESTION ORDERING IN FOOD FREQUENCY QUESTIONNAIRES

Doctoral Dissertation

PERSONALIZIRANA PREHRANSKA OBRAVNAVA: OPTIMIZACIJA ŠTEVILA VPRAŠANJ IN DINAMIČNO RAZVRŠČANJE VPRAŠANJ V PREHRANSKIH VPRAŠALNIKIH

Doktorska disertacija

Supervisor: Prof. Dr. Mitja Luštrek

Co-Supervisor: Prof. Dr. Barbara Koroušić Seljak

Ljubljana, Slovenia, January 2026

To my mom. *None of the following pages would have happened if it weren't for you. Not only because of your constant reminders, "Nina, don't forget about your PhD," but most importantly because you were always there: believing in me, putting me before anything or anyone else, including yourself, and helping me develop my full potential. You always reminded me that, while important, academic and other achievements matter far less than being a good and kind person.*

To my friends. *Tina, for every time you swept pieces of me into a pile and put me back together. Mojca and Staša, for each and every "Tuesday," whether it happened on a Tuesday or any other day. Dominik, because you are. And of course, because Mufi. And to so many more who stood by me, encouraged me, and believed in me. I hope I can be there for you too, even if all you need is an excuse: "I wouldn't myself, but you know, the doctor ordered it." After all, there's no need to disclose what kind of doctor, right?*

To me. *Let this be a reminder of what I am capable of. It took time, and over these years I lived through some of the hardest moments of my life. I was broken into pieces and came close to forgetting who I was. Yet every single time, I found my way back, got up, and fought on. After all, as Tina and I like to say: "First, we're going to cry a bit. Then we'll get up and do our thing. And we'll do it well. Really well."*

Acknowledgments

I would like to thank my supervisor, Mitja Luštrek, for supervising this PhD and for his guidance throughout the process. I am grateful for the care he took in ensuring that the thesis met the academic standards and requirements of the doctoral program, in particular through his careful reading of its many versions and the feedback he provided.

I am very grateful to my co-supervisor, Barbara Koroušić Seljak, for her support, encouragement, and thoughtful perspective throughout this PhD. I especially appreciate the opportunities she has given me to explore related topics beyond the immediate scope of this thesis. These opportunities broadened my perspective and helped me develop my own research interests.

I would also like to thank the members of my Evaluation Board, Prof. Dr. Matjaž Gams, Prof. Dr. Zoran Bosnić, and Assoc. Prof. Dr. Riste Stojanov, for their time, careful reading of this work, and constructive feedback. I am grateful for their thoughtful comments and suggestions, which helped me refine the thesis and present my work more clearly.

I gratefully acknowledge the financial support from the WellCo project, funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 769765, which made this research possible. Additional support from the Slovenian Research and Innovation Agency, through Research Core Funding No. P2-0209, is also gratefully acknowledged.

I am grateful to my colleagues and collaborators for the many discussions, ideas, and moments of support that shaped both this research and my experience throughout the PhD. I especially thank Eva Peklaj, whose deep knowledge of nutrition and enthusiasm for sharing her expertise helped me better understand which questions are most meaningful and why introducing AI into an often-overlooked field is important.

I am also deeply grateful to my friends, both within and outside of work, and to my extended family for their unwavering support, encouragement, and understanding throughout this journey.

Finally, and most importantly, I want to thank my mom. Her love, patience, encouragement, and sacrifices made everything in my life possible, but above all, her example shaped the person I am today. She taught me to work hard, stay humble, keep going, and remember that no achievement truly matters without kindness and compassion toward others.

Abstract

Food Frequency Questionnaires (FFQs) are widely used tools for dietary assessment in epidemiological studies and digital health applications. While efficient and low-cost, FFQs often suffer from incomplete responses, high respondent burden, and limited adaptability to individual users. These limitations reduce data quality and compromise the reliability of downstream dietary estimates. This PhD investigates how statistical and machine learning methods can be used to improve the efficiency, robustness, and adaptability of FFQs without sacrificing predictive accuracy.

The thesis focuses on three interconnected problems: reducing the number of required questions, mitigating the impact of missing responses, and dynamically selecting the most informative questions during questionnaire completion. First, feature selection and feature ranking methods are applied to identify compact subsets of FFQ questions that preserve predictive performance for dietary quality scores and nutrient intake estimates. Multiple approaches are evaluated, including filter, embedded, and wrapper methods, demonstrating that substantial questionnaire reduction is possible with minimal loss in accuracy.

Second, static question-ordering strategies are introduced to improve predictive quality under partial questionnaire completion. By ranking questions by their predictive importance, the most informative responses are collected early, thereby making dietary estimates more resilient to incomplete data. This approach consistently outperforms conventional missing-data handling techniques.

This thesis introduces a new method for dynamically ordering questionnaire items based on expected utility. Instead of a fixed order, each next question is chosen based on the answers given so far. The method combines predicted answer probabilities with measures of how much each question is expected to improve predictions and reduce uncertainty, allowing the questionnaire to adapt to each individual. The approach is further extended to predict multiple dietary outcomes simultaneously using a single model, ensuring consistent predictions across related targets. Weighted utilities are used to prioritise specific outcomes when required.

The proposed methods are tested on data from the Extended Short Form Food Frequency Questionnaire (ESFFFQ) and NHANES. Results on NHANES show the limits of predicting laboratory biomarkers using FFQs alone, but the consistent patterns in performance demonstrate that the approach is robust. Overall, this work shows that machine learning can be used to make FFQs shorter, adaptive, and more reliable, and provides a general framework for designing intelligent questionnaires in nutrition and other fields.

Povzetek

Vprašalniki o pogostosti uživanja živil (ang. *"Food Frequency Questionnaires – FFQ"*) so pogosto uporabljeno orodje za ocenjevanje prehranskih navad v epidemioloških raziskavah in različnih (mobilnih) aplikacijah. Kljub učinkovitosti pa so vprašalniki lahko predvsem dolgi vprašalniki lahko nekoliko problematični zaradi obremenitve za uporabnika in posledično delno odgovorjenih vprašalnikov. Te omejitve zmanjšujejo kakovost podatkov in ogrožajo zanesljivost nadaljnjih ocen prehrane. Doktorska disertacija preučuje, kako je mogoče s statističnimi metodami in metodami strojnega učenja izboljšati učinkovitost, robustnost in prilagodljivost vprašalnikov, ne da bi pri tem žrtvovali napovedno natančnost.

Disertacija se osredotoča na tri medsebojno povezane probleme: zmanjšanje števila potrebnih vprašanj, zmanjšanje vpliva manjkajočih odgovorov ter dinamično izbiro najbolj informativnih vprašanj med izpolnjevanjem vprašalnika.

V prvem koraku so uporabljene metode za zmanjšanje števila vprašanj ter izbiro značilk za identifikacijo podmnožic vprašanj vprašalnika, ki ohranjajo napovedno uspešnost pri ocenah kakovosti prehrane in vnosa hranil. Obravnavani so različni pristopi, vključno s filtrskimi, ovojnimi (ang. *"wrapper"*) metodami.

V drugem koraku so uvedene statične metode razvrščanja vprašanj za izboljšanje kakovosti napovedi ob delnem izpolnjevanju vprašalnika. Z razvrščanjem vprašanj glede na njihovo napovedno pomembnost se najbolj informativni odgovori zberejo že zgodaj, kar poveča odpornost prehranskih ocen na nepopolne podatke.

Osrednji prispevek disertacije je metoda za dinamično razvrščanje vprašanj, ki temelji na pričakovani uporabnosti (ang. *"expected utility"*). Predlagana metoda obravnava izpolnjevanje vprašalnika kot zaporedni odločitveni problem, pri katerem je vsako naslednje vprašanje izbrano na podlagi predhodno opaženih odgovorov. Pričakovana uporabnost združuje naučene pogojne verjetnosti odgovorov z modelno utemeljenimi merami spremembe napovedi in zmanjšanja negotovosti, kar omogoča večjo personalizacijo vprašalnikov. Metoda dinamičnega razvrščanja vprašanj je dodatno razširjena na večciljno napovedovanje, kar omogoča hkratno napovedovanje večih napovednih ciljev z enim samim modelom.

Predlagane metode so ovrednotene na podatkih zbranih z uporabo vprašalnika Extended Short Form Food Frequency Questionnaire (ESFFFQ) ter na podatkih prehranskega vprašalnika v okviru raziskave NHANES. Čeprav rezultati na podatkih NHANES kažejo omejeno napovedovanje laboratorijskih biomarkerjev zgolj na podlagi prehranskih vprašalnikov, pa se določeni trendi jasno nakazujejo tudi na teh podatkih. Doktorska dizertacija tako najprej pokaže, da je mogoče prehranske vprašalnike s pomočjo strojnega učenja skrajšati, prilagoditi in narediti bolj odporne, ter končno prispeva tudi metodologijo za personalizirano zasnovo vprašalnikov na področju prehrane (in širše).

Contents

List of Figures **xvii**

List of Tables **xxi**

List of Algorithms **xxiii**

Abbreviations **xxv**

1 Introduction **1**
1.1 Background . . . 1
1.2 Problem Description . . . 2
1.3 Goals of the Dissertation . . . 3
1.4 Hypotheses . . . 4

2 Related Work **5**
2.1 Dietary Assessment . . . 5
2.2 Food Frequency Questionnaires . . . 7
2.2.1 About FFQs . . . 7
2.2.2 FFQs in Mobile Applications . . . 9
2.2.3 FFQs in Relation to Health . . . 10
2.3 Personalisation of Questionnaires . . . 10
2.3.1 Prediction from Partially Completed Questionnaires . . . 11
2.3.2 Adaptive Questionnaires . . . 12
2.3.3 Dynamic Feature Acquisition for Online Questionnaires . . . 13
2.3.4 Multi-Target Prediction in Dynamic Question Ordering . . . 15

3 Datasets **17**
3.1 ESFFFQ Dataset . . . 17
3.1.1 Questionnaire Design and Targets . . . 17
3.1.1.1 Questionnaire Items and Response Format . . . 18
3.1.2 Scoring Methodology . . . 18
3.2 NHANES Dataset . . . 21
3.2.1 Overview and Motivation . . . 21
3.2.2 NHANES Dietary Questionnaires . . . 21
3.2.3 Feature Extraction . . . 22
3.2.4 Target Variables: Nutrition-Related Biomarkers . . . 22
3.2.5 Dataset Suitability . . . 24

4 Dietary Assessment: Predictive Modelling and Question Reduction **27**
4.1 Prediction in Questionnaires using Statistical and Machine Learning Methods 27
4.1.1 Statistical Methods for Handling Missing Data . . . 27
4.1.1.1 Zero Imputation . . . 28

4.1.1.2 Multiple Imputation by Chained Equations (MICE) 28
4.1.1.3 Comparison and Limitations 28
4.1.2 Machine Learning Methods 28
4.1.2.1 Linear Models 29
4.1.2.2 Ensemble Models 29
4.2 Optimising the Efficiency of FFQs through Question Reduction 29
4.2.1 Target-Independent Question Reduction Methods 30
4.2.1.1 Weighted Correlation for Dimensionality Reduction 30
4.2.1.2 SHAP for Feature Interactions 31
4.2.2 Target-Dependent Question Reduction Methods 33
4.2.2.1 Filter Methods 33
4.2.2.2 Embedded Methods 33
4.2.2.3 Wrapper Methods 34
4.2.2.4 Genetic Algorithms for Feature Selection 34

5 Optimisation of Question Ordering for Predictive Modelling 35
5.1 Static Question Ordering 35
5.1.1 Static Question Ordering for Single-Target Prediction Problems . . . 35
5.1.2 Static Question Ordering for Multi-Target Prediction Problems . . . 36
5.2 Dynamic Question Ordering 36
5.2.1 Dynamic Question Ordering for Single-Target Prediction Problems . 37
5.2.2 Dynamic Question Ordering for Multi-Target Prediction Problems . 41

6 Evaluation and Results 45
6.1 Evaluation 45
6.1.1 Evaluation of Predictive Models 45
6.1.2 PROMETHEE 46
6.1.3 Evaluation of Static and Dynamic Question Ordering 49
6.2 Results 50
6.2.1 Findings from Target-Independent Question Reduction Methods . . 50
6.2.1.1 Comparison of Question Reduction Methods 54
6.2.2 Findings from Target-Dependent Feature Selection Methods for FFQ Dimensionality Reduction 59
6.2.3 Findings from Single-Target Static Question Ordering Method 63
6.2.4 Findings from Multi-Target Static Question Ordering Method 67
6.2.5 Findings from Single-Target Dynamic Question Ordering 72
6.2.6 Findings from Multi-Target Dynamic Question Ordering 77
6.2.6.1 ESFFFQ 77
6.2.6.2 NHANES 79

7 Conclusion 81
7.1 Summary of Findings 81
7.1.1 Conclusion and Contributions to the Field 81
7.1.2 Practical Recommendations 83
7.1.3 Future Work 83

Appendix A Additional Results 85
A.1 Feature Reduction 85
A.1.1 Target-Independed Feature Reduction 85
A.1.2 Target Depended Feature Selection 92
A.2 Feature Ranking 100

A.2.1 Static Feature Ranking 100

References **107**

Bibliography **115**

Biography **119**

List of Figures

Figure 6.1: PROMETHEE pipeline for ranking feature subsets. F represents feature subsets; M denotes machine-learning methods. F_{w_i} is the best-performing subset for method M_i, $F_{w_{T_1}}$ is the best for target T_1, and F_w performs best across all targets. 49
Figure 6.2: Comparison of the F1-scores of different models on subsets created using SHAP and Weighted Correlation for target fat. 52
Figure 6.3: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target fat. 53
Figure 6.4: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Total Cholesterol. 54
Figure 6.5: Comparison of SHAP and Weighted Correlation for classification. . . . 55
Figure 6.6: Comparison of SHAP and Weighted Correlation for regression. 56
Figure 6.7: Comparison of SHAP and Weighted Correlation for regression. 58
Figure 6.8: Comparison of F1-score on subsets of features selected by different methods for target Fat. 60
Figure 6.9: Comparison of MAPE on subsets of features selected by different methods for target Fat. 61
Figure 6.10: Comparison of MAPE on subsets of features selected by different methods for target Total Cholesterol. 63
Figure 6.11: Comparison of static ranking algorithms for FAT: classification performance (F1 Score, top row) and regression performance (MAE in grams, bottom row). Column titles denote the feature representations. 64
Figure 6.12: Comparison of feature representations for FAT: classification performance (F1 Score, top row) and regression performance (MAE in grams, bottom row). 65
Figure 6.13: Comparison of static ranking algorithms for total cholesterol: regression performance MAE (mmol/L) for all available features and for 20 best features. 66
Figure 6.14: Comparison of feature representations for Total cholesterol: regression performance MAE (mmol/L) for all available features and for 20 best features. 66
Figure 6.15: Comparison of feature ranking approaches for classification performance F1-score for all available targets. 67
Figure 6.16: Comparison of feature ranking approaches for regression performance MAE (g/day) for all available targets. 68
Figure 6.17: Purple curves show single-target prediction (STP) rankings; green curves show multi-target prediction (MTP). Solid lines correspond to equal target weighting; dash–dot lines use weighted targets. The left panel shows results with raw features, and the right with frequency-transformed features. 68

Figure 6.18: Purple curves show single-target prediction rankings; green curves show multi-target prediction. Solid lines use equal-target weighting; dash-dot lines use weighted targets. The left panel shows results with raw features, the right with frequency-transformed features. Lower curves indicate more accurate nutrient predictions (MAE). 69
Figure 6.19: Comparison of feature ranking approaches for regression performance MAE (mmol/L) averaged over all available targets. 71
Figure 6.20: Purple curves show STP rankings; green curves show MTP. Solid lines use equal-target weighting; dash-dot lines use weighted targets. The left panel shows results with raw features, the right with frequency-transformed features. Lower curves indicate more accurate nutrient predictions (MAE). 72
Figure 6.21: Regression performance (MAE in grams) for static ranking (train and test) and dynamic ranking methods. 73
Figure 6.22: Regression performance (MAE in mmol/L) for static ranking (train and test) and dynamic ranking methods. 75
Figure 6.23: Regression performance (MAE in grams) for static ranking (train and test) and dynamic ranking methods in multi-target prediction problem. 77
Figure 6.24: Regression performance (MAE in grams) over all targets for static ranking and different weighted utility functions in dynamic ranking in multi-target prediction problem. 78
Figure 6.25: Regression performance (MAE in mmol/L) for static ranking (train and test) and dynamic ranking methods in multi-target prediction problem. 79
Figure 6.26: Regression performance (MAE in mmol/L) over all targets for static ranking and different weighted utility functions in dynamic ranking in MTP. 80

Figure A.1: Comparison of the F1-scores of different models on subsets created using SHAP and Weighted Correlation for target Sugar. 85
Figure A.2: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Sugar. 86
Figure A.3: Comparison of the F1-scores of different models on subsets created using SHAP and Weighted Correlation for target Fiber. 87
Figure A.4: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Fiber. 87
Figure A.5: Comparison of the F1-scores of different models on subsets created using SHAP and Weighted Correlation for target Protein. 88
Figure A.6: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Protein. 89
Figure A.7: Comparison of the F1-scores of different models on subsets created using SHAP and Weighted Correlation for target Salt. 89
Figure A.8: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Salt. 90
Figure A.9: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target HDL Cholesterol. 91
Figure A.10: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target LDL Cholesterol. 91
Figure A.11: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Glucose. 92
Figure A.12: Comparison of F1-score on subsets of features selected by different methods for target Sugar. 93

Figure A.13: Comparison of MAPE on subsets of features selected by different methods for target Sugar. . . . 93
Figure A.14: Comparison of F1-score on subsets of features selected by different methods for target Fibre. . . . 94
Figure A.15: Comparison of MAPE on subsets of features selected by different methods for target Fibre. . . . 95
Figure A.16: Comparison of F1-score on subsets of features selected by different methods for target Protein. . . . 95
Figure A.17: Comparison of MAPE on subsets of features selected by different methods for target Protein. . . . 96
Figure A.18: Comparison of F1-score on subsets of features selected by different methods for target Salt. . . . 97
Figure A.19: Comparison of MAPE on subsets of features selected by different methods for target Salt. . . . 97
Figure A.20: Comparison of MAPE on subsets of features selected by different methods for target LDL cholesterol. . . . 98
Figure A.21: Comparison of MAPE on subsets of features selected by different methods for target HDL cholesterol. . . . 99
Figure A.22: Comparison of MAPE on subsets of features selected by different methods for target Glucose. . . . 99
Figure A.23: Comparison of static ranking algorithms for sugar: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom). . . . 100
Figure A.24: Comparison of feature representations for sugar: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom). . . . 100
Figure A.25: Comparison of static ranking algorithms for fibre: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom). . . . 101
Figure A.26: Comparison of feature representations for fibre: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom). . . . 101
Figure A.27: Comparison of static ranking algorithms for salt: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom). . . . 102
Figure A.28: Comparison of feature representations for protein: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom). . . . 102
Figure A.29: Comparison of static ranking algorithms for salt: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom). . . . 103
Figure A.30: Comparison of feature representations for salt: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom). . . . 103
Figure A.31: Comparison of static ranking algorithms for LDL cholesterol: regression performance MAE (mmol/L) for all available features and for 20 best features. . . . 104
Figure A.32: Comparison of feature representations for LDL cholesterol: regression performance MAE (mmol/L) for all available features and for 20 best features. . . . 104

Figure A.33: Comparison of static ranking algorithms for HDL cholesterol: regression performance MAE (mmol/L) for all available features (top) and for 20 best features (bottom). 105
Figure A.34: Comparison of feature representations for LDL cholesterol: regression performance MAE (mmol/L) for all available features (top) and for 20 best features (bottom). 105
Figure A.35: Comparison of static ranking algorithms for glucose: regression performance MAE (mmol/L) for all available features (top) and for 20 best features (bottom). 106
Figure A.36: Comparison of feature representations for glucose: regression performance MAE (mmol/L) for all available features (top) and for 20 best features (bottom). 106

List of Tables

Table 3.2: Estimated vegetable intake (grams/day) based on reported frequency. . . 18
Table 3.3: Diet quality score thresholds based on Slovenian national dietary recommendations. 18
Table 3.1: List of questions and possible answers in ESFFFQ. 19
Table 3.4: Mapping of ESFFFQ questions to feature names. 20

Table 6.1: Feature inclusion across subsets grouped by Weighted Correlation and SHAP-based Interaction. 51
Table 6.2: Ranks for classification compared for target fat by the two feature reduction approaches, including average rank per method (rounded to one decimal). 51
Table 6.3: Ranks for regression compared for target fat by the two feature reduction approaches, including average rank per method (rounded to one decimal). 52
Table 6.4: Ranks for regression models compared for target total cholesterol by the two feature reduction approaches, including average rank per method (rounded to one decimal). 53
Table 6.5: Ranking of classification algorithms' performance on subsets of features selected by different methods for target Fat. 59
Table 6.6: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Fat. 60
Table 6.7: Final classification models' performance across all targets for target-dependent feature selection. 61
Table 6.8: Final Regression Rankings for Target-Dependent Feature Selection. . . . 62
Table 6.9: Ranking of algorithms' performance on subsets of features selected by different methods for target Total Cholesterol. 62
Table 6.10: Final Average Rankings for Cholesterol (Total, LDL, HDL) and Glucose. 63
Table 6.11: Timing Comparison for Feature Sets `X` and `X_freq`. 70
Table 6.12: Regression Timing Summary for Feature Sets `X` and `X_freq`. 71
Table 6.13: Timing results in seconds for different targets and methods. 74
Table 6.14: Timing for first feature selection for MPS and MUR methods. 75
Table 6.15: Similarity between static and dynamic question orderings across targets. 76
Table 6.16: Average computation time per feature for different utility functions. . . . 79

Table A.1: Ranks for classification compared for target Sugar by the two feature reduction approaches, including average rank per method (rounded to one decimal). 85
Table A.2: Ranks for regression compared for target Sugar by the two feature reduction approaches, including average rank per method (rounded to one decimal). 86

Table A.3: Ranks for classification compared for target Fibre by the two feature reduction approaches, including average rank per method (rounded to one decimal). . . . 86
Table A.4: Ranks for regression compared for target Fibre by the two feature reduction approaches, including average rank per method (rounded to one decimal). . . . 87
Table A.5: Ranks for classification compared for target Protein by the two feature reduction approaches, including average rank per method (rounded to one decimal). . . . 88
Table A.6: Ranks for regression compared for target Protein by the two feature reduction approaches, including average rank per method (rounded to one decimal). . . . 88
Table A.7: Ranks for classification compared for target Salt by the two feature reduction approaches, including average rank per method (rounded to one decimal). . . . 89
Table A.8: Ranks for regression compared for target Salt by the two feature reduction approaches, including average rank per method (rounded to one decimal). . . . 90
Table A.9: Ranks for regression models compared for target HDL cholesterol by the two feature reduction approaches, including average rank per method (rounded to one decimal). . . . 90
Table A.10: Ranks for regression models compared for target LDL cholesterol by the two feature reduction approaches, including average rank per method (rounded to one decimal). . . . 91
Table A.11: Ranks for regression models compared for target Glucose by the two feature reduction approaches, including average rank per method (rounded to one decimal). . . . 92
Table A.12: Ranking of classification algorithms' performance on subsets of features selected by different methods for target Sugar. . . . 92
Table A.13: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Sugar. . . . 93
Table A.14: Ranking of classification algorithms' performance on subsets of features selected by different methods for target Fibre. . . . 94
Table A.15: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Fibre. . . . 94
Table A.16: Ranking of classification algorithms' performance on subsets of features selected by different methods for target Protein. . . . 95
Table A.17: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Protein. . . . 96
Table A.18: Ranking of classification algorithms' performance on subsets of features selected by different methods for target Salt. . . . 96
Table A.19: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Salt. . . . 97
Table A.20: Ranking of regression algorithms' performance on subsets of features selected by different methods for target LDL cholesterol. . . . 98
Table A.21: Ranking of regression algorithms' performance on subsets of features selected by different methods for target HDL cholesterol. . . . 98
Table A.22: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Glucose. . . . 99

List of Algorithms

Algorithm 4.1: Weighted correlation analysis with multiple thresholds 31
Algorithm 4.2: SHAP-based feature interaction analysis 32

Algorithm 5.1: Expected-utility-based feature ranking for target variable 39
Algorithm 5.2: COnditional probability estimation subroutine for Algorithm 5.1 (predictProbabilities) 40

Abbreviations

AFA ... Active Feature Acquisition
DFA ... Dynamic Feature Acquisition
DQO ... Dynamic Question Ordering
ESFFFQ Extended Short Form Food Frequency Questionnaire
FFQ ... Food Frequency Questionnaire
FR ... Feature Reduction
FS ... Feature Selection
IPS ... International Postgraduate School
JSI ... Jožef Stefan Institute
kNN ... k-Nearest Neighbours
LinR ... Linear Regression
LogR ... Logistic Regression
MAPE... Mean Absolute Percentage Error
MAE ... Mean Absolute Error
MICE... Multiple Imputation by Chained Equations
MPS ... Maximum Prediction Shift
MMTPS Maximum Multi-Target Prediction Shift
MUR ... Maximum Uncertainty Reduction
MTUR... Multi-Target Uncertainty Reduction
MMTUR Maximum Multi-Target Uncertainty Reduction
WMTUR Weighted Multi-Target Uncertainty Reduction
STP ... Single-Target Prediction
MTP ... Multi-Target Prediction
MTR ... Multi-Target Regression
NHANES National Health and Nutrition Examination Survey
QE ... Quasi Entropy
RF ... Random Forest
SHAP... SHapley Additive exPlanations
SVC ... Support Vector Classification
SVR ... Support Vector Regression

Chapter 1

Introduction

1.1 Background

Maintaining a healthy diet is widely recognised as one of the most important factors affecting health. Unhealthy eating patterns substantially contribute to the global burden of non-communicable diseases. The Global Burden of Disease Study 2021 shows that low consumption of fruits, whole grains, and other nutrient-dense foods is a major cause of deaths and disability worldwide [1]. These findings are reinforced by a 2025 population-based study using Global Burden of Disease 2021 estimates, which shows that dietary risks such as low fruit and whole-grain consumption continue to account for a large share of disease burden among adults [2].

World Health Organisation and the US Centres for Disease Control and Prevention, therefore, place nutrition at the centre of their prevention strategies, emphasising that well-balanced dietary patterns can prevent or delay conditions such as type 2 diabetes and cardiovascular disease [3], [4]. Similar to these public health reports, user interest in following a healthy diet has increased. Nutrition remains among the most common health-related search topics online [5], and diet-tracking applications have been widely downloaded [6], highlighting the relevance of dietary assessment tools.

Effective lifestyle interventions rely on timely and accurate feedback on current behaviour. In behaviour-change research, dietary self-monitoring is a key part of weight-management strategies, and studies show that people who record their food intake more regularly tend to achieve greater weight loss and metabolic improvements [7], [8]. Maintaining dietary reporting during an intervention is also associated with better outcomes than periods of low engagement [7]. Similarly, feedback from continuous glucose monitoring has been shown to improve blood glucose control and dietary adherence [8]. Together, these findings underscore the importance of consistent dietary monitoring in the design and evaluation of nutrition interventions.

Food-Frequency Questionnaires (FFQs) meet many dietary monitoring needs because they are inexpensive and widely used in both research and practice. An FFQ asks people how often they consume a list of foods over a given time period, such as per day, week, or month [9]. Their simplicity and ability to capture dietary intake have led to their widespread use over time [10], [11].

As a result, FFQs are an important tool for assessing dietary patterns and calculating diet-quality scores [12], [13]. However, especially when questionnaires are repeated for self-monitoring, long FFQs can place a high burden on respondents, leading to lower engagement and more missing data [12]. These challenges have driven interest in automated methods that shorten or adapt FFQs while maintaining their usefulness.

From a computer science perspective, using FFQs for dietary assessment means infer-

ring information from data that are uncertain and incomplete. Dealing with missing or partially observed data is a common problem in machine learning, and both imputation-based and model-based methods are used to maintain predictive performance when inputs are missing [14]. At the same time, research in survey design and machine learning has explored adaptive question-ordering methods that reduce respondent burden while maximising information gain [15]. More broadly, this setting fits within active feature acquisition, where features are selected sequentially based on their expected usefulness under cost or budget constraints [16]. Together, these ideas place FFQ-based dietary assessment at the intersection of machine learning, optimisation, and decision-making under constraints.

To reduce respondent burden while still supporting personalised dietary feedback, many mobile health apps include digital FFQs [17]. Compared with paper-based questionnaires, digital tools such as the DIGIKOST-FFQ use features like skip logic, portion-size images, and real-time input checks to shorten completion time and reduce missing or implausible responses. Validation studies show that the DIGIKOST-FFQ agrees reasonably well with weighed diet records and standard FFQs, while taking much less time to complete. In addition to improving data collection, digital FFQs enable adaptive questioning, real-time feedback, and integration with other digital data sources, which are not possible with static questionnaires.

From this perspective, designing and administering FFQs can be seen as a problem of sequentially acquiring information under uncertainty. Although this type of problem has been studied in machine learning, it has received relatively little attention in dietary assessment.

1.2 Problem Description

This dissertation addresses the challenge of balancing questionnaire length and assessment accuracy in FFQs. FFQs estimate dietary intake by asking respondents about the frequency of food consumption over a defined period, but achieving sufficient coverage of dietary patterns often requires a large number of items. In practice, FFQs can range from approximately 20 questions to well over 100 questions, as in the NHANES FFQ [18]. While FFQs are easy to deploy in web and mobile applications [10], [11], [19], lengthy questionnaires increase the likelihood of partial completion, which can compromise the accuracy of derived dietary measures.

Dietary assessment methods—including FFQs, food diaries, and 24-hour recalls—produce quantitative estimates that can be aggregated into diet-quality metrics such as the Healthy Eating Index (HEI) [20] or the Mediterranean Diet Score (MDS) [21]. These metrics summarise adherence to dietary guidelines and are associated with a range of health outcomes. Accurately computing such scores requires sufficient information to capture both beneficial and adverse nutritional components (e.g., fruit and vegetable intake, added sugars, saturated fats), which contributes to the length and complexity of FFQs.

With many overlapping food items, FFQs can become unmanageable. Machine-learning methods offer a data-driven approach to reducing questionnaire length. For example, LASSO regression applies an ℓ_1 penalty that shrinks weak predictors to zero, effectively performing feature selection [22], while random-forest feature-importance measures rank questions by their contribution to predictive performance [23]. These approaches are well-suited to high-dimensional, correlated data. However, even shortened FFQs remain vulnerable to missing responses, and the absence of key items can disproportionately affect prediction accuracy. If the predictive value of each question is known—meaning how much it improves a prediction—questions can be ordered so that the most informative ones are asked first. This makes it more likely that essential information is collected even when a

questionnaire is not fully completed.

In this dissertation, two FFQs with distinct roles and prediction targets are used. The ESFFFQ was developed and validated within the WellCo project, a Horizon 2020 research project funded by the European Union that investigated how digital health technologies, including personalised nutrition and lifestyle interventions, can support the prevention and management of chronic diseases through improved self-monitoring and decision support [24]. The ESFFFQ is used to predict nutrient intakes (e.g., fat, sugar, protein, fibre, and salt). The NHANES FFQ [18], based on a 12-month dietary recall of 139 food and beverage items, is used to predict a set of biochemical biomarkers (e.g., fasting glucose and total cholesterol). Although NHANES includes a wide range of dietary and health-related variables, this work focuses on laboratory biomarkers with established links to dietary intake.

1.3 Goals of the Dissertation

The primary goal of this dissertation is to study optimisation strategies for FFQ-based dietary assessment to reduce respondent burden, including the use of machine-learning methods for inference from incomplete questionnaires, FFQ shortening, and information-driven question ordering.

- **Explore and evaluate methods for dietary assessment using partially answered questionnaires:** The goal is to develop and evaluate machine-learning methods for inferring from incomplete FFQ responses. Both single-target prediction (e.g., predicting one outcome at a time) and multi-target prediction (e.g., predicting multiple outcomes simultaneously) are considered, indicating the need to estimate several dietary outcomes within a single assessment. The target variables include estimated nutrient intake (e.g., fat and sugar intake) in the ESFFFQ and nutritional biomarkers (e.g., glucose and cholesterol) in the NHANES questionnaire.

- **Explore and evaluate methods for reducing questionnaire length:** The goal is to investigate and assess techniques for reducing the number of questions in FFQs while maintaining their ability to capture relevant dietary information, including identifying informative question subsets and evaluating trade-offs between questionnaire length and predictive performance.

- **Develop and evaluate static question ordering methods for single-target and multi-target prediction problems:** The goal is to design and assess static question-ordering strategies in which the question order is fixed in advance. These methods aim to prioritise the most informative questions to improve prediction performance when only a subset of questions is answered, in both single-target and multi-target settings.

- **Develop and evaluate a dynamic question ordering method for single-target and multi-target prediction problems:** The goal is to develop and assess a dynamic question-ordering approach that reorders questions in real time based on observed responses and a utility function. The method is applied to both single-target and multi-target prediction tasks. In the multi-target setting, target-specific weights are incorporated to capture varying importance across outcomes, and strategies for adapting these weights during inference are explored to allow the model's focus to evolve as predictions are refined.

1.4 Hypotheses

This research posits three core hypotheses to assess the performance of the proposed machine-learning methods, measured in terms of prediction accuracy, respondent burden reduction, and computational feasibility, for improving the effectiveness of dietary assessment through FFQs:

- **Hypothesis 1:** *Machine learning can (1) shorten FFQs by selecting the most informative questions and (2) deliver predictive accuracy on these reduced questionnaires that equals or exceeds that of standard methods for handling missing data on the same reduced set.* This hypothesis has two parts. First, machine learning should select a small set of questions from full-length FFQs that retain the primary predictive information, thereby reducing the questionnaire length and making it easier for participants. Second, models trained on these reduced FFQs should perform as well as standard methods that address missing data for the same questions, thereby demonstrating the benefits of ML for both question selection and prediction.

- **Hypothesis 2:** *Statistically prioritising questions based on their relevance can improve the accuracy of dietary assessments derived from partially completed FFQs, when the number of answered questions is held constant.* Machine-learning methods can rank FFQ items by predictive importance, such that, for the same number of answered questions, partially completed questionnaires yield more accurate dietary estimates than those with arbitrarily ordered questions.

- **Hypothesis 3:** *Dynamic question ordering, which adapts to previously gathered responses, provides higher predictive accuracy in comparison to static question ordering in FFQs, when the number of answered questions is held constant.* Machine-learning methods can use previously observed responses to select the next most informative question such that, for the same number of answered questions, dynamic ordering yields more accurate dietary predictions than any fixed question sequence under partial questionnaire completion.

Testing these hypotheses will introduce methods for shortening FFQs while preserving predictive accuracy, improve dietary estimates from incomplete data, and enable dynamic question ordering. The resulting evidence will help improve FFQs, making them more usable, robust, and efficient.

Chapter 2

Related Work

2.1 Dietary Assessment

Understanding what people eat is essential in both self-monitoring and nutritional epidemiology. Dietary assessment methods estimate food intake at the individual and population levels, enabling the study of dietary patterns, nutritional status, and links to chronic disease. Reliable assessment is therefore important for developing dietary guidelines, informing public health interventions, and supporting personalised nutrition [25].

From a computer science perspective, many dietary assessment tools face similar challenges: inputs are often noisy or incomplete, respondents are only willing to answer a limited number of questions, and the order in which questions are asked can affect both data quality and prediction accuracy. These constraints suggest that dietary assessment tools can be viewed as algorithmic systems whose design may be improved using machine learning and sequential decision-making methods.

Traditional dietary assessment methods include 24-hour recalls, food diaries, FFQs, and diet history interviews. Each method has unique advantages and limitations, which affect its suitability for specific study designs. The 24-hour recall, for example, involves a structured interview in which participants are asked to recall all foods and beverages consumed in the previous 24 hours. This method is advantageous for short-term dietary assessments and is widely used in large-scale studies due to its relatively low respondent burden and standardised data collection process [26]. However, it requires an expert involvement to accurately capture habitual intake, as it does not account for day-to-day dietary variability [27].

Food diaries, however, require participants to record their dietary intake over a set period of 3 to 7 days. Weighed food records, a specific type of food diary, are considered highly accurate since participants weigh each food item before consumption. Despite their accuracy, food diaries are limited by high respondent burden, potential underreporting due to self-consciousness, and the risk that participants may alter their usual dietary behaviour while recording [25], [28]. Given these challenges, food diaries are more commonly used in smaller studies or clinical settings where high precision is prioritised.

Diet history interviews aim to capture an individual's usual dietary intake over an extended period, often several months or a year, through an in-depth, interviewer-led assessment. This approach typically combines elements of 24-hour recalls, FFQs, and detailed probing about meal patterns, food preparation methods, and portion sizes. Diet history interviews can provide rich contextual information and are less affected by short-term intake variability than single recalls. However, they are time-consuming, require trained interviewers, and are challenging to scale to large populations, which limits their use in extensive epidemiological studies [25], [27].

Among these tools, self-assessment methods are particularly important for capturing long-term dietary intake while reducing participant burden. FFQs are typically semi-quantitative and capture average consumption frequencies of various food items over an extended period, such as a month or year. They are cost-effective and practical for large populations, but they rely heavily on the participant's memory and are subject to recall bias [29]. To enhance their accuracy, FFQs are often validated against biomarkers or more detailed dietary methods; however, they still face limitations in estimating portion sizes and accurately reflecting short-term dietary changes [30].

Recent technological advancements have led to the emergence of digital dietary assessment tools, particularly mobile applications and web-based platforms, that aim to address some of the limitations of traditional methods. Mobile applications have grown increasingly popular for dietary tracking due to their accessibility, ease of use, and ability to provide real-time data entry. Features such as barcode scanning, food databases, and portion-size estimation support more accurate and efficient dietary recording. Furthermore, mobile-based food records have demonstrated moderate to strong correlations with traditional assessment tools, indicating their potential for wide-scale public health use [31].

An innovative area within digital dietary assessment is image-assisted methods, in which users photograph their meals, and algorithms analyse the images to estimate portion sizes and nutrient content. Image-based methods can reduce reliance on memory and automate parts of dietary recording, but they still require user effort and remain technically challenging in free-living conditions. Although machine learning and AI play a growing role in food recognition and portion estimation, accuracy is sensitive to real-world imaging conditions and mixed dishes, and further validation is needed [32]. Systematic reviews note that many image-based approaches lack extensive validation across large, diverse populations, limiting their suitability for routine, longitudinal use [33], [34].

Despite these developments, dietary assessment still faces fundamental challenges. Recall bias and social desirability bias remain common, as people may forget foods or report what they believe they should have eaten. Estimating portion size is also difficult across all methods, particularly those that rely on self-reporting. While image-based approaches help address some of these issues, they are still prone to errors, especially for mixed or composite dishes [30].

Validation remains critical to ensuring the reliability and accuracy of dietary assessment methods. Biomarker validation compares dietary data to biological markers (e.g., blood, urine) that reflect nutrient intake, providing an objective benchmark for assessing method accuracy. Calibration studies further help to adjust for systematic errors in self-reported data, improving the comparability of dietary data across studies.

Overall, dietary assessment methods continue to evolve as digital health technology advances. Emerging tools offer promising solutions to enhance accuracy, reduce respondent burden, and expand accessibility, especially in large-scale studies. However, ongoing research and validation are needed to address challenges such as portion size estimation, cultural diversity in food databases, and technological accessibility in under-resourced settings.

While traditional dietary assessment methods have laid the foundation for modern nutritional epidemiology, Food-Frequency Questionnaires (FFQs) offer a particularly effective balance between practicality and informative value for large-scale and repeated dietary monitoring. Compared with more precise but labour-intensive approaches—such as weighed food records or repeated 24-hour recalls—FFQs impose substantially lower respondent burden and are therefore better suited to web- and mobile-based settings where sustained engagement is essential. This combination of scalability, low cost, and established validation makes FFQs a natural foundation for methodological development in

dietary assessment.

To connect this nutritional motivation to the dissertation's computational focus, it is useful to view FFQ-based assessment as an inference problem under practical constraints. In many real deployments, questionnaires are only partially completed, responses may be missing systematically, and burden constraints limit the amount of information that can be collected. This creates a setting in which questionnaire design choices (e.g., item selection and ordering) directly affect the quality of downstream prediction, and can therefore be treated as optimisation decisions.

Against this background, this dissertation focuses on optimising FFQ-based dietary assessment using machine-learning methods. Rather than replacing established instruments, the aim is to improve how FFQs are administered and analysed by better utilising partial responses, reducing questionnaire length, and prioritising questions to maximise information gain within respondent burden constraints.

2.2 Food Frequency Questionnaires

The FFQs involve a structured list of food items, asking participants to report the frequency and, in some cases, the portion size of each food item consumed over a specified period. FFQs are advantageous in long-term dietary assessments as they are cost-effective, relatively easy to administer, and less burdensome than food diaries or 24-hour recalls. From a computational standpoint, FFQs can also be viewed as structured, low-dimensional questionnaires with correlated features, which makes them suitable candidates for optimisation under partial observability.

2.2.1 About FFQs

FFQs were initially developed as a cost-effective and efficient method to assess habitual dietary intake across large populations, particularly in epidemiological studies investigating diet-disease relationships [25]. They are designed to capture dietary patterns over extended periods—typically weeks, months, or even years—by asking participants to report the frequency and portion sizes of specific food items they have consumed. These data are then used to estimate nutrient intake levels, allowing researchers to analyse associations between diet and health outcomes [35]. The long-term recall of dietary information makes FFQs particularly useful in identifying dietary patterns associated with chronic diseases, including cardiovascular disease, cancer, and diabetes [35].

FFQs vary widely in structure, scope, and detail, with typical variations including semi-quantitative and quantitative formats. Semi-quantitative FFQs focus on frequency of food consumption and approximate portion sizes, providing broad estimates of nutrient intake. Quantitative FFQs seek more precise portion estimates and may ask participants to specify typical portion sizes for each food item in greater detail. The choice between these formats depends on study objectives, available resources, and the required level of precision [27].

One key advantage of FFQs is their practicality for large-scale studies. They are relatively low-cost and impose a lower burden on participants than other dietary assessment tools, such as repeated recalls or weighed food records [25]. However, FFQs also have limitations due to their reliance on self-reported long-term recall, which can lead to recall and social desirability bias [36]. Accurately estimating portion sizes is another persistent challenge, and errors can lead to substantial misestimation of nutrient intake [27]. These limitations motivate both careful FFQ design and computational methods that can improve robustness under missing or noisy responses.

Digital versions of FFQs introduce interactive prompts, reference images, and automated checks that can improve usability and reduce certain types of data-entry error while preserving the simplicity and scalability of the FFQ format [29]. These properties make FFQs a practical foundation for algorithmic optimisation methods that aim to reduce burden and improve inference when full completion cannot be assumed.

Construction and Validation of Food Frequency Questionnaires

FFQs were initially developed as a cost-effective method for assessing habitual dietary intake in extensive epidemiological studies investigating diet–disease associations [25]. An FFQ consists of a predefined list of foods and beverages, each paired with frequency response categories (e.g. times per day, week, or month) and, in semi-quantitative versions, a standard portion size for each item [27], [37]. Participants report how often—and, when required, how much—they consume each item over a reference period, typically ranging from several months to a year, yielding estimates of average intake for foods, nutrients, or dietary patterns [35].

The construction of an FFQ involves selecting food items intended to capture the majority of between-person variation in dietary intake within a target population. Item lists are commonly derived from prior dietary surveys or food records, with foods included or grouped based on their contribution to key nutrients or nutritional components of interest [37]. The resulting item set is often refined through expert review and pilot testing to balance dietary coverage against respondent burden. Consequently, FFQ length is closely tied to study objectives: short screeners (20–30 items) target specific food groups or nutrients, whereas comprehensive FFQs (100–150 items or more) aim to capture total habitual diet [37]. For example, the Harvard/Willett FFQ includes approximately 130 food items with fixed reference portions and frequency-only responses [25], while the NCI-Block FFQ (110 items) incorporates portion-size categories relative to a stated reference [35]. The NHANES Food Propensity Questionnaire, by contrast, omits portion sizes and uses a reduced set of items to complement repeated 24-hour recalls with habitual frequency information [38].

FFQs are typically validated by comparing nutrient or food-group estimates against reference methods such as multiple 24-hour recalls, weighed food records, or—less commonly—recovery biomarkers (e.g. doubly labelled water for energy intake) [39], [40]. Validation studies generally report moderate correlations with reference methods and assess ranking agreement to confirm that FFQs can reliably order individuals by intake [29]. Portion-size estimation aids, including photographs and household measures, are frequently incorporated to reduce systematic error [30].

In digital settings, FFQs have been adapted for web and mobile platforms, incorporating automated prompts, consistency checks, and image-based portion guides to reduce respondent burden and improve data quality [29], [32]. Despite these enhancements, the underlying structure of most FFQs remains fixed: the full set of items is presented in a predefined order, and burden reduction is typically achieved by shortening the questionnaire at the design stage rather than during administration. **As a result, the optimisation focus in most digital FFQs remains largely offline (during design), rather than online (during completion).**

Recent work has therefore begun to explore more formal optimisation approaches to FFQ construction. Standard procedures for creating FFQs are time-consuming and rely heavily on expert judgment to select food items that adequately represent dietary intake in a given population. To address these limitations, Gerdessen et al. [13] proposed a Mixed Integer Linear Programming approach for optimising the selection of food items for FFQs. Their method demonstrated that optimisation techniques can speed up, standardise, and make the item-selection process more transparent. However, the approach still depends on

substantial domain expertise to define optimisation constraints and objectives, limiting its applicability without expert input.

