A blue and white logo

AI-generated content may be incorrect.  
UNIVERSITY OF HERTFORDSHIRE  
School of Physics, Engineering and Computer Science

MSc Artificial Intelligence and Robotics  
7COM1039-0509-2024 and MSc Advance Computer Science Masters Project  
Date: 14/07/2025

**Classification of Metabolic Disorders and Development of Personalised Nutrition Plans for Improved Health Management in the UK Population**

Name: Raj Ritesh Agrawal  
Student ID: 23028739  
Supervisor: Bente Riegler

**Table of Contents**

[Introduction 3](#_Toc203046041)

[Research questions 3](#_Toc203046042)

[Special Inquiry to be Undertaken 3](#_Toc203046043)

[Work and Methodology 4](#_Toc203046044)

[Tools and Techniques 4](#_Toc203046045)

[Deliverables 4](#_Toc203046046)

[Ethical, Legal, professional, and social considerations 5](#_Toc203046047)

[Progress to Date 7](#_Toc203046048)

[Introduction Chapter 7](#_Toc203046049)

[Literature Review Chapter 7](#_Toc203046050)

[Methodology Chapter 8](#_Toc203046051)

[EDA, Data Cleaning, and Data Collection 9](#_Toc203046052)

[Current Status and Ongoing Work 9](#_Toc203046053)

[Planned Work 11](#_Toc203046054)

[Task 1: Evaluation and Implementation of Models 11](#_Toc203046055)

[Task 2: System Design and System Integration - Recommendation 11](#_Toc203046056)

[Task 3: The Conclusion and Recommendations 12](#_Toc203046057)

[Task 4: Composition and Finishing of Dissertation 12](#_Toc203046058)

[Task 5: Tutor Check and Responses 12](#_Toc203046059)

[Task 6: Final Amendments and Poster Design 13](#_Toc203046060)

[Research Timeline 13](#_Toc203046061)

[References 16](#_Toc203046062)

**List of Figures**

[Figure 1:Gantt Chart 13](#_Toc202881201)

# Introduction

There are increased cases of metabolic disorders in the UK, characterized by type 2 diabetes, metabolic syndrome. As per the latest national surveys, they are associated with lifestyle issues such as poor diet, lack of physical exercise, and obesity (Rajput, Ashraff, and Siddiqui, 2022). Standardized dietary guidelines, on the one hand, can be quite useful in general to prevent National Prevention Week 2023: Mitigating the Impact of Influenza Pandemics Compendium preventative work, but are highly generalized and do not specifically work with the possible risk factors related to a particular person. The idea behind this project is based on the ability to integrate machine learning and evidence-based nutrition recommendations to develop custom interventions, which can be attributed to the metabolic status of an individual.

## Research questions

The following are the core research questions set to be addressed in this project:

1. Which model of supervised machine learning (e.g., Random Forest, XGBoost) is optimal in predicting metabolic disorders using structured health data?
2. What is the system of mapping classification results to evidence-based dietary plans based on the logic of rules in Python?
3. Which data preprocessing methods (imputation of missing values, feature scaling, dimensionality reduction) are likely to be most effective in preparing health datasets for successful classification?
4. What methods of interpretability, like SHAP, can be included to provide transparency to a given model and make a user have confidence in dietary recommendations?

## Special Inquiry to be Undertaken

The research will use a structured process that contains the following:

* Data Preparation: Sourcing, cleaning, and amalgamation of health data that acts as a representation of the UK population. This incorporates demographic data, rates (e.g., BMI, glucose, Lipid Tolerance Lists), and lifestyle factors.
* Model Development: Training of machine learning classifiers sequestered machine learning classifiers on processed datasets to assign metabolic cluster categories (Kyathanahally et al., 2021).
* Recommendation Logic: Developing an algorithm to match the output of classifications to evidence-based dietary recommendations based upon authoritative guidelines.
* Evaluation: Testing the predictive ability of the models and checking the suitability and clarity of recommendations on nutrition.

