



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

**Prediction of Early Gestational Diabetes Mellitus:
Using a Machine Learning Model.**

by

Md Mahadi Hasan

Supervisor: Mark Huntly

Second Marker: James Williams

Project submitted as part of the requirements for the
award of MSc Software Engineering and Artificial
Intelligence.

September 2026

DECLARATION

I, Md Mahadi Hasan declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.

Signed: ...MD Mahadi Hasan...

Date: ...07-09-2026...

FORM OF CONSENT

I Md Mahadi Hasan hereby consent that my Project, submitted in candidature for the M.Sc. Software Engineering and Artificial Intelligence_degree, if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.

Signed: ...Md Mahadi Hasan...

Date: ...07-09-2026...

Abstract

Gestational diabetes mellitus (GDM) is a frequent obstetric condition that has important maternal and fetal implications and early risk stratification is clinically useful. Unfortunately, many of the machine learning publications on GDM and diabetes prediction claim close-to-perfect accuracy that is hard to replicate, potentially due to class balancing before data splitting, lack of explainability, or failure to take into account clinical utility. In this research, a leakage-aware and explainable pipeline is created to compare the performance of three classical classifiers - Logistic Regression, Random Forest and XGBoost - for early prediction of GDM risk on the PIMA Indians Diabetes Dataset, which is reproduced. Physiologically impossible zeros were recorded and median imputed, features were min-max scaled, and the data were split 80/20 prior to any resampling. SMOTE was only applied to 10-fold stratified cross-validation folds and direct comparison revealed that the technique avoids an optimistic bias that existed in naive resampling. Tuning was done using cross-validated grid search and the evaluation was performed on a held-out test set with accuracy, precision, recall, F1-score, AUC-ROC, and with the help of confusion matrices, DCA, and SHAP explanations. XGBoost performed best overall with the highest recall (0.815) and lowest number of false negatives, and AUC (0.817) compared to Random Forest (0.819). The models consistently selected glucose and BMI as the top two predicting variables, in line with the clinical literature. Results are moderate but honest, avoiding the temptation to benchmark performance beyond the capabilities of the test, and provide a template that is defensible from a methodological perspective with regard to screening-oriented clinical prediction.

Table of Contents

Declaration	Error! Bookmark not defined.
Form of Consent	Error! Bookmark not defined.
Abstract	Error! Bookmark not defined.
Table of Contents	v
List of Figures	ix
List of Tables	xi
1 Chapter 1: Introduction	12
1.1 Context and Background	12
1.2 Statement of the Problem	13
1.3 Research Aim and Objectives	14
1.4 Significance of the Study	15
1.5 Structure of the Report	16
2 Chapter 2: Literature Review	17
2.1 Introduction to the Literature Review	17
2.2 Machine Learning Approaches to GDM Prediction	20
2.3 Deep Learning versus Classical Machine Learning for Tabular Clinical Data.....	21
2.4 Explainability Frameworks and Clinical Decision Support.....	23
2.5 Data Preprocessing, Class Imbalance and Dataset Limitations	24
2.6 Population Diversity, Novel Modalities and Emerging Directions.....	25
2.7 Literature Summary.....	25
3 Chapter 3: Research Methodology	27
3.1 Research Philosophy	27

3.2	Research Approach	28
3.3	Research Strategy: Design Science Research	29
3.4	Research Method and Strategy.....	30
3.5	Research Materials	32
3.5.1	Dataset Selection and Justification	32
3.5.2	Data Recording and Data Working Methods.....	33
3.5.3	Ethical Considerations	34
3.5.4	Required software, libraries and hardware.	34
3.6	Methods of Data Analysis and Evaluation.....	35
3.7	Chapter Summary.....	36
4	Chapter 4: Design and Implementation.....	38
4.1	Development Environment and Tools.....	38
4.2	Dataset.....	38
4.2.1	Class Balance	39
4.3	Exploratory Data Analysis	39
4.3.1	Implausible Zero Values	40
4.3.2	Correlation Structure.....	41
4.3.3	Association of Each Feature with the Outcome.....	42
4.4	Data Preparation.....	42
4.4.1	Missing Value Treatment.....	42
4.4.2	Feature Scaling.....	45
4.4.3	Training and Test Split.....	45
4.5	Leakage-Aware Cross-Validation and Resampling	46
4.6	Model Training and Hyperparameter Tuning	46
4.7	System Testing of the Real-Time Prediction Interface	47
4.7.1	TC01: Low-Risk Patient Prediction.....	48
4.7.2	TC02: High-Risk Patient Prediction	50

4.7.3	TC05: Boundary Values, Minimum	51
4.7.4	TC06: Boundary Values, Maximum.....	51
4.7.5	TC07: About This Study Tab.....	52
4.7.6	Outstanding Test Cases.....	53
4.8	Chapter Summary.....	53
5	Chapter 5: Evaluation of Results.....	55
5.1	Headline Performance	55
5.2	Confusion Matrices	56
5.3	Discrimination: ROC Curves	57
5.4	Clinical Usefulness: Decision Curve Analysis.....	58
5.5	Explainability: SHAP Analysis	59
5.5.1	Feature Importance Across Models	60
5.5.2	Explaining a Single Prediction.....	62
5.6	Overall Model Comparison.....	63
5.7	Benchmarking Against the Literature	64
5.8	Chapter Summary.....	65
6	Chapter 6: Conclusions and Recommendations.....	66
6.1	Overview	66
6.2	Summary of Key Findings	66
6.3	Contribution	67
6.4	Limitations	67
6.5	Recommendations for Future Work.....	68
6.6	Concluding Remarks	68
7	Chapter 7: Reflection	69
7.1	Introduction	69
7.2	Technical Learning.....	69
7.3	Research and Methodological Learning.....	70