Machine-learning methods offer a complementary way to reduce reliance on manual expertise by learning predictive structure directly from data. Prior work has shown that machine learning can outperform traditional statistical approaches when evaluating dietary patterns and their associations with health outcomes [41]. Machine-learning techniques have also been applied to identify dietary patterns [42], [43] and to estimate nutrient intake from dietary data [44]. However, most existing applications focus on analysing dietary data after collection rather than supporting FFQ construction or administration. In particular, the use of machine learning to identify informative subsets of FFQ items—while explicitly accounting for partial responses and respondent burden—remains relatively underexplored. This research gap underscores the need for methodologies that integrate machine learning into the FFQ design process to improve adaptability and accuracy in dietary assessments. By doing so, our work aims to develop novel approaches to improve the practical use of FFQs across various settings.

Building on these ideas, this dissertation treats the construction and administration of FFQs as an optimisation problem under incomplete data. Rather than defining a single reduced questionnaire in advance, the proposed approach first identifies the most informative questions for prediction and dynamically prioritises them during questionnaire completion. This two-step framework—feature selection followed by adaptive feature acquisition—extends existing FFQ design strategies by explicitly accounting for missing responses and respondent burden at inference time, while retaining the simplicity and validation advantages of the FFQ format.

2.2.2 FFQs in Mobile Applications

Integrating FFQs into mobile applications has transformed dietary assessment by making it more accessible, user-friendly, and interactive. By moving from traditional paper-based FFQs to digital versions, mobile applications enhance data collection accuracy and reliability through real-time data entry, interactive feedback, and greater user engagement [29]. Mobile FFQs have shown promise in improving user compliance by making the reporting process faster and less intrusive, particularly advantageous in large-scale studies. Several studies report that mobile-based FFQs reduce reporting errors, encourage consistent data input, and lower the overall burden of dietary assessment compared to traditional FFQs [31]. This shift to digital platforms also facilitates the integration of real-time feedback mechanisms, which may improve dietary self-awareness and motivate healthier dietary choices among users.

While mobile FFQs offer improved accessibility and user engagement, a standard limitation is the use of fixed question ordering, which mirrors traditional paper-based FFQs. This rigid design—though familiar—may be suboptimal in digital settings. Survey research consistently demonstrates that fixed question sequences can produce *order effects*, including priming, anchoring, or response bias, especially when questions are displayed one at a time on mobile devices—as opposed to grid layouts where responses may be less susceptible to ordering effects [45]. Dense grid presentations of long item lists—common in FFQs—may also increase respondent fatigue and burden. In other domains, adaptive frameworks, such as *Dynamic Question Ordering (DQO)*, have been proposed for online surveys to tailor the question flow based on prior responses, thereby enhancing engagement and completion [15]. Although DQO has not yet been implemented in dietary assessment tools or FFQs, applying such adaptive sequencing in mobile FFQs could substantially improve usability and data quality.

The future of mobile FFQs appears promising as technological advancements in artifi-

cial intelligence (AI) and machine learning (ML) continue to improve the accuracy of food recognition and portion-size estimation. These advancements are expected to make mobile FFQs an increasingly valuable tool for dietary evaluation in both research and clinical settings.

2.2.3 FFQs in Relation to Health

FFQs are instrumental in epidemiological research, enabling researchers to identify dietary patterns linked to health outcomes such as obesity, cardiovascular disease, type 2 diabetes, and cancer [25]. These tools are often validated against biomarkers—such as blood, urine, or tissue measures—to assess intake reliability and support evidence-informed dietary guidelines [35].

Beyond population studies, digital FFQs also enable *self-assessment* and support *behavioural interventions* at the individual level. For example, the *e-12HR* mobile app—which combines daily recalls with periodic FFQ-based feedback—has shown good agreement with traditional FFQs and 4-day food records, and has been validated for estimating habitual intake of key food groups [46]. Similarly, the *Diet-A* app provides users with real-time nutrient feedback and dietary summaries to facilitate self-monitoring. However, some users report challenges related to logging burden and memory lapses [47].

Digital FFQs and self-monitoring tools thus function as more than data-collection instruments; they provide *actionable feedback* and help reinforce healthy eating behaviours. A recent randomised trial using the *e-12HR* feedback system demonstrated improved adherence to the Mediterranean diet among university students, indicating its promise as a short-term intervention tool [48].

As digital health tools evolve, FFQs continue to integrate AI-based analysis and cross-referencing with user health data to generate more precise dietary assessments and individualised insights. These innovations enhance their role in both public health and personal dietary management, helping overcome barriers related to dietary variation, cultural context, and accessibility. By supporting both large-scale research and individualised interventions, modern FFQs contribute to more accurate data collection, *personalised nutrition*, and preventive health strategies.

The remainder of this chapter reviews computational approaches relevant to optimising FFQ-based assessment, focusing on questionnaire personalisation, adaptive ordering, and dynamic feature acquisition under partial completion.

2.3 Personalisation of Questionnaires

The literature reviewed in this section highlights four interrelated challenges that motivate the methodological contributions of this dissertation: (i) predicting dietary outcomes from partially completed questionnaires, (ii) reducing questionnaire length while preserving predictive accuracy, (iii) improving robustness to partial completion through static question ordering, and (iv) further enhancing efficiency and accuracy through dynamic, adaptive question ordering. While existing work addresses some of these challenges individually, they are rarely considered jointly. In particular, prediction under systematic partial completion—and its interaction with question ordering and multi-target inference—remains underexplored. The remainder of this section reviews prior approaches to questionnaire personalisation in light of these challenges and positions the proposed methods as a unified framework that integrates all four.

The need for accurate, respondent-friendly data collection methods has driven growing interest in questionnaire personalisation. Traditionally, questionnaires have been static and

uniform, presenting the same set of questions in a fixed order to all respondents. While this simplifies administration and ensures comparability across respondents, it does not account for differences in respondent engagement, cognitive strategies, or partial completion. As a result, respondent fatigue and satisficing behaviour are common, often leading to reduced data quality. Krosnick [49] describes how respondents adapt their answering behaviour to manage cognitive burden, frequently at the expense of response accuracy.

One line of work has focused on improving questionnaire performance through static design choices, including the prioritisation and ordering of questions. In survey methodology, question order has long been recognised as influencing response quality and completion rates, particularly when respondents do not complete the full questionnaire [50], [51]. In the context of dietary assessment, static question ordering has been explored as a means of maximising the utility of partially completed FFQs by placing the most informative questions earlier in the questionnaire. In prior work, we demonstrated that ranking FFQ items by their predictive relevance can improve dietary assessment outcomes under partial completion, while preserving a fixed questionnaire structure that remains compatible with standard FFQ validation practices [52]. This approach directly addresses the problem of incomplete responses without altering the questionnaire during administration.

Beyond static ordering, more adaptive forms of personalisation have been investigated. In educational and psychological assessment, computerised adaptive testing dynamically selects items based on an individual's estimated ability, reducing test length while maintaining measurement precision [53]. Related ideas have also been studied in machine learning under the framework of dynamic or active feature acquisition, where features are selected sequentially based on their expected contribution to prediction under cost or budget constraints [16]. These approaches introduce adaptivity at inference time, allowing the system to condition future questions on previously observed responses.

Within dietary assessment, however, dynamic question ordering has only recently begun to be explored. Prior work investigated single-target dynamic ordering in FFQs, showing that adaptively selecting questions based on previous responses can improve the prediction of a specific dietary outcome compared with static ordering [54]. While promising, such approaches are typically limited to single prediction targets and do not fully address the broader challenges of multi-target inference or the systematic partial completion that arises in real-world FFQ deployment.

Building on this line of research, the present dissertation studies both static and dynamic question ordering as optimisation problems tailored to FFQ-based dietary assessment. Static ordering is treated as a principled and practical baseline, well-suited to FFQs where fixed structure and comparability across respondents are essential. This is then extended to dynamic ordering strategies that operate explicitly under partial completion and support both single- and multi-target prediction. Together, these approaches aim to improve predictive accuracy and robustness while explicitly accounting for respondent burden and the realities of incomplete questionnaire data.

2.3.1 Prediction from Partially Completed Questionnaires

A fundamental challenge in questionnaire-based assessment is the need to make reliable predictions when questionnaires are only partially completed. While many predictive models implicitly assume complete input data, partial completion is common in practice, particularly in web- and mobile-based questionnaires, where respondents may stop responding before completing the survey. This predicts incomplete responses as a central methodological problem rather than a marginal data-quality issue.

Existing approaches typically address incomplete questionnaires using one of several strategies. A common practice is complete-case analysis, in which only fully completed

questionnaires are retained for analysis. While this simplifies modelling, it leads to substantial data loss and systematic bias when partial completion is not random [55]. Alternatively, missing responses are often handled through post hoc imputation, in which unobserved values are filled in using statistical techniques such as mean imputation or more advanced methods, including Multiple Imputation by Chained Equations (MICE) [56]. These approaches treat missingness as a preprocessing problem and assume that unobserved responses can be reliably reconstructed from observed data.

A third line of work focuses on model-level robustness to missing values, using predictive models that implicitly account for incomplete inputs. Tree-based methods and ensemble models, for example, are often used in dietary and epidemiological studies for their ability to handle missing values without explicit imputation [23]. However, such approaches do not control which questionnaire items are observed and provide no mechanism for prioritising informative questions early in the questionnaire.

While effective in some domains, these strategies are poorly aligned with FFQ-based dietary assessment. Dietary data are characterised by strong correlations, behavioural variability, and recall-dependent responses, assuming that missing answers can be accurately imputed, problematic [40]. Moreover, in FFQs, partial completion is often systematic rather than incidental: respondents tend to answer early questions more frequently than later ones, especially in long questionnaires. Traditional missing-data approaches do not account for this structured form of incompleteness and therefore fail to exploit the information in partially completed questionnaires effectively.

Importantly, most existing methods treat partial completion as an undesirable artefact to be corrected after data collection, rather than as a core constraint to be addressed during questionnaire design and administration. This motivates a shift toward methods that explicitly account for incomplete responses during inference. The remainder of this section therefore reviews approaches that aim to reduce the impact of partial completion through questionnaire reduction, static question ordering, and dynamic question ordering, positioning prediction under incomplete data as a central design objective rather than a secondary concern.

2.3.2 Adaptive Questionnaires

Adaptive questionnaires aim to personalise data collection by adjusting which questions are asked and in what order, based on information gathered during the assessment. The primary motivation for adaptivity is to reduce respondent burden while maintaining sufficient information for inference. Rather than administering a fixed sequence of questions to all respondents, adaptive systems attempt to prioritise questions that are expected to be most informative given the responses observed so far.

The most established form of adaptive questionnaires originates from computerised adaptive testing (CAT) and item response theory (IRT). In these frameworks, the goal is to estimate a single latent trait—such as ability, disease severity, or psychological state—with a specified precision. After each response, the respondent's latent trait estimate is updated, and the next question is selected to maximise information about that same trait [53], [57]. The assessment typically continues until a stopping criterion is reached, such as achieving sufficient certainty in the latent trait estimate. These approaches have proven highly effective in educational testing and clinical assessments, where item banks are carefully calibrated, and the objective is narrowly defined [58].

Beyond CAT, survey methodology has long recognised that question order influences response behaviour, completion rates, and data quality [50], [51]. In this context, adaptivity is often used to improve engagement or mitigate order effects rather than to optimise predictive inference. More recently, Early formalised adaptive questionnaire design as

an optimisation problem, demonstrating that dynamically selecting questions based on expected information gain can improve inference under respondent burden constraints [15]. However, such approaches are typically formulated for single outcomes and do not explicitly address systematic partial completion or multi-target prediction.

Related ideas have also been studied in machine learning under the framework of dynamic or active feature acquisition. In this setting, features are selected sequentially at inference time based on their expected contribution to prediction accuracy, given the features already observed [16]. When applied to questionnaires, this corresponds to selecting the question with the highest expected utility for prediction under partial observability. Information-theoretic approaches quantify this utility using entropy or uncertainty reduction [59], while other methods rely on greedy heuristics or learned acquisition policies [60]. These approaches provide a principled basis for dynamic question ordering, but are most often evaluated in single-target settings and on general-purpose datasets rather than structured questionnaires.

Despite their success in other domains, existing adaptive questionnaire approaches are not directly suitable for FFQ-based dietary assessment. First, most CAT and IRT-based methods are designed to estimate a single latent trait, whereas FFQs are used to predict multiple observable outcomes, such as nutrient intakes or biomarkers. Second, adaptive testing frameworks typically assume that respondents continue answering until a formal stopping criterion is met. At the same time, in FFQ deployment—especially in web- and mobile-based settings—partial completion is typical and often unplanned. Third, many adaptive feature acquisition methods rely on complex models or large datasets to estimate conditional distributions, which is less compatible with the low-dimensional, structured nature of FFQ data and the modest sample sizes typical of dietary studies.

In contrast, the approach taken in this dissertation treats adaptivity as an optimisation problem under incomplete data. Rather than selecting questions to estimate a latent trait, questions are prioritised based on their expected contribution to predicting one or more dietary outcomes, given the answers already provided. This formulation supports both static and dynamic question ordering, remains compatible with existing FFQ structures and validation practices, and explicitly accounts for partial questionnaire completion as a core design constraint. By adapting ideas from dynamic feature acquisition to the specific characteristics of FFQs, the proposed framework provides a practical and interpretable basis for adaptive dietary assessment.

2.3.3 Dynamic Feature Acquisition for Online Questionnaires

Dynamic Feature Acquisition (DFA), also referred to as Active Feature Acquisition (AFA), addresses the problem of selecting features sequentially during inference to maximise predictive performance while limiting the number of acquired features or their associated costs [16]. In questionnaire settings, DFA can be interpreted as deciding which questions to ask next based on previously observed responses. Existing DFA methods have primarily been developed for general structured datasets and can be broadly grouped into several methodological categories.

- **Information-theoretic methods** score candidate features using measures such as conditional entropy or mutual information, selecting the feature expected to reduce uncertainty about the prediction target maximally [59], [61]. These approaches provide a principled and model-agnostic way to quantify feature utility and have been widely used in active feature acquisition. However, they typically assume that joint or conditional feature distributions can be estimated reliably, an assumption that is more realistic in high-dimensional or large-sample settings than in FFQs, where

the number of questions is limited, and dependencies between dietary variables are strong and domain-specific.

- **Greedy or decision-tree–style policies** select features sequentially using heuristic criteria or learned policies conditioned on the current set of observed features [60], [62]. These approaches are computationally efficient and effective in many classification tasks, but they are often designed for single-target prediction and can be sensitive to the training distribution. In the context of FFQs, such methods offer limited interpretability and do not naturally accommodate multi-target prediction or partial questionnaire completion as a core design constraint.

- **Generative surrogate models**, such as ACFlow, estimate the importance of missing features by modelling the joint distribution of features and targets [61]. This reduces the need for task-specific selection rules, but comes at the cost of complex training and high data requirements. Given the relatively low dimensionality and structure of FFQ data, and the small sample sizes common in dietary studies, these models provide limited practical benefits relative to simpler approaches.

Some DFA approaches additionally incorporate *feature cost sensitivity*, explicitly balancing predictive benefit against the cost of acquiring each feature [63]. For example, the FOCUS framework optimises acquisition decisions by trading off accuracy and cost. In FFQs, however, most questions are comparable in acquisition cost, and respondent burden is driven primarily by the total number of questions answered rather than by differential item costs. Consequently, cost-sensitive formulations provide limited additional value in this setting.

More recently, DFA research has increasingly explored neural network–based approaches that learn acquisition policies end-to-end, often using deep policy networks or transformer-style architectures to guide sequential selection decisions [64]. These methods have demonstrated strong performance in large-scale, high-dimensional problems, where complex nonlinear dependencies can be learned from abundant training data.

However, the assumptions underlying such approaches are less compatible with FFQ-based dietary assessment. FFQ data are typically low-dimensional, highly structured, and collected in relatively small to moderate cohorts, limiting the practical advantages of deep learning models that require large datasets for stable training. In addition, health-related assessment tools place strong emphasis on interpretability, robustness, and reproducibility, as adaptive decisions must be transparent and amenable to validation by domain experts. These requirements favour simpler, more interpretable acquisition strategies over highly parameterised neural architectures, particularly in settings where partial questionnaire completion is common and predictions must remain reliable at any stage of data collection.

While existing DFA approaches provide valuable general strategies for sequential feature selection, few have been applied to questionnaire-based dietary assessment, and fewer still explicitly address multi-target prediction and systematic partial completion. In prior work, we proposed a dynamic feature ordering method for FFQs that combined conditional entropy and expected utility to personalise question sequences in a single-target setting [54]. That work demonstrated improved predictive performance compared with static question ordering, but did not consider multiple prediction targets or goal-aware adaptivity. The present dissertation builds directly on this foundation by extending dynamic feature acquisition to multi-target dietary assessment and by introducing mechanisms for dynamically reweighting prediction objectives during questionnaire completion.

In questionnaire-based applications, dynamic feature acquisition naturally gives rise to the problem of dynamic question ordering (DQO), where features correspond to questionnaire items and acquisition decisions determine which question is asked next. While

DFA provides a general framework for sequential feature selection under uncertainty, DQO focuses specifically on interactive survey settings, where responses are collected incrementally, and respondent burden, interpretability, and partial completion are central concerns. In this sense, DQO can be viewed as a domain-specific DFA tailored to the constraints and objectives of questionnaire-based data collection.

2.3.4 Multi-Target Prediction in Dynamic Question Ordering

Most existing dynamic question ordering (DQO) frameworks are formulated and evaluated for single-target prediction, where the objective is to optimise inference for a single outcome variable. In contrast, the setting considered in this dissertation involves multi-target prediction, where a single questionnaire is used to estimate several numeric outcomes simultaneously, such as multiple dietary intake variables or biomarkers. Multi-output learning and multi-target regression are well-studied settings in machine learning, motivated by the fact that targets are often statistically dependent and can be predicted more accurately when modelled jointly rather than independently [65], [66].

Multi-target prediction introduces additional complexity for dynamic question ordering. Rather than selecting questions solely based on their expected utility for a single outcome, the system must account for how each question contributes to multiple, potentially correlated targets, each with different relative importance. A question that is highly informative for one target may provide little benefit for another, making trade-offs between objectives unavoidable during acquisition.

This distinction is particularly important in FFQ-based dietary assessment. Dietary outcomes, such as nutrient intakes or biomarkers, are not only correlated; practical use cases typically require obtaining a useful profile across several outcomes with partial questionnaire completion, rather than optimising a single variable in isolation. However, many existing DQO and DFA methods either optimise a single prediction objective, assume a fixed notion of utility across all targets, or do not treat systematic partial completion as a core design constraint.

In particular, explicit formulations of dynamic question ordering that jointly optimise multiple numeric targets and incorporate differing target priorities remain uncommon in the questionnaire literature, despite being a natural requirement in dietary assessment settings.

To address this gap, we employ multi-output regression models that jointly predict multiple dependent outcomes. By modelling targets together rather than independently, these approaches exploit shared structure across outcomes, allowing individual questions to reduce prediction error for several targets simultaneously. This is especially relevant in dietary assessment, where many intake variables are linked through common food sources and consumption patterns. Established multi-target regression methods explicitly leverage such dependencies, for example, through chaining or stacking strategies [67].

In addition, we introduce a target-weighting mechanism that prioritises different prediction targets during question ordering. Rather than treating all outcomes as equally important, the framework assigns weights that reflect their relative relevance. Weighted objectives are commonly used to reconcile competing goals in multi-task and multi-objective learning [68], and adaptive weighting strategies have been explored to balance objectives of differing difficulty or scale [69], [70]. While our setting differs from typical neural multi-task learning, the underlying motivation is the same: priorities between multiple objectives may need to shift as inference progresses.

By combining multi-output prediction with target weighting, the proposed framework extends dynamic question ordering beyond single-target optimisation and uniform objective functions. This allows DQO strategies to better reflect the inherently multi-dimensional

nature of dietary assessment, where multiple outcomes must be estimated concurrently, and trade-offs between targets are intrinsic to the task. More broadly, this formulation generalises dynamic question ordering to settings in which prediction objectives are both multi-dimensional and prioritised, providing a flexible foundation for adaptive questionnaires under complex inference goals.

Chapter 3

Datasets

Current dietary assessment methods face significant challenges in capturing accurate, comprehensive data due to limitations in traditional questionnaire formats, which often rely on self-reporting and are prone to bias and inaccuracies. Addressing these challenges is crucial for improving nutritional epidemiology and public health outcomes. This work uses two distinct but complementary datasets to explore and evaluate adaptive questionnaire methods in dietary assessment. The first dataset consists of response data collected using the Extended Short-Form Food Frequency Questionnaire (ESFFFQ), a curated dietary questionnaire developed in collaboration with nutrition experts for use in digital self-assessment contexts. The second dataset is derived from the National Health and Nutrition Examination Survey (NHANES), a large-scale, publicly available study widely recognised for its methodological rigour and population representativeness.

While both datasets are based on food frequency questionnaires, they differ in scope, target variables, and complexity. The ESFFFQ response data provide a compact, interpretable dataset that is well-suited for developing and analysing adaptive questionnaire models. In contrast, the NHANES data enable evaluation of generalisability and robustness in a more complex, real-world dietary assessment scenario, where questionnaire responses are linked to laboratory-measured health outcomes.

3.1 ESFFFQ Dataset

3.1.1 Questionnaire Design and Targets

The Extended Short-Form Food Frequency Questionnaire (ESFFFQ) was developed and validated within the WellCo project, a Horizon 2020 research initiative funded by the European Union that investigated how digital health technologies can support the prevention and management of chronic diseases through improved self-monitoring and decision support [24]. The WellCo project addressed multiple aspects of lifestyle behaviour, including nutrition, physical activity, and wellbeing, to integrate these domains into personalised digital health interventions.

Within this broader framework, the ESFFFQ represents the nutrition-specific component. Its purpose is not an exhaustive dietary assessment, but low-burden monitoring of habitual intake relevant for personalised feedback in mobile and web-based applications. It captures dietary intake across eight nutrition goals identified by expert nutritionists: *fruit*, *vegetables*, *fish*, *fat*, *sugar*, *protein*, *fibre*, and *salt*. These goals reflect dietary components with established relevance for chronic disease prevention and are well-suited for repeated self-assessment in digital health settings.

To balance brevity with coverage, the ESFFFQ builds upon the validated Short Form

Food Frequency Questionnaire (SFFFQ) [71], which initially targeted five of these goals (fruit, vegetables, oily fish, fat, and non-milk extrinsic sugar). The SFFFQ consists of 20 items and produces diet-quality scores based on UK dietary recommendations, assigning values of 1-3 to each component.

The ESFFFQ extends the SFFFQ by introducing additional items to capture fibre, salt, and protein intake. These additions were developed in consultation with nutrition experts and adapted to align with Slovenian dietary guidelines. The resulting questionnaire retains a short, checklist-style format while expanding coverage of key dietary dimensions, making it suitable for use in adaptive, low-burden digital assessment tools.

3.1.1.1 Questionnaire Items and Response Format

The ESFFFQ presents respondents with a structured list of food items and asks how frequently each item is consumed. Response options span nine ordered categories, from "rarely or never" to "5+ times per day." To illustrate, if a respondent consumes fruit 2-3 times a week, this corresponds to category 4. The original SFFFQ items are supplemented with ESFFFQ-specific additions. Table 3.1 lists all items and response categories.

3.1.2 Scoring Methodology

Each ESFFFQ item contributes to one or more nutrition goals. Reported frequencies are first converted into estimated intake amounts (grams) using standard portion sizes and reference conversion factors. These estimates are then mapped to diet-quality scores (1 = bad, 2 = medium, 3 = good) according to national dietary recommendations.

For illustration, Table 3.2 shows estimated vegetable intake derived from reported frequencies, while Table 3.3 summarises scoring thresholds for all nutrition goals. This two-stage mapping—from frequency to amount, and from amount to score—supports both quantitative intake estimation and qualitative dietary assessment.

Table 3.2: Estimated vegetable intake (grams/day) based on reported frequency.

	Rarely	<1/wk	1/wk	2–3/wk	4–6/wk	1–2/day	3–4/day	5+/day
Salad	0	4	11.2	28.8	56.8	120	280	480
Vegetables	0	4	11.2	28.8	56.8	120	280	480

Table 3.3: Diet quality score thresholds based on Slovenian national dietary recommendations.

	Diet quality scores		
	1 - Bad	2 - Medium	3 - Good
Fruit	$\leq$ 2 serv/week	$>$ 2 serv/week and $\leq$ 2 serv/day	$>$ 2 serv/day
Vegetables	$\leq$ 2 serv/day	1 – 3 serv/day	$>$ 3 serv/day
(Oily) Fish	No intake	1 serv/week	2 serv/week
Fat	$\geq$ 111 g/day	$<$ 111 g/day and $\geq$ 74 g/day	$<$ 74 g/day
Sugar	$\geq$ 82.5 g/day	$<$ 82.5 g/day and $\geq$ 55 g/day	$<$ 55 g/day
Protein	$<$ 0.8·BM	$\leq$ 0.8·BM and $\geq$ 1.0·BM g/day	$>$ 1.0·BM g/day
Fibre	$\leq$ 25 g/day	$>$ 25 g/day and $\leq$ 30g/day	$>$ 30 g/day
Salt	$>$ 6g/day	/	$\leq$ 6 g/day

Dataset Description and Feature Extraction

The final ESFFFQ dataset includes responses from 1131 participants. Age information was available for a subset of participants and exhibited a skewed distribution toward older

Table 3.1: List of questions and possible answers in ESFFFQ.

How often do you eat or drink:	Rarely or never	Less than once a week	Once a week	2-3 times a week	4-6 times a week	7+ times a week	Once/twice a day	3-4 times a day	5+ times a day
Questions included in SFFFQ									
Fruit (tinned/fresh)	✓	✓	✓	✓	✓		✓	✓	✓
Juice (not cordial or squash)	✓	✓	✓	✓	✓		✓	✓	✓
Salad (not garnish or added to sandwiches)	✓	✓	✓	✓	✓		✓	✓	✓
Vegetables (tinned/frozen/fresh, NOT potatoes)	✓	✓	✓	✓	✓		✓	✓	✓
Chips / Fried potatoes	✓	✓	✓	✓	✓		✓	✓	✓
Beans or pulses (baked beans, chickpeas, dahl)	✓	✓	✓	✓	✓		✓	✓	✓
Fibre-rich breakfast cereal (porridge, muesli)	✓	✓	✓	✓	✓		✓	✓	✓
Wholemeal bread, chapattis	✓	✓	✓	✓	✓		✓	✓	✓
Cheese / Yogurt	✓	✓	✓	✓	✓		✓	✓	✓
Crisps / Savoury snacks	✓	✓	✓	✓	✓		✓	✓	✓
Sweet biscuits, cakes, chocolate, sweets	✓	✓	✓	✓	✓		✓	✓	✓
Ice cream / Cream	✓	✓	✓	✓	✓		✓	✓	✓
Non-alcoholic fizzy drinks (NOT sugar free)	✓	✓	✓	✓	✓		✓	✓	✓
Beef, lamb, pork, ham (steaks, roasts, etc.)	✓	✓	✓	✓	✓	✓			
Chicken, turkey (NOT in batter/breadcrumbs)	✓	✓	✓	✓	✓	✓			
Sausages, bacon, corned beef, meat pies/pasties	✓	✓	✓	✓	✓	✓			
Chicken, turkey (in batter/breadcrumbs)	✓	✓	✓	✓	✓	✓			
White fish in batter or breadcrumbs	✓	✓	✓	✓	✓	✓			
White fish NOT in batter/breadcrumbs	✓	✓	✓	✓	✓	✓			
Questions included in ESFFFQ									
SALTED nuts, peanuts or seeds	✓	✓	✓	✓	✓		✓	✓	✓
UNSALTED nuts, peanuts or seeds	✓	✓	✓	✓	✓		✓	✓	✓
Grains (pasta, rice, couscous, bulgur)	✓	✓	✓	✓	✓		✓	✓	✓
Pre-prepared sauces, gravies, dry soup mixes	✓	✓	✓	✓	✓		✓	✓	✓
Pizza, pasta/noodle dishes with cheese sauce	✓	✓	✓	✓	✓		✓	✓	✓
Bread, buns and other bread pastries (non-sweet)	✓	✓	✓	✓	✓		✓	✓	✓
Potato (mashed, baked, cooked; NOT fried)	✓	✓	✓	✓	✓		✓	✓	✓
Eggs (boiled, fried, scrambled,...)	✓	✓	✓	✓	✓	✓			

individuals, with ages ranging from 18 to 74 years (mean 57.0 years; median 65 years). Demographic variables such as age, age group, sex, and body mass index (BMI) were collected; however, due to substantial missingness and the primary methodological focus of this dissertation, only dietary responses were used as predictive features.

Age and age-group information are available for 842 participants (approximately 74

The focus of this work is on methodological evaluation of adaptive questionnaire strategies rather than population-level inference. Consequently, all experiments are conducted using only dietary variables, ensuring that model behaviour reflects properties of the questionnaire structure and response patterns rather than demographic differences between participants.

The mapping of food items to feature names is shown in Table 3.4.

Table 3.4: Mapping of ESFFFQ questions to feature names.

How often do you eat or drink:	**Feature name**
Questions included in SFFFQ	
Fruit (tinned/fresh)	fruit
Juice (not cordial or squash)	juice
Salad (not garnish or added to sandwiches)	salad
Vegetables (tinned/frozen/fresh, NOT potatoes)	vegetables
Chips / Fried potatoes	chips
Beans or pulses (baked beans, chickpeas, dahl)	beans
Fibre-rich breakfast cereal (porridge, muesli)	fibre
Wholemeal bread, chapattis	wholebread
Cheese / Yogurt	cheese
Crisps / Savoury snacks	crisps
Sweet biscuits, cakes, chocolate, sweets	cakes
Ice cream / Cream	cream
Non-alcoholic fizzy drinks (not sugar free)	pop
Beef, lamb, pork, ham (steaks, roasts, etc.)	redmeat
Chicken, turkey (NOT in batter/breadcrumbs)	whitemeat
Sausages, bacon, corned beef, meat pies/pasties	sausages
Chicken, turkey (in batter/breadcrumbs)	nuggets
White fish in batter or breadcrumbs	fish
White fish NOT in batter/breadcrumbs	batterfish
Questions included in ESFFFQ	
SALTED nuts, peanuts or seeds	nuts
UNSALTED nuts, peanuts or seeds	nuts_salt
Grains (pasta, rice, couscous, bulgur)	grains
Pre-prepared sauces, gravies, dry soup mixes	pre-processed
Pizza, pasta/noodle dishes with cheese sauce	pizza
Bread, buns and other bread pastries (non-sweet)	bread
Potato (mashed, baked, cooked; NOT fried)	potato
Eggs (boiled, fried, scrambled,...)	eggs

- **Categorical responses:** Equidistant integers representing response categories (e.g., 1 = rarely/never to 9 = 5+ times/day).

- **Weekly consumption frequencies:** Converted to continuous values (e.g., 0.14 = once per week).

- **Estimated nutrient intake:** Grams per day/week of specific nutrients, computed using frequency and portion size data.

Target Variable Construction

To support the evaluation of different modelling tasks, the ESFFFQ data were transformed into target variables that reflect both quantitative intake estimation and qualitative dietary assessment. This formulation bridges questionnaire design with the computational methods introduced later in the dissertation.

Specifically, two complementary types of prediction targets are defined:

- **Amounts:** Continuous intake estimates (grams/day), corresponding to regression tasks.
- **Scores:** Discrete diet-quality scores on a three-level scale (1–3), corresponding to classification tasks.

Both target types are derived from the same underlying questionnaire responses using the conversion and scoring procedures described in the previous sections. All eight nutrition goals are represented in both forms, allowing predictive methods to be evaluated across tasks that differ in output structure and complexity while relying on identical input data.

3.2 NHANES Dataset

3.2.1 Overview and Motivation

The National Health and Nutrition Examination Survey (NHANES) is one of the most comprehensive and widely used sources of publicly available health and nutrition data. Conducted by the Centres for Disease Control and Prevention (CDC), NHANES combines interviews, physical examinations, and laboratory tests to assess the health and nutritional status of the U.S. population. It is considered a gold standard for epidemiological and public health research due to its rigorous methodology, large sample size, and representativeness.

In this dissertation, NHANES serves as a benchmark dataset for evaluating adaptive dietary assessment models in a large-scale, real-world context. Unlike ESFFFQ targets, which are derived directly from questionnaire responses, NHANES enables prediction of laboratory-measured biomarkers that reflect downstream physiological processes influenced—but not fully determined—by diet.

3.2.2 NHANES Dietary Questionnaires

The National Health and Nutrition Examination Survey (NHANES) collects dietary intake data using several complementary instruments. In this dissertation, analyses are based exclusively on the NHANES food frequency questionnaire (FFQ), formally referred to within NHANES as the *Food Propensity Questionnaire (FPQ)*. For consistency with earlier chapters, this instrument is referred to as an FFQ throughout the dissertation.

The NHANES FFQ is a frequency-based dietary assessment instrument designed to capture habitual food and beverage consumption over the previous 12 months. It includes 139 food and beverage items, each assessed using a frequency scale ranging from never" to 6 or more times per day.'' Unlike some semi-quantitative FFQs, the NHANES FFQ does not collect portion size information and focuses solely on intake frequency. NHANES also administers repeated 24-hour dietary recalls to collect detailed short-term intake data; however, these data are *not* used in the analyses presented in this dissertation. The exclusive focus on the FFQ is intentional, as the methodological contributions of this

work concern inference from partially completed frequency-based questionnaires and the optimisation of question ordering within FFQ-style instruments.

In addition to dietary instruments, NHANES includes a wide range of health- and lifestyle-related questionnaires that provide contextual information about the study population. These include:

- **Demographics and Socioeconomic Status (DEMO):** Age, sex, race/ethnicity, education, income, and related background variables.
- **Physical Activity Questionnaire (PAQ):** Measures frequency and intensity of physical activity.
- **Medical Conditions and Prescription Use**: Information on chronic diseases, medications, and dietary supplements.
- **Dietary Supplement Questionnaire (DSQ)**: Captures intake not reflected in food-based questionnaires.
- **Food Security Questionnaire**: Assesses household access to sufficient and nutritious food.

While these additional questionnaires are not directly used for feature construction in this work, they provide essential contextual information and highlight the complexity of the NHANES dataset, motivating its use as a benchmark for evaluating FFQ-based prediction methods in a heterogeneous population.

3.2.3 Feature Extraction

The full FFQ dataset contains over 10,000 entries. Because many of the laboratory target variables were missing for a large portion of participants, we applied filtering to retain only entries with complete target information. This was necessary for developing multi-target prediction algorithms, which require all target variables to be available for every instance. After filtering, the final dataset included 2086 individuals and 200 features.

We extracted two types of features from the NHANES dietary FFQ:

- **Numerical categorical features**: Coded as ordered integers representing discrete frequency categories for each food item.
- **Estimated consumption frequencies**: Converted to continuous weekly frequencies using the `Diet*Calc` analysis software developed by the National Cancer Institute (NCI), which standardises frequency interpretation across FFQ items [72].

Each food item was processed into a standardised numerical representation suitable for input into machine learning models. As with the ESFFFQ, both discrete and continuous forms were retained to allow flexible feature encoding depending on the model.

3.2.4 Target Variables: Nutrition-Related Biomarkers

In addition to dietary intake data, NHANES includes a wide range of laboratory measurements that reflect physiological status. In this dissertation, selected nutrition-related laboratory biomarkers are used as *prediction targets* to examine the extent to which patterns captured by FFQ responses are informative of downstream biological measures. Significantly, these biomarkers are influenced by multiple factors beyond diet, including genetics, physical activity, health status, and environmental exposures. As such, the goal is not to

assume that all biomarkers are entirely predictable from FFQ data, but rather to evaluate whether meaningful associations can be learned under realistic conditions of partial and noisy dietary information.

The biomarkers considered were selected based on prior evidence linking them, at least partially, to dietary intake. They span metabolic, hematological, and micronutrient-related domains and represent outcomes commonly studied in nutritional epidemiology:

- **Cholesterol, LDL, HDL, and triglycerides**: Blood lipid profiles are known to be influenced by dietary composition, particularly intake of saturated fats, simple carbohydrates, fibre, and overall dietary patterns. Diets high in processed foods and added sugars and low in fibre are associated with adverse lipid profiles [73]. While non-dietary factors such as physical activity and energy balance also play a role, these markers provide a well-established link between diet and cardiometabolic risk [74].

- **Plasma fasting glucose**: Fasting glucose levels are associated with carbohydrate quality and quantity, particularly intake of refined sugars and glycaemic load. Dietary patterns high in simple carbohydrates are linked to impaired glucose regulation and increased risk of type 2 diabetes [75]. At the same time, glucose homeostasis is influenced by physical activity and metabolic health, making this marker both challenging and informative.

- **White blood cell count (WBC) and lymphocytes**: These hematological markers reflect immune and inflammatory status and have been associated with overall diet quality. Diets rich in fruits, vegetables, and adequate protein intake provide micronutrients and antioxidants that support immune function [76], [77]. Dietary influences are indirect but relevant in population-level analyses.

- **Hemoglobin and ferritin**: Hemoglobin concentration and ferritin levels are closely related to iron status. Dietary iron intake—particularly heme iron from meat—is a key determinant of iron stores, and insufficient intake is a common cause of iron-deficiency anaemia [78], [79].

- **Serum iron and folate**: Circulating iron and folate levels reflect recent intake and longer-term dietary patterns. Folate is primarily obtained from leafy green vegetables and fortified foods and is essential for erythropoiesis and cellular metabolism [80].

- **Vitamin D**: Vitamin D status is influenced by both dietary intake (e.g. fortified dairy products and fatty fish) and non-dietary factors such as sun exposure. Although diet alone does not fully determine vitamin D levels, FFQ data may still capture a partial signal related to intake patterns [81].

- **Vitamins A and E**: These fat-soluble vitamins depend on both dietary sources and fat intake for absorption. Foods such as leafy vegetables, dairy products, and liver contribute to their intake, and dietary fat facilitates their bioavailability [82].

- **Phytoestrogens**: Phytoestrogens are plant-derived compounds found primarily in soy-based foods and legumes. Their circulating levels are strongly diet-dependent and provide a clear example of a biomarker with a direct link to specific food groups [83].

- **Nitrates**: Nitrate exposure can be influenced by dietary intake, particularly consumption of processed meats and certain vegetables. High intake of processed meats

has been associated with increased exposure to nitrate and nitrite and related health risks [84].

- **Metals in urine**: Urinary concentrations of metals such as cadmium, mercury, and lead can reflect dietary exposure through foods like fish and produce, although environmental and occupational factors also contribute substantially [85].
- **Homocysteine**: Homocysteine levels are influenced by protein intake and B-vitamin status, particularly folate and vitamin B_{12}. Elevated levels are associated with cardiovascular risk and may partially reflect dietary patterns [86].

Taken together, these biomarkers provide a diverse set of targets for evaluating the extent to which FFQ-derived dietary patterns can support the prediction of biologically meaningful outcomes. Their inclusion allows assessment of both relatively direct diet–biomarker relationships and more complex cases where multiple interacting factors mediate dietary effects.

3.2.5 Dataset Suitability

The NHANES dataset complements the ESFFFQ in several important ways that make it well-suited for evaluating the methods proposed in this dissertation:

- It provides a substantially larger and more heterogeneous sample than the ESFFFQ, enabling evaluation of model behaviour across a broader range of dietary patterns and participant characteristics.
- It supports prediction tasks with fundamentally different target types. In the ESFFFQ dataset, targets correspond to estimated nutrient intakes derived directly from questionnaire responses. In contrast, in NHANES, the targets are laboratory-measured biomarkers that reflect downstream physiological processes influenced by diet.
- It enables robustness and generalisability analyses by testing whether methods developed for predicting nutrient intake from a compact FFQ remain effective when applied to a more complex FFQ-style instrument and to biologically measured outcomes.

Taken together, the two datasets form a complementary experimental setup. The ESFFFQ represents a controlled setting in which questionnaire responses are mapped to derived dietary intake estimates, allowing methodological development under relatively direct diet–target relationships. In contrast, NHANES introduces a more challenging scenario in which FFQ responses are used to predict laboratory biomarkers that are only partially determined by diet. This distinction allows the proposed methods to be evaluated across different levels of target complexity, from direct dietary estimates to indirect, biologically mediated outcomes, providing a more comprehensive assessment of their applicability.

Data Availability

The code used for data preprocessing, modelling, and analysis in this dissertation will be made publicly available via a GitHub repository upon publication of the thesis. A placeholder for the repository is provided below and will be updated with the final link before submission:

`https://github.com/ninarescic/adaptive-ffq`

The raw NHANES data used in this work are publicly available from the U.S. Centres for Disease Control and Prevention (CDC) through the NHANES website. To support reproducibility, the repository will additionally include the processed NHANES datasets and derived features used in this dissertation, enabling direct replication of the analyses without requiring reprocessing of the raw data.

Processed ESFFFQ data used in this work will also be shared through the same repository. These data are not directly retrievable from external sources and are provided to facilitate transparency and reproducibility of the proposed methods.

Together, the shared code and processed datasets are intended to enable full replication of the experimental pipeline and to support future research on adaptive questionnaires and dietary assessment.

Chapter 4

Dietary Assessment: Predictive Modelling and Question Reduction

4.1 Prediction in Questionnaires using Statistical and Machine Learning Methods

Food Frequency Questionnaires (FFQs) are frequently only partially completed in real-world settings, leading to missing responses that introduce *measurement error* and uncertainty into dietary assessment. When unanswered questions are treated naively, these errors can propagate to downstream analyses, affecting the reliability of estimated dietary intake and derived diet quality scores.

In the context of this dissertation, the prediction task is defined as follows: given a partially completed FFQ, the goal is to predict dietary quality scores that would otherwise be computed deterministically if all questionnaire responses were available. This framing allows dietary assessment to be treated as a supervised learning problem, where observed responses serve as input features and fully computed quality scores act as targets.

To address this problem, we investigate both statistical imputation techniques and machine learning–based prediction models. Statistical methods aim to reconstruct missing responses by exploiting correlations within the questionnaire data, thereby approximating complete FFQs. In contrast, machine learning models directly predict dietary quality outcomes from incomplete response patterns, learning mappings that are robust to missingness and partial information.

As a baseline, two widely used statistical approaches—zero imputation and Multiple Imputation by Chained Equations (MICE)—are examined for handling missing questionnaire data by generating plausible estimates for absent responses. These methods preserve the dataset structure and enable downstream prediction, while providing a reference point for evaluating the added value of machine–learning–based models.

4.1.1 Statistical Methods for Handling Missing Data

Missing or incomplete responses are common in Food Frequency Questionnaires (FFQs), particularly in self-administered and digital settings. In this work, missing data are treated as a source of *prediction error* rather than bias per se, as unanswered questions lead to incomplete information about dietary intake. A standard way to address this issue is to estimate plausible values for missing responses before performing dietary assessment or prediction.

Two widely used statistical approaches for handling missing questionnaire data are *zero imputation* and *Multiple Imputation by Chained Equations (MICE)*. These methods aim

to reconstruct incomplete FFQs by filling in missing responses, effectively approximating the diet quality scores that would be obtained if all questions had been answered. In this dissertation, they serve as baseline techniques against which more flexible machine learning–based prediction methods are compared.

4.1.1.1 Zero Imputation

Zero imputation replaces missing values with zeros, assuming missing values indicate no consumption or intake. It is unrelated to zero being a natural number, which I firmly believe. This simple and computationally efficient approach makes it attractive for handling missing dietary intake data in FFQs. However, it is reasonable to expect that zero imputation will underestimate true intake, particularly when the missing data may reflect non-response rather than zero consumption. As a result, zero imputation may lead to an underestimation of nutrient intake, resulting in incorrect diet quality scores. For example, individuals who skipped questions might have consumed those foods, but the zero imputation would falsely suggest they did not, ultimately skewing the dietary assessment. This underestimation effect is particularly problematic when assessing key nutritional components, such as fat or sugar intake, essential for health-related outcomes [87].

4.1.1.2 Multiple Imputation by Chained Equations (MICE)

Multiple Imputation by Chained Equations (MICE) is a model-based approach for handling missing data that explicitly accounts for uncertainty in the imputation process. Instead of producing a single completed dataset, MICE generates multiple plausible versions of the dataset by replacing missing values with predictions drawn from statistical models fitted to the observed data [55], [56].

The method proceeds iteratively by modelling each variable with missing values as a function of the remaining variables. For a given feature, missing entries are imputed using an appropriate regression model (e.g., linear or logistic regression), conditioned on the available data. Once imputed, these values are used as inputs in subsequent modelling. This process cycles through all variables with missing data multiple times until the imputations stabilise.

The outcome is a set of fully imputed datasets, each representing a plausible completion of the original questionnaire. Analyses are performed separately on each dataset and then combined using standard pooling rules, allowing downstream estimates to reflect both data variability and uncertainty due to missingness [55].

4.1.1.3 Comparison and Limitations

While MICE offers a statistically more robust method for handling missing data than zero imputation, both methods have limitations. Zero imputation, as mentioned, tends to underestimate intake and can lead to erroneous conclusions about dietary quality. MICE, on the other hand, requires more computational resources and may be sensitive to model specification. Additionally, both methods assume that missing data are missing at random (MAR), a condition that may not always hold in dietary assessments. Therefore, carefully considering the mechanism underlying missing data is essential when selecting the most appropriate imputation method.