## Work and Methodology

In the technical work, several significant artefact elements are constructed and adjusted:

* A pipeline for data preprocessing that is a Python (pandas, NumPy) library that can process missing values, outliers, and feature engineering.
* A set of supervised learning algorithms trained with scikit-learn and XGBoost that are tested via k-fold cross-validation.
* As a system, an algorithmic recommendation module provides the output of the model in terms of dietary plans.
* Visual reporting and interpretation measures to explicate predictions, such as matplotlib and SHAP values (Mia et al., 2024).
* A demonstration interface, which allows the user to input sample data and outputs predictions and dietary advice.

## Tools and Techniques

Important libraries and tools are:

* Python (outputs)
* pandas, NumPy (arrays)
* XGBoost (machine learning), scikit-learn
* matplotlib, seaborn (visualization)
* SHAP (defensibility)
* Jupyter Notebook (development environment)

To measure the performance, the project will apply some of the established evaluation metrics, which include Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

## Deliverables

The next deliverables are intended to:

* Processed datasets cleaned up to be utilized in model training
* Validated models of classification
* The code of recommendation logic
* transcription of the preprocessing and modeling steps
* Graphical representations of the trend in data and the model performance
* A conclusion report about the process and results of the research
* A presentation in poster form with clear descriptions of aims, procedures, and findings
* An artefact of demonstration of the model and recommendation results

## Ethical, Legal, professional, and social considerations

**Ethical Considerations**

The project is based on the data available publicly and anonymised, and no clinical intervention or real-time monitoring of patients is carried out.

An important ethical issue is:

* **Model Bias:** It is possible that the machine learning algorithms will not perform that well on underrepresented groups in the data (eg, by age group, gender, ethnicity). To reduce this, the distribution of data is minimised and equilibrated with the help of over-sampling methods, like SMOTE (Synthetic Minority Over-sampling Technique), to make it fair.
* **Transparency and Trust:** Transparency and trust may be an issue where the black-box model can provide predictions that countries do not easily explain. Thus, SHAP (Shapley Additive explanations) is going to be used to give a graphical and textual representation of the justification of model outputs.
* **Recommendations Misuse or Misinterpretation:** All the dietary plans are created by the system; thus, there is a likelihood that the user may interpret them as clinical prescriptions.

**Legal Considerations**

The work is performed in the UK GDPR (General Data Protection Regulation) guidelines. There is no personally identifiable information (PII) in the dataset, and locally stored and processed data is processed through secured environments (Jørgensen et al., 2025).

The system design follows the general principles of responsible AI as well and aims to prevent algorithmic discrimination in its results, especially those related to health.

**Professional and Social Considerations**

The work follows the ACM Code of Ethics and the IEEE Code of Conduct, especially on the counts of:

* Public good and well-being
* Data responsibility
* Equity and equality
* Maintaining technical competence

The project has social initiatives to enhance positive health outcomes by increasing access to individualised nutritional consultations and rendering them data-driven.

**Ethics Approval**

At this point, there’s no need to apply the ethics approval due to the following reasons:

* The publicly accessible and anonymised datasets are used.
* Personal data will be neither gathered nor produced.
* No user or human study is carried out.

# Progress to Date

In this section, the progress of the MSc dissertation project to date is outlined, and four deliverables have already been delivered (Introduction, Literature Review, Methodology, and Data Preparation (including data collection, cleaning, and exploratory data analysis)).

## Introduction Chapter

The introductory chapter of the dissertation is done. It provides an evident background where it describes the increasing rates of occurrence of metabolic disorders like type 2 diabetes and metabolic syndrome in the UK population (Yun et al., 2022). It also throws light on the drawbacks of the common dietary treatment, especially the inability to personalise nutritional treatment.

The chapter constitutes the problem statement of the project, formulating the problem where there is a requirement to have data-driven and individualised naturalistic solutions to eating. It depicts the purpose of the project, which would be to classify metabolic disorders through the use of machine learning and to produce a personalised diet plan that fits each identified condition.