7.4	Project and Time Management	70
7.5	What I Would Do Differently	71
7.6	Concluding Reflection.....	71
8	References	72

List of Figures

Figure 3.1 Research design of this study, showing the position adopted at each layer	28
Figure 3.2 Design Science Research (DSR) Methodology Flowchart	29
Figure 3.3 ML System Architecture Diagram	31
Figure 4.1 Class balance of the Outcome variable	39
Figure 4.2 Count of physiologically implausible zero values before cleaning.....	40
Figure 4.3 Correlation heatmap of the raw features and Outcome	41
Figure 4.4 Absolute correlation of each feature with the Outcome, after scaling	42
Figure 4.5 Distribution of each affected feature before and after median imputation.....	44
Figure 4.6 TC01, low-risk prediction result.....	49
Figure 4.7 TC01, SHAP explanation and patient values used.....	49
Figure 4.8 TC02, high-risk prediction result and SHAP explanation.....	50
Figure 4.9 TC05, minimum boundary values	51
Figure 4.10 TC06, maximum boundary values	52
Figure 4.11 TC07, About this Study tab.....	53
Figure 5.1 Confusion matrices for the three classifiers on the test set	56
Figure 5.2 ROC curves for the three classifiers on the test set.....	57
Figure 5.3 Decision curve analysis showing net benefit against threshold probability	58
Figure 5.4 SHAP summary for XGBoost, features ordered by mean absolute impact	60
Figure 5.5 SHAP summary for Random Forest.....	61
Figure 5.6 SHAP summary for Logistic Regression	61
Figure 5.7 SHAP waterfall for a single test patient under the XGBoost model	62
Figure 5.8 Model comparison across all five metrics	63

Figure 5.9 Radar view of model comparison across all five metrics.....	63
--	----

List of Tables

Table 2.1 Summary of Reviewed Studies Dataset, Tools/Methods, and Problem-Solving Approach.....	17
Table 2.2 Comparative performance of XGBoost across the reviewed studies	22
Table 4.1 Tuned hyperparameters and cross-validation AUC for each classifier	47
Table 4.2 System Testing Summary	48
Table 5.1 Test-set performance of the three classifiers across all metrics.....	55
Table 5.2 This study's results in the context of comparable PIMA-based studies.....	64

1 Chapter 1: Introduction

1.1 Context and Background

Gestational diabetes mellitus (GDM) is a very important metabolic disorder in contemporary obstetric practice. Given the high prevalence of GDM around the world (e.g., approximately 14 percent of pregnancies meet IADPSG criteria [1]), and the known links between GDM and a variety of adverse maternal and fetal outcomes [2], early and accurate risk identification is of clinical significance. The prediction of GDM using artificial intelligence (AI) and machine learning (ML) has thus emerged as a vast and expanding area of research; a recent systematic review found 126 studies related to screening, diagnosis, management, and prediction of the outcomes of GDM [3]. Despite this activity, clinically interpretable ML tools for GDM are still limited and sparse of tools have been introduced into routine clinical care [4]. This disconnect between growing evidence and an enduring translation issue is set against the backdrop of the current study.

There have been different ways of approaching prediction in the context of GDM. Dietary data have been integrated with clinical and demographic data: XGBoost performed better than Logistic Regression and LightGBM with glucose, age, BMI, triglycerides and HDL as the top features [5]. At national scale, birth-cohort information has been utilized: for instance, 82,698 women with pregnancy in Japan, where a Gradient Boosting Decision Tree gave the best performance, and led to the discovery of factors that were under-reported, such as maternal birthweight [6]. Kang et al. [7] used SHAP to select features at various stages of pregnancy across multiple centres in South Korea, employing LightGBM and XGBoost. In other areas, the data from wearables has been used alongside matrix tensor factorisation to predict GDM-related biomarkers before traditional screening [8], and the use of wearables data in conjunction with work in NHS trusts has been undertaken for risk stratification of confirmed GDM, with internal and external validation across two NHS trusts [9].

In addition to the work presented in this paper, there exist many other papers that predict diabetes based on the same PIMA Indians Diabetes Dataset, but on a much larger scale. Some of these studies discuss class balancing using SMOTE, explainability using SHAP [10], AutoML pipelines using SHAP, LIME and Integrated Gradients [11], feature engineering for recursive feature elimination [12], and boosting ensembles like AdaBoost and XGBoost [13].

One of the frequent questions is if deep learning provides any benefit on top of tabular clinical data: In the GDM domain, XGBoost proved to perform better than the proposed deep architectures in a systematic benchmark with fewer tuning parameters [14], and, with nuances, in an RNN-LSTM comparison [15].

1.2 Statement of the Problem

Despite this voluminous body of literature, there are still some methodical trends that limit the scientific rigor and clinical maturity of this field. Abousaber et al. [16] compared the imbalance-handling techniques among the three diabetes datasets and found that some of the PIMA-based results had perfect or near-perfect results and explained that these might be due to overfitting or used SMOTE before the train-test split. Similarly, Altamimi et al. [17] achieved an accuracy of 98.59 percent with a stacked ensemble model on a small data set that was used repeatedly. It is also true that the same concern can be made for ensemble study (AdaBoost+XGBoost) that achieves high AUC score without using any explainability method [13] and for feature-engineering study that achieves 94 percent accuracy without conducting any explainability analysis and without paying attention to the class imbalance of the PIMA data [12].