4.1.2 Machine Learning Methods

In this study, we built eight classification models and eight regression models, ranging from simple methods to more complex techniques. These models included basic algorithms

such as *Linear/Logistic Regression*, *Decision Tree*, and *K-Nearest Neighbours*, as well as more advanced models like *Support Vector Machine*, *Random Forest*, *Gradient Boosting Classifier/Regressor*, and the *Voting Classifier/Regressor*, which combines all the models above.

4.1.2.1 Linear Models

Linear models such as *Linear Regression* and *Logistic Regression* serve as foundational algorithms in machine learning. These models are simple and interpretable, providing a baseline for comparison with more complex techniques. While they are less powerful in handling non-linear relationships, they remain highly effective for problems where the relationship between input variables and the target variable is approximately linear.

4.1.2.2 Ensemble Models

Ensemble methods, such as *Random Forest*, *Gradient Boosting*, and the *Voting Classifier/Regressor*, combine multiple weak models to form a stronger predictive model. These methods typically provide higher accuracy than individual models by reducing the risk of overfitting and bias. In this study, the *Voting Classifier/Regressor* proved particularly useful, leveraging the strengths of multiple models to make more robust predictions by combining results from simpler classifiers, such as decision trees, and more complex ones, such as support vector machines.

All machine learning models were implemented using the `scikit-learn` library [88] with default hyperparameter settings. Although model performance can often be improved through hyperparameter tuning—such as adjusting the number of trees, tree depth, regularisation strength, or learning rates—we deliberately avoided tuning to ensure a consistent and fair comparison across models and experimental conditions.

Across the different experimental setups considered in this work, the primary objective was not to optimise individual model performance, but to analyse how predictive performance changes as the number of available FFQ questions is reduced. In particular, we aimed to identify subsets of questions for which a broad range of models achieve comparable or better performance than using the full questionnaire. Demonstrating that machine learning models retain strong predictive accuracy even with substantial question reduction and outperform simple baselines in these settings is therefore a central outcome evaluated throughout this chapter.

4.2 Optimising the Efficiency of FFQs through Question Reduction

This section explores methodologies to enhance the efficiency of Food Frequency Questionnaires (FFQs) by reducing their dimensionality while maintaining data quality. We investigate two primary approaches: target-independent methods that rely on the intrinsic properties of questionnaire items and predictive model-based feature selection techniques. Methods such as weighted correlation analysis [37], [89] and SHAP interaction values [90] are discussed, alongside machine learning-based approaches including filter, embedded, wrapper methods, and genetic algorithms. A comparison of these strategies concludes the section, providing insights into their suitability for FFQ optimisation.

4.2.1 Target-Independent Question Reduction Methods

The need for dimensionality reduction in Food Frequency Questionnaires (FFQs) arises from their extensive nature, which can reduce user engagement and make them cumbersome for large-scale use. This section explores methods that streamline FFQs without requiring predictive models, emphasising data-driven approaches such as weighted correlation [37], [89] and SHAP for feature interactions [90].

4.2.1.1 Weighted Correlation for Dimensionality Reduction

Weighted Correlation Analysis enhances traditional correlation-based feature selection by introducing weights to prioritise features based on their importance or variability. This approach ensures that critical dietary questions in FFQs are preserved while reducing redundancy.

The following pipeline describes the process of weighted correlation-based feature selection from Algorithm 4.1:

1. **Assign Weights to Features:** Each feature (f_i) is assigned a weight based on its variability. Features with greater variability across the dataset are considered more informative. The weight is calculated as: [$\text{Weight}(f_i) = \text{Var}(f_i)$] Features with higher weights are prioritised during feature selection.

2. **Calculate Pearson Correlation Coefficients:** Compute the Pearson correlation coefficient r between all pairs of features to identify highly correlated pairs. A high value (e.g., $r > 0.8$) indicates strong linear dependence between two features.

3. **Threshold-Based Feature Selection:** Iteratively apply correlation thresholds $TH_1, TH_2, \ldots, TH_n$, where thresholds are initially derived from the empirical distribution of correlation coefficients and then manually adjusted to obtain feature subsets of approximately linearly increasing size. Threshold selection follows a data-driven but supervised procedure:

    - *Median and quantiles:* Initial thresholds were based on summary statistics of the correlation distribution, with the median used as the starting threshold TH_1. Subsequent thresholds were chosen from higher quantiles of the distribution.
    - *Manual refinement:* Threshold values were subsequently fine-tuned by inspection to ensure that the resulting subsets $P_0, P_1, \ldots, P_n$ increased in size in an approximately linear manner, facilitating systematic comparison across different levels of question reduction.

4. **Remove Redundant Features Using Weights:** For each pair of features with $r > TH_i$, compare their weights and retain the feature with the higher weight:

    - Drop the feature with the lower weight to minimise redundancy.
    - Dynamically update the retained feature set, ensuring the remaining features maintain variability and relevance.

The weighted correlation method prioritises critical features in FFQs, ensuring that the final subset retains high variability and relevance. By integrating feature weights, the approach aligns statistical reduction techniques with domain-specific dietary priorities.

Algorithm 4.1: Weighted correlation analysis with multiple thresholds

Input: Scaled dataset X_{scale}, initial feature variances var, ordered correlation thresholds $\mathcal{TH}$

Output: Retained features for each threshold weighted_corr_feats

Note: $\mathcal{TH} = \{TH_1, \ldots, TH_n\}$ is ordered such that $TH_1 > TH_2 > \cdots > TH_n$, producing nested feature sets with increasing size:
$P(TH_1) \subseteq P(TH_2) \subseteq \cdots \subseteq P(TH_n)$.
Step 1: Initialize variables
var $\leftarrow$ Var(X_{scale})
weighted_corr_feats $\leftarrow$ {} {Stores retained features per threshold}
retained_features $\leftarrow$ Columns(X_{scale})
Step 2: Weighted correlation filtering for each threshold
for $TH \in \mathcal{TH}$ **do**
  to_remove $\leftarrow \emptyset$ {Features marked for removal at this threshold}
  for each unordered pair (f_i, f_j) with $f_i \neq f_j$ and $f_i, f_j \in$ retained_features **do**
    $r \leftarrow \text{corr}(X_{\text{scale}}[f_i], X_{\text{scale}}[f_j])$ {Pearson correlation}
    if $|r| > TH$ **then**
      if $\text{var}[f_i] \geq \text{var}[f_j]$ **then**
        to_remove $\leftarrow$ to_remove $\cup \{f_j\}$
      else
        to_remove $\leftarrow$ to_remove $\cup \{f_i\}$
      end if
    end if
  end for
  retained_features $\leftarrow$ retained_features $\setminus$ to_remove
  weighted_corr_feats[TH] $\leftarrow$ retained_features
  var $\leftarrow$ Var(X_{scale}[retained_features])
end for
return weighted_corr_feats

4.2.1.2 SHAP for Feature Interactions

SHAP (SHapley Additive exPlanations) is most commonly used as a post-hoc interpretability tool for analysing the contribution of individual features to a trained predictive model [90]. In this work, SHAP is instead used as a mechanism for *feature reduction*, aiming to identify a compact subset of questionnaire items that capture the dominant structure and variability in the data. Crucially, this is done without requiring true prediction targets, which makes the approach suitable for target-independent questionnaire optimisation.

To enable SHAP use in the absence of true outcomes, we introduce a proxy learning task that captures broad patterns in the questionnaire responses. The central assumption is that features which consistently contribute to separating respondents into distinct groups are informative and non-redundant, and can therefore be prioritised when reducing the number of questions. Rather than modelling a specific dietary outcome, the method focuses on explaining variation within the response space itself.

A SHAP-based framework is employed to quantify feature contributions in this target-independent setting. Although a predictive model is used internally, it serves only as a proxy to induce structure in the data; the resulting SHAP values are interpreted as measures of feature relevance rather than as explanations of a real-world prediction task. The procedure is designed for unsupervised use and provides a systematic way to uncover informative features and redundancies among questionnaire items. The full pipeline is

described below and formalised in Algorithm 4.2.

1. **Pseudo-target generation.** The feature reduction pipeline begins by constructing pseudo-target labels using clustering. The KMeans algorithm is applied to the scaled dataset (X_{scale}) to partition respondents into n clusters. Each respondent is assigned a cluster label, which serves as a pseudo-target capturing coarse dietary response patterns.

2. **Training a proxy model.** A Random Forest classifier is trained to predict the pseudo-targets from the questionnaire features. The purpose of this model is not accurate classification per se, but to learn decision structures that distinguish between different response profiles. Random Forest is chosen for its robustness and its ability to model non-linear relationships between questionnaire items.

3. **SHAP-based feature contribution analysis.** SHAP values are computed for the trained proxy model using the TreeExplainer. These values quantify the extent to which each feature contributes to assigning respondents to different clusters. Absolute SHAP values are averaged across individuals and, in the multi-class setting, aggregated across clusters to produce a global importance score for each feature.

4. **Feature ranking and selection.** Features are ranked according to their aggregated SHAP importance scores. The top k features are then selected, yielding a reduced questionnaire that preserves the most informative items while removing those that contribute little to explaining the data's structure.

Algorithm 4.2: SHAP-based feature interaction analysis

Input: Scaled dataset X_{scale}, number of clusters n, number of trees n_{trees}, top-k value
Output: Selected top-k features selected_features

Step 1: Pseudo-target generation
Initialize KMeans with n clusters
$pseudo_target \leftarrow$ KMeans.fit_predict(X_{scale})
Step 2: Train proxy model
Initialize Random Forest classifier with n_{trees} trees
$model \leftarrow$ RandomForestClassifier.fit($X_{\text{scale}}, pseudo_target$)
Step 3: Compute SHAP values
$shap_values \leftarrow$ TreeExplainer($model$).shap_values(X_{scale})
if $shap_values$ is a list **then**
    $shap_values_combined \leftarrow \text{mean}(|shap_values|)$ {Mean absolute SHAP values aggregated column-wise across classes}
else
    $shap_values_combined \leftarrow |shap_values|$
end if
$feature_importance \leftarrow \text{mean}(|shap_values_combined|)$ {Mean absolute SHAP values aggregated row-wise across samples}
Step 4: Feature ranking and selection
Sort features by $feature_importance$ in descending order
$selected_features \leftarrow$ Top-k ranked features
return $selected_features$

4.2.2 Target-Dependent Question Reduction Methods

Feature selection is the process of selecting variables that provide the most accurate predictions relative to the true outcome. In dietary studies, it is essential to ask as few questions as possible while still obtaining sufficient information to assess individuals' nutritional intake. The ESFFFQ [24] was developed to compute diet quality scores for eight dietary components: fruit, vegetables, fish, salt, sugar, fat, fibre, and protein. These scores are derived from estimated nutrient intakes. They are used to provide targeted dietary feedback to users, such as recommending reduced sugar intake when a low-quality score is detected.

Because ESFFFQ items contribute differently to each dietary score, some questions are redundant for specific targets. For three scores (fruit, vegetables, and fish), the relevant information is captured by one or two questions, making feature reduction trivial. For the remaining five scores—fat, sugar, fibre, protein, and salt—multiple food items jointly influence the outcome, introducing redundancy and overlap. We therefore formulate question reduction for the ESFFFQ as a set of single-target feature selection problems, where the goal is to identify minimal subsets of questions that preserve predictive performance for each dietary score. This framing allows machine learning–based feature selection methods to be applied in a way that directly reflects the structure and intended use of the ESFFFQ.

We explored several feature selection techniques:

4.2.2.1 Filter Methods

- *Feature importance from machine learning models* — In this approach, features are ranked based on importance scores derived from models trained on the complete feature set, without iterative retraining on reduced subsets. Specifically, we use `feature_importances_` from *Random Forest* models and `coef_` from *Logistic Regression*, as implemented in `sklearn`.

  For Logistic Regression, coefficients represent the contribution of each feature to the linear decision boundary, with larger absolute values indicating greater importance. For Random Forests, feature importance reflects the average decrease in impurity (e.g. Gini) contributed by each feature across all trees, capturing both linear and non-linear effects. Features whose importance exceeded a predefined threshold were retained, while those below the threshold were discarded. All models were trained using default parameters.

4.2.2.2 Embedded Methods

- *Boruta* — Boruta is a feature selection method built around Random Forests that embeds feature evaluation within the model training process. It augments the dataset with permuted versions of each feature (shadow features) and trains a Random Forest on the extended feature set. Features whose importance consistently exceeds that of the best-performing shadow feature are retained as relevant. We applied Boruta using default parameters from the `BorutaPy` package [91].

- *Least Absolute Shrinkage and Selection Operator (LASSO)* — LASSO regression [22] performs feature selection implicitly through L1 regularisation. By penalising the absolute magnitude of coefficients, LASSO drives less informative coefficients exactly to zero, thereby removing the corresponding features during model fitting.

- *SHapley Additive exPlanations (SHAP)* — SHAP values were used to rank feature importance for trained Random Forest and Logistic Regression/Ridge models [90].

Although SHAP is primarily an interpretability framework, in this work, it is employed as a feature selection mechanism by coupling importance ranking with iterative model evaluation. Features are ordered by their average absolute SHAP values, and predictive models are re-trained on progressively smaller subsets of top-ranked features. Selection proceeds until predictive performance no longer improves. Feature stability is assessed via 5-fold cross-validation, and features retained in at least three folds are included in the final subset. In this formulation, SHAP-based selection behaves similarly to a wrapper method, as feature subsets are evaluated based on model performance.

4.2.2.3 Wrapper Methods

- *Recursive Feature Elimination (RFE)* — RFE is an iterative wrapper method that repeatedly trains a predictive model and removes the least important features based on the model's internal importance estimates. In this work, we applied RFE using Logistic Regression and Random Forest as base estimators, implemented in `sklearn`. For Logistic Regression, feature importance is derived from the magnitude of the coefficients, while for Random Forests, it is based on impurity-based importance. At each iteration, a subset of features is removed, and the model is retrained until the desired number of features remains.

4.2.2.4 Genetic Algorithms for Feature Selection

- *Genetic Algorithm (GA)* — Genetic Algorithms [92] are population-based heuristic optimisation methods inspired by evolutionary processes. In our implementation, candidate feature subsets are encoded as binary chromosomes and evolved over multiple generations. Each subset is evaluated by training a predictive model (Random Forest or Logistic Regression/Ridge) using cross-validation, and its fitness is measured using a task-specific scoring metric.

  In contrast to the other feature selection methods explored in this chapter, which rely on default or generic hyperparameter settings, the genetic algorithm requires explicit control of its evolutionary dynamics. Due to the relatively low dimensionality of the dataset and the moderate number of questionnaire items, we used a population size of 20 individuals evolved over 10 generations, with a crossover probability of 0.5 and a mutation probability of 0.2. These parameters control the balance between exploration and exploitation during the search process and were kept fixed across all experiments unless stated otherwise.

  Feature selection is performed across 5-fold cross-validation, and features appearing in more than 50% of the folds are retained in the final subset. While the GA relies on model-based evaluation, its stochastic search strategy distinguishes it from deterministic wrapper methods.

Chapter 5

Optimisation of Question Ordering for Predictive Modelling

5.1 Static Question Ordering

Accurate feature (question) ranking is essential in dietary assessment when FFQs are only partially completed. In this context, each question corresponds to a feature whose importance can be evaluated based on its contribution to predicting dietary outcomes. *Static ranking* determines a fixed, global order of questions before deployment, based on their relevance to the target variable. Presenting more informative questions earlier improves the reliability of prediction under partial completion, because useful information is collected sooner, even if respondents stop before finishing the questionnaire.

This perspective is particularly relevant in mobile and web-based settings, where responses are often provided incrementally and missingness arises naturally from drop-off rather than from an explicit attempt to shorten the questionnaire. A well-ranked sequence prioritises questions that yield the most significant performance gains early on, ensuring that the prediction improves as much as possible from the very first responses. As a result, each additional answered question contributes meaningful new information, leading to the most significant achievable improvement in predictive accuracy at every stage of partial questionnaire completion. While the same ranking can also support deliberate burden reduction by enabling shorter questionnaires, the primary objective here is to improve inference when full completion cannot be assumed.

5.1.1 Static Question Ordering for Single-Target Prediction Problems

For each dietary component targeted by the ESFFFQ and NHANES FFQ, we formulated a single-target prediction problem using machine-learning models. We applied two machine learning algorithms: Logistic/Linear Regression and Random Forest Classifier/Regressor. These models were selected based on prior experiments demonstrating reliable, competitive performance in dietary prediction and compatibility with feature importance extraction. Feature importance for each model was derived from the absolute value of coefficients in Logistic/Linear Regression models and the feature importance attribute in Random Forest models within the `sklearn` library. This ranking guided the sequential addition of questions, allowing us to evaluate how each incrementally contributed to predictive accuracy.

To further enhance our analysis, we experimented with different feature representations: nominal discrete values, frequency-based transformations, and nutrient-based transformations. These representations were selected based on prior experiments indicating that input

encoding significantly influences model performance in dietary prediction tasks. By adding questions in order of their ranked importance, we can progressively evaluate their impact on the model's predictive performance across both classification and regression tasks. This approach enabled us to refine the question ordering, ensuring that the most informative questions appeared early in the sequence, thereby enhancing the practical utility of both FFQs in dietary assessment applications.

5.1.2 Static Question Ordering for Multi-Target Prediction Problems

While single-target prediction provides a focused view on how specific questions relate to individual dietary outcomes, many FFQ items are inherently informative across multiple outcomes. For instance, a question about sweet foods—such as cakes or pastries—can simultaneously influence predictions of both sugar and fat intake, given their high content of both nutrients. Similarly, questions about fruit consumption are predictive of both fibre intake and sugar intake, since fruits are rich in dietary fibre and naturally occurring sugars. Ignoring these shared associations by treating targets independently may result in inefficient question orderings.

Multi-target (or multi-output) prediction addresses this limitation by learning to predict several dietary outcomes simultaneously. This approach is especially suitable for dietary assessment, where many nutritional variables are interrelated and influenced by overlapping dietary behaviours. By capturing the commonalities among targets, multi-target models can exploit underlying data structure to improve both predictive performance and model interpretability [65], [93]. Moreover, joint modelling reduces the risk of redundancy and overfitting that can arise from learning separate models for each target [94].

In practice, this yields more compact and informative FFQs. Questions that predict multiple outcomes are prioritised, enabling accurate multidimensional dietary profiling with fewer items. This efficiency is particularly valuable in mobile or digital health applications, where minimising user burden is crucial. Question sequences generated from multi-target feature rankings ensure that each response contributes value across multiple dietary dimensions, maximising the usefulness of each user interaction.

In our work, we extended static question ordering to the multi-target setting using multi-output regression and classification models. This allowed us to create unified question orderings that reflect global predictive relevance. However, in static question-ordering scenarios—where the same fixed sequence of questions is used for all individuals and targets—it may also be necessary to weight targets differently. Some outcomes (e.g., sugar intake for diabetic users, or salt intake for those with hypertension) may be more clinically relevant or time-sensitive than others. In such cases, target-specific weighting strategies can be applied during feature ranking to emphasise the importance of higher-priority outcomes, balancing efficiency with real-world relevance.

5.2 Dynamic Question Ordering

Dynamic question ordering provides a personalised, adaptive approach to dietary assessment by tailoring the question sequence in real time based on each user's responses. Unlike static ordering, which follows a fixed, pre-ranked sequence, dynamic ordering adapts to user input, prioritising the most informative questions to improve efficiency and reduce burden.

This approach is especially valuable in mobile health applications, where user engagement may vary. By continuously learning from responses, the system refines its understanding of individual dietary patterns, focusing on areas of most significant uncertainty or predictive value to maximise assessment accuracy with fewer questions.

Dynamic question ordering is thus a promising approach to dietary assessment, especially in contexts where users may have variable engagement or dietary behaviours that do not align well with a one-size-fits-all question sequence. By combining real-time adaptation with predictive accuracy, this method enhances both the precision and user experience of dietary assessment tools, making it an essential advancement in personalised health and nutrition monitoring.

5.2.1 Dynamic Question Ordering for Single-Target Prediction Problems

In dynamic feature ranking, the next most informative feature is selected based on the values of previously acquired features, allowing the question sequence to adapt progressively to the individual respondent. In this work, dynamic question ordering is formulated within an active feature acquisition framework [59], combining information-theoretic measures with task-specific utility to guide sequential feature selection.

We begin by quantifying uncertainty using entropy, a fundamental concept from information theory. For a discrete random variable X taking values $x \in \mathcal{X}$ with probabilities $p(x)$, entropy is defined as:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x). \tag{5.1}$$

Entropy is highest when all outcomes are equally likely and lowest when one outcome is highly probable. In the context of questionnaires, entropy captures how uncertain we are about the value of a question before it is answered.

As questionnaire completion progresses, some features are already observed. To account for this, uncertainty is measured conditionally on the set of known feature values. For a candidate feature f, conditional entropy given the currently observed answers $x_{\mathcal{F}_B}$ is defined as:

$$H(f \mid x_{\mathcal{F}_B}) = -\sum_{r \in R(f)} p(f = r \mid x_{\mathcal{F}_B}) \log p(f = r \mid x_{\mathcal{F}_B}). \tag{5.2}$$

This conditional entropy is evaluated at a single questionnaire state and corresponds to the remaining uncertainty about feature f for a specific respondent. From an information-theoretic perspective, features with higher conditional entropy are less predictable given the current information and therefore offer greater potential for uncertainty reduction when acquired.

Minimising conditional entropy, therefore, yields a purely information-theoretic acquisition strategy: features are selected because their values are uncertain given the current information. This corresponds to maximising expected information gain about the feature itself, independent of any downstream prediction task.

Based on this principle, a purely information-theoretic strategy would select the next feature by minimising conditional entropy over the set of available features $\mathcal{F}_\mathcal{A}$:

$$f^* = \arg\min_{f \in \mathcal{F}_\mathcal{A}} H(f \mid x_{\mathcal{F}_B}). \tag{5.3}$$

While this criterion optimally reduces uncertainty about feature values, it is agnostic to the downstream prediction task. In particular, it treats all uncertainty reductions as equally valuable, regardless of whether resolving that uncertainty improves the prediction of the dietary outcome of interest. In FFQs, this distinction is essential: a question may be highly uncertain yet have negligible influence on the prediction target, while another may be relatively predictable but strongly informative for improving the estimate.

To incorporate task relevance into the acquisition decision, we extend the use of conditional entropy within an expected-utility framework. Rather than selecting features solely based on uncertainty reduction, we weight the uncertainty associated with each feature value by a utility term that reflects its expected contribution to predictive performance. This leads to the following expected utility formulation:

$$\mathbb{E}(U_f \mid x_{\mathcal{F}_B}) = -\sum_{r \in R(f)} p(f = r \mid x_{\mathcal{F}_B}) \log p(f = r \mid x_{\mathcal{F}_B}) \cdot U(f \coloneqq r). \tag{5.4}$$

Here, $x_{\mathcal{F}_B}$ denotes the values of the already observed features, and $U(f \coloneqq r)$ is a task-specific utility that quantifies how informative observing the value r of feature f would be for predicting the target variable. The entropy term captures the uncertainty in the feature value given the current information, while the utility term modulates this uncertainty by its expected impact on prediction quality.

This formulation can be interpreted as a utility-weighted extension of conditional entropy. When $U(f \coloneqq r)$ is constant, the expected utility reduces to standard conditional entropy; when utilities vary, the criterion prioritises uncertainty reductions that are most relevant for the prediction task. By minimising $\mathbb{E}(U_f)$, the algorithm selects questions that balance uncertainty reduction with predictive relevance, yielding a personalised and outcome-driven question sequence.

At each step, the next question is selected as the feature with the highest expected utility. The overall procedure is summarised in Algorithm 5.1, with conditional probability estimation handled by Algorithm 5.2.

Algorithm 5.1 produces a ranked sequence of questions for each target and for each respondent instance. It starts by initialising the output container (lines 1–2) and iterating over prediction targets $t \in T$ (line 4). For each target, a precomputed static ranking is used as an initial seed (lines 6–7), ensuring a consistent starting order before personalisation.

For each validation instance $x \in X_{\text{valid}}$ (line 8), missing feature values are first imputed using a KNN imputer fitted on the training data X (line 9). This enables consistent evaluation of predictions during ranking. The set $\mathcal{F}_A$ contains candidate features that have not yet been selected (line 10), while $\mathcal{F}_B$ stores the already ranked features (line 11). The algorithm continues until no candidates remain (line 12).

For each candidate feature $f \in \mathcal{F}_A$ (line 13), the algorithm: (i) estimates the conditional distribution $p(f = r \mid x_{\mathcal{F}_B})$ over possible answers r (line 14); (ii) trains a prediction model using the currently selected features $X[\mathcal{F}_B]$ and computes the current prediction $\hat{y}_{\text{prev}}$ (line 15); (iii) computes the expected utility $\mathbb{E}(U_f)$ by aggregating over all possible answers r (lines 16–17).

The probability term $p(f = r \mid x_{\mathcal{F}_B})$ required in line 14 is obtained using Algorithm 5.2. This routine sets up a surrogate supervised task in which the candidate feature f_{pred} serves as the target. The input data are restricted to the currently known features $X_{\text{part}} \leftarrow X[\mathcal{F}_B]$ (Algorithm 5.2, line 2), and the labels are given by $y_{\text{pred}} \leftarrow X[f_{\text{pred}}]$ (line 3). A Random Forest classifier is trained on $(X_{\text{part}}, y_{\text{pred}})$ (line 5), and `predict_proba` returns a probability vector for the current instance x (line 7). These probabilities are then mapped to the distinct answer categories R and stored in `dict_prob` (lines 8–12), yielding $p(f = r \mid x_{\mathcal{F}_B})$.

In Algorithm 5.1, line 17 combines the predicted probabilities with a utility function. The term

$$p(f = r \mid x_{\mathcal{F}_B}) \cdot \log p(f = r \mid x_{\mathcal{F}_B})$$

is inspired by conditional entropy: it gives higher weight to features whose answers are informative given the current context. Because these probabilities are learned from data

using Algorithm 5.2, rather than computed from an explicit probabilistic model, we refer to this component as *Quasi Entropy (QE)*.

The utility function $U(f \coloneqq r)$ used in line 17 of Algorithm 5.1 specifies how relevant a particular answer value r of feature f is for the prediction target. The formulation shown in the algorithm corresponds to the *Maximum Prediction Shift (MPS)* utility, which is formally defined and discussed in the following subsection. Other utility functions considered in this work are also defined there and can be substituted into line 17 without changing the ranking algorithm's structure.

After computing $\mathbb{E}(U_f)$ for all candidate features (lines 13–18), the algorithm selects the feature with the minimum expected utility value (line 19), adds it to the ranked set $\mathcal{F}_B$ (line 20), and removes it from $\mathcal{F}_A$ (line 21). The final ranked list, for instance x, is saved (line 23). Repeating this process for all instances yields personalised question orders, while repeating it across targets produces one ranked sequence per target (lines 24–26).

Algorithm 5.1: Expected-utility-based feature ranking for target variable

Input: Training data X, validation data X_{valid}, targets T, target values Y, precomputed static rankings static_ranks

Output: Ranked feature sequences per instance and target best_ranks_target

1: **Step 1: Initialize output container**
2: best_ranks_target $\leftarrow$ {} {Stores ranked lists per target}
3: **Step 2: Rank features for each target**
4: **for** each target $t \in T$ **do**
5: $y_t \leftarrow Y[t]$
6: static_rank $\leftarrow$ static_ranks$[t]$
7: best_rank $\leftarrow$ first two features of static_rank {Initial seed ranking}
8: **for** each instance $x \in X_{\text{valid}}$ **do**
9: Impute missing values in x using a KNN Imputer fit on X
10: $\mathcal{F}_\mathcal{A} \leftarrow$ Features$(X) \setminus$ best_rank {Candidate features not yet selected}
11: $\mathcal{F}_\mathcal{B} \leftarrow$ best_rank {Current ranked feature set}
12: **while** $\mathcal{F}_\mathcal{A} \neq \emptyset$ **do**
13: **for** each $f \in \mathcal{F}_\mathcal{A}$ **do**
14: $p(f = r \mid x_{\mathcal{F}_\mathcal{B}}) \leftarrow$ predictProbabilities$(X, \mathcal{F}_\mathcal{B}, f, x)$ {Predictive distribution of f given selected features}
15: Train model on $X[\mathcal{F}_\mathcal{B}]$ to predict y_t and compute baseline prediction $\hat{y}_{\text{prev}}$
16: Compute expected utility:
17: $\mathbb{E}(U_f) \leftarrow \sum_r p(f = r \mid x_{\mathcal{F}_\mathcal{B}}) \cdot \log p(f = r \mid x_{\mathcal{F}_\mathcal{B}}) \cdot |\hat{y}_r - \hat{y}_{\text{prev}}|$
18: **end for**
19: $f^* \leftarrow \arg\min_{f \in \mathcal{F}_\mathcal{A}} \mathbb{E}(U_f)$ {Select feature with minimum expected utility}
20: $\mathcal{F}_\mathcal{B} \leftarrow \mathcal{F}_\mathcal{B} \cup \{f^*\}$ {Add to ranked features}
21: $\mathcal{F}_\mathcal{A} \leftarrow \mathcal{F}_\mathcal{A} \setminus \{f^*\}$
22: **end while**
23: Save $\mathcal{F}_\mathcal{B}$ as final ranked list for instance x
24: **end for**
25: best_ranks_target$[t] \leftarrow$ ranked lists for all instances
26: **end for**
27: **return** best_ranks_target

Algorithm 5.2: COnditional probability estimation subroutine for Algorithm 5.1 (predictProbabilities)

Input: Training dataset X, known feature set $\mathcal{F}_\mathcal{B}$, candidate feature f_{pred}, target instance x

Output: Predicted probabilities dict_prob such that dict_prob$[r] = p(f_{\text{pred}} = r \mid x_{\mathcal{F}_\mathcal{B}})$

Step 1: Prepare training data
$X_{\text{part}} \leftarrow X[\mathcal{F}_\mathcal{B}]$ {Subset restricted to selected features}
$y_{\text{pred}} \leftarrow X[f_{\text{pred}}]$ {Candidate feature as categorical labels}
Step 2: Train classifier
Train Random Forest classifier on $(X_{\text{part}}, y_{\text{pred}})$
Step 3: Predict class probabilities
$P \leftarrow$ predict_proba$(x_{\mathcal{F}_\mathcal{B}})$ {Probability vector for instance x}
$\mathcal{R} \leftarrow$ Classes(f_{pred}) {Distinct values of the candidate feature}
dict_prob $\leftarrow \{\}$ {Initialize probability dictionary}
for each $r \in \mathcal{R}$ **do**
    dict_prob$[r] \leftarrow P[r]$
end for
return dict_prob

Examples of Utility Functions

The utility function $U(f \coloneqq r)$ quantifies how informative a candidate answer value r of feature f is for a given prediction task. Different utility definitions capture different notions of informativeness, such as how much a prediction changes, how relevant a feature is from a nutritional perspective, or how much uncertainty is reduced. Below we summarise the utility functions used in this work.

- **Maximum Prediction Shift (MPS)**:

$$U(f \coloneqq r) = |\hat{y}_r - \hat{y}_{\text{prev}}|,$$

  where $\hat{y}_{\text{prev}}$ is the model prediction based on the currently selected features $\mathcal{F}_B$, and $\hat{y}_r$ is the prediction obtained when the candidate feature f is hypothetically set to value r. This utility favours questions whose answers are likely to cause a large change in the current prediction, indicating high influence.

- **Nutrient Contribution (NC)**:

$$U(f \coloneqq r) = g(f_r),$$

  where $g(f_r)$ is a domain-informed function that maps the answer value r of feature f to an estimated contribution (e.g., grams per day) of the target nutrient. This utility directly encodes expert knowledge, prioritising questions that are known to be nutritionally important for the target outcome.

- **Maximum Prediction Shift Scaled by Nutrient Contribution (MPSSNC)**:

$$U(f \coloneqq r) = |\hat{y}_r - \hat{y}_{\text{prev}}| \cdot g(f_r),$$

  which combines model sensitivity (MPS) with nutritional relevance (NC). This hybrid utility prioritises questions whose answers are expected to both strongly affect the prediction and correspond to substantial nutrient intake.

- **Maximum Uncertainty Reduction (MUR)**:

$$U(f \coloneqq r) = \sigma_{\text{prev}} - \sigma_r,$$

where σ_{prev} is the predictive uncertainty given the current feature set $\mathcal{F}_B$, and σ_r is the uncertainty after assuming $f = r$. This utility prioritises questions that reduce prediction uncertainty.

The uncertainty is estimated as:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2},$$

where $\hat{y}_i$ are predictions obtained from multiple stochastic model evaluations (e.g., different trees in an ensemble), and $\bar{\hat{y}}$ is their mean. A larger reduction in uncertainty indicates a more informative question.

5.2.2 Dynamic Question Ordering for Multi-Target Prediction Problems

In many real-world applications, a single input is associated with multiple dependent outcomes. Traditional predictive modelling often adopts a single-target approach, where each outcome variable is predicted independently using a separate model. While this is conceptually simple, it ignores the fact that targets may be related and influenced by shared input patterns. Multi-Target Prediction (MTP) addresses this setting by learning a single model to predict multiple targets jointly.

In the context of Food Frequency Questionnaires (FFQs), this formulation is particularly natural. A single dietary habit often contributes to multiple nutrient estimates. For example, frequent consumption of sweets increases both sugar and fat intake, while fruit consumption affects both fruit and fibre intake. These relationships do not imply that one target determines another; instead, they reflect shared explanatory structure in the questionnaire responses. Treating each nutrient independently therefore duplicates modelling effort and prevents the model from exploiting common dietary patterns.

For this reason, the motivation for using Multi-Target Regression (MTR) in this thesis is not to claim guaranteed improvements in predictive accuracy, but to provide a coherent modelling framework in which related dietary outcomes are learned jointly. By training a single model to predict multiple nutrient targets simultaneously, MTR allows shared sensitivities in the input data to be captured explicitly and ensures that predictions for different nutrients are consistent with the same underlying responses. This is particularly important when predictions are made from partial questionnaires, where limited information must be used efficiently across all targets.

This thesis focuses on regression-based MTR because earlier experiments showed that classification-based formulations on the same targets performed poorly. Regression further provides a continuous prediction space, which is better suited to modelling gradual changes in nutrient intake as additional questions are answered.

The dynamic question ordering procedure for the multi-target setting builds directly on the single-target algorithm described in Algorithm 5.1. For this reason, a separate algorithmic listing is not introduced. Instead, we describe the key conceptual and computational differences that arise when extending the method to multi-target regression.

The most crucial distinction is that a single predictive model is trained to jointly predict all targets. As a result, the algorithm no longer iterates over targets. At each step, predictions, uncertainties, and utility values are computed simultaneously for the whole

target vector. This allows the model to capture dependencies among targets implicitly, rather than through separate prediction pipelines.

The expected utility computation is adapted accordingly. In the single-target setting, utility is defined with respect to one outcome variable. In the multi-target setting, the utility of acquiring a candidate feature value is computed by aggregating its effect across all targets. Utility functions such as Maximum Multi-Target Prediction Shift (MMTPS) or Maximum Multi-Target Uncertainty Reduction (MMTUR) quantify how informative a question is when its expected impact is evaluated jointly over the target set $\mathcal{T}$. When target prioritisation is required, weighted variants (e.g. WMTUR) can be used, assigning different importance to different targets.

Apart from these changes, all other components of the procedure remain unchanged. Conditional answer probabilities are estimated in the same way, quasi-entropy is used to weight expected informativeness, and features are selected iteratively by minimising expected utility. As in the single-target case, the outcome is an instance-specific ranking that prioritises questions expected to produce the largest improvement in the current prediction, now evaluated jointly across multiple targets.

Utility Functions for Multi-Target Regression

In the multi-target setting, utility functions for dynamic question ordering are natural extensions of those defined for the single-target problem. Rather than computing utilities separately for each outcome, a single utility value is obtained by aggregating information across all targets $t \in \mathcal{T}$. This allows the algorithm to prioritise questions that are informative across multiple related prediction targets simultaneously.

We focus on model-based utility functions derived from prediction behaviour and predictive uncertainty. In contrast to the single-target setting, nutrient-contribution-based utilities are not used here. Because multi-target regression relies on a single predictive model that jointly outputs all targets, assigning separate nutrient-specific contributions would no longer be consistent with the model's structure or the prediction task.

The following utility functions are used in the multi-target setting:

- **Multi-Target Uncertainty Reduction (MTUR)**:

$$U(f := r) = \sum_{t \in \mathcal{T}} \left(\sigma_{\text{prev},t} - \sigma_{f:=r,t}\right),$$

  where $\sigma_{\text{prev},t}$ denotes the predictive uncertainty for target t given the currently selected features, and $\sigma_{f:=r,t}$ is the uncertainty after assuming that feature f takes value r.

- **Multi-Target Maximum Prediction Shift (MTMPS)**:

$$U(f := r) = \sum_{t \in \mathcal{T}} \left|y_{\text{pred},f:=r,t} - y_{\text{pred, prev},t}\right|,$$

  where $y_{\text{pred, prev},t}$ is the current prediction for target t, and $y_{\text{pred},f:=r,t}$ is the prediction after incorporating the hypothetical answer $f = r$.

- **Weighted Multi-Target Uncertainty Reduction (WMTUR)**:

$$U(f := r) = \sum_{t \in \mathcal{T}} w_t \left(\sigma_{\text{prev},t} - \sigma_{f:=r,t}\right),$$

  where $w_t \geq 0$ is a weight assigned to target t. When all weights are equal, WMTUR reduces to MTUR.

These utility functions are used within the expected utility framework of Algorithm 5.1, where the per-answer utility $U(f \coloneqq r)$ is combined with the estimated conditional probabilities $p(f = r \mid x_{\mathcal{F}_B})$ to determine the next most informative question.

Target Weighting Strategies

In multi-target regression, not all targets necessarily contribute equally to uncertainty reduction or prediction shifts. Some outcomes may be more critical due to domain considerations, higher variance, or downstream relevance. Different weighting strategies can therefore be applied when computing utility functions.

- **Equal Weights (EW)**:
$$w_t = \frac{1}{|\mathcal{T}|},$$
assigning equal importance to all targets.

- **Variance-Based Weights (VBW)**:
$$w_t = \frac{\sigma_{\text{prev},t}}{\sum_{t' \in \mathcal{T}} \sigma_{\text{prev},t'}},$$
emphasising targets with higher current uncertainty.

- **Inverse Variance Weights (IVW)**:
$$w_t = \frac{1/\sigma_{\text{prev},t}}{\sum_{t' \in \mathcal{T}} 1/\sigma_{\text{prev},t'}},$$
prioritising more stable targets.

- **Manually Assigned Weights (MAW)**: where weights w_t are specified based on domain expertise.

- **Prediction Sensitivity-Based Weights (PSW)**:
$$w_t = \frac{\sum_{r \in R} |y_{\text{pred},r,t} - y_{\text{true},r,t}|}{\sum_{t' \in \mathcal{T}} \sum_{r \in R} \left| y_{\text{pred},r,t'} - y_{\text{true},r,t'} \right|},$$
assigning a higher weight to targets with larger overall prediction error across feature-value scenarios.

Chapter 6

Evaluation and Results

6.1 Evaluation

In this chapter, we describe the machine learning evaluation methods applied to assess the performance of the predictive models used in this study. These methods were designed to rigorously evaluate static and dynamic question-ordering approaches across various dietary targets, ensuring that the models were effective, reliable, and suitable for real-world dietary assessment applications.

6.1.1 Evaluation of Predictive Models

The evaluation of predictive models in machine learning focuses on examining the models' behaviour, interpretability, and reliability beyond raw accuracy scores. In this study, we used several machine-learning-based evaluation techniques to assess the strengths and limitations of models developed for dietary prediction tasks. The main approaches employed are as follows:

- **Cross-Validation:** To assess model robustness and generalisability, we used 5-fold cross-validation. The training dataset was partitioned into five equal subsets, with each iteration using four folds for training and one for validation. This process was repeated five times, ensuring that each subset served as a validation set once. The averaged performance across folds provided a more stable estimate and reduced the risk of overfitting. Cross-validation was applied to both static and dynamic question-ordering models to ensure a fair and consistent evaluation of their predictive performance.

- **Feature Importance Analysis:** We analysed feature importance to identify the most informative questions for each dietary prediction task. The method of importance extraction depended on the model type: for linear models (Logistic Regression, Linear Regression), feature importance was derived from the absolute values of the learned coefficients; for ensemble models such as Random Forest, we used the `feature_importances_` attribute from `sklearn`. These importance scores were used to rank features, informing both the fixed (static) question sequence and the adaptive selection logic in dynamic ordering.

- **Evaluation Metrics:** Model performance was evaluated using task-specific metrics. For classification tasks (e.g., predicting dietary quality scores), we used accuracy, precision, recall, and F1-score to assess classification performance. For regression tasks (e.g., estimating nutrient intake), we employed Mean Absolute Error (MAE), Mean

Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2) to measure predictive accuracy. These metrics enabled a comprehensive comparison across diverse dietary targets and output types.

These machine learning evaluation methods provided a comprehensive framework for testing predictive models in dietary assessment, enabling us to examine both static and dynamic question-ordering approaches rigorously. By employing cross-validation, feature-importance analysis, and task-appropriate performance metrics, we ensured that our evaluation was robust, interpretable, and consistent across the different experimental settings considered in this thesis.

6.1.2 PROMETHEE

Evaluating machine-learning models across multiple feature subsets and dietary targets typically involves a range of performance metrics. While this multi-metric evaluation increases robustness and fairness, it also complicates the comparison process. With each model producing a mix of results (e.g., some with higher accuracy, others with lower error), drawing clear conclusions becomes challenging—especially as the number of combinations grows.

To address this, we adopt a multi-criteria decision-making (MCDM) approach that integrates multiple performance metrics into a single, interpretable ranking. Specifically, we use the PROMETHEE method—short for Preference Ranking Organization METHod for Enrichment Evaluations—proposed initially by Jean-Pierre Brans and later developed with Bertrand Mareschal [95], [96]. Although initially designed for general multi-criteria decision problems, PROMETHEE is flexible and well-suited to both classification and regression tasks. We follow the approach of Eftimov and Kocev [97], who use PROMETHEE to combine multiple evaluation metrics into a single ranking for comparing models across targets.

In our setting, PROMETHEE helps rank machine-learning models and feature subsets across multiple dietary targets by combining multiple evaluation metrics into a single decision-making framework. This is particularly advantageous during the exploratory phase of our analysis, where we compare many models and feature combinations. PROMETHEE serves as a high-level filter, highlighting promising candidates without oversimplifying results. Later in the analysis, when fewer models are under consideration, we revert to more conventional metrics (e.g., accuracy, precision, MAE) for interpretability and detailed reporting.

PROMETHEE Framework

Figure 6.1 shows the PROMETHEE pipeline for ranking feature subsets; the same approach applies to model comparison by simply switching the roles of subsets (F) and methods (M).

PROMETHEE (Preference Ranking Organization METHod for Enrichment Evaluations) is a multi-criteria decision-making (MCDM) method used to compare and rank different alternatives when multiple, often conflicting, criteria must be considered. Instead of relying on a single performance measure, PROMETHEE evaluates all relevant criteria together to provide a balanced overall assessment.

The method works by performing pairwise comparisons across all alternatives (e.g., model M_1 vs. M_2) for each criterion. For every comparison, it calculates the degree of preference—that is, how much one alternative is favoured over another based on that specific criterion. These preferences are then aggregated across all criteria, weighted by

their relative importance, to produce a comprehensive ranking of alternatives. In essence, PROMETHEE helps identify the best overall options when multiple evaluation criteria are considered simultaneously, providing a structured, transparent approach to complex decision-making.

We begin by training k different models $M_1, M_2, \ldots, M_k$ on each of n feature subsets $F_1, F_2, \ldots, F_n$, for each of m dietary targets $T_1, T_2, \ldots, T_m$. For each combination of target and feature subset, we compute a decision matrix D, where each entry $q_j(M_i)$ is a performance score from metric j (e.g., precision, recall, or MSE) for model M_i:

$$D = \begin{bmatrix} q_1(M_1) & q_2(M_1) & \ldots & q_N(M_1) \\ q_1(M_2) & q_2(M_2) & \ldots & q_N(M_2) \\ \vdots & \vdots & \ddots & \vdots \\ q_1(M_k) & q_2(M_k) & \ldots & q_N(M_k) \end{bmatrix} \tag{6.1}$$

Next, for each metric q_j, we perform pairwise comparisons between models to determine which one performs better and by how much. To express these differences consistently, PROMETHEE uses a V-shaped preference function that smoothly maps differences in metric values to a preference degree ranging from 0 (no preference) to 1 (strong preference).

An indifference threshold Q defines the smallest performance difference considered relevant or practically significant for a given metric [97]. Minor differences (below Q) are treated as negligible, meaning the two models are regarded as equivalent in practical terms. As the difference increases beyond Q, the preference value grows linearly until it reaches 1, indicating a clear advantage of one model over the other. For example, if accuracy is the criterion and $Q = 0.02$, two models differing by less than 2% in accuracy are considered practically equivalent. However, as the gap exceeds this threshold, PROMETHEE assigns a stronger preference to the model with higher accuracy, reflecting a more meaningful improvement.

Formally, the preference function is defined as:

$$P_j(M_1, M_2) = \begin{cases} p_j(q_j(M_1) - q_j(M_2)), & \text{if } q_j \text{ is to be maximised,} \\ p_j(q_j(M_2) - q_j(M_1)), & \text{if } q_j \text{ is to be minimised,} \end{cases}$$

$$p(x) = \begin{cases} 0, & x \leq 0, \\ \frac{x}{Q}, & 0 < x \leq Q, \\ 1, & x > Q. \end{cases}$$

Pairwise preferences are then aggregated into an average preference index:

$$\pi(M_1, M_2) = \frac{1}{N} \sum_{j=1}^{N} \omega_j P_j(M_1, M_2),$$

where ω_j is the relative weight of metric q_j, reflecting its importance in the overall decision.