The aims, research questions, and an elaborate rationale have been articulated and introduced by conventional research practices. This chapter meets Objective 1 stated in Section 1 and is a solid basis for the rest of the dissertation.

## Literature Review Chapter

The second section completed is the Literature Review. It reads and compares numerous pertinent scholarly sources to determine principal patterns, methods, and knowledge gaps within the discipline of answers to the issue of metabolic health classification and personalised nutrition.

Three fundamental research studies have been exploited as main references to determine the methodological direction of this project:

* Sghaireen et al. (2022) proved that reliable predictive models of ensemble learning methods, such as Random Forest and XGBoost, are reliable in estimating risk factors of metabolic syndrome (Sghaireen et al., 2022).
* Kassem et al. (2025): The opportunity of using artificial intelligence in personal dietary planning was also discussed, which focused on such importance as the importance of employing interpretable models (Kassem et al., 2025).
* White et al. (2024): a clinical evidence-based diet because to manage such conditions as type 2 diabetes and hyperlipidemia. This project uses the nutritional backbone of a clinical evidence-based diet to provide evidence-based recommendations, detecting difficulties in care (White et al., 2024).
* Advanced hybrid data-balancing and counterfactual techniques are emphasized by Yang et al. (2024) to manage the problem of class imbalance and achieve enhancement of the target prediction accuracy of metabolic syndrome (Yang et al., 2024).
* Kibria et al. (2022) reveal that Random Forest combined with an explainable AI method, such as SHAP, had high accuracy in predicting diabetes at an early stage (Kibria et al., 2022).
* ChatDiet (Yang et al., 2024) is a chatbot that supports customized nutrition based on LLMs, bringing attention to interpretability by giving personalized data-driven diet recommendations through dialogue.
* Lin et al. (2021): use a clustering and ensemble technique to discriminate between obesity subtypes, which makes metabolic classification model interpretation and clinical application significantly more understandable (Li et al., 2021).

The comparison of previous studies of feature sets, the type of dataset, and the evaluation strategy was also conducted in the literature review. The given chapter leads directly to the achievement of Objective 2, which would ensure the scholarly rationale behind the selection of methods.

## Methodology Chapter

The third milestone realized is the completion of the Methodology chapter.

It describes:

* The recommendation logic is a rule-based system that connects the type of ordered disorder and particular dietary recommendations (based on the NHS and WHO recommendations).
* The metrics of evaluation: Accuracy, Precision, Recall, F1-score, and ROC-AUC.
* The libraries and tools that were utilized: Python, pandas, NumPy, scikit-learn, XGBoost, matplotlib, seaborn, and SHAP.

This chapter meets Objective 3, which describes the steps that will be used to execute the research and assess it.

## EDA, Data Cleaning, and Data Collection

The final important item that was achieved was the gathering and processing of datasets. The most important dataset is assembled with anonymised data of the UK Biobank and National Diet and Nutrition Survey (NDNS).

**Data Cleaning:**

* Missing values were imputed using applied median imputation
* Capped extreme outliers of the 99th percentile
* Categorical features, which were encoded using one-hot encoding
* Standardised numerical characteristics with the z-score normalisation

**Exploratory Data Analysis (EDA):**

* Histograms, boxplots, and density plots used as a visualisation tool
* Correlation matrices indicated associations to be great (e.g., BMI and triglycerides: r = 0.61)
* The dimensionality was reduced, and clustering was observed using Principal Component Analysis (PCA) (Hasan and Abdulazeez, 2021)
* Insights also validated model inputs and proved the relevance of selected features

The Python code was used to modularise all of the preprocessing scripts, and they were stored in Jupyter notebooks. This development fulfils Objectives 1 and 2 by giving us a clean, ready-to-use dataset to be used in modelling and the generation of recommendations.

## Current Status and Ongoing Work

At this point, the work actively stays at the stage of the model implementation and assessment. The pre-processed and cleaned dataset has been imported into a Jupyter Notebook workspace. Random Forest and XGBoost, as two of the ensemble classifiers, will be trained on a stratified training-test split and tested with a k-fold cross-validation (k=5).