A second frequent limitation is the sample size, and the generalisability to one site. There are several studies specific to GDM with small cohorts of women from a single centre, for example, a cohort of 554 women from China [5], 253 women in a number of hospitals in Nigeria [18] and a small pilot study in South Africa [8], making it difficult to generalise their results to other populations. This is in contrast to the lesser discrimination (AUC 0.67 to 0.80) but presumably more realistic results in the larger multi-centre cohorts mentioned above as compared to some smaller single-centre or poorly validated studies reporting near-perfect discrimination [6], [7]. Throughout the field this is a tension between what is achieved in the benchmark and what is actually happening in the field, as well as what is driving a ‘cautious, appropriately validated approach’ in the present study.

A third constraint is the lack of uniformity in adopting explainability. In contrast, SHAP has not been applied in a several other methodologically similar studies [12], [18] yet has been applied in some recent studies [10], [11], including one using a much larger sample size of 3,649 women from NHANES to predict GDM [19]. Nor have there been broadly adopted tools

of evaluation, such as Decision Curve Analysis (DCA) that has assessed if the model would have had a clinically relevant performance advantage over a 'treat all' or 'treat none' hypothesis beyond discrimination and calibration assessment [20].

Overall these patterns illustrate a clear and tractable problem: that of having no existing repeatable classical ML pipeline for early GDM prediction, with principled, leakage-free class balancing, systematic comparison of multiple algorithms, and evaluation of these algorithms with realistic performance benchmarks instead of potentially overfitted benchmarks.

This initial research question is broken down into four sub-questions to ensure that the problem identified above is directly related to the aim and the four objectives of this research;

- RQ1 (Objective 1): What trends in methodology, comparative performance, and gaps in research and development is seen in the current body of machine learning and deep learning-based literature on GDM prediction?
- RQ2: How can Logistic Regression, Random Forest and XGBoost be built and trained on leakage-free, SMOTE-balanced PIMA data to create a reproducible prediction artefact for GDM?
- RQ3 What are the performances of the three classifiers based on the following evaluation metrics: Accuracy, F1-score, AUC-ROC, Precision and Recall; What are the most influential individual clinical features in their predictions, according to SHAP analyses?
- RQ4: What is the best balance between predictive accuracy and clinical interpretability when using the three classifiers to stratify the risk of early GDM?

1.3 Research Aim and Objectives

The aim of this study is to develop, train and compare a machine learning model based on clinical and demographic data for early prediction of GDM, with the goal of identifying the best predictive model for early clinical risk stratification, which is both accurate and interpretable.

This is achieved by four objectives:

1. To systematically review the published literature to critically evaluate the current machine learning and deep learning methods used for the prediction of GDM, determine trends in methodology, comparative performance benchmarks, and unmet research gaps.
2. To create and train an ML-based prediction artefact using the Logistic Regression, Random Forest, and XGBoost classifiers on the pre-processed and SMOTE-balanced clinical dataset from the PIMA Indians Diabetes Dataset, with class balancing strictly performed within cross-validation folds to avoid the leakage issues identified in the literature [16].
3. To test the developed models with standard performance metrics: Accuracy, F1-score, AUC-ROC, Precision, and Recall, and use SHAP analysis to interpret the impact of individual clinical features on each model's predictions.
4. To compare the results of the evaluated models and determine which one is the best prediction classifier of the GDM, explicitly considering the balance between accuracy in the prediction and clinical interpretability, and where possible, benchmarking with the more conservative range of AUCs reported across the broader literature [6], [7].

1.4 Significance of the Study

The methodological and clinical importance of the study. Technically, it addresses a need across the field of interpretable ML approaches in GDM [3], [4] by providing only one transparent and reproducible comparative pipeline instead of one model result. It also takes a clinical-utility viewpoint, based on Decision Curve Analysis [20] to ensure that assessment is focused on real decision support and not just discrimination.

Clinically, early risk stratification of GDM is important for mother and baby, as there has been ample evidence of adverse outcomes [2]. A few examples of ML-based prenatal modelling are early prediction of women becoming Type 2 Diabetes after pregnancy on the basis of mid-gestation features [21], or the use of data from fetal imaging to predict growth conditions for the child after birth. However, the responsible application of AI in healthcare requires careful consideration of privacy and data security, algorithmic bias, transparency and comprehensive

clinical validation prior to implementation, which are addressed in the design phase and not an afterthought [22]. This study is conducted from the beginning with these considerations in mind, and further, involves analysis of an anonymous, publicly available secondary data set, thus avoiding the potential for ethical issues in primary data collection.

1.5 Structure of the Report

The rest of the report proceeds in the following manner. Chapter 2, Literature Review, builds the evidence base in thematic sections covering classical ML methods for GDM prediction, deep learning vs classical ML, explainability and clinical decision support, and dataset and class-imbalance challenges. Research methodology is explained in Chapter 3, which includes the research philosophy, approach and Design Science Research strategy and the justification for Logistic Regression, Random Forest and XGBoost. The ML system design, the pre-processing, the class balancing and features decisions are described in Chapter 4, Design and Implementation, along with the real-life implementation of the artefact. Evaluation of Results (Chapter 5): The results of the three classifiers are reported and analysed based on the metrics outlined in Objective 3. The results are presented and related to the aim and the objectives in Chapter 6, Conclusions and Recommendations, while the future work is identified. In the chapter Reflection, the researcher reflects on his or her own learning and development.