Weight Computation. The weights ω_j determine how strongly each metric influences the final ranking. In the classical PROMETHEE framework, these can be assigned manually, based on expert judgment or domain knowledge. However, following the data-driven approach proposed by Eftimov and Kocev [97], we automatically compute the weights from the performance data to ensure objectivity and reproducibility.

Specifically, Eftimov and Kocev [97] suggests assigning higher weights to metrics with greater discriminative power across models — that is, metrics that better distinguish model performance. For each metric q_j, the variability of its values (e.g., the standard deviation or entropy across all models) is used as an indicator of how informative that metric is. Metrics with larger variance or higher information gain contribute more to distinguishing between models and therefore receive higher weights. Formally, each weight can be defined as:

$$\omega_j = \frac{v_j}{\sum_{l=1}^{N} v_l},$$

where v_j is a measure of variability (such as standard deviation) of metric q_j across all models. This normalisation ensures that the weights sum to one: $\sum_{j=1}^{N} \omega_j = 1$.

This data-driven weighting approach reduces subjective bias and emphasises the metrics most useful for ranking.

After weighting and aggregating preferences, PROMETHEE computes three preference flows for each method:

- **Positive flow** (ϕ^+) — how much a method is preferred over others.

$$\phi(M_i^+) = \frac{1}{k-1} \sum_{x \in M} \pi(M_i, x)$$

- **Negative flow** (ϕ^-) — how much other methods are preferred over it.

$$\phi(M_i^-) = \frac{1}{k-1} \sum_{x \in M} \pi(x, M_i)$$

- **Net flow** (ϕ) — the overall preference strength, defined as the difference between positive and negative flows:

$$\phi(M_i) = \phi(M_i^+) - \phi(M_i^-)$$

The resulting $\phi(M_i)$ values provide an interpretable ranking of models or feature subsets: higher net flow indicates better overall performance when all metrics are considered simultaneously. Rankings are computed separately for each dietary target and can be averaged to identify the most robust feature subset across all targets. This structured process enables consistent, transparent, and interpretable model and feature subset evaluation by integrating multiple performance criteria into a single, data-informed ranking.

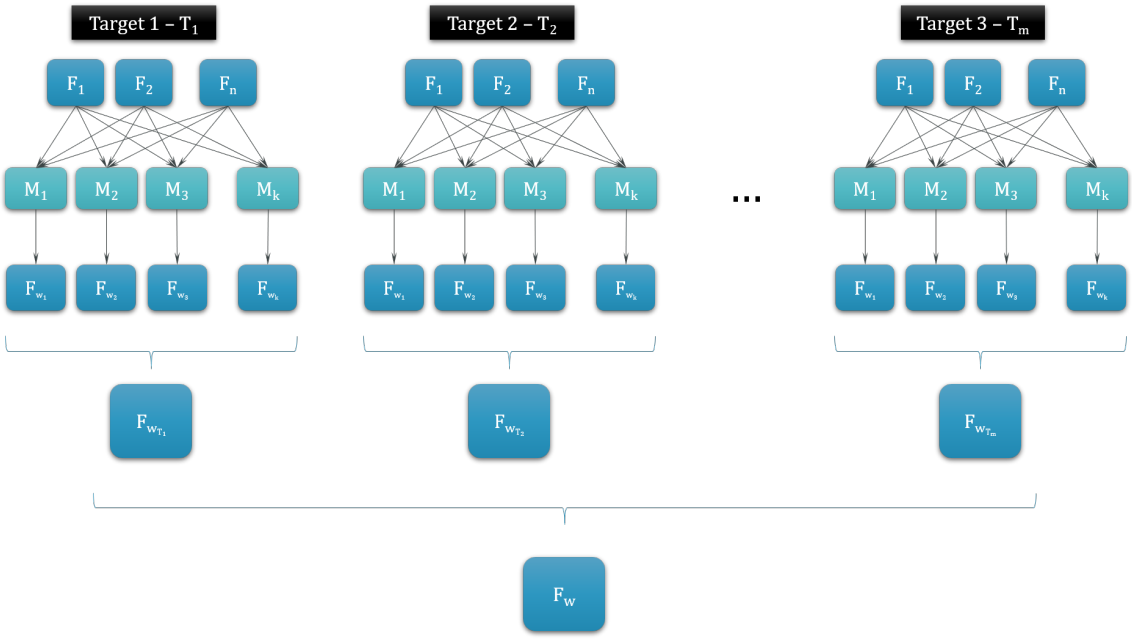


Figure 6.1: PROMETHEE pipeline for ranking feature subsets. F represents feature subsets; M denotes machine-learning methods. F_{w_i} is the best-performing subset for method M_i, $F_{\mathrm{w}_{T_1}}$ is the best for target T_1, and F_{w} performs best across all targets.

6.1.3 Evaluation of Static and Dynamic Question Ordering

The evaluation of static and dynamic question-ordering approaches focuses on predictive accuracy and response efficiency.

- **Predictive Accuracy:** The predictive accuracy of static and dynamic ordering was evaluated for each dietary target using the metrics appropriate for the prediction type—classification metrics (accuracy, precision, recall, and F1-score) for categorical dietary scores and regression metrics (Mean Absolute Error, Mean Absolute Percentage Error, Mean Squared Error, and R^2) for continuous nutrient intake values. For static ordering, the model followed a predetermined sequence based on global feature importance, whereas in dynamic ordering, the question sequence was adapted in real time based on user responses. By comparing the accuracy metrics of the two methods, we found that one method produced more precise predictions for specific dietary targets.
- **Response Efficiency and Question Count:** An essential aspect of the evaluation involved analysing the efficiency of each question ordering method in terms of the number of questions required to achieve an optimal prediction. In static ordering, questions were asked in a fixed order, which could lead to redundancy if high-importance questions were placed later in the sequence. In contrast, dynamic ordering prioritised questions adaptively, selecting the most relevant question based on prior responses, potentially reducing the number of questions needed. We measured the average number of questions required for each approach to achieve a specified level of predictive accuracy, highlighting the efficiency gains of.

Through this comprehensive evaluation of static and dynamic question ordering, we gained insights into the effectiveness and applicability of each method in the assessment.

This analysis highlighted the strengths of each approach in specific contexts. It provided a balanced understanding of the trade-offs, informing decisions on the optimal ordering method for different dietary assessment applications.

6.2 Results

Across ESFFFQ, the proposed reduction and ordering methods produce clear and consistent gains. On NHANES, trends are similar, but the overall prediction task is more challenging because laboratory biomarkers are only partially explained by FFQ responses alone.

6.2.1 Findings from Target-Independent Question Reduction Methods

In this section, we present and analyse the performance of the methods introduced in Section 4.2.1. To evaluate and compare question-ranking approaches across multiple dietary targets and feature subsets, we apply the PROMETHEE multi-criteria decision-making framework. This method enables us to integrate multiple evaluation metrics into a unified ranking, providing a robust, interpretable comparison across a wide range of model–feature configurations.

For this evaluation, we focus on frequency-based feature representations based on findings in previous experiments. This trend is supported by results later in section 6.2.3.

To ensure clarity and reduce redundancy, we present detailed results for a single representative target from each FFQ: fat for the ESFFFQ and total cholesterol for the NHANES FFQ. Full results for all other dietary targets are provided in the Appendix.

ESFFFQ

Table 6.1 summarises the composition of progressively larger feature subsets obtained using weighted correlation and SHAP-based interaction methods, illustrating how individual FFQ items are introduced as subset size increases and highlighting differences in selection order between the two reduction strategies.

Table 6.2 and Figure 6.2 compare model performance on SHAP-based and Weighted Correlation–based feature subsets for the ESFFFQ target fat. Model rankings reported in the tables are obtained using the PROMETHEE multi-criteria decision analysis method, which combines multiple evaluation metrics into a single overall ranking. In contrast, the figures visualise performance using a single representative metric for clarity of presentation. This distinction applies consistently to all corresponding pairs of ranking tables and performance figures throughout this chapter.

Overall performance improves as more features are added. Bagging, Logistic Regression, and Decision Tree models frequently rank among the top performers, particularly on smaller feature subsets (P0–P2, S0–S2). For example, Bagging ranks first at P0 and remains competitive across larger subsets.

Imputation-based baselines, especially MICE, perform well on later subsets, where fewer missing values reduce the reliance on imputation. This suggests that simpler methods can sometimes match or outperform more complex models when feature sets are small or moderately sized, likely due to reduced overfitting.

F1-score plots show SHAP-based subsets offer stable performance, while Weighted Correlation creates more variation—often benefiting top models like Bagging and Logistic Regression. Overall, Weighted Correlation shows a slight edge in producing more discriminative subsets, especially when model complexity is low.

Table 6.1: Feature inclusion across subsets grouped by Weighted Correlation and SHAP-based Interaction.

Feature	**Weighted Correlation**						**SHAP-based Interaction**					
	P0	P1	P2	P3	P4	P5	S0	S1	S2	S3	S4	S5
batterfish		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
beans		✓	✓	✓	✓	✓						✓
bread						✓			✓	✓	✓	✓
cakes	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓
cheese			✓	✓	✓	✓				✓	✓	✓
chips				✓	✓	✓		✓	✓	✓	✓	✓
cream				✓	✓	✓					✓	✓
crisps					✓	✓			✓	✓	✓	✓
eggs						✓	✓	✓	✓	✓	✓	✓
fibre	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓
fish			✓	✓	✓	✓			✓	✓	✓	✓
fruit			✓	✓	✓	✓				✓	✓	✓
grains			✓	✓	✓	✓						✓
juice						✓			✓	✓	✓	✓
nuts		✓	✓	✓	✓	✓					✓	✓
nuts_salt			✓	✓	✓	✓						✓
nuggets						✓		✓	✓	✓	✓	✓
pizza			✓	✓	✓	✓				✓	✓	✓
pop			✓	✓	✓	✓			✓	✓	✓	✓
potato				✓	✓	✓						✓
pre-processed					✓	✓					✓	✓
redmeat					✓	✓		✓	✓	✓	✓	✓
salad					✓	✓			✓	✓	✓	✓
sausage						✓		✓	✓	✓	✓	✓
vegetables				✓	✓	✓		✓	✓	✓	✓	✓
whitemeat		✓	✓	✓	✓	✓						✓
wholebread		✓	✓	✓	✓	✓					✓	✓
Total Features	2	7	13	18	22	27	2	7	13	18	22	27

Table 6.2: Ranks for classification compared for target fat by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	2	7	13	18	22	27		2	7	13	18	22	27	
RF	7	4	6	5	7	7	6.0	7	4	5	9	7	7	6.5
Bagging	**1**	3	2	7	4	3	**3.3**	**1**	5	**1**	4	9	3	3.8
GB	4	7	7	9	9	10	7.7	4	3	10	10	10	10	7.8
LogR	6	8	5	6	**1**	4	5.0	6	6	4	**1**	6	4	4.5
kNN	3	**1**	4	2	6	6	3.7	3	**1**	2	2	3	6	**2.8**
DT	2	6	**1**	4	2	5	**3.3**	2	9	6	6	5	5	5.5
SVM	8	9	10	10	10	8	9.2	8	2	7	8	**1**	8	5.7
VC	5	5	9	8	3	9	6.5	5	8	9	7	2	9	6.7
Zero	9	10	8	**1**	5	**1**	5.7	9	10	8	5	4	**1**	6.2
MICE	9	2	3	3	8	**1**	4.3	9	7	3	3	8	**1**	5.2

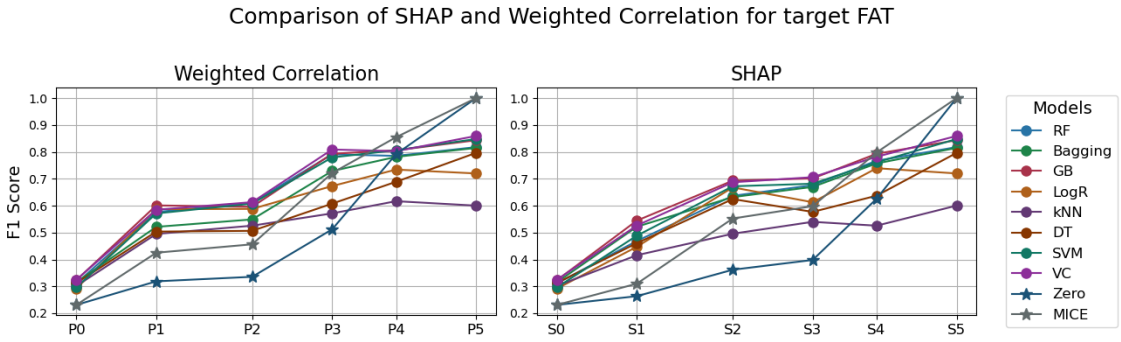


Figure 6.2: Comparison of the F1-scores of different models on subsets created using SHAP and Weighted Correlation for target fat.

Table 6.3 and Figure 6.3 present the regression performance of models using subsets selected by SHAP-based Interaction Analysis and Weighted Correlation Analysis for the ESFFFQ target fat. The table shows model rankings across subsets, while the figure displays corresponding MAE values. Linear Regression and Gradient Boosting are among the top-performing models across most subsets. For example, Linear Regression consistently ranks first or second in the early Weighted Correlation subsets (P0–P2), while Gradient Boosting dominates in SHAP-based subsets (S1–S4). These models benefit from fewer features, where simpler models are less prone to overfitting—a common trend when working with smaller, cleaner subsets.

MICE and Zero-imputation baselines show better rankings on later subsets (e.g., P4–P5, S5), where the number of missing values is minimal. As expected, these imputation-based approaches perform better when inference is less necessary, and the target can be estimated more directly from the available data. MAE plots show that Weighted Correlation yields sharper declines in error and greater separation between top- and lower-performing models. In contrast, SHAP subsets offer more gradual improvements. Overall, both feature selection methods are effective. Still, Weighted Correlation appears to yield more discriminative subsets for regression on the target *fat*, especially in the early stages and when paired with Linear Regression or Gradient Boosting.

Table 6.3: Ranks for regression compared for target fat by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	2	7	13	18	22	27		2	7	13	18	22	27	
RF	6	4	4	2	3	4	3.8	6	4	2	3	3	4	3.7
Bagging	3	7	6	5	5	5	5.2	3	5	4	4	4	5	4.2
GB	2	2	3	**1**	2	3	**2.2**	2	**1**	**1**	**1**	**1**	3	**1.5**
LinR	**1**	**1**	**1**	6	8	8	4.2	**1**	3	5	5	8	8	5.0
kNN	8	5	5	9	9	9	7.5	8	6	6	7	9	9	7.5
DT	4	9	9	7	7	7	7.2	4	8	8	8	7	7	7.0
SVR	7	6	7	10	10	10	8.3	7	7	9	9	10	10	8.7
VR	5	3	2	3	6	6	4.2	5	2	3	2	5	6	3.8
Zero	10	10	10	8	4	**1**	7.2	10	10	10	10	6	**1**	7.8
MICE	9	8	8	4	**1**	**1**	5.2	9	9	7	6	2	**1**	5.7

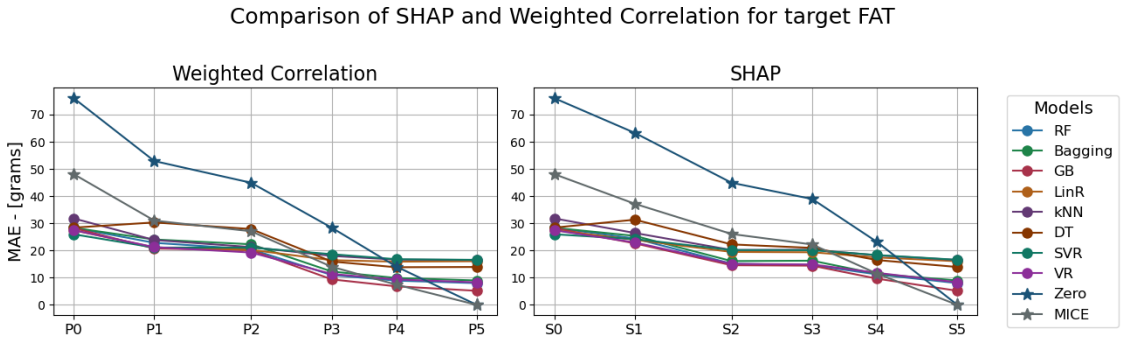


Figure 6.3: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target fat.

NHANES

Table 6.4: Ranks for regression models compared for target total cholesterol by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
#	9	49	75	127	156	200		9	49	75	127	156	200	
RF	5	3	3	**1**	**1**	**1**	2.3	5	3	2	3	4	**1**	3.0
Bagging	6	7	6	5	5	5	5.7	6	6	6	6	6	5	5.8
GB	2	2	**1**	2	2	2	**1.8**	2	**1**	**1**	**1**	2	2	**1.5**
LinR	**1**	**1**	2	6	6	6	3.7	**1**	2	5	5	5	6	4.0
kNN	7	6	7	7	7	7	7.0	7	7	7	7	7	7	7.0
DT	8	8	8	8	8	8	8.0	8	8	8	8	8	8	8.0
SVR	3	5	5	3	3	3	3.7	4	5	4	2	**1**	3	3.2
VR	4	4	4	4	4	4	4.0	3	4	3	4	3	4	3.5

Table 6.4 and Figure 6.4 show the performance of regression models for total cholesterol, using feature subsets created via SHAP-based Interaction Analysis and Weighted Correlation Analysis. As with the ESFFFQ results, both tables and plots show a general improvement in performance as more features are added. Linear Regression and Random Forest models consistently achieve top rankings, especially in the Weighted Correlation subsets. For example, Linear Regression ranks first in P0–P2, and Random Forest dominates from P3 onward. These trends are supported by the MAE plots, which show that both models achieve relatively low, stable error rates across subsets. In contrast to the previous analysis on *fat*, the SHAP-based feature subsets yield more consistent MAE performance across most models, including Random Forest and Linear Regression.

Overall, Weighted Correlation Analysis appears to offer a slight edge in producing subsets that lead to lower MAE for total cholesterol, especially when paired with Random Forest or Linear Regression. However, SHAP-based subsets exhibit more uniform model behaviour, which may be preferable when interpretability and consistency across methods are priorities.

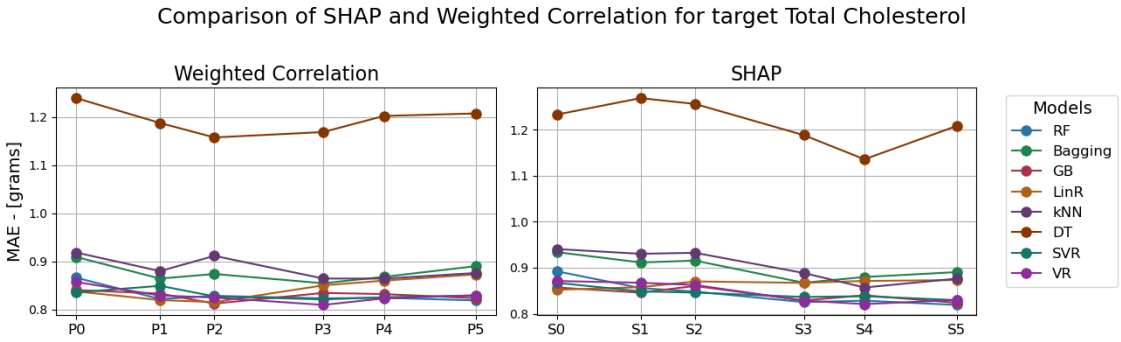


Figure 6.4: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Total Cholesterol.

6.2.1.1 Comparison of Question Reduction Methods

ESFFFQ

Figure 6.2 presents a comparison of the SHAP and Weighted Correlation methods across all targets. The Weighted Correlation method consistently outperforms SHAP across most targets.

Target-Wise Observations:

- **Fat:** Weighted Correlation achieves significantly higher F1 scores compared to SHAP, particularly in the later scenarios. The SHAP method shows relatively slow improvement, indicating limitations in its ability to adapt to complex scenarios.

- **Sugar:** SHAP outperforms Weighted Correlation in all metrics, with sharp improvements from P1 to P3. While Weighted Correlation performs slightly better for smaller sets, the performance stagnates for larger sets

- **Fibre :** For Fibre, Weighted Correlation demonstrates clear dominance, especially for the MAX and AVG F1 scores. The gap between the two methods widens as the scenarios progress, showcasing Weighted Correlation's consistent advantage.

- **Protein:** Similar to other targets, Weighted Correlation maintains superior performance across all metrics. SHAP exhibits minor improvements in later scenarios but fails to close the performance gap.

- **Salt:** Both methods perform poorly in the initial scenarios (P0 and P1). However, Weighted Correlation achieves rapid improvements from P2 onwards, outperforming SHAP, which lags throughout.

The analysis shows that Weighted Correlation is a stronger and more reliable method for classification compared to SHAP. It consistently achieves higher F1 scores across all nutritional targets and performance metrics, particularly in more complex scenarios. Weighted Correlation's steady improvement and adaptability make it a dependable choice for classification tasks. On the other hand, while SHAP shows some progress, it struggles to match the effectiveness of Weighted Correlation as tasks become more complex.

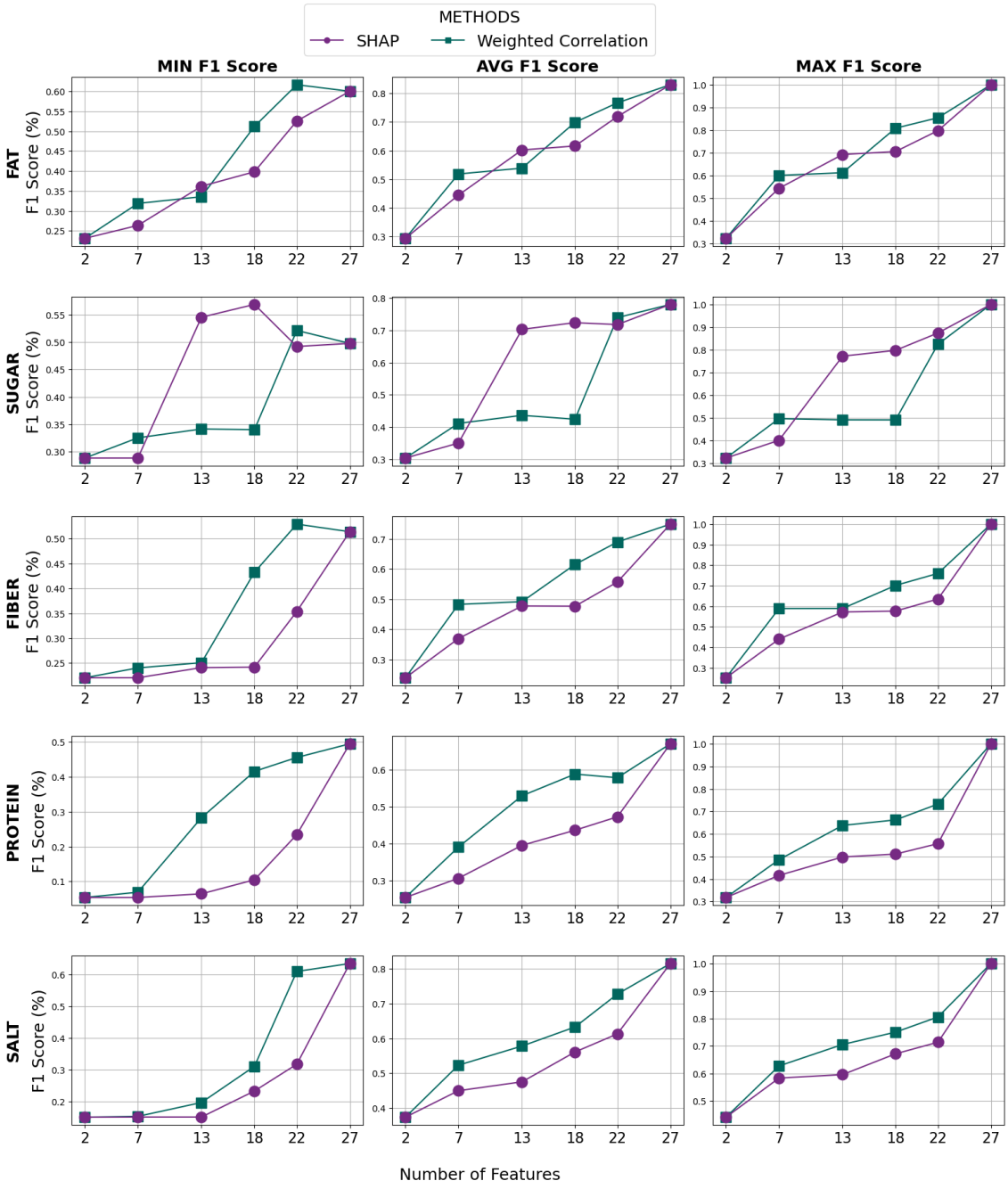


Figure 6.5: Comparison of SHAP and Weighted Correlation for classification.

Figure 6.6 illustrates the comparative performance of SHAP and Weighted Correlation methods for regression. The performance is evaluated using the Mean Absolute Percentage Error (MAPE) across three metrics: Minimum (MIN), Average (AVG), and Maximum (MAX).

The results show that Weighted Correlation consistently achieves lower MAPE values compared to SHAP across most nutritional targets and metrics. Lower MAPE values indicate better regression performance, suggesting that Weighted Correlation yields more accurate predictions.

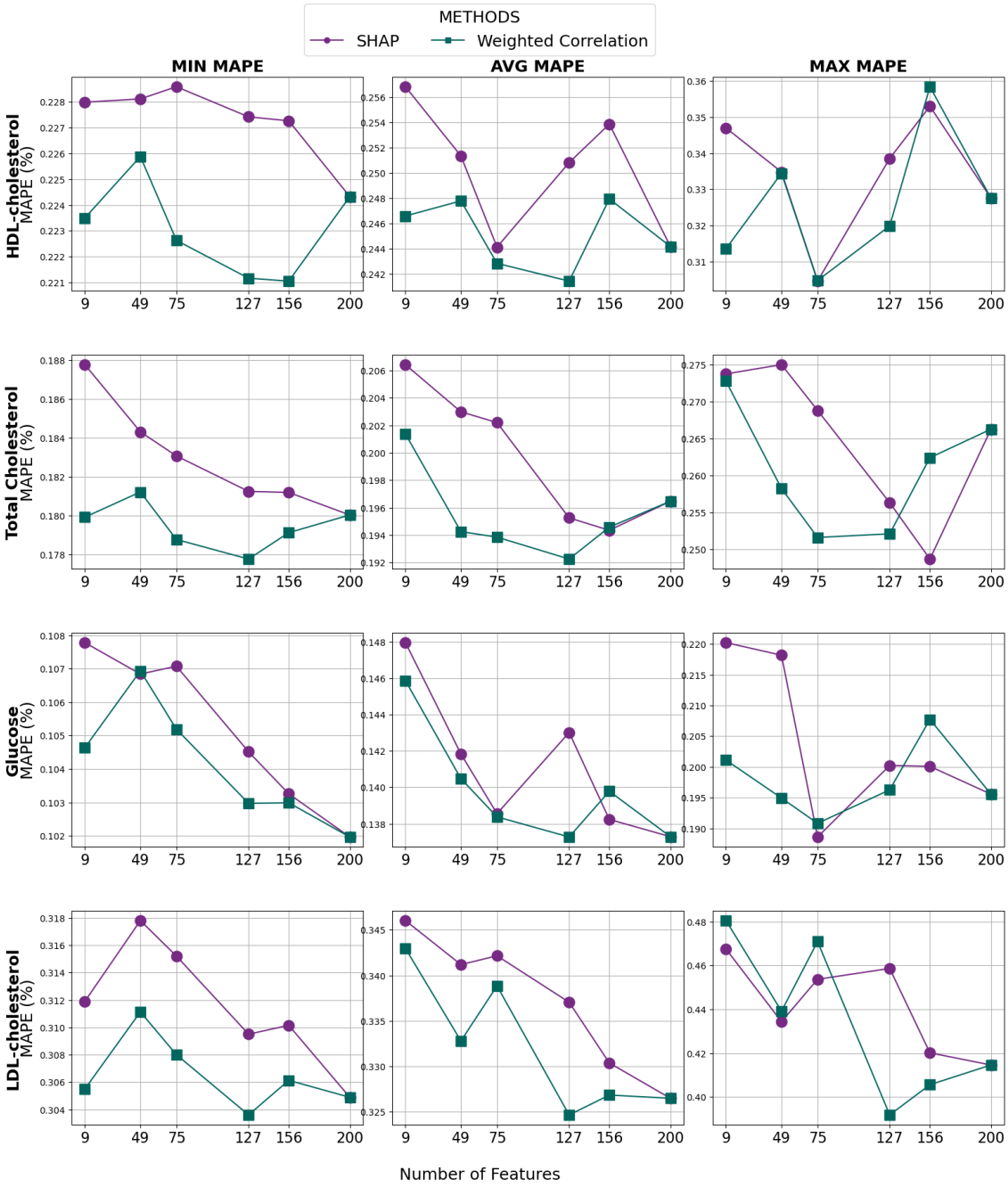


Figure 6.6: Comparison of SHAP and Weighted Correlation for regression.

Target-Wise Observations:

- **Fat:** Weighted Correlation achieves significantly lower MAPE values across all metrics compared to SHAP, particularly in the later scenarios (P3–P5). SHAP demonstrates a slower reduction in error, indicating its limitations in handling more complex tasks.
- **Sugar:** Weighted Correlation shows sharper improvements and consistently outperforms SHAP across all scenarios and metrics. SHAP's performance stagnates in earlier scenarios, widening the performance gap in later stages.

- **Fiber:** Weighted Correlation maintains a clear advantage over SHAP, with significantly lower MAPE values in all metrics. The performance gap becomes more pronounced in advanced scenarios.
- **Protein:** Weighted Correlation consistently outperforms SHAP across all scenarios and metrics. While SHAP shows gradual improvement, it struggles to reach the lower error rates achieved by Weighted Correlation.
- **Salt:** Both methods show improvement across scenarios, but Weighted Correlation demonstrates a more substantial reduction in error compared to SHAP, especially in scenarios P3–P5.

The results indicate that Weighted Correlation is a more effective and reliable method for regression tasks compared to SHAP. It consistently achieves lower MAPE values across all targets and metrics, demonstrating superior accuracy and adaptability. SHAP, while showing some improvements, is less effective in reducing prediction errors, particularly in more complex scenarios. This makes Weighted Correlation a more dependable approach for regression problems.

NHANES

Note on NHANES. Across the NHANES experiments, the same qualitative trends as in ESFFFQ are somehow visible (e.g., in some targets performance improves with more questions and benefits from better ordering), but the absolute error levels remain higher. This suggests that predicting laboratory biomarkers from FFQ responses alone may be a weak-signal problem: diet patterns captured by FFQs explain only part of the variation in biomarkers, and substantial variance is likely driven by factors not represented in the FFQ. This does not invalidate the methods; rather, it highlights the limits of the available inputs in this setting. A natural next step is to combine FFQ responses with additional variables (e.g., demographics, anthropometrics, lifestyle factors, and medical history) to strengthen the prediction signal.

In addition, although a broad set of nutrition-related biomarkers was introduced in Section 3.2.4, the experimental analysis in this section focuses on a smaller subset. This restriction is intentional: only biomarkers with sufficient coverage across participants were included in the regression experiments. Limiting the analysis in this way reduces the risk that performance differences are driven by sparsity in the laboratory data rather than by the modelling and question-ordering methods.

Target-wise observations:

- **HDL Direct (mmol/L):** Weighted Correlation maintains lower MAPE values across all feature sets. SHAP shows a less stable trend, with noticeable peaks in error values, particularly for AVG and MAX MAPE. The performance gap between the methods becomes more evident as the number of features increases.
- **Cholesterol Total (mmol/L):** Weighted Correlation achieves lower error values across all metrics. SHAP struggles to maintain consistency, showing higher MAX MAPE values at multiple feature counts. The benefits of increasing features are more pronounced for Weighted Correlation, leading to a steady decline in MAPE.
- **Glucose (mmol/L):** Both methods improve as the feature count increases, but Weighted Correlation consistently exhibits lower errors. SHAP shows significant fluctuations in AVG and MAX MAPE, while Weighted Correlation maintains a more stable trend. The performance difference is especially clear in scenarios with fewer features (9 and 49), where Weighted Correlation significantly outperforms SHAP.

- **LDL Direct (mmol/L):** Weighted Correlation consistently results in lower MAPE values, particularly in MIN and AVG MAPE. SHAP displays inconsistent behavior, with noticeable peaks in error, especially for AVG and MAX MAPE. The performance advantage of Weighted Correlation is more evident as the number of features increases.

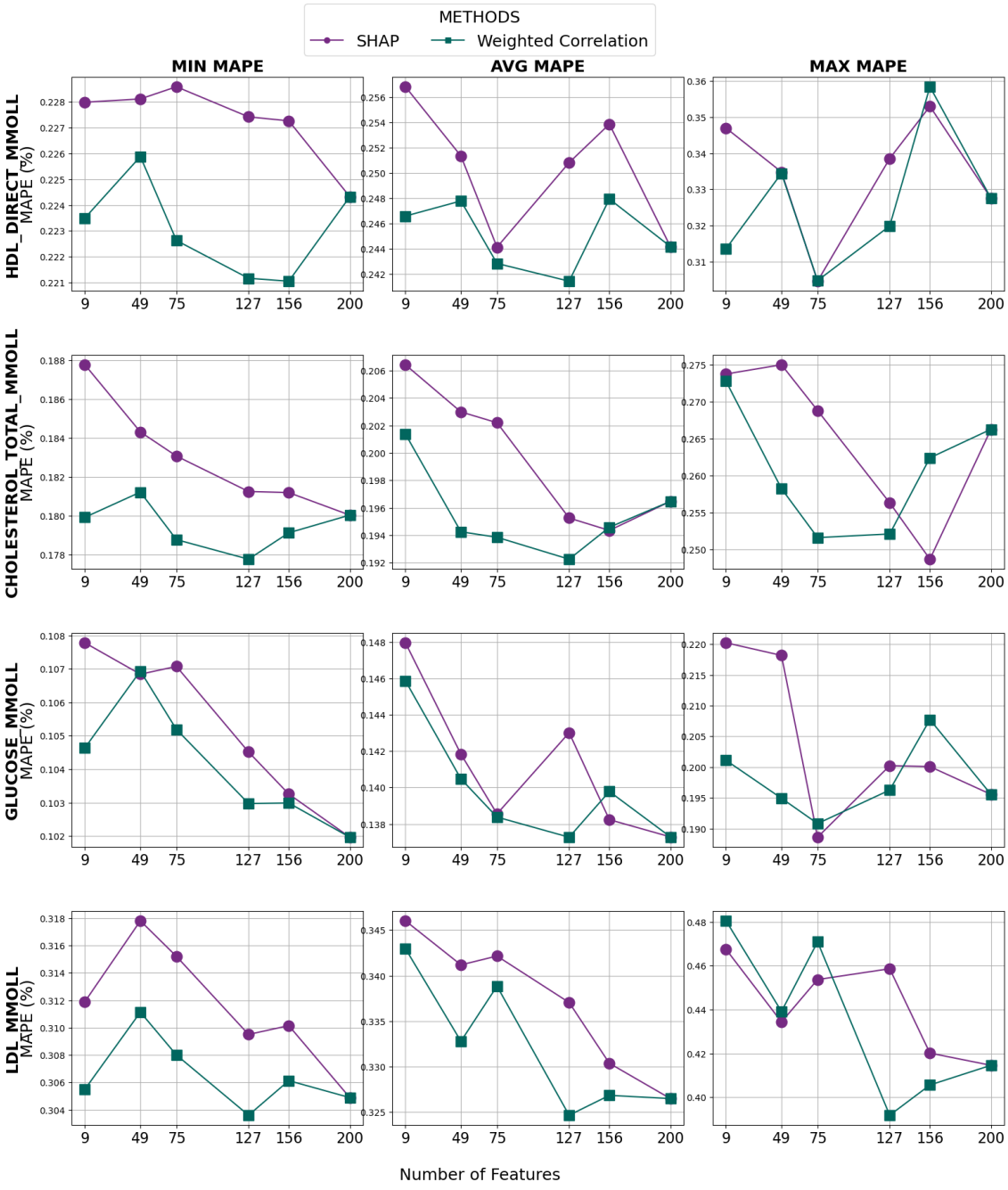


Figure 6.7: Comparison of SHAP and Weighted Correlation for regression.

Once again, the results indicate that Weighted Correlation is a more effective and stable approach for regression tasks in the NHANES dataset compared to SHAP.

6.2.2 Findings from Target-Dependent Feature Selection Methods for FFQ Dimensionality Reduction

In this section, we present and analyse the performance of the feature selection methods introduced in Section 4.2.2, focusing specifically on their effectiveness in FFQ dimensionality reduction. To systematically evaluate these methods across diverse dietary targets and feature subsets, we employ the PROMETHEE multi-criteria decision-making framework. This approach integrates multiple evaluation metrics into a single, interpretable ranking, allowing for fair and comprehensive comparison of different model–feature subset configurations.

For consistency, all evaluations are conducted using frequency-based feature representations, as they have been shown to outperform nominal and amount-based representations—an observation further confirmed by results presented later in section 6.2.3.

To maintain clarity while avoiding excessive repetition, we report detailed results for one representative target per FFQ: *fat intake* for the ESFFFQ and *total cholesterol* for the NHANES FFQ. Results for all other dietary targets are included in the Appendix.

The feature selection methods include ml_rf (integrated function `feature_imporatance` within Random Forest), ml_lr (integrated function `coefficient` within Linear Regression), lasso (Least Absolute Shrinkage and Selection Operator), boruta (Boruta algorithm), shap (SHapley Additive exPlanations), rfe_rf (Recursive Feature Elimination with Random Forest), rfe_lr (Recursive Feature Elimination with Logistic Regression), ga_rf (Genetic Algorithm with Random Forest), and ga_lr (Genetic Algorithm with Logistic Regression).

ESFFFQ

Table 6.5: Ranking of classification algorithms' performance on subsets of features selected by different methods for target Fat.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	6	5	27	13	3	13	13	19	16
RF	6	2	5	7	2	8	8	2	5
Bagging	4	4	**1**	5	4	6	3	7	3
GB	**1**	6	8	8	7	3	7	6	7
LogR	2	3	3	2	5	**1**	**1**	4	2
kNN	3	8	4	6	**1**	4	4	3	6
DT	5	5	2	**1**	3	2	2	**1**	**1**
SVM	8	7	6	4	8	7	5	8	4
VC	7	**1**	7	3	6	5	6	5	8

The results in Table 6.5 and Figure 6.8 evaluate the classification performance of different algorithms for the target Fat on feature subsets generated using various feature selection methods. Overall, LogR and SVM achieved the best performance across most feature selection approaches, often obtaining the lowest values. In contrast, models such as GB and RF generally showed higher values, indicating weaker performance under these settings. Among the feature selection methods, RFE (both rfe_rf and rfe_lr) and GA-based methods (ga_rf and ga_lr) provided consistently good results across multiple models, while LASSO and Boruta tended to yield less favorable outcomes. These findthan LASSO and SHAP, while Decision Tree shows strong performance onfeature selection approaches lead to more efficient performance in this context.

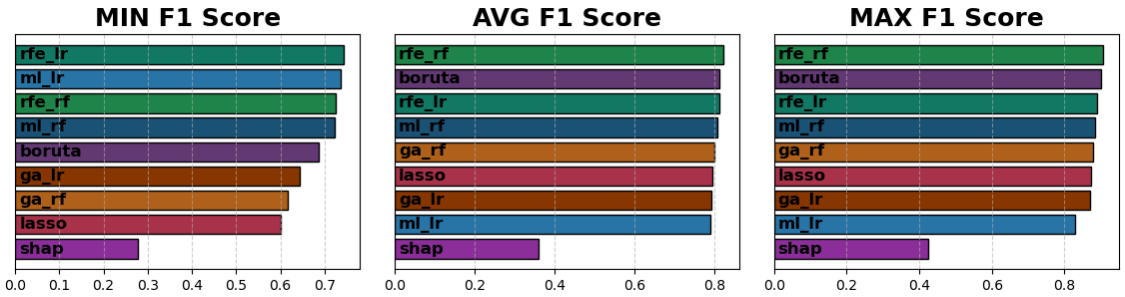


Figure 6.8: Comparison of F1-score on subsets of features selected by different methods for target Fat.

Figure 6.8 illustrates the comparison of F1-scores (minimum, average, and maximum) for the FAT target across different feature selection methods. Overall, RFE-based approaches (rfe_rf and rfe_lr) and GA-based methods (ga_rf and ga_lr) achieved the best results. Boruta also performed competitively, particularly for the average and maximum F1-scores. In contrast, lasso and especially shap exhibited noticeably lower F1-scores, indicating weaker predictive performance. These results reinforce the effectiveness of the Gradient Boosting as the top-performing model and underscore s target.

Table 6.6: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Fat.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	8	5	20	12	7	13	13	13	17
RF	2	5	2	2	2	2	2	2	2
Bagging	4	7	3	4	4	4	3	4	3
GB	**1**	2	**1**	**1**	**1**	**1**	**1**	**1**	**1**
LinR	6	**1**	6	6	7	6	6	6	6
kNN	7	4	7	7	6	7	7	7	7
DT	5	8	5	5	5	5	5	5	5
SVR	8	6	8	8	8	8	8	8	8
VR	3	3	4	3	3	3	4	3	4

The results in Table 6.6 and Figure 6.9 evaluate the regression results for the Fat target. The ranking results show that Gradient Boosting (GB) consistently achieved the lowest values across nearly all feature selection methods, indicating superior performance. RF also performed well, maintaining stable and low ranks, while LinR, kNN, and SVR generally showed weaker results with higher values. Among the feature selection techniques, the RFE and GA-based approaches again produced strong and consistent outcomes, confirming their reliability across different models. In contrast, LASSO and Boruta yielded slightly higher rankings, suggesting that they were less effective for this particular target variable.

Figure 6.9 presents the comparison of MAPE values (minimum, average, and maximum) for the target Fat across different feature selection methods. Overall, GA-based methods (ga_rf and ga_lr) and RFE approaches (rfe_rf and rfe_lr) achieved the best performance, with consistently shorter bars indicating lower prediction error. Boruta and LASSO also performed competitively, maintaining relatively low MAPE values. In contrast, ml_lr exhibited the highest MAPE across all three metrics, reflecting weaker accuracy in this

regression context. These results confirm that evolutionary and recursive feature selection strategies are more effective for minimizing prediction error in this target variable.

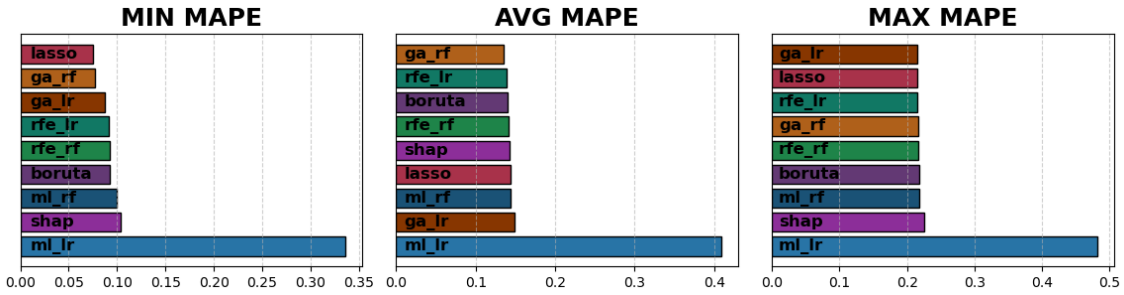


Figure 6.9: Comparison of MAPE on subsets of features selected by different methods for target Fat.

Overall results for ESFFFQ

Table 6.7 summarizes the final classification performance of all models across the nutritional targets using target-dependent feature selection. DT achieved the best overall performance, with the lowest average rank and top results for Fat, Fiber, and Protein. LogR also performed strongly, maintaining consistently low ranks across most targets. Bagging and kNN showed competitive results, while RF, GB, SVM, and VC generally exhibited higher ranks, indicating weaker performance. Overall, DT emerged as the most effective classifier in this setting, followed closely by LogR.

Table 6.8 presents the final regression rankings for target-dependent feature selection. GB achieved the best overall performance, ranking first across all targets with the lowest average values. RF also performed strongly, maintaining stable and low ranks, while VR showed competitive results in several cases. In contrast, LinR, kNN, DT, and SVR generally exhibited higher rankings, indicating weaker regression accuracy. Overall, GB proved to be the most effective and consistent regression model across all nutritional targets.

Table 6.7: Final classification models' performance across all targets for target-dependent feature selection.

	Fat	Sugar	Fiber	Protein	Salt	Overall Rank
RF	5.0	6.3	4.7	5.9	5.8	5.5
Bagging	4.1	3.9	3.0	3.1	5.0	3.8
GB	5.9	5.0	5.8	5.1	3.4	5.0
LogR	2.6	**2.3**	2.8	4.2	6.4	3.7
kNN	4.3	5.9	5.8	2.3	**2.2**	4.1
DT	**2.4**	3.1	**2.3**	**2.2**	4.7	**2.9**
SVM	6.3	3.6	5.2	6.6	4.1	5.2
VC	5.3	5.8	4.7	5.7	4.3	5.2

Table 6.8: Final Regression Rankings for Target-Dependent Feature Selection.