GridSearchCV is implemented to tune hyperparameters, and SHAP (Shapley Additive Explanations) has been used to visualize the contribution of the features to the predictions, as the model fidelity criterion has been the visual requirements of the model. The development is in Python with the libraries of scikit-learn and XGBoost. It is measuring performance in terms of accuracy, precision, recall, F1-score, and ROC-AUC.

At the same time, the proposal has started to integrate the rule-based recommendation system. A logic mapping has been drawn to integrate each disorder d predicted (e.g., insulin resistance, dyslipidemia) with an associated nutrition plan according to NHS and WHO recommendations. Such a mapping is being realized with the help of the conditional logic in Python, and the examples of the output that are being produced are checked to verify the system's consistency and safety.

# Planned Work

This section offers the rest of activities that will be accomplished to complete the MSc dissertation project. All the tasks are charted out with purpose, output, and target schedule.

## Task 1: Evaluation and Implementation of Models

**Description:**

Accuracy, Precision, Recall, F1-score, and ROC-AUC will be applied in the evaluation of the performance of the classification models according to the assessment frameworks applied in the study by Sghaireen et al. (2022). These were the metrics in their study that were employed in validating the ensemble models in identifying metabolic syndrome risk factors. Moreover, it will be ensured that the model is interpretable since it will be done through SHAP (Shapley Additive Explanations) as it was done in the case of Kassem et al. (2025), where SHAP played an important role in increasing user trust and model transparency. Interpretability and clinical alignment, alongside the accuracy of models, will represent a variable against which they will be compared.

The cross-validation will be used to test the developed and built classifiers, like Random Forest, XGBoost, and Support Vector Machines, using scikit-learn (Mohammad agha, 2025).

**Deliverables:**

* Classification models finished
* Evaluation reports (confusion matrices, ROC curves)
* SHAP feature contributions visualisation

## Task 2: System Design and System Integration - Recommendation

**Description:**

It will be the step where the rule-based recommendation engine is implemented. Its logic will be designed to translate outputs of the classification into personalised diets in line with the recommendations of the NHS and WHO (Sawicki, Haslam, and Bhupathiraju, 2023). This system will be linked to the type of classification models to get a complete artefact.

**Deliverables:**

* Final model integrated artefact (+ reasoning logic)
* Interface of interaction with the user
* Examples of functional output

## Task 3: The Conclusion and Recommendations

**Description:**

In this dissertation section, the researcher will provide a synopsis of essential findings and analysis, comment on research questions and goals, and give implications of broader importance. The limitations of the project will also be indicated in the chapter, as well as recommendations about the future development or research.

**Deliverables:**

* Chapter on conclusion and recommendations
* Backing illustrations or statistics overview

## Task 4: Composition and Finishing of Dissertation

**Description:**

This is whereby all the sections that had been filled in (Introduction, Literature Review, Methodology, Results, Discussion, Conclusion, References) are compiled into the end dissertation document. There will also be time to do proofreading, formatting, and making it compatible with the submission requirements.

**Deliverables:**

* Full MSc Dissertation (10,000 words or more)
* Method of internal table of contents, appendices, and references list

## Task 5: Tutor Check and Responses

**Description:**

After it is done, the dissertation will be sent to the academic supervisor. The feedback will be collected and documented so that areas of improvement may be identified. Depending on the availability of the supervisor, an additional meeting or email review may be scheduled.

**Deliverables:**

* Draft dissertation is presented to the supervisor
* Notes on feedback

## Task 6: Final Amendments and Poster Design

**Description:**

The dissertation will undergo final edits based on the supervisors' comments. In continuation, a professional academic poster will be designed to give an overview of the goals and the approach taken in the project, as well as the results. The poster will be created as an aid to the final project demonstration and presentation.