2 Chapter 2: Literature Review

2.1 Introduction to the Literature Review

In this chapter, a critical review of many published works between 2022 and 2026 was used to discuss the application of machine learning in predicting gestational diabetes mellitus (GDM). In a recently published systematic review, the researchers mapped study designs, types of input data, model types, and validation strategies for the entire range of AI applications in GDM care, including screening and diagnosis, management and prediction of outcomes [3]. Since the key methodological questions (which classifiers are best suited, explainability, and how to deal with the limitations of the datasets) are more natural themes than years, the review is organized thematically, not chronologically. It lays out classic machine learning methods, the debate between deep learning and classic machine learning, explains machine learning explainability and clinical decision support, data pre-processing and class imbalance, population diversity and emerging directions, and ends with a summary of the gap this study fills. Table 2.1 summarizes the data, tools, and methods, and problem-solving process used for each of the empirical sources discussed.

Table 2.1 Summary of Reviewed Studies Dataset, Tools/Methods, and Problem-Solving Approach

Ref	Study	Dataset	Tools / Methods Used	Approach and Key Result
[5]	Ding et al., 2024	554 women, single-centre, China	Logistic Regression, XGBoost, LightGBM, Random Forest-based feature selection	Used Random Forest to select dietary and clinical predictors, then compared three classifiers; XGBoost best (AUC 0.788)
[6]	Watanabe et al., 2023	82,698 women, Japan (JECS cohort)	Random Forest, Gradient Boosting, Decision Tree, SVM, Logistic Regression	Compared four classifiers on pre-pregnancy and first-trimester data; GBDT performed best (AUC 0.67-0.76)

Ref	Study	Dataset	Tools / Methods Used	Approach and Key Result
[7]	Kang et al., 2023	34,387 pregnancies, multi-centre, South Korea	LightGBM, XGBoost; SHAP-based feature selection	Used SHAP to select variables at each gestational timepoint, then trained both classifiers; AUC 0.71-0.80
[18]	Basil et al., 2024	253 women, multi-hospital, Nigeria	Multivariable logistic regression; nomogram; Hosmer-Lemeshow calibration	Built a 4-variable clinical nomogram for low-resource settings; AUC 0.814
[14]	Shwartz-Ziv and Armon, 2022	Multiple public tabular datasets	Systematic deep learning vs. XGBoost benchmarking	Benchmarked several DL architectures against XGBoost across datasets; XGBoost consistently won with less tuning
[15]	Kurt et al., 2023	489 patients, prospective, Turkey	RNN-LSTM with Bayesian hyperparameter optimisation; Random Forest, SVM	Tuned an RNN-LSTM via Bayesian search and compared it to classical models; DL reached AUC 0.98 (small-sample caution)
[10]	Netayawijit et al., 2025	Kaggle diabetes dataset	Logistic Regression, SVM, Random Forest, Gradient Boosting, XGBoost; SMOTE fitted within CV folds; SHAP	Balanced classes with leakage-aware SMOTE, then applied SHAP; RF with SMOTE best (AUC 0.998)
[11]	Hasan et al., 2025	PIMA Indians Diabetes Dataset	AutoGluon AutoML; SHAP, LIME, Integrated Gradients, counterfactual analysis	Used AutoML to build an ensemble, then applied four explainability methods to cross-check feature importance (76.6% accuracy)

Ref	Study	Dataset	Tools / Methods Used	Approach and Key Result
[9]	Yang et al., 2022	1,148 + 709 pregnancies, two NHS trusts, UK	Tree-based regression models; EHR and daily glucose data	Trained a risk-stratification model internally, then externally validated it on a second NHS trust's data
[21]	Kumar et al., 2022	561 pregnancies, Singapore	CatBoost; SHAP	Built a 2-feature CatBoost model to predict progression to T2DM, interpreted via SHAP; AUC 0.86
[8]	Kolozali et al., 2024	Small pilot cohort, South Africa	Decision Tree and Random Forest regressors; Coupled-Matrix Tensor Factorisation; Elastic Net feature selection	Combined wearable sensor data with tensor factorisation to forecast biomarkers 13-16 weeks early
[13]	Sampath et al., 2024	PIMA-derived, 829 records after pre-processing	Mean imputation, IQR outlier rejection, correlation-based feature selection, SMOTE, AdaBoost + XGBoost ensemble	Cleaned outliers and balanced classes, then ensembled two boosting methods; AUC 0.968
[12]	El Massari et al., 2024	PIMA Indians Diabetes Dataset	Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost; RFECV and four other feature-selection methods	Applied five feature-selection techniques and hyperparameter tuning across five classifiers; best 94% accuracy
[16]	Abousaber et al., 2025	Three diabetes datasets incl. PIMA	SMOTE, ADASYN, Borderline-SMOTE, RandomUnderSampler, SMOTEENN; 12	Statistically compared five imbalance-handling techniques across twelve