	Fat	Sugar	Fiber	Protein	Salt	Overall Rank
RF	2.3	2.4	2.0	2.3	4.2	2.6
Bagging	4.0	3.7	3.6	4.6	4.7	4.1
GB	**1.1**	**1.1**	**1.0**	**1.3**	**1.0**	**1.1**
LinR	5.6	6.0	5.1	4.3	5.7	5.3
kNN	6.4	6.2	7.8	6.2	7.1	6.7
DT	5.3	5.0	5.9	7.6	7.8	6.3
SVR	7.8	7.4	7.3	7.3	2.3	6.4
VR	3.3	4.1	3.4	2.4	3.2	3.3

NHANES

For the Total Cholesterol, the results in Table 6.9 show that GB and RF achieved the lowest ranks across most feature selection methods, indicating the best overall performance. LinR also performed well in several cases, while DT and kNN generally performed the worst. Among the feature selection techniques, rfe_rf, rfe_lr, ga_rf, and ga_lr provided the most consistent and accurate results, followed closely by boruta.

The MAPE comparison in Figure 6.10 confirms these findings, with GA- and RFE-based methods producing the lowest error, whereas lasso and shap showed higher MAPE values. Overall, combining GB or RF with RFE or GA feature selection leads to the most reliable regression performance for this target.

Table 6.9: Ranking of algorithms' performance on subsets of features selected by different methods for target Total Cholesterol.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	6	5	148	3	1	101	97	92	78
RF	5	6	**1**	6	2	**1**	2	**1**	2
Bagging	7	7	4	7	5	4	5	6	6
GB	3	2	3	2	4	3	3	2	**1**
LinR	**1**	**1**	6	**1**	**1**	6	6	5	4
kNN	6	5	7	5	8	7	7	7	7
DT	8	8	8	8	3	8	8	8	8
SVR	2	3	2	3	7	5	4	3	3
VR	4	4	5	4	6	2	**1**	4	5

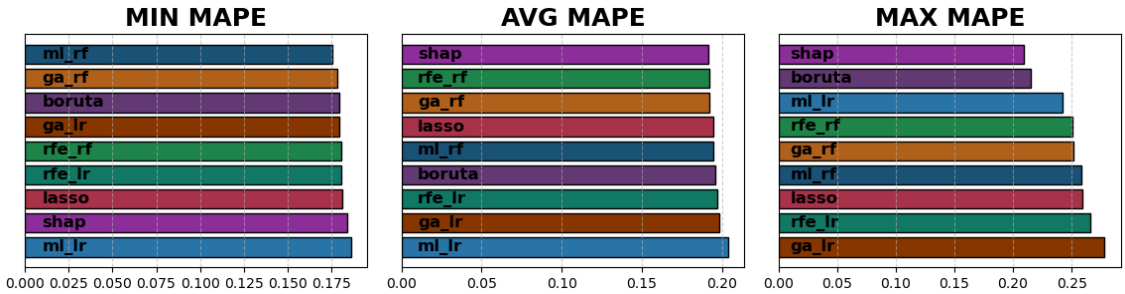


Figure 6.10: Comparison of MAPE on subsets of features selected by different methods for target Total Cholesterol.

Table 6.10 summarizes the final average rankings of the algorithms for the cholesterol-related targets and glucose. RF achieved the best overall performance, obtaining the lowest average rank across all targets. GB and LinR followed closely, showing strong and consistent results, particularly for Glucose, where LinR ranked first. VR also performed competitively, maintaining moderate ranks across all variables. In contrast, DT, kNN, and SVR exhibited higher average rankings, indicating comparatively weaker performance. Overall, RF demonstrated the most stable and accurate results across cholesterol and glucose predictions.

Table 6.10: Final Average Rankings for Cholesterol (Total, LDL, HDL) and Glucose.

Algorithm	Total Cholesterol	LDL	HDL	Glucose	Overall Rank
RF	2.9	2.9	**2.4**	2.2	**2.6**
Bagging	5.7	5.7	3.8	4.3	4.9
GB	**2.6**	**2.4**	3.4	3.7	3.0
LinR	3.4	3.4	4.3	**1.9**	3.3
kNN	6.6	6.9	6.0	6.5	6.5
DT	7.4	7.4	5.2	6.9	6.7
SVR	3.6	4.2	7.8	6.6	5.6
VR	3.9	2.8	3.0	3.8	3.4

6.2.3 Findings from Single-Target Static Question Ordering Method

This section focuses on the evaluation of single-target static question ordering methods as described in Section 5.1.1. Unlike previous sections that relied on the PROMETHEE multi-criteria decision-making framework, we now apply more commonly used evaluation metrics such as mean absolute error (MAE) and mean absolute percentage error (MAPE). This approach allows for clearer interpretation and direct comparison, especially as we narrow the scope to individual targets and a smaller set of models.

As in the previous section, results are shown only for one target per dataset: target Fat for the ESFFFQ and target Total cholesterol for NHANES. The remaining results are in Appendix A.

ESFFFQ

Figure 6.11 compares equidistant, frequency-based, and amounts-based feature representations using Logistic Regression, Random Forest, and Gradient Boosting. In all ML-based

settings, the same models are used consistently in a pairwise manner: each algorithm is employed both to derive feature rankings and to evaluate predictive performance on progressively revealed feature subsets. This ensures that differences between rankings reflect the behaviour of the ranking strategy rather than mismatches between ranking and prediction models. In addition to ML-based feature rankings, the figure includes a baseline corresponding to the original FFQ question order (grey curves), which serves as a reference for an unoptimised feature sequence. For this FFQ-based ordering, prediction is performed using Linear Regression (for regression tasks) or Logistic Regression (for classification tasks), reflecting standard practice in FFQ analysis and ensuring comparability across feature subsets.

For classification, amount-based features achieve the highest training performance, particularly with Random Forest and Gradient Boosting, but exhibit clear signs of overfitting. Across all feature representations, the FFQ-based ordering consistently underperforms ML-based rankings, requiring more questions to achieve comparable F1 scores. For regression, the lowest MAE is achieved with Linear Regression on frequency-based features, whereas Random Forest again exhibits substantial overfitting, particularly for amount-based features. In all cases, FFQ-based ordering yields higher errors for a given number of questions, confirming that the original questionnaire order is suboptimal relative to data-driven ranking approaches.

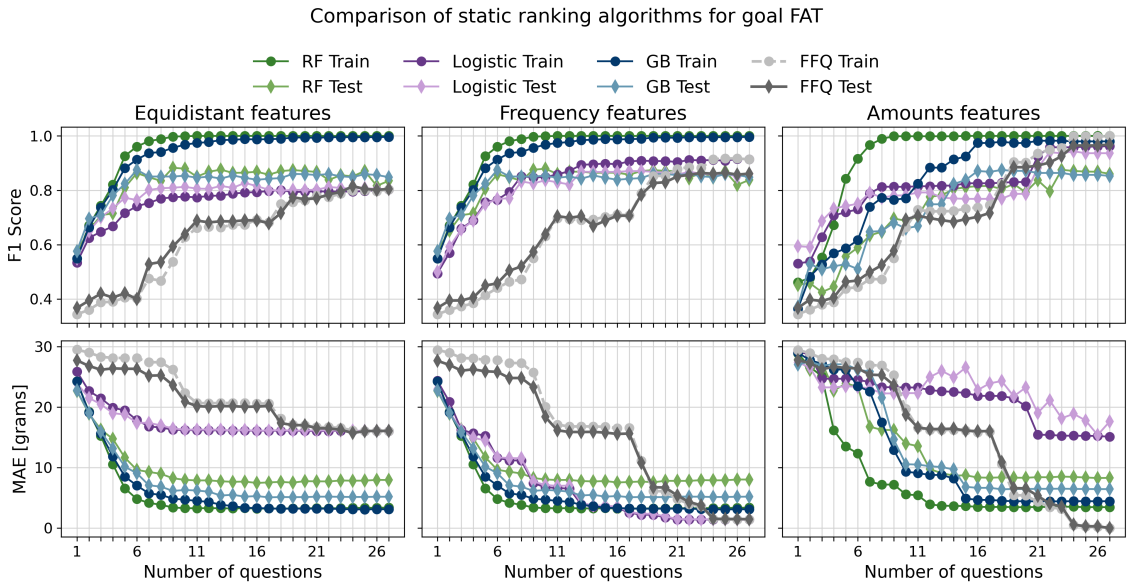


Figure 6.11: Comparison of static ranking algorithms for FAT: classification performance (F1 Score, top row) and regression performance (MAE in grams, bottom row). Column titles denote the feature representations.

Figure 6.12 focuses exclusively on ML-based feature rankings and therefore does not include the FFQ question order, as the previous figure already demonstrates its inferior performance. Logistic Regression performs best on classification tasks with amount-based features but poorly on other representations. Random Forest achieves the lowest MAE with frequency-based features, whereas amount-based features yield higher regression errors. Both Random Forest and Gradient Boosting show signs of overfitting and behave similarly for equidistant and frequency-based features.

Overall, frequency-based features are the most effective for regression. However, when

treating the problem as a classification task, this either leads to overfitting (as seen with Random Forest) or to poor performance on the training set, which degrades further during testing.

Similar conclusions were drawn across all targets, so using a frequency-based representation with Linear/Logistic Regression (to avoid overfitting) in subsequent experiments is a reasonable choice.

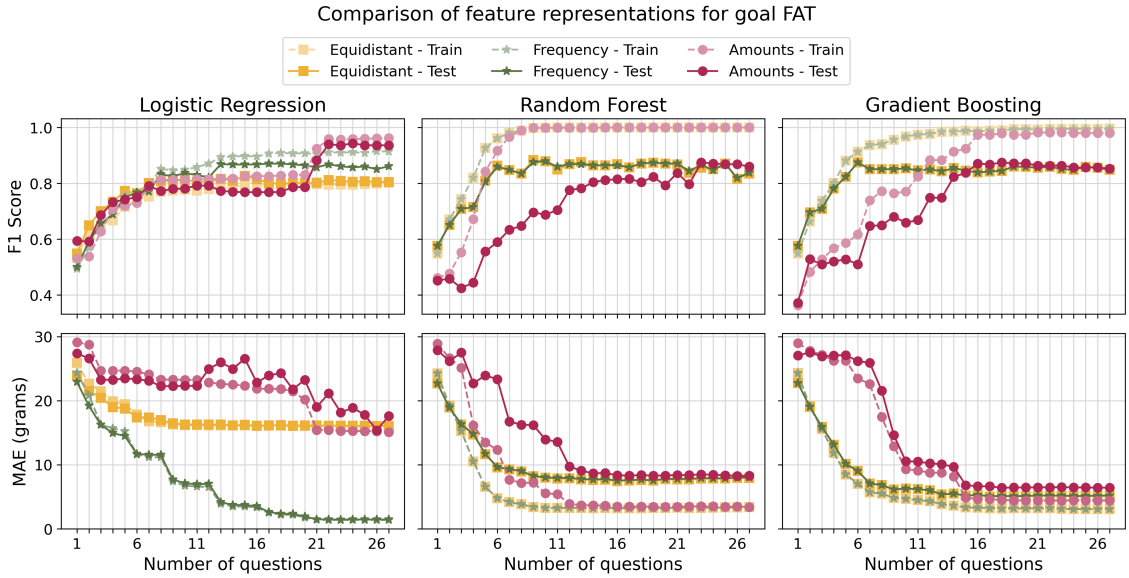


Figure 6.12: Comparison of feature representations for FAT: classification performance (F1 Score, top row) and regression performance (MAE in grams, bottom row).

NHANES

Figures 6.13 and 6.14 illustrate the regression performance of different models and feature selection strategies for predicting total cholesterol levels. The performance is evaluated using the Mean Absolute Error (MAE) metric across varying numbers of selected features.

Figure 6.13 compares the effectiveness of static ranking algorithms based on two feature selection strategies: equidistant features and frequency-based features. The results are presented for three regression models: Linear Regression (LR), Random Forest (RF), and Gradient Boosting (GB). The top row displays performance trends when incorporating all available features, while the bottom row focuses on the 20 most informative features. The results indicate that RF and GB consistently outperform LR, particularly when fewer features are used. The frequency-based feature selection approach significantly improves the performance of RF, achieving a lower MAE as the number of features increases.

Figure 6.14 further examines the impact of feature representation using regression model. The top row shows MAE trends when using all available features, while the bottom row highlights performance with only the top 20 features. The results confirm that RF benefits the most from feature selection, demonstrating a sharp MAE reduction when using the top-ranked features, and GB also exhibits strong performance improvements. However, both are heavily overfitting on the train set, while LR remains less sensitive.

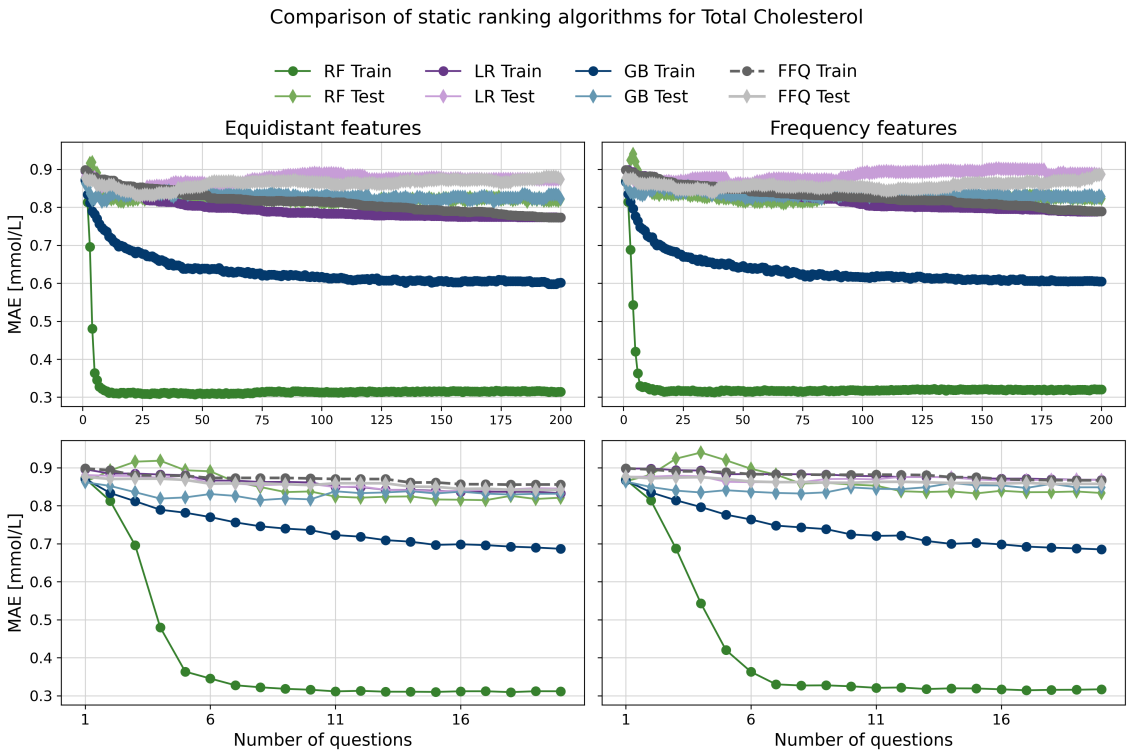


Figure 6.13: Comparison of static ranking algorithms for total cholesterol: regression performance MAE (mmol/L) for all available features and for 20 best features.

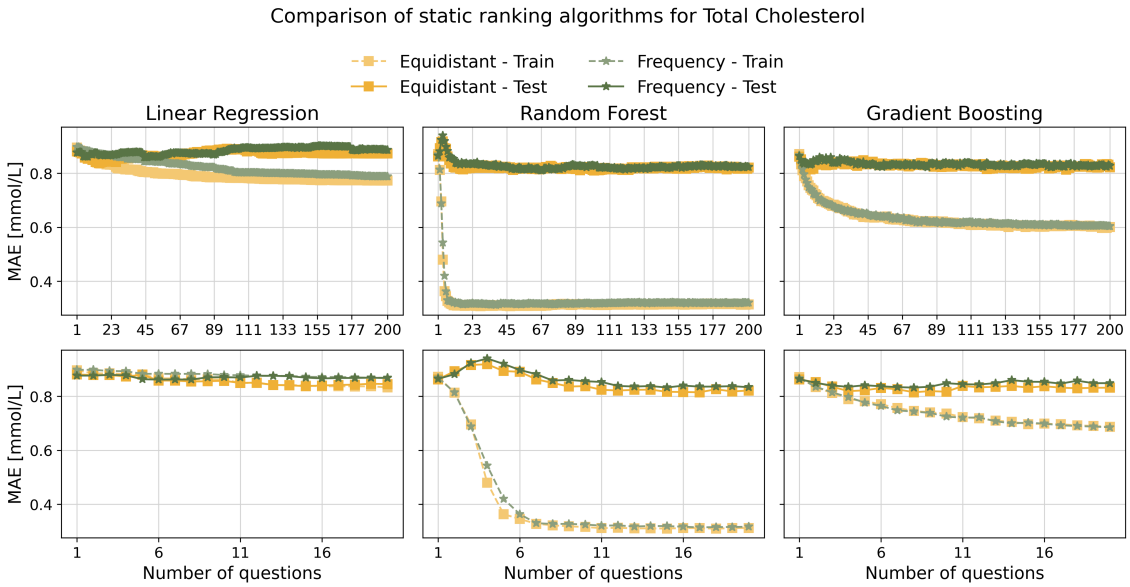


Figure 6.14: Comparison of feature representations for Total cholesterol: regression performance MAE (mmol/L) for all available features and for 20 best features.

6.2.4 Findings from Multi-Target Static Question Ordering Method

This section presents the evaluation of multi-target static question ordering methods as described in section 5.1.2. As in the previous section, we move away from PROMETHEE and focus on the standard evaluation metric of mean absolute error (MAE), which allows for clearer interpretation and more direct comparison.

We compare performance under both unweighted and weighted configurations of target importance. The analysis includes three approaches: (1) targets-specific single-target static question ordering applied independently for each dietary target, (2) aggregated static ordering where feature rankings are first averaged across all targets and then used for prediction, and (3) multi-target static question ordering optimised jointly across all targets. For each method, we report the average MAE to evaluate performance in a multi-target prediction setting.

ESFFFQ

In the unweighted multi-target setting, the differences between the three evaluated strategies — target-specific static question ordering, aggregated ranking, and multi-target ranking are minimal. As shown in Figure 6.15 and Figure 6.16, all three approaches achieve very similar mean absolute errors (MAEs) across targets. This indicates that, in the absence of target prioritisation, all strategies perform comparably in terms of overall predictive accuracy.

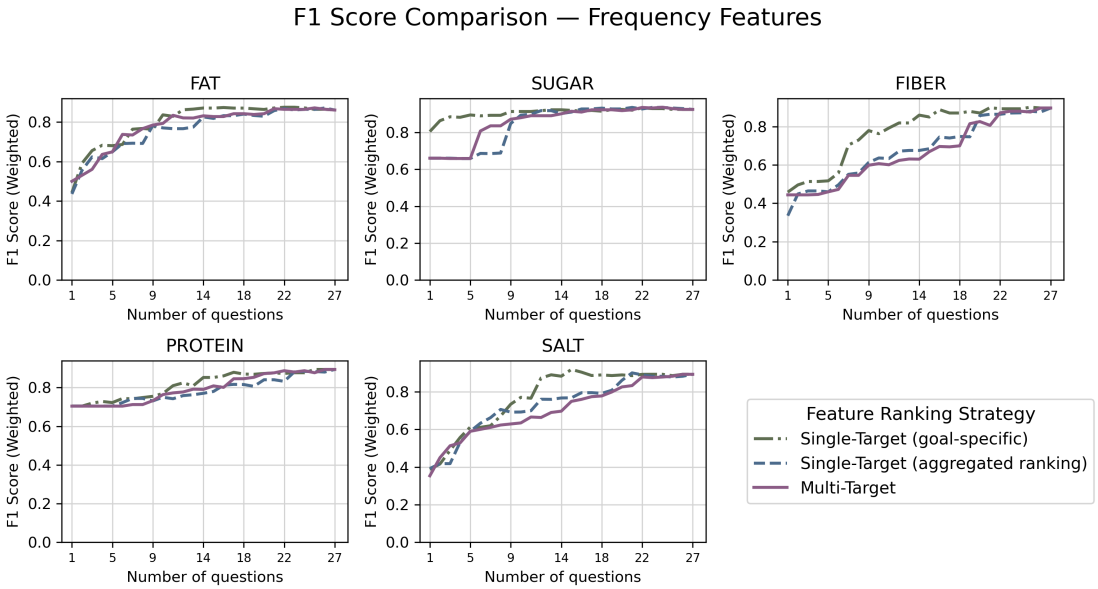


Figure 6.15: Comparison of feature ranking approaches for classification performance F1-score for all available targets.

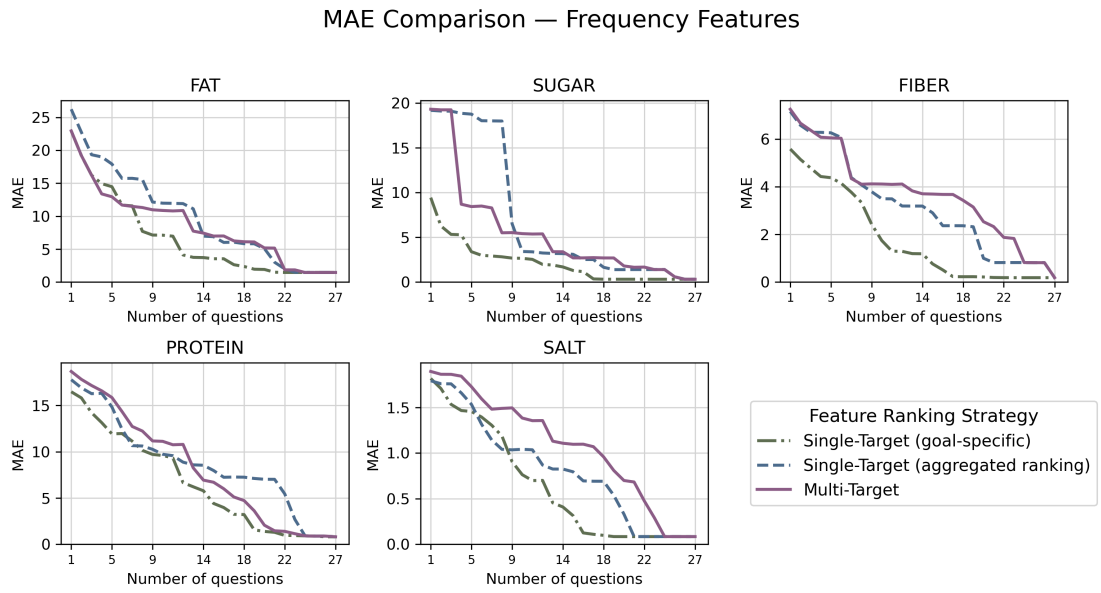


Figure 6.16: Comparison of feature ranking approaches for regression performance MAE (g/day) for all available targets.

Weighted targets

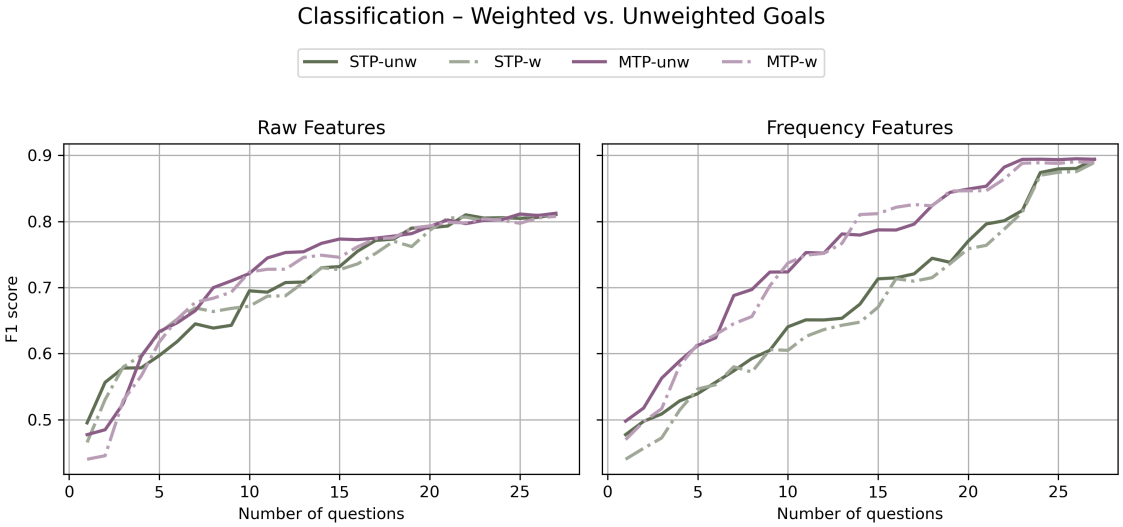


Figure 6.17: Purple curves show single-target prediction (STP) rankings; green curves show multi-target prediction (MTP). Solid lines correspond to equal target weighting; dash–dot lines use weighted targets. The left panel shows results with raw features, and the right with frequency-transformed features.

To investigate the effect of target prioritisation, we additionally evaluate multi-target static question ordering under weighted targets. The target weights used in these experiments are assigned randomly. Exhaustively exploring all possible weighting configurations is in-

feasible due to the combinatorial number of weight assignments, particularly as the number of targets grows. Instead, this experiment is designed to test whether the proposed multi-target framework remains effective under arbitrary, non-uniform target importance.

The results show that, even under randomly assigned weights, multi-target ordering consistently outperforms both single-target and unweighted multi-target strategies. As illustrated in Figure 6.17 and Figure 6.18, weighted multi-target ordering achieves lower error values and higher predictive performance across a range of feature counts. This effect is especially pronounced when using frequency-based feature representations.

These findings demonstrate that the proposed multi-target approach is robust to uneven and non-uniform target importance and can effectively balance predictive performance across multiple outcomes, even when target priorities are not optimally chosen. This supports the use of weighted multi-target ordering as a flexible and practical strategy in scenarios where different dietary outcomes must be prioritised.

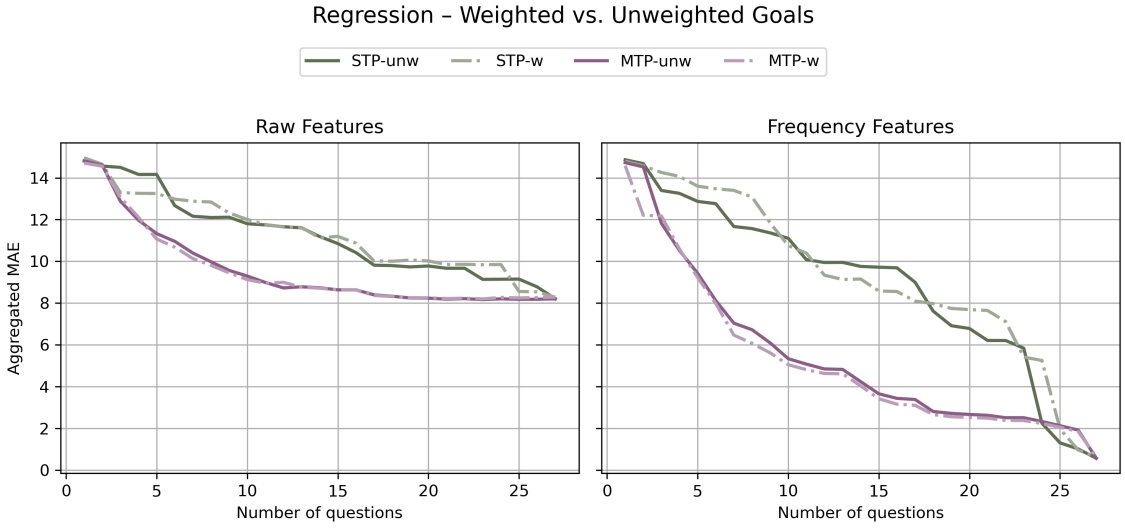


Figure 6.18: Purple curves show single-target prediction rankings; green curves show multi-target prediction. Solid lines use equal-target weighting; dash-dot lines use weighted targets. The left panel shows results with raw features, the right with frequency-transformed features. Lower curves indicate more accurate nutrient predictions (MAE).

Runtime Analysis

Table 6.11 compares the runtime of single-target prediction (STP) and multi-target prediction (MTP) for regression and classification using feature sets `X` and `X_freq`. In contrast to earlier experiments that combined multiple model types, the results shown here isolate linear models in STP and Lasso-based models in MTP, enabling a clearer interpretation of computational behaviour.

For regression, STP is faster than MTP. Each single-target Linear Regression model is inexpensive to train, and the total STP runtime grows linearly with the number of targets. MTP regression, in contrast, relies on a single multi-output Lasso model with cross-validation, which introduces additional optimisation overhead despite fitting only one model. For classification, the opposite trend is observed. STP classification is substantially slower, particularly for the raw feature set `X`. This is primarily due to Logistic Regression, which requires iterative numerical optimisation and must be trained separately for each

target. As a result, the overall STP classification time increases with the number of targets.

MTP classification uses a single Lasso-based multi-target model, so the optimisation cost is incurred only once for all targets. Consequently, MTP classification remains computationally stable as the number of targets grows. Using frequency-transformed features reduces runtimes for both approaches, but the relative trends remain consistent.

Table 6.11: Timing Comparison for Feature Sets `X` and `X_freq`.

Metric	**X (s)**	**X_freq (s)**
MTP Regression	0.63	0.46
MTP Classification	0.66	0.54
STP Total Regression	0.15	0.13
STP Total Classification	2.71	0.43
Per-target STP Times (Regression | Classification)		
Fat	0.06 | 0.53	0.03 | 0.09
Sugar	0.02 | 0.61	0.02 | 0.08
Fiber	0.02 | 0.53	0.04 | 0.09
Protein	0.02 | 0.50	0.03 | 0.07
Salt	0.02 | 0.55	0.02 | 0.09

NHANES

In the unweighted multi-target setting for NHANES, the differences among the three evaluated strategies - target-specific static question ordering, aggregated ranking, and multi-target ranking—are minimal. As shown in Figure 6.19, all three approaches achieve very similar mean absolute errors (MAEs) across targets. This suggests that, in the absence of target prioritization, all strategies perform comparably in terms of overall predictive accuracy. However, in NHANES, the prediction error initially decreases but then increases once all questions are included (top row). When considering the top 20 features across all targets, the errors are nearly identical, though the multi-target approach slightly outperforms the aggregated STP.

Weighted targets

Figure 6.20 compares prediction performance across different feature selection strategies and target-weighting schemes for nutrient regression using the NHANES dataset. The plots show aggregated mean absolute error (MAE) as a function of the number of selected questions (features). Purple curves represent STP rankings, while green curves show MTP rankings. Solid lines correspond to unweighted (equal-target) settings, and dash-dot lines indicate target-weighted settings.

The top row illustrates performance when up to 200 features are considered, while the bottom row focuses on the top 20 features. The left column shows results with raw feature values, and the right column uses frequency-transformed features. In the unweighted setting (solid lines), all methods yield similar MAEs, particularly in the top 20 feature range. The MTP runtimes are consistent with the corresponding single-target timings, in the sense that per-target STP costs remain comparable across targets, while MTP avoids repeating the full training and evaluation cycle separately for each target.

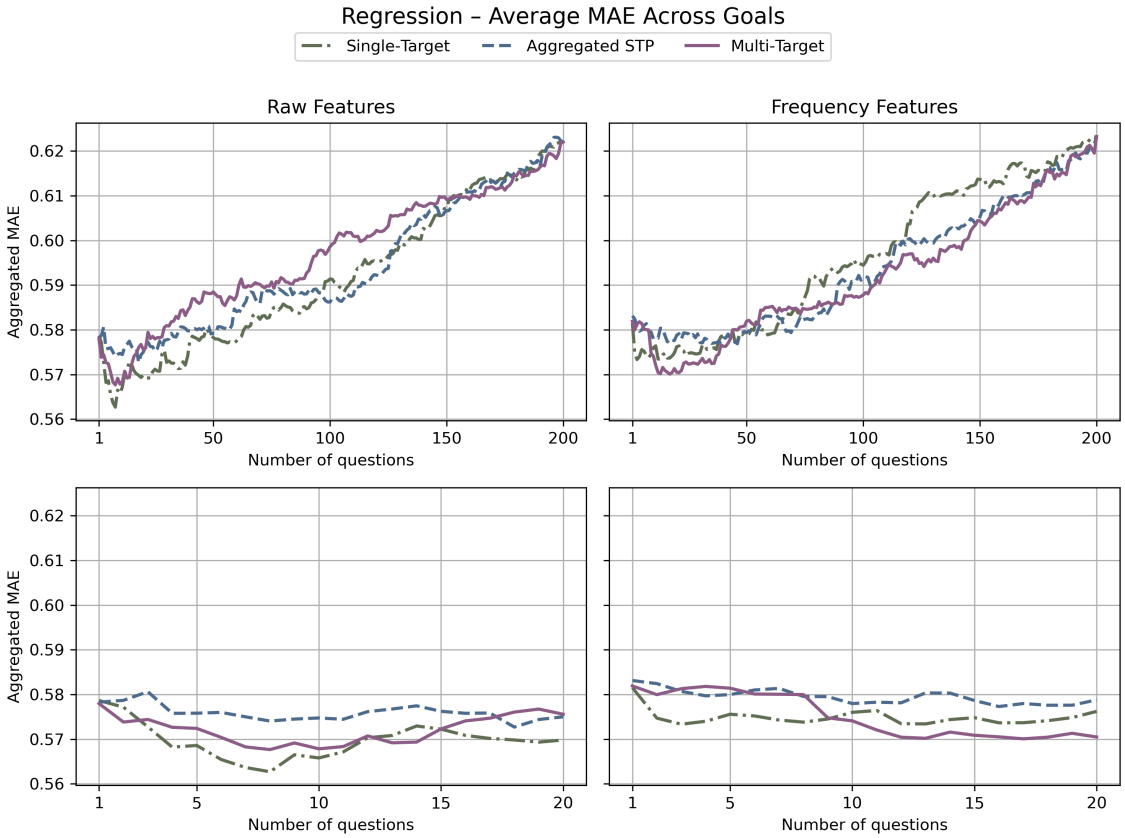


Figure 6.19: Comparison of feature ranking approaches for regression performance MAE (mmol/L) averaged over all available targets.

Runtime Analysis

Table 6.12: Regression Timing Summary for Feature Sets `X` and `X_freq`.

Metric	**X (s)**	$X_{\mathbf{freq}}$ **(s)**
MTP Regression Time	17.39	28.49
STP Total Regression Time	0.37	0.48
Per-target STP Times (Regression)		
HDL Direct (mmol/L)	0.11	0.12
Total Cholesterol (mmol/L)	0.08	0.12
Glucose (mmol/L)	0.09	0.11
LDL Direct (mmol/L)	0.10	0.13

Table 6.12 presents a runtime comparison between multi-target prediction (MTP) and single-target prediction (STP) for two feature representations: the original input features (X) and frequency-transformed features (X_{freq}). In contrast to earlier experiments, MTP is substantially slower in this setting, requiring 17.39 seconds for X and 28.49 seconds for X_{freq}, while the total STP regression time remains below half a second for both feature sets. The lower part of the table reports per-target STP runtimes, showing that each individual regression model requires approximately 0.08–0.13 seconds. Since STP models

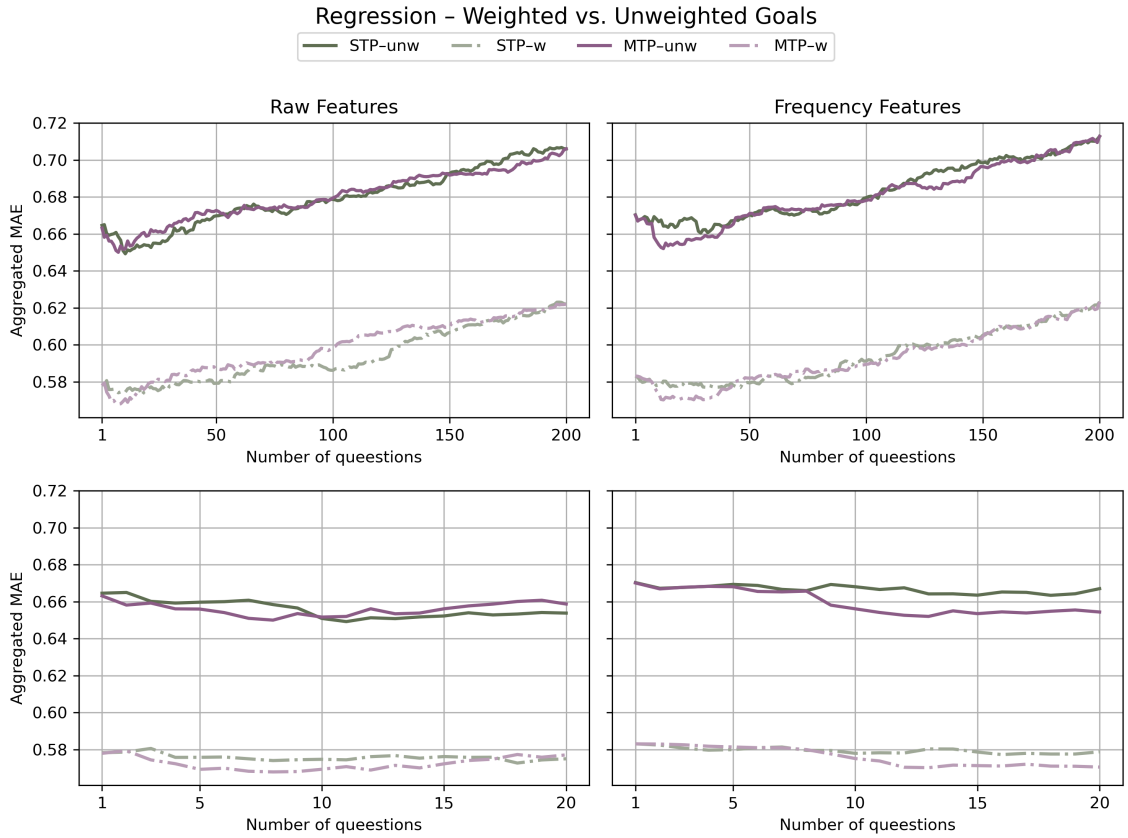


Figure 6.20: Purple curves show STP rankings; green curves show MTP. Solid lines use equal-target weighting; dash-dot lines use weighted targets. The left panel shows results with raw features, the right with frequency-transformed features. Lower curves indicate more accurate nutrient predictions (MAE).

are trained independently and use simple linear regression, the total STP runtime scales almost linearly with the number of targets and remains low in absolute terms.

As with ESFFFQ, the results indicate that the relative runtime of MTP and STP depends strongly on the underlying model class and optimisation procedure, rather than on the question-ordering framework itself. While MTP offers conceptual advantages in joint modelling and target coupling, its computational cost can exceed that of STP when more complex regularised models are employed. This observation underscores the importance of considering model choice and optimisation strategy when deploying multi-target dynamic question ordering in practice.

6.2.5 Findings from Single-Target Dynamic Question Ordering

In the dynamic question ordering algorithm (Algorithm 5.1), the next best feature is selected based on its expected utility value $\mathbb{E}(U_f)$, which is determined by the utility function $U(_{f:=r})$. Several options are available for defining the utility function, and some of these were explored in this thesis. In certain cases, the cost of acquiring a feature may also be considered as part of the utility function. However, in the FFQ problem, all questions carry the same cost, so alternative options had to be explored.

The expression in Algorithm 5.1, $p(f = r \mid x_{\mathcal{F}_B}) \log p(f = r \mid x_{\mathcal{F}_B})$, is built upon the

idea of conditional entropy but utilises machine learning to approximate chain probability calculations. We refer to this approach as Quasi Entropy (QE).

ESFFFQ

Figure 6.21 compares the regression performance—measured as Mean Absolute Error (MAE) in grams—for predicting the intake of five nutrients: fat, sugar, fiber, protein, and salt. The performance is shown for static ranking (on both train and test set) and four dynamic ranking methods: MPS, NC, MPSNC, and MUR. The accompanying table reports the average inference time for each method, highlighting the computational efficiency of MUR.

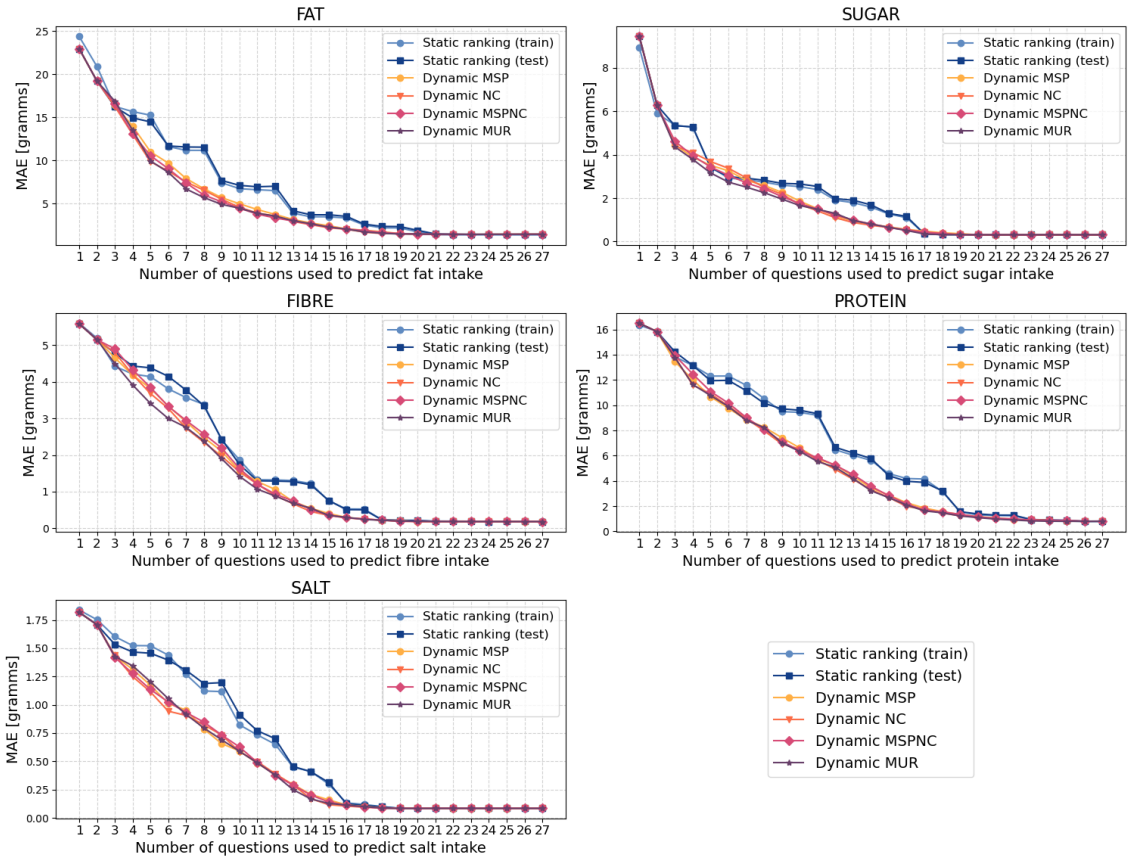


Figure 6.21: Regression performance (MAE in grams) for static ranking (train and test) and dynamic ranking methods.

Across all five nutrients, dynamic ranking methods outperform static ranking, particularly in the early stages of question acquisition. While static ranking (test) eventually converges to similar MAE values after more than 15 questions, the dynamic approaches—especially MUR and NC—achieve lower error rates with fewer questions. For fat intake, dynamic MUR consistently achieves the lowest MAE across almost all question counts, followed closely by NC and MPS. Static methods lag, especially in the early and mid-range stages. The trend is similar for sugar and fiber, where dynamic methods show rapid improvement in MAE within the first 10 questions. Again, MUR leads in accuracy

and is closely followed by NC, while static test performance improves more gradually. In the case of protein and salt, the dynamic methods outperform static ones even more clearly. MUR achieves the lowest error across nearly all question counts, showing its robustness and adaptability across nutrient types.

	MSP	NC	MSPNC	MUR
Fat	3.0 s	12.0 s	24.5 s	38.0 s
Sugar	2.9 s	11.8 s	24.8 s	36.5 s
Fiber	2.6 s	11.6 s	24.7 s	37.0 s
Protein	2.9 s	12.2 s	25.0 s	35.5 s
Salt	2.6 s	11.7 s	24.0 s	38.5 s

Table 6.13: Timing results in seconds for different targets and methods.

The runtime results in Table 6.13 further clarify the computational trade-offs between the dynamic ranking strategies. The MUR method incurs the highest inference time (approximately 35–39 seconds across targets), reflecting the additional computation required to evaluate utility-based updates at each step. In contrast, NC and MSPNC are considerably faster (approximately 11–25 seconds), while the lightweight MSP approach remains the fastest overall (approximately 2.6–3.0 seconds) due to its simplified ranking mechanism.

These results indicate that MUR does not minimise runtime, but instead trades additional computation for improved ranking quality. In practical terms, this overhead remains moderate and predictable, and may be acceptable in settings where improved early predictive performance is prioritised over minimal inference time. At the same time, further runtime reductions would likely be necessary for fully interactive or mobile deployments. One natural direction is to limit the candidate feature set considered during dynamic ordering—for example, by applying one of the feature reduction strategies introduced earlier in the chapter—thereby reducing the number of utility evaluations required at each step.