# Research Timeline

|  |  |  |
| --- | --- | --- |
| **Week** | **Task** | **Deliverables** |
| **Week 1** | **Model Implementation** Train and evaluate classification models (Random Forest, XGBoost) using GridSearchCV and cross-validation. Integrate SHAP for model interpretability. | Trained models, evaluation reports (accuracy, F1, ROC-AUC), SHAP plots |
| **Week 2** | **Design Integration** Develop a logic-based recommendation engine using NHS/WHO dietary rules. Combine a prediction + recommendation pipeline in Python. | Finalised recommendation engine, test outputs, integrated artefact |
| **Week 3** | **Writing the Conclusion and Recommendations** Summarise findings, reflect on research questions, and suggest future improvements. | Completed “Conclusion” and “Recommendations” dissertation chapters |
| **Week 4** | **Dissertation Compilation** Merge all chapters, perform editing and formatting according to university submission standards. | Full dissertation draft with ToC, references, and appendices |
| **Week 5** | **Tutor Review and Final Edits** Submit the draft to the supervisor and apply the received feedback. Begin poster design. | Revised dissertation draft; poster outline |
| **Week 6** | **Poster Design and Final Submission** Complete A1-size academic poster, prepare slides for demonstration/presentation. Submit all components. | Final dissertation PDF, poster (digital), presentation slides, and notes |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task Name | Duration | Start | Finish | Predecessors |
| Project Initiation & First Class | 1 day | Tue 20-05-25 | Tue 20-05-25 |  |
| Drafting Dissertation Project Proposal (DPP) | 10 days | Wed 21-05-25 | Tue 03-06-25 | 1 |
| DPP Submission | 1 day | Wed 04-06-25 | Wed 04-06-25 | 2 |
| Supervisor Feedback on DPP | 8 days | Thu 05-06-25 | Mon 16-06-25 | 3 |
| Data Collection and Cleaning | 5 days | Tue 17-06-25 | Mon 23-06-25 | 4 |
| Exploratory Data Analysis (EDA) | 4 days | Tue 24-06-25 | Fri 27-06-25 | 5 |
| Model Design and Selection | 5 days | Mon 30-06-25 | Fri 04-07-25 | 6 |
| IPR Writing | 7 days | Mon 30-06-25 | Tue 08-07-25 | 6 |
| IPR Submission | 1 day | Wed 09-07-25 | Wed 09-07-25 | 8 |
| Submission to Supervisor for Feedback (Full Draft) | 1 day | Thu 10-07-25 | Thu 10-07-25 | 9,7,8 |
| Report Revision Based on Supervisor Feedback | 2 days | Fri 11-07-25 | Mon 14-07-25 | 10 |
| Final Formatting and Submission + Viva Prep | 1 day | Tue 15-07-25 | Tue 15-07-25 | 11 |



Figure 1: Gantt Chart

# References

Hasan, B.M.S. and Abdulazeez, A.M. (2021) ‘A review of principal component analysis algorithm for dimensionality reduction’, *Journal of Soft Computing and Data Mining*, 2(1), pp.20-30. <https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/download/8032/4199>

Jørgensen, B.N., Gunasekaran, S.S. and Ma, Z.G. (2025) ‘Impact of EU Laws on AI Adoption in Smart Grids: A Review of Regulatory Barriers, Technological Challenges, and Stakeholder Benefits’, *Energies*, 18(12), p.3002. <https://www.mdpi.com/1996-1073/18/12/3002>

Kassem, H., Beevi, A.A., Basheer, S., Lutfi, G., Cheikh Ismail, L. and Papandreou, D. (2025) ‘Investigation and Assessment of AI’s Role in Nutrition—An Updated Narrative Review of the Evidence’, *Nutrients*, 17(1), p.190. <https://www.mdpi.com/2072-6643/17/1/190>

Kibria, H.B., Nahiduzzaman, M., Goni, M.O.F., Ahsan, M. and Haider, J. (2022) ‘An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI’, *Sensors*, 22(19), p.7268. <https://www.mdpi.com/1424-8220/22/19/7268>