Ref	Study	Dataset	Tools / Methods Used	Approach and Key Result
			classifiers; t-test validation	classifiers; flagged leakage risk in PIMA results
[17]	Altamimi et al., 2024	PIMA Indians Diabetes Dataset	KNN Imputer; stacked ensemble voting classifier	Used KNN imputation to address zero-values, then stacked three classifiers by vote; 98.59% accuracy
[19]	Xiong et al., 2026	3,649 women, NHANES, USA	XGBoost, LightGBM, Random Forest, SVM, Gaussian Naive Bayes, KNN; VIF and Boruta feature selection; SHAP	Extended dietary-nutrient prediction to a large national sample using SHAP-guided feature selection across six classifiers
[20]	Piovani et al., 2023	Not applicable (methodological tutorial)	Decision Curve Analysis, net benefit calculation	Explains how to assess a model's clinical utility beyond discrimination/calibration, via net benefit versus threshold probability

2.2 Machine Learning Approaches to GDM Prediction

The literature closest to this problem uses classical ML classifiers to predict early GDM from clinical and demographic data. Ding et al. [5], in 554 pregnant women, used Random Forest for feature selection and then compared Logistic Regression, XGBoost and LightGBM; XGBoost performed best (AUC 0.788 versus 0.712 for Logistic Regression), confirming glucose, age, BMI, triglycerides and HDL as key features. The study is methodologically sound but single-centre and omits explainability, a gap the present work addresses through integrated SHAP analysis. At much larger scale, Watanabe et al. [6] compared Random Forest, a Gradient Boosting Decision Tree, SVM and Logistic Regression on pre-pregnancy and first-trimester data from 82,698 women in the Japan Environment and Children's Study, finding the Gradient

Boosting Decision Tree best and surfacing under-reported predictors such as maternal birthweight. Notably, the AUCs there (0.67 to 0.76) are well below those of smaller single-centre studies, a recurring pattern that raises doubt about how much apparent high performance reflects smaller, less heterogeneous samples.

A similarly large multi-centre study conducted by Kang et al. applied SHAP-based feature selection to determine the most informative variables at each time point, and then trained LightGBM and XGBoost on the variable sets at each time point using 34,387 pregnancies in South Korea, reporting AUCs ranging from 0.71 to 0.80 for each time point [7]. It's interesting that this is a convergence of two large, independent, multi-centre studies on a similar AUC range, and not the higher numbers reported by smaller numbers of patients, and it's one of the more important patterns this study considers when setting realistic performance expectations. In contrast, in a resource poor setting in Nigeria, Basil et al. created a four variable logistic regression nomogram using the Hosmer-Lemeshow test, which was validated internally with an AUC of 0.814, and included only age, first trimester BMI, family history of diabetes and history of foetal macrosomia [18].

2.3 Deep Learning versus Classical Machine Learning for Tabular Clinical Data

A key question, directly relevant to the classifiers used here, is whether deep learning offers any benefit over classical ML on tabular clinical data. Shwartz-Ziv and Armon [14] benchmarked several deep architectures against XGBoost across public tabular datasets and found XGBoost consistently superior and far less demanding to tune. Although not medical-specific, this mirrors the GDM literature. Kurt et al. [15] trained an RNN-LSTM with Bayesian optimisation against Random Forest and SVM on 489 Turkish patients, reporting 95 percent sensitivity, 99 percent specificity and AUC 0.98, figures far above any classical result in this review; given the small sample and absence of external validation, these are plausibly optimistic rather than evidence of a real architectural advantage.

Combined, these two sources justify the methodological choice of this study to use classical ML classifiers instead of deep learning architectures, as discussed in Chapter 1. The 768 entries in the PIMA Indians Diabetes Dataset are much smaller than the number of records needed to make deep learning architectures truly better, and the almost flawless metrics reported by Kurt

et al. on a similarly small number of records point exactly towards the danger of overfitting, which this study takes pains to prevent by using cross validation and realistic benchmarks [14], [15].

Having established that classical ML suits a dataset of this size, the evidence points to XGBoost as the strongest anchor classifier, for four reasons. First, XGBoost applies sequential gradient boosting, each tree fitting the residual error of the ensemble so far [23], capturing non-linear feature interactions that a linear baseline cannot. Second, it repeatedly led the reviewed cohorts: it beat Logistic Regression and LightGBM in Ding et al. [5] (AUC 0.788 versus 0.712), reached 94 percent accuracy after recursive feature elimination in El Massari et al. [12], and was among the two leading classifiers in Kang et al. [7]. Third, on comparable tabular data it outperforms purpose-built deep architectures with far less tuning [14], decisive at PIMA's modest scale. Fourth, it is natively compatible with tree-based SHAP [24], so the feature-importance analysis central to Objective 3 follows directly from the trained model. XGBoost is therefore retained as the primary gradient-boosted classifier, with Logistic Regression as an interpretable linear baseline and Random Forest as an alternative ensemble comparator.

Table 2.2 Comparative performance of XGBoost across the reviewed studies

Study	Classifiers compared	Best performer and result	Relevance to XGBoost choice
Ding et al. [5]	Logistic Regression, XGBoost, LightGBM	XGBoost best (AUC 0.788 vs 0.712 for Logistic Regression)	XGBoost beat both the linear baseline and LightGBM on clinical and dietary features
Kang et al. [7]	LightGBM, XGBoost (SHAP-selected features)	AUC up to 0.804 across gestational timepoints	XGBoost among the two leading classifiers on a large multi-centre cohort
El Massari et al. [12]	Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost	XGBoost best (94% accuracy after RFECV)	XGBoost led five classifiers on the PIMA dataset used in this study