Overall, the results highlight a spectrum of dynamic ranking methods, ranging from highly efficient but simpler strategies (MSP) to more computationally demanding yet more informative approaches (MUR). This provides flexibility in choosing an appropriate balance between predictive quality and computational cost, depending on the constraints of the target application.

NHANES

Figure 6.22 presents regression performance—measured as Mean Absolute Error (MAE) in mmol/L—for four biomarkers: HDL Direct, LDL, Total Cholesterol, and Glucose. The figure compares static ranking (evaluated on both training and test sets) with two dynamic ranking approaches: dynamic MPS and dynamic MUR. Overall, the absolute MAE values remain relatively high across all methods, reflecting the inherent difficulty of predicting laboratory biomarkers using FFQ data alone. This is consistent with previous observations in this thesis and with the broader literature: dietary frequency information captures only part of the biological variability in blood-based markers, which are also strongly influenced by factors such as age, genetics, medication use, metabolic status, and other lifestyle variables not represented in the FFQ.

Within this constrained predictive setting, dynamic ranking methods still show modest but consistent advantages in the early stages of question acquisition. Across most biomarkers, both dynamic approaches outperform static ranking when only a small number of questions are available, particularly within the first 5–10 questions. For HDL Direct,

dynamic methods reduce MAE more rapidly than static ranking in the early phase. Similarly, for LDL and Total Cholesterol, dynamic MPS achieves the lowest errors among the compared methods in the first few questions, indicating that adaptive selection can prioritise more informative items when information is limited. For Glucose, early gains are also visible, with dynamic MUR achieving the lowest MAE at very small question counts. As more questions are added, the differences between methods diminish, and performance becomes more stable across all approaches. Dynamic methods also exhibit greater variability as the sequence progresses, which is expected given the weak overall signal and the sensitivity of adaptive strategies to early responses in this setting.

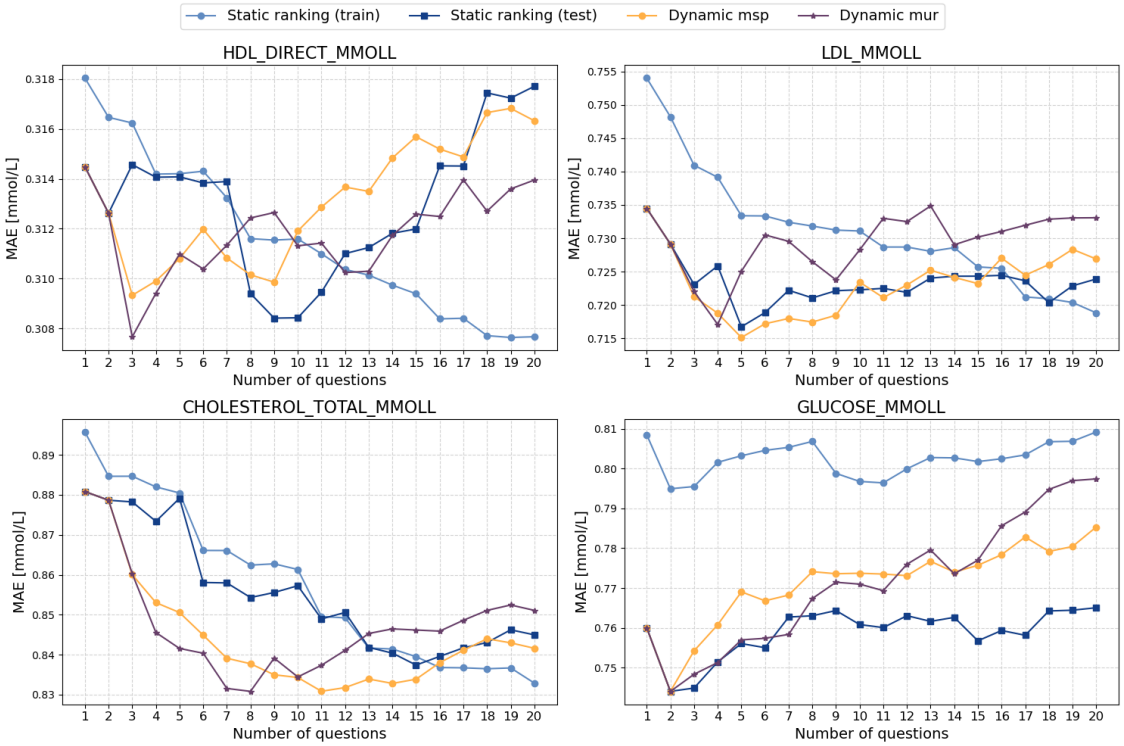


Figure 6.22: Regression performance (MAE in mmol/L) for static ranking (train and test) and dynamic ranking methods.

	MPS	MUR
Cholesterol	29.4s	23.7s
HDL cholesterol	28.9s	23.7s
LDL cholesterol	29.8s	24.6s
Glucose	30.0s	24.2s

Table 6.14: Timing for first feature selection for MPS and MUR methods.

Taken together, these results should be interpreted cautiously. They do not suggest that FFQ-based prediction of biomarkers is highly accurate; rather, they indicate that, even in a challenging, low-signal scenario, dynamic question ordering can extract at least some useful information more efficiently than static ranking when only a limited number of responses are available. This supports the use of dynamic ordering primarily as a burden-

reduction and early-information strategy, rather than as a solution to the fundamental limitations of FFQ-only biomarker prediction.

Table 6.14 reports the time to first feature selection. Across all four biomarkers, MUR outperforms MPS by approximately $5-6$ seconds, consistently selecting informative questions faster (e.g., 23.7s vs 29.4s for cholesterol). In summary, dynamic ranking methods offer early predictive accuracy and out of the two methods MUR is faster, making it a better candidate for further exploration of MTP where all targets are not of the same importance.

Comparison of Static and Dynamic Question Orders

To further analyse the effect of the proposed dynamic procedure, we compared the instance-specific dynamic rankings with the corresponding static baseline rankings. This analysis was carried out on the ESFFFQ data, where dynamic question ordering showed better performance, and for the MUR utility, which was also used later in the multi-target regression setting. Predictive performance shows whether the method improves estimation quality, but it does not directly show how much the question sequence changes from one respondent to another. For this reason, we additionally evaluated the similarity between static and dynamic question orders.

For each prediction target, the static ranking is fixed and shared across all respondents, whereas the dynamic ranking is generated separately for each validation instance. We quantified the difference between these rankings using normalised Kendall distance, which measures disagreement in the relative ordering of question pairs. A value of 0 indicates identical rankings, while larger values indicate a greater departure from the static baseline. Because the beginning of the questionnaire is especially important in partially completed FFQs, we also considered Top-5 overlap and the position of the first divergence. Top-5 overlap measures how many of the first five questions are shared between the static and dynamic rankings, while first divergence records the earliest position at which the two sequences differ.

Across all targets, the mean normalised Kendall distance was 0.167, indicating a moderate departure from the static baseline rather than a complete reordering. The mean Top-5 overlap was 0.618, showing that approximately three of the first five questions were shared on average, while the mean first-divergence position was 3.30. The smallest deviation from the static ranking was observed for fibre (0.142), and the largest for protein (0.215). Agreement among the earliest questions was highest for sugar (Top-5 overlap 0.792) and lowest for salt (Top-5 overlap 0.466). Overall, these results show that the dynamic procedure introduces meaningful respondent-specific personalisation while still preserving a substantial part of the global static structure. The results of this analysis are summarised in Table 6.15.

Target	**Kendall distance**	**Top-5 overlap**	**First divergence**
Fat	0.164	0.600	3.86
Fibre	0.142	0.608	3.45
Protein	0.215	0.626	3.19
Salt	0.157	0.466	3.00
Sugar	0.155	0.792	3.00
Overall	0.167	0.618	3.30

Table 6.15: Similarity between static and dynamic question orderings across targets.

6.2.6 Findings from Multi-Target Dynamic Question Ordering

We investigated the effectiveness of dynamic question ordering in multi-target prediction scenarios, where multiple outcomes are predicted in parallel during a single interaction. Two primary settings were explored: one where all targets are considered equally important, and another where target importance is explicitly weighted. In the equal-weight setting, the system aims to minimize the average error across all targets, treating them uniformly during feature selection.

In the weighted setting, importance values are assigned to targets either through static weights—manually defined and fixed—or through dynamic weights, which are recalculated at each step based on current prediction uncertainty or model confidence. This dynamic approach allows the system to focus its questioning strategy on more uncertain or higher-priority targets, improving predictive performance when question budgets are limited.

6.2.6.1 ESFFFQ

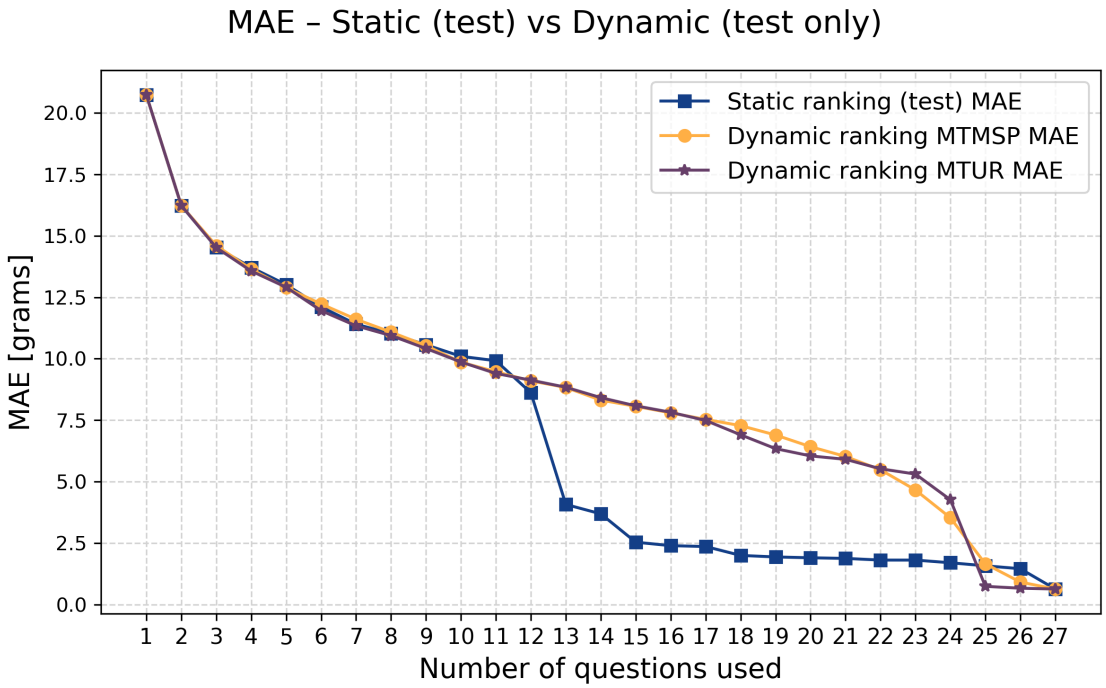


Figure 6.23: Regression performance (MAE in grams) for static ranking (train and test) and dynamic ranking methods in multi-target prediction problem.

Figures 6.23 and 6.24 show the regression performance of multi-target dynamic question ordering on the ESFFFQ dataset, where prediction accuracy is evaluated using Mean Absolute Error (MAE) in grams. These figures compare static ranking (test only) with a variety of dynamic ranking methods, under both unweighted and weighted multi-target prediction settings.

Figure 6.23 presents results for the unweighted case, comparing static ranking to two dynamic approaches: MTMPS (multi-target mean selection probability) and MTUR (multi-target uncertainty reduction). The results indicate that dynamic methods offer only a slight advantage over static ranking in the early stages, particularly within the first 10–15

questions. However, after around question 15, static ranking consistently outperforms both dynamic strategies, achieving lower MAE by the end of the sequence. While dynamic approaches show potential for early-stage efficiency, they are less stable as more questions are asked, and do not maintain their initial advantage in the long run.

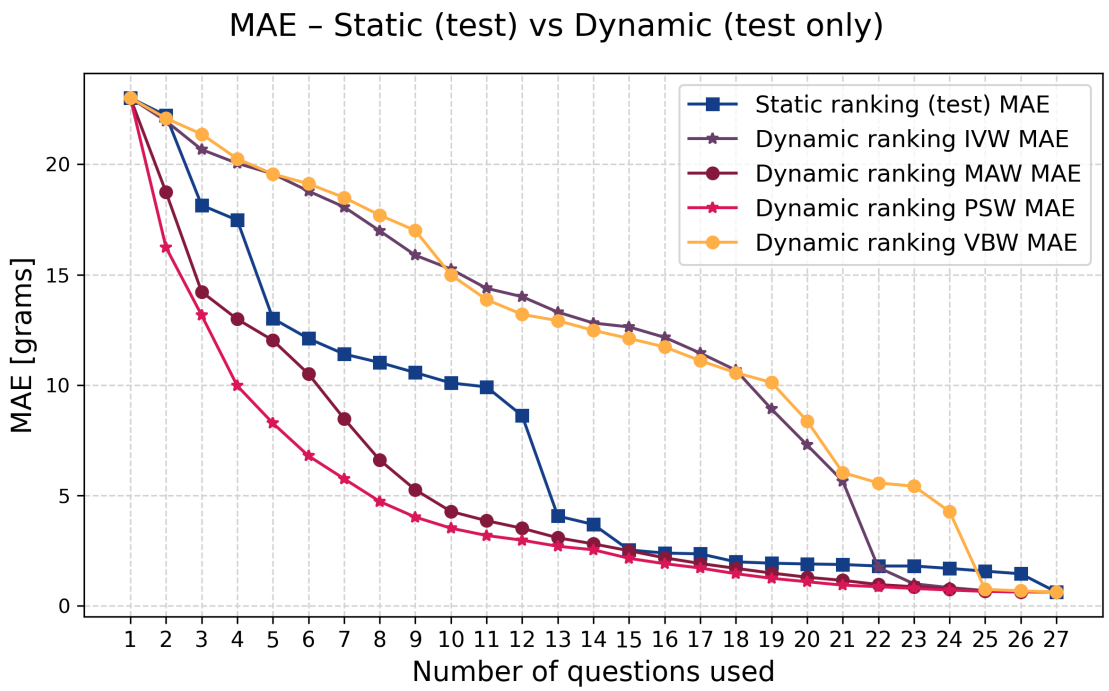


Figure 6.24: Regression performance (MAE in grams) over all targets for static ranking and different weighted utility functions in dynamic ranking in multi-target prediction problem.

In contrast, 6.24 explores the weighted multi-target prediction setting, where MTUR incorporates different weighting strategies to prioritize targets. These include manually assigned weights (MAV), which are set statically before the algorithm is run, and dynamically calculated weights – inverse variance weighting (IVW), prediction sensitivity-based weights (PSW), and variance-based dynamic weighting (VBW). Among these, PSW performs best overall, showing clear improvements over static ranking throughout most of the question range. MAV, in particular, achieves lower MAE in the early and middle stages and remains competitive until the final questions. VBW demonstrates more gradual improvement, converging with the best-performing methods only after 20 questions. IVW performs poorly across most of the sequence, especially in the early stages.

These findings suggest that while unweighted dynamic multi-target prediction offers limited benefits, incorporating target-specific weighting strategies can significantly enhance performance. The superior results of MAV and PSW demonstrate that introducing prior knowledge or error-based prioritization allows the model to ask more informative questions earlier, reducing the number of questions needed to reach a desired accuracy. In practical applications, where minimizing interaction length is essential, weighted dynamic methods—particularly MAV—offer the most promise for efficient and accurate multi-target dietary intake estimation.

Table 6.16: Average computation time per feature for different utility functions.

Metric	MTMPS	MTUR
Time per feature (s)	11.38	189.53

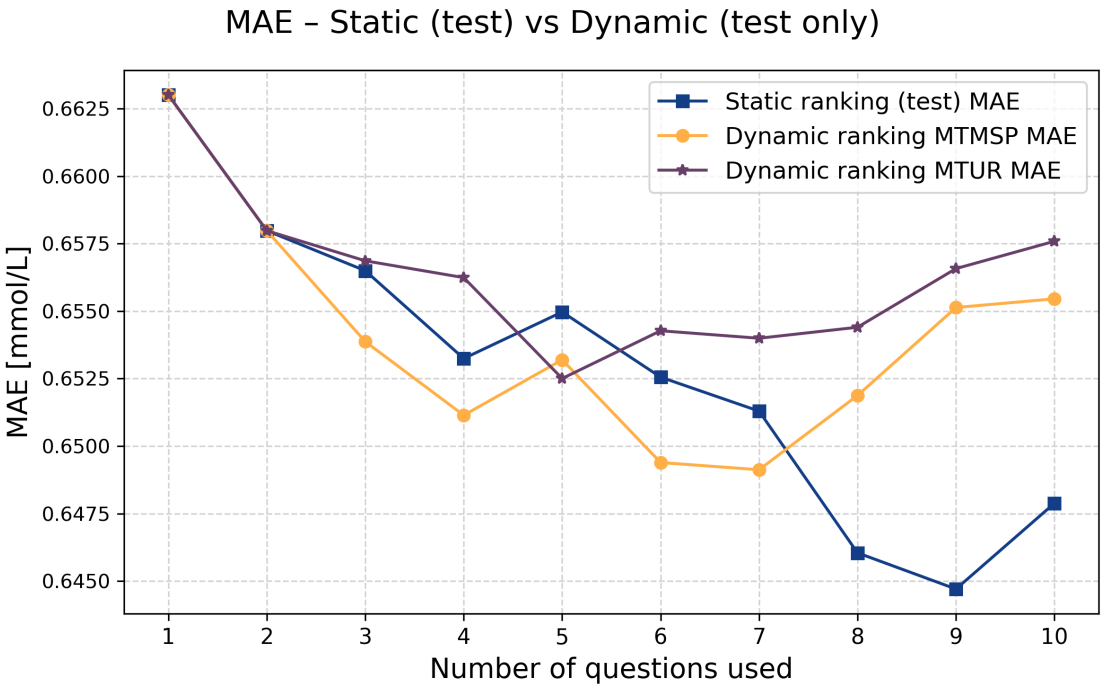


Figure 6.25: Regression performance (MAE in mmol/L) for static ranking (train and test) and dynamic ranking methods in multi-target prediction problem.

6.2.6.2 NHANES

Figures 6.25 and 6.26 illustrate the regression performance of multi-target dynamic question ordering on the glucose dataset, where prediction accuracy is evaluated using Mean Absolute Error (MAE) in mmol/L. Both figures compare static ranking (test only) against different dynamic ranking methods, under unweighted and weighted prediction settings.

Figure 6.25 shows results for the unweighted case, where static ranking is compared with two dynamic strategies: MTMPS (multi-target mean selection probability) and MTUR (multi-target uncertainty reduction). Across the first ten questions, MTMPS consistently outperforms static ranking, achieving lower MAE in the early and middle stages. MTUR, however, performs less effectively, offering only minor improvements initially and lagging behind both MTMPS and static ranking after the fifth question. Overall, static ranking eventually converges toward the best performance, but MTMPS demonstrates a clear advantage in efficiency during the initial stages.

In contrast, Figure 6.26 examines the weighted multi-target prediction setting, where MTUR incorporates alternative weighting schemes: equal weights (EW), inverse variance weighting (IVW), manually assigned weights (MAW), prediction sensitivity-based weights (PSW), and variance-based weighting (VBW). Among these, PSW achieves the most consistent gains, outperforming static ranking across nearly the entire sequence of questions. Both MAW and VBW also perform competitively, with MAW providing an early advantage

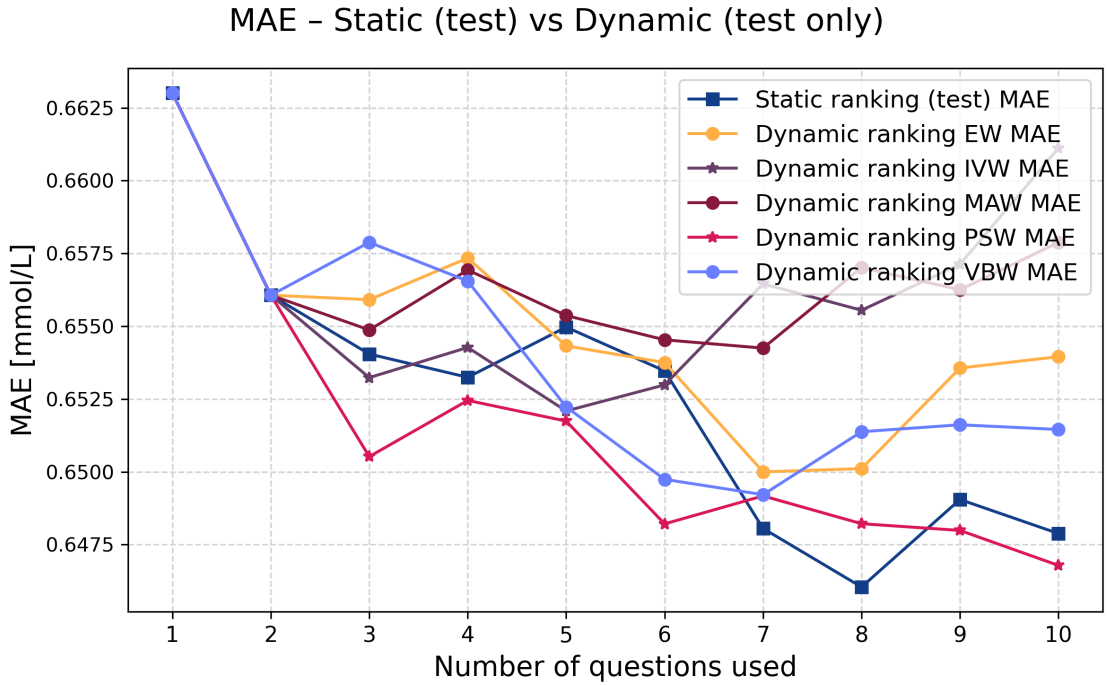


Figure 6.26: Regression performance (MAE in mmol/L) over all targets for static ranking and different weighted utility functions in dynamic ranking in MTP.

and VBW showing stronger improvements in later questions. IVW, by contrast, performs relatively poorly throughout, failing to match the accuracy of the other methods.

Together, these findings indicate that while unweighted dynamic approaches like MTMPS can offer early improvements, incorporating informed weighting strategies—particularly PSW—substantially enhances predictive efficiency. Weighted dynamic methods enable the model to prioritize more informative questions earlier in the process, reducing overall error more quickly than static ranking and thereby supporting more efficient and accurate multi-target glucose prediction.

Chapter 7

Conclusion

7.1 Summary of Findings

7.1.1 Conclusion and Contributions to the Field

This research set out to investigate how machine learning can improve the effectiveness and efficiency of Food Frequency Questionnaires (FFQs) through three core hypotheses:

- **Hypothesis 1** posited that FFQs could be significantly shortened using machine learning while preserving or even improving predictive accuracy relative to standard methods for handling missing data. This hypothesis had two components: selecting a reduced subset of informative questions and achieving comparable or superior predictive performance on that subset. The hypothesis is supported by the results presented in Section 6.2.1 and Section 6.2.2.
- **Hypothesis 2** proposed that ordering questions by their predictive importance would improve dietary estimates derived from partially completed FFQs. The assumption was that presenting the most informative questions earlier would mitigate the impact of incomplete responses. This hypothesis is supported by the results in Section 6.2.3.
- **Hypothesis 3** introduced a dynamic element, suggesting that adaptive question ordering—where each subsequent question is selected based on previously provided answers—would further improve predictive performance over static question sequences. This hypothesis is supported by the results in Section 6.2.5 and Section 6.2.6.

Empirical results from extensive experiments using the ESFFFQ dataset support all three hypotheses. When the methods were applied to the NHANES dataset, the observed effects were less pronounced, although the overall trends remained consistent. This difference does not primarily reflect a limitation of the proposed methods, but rather highlights the inherent difficulty of predicting laboratory-based biomarkers from self-reported FFQ data alone. Biomarker outcomes capture physiological processes that are only indirectly related to dietary intake and are influenced by many additional factors not present in FFQs. These results suggest that accurate prediction of laboratory measures is unlikely to be achievable using FFQ data in isolation, and that future improvements will likely require the integration of complementary information such as demographic variables, anthropometric measures, lifestyle factors, or other health-related data.

Despite these challenges, machine learning models successfully identified compact sets of questions that achieved predictive performance comparable to, or better than, full-length FFQs. Prioritising questions based on their predictive importance consistently

improved accuracy in scenarios with partial responses. Furthermore, dynamically adapting the question sequence in real time yielded the most accurate predictions, particularly when the level of response completeness varied.

A key novel contribution of this work is the development and evaluation of a *multi-target dynamic question ordering strategy*, which tailors the sequence of questions based on both previous answers and multiple predictive targets simultaneously. When combined with *dynamically adjusted target weights*, this approach consistently outperformed static multi-target orderings. This result demonstrates that adaptivity not only at the question level but also at the goal-weighting level leads to more accurate and efficient dietary assessments, especially in settings with diverse prediction objectives.

Together, these findings lead to several important contributions:

1. **Data-Driven Questionnaire Reduction**
   This thesis provides a validated framework for using supervised learning to design reduced FFQs that retain predictive power. This contributes a scalable framework for reducing respondent burden and improving compliance in dietary surveys.

2. **Resilience to Missing Data via Static Question Prioritisation**
   By demonstrating the benefit of ordering questions by predictive relevance, this work introduces a robust strategy for mitigating data loss in incomplete surveys. This improves the quality of dietary assessments in settings where full participation cannot be guaranteed.

3. **Dynamic, Personalised Questionnaire Design**
   The integration of dynamic question ordering establishes an adaptive model for FFQs, where the information value of each subsequent question is conditioned on prior responses. This innovation leads to more efficient and personalised data collection and sets the foundation for intelligent, real-time survey systems.

4. **Multi-Target Adaptive Learning with Weighted Objectives**
   The proposed multi-target dynamic ordering approach, particularly when paired with dynamic target reweighting, introduces a new paradigm for dietary assessment design. This combination enhances prediction quality across multiple outcomes and accommodates shifting goal priorities, offering a flexible tool for complex, multi-objective survey applications.

5. **Generalisable Methodological Innovations**
   Beyond dietary assessment, this work contributes transferable machine learning methods—including question ranking, dynamic selection, and multi-target prediction—that can be adapted to a wide range of domains requiring efficient and adaptive survey-based data collection, such as public health monitoring, behavioural research, and clinical screening.

In summary, this thesis demonstrates that FFQs can be made shorter, more robust to missing data, and more adaptive—without sacrificing predictive performance. The results support the development of intelligent, user-centred tools for dietary self-assessment, personalised nutrition interventions, and mobile health applications. More broadly, the methods presented here inform the design of adaptive and predictive questionnaires across domains where efficient and individualised data collection is essential.

7.1.2 Practical Recommendations

The findings of this research translate into several actionable recommendations for the design and deployment of FFQs and other survey-based tools, particularly in digital health and mobile settings:

1. **Leverage Machine Learning to Shorten Questionnaires**
   Supervised learning methods can identify the most informative questions, enabling shorter FFQs with minimal loss in predictive accuracy. This trade-off is particularly favourable in mobile and self-administered contexts, where user engagement and completion rates are critical.

2. **Order Questions to Maximise Accuracy under Incomplete Responses**
   When survey completion cannot be guaranteed, ordering questions by predictive importance ensures that the most informative responses are collected early, improving robustness to missing data.

3. **Adopt Dynamic, Adaptive Question Ordering**
   Dynamically selecting questions based on previously observed answers improves prediction accuracy and personalises the assessment process. This approach is especially well suited to mobile health applications.

4. **Combine Multi-Target Learning with Dynamic Weighting**
   For problems involving multiple outcomes, combining multi-target prediction with dynamic target weighting allows questionnaires to adapt to shifting priorities and improves overall predictive performance.

5. **Extend Methods Beyond Nutrition**
   The proposed techniques are generalisable and can be applied to other survey-based domains, including behavioural assessment, mental health screening, and clinical triage systems.

7.1.3 Future Work

While this research demonstrates the value of machine learning for optimising FFQs through dynamic and multi-target approaches, several directions for future investigation remain open. One important avenue is improving the computational efficiency of dynamic question ordering. Although effective, the current approach can be computationally demanding, particularly in multi-target settings. Developing approximate or heuristic variants could make real-time deployment more practical, especially on resource-constrained devices.

Another promising direction is to integrate cost-awareness into the multi-target framework. Although all questions in FFQs have a similar nominal cost, factors such as cognitive effort, response time, or dropout risk could be incorporated to further improve usability. Extending cost-sensitive formulations to the multi-target setting remains an open challenge.

Future work should also explore integrating additional data sources. The NHANES results suggest that predicting biomarker outcomes from FFQs alone is inherently limited. Incorporating demographic information, anthropometric measures, lifestyle variables, or longitudinal data may substantially improve performance for such targets.

Another direction for future work is to test the proposed methods in real user-facing settings. In this thesis, the evaluation was performed offline, where partial completion was simulated from existing questionnaire data. This allowed a controlled comparison of the

proposed methods, but it did not capture how users actually interact with an adaptive questionnaire, including engagement, fatigue, response behaviour, or dropout. Testing dynamic question ordering in mobile or self-administered FFQ applications would therefore be an important next step. Such an evaluation would show whether the improvements observed offline also translate into better usability, higher completion rates, and improved data quality in practice. This is particularly relevant because the proposed methods are intended for settings in which users may not answer all questions, making real user behaviour central to their practical effectiveness.

An important direction for future work is to evaluate the proposed methods in domains beyond dietary assessment. Although this thesis focuses on FFQs, the underlying task of sequentially selecting the most informative inputs is not specific to nutrition. Similar problems arise in other settings where questions, tests, or measurements must be prioritised under limited resources. One example is the RED-S screening tool developed as part of our related research, in which dynamic question ordering could improve screening efficiency while reducing respondent burden. Another example is battery degradation prediction, where the same framework could be used to determine which diagnostic tests to perform first to estimate future battery health as accurately as possible, given partial information. Such applications would provide a clearer assessment of the generalisability of the proposed framework and help identify which components transfer directly and which require adaptation to the specific characteristics of a new domain.

Finally, deeper personalisation represents a natural extension of this work. By incorporating user-specific context—such as age, health history, or prior responses—future systems could dynamically adapt both question selection and target prioritisation, enabling fully personalised assessment tools that optimise prediction quality while minimising user burden in real-world settings.

Appendix A

Additional Results

A.1 Feature Reduction

A.1.1 Target-Independed Feature Reduction

ESFFFQ

Table A.1: Ranks for classification compared for target Sugar by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	2	7	13	18	22	27		2	7	13	18	22	27	
RF	5	2	9	8	6	7	6.2	5	2	7	7	7	7	5.8
Bagging	**1**	**1**	**1**	7	3	10	**3.8**	**1**	5	3	**1**	4	10	4.0
GB	4	6	3	**1**	4	9	4.5	4	4	5	6	5	9	5.5
LogR	7	9	5	9	**1**	3	5.7	7	7	2	2	2	3	3.8
kNN	2	7	4	3	8	6	5.0	2	3	9	9	8	6	6.2
DT	2	3	6	2	7	4	4.0	2	**1**	**1**	4	**1**	4	**2.2**
SVM	7	4	2	10	2	5	5.0	7	10	4	3	3	5	5.3
VC	6	5	7	6	5	8	6.2	6	6	6	5	6	8	6.2
Zero	7	8	8	4	10	**1**	6.3	7	8	8	8	9	**1**	6.8
MICE	7	10	10	5	9	**1**	7.0	7	8	10	10	10	**1**	7.7

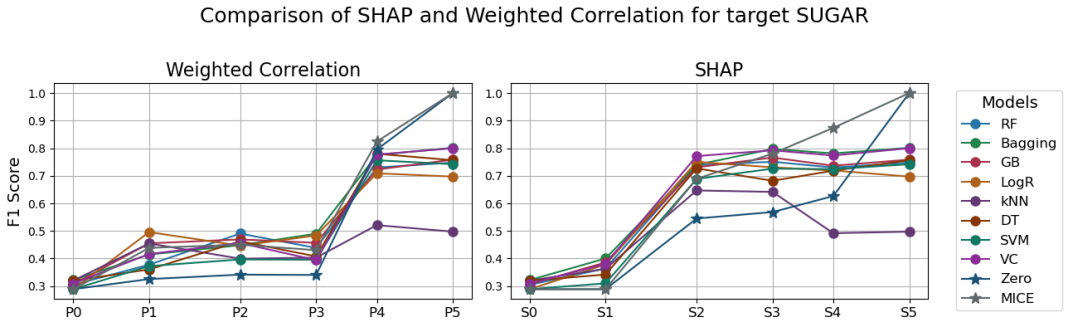


Figure A.1: Comparison of the F1-scores of different models on subsets created using SHAP and Weighted Correlation for target Sugar.

Table A.2: Ranks for regression compared for target Sugar by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	2	7	13	18	22	27		2	7	13	18	22	27	
RF	4	5	3	3	4	4	3.8	4	6	2	2	3	4	3.5
Bagging	6	7	6	6	5	5	5.8	6	5	3	3	4	5	4.3
GB	2	2	4	4	2	3	**2.8**	2	4	**1**	**1**	**1**	3	**2.0**
LinR	**1**	**1**	**1**	**1**	8	8	3.3	**1**	**1**	8	8	8	8	5.7
kNN	7	4	5	5	9	9	6.5	7	7	9	9	9	9	8.3
DT	5	10	10	10	6	6	7.8	5	9	6	6	7	6	6.5
SVR	8	6	7	7	10	10	8.0	8	3	10	10	10	10	8.5
VR	3	3	2	2	7	7	4.0	3	2	4	4	5	7	4.2
Zero	10	9	9	9	3	**1**	6.8	10	10	7	7	6	**1**	6.8
MICE	9	8	8	8	**1**	**1**	5.8	9	8	5	5	2	**1**	5.0

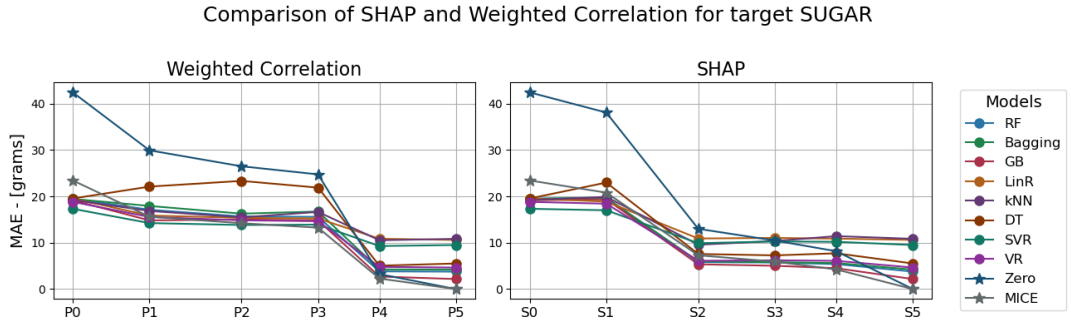


Figure A.2: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Sugar.

Table A.3: Ranks for classification compared for target Fibre by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	2	7	13	18	22	27		2	7	13	18	22	27	
RF	4	2	8	6	7	6	5.5	4	3	5	3	9	6	5.0
Bagging	2	**1**	**1**	2	6	5	**2.8**	2	2	**1**	**1**	2	5	**2.2**
GB	7	4	2	3	5	8	4.8	7	7	2	2	3	8	4.8
LogR	8	8	9	5	**1**	3	5.7	8	5	8	7	5	3	6.0
kNN	**1**	5	5	8	4	7	5.0	**1**	4	3	6	6	7	4.5
DT	4	3	3	**1**	2	4	**2.8**	4	**1**	6	4	**1**	4	3.3
SVM	3	9	6	4	3	9	5.7	3	6	9	9	8	9	7.3
VC	4	7	7	7	8	10	7.2	4	8	7	8	7	10	7.3
Zero	8	10	10	9	9	**1**	7.8	8	9	10	10	10	**1**	8.0
MICE	8	6	4	10	10	**1**	6.5	8	10	4	5	4	**1**	5.3

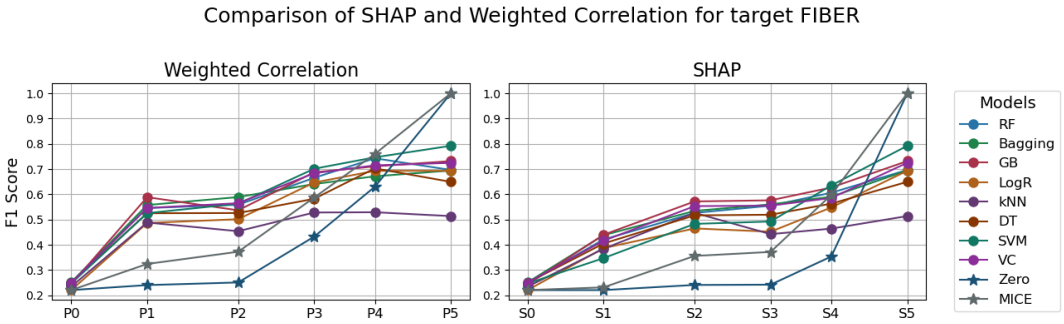


Figure A.3: Comparison of the F1-scores of different models on subsets created using SHAP and Weighted Correlation for target Fiber.

Table A.4: Ranks for regression compared for target Fibre by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	2	7	13	18	22	27		2	7	13	18	22	27	
RF	4	3	**1**	3	4	4	3.2	4	3	4	2	2	4	3.2
Bagging	2	4	3	5	5	5	4.0	2	5	7	4	3	5	4.3
GB	3	**1**	2	**1**	2	3	**2.0**	3	4	3	**1**	**1**	3	**2.5**
LinR	**1**	5	5	6	7	7	5.2	**1**	2	2	5	5	7	3.7
kNN	8	6	6	10	10	10	8.3	8	7	5	6	9	10	7.5
DT	5	9	8	8	8	8	7.7	5	8	9	9	10	8	8.2
SVR	7	7	7	9	9	9	8.0	7	6	6	7	7	9	7.0
VR	6	2	4	4	6	6	4.7	6	**1**	**1**	3	4	6	3.5
Zero	10	10	10	7	3	**1**	6.8	10	10	10	10	8	**1**	8.2
MICE	9	8	9	2	**1**	**1**	5.0	9	9	8	8	6	**1**	6.8

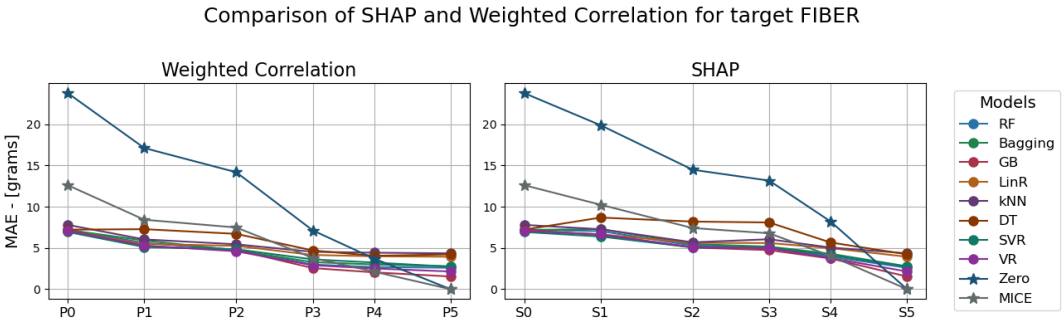


Figure A.4: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Fiber.

Table A.5: Ranks for classification compared for target Protein by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	2	7	13	18	22	27		2	7	13	18	22	27	
RF	3	3	10	9	9	9	7.2	3	3	5	10	10	9	6.7
Bagging	3	4	3	2	4	8	4.0	3	2	**1**	6	8	8	4.7
GB	3	6	7	6	7	7	6.0	3	6	6	5	7	7	5.7
LogR	3	7	4	3	3	5	4.2	3	8	4	3	3	5	4.3
kNN	**1**	2	2	5	2	4	2.7	**1**	4	2	2	**1**	4	2.3
DT	3	**1**	**1**	**1**	**1**	3	**1.7**	3	**1**	3	**1**	2	3	**2.2**
SVM	3	9	6	8	6	6	6.3	3	9	8	4	6	6	6.0
VC	3	5	9	10	8	10	7.5	3	7	7	8	5	10	6.7
Zero	2	8	8	4	5	**1**	4.7	2	5	10	9	9	**1**	6.0
MICE	10	10	5	7	10	**1**	7.2	10	10	9	7	4	**1**	6.8

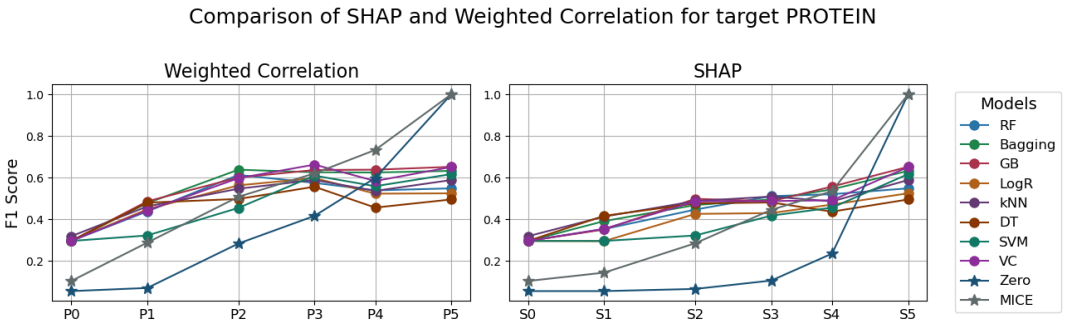


Figure A.5: Comparison of the F1-scores of different models on subsets created using SHAP and Weighted Correlation for target Protein.

Table A.6: Ranks for regression compared for target Protein by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	2	7	13	18	22	27		2	7	13	18	22	27	
RF	4	4	2	3	5	5	3.8	4	3	4	2	3	5	3.5
Bagging	5	5	4	5	6	6	5.2	5	4	5	5	2	6	4.5
GB	2	**1**	**1**	**1**	2	3	**1.7**	2	5	2	**1**	**1**	3	**2.3**
LinR	**1**	2	5	6	7	7	4.7	**1**	**1**	**1**	4	4	7	3.0
kNN	8	7	7	9	10	10	8.5	8	7	6	6	7	10	7.3
DT	3	8	8	10	8	8	7.5	3	8	8	9	9	8	7.5
SVR	6	6	9	8	9	9	7.8	6	6	7	7	8	9	7.2
VR	7	3	3	4	4	4	4.2	7	2	3	3	5	4	4.0
Zero	10	10	10	7	3	**1**	6.8	10	10	10	10	10	**1**	8.5
MICE	9	9	6	2	**1**	**1**	4.7	9	9	9	8	6	**1**	7.0

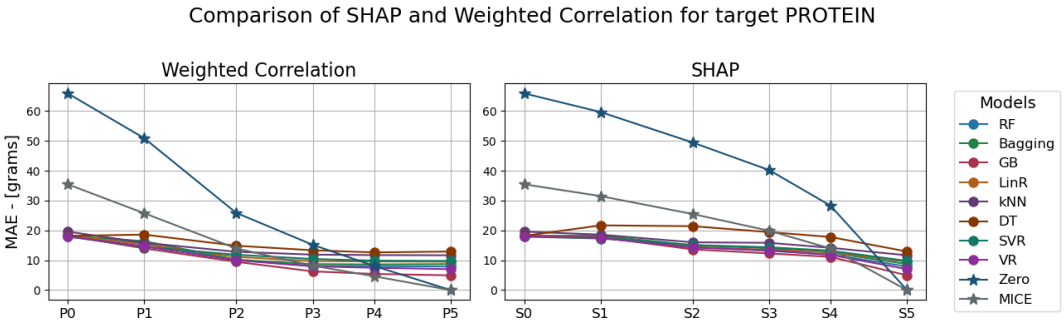


Figure A.6: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Protein.

Table A.7: Ranks for classification compared for target Salt by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	2	7	13	18	22	27		2	7	13	18	22	27	
RF	8	5	5	5	10	6	6.5	8	2	8	7	7	6	6.3
Bagging	8	10	4	10	9	9	8.3	8	**1**	2	10	4	9	5.7
GB	7	9	8	7	5	4	6.7	7	7	10	9	10	4	7.8
LogR	4	7	10	8	6	10	7.5	4	5	7	8	9	10	7.2
kNN	3	3	2	2	2	3	2.5	3	3	4	3	5	3	3.5
DT	6	4	**1**	6	8	5	5.0	6	8	**1**	**1**	2	5	3.8
SVM	5	6	9	9	4	8	6.8	5	9	6	6	8	8	7.0
VC	10	8	7	4	7	7	7.2	10	10	9	5	6	7	7.8
Zero	**1**	2	6	3	3	**1**	2.7	**1**	6	5	4	3	**1**	3.3
MICE	**1**	**1**	3	**1**	**1**	**1**	**1.3**	**1**	4	3	2	**1**	**1**	**2.0**

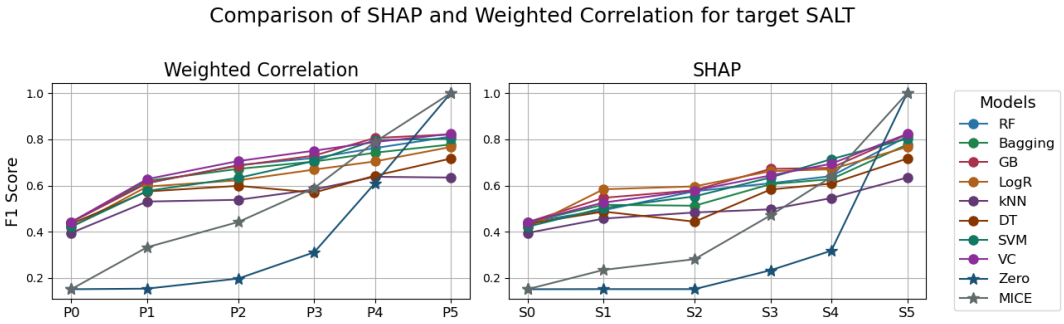


Figure A.7: Comparison of the F1-scores of different models on subsets created using SHAP and Weighted Correlation for target Salt.