Kyathanahally, S.P., Hardeman, T., Merz, E., Bulas, T., Reyes, M., Isles, P., Pomati, F. and Baity-Jesi, M. (2021) ‘Deep learning classification of lake zooplankton’, *Frontiers in microbiology*, 12, p.746297. <https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.746297/pdf>

Lin, Z., Feng, W., Liu, Y., Ma, C., Arefan, D., Zhou, D., Cheng, X., Yu, J., Gao, L., Du, L. and You, H. (2021) ‘Machine learning to identify metabolic subtypes of obesity: a multi-center study’, *Frontiers in Endocrinology*, 12, p.713592. <https://www.frontiersin.org/articles/10.3389/fendo.2021.713592/pdf>

Mia, M., Pritom, M.M.A., Islam, T. and Hasan, K. (2024) ‘Visually Analyze SHAP Plots to Diagnose Misclassifications in ML-based Intrusion Detection’, *arXiv preprint arXiv:2411.02670*. <https://arxiv.org/pdf/2411.02670>

Mohammadagha, M. (2025) ‘Hyperparameter Optimization Strategies for Tree-Based Machine Learning Models Prediction: A Comparative Study of AdaBoost, Decision Trees, and Random Forest’, *Decision Trees, and Random Forest (April 11, 2025)*. <https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=5226457>

Rajput, S.A., Ashraff, S. and Siddiqui, M. (2022) ‘Diet and management of type II diabetes mellitus in the United Kingdom: a narrative review’, *Diabetology*, 3(1), pp.72-78. <https://www.mdpi.com/2673-4540/3/1/6>

Sawicki, C., Haslam, D. and Bhupathiraju, S. (2023) ‘Utilising the precision nutrition toolkit in the path towards precision medicine’, *Proceedings of the Nutrition Society*, 82(3), pp.359-369. <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/2F1211DAA712C79FECD8B7CEA84D6D08/S0029665123003038a.pdf/utilising_the_precision_nutrition_toolkit_in_the_path_towards_precision_medicine.pdf>

Sghaireen, M.G., Al-Smadi, Y., Al-Qerem, A., Srivastava, K.C., Ganji, K.K., Alam, M.K., Nashwan, S., and Khader, Y. (2022) ‘Machine learning approach for metabolic syndrome diagnosis using explainable data-augmentation-based classification’, *Diagnostics*, 12(12), p.3117. <https://www.mdpi.com/2075-4418/12/12/3117>

White, B., Ng, S.M., Agwu, J.C., Barrett, T.G., Birchmore, N., Kershaw, M., Drew, J., Kavvoura, F., Law, J., Moudiotis, C., and Procter, E. (2024) ‘A practical evidence-based approach to management of type 2 diabetes in children and young people (CYP): UK consensus’, *BMC medicine*, 22(1), p.144. <https://link.springer.com/content/pdf/10.1186/s12916-024-03349-4.pdf>

Yang, F., Qiao, Y., Hajek, P. and Abedin, M.Z. (2024) ‘Enhancing cardiovascular risk assessment with advanced data balancing and domain knowledge-driven explainability’, *Expert Systems with Applications*, 255, p.124886. <https://www.sciencedirect.com/science/article/pii/S0957417424017536>

Yang, Z., Khatibi, E., Nagesh, N., Abbasian, M., Azimi, I., Jain, R. and Rahmani, A.M. (2024) ‘ChatDiet: Empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework’, *Smart Health*, 32, p.100465. <https://www.sciencedirect.com/science/article/pii/S2352648324000217>

Yun, J.S., Jung, S.H., Shivakumar, M., Xiao, B., Khera, A.V., Won, H.H., and Kim, D. (2022). Polygenic risk for type 2 diabetes, lifestyle, metabolic health, and cardiovascular disease: a prospective UK Biobank study. *Cardiovascular diabetology*, 21(1), p.131. <https://link.springer.com/content/pdf/10.1186/s12933-022-01560-2.pdf>