Sampath et al. [13]	AdaBoost ensembles with XGBoost and Random Forest, plus SMOTE	AdaBoost with XGBoost best (AUC 0.968)	Gradient boosting central to the strongest ensemble on PIMA-derived data
Shwartz-Ziv and Armon [14]	Multiple deep learning architectures vs XGBoost	XGBoost consistently best, with less tuning	Confirms XGBoost outperforms deep learning on tabular data of this kind

2.4 Explainability Frameworks and Clinical Decision Support

Explainability is a growing focus in this literature and directly informs the SHAP-based approach taken here. Netayawijit et al. [10] trained five classifiers with SMOTE applied strictly within each CV fold to prevent leakage, then applied SHAP [24]; Random Forest with SMOTE was best (96.9 percent accuracy, AUC 0.998), and SHAP consistently identified glucose and BMI as the most important features. This leakage-aware SMOTE safeguard is adopted here and discussed in Section 2.5. Hasan et al. [11] built an AutoGluon ensemble on the same dataset used here and applied SHAP, LIME, Integrated Gradients and counterfactual analysis, reaching 76.6 percent accuracy while confirming glucose and BMI as the dominant predictors across all four methods. Although its accuracy trails several standalone classical studies, the agreement of feature importance across four independent methods supports the reliability of SHAP, consistent with Salih et al. [25].

Several studies extend explainability into clinical decision support and risk stratification. Yang et al. [9] built tree-based regression models from daily glucose readings and EHR data for GDM management, trained and internally tested on 1,148 pregnancies at Oxford University Hospitals NHS Trust and externally validated on 709 patients at a second trust, a level of external validation rare in this field. In line with the early-stratification argument of Chapter 1, Kumar et al. [21] used a CatBoost model with only two mid-gestation features to predict progression from GDM to postpartum Type 2 Diabetes (AUC 0.86). Kolozali et al. [8] pushed early prediction furthest, using wearable IoT sensor data with Decision Tree and Random Forest regressors, Coupled-Matrix Tensor Factorisation and Elastic Net selection to forecast

GDM-associated biomarkers 13 to 16 weeks before standard screening. This pilot work shows the classification approach used here could be extended with novel data modalities.

2.5 Data Pre-processing, Class Imbalance and Dataset Limitations

Class imbalance and pre-processing recur throughout this work and shape the pipeline used here. Sampath et al. [13] combined mean imputation, IQR-based outlier rejection and correlation-based selection with an AdaBoost-XGBoost ensemble and SMOTE, reaching AUC 0.968 (AdaBoost with Random Forest gave 0.962, Random Forest alone 0.963); the study omits explainability, and its outlier step cut the usable sample to 829 records, which the present work avoids by using median imputation to retain the full sample. El Massari et al. [12] applied recursive feature elimination with cross-validation plus four further selection methods and tuning across five classifiers, with XGBoost reaching 94 percent accuracy after RFECV, but used neither SHAP nor any imbalance handling, both of which this pipeline supplies given PIMA's well-established imbalance.

To underline the risk of overstated performance from inappropriate pre-processing, Abousaber et al. [16] statistically compared five imbalance-handling techniques (SMOTE, ADASYN, Borderline-SMOTE, RandomUnderSampler and SMOTEENN) across twelve classifiers and three diabetes datasets, using t-tests. Random Forest with SMOTE was the best statistically validated combination, but the authors warn that some PIMA combinations produced perfect or near-perfect scores caused by overfitting or by applying SMOTE before the train-test split. This directly motivates the leakage-aware SMOTE design applied strictly within cross-validation folds in this study (Objective 2), a concern also raised in broader resampling reviews of imbalanced medical data [26]. Relatedly, PIMA's missing and physiologically implausible zero values were handled by Altamimi et al. [17] with a KNN imputer and a stacked ensemble, reaching 98.59 percent accuracy, 99.26 percent precision and 99.75 percent recall; as with similar PIMA results, such near-perfect figures on a heavily reused benchmark are treated here as a warning sign rather than a target, echoing the gap between reported ML performance and clinical readiness noted by Lu et al. [4].

2.6 Population Diversity, Novel Modalities and Emerging Directions

Beyond these themes, some broader directions frame this study and future work. Xiong et al. [19] applied VIF and Boruta selection across six classifiers to 3,649 women in NHANES, scaling up dietary-nutrient prediction of GDM and identifying race and ethnicity, BMI, protein, fibre and micronutrients as core predictors. This shows how a small-scale finding, such as the dietary approach of Ding et al. [5], can be corroborated at larger scale, while illustrating the population-diversity concern also addressed by the Nigerian nomogram of Basil et al. [18], where leading predictors vary across populations. A directly relevant limitation, revisited in Chapter 7, is that PIMA covers only Pima Indian women over 21.

Another new direction is the final way in which model performance should be evaluated, this time not statistically, but clinically. Decision Curve Analysis (DCA) is explained by Piovani et al. as a way to determine whether a prediction model would be clinically useful compared to treating everyone or no one, and is based on net benefit and threshold probabilities instead of discrimination and calibration metrics [20]. This is a general methodological tutorial and not a study specific to GDM, but the logic involved in this tutorial is directly relevant to the evaluation framework developed in Chapter 3, which is part of the approach used in this study to evaluate the clinical utility. It is reviewed again in Chapter 3, where it is part of the clinical-utility-oriented approach used to evaluate the clinical utility.