Table A.8: Ranks for regression compared for target Salt by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	2	7	13	18	22	27		2	7	13	18	22	27	
RF	3	7	2	3	3	6	4.0	3	5	3	4	4	6	4.2
Bagging	4	6	4	2	7	5	4.7	4	6	6	5	7	5	5.5
GB	2	**1**	**1**	**1**	2	3	**1.7**	2	2	4	**1**	**1**	3	**2.2**
LinR	6	3	6	6	8	8	6.2	6	**1**	5	6	6	8	5.3
kNN	8	5	7	9	9	9	7.8	8	7	7	8	8	9	7.8
DT	**1**	9	8	8	10	10	7.7	**1**	8	8	9	9	10	7.5
SVR	5	4	5	5	5	4	4.7	5	4	2	3	3	4	3.5
VR	7	2	3	4	4	7	4.5	7	3	**1**	2	2	7	3.7
Zero	10	10	10	10	6	**1**	7.8	10	10	10	10	10	**1**	8.5
MICE	9	8	9	7	**1**	**1**	5.8	9	9	9	7	5	**1**	6.7

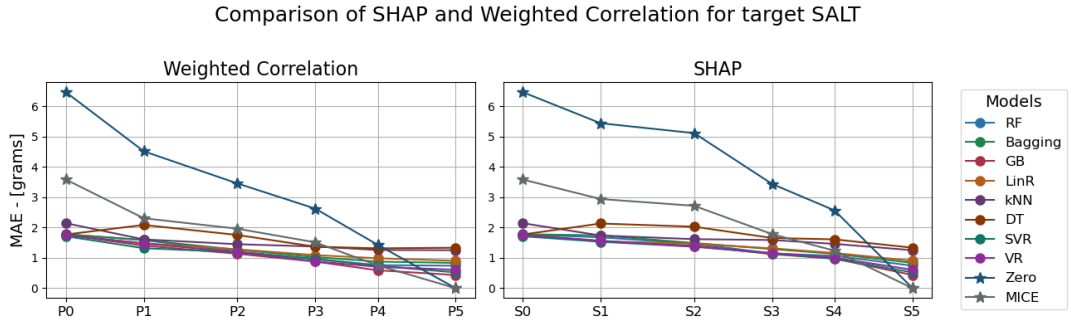


Figure A.8: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Salt.

NHANES

Table A.9: Ranks for regression models compared for target HDL cholesterol by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
Subset	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	9	49	75	127	156	200		9	49	75	127	156	200	
RF	**1**	**1**	**1**	**1**	**1**	**1**	**1.0**	**1**	**1**	**1**	**1**	**1**	**1**	**1.0**
Bagging	5	6	3	7	5	3	4.8	7	6	2	7	8	3	5.5
GB	4	3	6	4	4	4	4.2	3	3	4	3	4	4	3.5
LinR	3	4	7	6	7	6	5.5	2	4	5	6	6	6	4.8
kNN	6	7	4	2	2	5	4.3	6	7	6	5	3	5	5.3
DT	8	5	5	5	6	7	6.0	5	5	7	4	5	7	5.5
SVR	7	8	8	8	8	8	7.8	8	8	8	8	7	8	7.8
VR	2	2	2	3	3	2	2.3	4	2	3	2	2	2	2.5

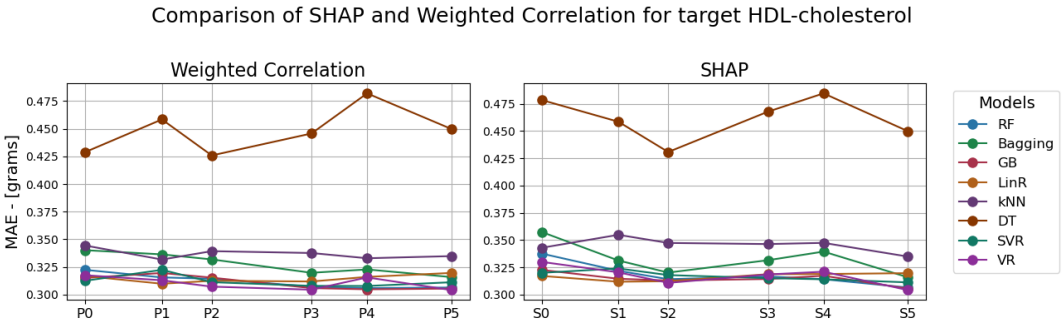


Figure A.9: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target HDL Cholesterol.

Table A.10: Ranks for regression models compared for target LDL cholesterol by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
Subset	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	9	49	75	127	156	200		9	49	75	127	156	200	
RF	5	4	**1**	**1**	**1**	**1**	2.2	5	**1**	**1**	2	3	**1**	2.2
Bagging	6	5	6	7	5	5	5.7	6	6	6	6	6	5	5.8
GB	**1**	**1**	2	2	2	3	**1.8**	2	2	2	**1**	2	3	**2.0**
LinR	3	3	3	6	6	7	4.7	**1**	3	4	5	5	7	4.2
kNN	7	7	7	5	7	6	6.5	7	7	7	7	7	6	6.8
DT	8	8	8	8	8	8	8.0	8	8	8	8	8	8	8.0
SVR	2	6	5	4	4	4	4.2	3	5	5	4	4	4	4.2
VR	4	2	4	3	3	2	3.0	4	4	3	3	**1**	2	2.8

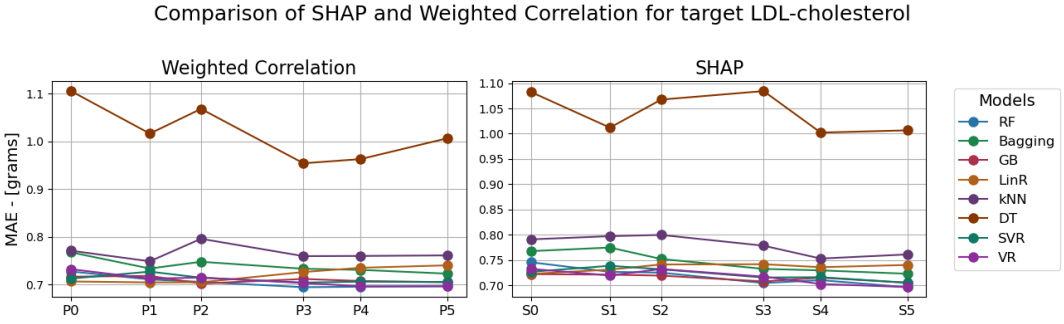


Figure A.10: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target LDL Cholesterol.

Table A.11: Ranks for regression models compared for target Glucose by the two feature reduction approaches, including average rank per method (rounded to one decimal).

	Weighted Correlation Analysis							SHAP-based Interaction Analysis						
Subset	P0	P1	P2	P3	P4	P5	Avg	S0	S1	S2	S3	S4	S5	Avg
# of questions	9	49	75	127	156	200		9	49	75	127	156	200	
RF	**1**	3	2	2	3	3	2.3	4	2	4	3	4	3	3.3
Bagging	3	7	5	4	2	4	4.2	7	4	5	4	2	4	4.3
GB	5	6	4	5	4	5	4.8	5	5	3	5	5	5	4.7
LinR	2	**1**	**1**	**1**	**1**	**1**	**1.2**	**1**	**1**	**1**	**1**	**1**	**1**	**1.0**
kNN	6	2	6	6	7	7	5.7	3	6	6	7	6	7	5.8
DT	8	8	8	8	8	8	8.0	8	8	8	8	8	8	8.0
SVR	7	5	7	7	6	6	6.3	6	7	7	6	7	6	6.5
V	4	4	3	3	5	2	3.5	2	3	2	2	3	2	2.3

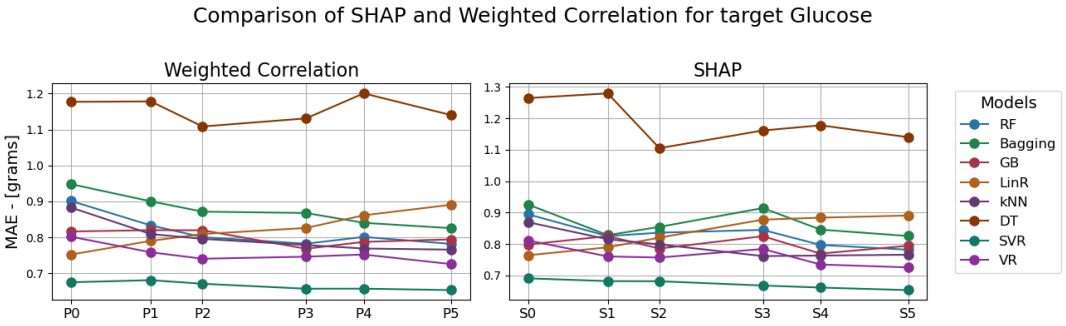


Figure A.11: Comparison of MAE of different models on subsets created using SHAP and Weighted Correlation for target Glucose.

A.1.2 Target Depended Feature Selection

ESFFFQ

Table A.12: Ranking of classification algorithms' performance on subsets of features selected by different methods for target Sugar.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	6	5	27	5	3	13	11	6	13
RF	8	3	5	7	4	7	8	8	7
Bagging	2	5	8	4	3	3	4	2	4
GB	6	7	7	3	5	4	5	5	3
LogR	**1**	4	**1**	**1**	6	**1**	3	3	**1**
kNN	4	6	4	8	**1**	8	7	7	8
DT	5	2	2	5	2	5	**1**	**1**	5
SVM	3	8	3	2	6	2	2	4	2
VC	7	**1**	6	6	8	6	6	6	6

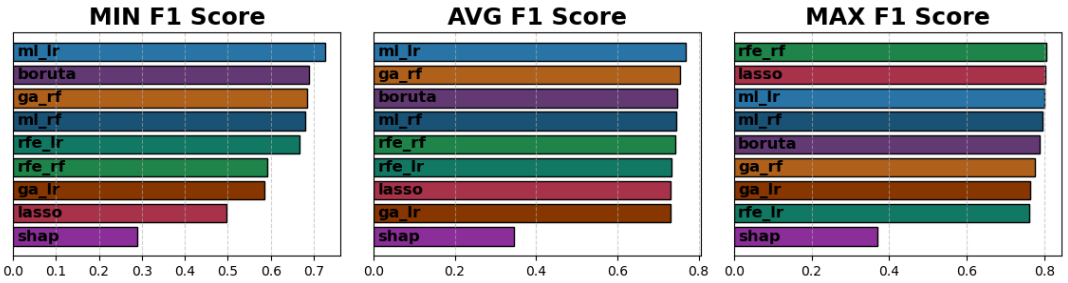


Figure A.12: Comparison of F1-score on subsets of features selected by different methods for target Sugar.

Table A.13: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Sugar.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	5	7	19	5	5	12	12	11	9
RF	2	6	2	2	2	2	2	2	2
Bagging	3	7	3	3	3	3	4	3	3
GB	**1**	2	**1**	**1**	**1**	**1**	**1**	**1**	**1**
LinR	7	**1**	6	7	7	6	6	6	7
kNN	6	4	7	6	6	7	7	7	6
DT	5	8	4	5	5	5	3	5	5
SVR	8	5	8	8	8	8	8	8	8
VR	4	3	5	4	4	4	5	4	4

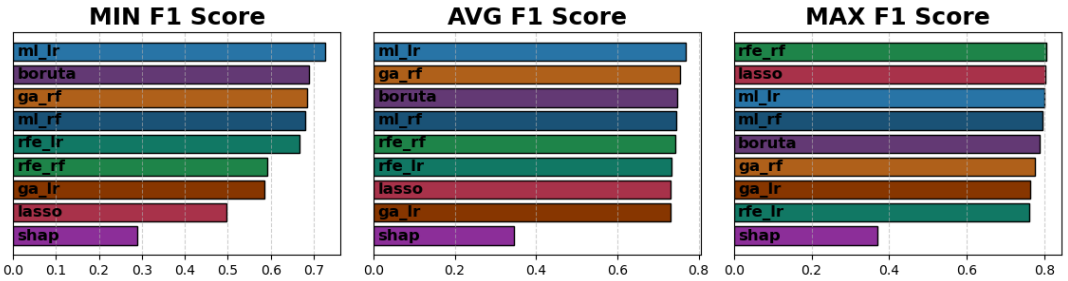


Figure A.13: Comparison of MAPE on subsets of features selected by different methods for target Sugar.

Table A.14: Ranking of classification algorithms' performance on subsets of features selected by different methods for target Fibre.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	6	5	27	11	1	13	13	15	20
RF	3	3	5	8	3	5	6	4	5
Bagging	**1**	**1**	3	5	3	3	3	2	6
GB	7	8	6	6	3	4	7	8	3
LogR	6	7	**1**	2	3	2	**1**	**1**	2
kNN	8	2	4	7	**1**	7	8	7	8
DT	4	4	2	**1**	3	**1**	2	3	**1**
SVM	2	5	8	3	3	8	5	6	7
VC	5	6	7	4	1	6	4	5	4

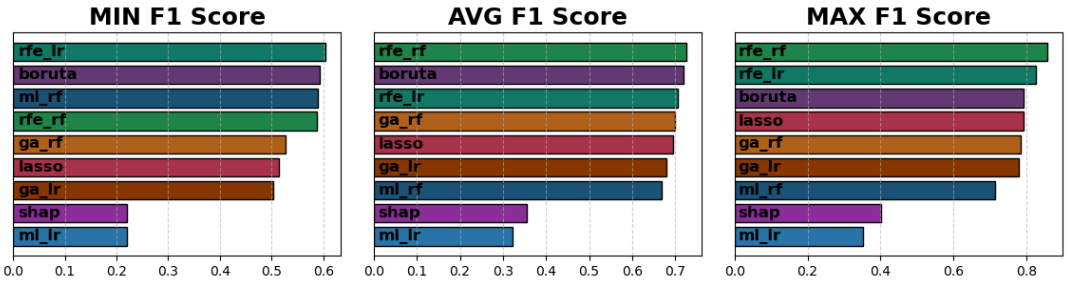


Figure A.14: Comparison of F1-score on subsets of features selected by different methods for target Fibre.

Table A.15: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Fibre.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	10	12	23	13	9	13	13	17	20
RF	2	2	2	2	2	2	2	2	2
Bagging	4	4	3	3	3	3	4	4	4
GB	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
LinR	5	5	5	5	6	5	5	5	5
kNN	7	8	8	8	7	8	8	8	8
DT	6	6	6	6	5	6	6	6	6
SVR	8	7	7	7	8	7	7	7	7
VR	3	3	4	4	4	4	3	3	3

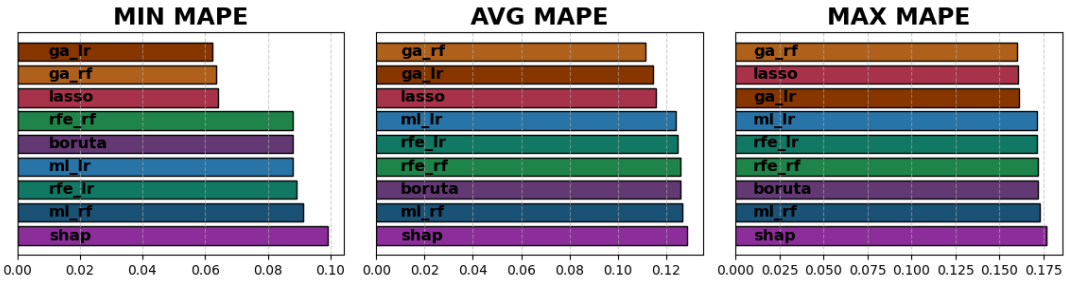


Figure A.15: Comparison of MAPE on subsets of features selected by different methods for target Fibre.

Table A.16: Ranking of classification algorithms' performance on subsets of features selected by different methods for target Protein.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	5	6	27	9	2	14	13	18	19
RF	3	4	8	5	2	8	8	7	8
Bagging	**1**	2	6	3	2	5	2	3	4
GB	5	8	5	7	2	4	4	5	6
LogR	8	6	3	**1**	7	2	5	4	2
kNN	2	3	2	4	6	**1**	**1**	**1**	**1**
DT	4	**1**	**1**	2	**1**	3	3	2	3
SVM	7	6	4	6	7	7	7	8	7
VC	6	5	7	8	2	6	6	6	5

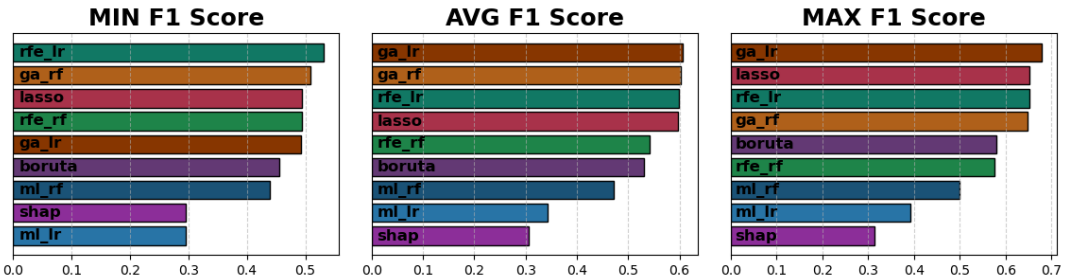


Figure A.16: Comparison of F1-score on subsets of features selected by different methods for target Protein.

Table A.17: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Protein.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	7	8	23	12	3	13	12	16	16
RF	3	3	3	2	2	2	2	2	2
Bagging	5	4	5	5	3	4	5	5	5
GB	**1**	**1**	**1**	**1**	4	**1**	**1**	**1**	**1**
LinR	4	5	4	4	5	5	4	4	4
kNN	6	6	8	6	6	6	6	6	6
DT	8	8	6	7	8	8	8	8	7
SVR	7	7	7	8	7	7	7	7	8
VR	**2**	**2**	**2**	3	**1**	3	3	3	3

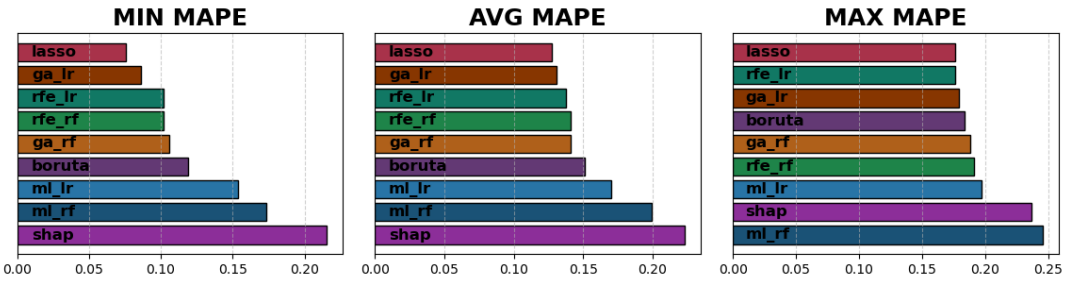


Figure A.17: Comparison of MAPE on subsets of features selected by different methods for target Protein.

Table A.18: Ranking of classification algorithms' performance on subsets of features selected by different methods for target Salt.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	7	8	27	13	3	13	13	19	21
RF	7	6	4	7	3	8	6	4	7
Bagging	2	2	7	2	5	6	8	7	6
GB	8	5	3	**1**	6	5	**1**	**1**	**1**
LogR	6	8	8	8	8	**1**	3	8	8
kNN	**1**	3	**1**	3	**1**	2	2	3	4
DT	4	7	5	6	2	3	7	5	3
SVM	3	4	6	4	4	4	5	2	5
VC	5	**1**	2	5	7	7	4	6	2

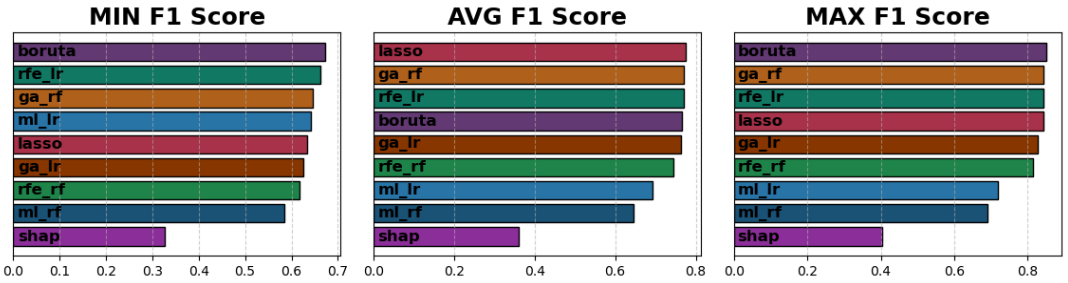


Figure A.18: Comparison of F1-score on subsets of features selected by different methods for target Salt.

Table A.19: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Salt.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	9	5	24	17	12	12	13	18	16
RF	4	4	3	5	5	5	4	5	3
Bagging	5	7	4	4	4	4	5	4	5
GB	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
LinR	6	3	6	6	6	6	6	6	6
kNN	7	6	7	7	8	8	7	7	7
DT	8	8	8	8	7	7	8	8	8
SVR	2	5	2	2	2	2	2	2	2
VR	3	2	5	3	3	3	3	3	4

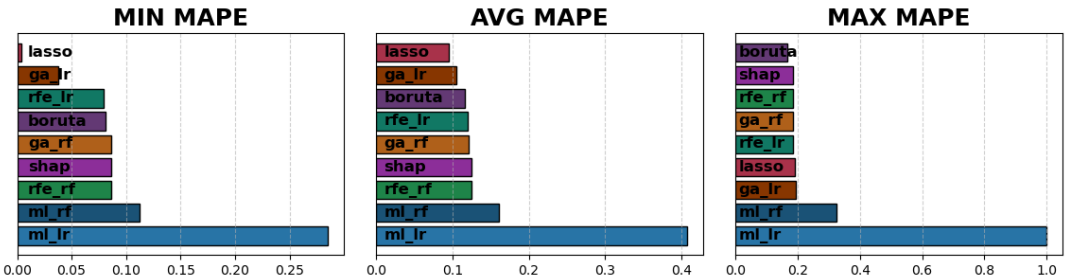


Figure A.19: Comparison of MAPE on subsets of features selected by different methods for target Salt.

NHANES

Table A.20: Ranking of regression algorithms' performance on subsets of features selected by different methods for target LDL cholesterol.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	6	5	135	2	1	104	104	97	76
RF	5	6	**1**	5	2	2	2	2	**1**
Bagging	7	7	5	6	5	5	4	6	6
GB	2	2	4	2	4	3	**1**	**1**	3
LinR	**1**	**1**	6	**1**	**1**	6	6	5	4
kNN	6	5	7	8	8	7	7	7	7
DT	8	8	8	7	6	8	8	8	8
SVR	3	3	3	4	7	4	5	4	5
VR	4	4	2	3	3	**1**	3	3	2

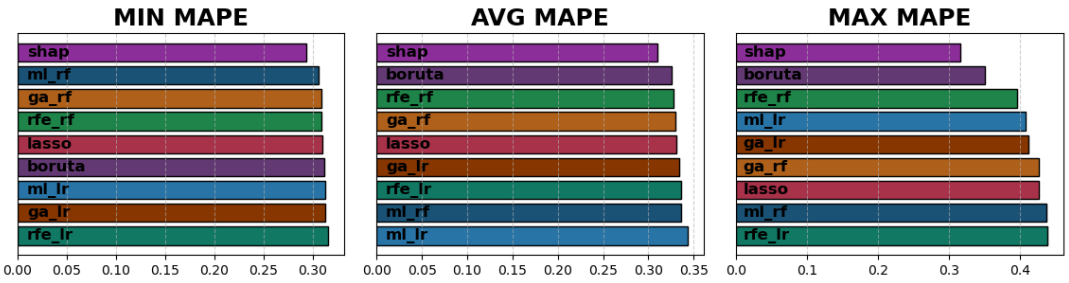


Figure A.20: Comparison of MAPE on subsets of features selected by different methods for target LDL cholesterol.

Table A.21: Ranking of regression algorithms' performance on subsets of features selected by different methods for target HDL cholesterol.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	6	7	56	1	1	105	91	100	88
RF	4	4	**1**	2	6	**1**	**1**	2	**1**
Bagging	5	6	4	**1**	3	5	4	3	3
GB	3	2	2	3	4	3	3	4	6
LinR	**1**	**1**	5	5	2	7	6	7	5
kNN	6	8	6	8	**1**	6	7	6	7
DT	7	5	8	4	5	4	5	5	4
SVR	8	7	7	7	8	8	8	8	8
VR	2	3	3	6	7	2	2	**1**	2

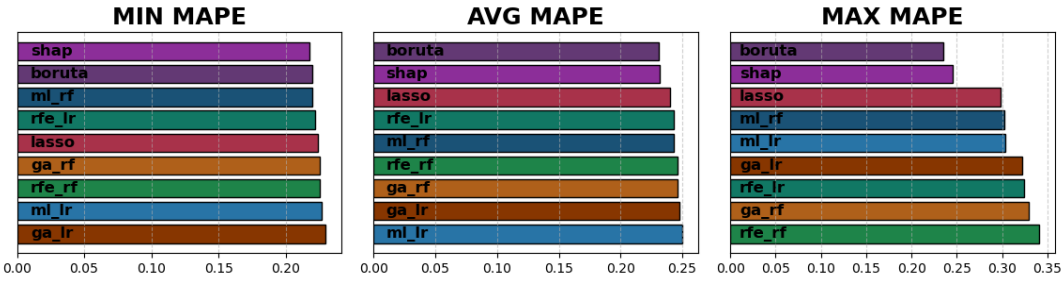


Figure A.21: Comparison of MAPE on subsets of features selected by different methods for target HDL cholesterol.

Table A.22: Ranking of regression algorithms' performance on subsets of features selected by different methods for target Glucose.

	ml_rf	ml_lr	lasso	boruta	shap	rfe_rf	rfe_lr	ga_rf	ga_lr
# of Features	7	9	173	1	1	104	95	101	87
RF	3	5	2	**1**	**1**	4	**1**	2	2
Bagging	4	4	4	5	4	3	5	5	5
GB	6	3	5	2	2	5	3	3	4
LinR	**1**	**1**	**1**	4	5	**1**	2	**1**	**1**
kNN	5	6	7	8	7	7	7	6	6
DT	8	8	8	3	3	8	8	8	8
SVR	7	7	6	7	8	6	6	7	7
VR	2	2	3	6	6	2	4	4	3

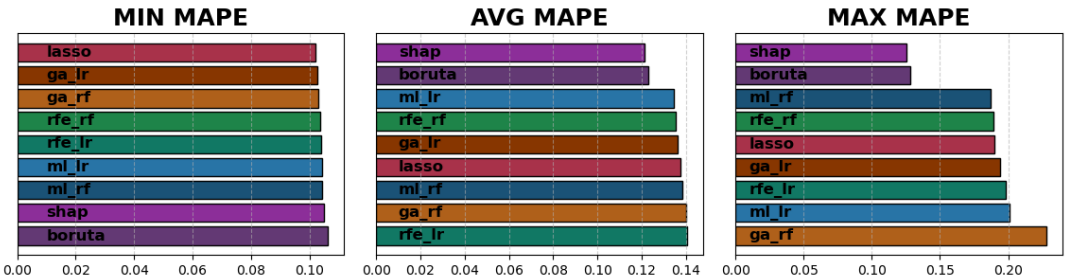


Figure A.22: Comparison of MAPE on subsets of features selected by different methods for target Glucose.

A.2 Feature Ranking

A.2.1 Static Feature Ranking

ESFFFQ

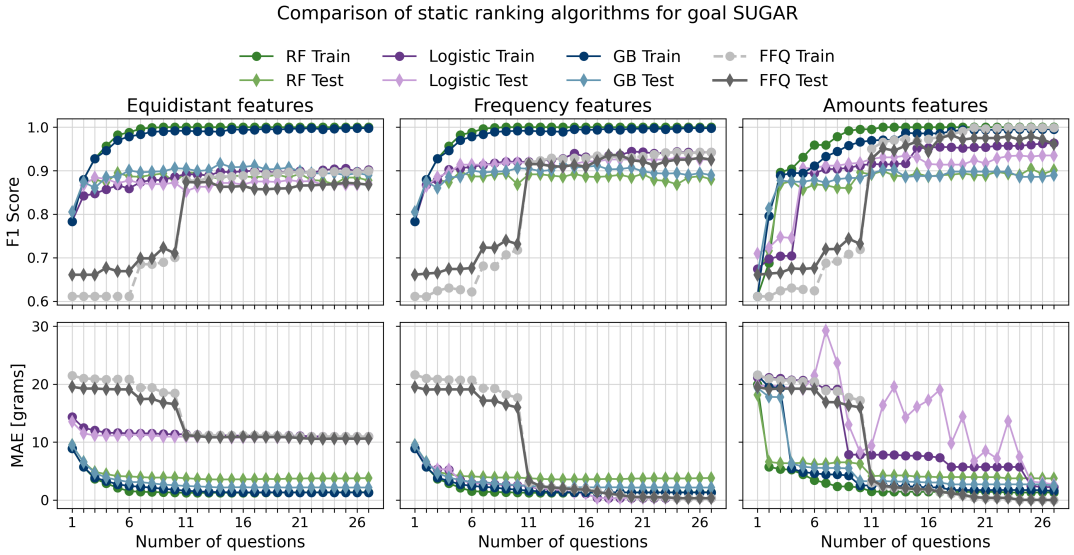


Figure A.23: Comparison of static ranking algorithms for sugar: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom).

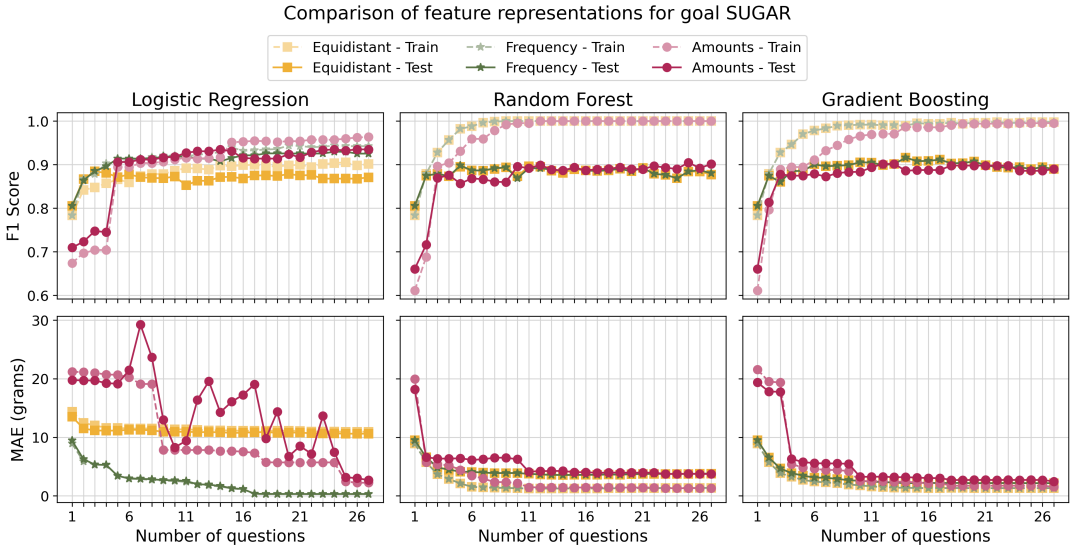


Figure A.24: Comparison of feature representations for sugar: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom).

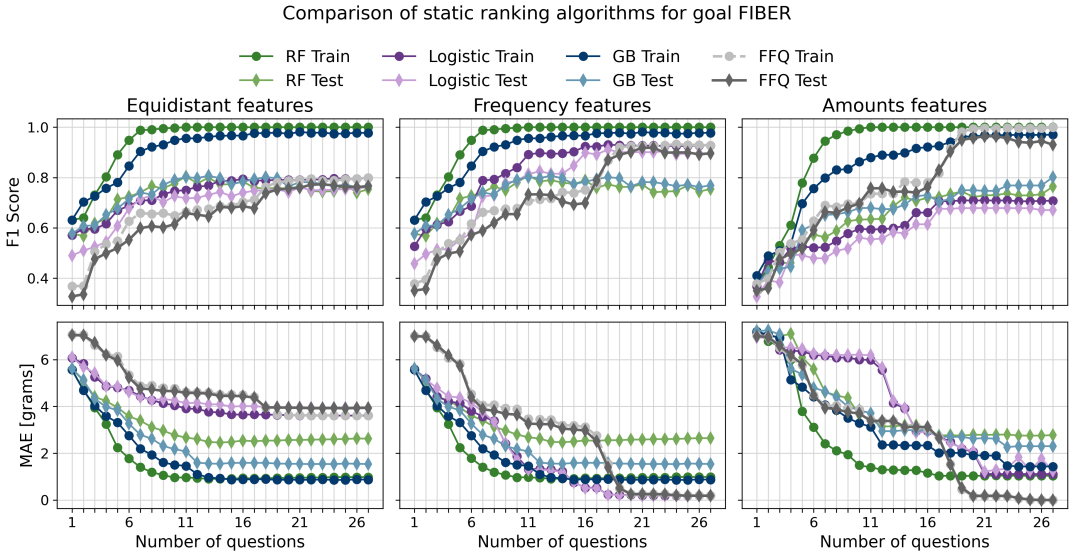


Figure A.25: Comparison of static ranking algorithms for fibre: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom).

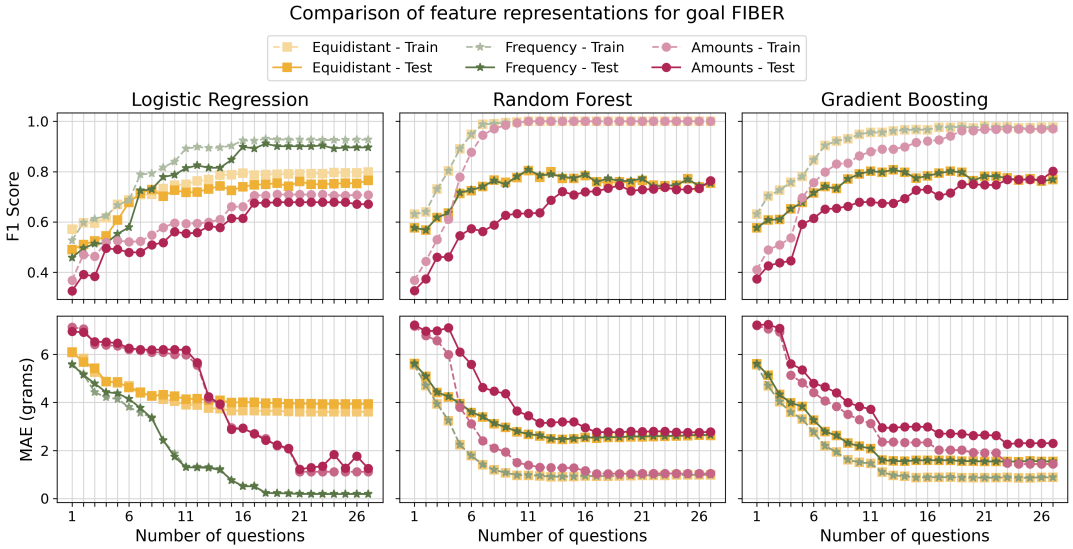


Figure A.26: Comparison of feature representations for fibre: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom).

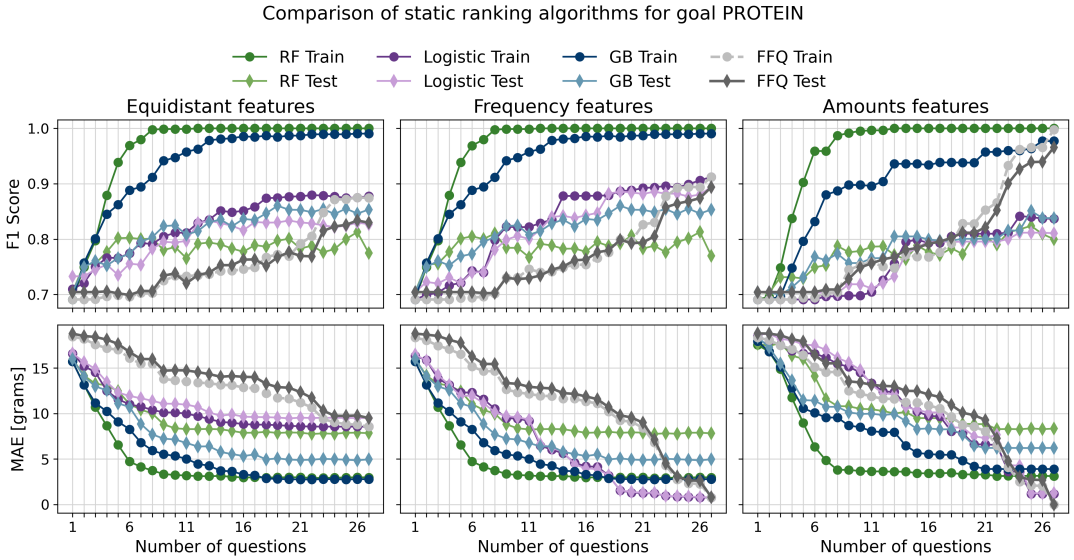


Figure A.27: Comparison of static ranking algorithms for salt: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom).



Figure A.28: Comparison of feature representations for protein: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom).

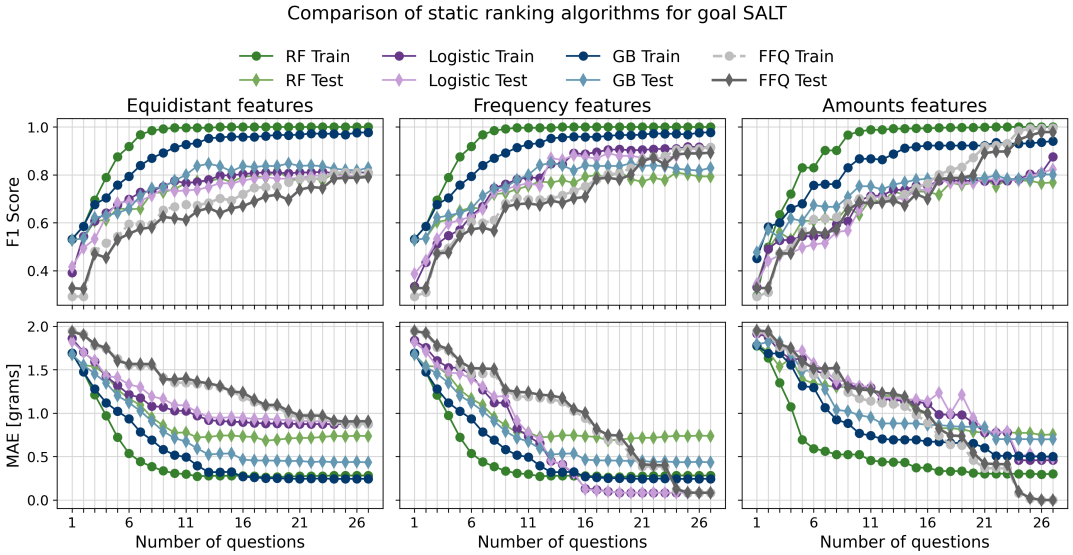


Figure A.29: Comparison of static ranking algorithms for salt: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom).

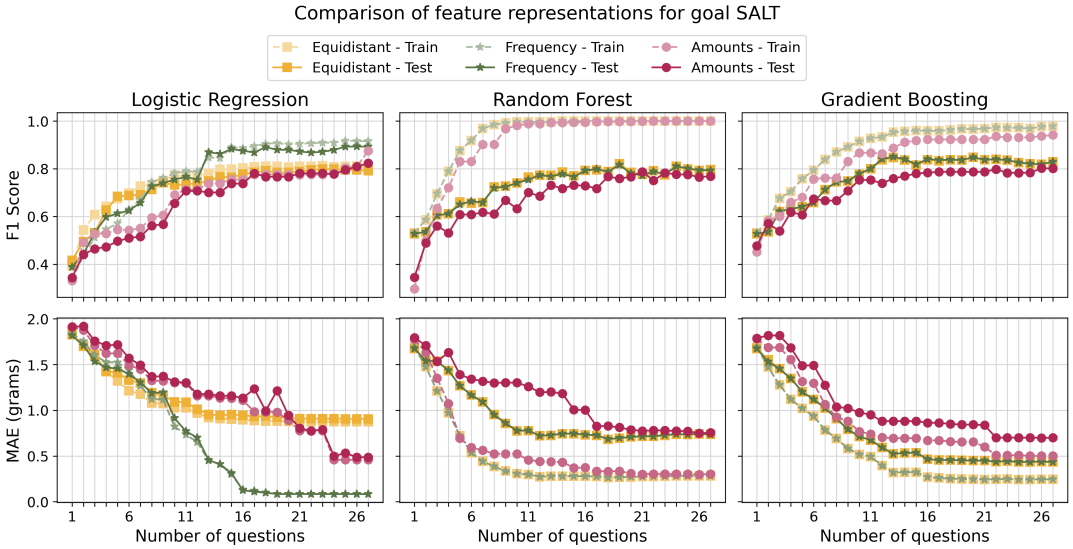


Figure A.30: Comparison of feature representations for salt: classification performance (F1 Score, top) and regression performance (MAE in grams, bottom).

NHANES

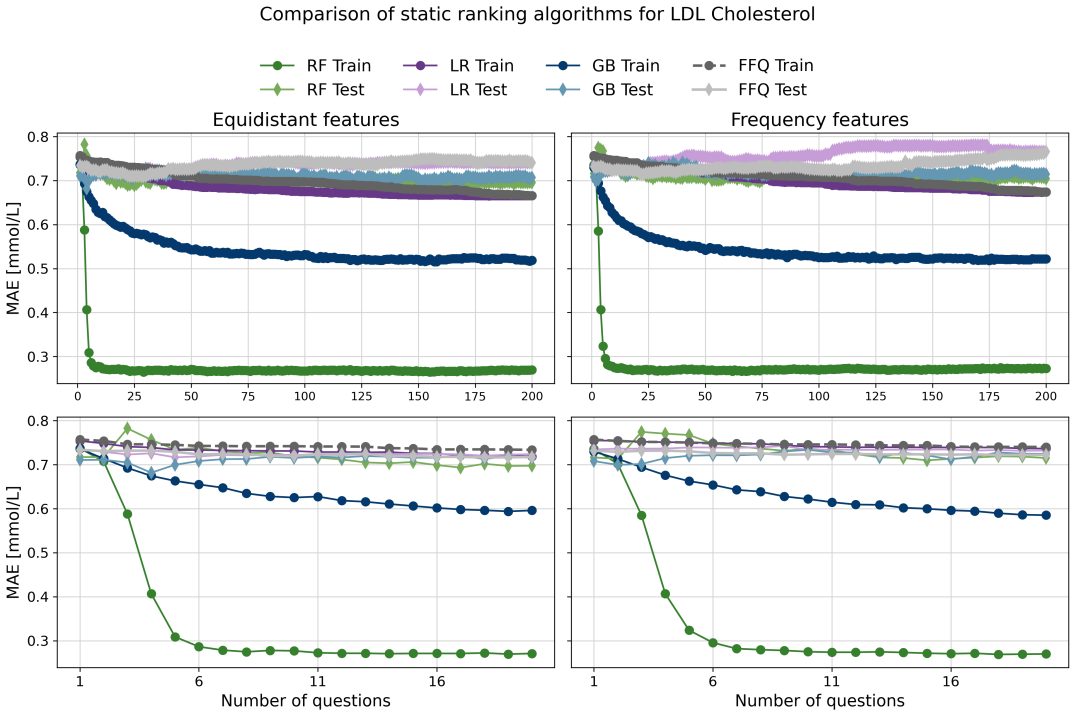


Figure A.31: Comparison of static ranking algorithms for LDL cholesterol: regression performance MAE (mmol/L) for all available features and for 20 best features.

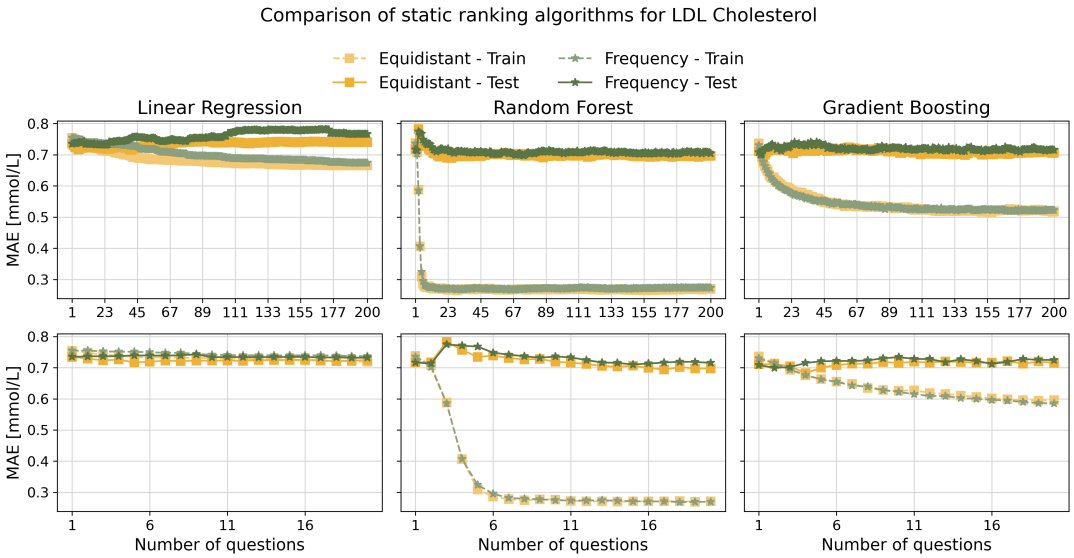


Figure A.32: Comparison of feature representations for LDL cholesterol: regression performance MAE (mmol/L) for all available features and for 20 best features.