2.7 Literature Summary

Table 2.1 highlights several trends across the twenty sources reviewed. Random Forest and XGBoost are repeatedly the strongest classical classifiers for GDM [6], [7], with AUCs typically 0.70 to 0.80 in realistic large-cohort and multi-centre settings and much higher, sometimes implausibly so, in smaller single-centre or weakly validated PIMA studies [13], [15], [16], [17]. Glucose and BMI emerge as the two dominant predictors regardless of classifier, tool or explainability method [10], [11]. Many studies, however, omit explainability or apply SMOTE in ways that risk leakage and over-optimistic evaluation [12], [13], [16]. Deep learning, despite occasional extreme results, consistently needs more data than GDM prediction usually provides and overfits on small clinical cohorts [14], [15].

This study is well-suited to directly fill these gaps. This study is designed to produce a methodologically transparent and clinically credible contribution to a field, as the systematic

review mapping of 126 published studies clearly indicates, yet remains lacking in many of its published contributions in terms of methodological rigour [3]; specifically, by using Logistic Regression, Random Forest and XGBoost in one reproducible pipeline, by applying SMOTE only within the folds of the cross-validation, to prevent the leakage concern raised explicitly in this literature [16], by integrating SHAP explainability consistently across all three classifiers, and by evaluating results against the more conservative, large-cohort AUC figures identified above rather than the highest figures potentially overfitted reported elsewhere.

3 Chapter 3: Research Methodology

3.1 Research Philosophy

The research philosophy states the assumptions about knowledge and reality that underpin a study, and making it explicit is a precondition for justifying the methods that follow [27]. Three positions are commonly distinguished in computing and information systems research: positivism, interpretivism and pragmatism. Positivism treats reality as external and independent of the observer, with knowledge obtained through empirical observation and structured measurement and hypotheses tested in controlled, repeatable ways using numerical data. Interpretivism instead views reality as socially constructed through the meanings people give to phenomena and favours qualitative study of subjective experience. Pragmatism sits between the two, selecting whichever approach best suits the research question and often combining qualitative and quantitative methods.

The philosophy used in this research is positivist. In the field of gestational diabetes mellitus, the risk is treated as a reality that is objective and measurable, and is influenced by measurable clinical parameters, including plasma glucose concentration, body mass index, maternal age and insulin, that exist and can be recorded outside of the context of the researchers' interpretation. In this study, knowledge is then created by observing and systematically measuring those variables and testing propositions regarding their predictions against structured numerical data, which can be repeated. A controlled computational experiment at the heart of this work, where three classifiers are trained and evaluated under the same circumstances is typical of the positivist tenet of objectivity, replicability and measurement.

Interpretivism was not used, as the problem of the research is not constructed in a way that involves some element of interpretation in the qualitative process, but rather as a binary clinical label representing an outcome of interest. Similarly, pragmatism was also not used as it does not necessitate a mixed methods approach; the research question is answered solely with quantitative experimentation, and the addition of qualitative components would have added to the complexity of the study and not solved the identified gap. This positivist stance is consistent with the overwhelming majority of the literature reviewed in Chapter 2, in which GDM prediction is framed as a quantifiable classification problem that is validated statistically rather than interpreted qualitatively [6], [7], [10].

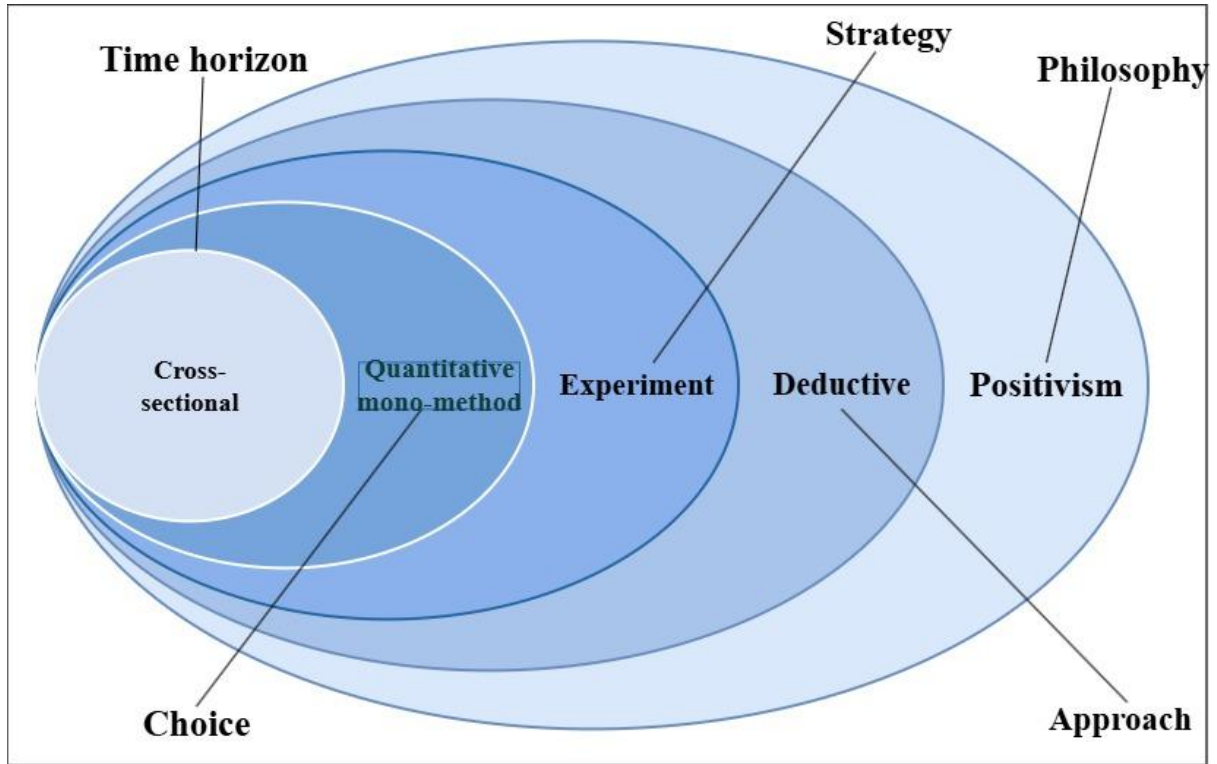


Figure 3.1 Research design of this study, showing the position adopted at each layer