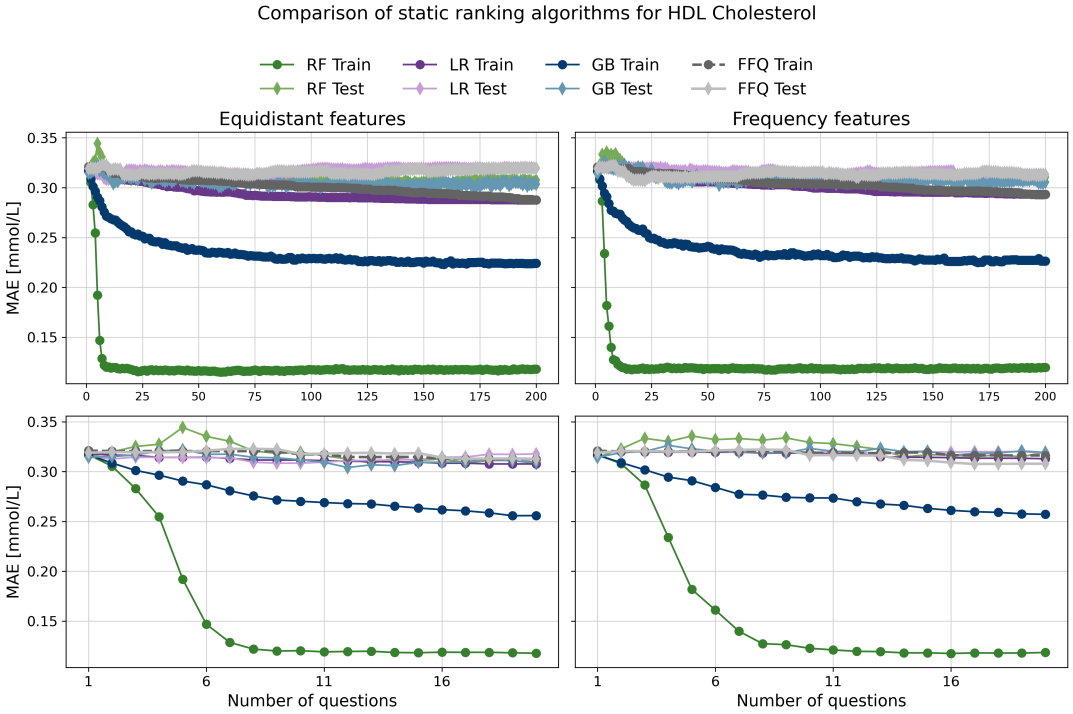


Figure A.33: Comparison of static ranking algorithms for HDL cholesterol: regression performance MAE (mmol/L) for all available features (top) and for 20 best features (bottom).

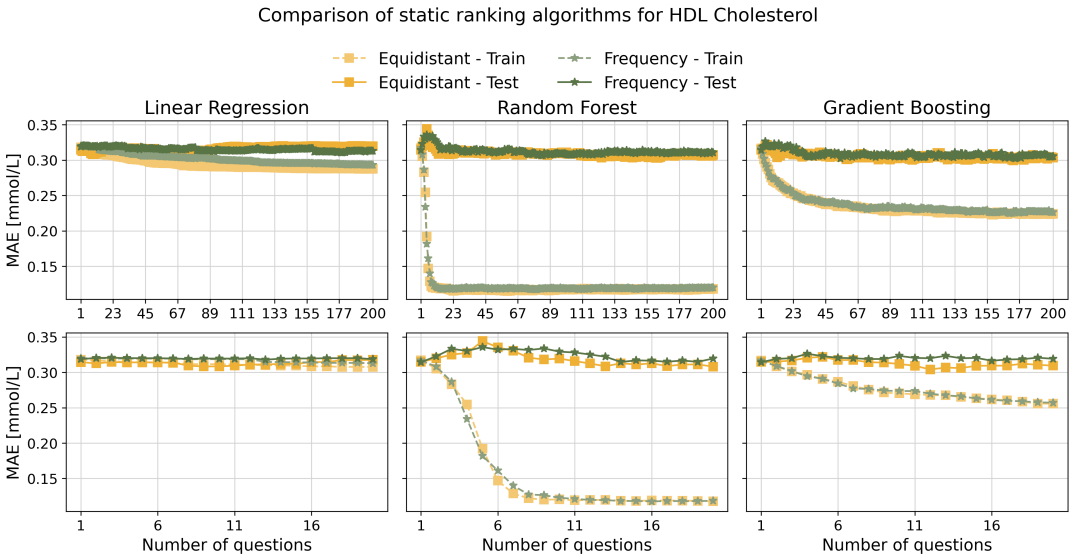


Figure A.34: Comparison of feature representations for LDL cholesterol: regression performance MAE (mmol/L) for all available features (top) and for 20 best features (bottom).

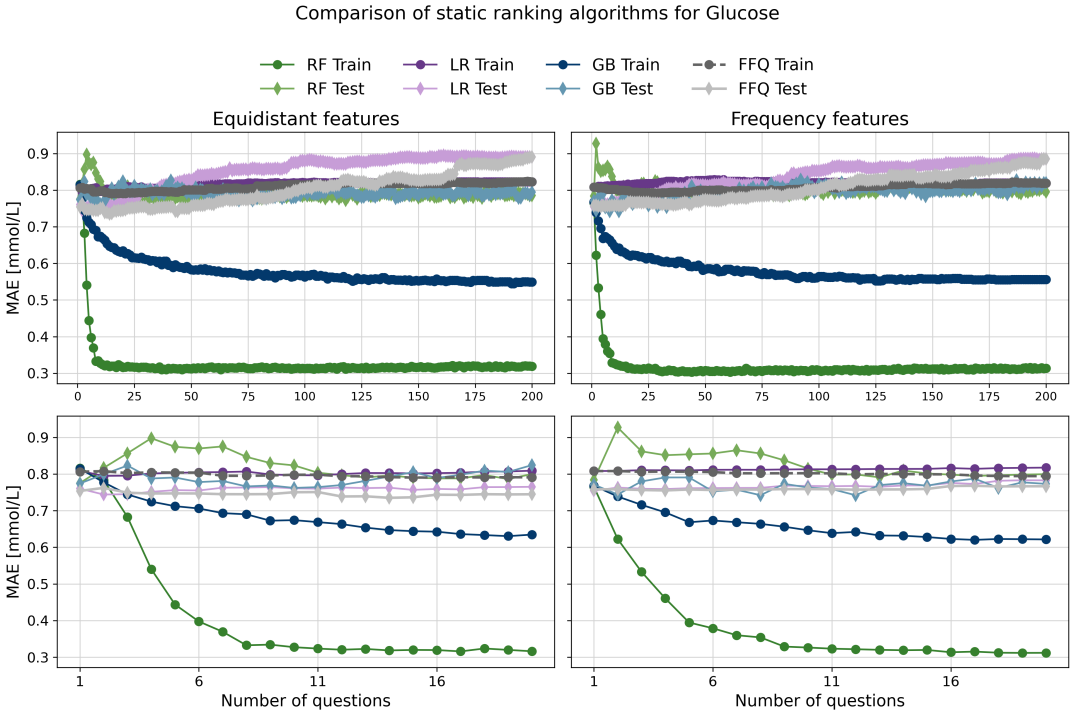


Figure A.35: Comparison of static ranking algorithms for glucose: regression performance MAE (mmol/L) for all available features (top) and for 20 best features (bottom).

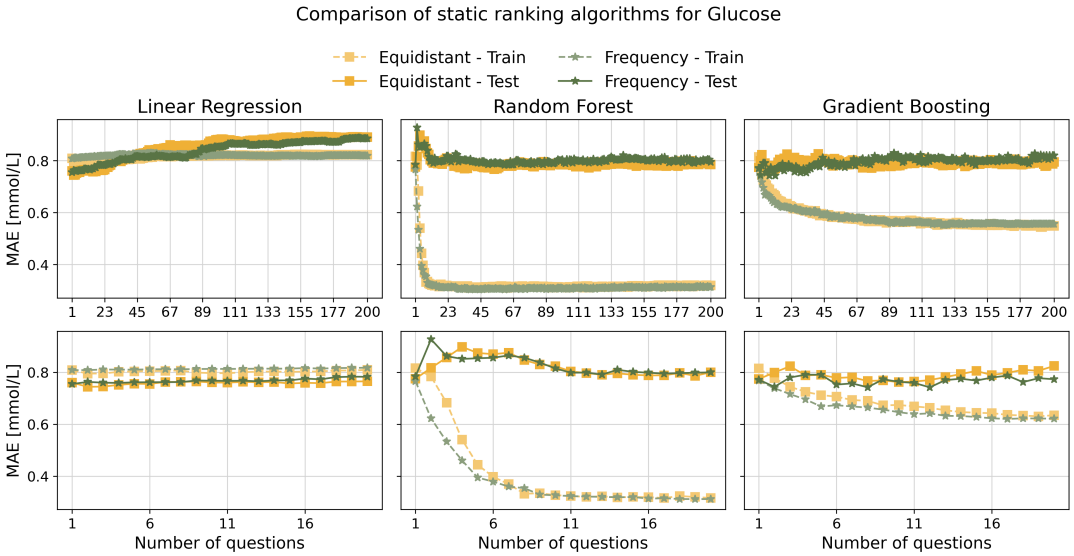


Figure A.36: Comparison of feature representations for glucose: regression performance MAE (mmol/L) for all available features (top) and for 20 best features (bottom).

References

[1] M. Brauer *et al.*, "Global burden and strength of evidence for 88 risk factors in 204 countries and territories, 1990–2021: A systematic analysis for the global burden of disease study 2021," *The Lancet*, vol. 403, no. 10428, pp. 2162–2203, 2024. DOI: 10.1016/S0140-6736(24)00933-4.

[2] Y. Xin *et al.*, "Dietary risk factors and associated disease burden among adults in china using global burden of disease 2021 estimates," *JMIR Public Health and Surveillance*, vol. 11, e72978, 2025. DOI: 10.2196/72978.

[3] World Health Organization, *Healthy diet*, https://www.who.int/health-topics/healthy-diet, Accessed 29 Apr 2025, 2024.

[4] Centers for Disease Control and Prevention, *Healthy eating for a healthy weight*, https://www.cdc.gov/healthyweight/healthy_eating/index.html, Accessed 29 Apr 2025, 2024.

[5] WordStream, *Top health and fitness keywords for search engine marketing*, Accessed: 2025-08-03, 2024. [Online]. Available: https://www.wordstream.com/popular-keywords/health-fitness-keywords.

[6] B. of Apps et al., *Diet & nutrition app download statistics: Myfitnesspal over 200 million downloads*, Reports MyFitnessPal surpassing 200 million downloads and health app downloads of 388 million in 2024; accessed 2025-08-03, 2025. [Online]. Available: https://www.businessofapps.com/data/health-app-market/.

[7] M. Raber, Y. Liao, A. Rara, *et al.*, "A systematic review of the use of dietary self-monitoring in behavioural weight loss interventions: Delivery, intensity and effectiveness," *Public Health Nutrition*, vol. 24, no. 17, pp. 3613–3625, 2021. DOI: 10.1017/S136898002100358X.

[8] K. M. Richardson, M. R. Jospe, L. C. Bohlen, J. Crawshaw, A. A. Saleh, and S. M. Schembre, "The efficacy of using continuous glucose monitoring as a behaviour change tool in populations with and without diabetes: A systematic review and meta-analysis of randomised controlled trials," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 21, no. 1, p. 145, 2024. DOI: 10.1186/s12966-024-01692-6.

[9] B. Margetts and M. Nelson, *Design Concepts in Nutritional Epidemiology*. Sep. 2009, pp. 1–472, ISBN: 9780192627391. DOI: 10.1093/acprof:oso/9780192627391.001.0001.

[10] M. Gonzalez-Ramirez, R. Sanchez-Carrera, A. Cejudo-Lopez, *et al.*, "Short-term pilot study to evaluate the impact of salbi educa nutrition app in macronutrients intake and adherence to the mediterranean diet: Randomized controlled trial," *Nutrients*, vol. 14, no. 10, 2022, ISSN: 2072-6643. DOI: 10.3390/nu14102061. [Online]. Available: https://www.mdpi.com/2072-6643/14/10/2061.

[11] J. D. Kusuma, H.-L. Yang, Y.-L. Yang, Z.-F. Chen, and S.-Y. P. K. Shiao, "Validating accuracy of a mobile application against food frequency questionnaire on key nutrients with modern diets for mhealth era," *Nutrients*, vol. 14, no. 3, 2022, ISSN: 2072-6643. DOI: 10.3390/nu14030537. [Online]. Available: https://www.mdpi.com/2072-6643/14/3/537.

[12] J. Cade, R. Thompson, V. Burley, and D. Warm, "Development, validation and utilisation of food-frequency questionnaires – a review," *Public Health Nutrition*, vol. 5, no. 4, pp. 567–587, 2002.

[13] J. C. Gerdessen, O. W. Souverein, P. van 't Veer, and J. H. de Vries, "Optimising the selection of food items for ffqs using mixed integer linear programming," *Public Health Nutrition*, vol. 18, no. 1, pp. 68–74, 2015. DOI: 10.1017/S1368980013003479.

[14] E. Tlamelo, T. Maupong, D. Mpoeleng, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, p. 140, 2021, Survey on machine learning approaches to missing data.

[15] K. Early, J. Mankoff, and S. E. Fienberg, "Dynamic question ordering in online surveys," *Journal of Official Statistics*, vol. 33, no. 3, pp. 625–657, 2017. DOI: 10.1515/jos-2017-0030.

[16] A. Rahbar, L. Aronsson, and M. H. Chehreghani, "A survey on active feature acquisition strategies," *arXiv preprint*, 2025. DOI: 10.48550/arXiv.2502.11067.

[17] N. Reščič, Ó. Mayora, C. Eccher, and M. Luštrek, "Food frequency questionnaire personalisation using multi-target regression," *Nutrients*, vol. 14, no. 19, p. 3943, 2022. DOI: 10.3390/nu14193943.

[18] Centers for Disease Control and Prevention, *Food frequency questionnaire, nhanes 2005-2006*, Accessed: 2024-07-03, 2023. [Online]. Available: https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/FFQDC_D.htm.

[19] M. Molag, *Towards Transparent Development of Food Frequency Questionnaires: Scientific Basis of the Dutch FFQ-TOOL: a Computer System to Generate, Apply and Process FFQs*. Wageningen Universiteit, 2010.

[20] P. M. Guenther, S. I. Kirkpatrick, J. Reedy, and et al., "The healthy eating index–2010 is a valid and reliable measure of diet quality according to the 2010 dietary guidelines for americans," *Journal of Nutrition*, vol. 144, no. 3, pp. 399–407, 2014. DOI: 10.3945/jn.113.183079.

[21] A. Trichopoulou, T. Costacou, C. Bamia, and D. Trichopoulos, "Adherence to a mediterranean diet and survival in a greek population," *New England Journal of Medicine*, vol. 348, no. 26, pp. 2599–2608, 2003. DOI: 10.1056/NEJMoa025039.

[22] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[24] N. Reščič, E. Valenčič, E. Mlinarič, B. Koroušić-Seljak, and M. Luštrek, "Mobile nutrition monitoring for well-being," ser. UbiComp/ISWC '19 Adjunct, London, United Kingdom: Association for Computing Machinery, 2019, pp. 1194–1197. DOI: 10.1145/3341162.3347076.

[25] W. Willett, *Nutritional Epidemiology*, 3rd. Oxford University Press, 2013.

[26] J. M. Conway, L. A. Ingwersen, and A. J. Moshfegh, "Accuracy of dietary recall using the usda five-step multiple-pass method in men: An observational validation study," *Journal of the American Dietetic Association*, vol. 104, no. 4, pp. 595–603, 2004.

[27] F. E. Thompson and A. F. Subar, "Dietary assessment methodology," in *Nutrition in the Prevention and Treatment of Disease*, Elsevier, 2013, pp. 5–46.

[28] B. Liu, H. Young, F. L. Crowe, *et al.*, "Development and evaluation of the oxford webq, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies," *Public Health Nutrition*, vol. 14, no. 11, pp. 1998–2005, 2011.

[29] A.-K. Illner, H. Freisling, H. Boeing, I. Huybrechts, S. P. Crispim, and N. Slimani, "Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology," *International Journal of Epidemiology*, vol. 41, no. 4, pp. 1187–1203, 2012. DOI: 10.1093/ije/dys105.

[30] J.-S. Shim, K. Oh, and H. C. Kim, "Dietary assessment methods in epidemiologic studies," *Epidemiology and Health*, vol. 36, e2014009, 2014.

[31] L. Gemming, J. Utter, and C. Ni Mhurchu, "Image-assisted dietary assessment: A systematic review of the evidence," *Journal of the Academy of Nutrition and Dietetics*, vol. 115, no. 1, pp. 64–77, 2015.

[32] F. Zhu, M. Bosch, I. Woo, *et al.*, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756–766, 2010.

[33] K. V. Dalakleidi, M. Papadelli, I. Kapolos, and K. Papadimitriou, "Applying image-based food-recognition systems on dietary assessment: A systematic review," *Advances in Nutrition*, vol. 13, no. 6, pp. 2590–2619, 2022, ISSN: 2161-8313. DOI: https://doi.org/10.1093/advances/nmac078. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2161831323000935.

[34] C. Höchsmann and C. K. Martin, "Review of the validity and feasibility of image-assisted methods for dietary assessment," *International Journal of Obesity*, vol. 44, no. 12, pp. 2358–2371, 2020. DOI: 10.1038/s41366-020-00693-2.

[35] A. F. Subar, F. E. Thompson, V. Kipnis, *et al.*, "Comparative validation of the block, willett, and national cancer institute food frequency questionnaires," *American Journal of Epidemiology*, vol. 154, no. 12, pp. 1089–1099, 2001.

[36] A. F. Subar, L. S. Freedman, J. A. Tooze, *et al.*, "Addressing current criticism regarding the value of self-report dietary data," *Journal of Nutrition*, vol. 145, no. 12, pp. 2639–2645, 2015.

[37] J. Cade, V. Burley, D. Warm, R. Thompson, and B. Margetts, "Food-frequency questionnaires: A review of their design, validation and utilisation," *Nutrition Research Reviews*, vol. 17, no. 1, pp. 5–22, 2004. DOI: 10.1079/NRR200370.

[38] National Center for Health Statistics, "Nhanes food propensity questionnaire," 2021, Accessed 27 May 2025.

[39] A. F. Subar, S. I. Kirkpatrick, B. Mittl, *et al.*, "The automated self-administered 24-hour dietary recall (asa24): A resource for researchers, clinicians, and educators from the national cancer institute," *Journal of the Academy of Nutrition and Dietetics*, vol. 115, no. 8, pp. 113–118, 2015. DOI: 10.1016/j.jand.2014.08.004.

[40] S. I. Kirkpatrick, J. Reedy, S. M. Krebs-Smith, T. E. Pannucci, and A. F. Subar, "Application of the dietary assessment primer to describe the validity and measurement error of self-reported dietary assessment instruments," *American Journal of Clinical Nutrition*, vol. 109, no. 7 Suppl, 50S–71S, 2019. DOI: 10.1093/ajcn/nqy036.

[41] D. Panaretos, E. Koloverou, A. C. Dimopoulos, *et al.*, "A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002–2012): The attica study," *British Journal of Nutrition*, vol. 120(3), pp. 326–334, 2018. DOI: 10.1017/S0007114518001150.

[42] H. Uemura, A. Ghaibeh, S. Katsuura-Kamano, *et al.*, "Systemic inflammation and family history in relation to the prevalence of type 2 diabetes based on an alternating decision tree," *Scientific reports*, vol. 7, Mar. 2017. DOI: 10.1038/srep45502.

[43] M. Gjoreski, S. Kochev, N. Reščič, M. Gregorič, T. Eftimov, and B. K. Seljak, "Exploring dietary intake data collected by fpq using unsupervised learning," in *2019 IEEE International Conference on Big Data*, 2019, pp. 5126–5130. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9005715.

[44] E. L. Chin, G. Simmons, Y. Y. Bouzid, *et al.*, "Nutrient estimation from 24-hour food recalls using machine learning and database mapping: A case study with lactose," *Nutrients*, vol. 11, no. 12, 2019, ISSN: 2072-6643. DOI: 10.3390/nu11123045.

[45] Á. Stefkovics and Z. Kmetty, "A comparison of question order effects on item-by-item and grid formats: Visual layout matters," *Measurement Instruments for the Social Sciences*, vol. 4, no. 1, p. 8, 2022. DOI: 10.1186/s42409-022-00036-z.

[46] L. M. Béjar, M. D. García-Perea, and Ó. A. Reyes, "Relative validity of a smartphone app (e-12hr) for collecting dietary intake data compared to a food frequency questionnaire and dietary records: A cross-sectional study," *JMIR mHealth and uHealth*, vol. 7, no. 4, e11531, 2019. DOI: 10.2196/11531.

[47] E.-K. Kim, H.-S. Park, H.-J. Lee, and N.-S. Joo, "Development of a mobile application for dietary self-monitoring and evaluation based on the korean healthy eating index: Diet-a," *Nutrition Research and Practice*, vol. 11, no. 6, pp. 622–628, 2017. DOI: 10.4162/nrp.2017.11.6.622.

[48] L. M. Béjar, M. D. García-Perea, and P. Mesa-Rodríguez, "Effectiveness of the e-12hr mobile app in improving adherence to the mediterranean diet in university students: A randomized controlled trial," *Nutrients*, vol. 14, no. 19, p. 4196, 2022. DOI: 10.3390/nu14194196.

[49] J. A. Krosnick, "Response strategies for coping with the cognitive demands of attitude measures in surveys," *Applied Cognitive Psychology*, vol. 5, no. 3, pp. 213–236, 1991. DOI: 10.1002/acp.2350050305.

[50] R. Tourangeau, L. J. Rips, and K. Rasinski, *The Psychology of Survey Response*. Cambridge University Press, 2000.

[51] S. Sudman, N. M. Bradburn, and N. Schwarz, *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, 1996.

[52] N. Reščič and M. Luštrek, "Question ranking for food frequency," in *Information Society - IS 2021, Volume A*, Ljubljana, Slovenia: Institut "Jožef Stefan", 2021, pp. 39–42. [Online]. Available: http://library.ijs.si/Stacks/Proceedings/InformationSociety/2021/IS2021_Volume_A.pdf.

[53] H. Wainer, *Computerized Adaptive Testing: A Primer*. Lawrence Erlbaum Associates, 2000.

[54] N. Reščič and M. Luštrek, "Dynamic feature ordering for personalised assessment of diet quality in ffqs," in *IEEE Big Data*, Sorrento, Italy: IEEE, 2023.

[55] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.

[56] S. van Buuren, *Flexible Imputation of Missing Data*, 2nd. CRC Press, 2018.

[57] H.-H. Chang and Z. Ying, "Computerized adaptive testing: A comprehensive introduction," in *Handbook of Modern Item Response Theory*, Springer, 2019, pp. 141–168. DOI: 10.1007/978-3-319-69218-0_3.

[58] D. Khanna, D. E. Furst, and W. K. Wong, "Computerized adaptive testing in clinical assessments: Accuracy and efficiency," *The Journal of Rheumatology*, vol. 41, no. 1, pp. 153–159, 2014. [Online]. Available: https://www.jrheum.org/content/41/1/153.

[59] P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney, "An expected utility approach to active feature-value acquisition," in *Proceedings of the IEEE International Conference on Data Mining*, 2005. DOI: 10.1109/ICDM.2005.23.

[60] F. Nan and V. Saligrama, "Adaptive classification for prediction under a budget," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[61] Y. Li and J. B. Oliva, "Dynamic feature acquisition with arbitrary conditional flows," *Proceedings of NeurIPS*, 2020. DOI: 10.48550/arXiv.2006.07701.

[62] H. Shim, S. J. Hwang, and E. Yang, "Why pay more when you can pay less: A joint learning framework for active feature acquisition and classification," in *Proceedings of ICML Workshops*, 2017.

[63] G. Contardo, L. Denoyer, and T. Artières, "Test-time feature ordering with focus: Interactive predictions with minimal user burden," pp. 329–338, 2016. DOI: 10.1145/2971648.2971748.

[64] O. B. Guney, K. S. Saichandran, K. Elzokm, Z. Zhang, and V. B. Kolachalama, "Active feature acquisition via explainability-driven ranking," in *Proceedings of the International Conference on Machine Learning (ICML) 2025*, Poster / conference paper on neural network–based policy networks for AFA, 2025. [Online]. Available: https://openreview.net/forum?id=J8YRdm39jn.

[65] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 5, pp. 216–233, 2015. DOI: 10.1002/widm.1157.

[66] D. Xu, W. Shi, I. W. Tsang, Y.-S. Ong, and C. Gong, "A survey on multi-output learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2409–2429, 2020. DOI: 10.1109/TNNLS.2019.2948133.

[67] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-target regression via input space expansion: Treating targets as inputs," *Machine Learning*, vol. 104, no. 1, pp. 55–98, 2016. DOI: 10.1007/s10994-016-5546-z.

[68] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. [Online]. Available: https://papers.nips.cc/paper/7334-multi-task-learning-as-multi-objective-optimization.

[69] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7482–7491. DOI: 10.1109/CVPR.2018.00781.

[70] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80, 2018, pp. 794–803. [Online]. Available: https://proceedings.mlr.press/v80/chen18a.html.

[71] C. L. Cleghorn, R. A. Harrison, J. K. Ransley, S. Wilkinson, J. Thomas, and J. E. Cade, "Can a dietary quality score derived from a short-form ffq assess dietary quality in uk adult population surveys?" *Public Health Nutrition*, vol. 19, no. 16, pp. 2915–2923, 2016. DOI: 10.1017/S1368980016001099.

[72] Centers for Disease Control and Prevention, *NHANES Food Frequency Questionnaire Data Documentation*, https://wwwn.cdc.gov/Nchs/Nhanes/, Dietary frequency data processed using NCI Diet*Calc; accessed: 2025-01-10, 2023.

[73] A. H. Lichtenstein *et al.*, "Dietary fat intake and cardiovascular disease," *Current Opinion in Lipidology*, vol. 17, no. 1, pp. 3–11, 2006. DOI: 10.1097/01.mol.0000204814.41794.c1.

[74] M. Mountjoy *et al.*, "Relative energy deficiency in sport (red-s)," *British Journal of Sports Medicine*, vol. 52, no. 7, pp. 434–444, 2018. DOI: 10.1136/bjsports-2017-098148.

[75] P. Zimmet *et al.*, "The global epidemiology of type 2 diabetes and its cardiovascular risks," *Nature Reviews Cardiology*, vol. 12, no. 3, pp. 107–118, 2015. DOI: 10.1038/nrcardio.2014.198.

[76] R. H. Liu *et al.*, "Antioxidant properties of fruits and vegetables," *Current Opinion in Food Science*, vol. 2, no. 1, pp. 50–57, 2007. DOI: 10.1016/j.cofs.2017.10.003.

[77] R. S. Gibson, *Principles of Nutritional Assessment*. Oxford University Press, 2005.

[78] S. Hercberg *et al.*, "Iron deficiency and iron deficiency anemia in france: Relevance to public health," *Public Health Nutrition*, vol. 4, no. 2, pp. 117–126, 2001. DOI: 10.1079/PHN2001117.

[79] S. R. Lynch *et al.*, "Iron deficiency and iron deficiency anaemia," *Best Practice & Research Clinical Haematology*, vol. 21, no. 3, pp. 193–202, 2008. DOI: 10.1016/j.beha.2008.01.002.

[80] L. B. Bailey, *Folate in Health and Disease*. CRC Press, 2009.

[81] M. F. Holick, "Vitamin d deficiency," *New England Journal of Medicine*, vol. 357, no. 3, pp. 266–281, 2007. DOI: 10.1056/NEJMra070553.

[82] R. S. Parker *et al.*, "Dietary fat and the absorption of fat-soluble vitamins," *The American Journal of Clinical Nutrition*, vol. 95, no. 4, pp. 1003–1012, 2012. DOI: 10.3945/ajcn.111.029320.

[83] M. Messina, "Soy and health update: Evaluation of the latest research," *Nutrition in Clinical Practice*, vol. 25, no. 1, pp. 9–19, 2010. DOI: 10.1177/0884533609356051.

[84] V. Bouvard *et al.*, "Carcinogenicity of consumption of red and processed meat," *The Lancet Oncology*, vol. 16, no. 16, pp. 1599–1600, 2015. DOI: 10.1016/S1470-2045(15)00444-1.

[85] M. López *et al.*, "Bioavailability of heavy metals and other toxic elements in the human diet," *Environmental Toxicology and Chemistry*, vol. 22, no. 4, pp. 817–823, 2003. DOI: 10.1897/01-586.

[86] A. G. Bostom *et al.*, "Homocysteine and risk of cardiovascular disease," *Journal of the American Medical Association*, vol. 281, no. 17, pp. 1638–1645, 1999. DOI: 10.1001/jama.281.17.1638.

[87] T. VanderWeele, "Imputation of missing dietary data," *Public Health Nutrition*, vol. 16, no. 1, pp. 74–82, 2013.

[88] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[89] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895. DOI: 10.1098/rspl.1895.0041.

[90] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[91] "Borutapy." Accessed: 2024-11-14. (), [Online]. Available: http://github.com/scikit-learn-contrib/boruta_py.

[92] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[93] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2022. DOI: 10.1109/TKDE.2021.3070203.

[94] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997. DOI: 10.1023/A:1007379606734.

[95] J.-P. Brans, "How to select and how to rank projects: The promethee method," *European Journal of Operational Research*, vol. 24, no. 2, pp. 228–238, 1982.

[96] J.-P. Brans and B. Mareschal, "The promethee 2 method: How to select and how to rank," *European Journal of Operational Research*, vol. 24, no. 2, pp. 228–238, 1986.

[97] T. Eftimov and D. Kocev, "Performance measures fusion for experimental comparison of methods for multi-label classification," in *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*, Mar. 2019.

Bibliography

Publications Related to the Thesis

Journal Articles

N. Reščič, T. Eftimov, B. Koroušić Seljak, and M. Luštrek, "Optimising an ffq using a machine learning pipeline to teach an efficient nutrient intake predictive model," *Nutrients*, vol. 12, no. 12, 2020, ISSN: 2072-6643. DOI: 10.3390/nu12123789. [Online]. Available: https://www.mdpi.com/2072-6643/12/12/3789.

N. Reščič, Ó. Mayora, C. Eccher, and M. Luštrek, "Food frequency questionnaire personalisation using multi-target regression," *Nutrients*, vol. 14, no. 19, p. 3943, 2022. DOI: 10.3390/nu14193943.

Conference Papers

N. Reščič, T. Eftimov, and B. K. Seljak, "Comparison of feature selection algorithms for minimization of target specific ffqs," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 3592–3595. DOI: 10.1109/BigData50022.2020.9378246.

N. Reščič, E. Valenčič, E. Mlinarič, B. Koroušić-Seljak, and M. Luštrek, "Mobile nutrition monitoring for well-being," ser. UbiComp/ISWC '19 Adjunct, London, United Kingdom: Association for Computing Machinery, 2019, pp. 1194–1197. DOI: 10.1145/3341162.3347076.

N. Reščič and M. Luštrek, "Dynamic feature ordering for personalised assessment of diet quality in ffqs," in *IEEE Big Data*, Sorrento, Italy: IEEE, 2023.

N. Reščič and M. Luštrek, "Question ranking for food frequency," in *Information Society - IS 2021, Volume A*, Ljubljana, Slovenia: Institut "Jožef Stefan", 2021, pp. 39–42. [Online]. Available: http://library.ijs.si/Stacks/Proceedings/InformationSociety/2021/IS2021_Volume_A.pdf.

Other Publications

Journal Articles

M. Gams, A. Marinko, N. Reščič, *et al.*, "Balancing health and socioeconomic impacts: A uniform framework for evaluating non-pharmaceutical interventions," *PLOS ONE*, pp. 1–25, 2025. DOI: 10.1371/journal.pone.0324232. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0324232.

V. Janko, N. Reščič, A. Vodopija, *et al.*, "Optimizing non-pharmaceutical intervention strategies against covid-19 using artificial intelligence," *Frontiers in Public Health*, vol. 11, pp. 1–11, 2023. DOI: 10.3389/fpubh.2023.1073581. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1073581/full.

J. Rufino, C. Baquero, N. Reščič, and A. Fernández Anta, "Using survey data to estimate the impact of the omicron variant on vaccine efficacy against covid-19 infection," *Scientific Reports*, vol. 13, p. 900, 2023. DOI: 10.1038/s41598-023-27951-3. [Online]. Available: https://www.nature.com/articles/s41598-023-27951-3.

E. Peklaj, N. Reščič, B. Korousić-Seljak, and N. Rotovnik-Kozjek, "New epidemic of malnutrition in young slovenian athletes," *Zdravstveno Varstvo: Slovenian Journal of Public Health*, vol. 62, no. 3, pp. 121–128, 2023. DOI: 10.2478/sjph-2023-0017. [Online]. Available: https://sciendo.com/article/10.2478/sjph-2023-0017.

E. Peklaj, N. Reščič, B. Korousić-Seljak, and N. Rotovnik-Kozjek, "Is red-s in athletes just another face of malnutrition?" *Clinical Nutrition ESPEN*, vol. 48, pp. 298–307, 2022. DOI: 10.1016/j.clnesp.2022.01.031.

S. Kalabakov, S. Stankoski, I. Kiprijanovska, *et al.*, "What actually works for activity recognition in scenarios with significant domain shift: Lessons learned from the 2019 and 2020 sussex-huawei challenges," *Sensors*, vol. 22, no. 10, p. 3613, 2022. DOI: 10.3390/s22103613. [Online]. Available: https://www.mdpi.com/1424-8220/22/10/3613.

V. Janko, G. Slapničar, E. Dovgan, *et al.*, "Machine learning for analyzing non-countermeasure factors affecting early spread of covid-19," *International Journal of Environmental Research and Public Health*, vol. 18, no. 13, pp. 6750-1–6750-33, 2021. DOI: 10.3390/ijerph18136750.

M. Gjoreski, V. Janko, G. Slapničar, *et al.*, "Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors," *Information Fusion*, vol. 62, pp. 47–62, 2020. DOI: 10.1016/j.inffus.2020.04.004.

Conference Papers

M. Jordan, N. Reščič, S. Kramar, M. Založnik, and M. Luštrek, "Smartchange risk prediction tool : Demonstrating risk assessment for children and youth," in *Slovenska konferenca o umetni inteligenci. Zv. A : zbornik 27. mednarodne multikonference Informacijska družba - IS 2024*, Ljubljana, Slovenia: Institut "Jožef Stefan", 2024, pp. 71–74. [Online]. Available: https://is.ijs.si/wp-content/uploads/2024/11/IS2024_Volume-A-1.pdf.

N. Reščič, J. Alberts, T. M. Altenburg, *et al.*, "Smartchange : Ai-based long-term health risk evaluation for driving behaviour change strategies in children and youth," in *2023 International Conference on Applied Mathematics & Computer Science (ICAMCS)*, Safi: ICAMS, 2023. DOI: 10.1109/ICAMCS59110.2023.00020.

J. Rufino, C. Baquero, N. Reščič, A. F. Anta, *et al.*, "Using survey data to estimate the impact of the omicron variant on vaccine efficacy against covid-19 infection," in *epiDAMIK 5.0 : The 5th International Workshop on Epidemiology Meets Data Mining and Knowledge Discovery at KDD 2022*, Washington: The Association for Computing Machinery, 2022. [Online]. Available: https://openreview.net/pdf?id=zBeXC8DcnSK.

V. Janko, N. Reščič, T. Tušar, M. Luštrek, and M. Gams, "What-if analysis of countermeasures against covid-19 in november 2020 in slovenia," in *Informacijska družba - IS 2021: zbornik 24. mednarodne multikonference*, 2021, pp. 42–45. [Online]. Available: http://library.ijs.si/Stacks/Proceedings/InformationSociety/2021/IS2021_Volume_D.pdf.

C. Picard, V. Janko, N. Reščič, M. Gjoreski, and M. Luštrek, "Identification of cooking preparation using motion capture data: A submission to the cooking activity recognition challenge," in *Human Activity Recognition Challenge*, 2021, pp. 103–113. DOI: 10.1007/978-981-15-8269-1_9.

M. Luštrek, N. Reščič, V. Janko, *et al.*, "Napovedovanje trendov in optimiziranje ukrepov v boju proti pandemiji covid-19: Tekmovanje xprize in naslednji koraki," in *Informacijska družba - IS 2021: zbornik 24. mednarodne multikonference*, 2021, pp. 52–55. [Online]. Available: http://library.ijs.si/Stacks/Proceedings/InformationSociety/2021/IS2021_Volume_D.pdf.

N. Reščič, V. Janko, D. Susič, *et al.*, "Finding efficient intervention plans against covid-19: Second place at the xprize pandemic response challenge," in *XV International Conference ETAI 2021 Proceedings*, ETAI Society of Macedonia, 2021, pp. 139–143. [Online]. Available: https://etai.org.mk/wp-content/uploads/2021/10/ETAI%202021%20Proceedings.pdf.

S. Stankoski, N. Reščič, G. Mežič, and M. Luštrek, "Real-time eating detection using a smartwatch," in *International Conference on Embedded Wireless Systems and Networks (EWSN) 2020*, ACM, 2020, pp. 247–252. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/3400306.3400354.

N. Reščič, M. Jordan, J. de Boer, I. Bierhoff, and M. Luštrek, "Mobile nutrition monitoring system: Qualitative and quantitative monitoring," in *Slovenska konferenca o umetni inteligenci: Informacijska družba - IS 2020*, 2020, pp. 80–83. [Online]. Available: http://library.ijs.si/Stacks/Proceedings/InformationSociety/2020/IS2020_Volume_A%20-%20SLAIS.pdf.

S. Kalabakov, S. Stankoski, N. Reščič, *et al.*, "Tackling the shl challenge 2020 with person-specific classifiers and semi-supervised learning," in *UbiComp/ISWC '20 Adjunct*, 2020, pp. 323–328. DOI: 10.1145/3410530.3414848.

M. Gjoreski, S. Kochev, N. Reščič, M. Gregorič, T. Eftimov, and B. K. Seljak, "Exploring dietary intake data collected by fpq using unsupervised learning," in *2019 IEEE International Conference on Big Data*, 2019, pp. 5126–5130. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9005715.

M. Luštrek, B. Fele, N. Reščič, and V. Janko, "Counting bites with a smart watch," in *Slovenska konferenca o umetni inteligenci: Informacijska družba - IS 2019*, 2019, pp. 49–52. [Online]. Available: http://library.ijs.si/Stacks/Proceedings/InformationSociety/2019/IS2019_Volume_A.pdf.

V. Janko, M. Gjoreski, C. D. Masi, N. Reščič, M. Luštrek, and M. Gams, "Cross-location transfer learning for the sussex-huawei locomotion recognition challenge," in *UbiComp/ISWC '19 Adjunct*, 2019, pp. 730–735. DOI: 10.1145/3341162.3344856.

N. Reščič and M. Luštrek, "Hierarchical multi-label classification for activity recognition," in *Slovenska konferenca o umetni inteligenci: Informacijska družba - IS 2018*, 2018, pp. 53–56. [Online]. Available: https://is.ijs.si/archive/proceedings/2018/files/Zbornik%20-%20A.pdf.

M. Gjoreski, V. Janko, N. Reščič, *et al.*, "Applying multiple knowledge to sussex-huawei locomotion challenge," in *UbiComp/ISWC'18 Adjunct*, 2018, pp. 1488–1496. [Online]. Available: https://dl.acm.org/citation.cfm?id=3267515.

V. Janko, N. Reščič, M. Mlakar, *et al.*, "A new frontier for activity recognition: The sussex-huawei locomotion challenge," in *UbiComp/ISWC'18 Adjunct*, 2018, pp. 1511–1520. [Online]. Available: https://dl.acm.org/citation.cfm?id=3267518.

E. Peklaj, N. Reščič, B. Koroušić-Seljak, and N. Rotovnik-Kozjek, "Red-simptomi pri mladih slovenskih športnikih," in *Zdravje otrok in mladostnikov : 6. znanstvena in strokovna konferenca z mednarodno udeležbo : zbornik povzetkov z recenzijo*, Založba Univerze na Primorskem, 2022, pp. 70–71. DOI: 10.26493/978-961-293-167-4. [Online]. Available: http://www.hippocampus.si/ISBN/978-961-293-167-4.pdf.

M. Ratajcec, A. Gradišek, and N. Reščič, "Mapping medical procedure codes using language models," in *Slovenska konferenca o umetni inteligenci = Slovenian Conference on Artificial Intelligence : Informacijska družba – IS 2025 = Information Society – IS 2025 : zbornik 28. mednarodne multikonference*, Ljubljana, 2025, pp. 63–66. [Online]. Available: https://is.ijs.si/wp-content/uploads/2025/11/IS2025_Volume-A.pdf.

N. Reščič, M. Jordan, S. Kramar, A. Krstevska, M. Založnik, M. Luštrek, *et al.*, "Smartchange risk prediction tool : Next-generation risk assessment for children and youth," in *Slovenska konferenca o umetni inteligenci = Slovenian Conference on Artificial Intelligence : Informacijska družba – IS 2025 = Information Society – IS 2025 : zbornik 28. mednarodne multikonference*, Ljubljana: Institut "Jožef Stefan", 2025, pp. 71–74. [Online]. Available: https://is.ijs.si/wp-content/uploads/2025/11/IS2025_Volume-A.pdf.

Book Chapters

V. Janko, M. Gjoreski, G. Slapničar, *et al.*, "Winning the sussex-huawei locomotion-transportation recognition challenge," in *Human Activity Sensing: Corpus and Applications* (Springer Series in Adaptive Environments), N. Kawaguchi, Ed., Springer Series in Adaptive Environments. Cham: Springer, 2019, pp. 233–250.

Edited conference proceedings

M. Luštrek, M. Gams, R. Piltaver, *et al.*, Eds., *Informacijska družba – IS 2022 = Information Society – IS 2022 : zbornik 25. mednarodne multikonference = proceedings of the 25th International Multiconference*, Ljubljana: Institut "Jožef Stefan", 2022, ISBN: 978-961-264-255-6. [Online]. Available: http://library.ijs.si/Stacks/Proceedings/InformationSociety/.

N. Reščič, O. Mayora, and D. Denkovski, Eds., *Vseprisotne zdravstvene storitve in pametni senzorji = Pervasive Health and Smart Sensing : Informacijska družba – IS 2022 = Information Society – IS 2022 : zbornik 25. mednarodne multikonference : zvezek H = volume H*, Ljubljana: Institut "Jožef Stefan", 2022, ISBN: 978-961-264-247-1. [Online]. Available: http://library.ijs.si/Stacks/Proceedings/InformationSociety/2022/IS2022_Volume-H%20-%20PHSS.pdf.

Biography

Nina Reščič received her degree in Applied Mathematics from the Faculty of Mathematics and Physics at the University of Ljubljana. She is enrolled in a PhD programme in Information and Communication Technologies (ICT) at the International Postgraduate School Jožef Stefan. Since 2017, she has been a researcher at the Jožef Stefan Institute, affiliated with the Department of Intelligent Systems. She is currently a researcher in the Complex Networks and Brain Dynamics Department (COBRA) of the Czech Academy of Sciences in Prague, Czech Republic. In addition, she is a part-time researcher at the Laboratory for Internal Combustion Engines and Electromobility (LICeM) at the Faculty of Mechanical Engineering, University of Ljubljana.

At the Jožef Stefan Institute, Nina's research focused on machine learning, artificial intelligence, and data analysis, with particular emphasis on healthcare applications and human-centred data. This work involved developing and evaluating models for sensor data, questionnaire-based data, and health-related information, including applications in human activity recognition using multimodal sensor data. Her research has also included active participation in international competitions and large-scale evaluation challenges. As a member of the winning team of the Sussex–Huawei Locomotion and Transportation Recognition Challenge, she contributed to a major benchmark competition in smartphone-based human activity recognition. She was also involved in a successful team submission to the Cooking Activity Recognition Challenge and participated in the JSIvsCOVID team, which won second place in the XPRIZE Pandemic Response Challenge, where her contribution primarily focused on epidemiological modelling.

At the Czech Academy of Sciences, Nina's research focuses on the analysis and modelling of complex networks, particularly epidemiological modelling and the study of spreading processes in complex systems. At LICeM, she applies artificial intelligence methods to energy-related problems, with a focus on predicting battery remaining useful life (RUL) using data-driven and machine learning approaches.

In addition to her scientific work, Nina is an experienced musician. She studied jazz flute at Gustav Mahler Privatuniversität für Musik in Klagenfurt. She has collaborated with wind and symphonic orchestras in Slovenia and abroad and has performed with a variety of ensembles spanning classical and popular music genres.