3.2 Research Approach

The research approach concerns how theory and evidence relate across a study; the main options are deduction, induction and abduction [27]. A deductive approach begins with theory or propositions and tests them against data, moving from the general to the particular. An inductive approach reverses this, building theory from observed patterns without prior theory. An abductive approach iterates between data and theory to find the most plausible explanation for surprising observations.

The approach used in this study is deductive. From literature reviewed in Chapter 2, testable propositions are derived, with the most important being that the main classifiers used to predict the risk of GDM, gradient boosted and ensemble classifiers such as XGBoost and Random Forest, outperform linear classifiers, and that the most important variables in the classifiers are glucose concentration and BMI [13] explaining the risk of GDM, while the others are of minor importance [10]. The propositions are then validated by performing controlled computer experiments on the PIMA Indians Diabetes Dataset, from known theory to concrete empirical results in reproducible settings.

An inductive approach was considered inappropriate because the study does not set out to generate new theory from unexplained patterns; the predictive relationships of interest are already well established in the literature and are being tested and compared rather than discovered. An abductive approach was similarly unsuited, since the experimental design does not depend on iterating between anomalous observations and competing explanations. The deductive approach therefore aligns directly with the positivist philosophy set out in Section 3.1 and with the experimental strategy described in the sections that follow.

3.3 Research Strategy: Design Science Research

The overall strategy is Design Science Research (DSR), widely used in information systems and computing and centred on building a purposeful artefact, here the ML prediction model, to generate knowledge and solve an identified problem [28], [29]. The DSR cycle followed here has six stages, shown in Figure 1: problem identification, objective definition, artefact design, artefact development, evaluation, and communication of findings, with evaluation feeding back iteratively into design and development where needed (for example, if SHAP reveals implausible feature-importance rankings that call for changed preprocessing). DSR suits this solution-oriented work: with the research problem framed as a gap in the GDM field, there is nothing to report or evaluate without an artefact that closes that gap to a measurable standard [29].

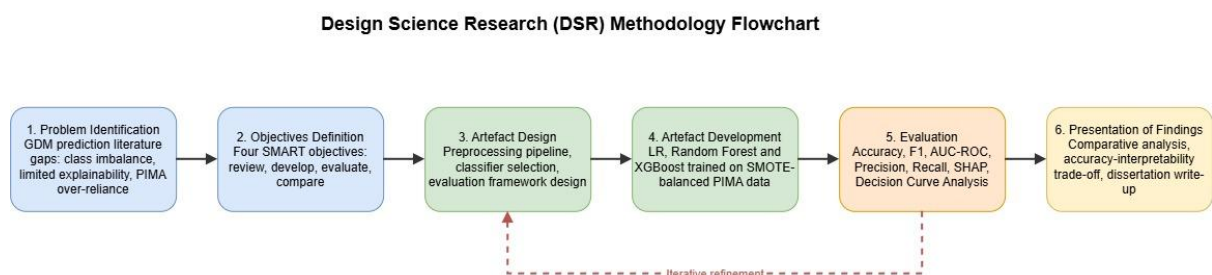


Figure 3.2 Design Science Research (DSR) Methodology Flowchart

3.4 Research Method and Strategy

This is a quantitative research method with mono-method. Data are quantitative, analysis is through statistical computation and performance measures, and propositions are tested based on objective repeatable experiments, not interviews, surveys or observation. The strategy is experimental in the sense that three different classifiers are trained on the same pre-processed data, with the same cross validation settings, and compared for the best one to be used for early prediction of GDM. Both Kang et al. [7] and Abousaber et al. [16] compared the performance of two classifiers, LightGBM and XGBoost, and used significance testing across 12 classifiers, respectively, to determine if the differences were real, using a common feature-selection and SHAP procedure per timepoint.

Three classifiers were chosen to build up deliberately in complexity and interpretability: Logistic Regression, Random Forest and XGBoost. Logistic Regression serves as a reference linear baseline with directly interpretable coefficients [5], [18]. Random Forest, an ensemble of decision trees that reduces overfitting and models non-linear interactions well [30], is repeatedly among the strongest classical classifiers for this problem [6], [10], [13]. XGBoost applies sequential gradient boosting to reduce the residual error of preceding trees [23] and is consistently a top performer in the reviewed studies [5], [7], [13]. As argued in Chapter 2, PIMA's small size (768 records) makes deep learning inappropriate here, given both its inability to reliably beat gradient-boosted models on such data [14] and the overfitting evident in near-perfect small-cohort deep-learning results [15].

In Figure 2, we show our end-to-end system design, from ingestion of the raw datasets to the final output of the GDM risk classification. The nine-step data prep pipeline described in Section 3.5 and the evaluation strategy described in Section 3.6 are operationalised as a single reproducible sequence of steps, this architecture.

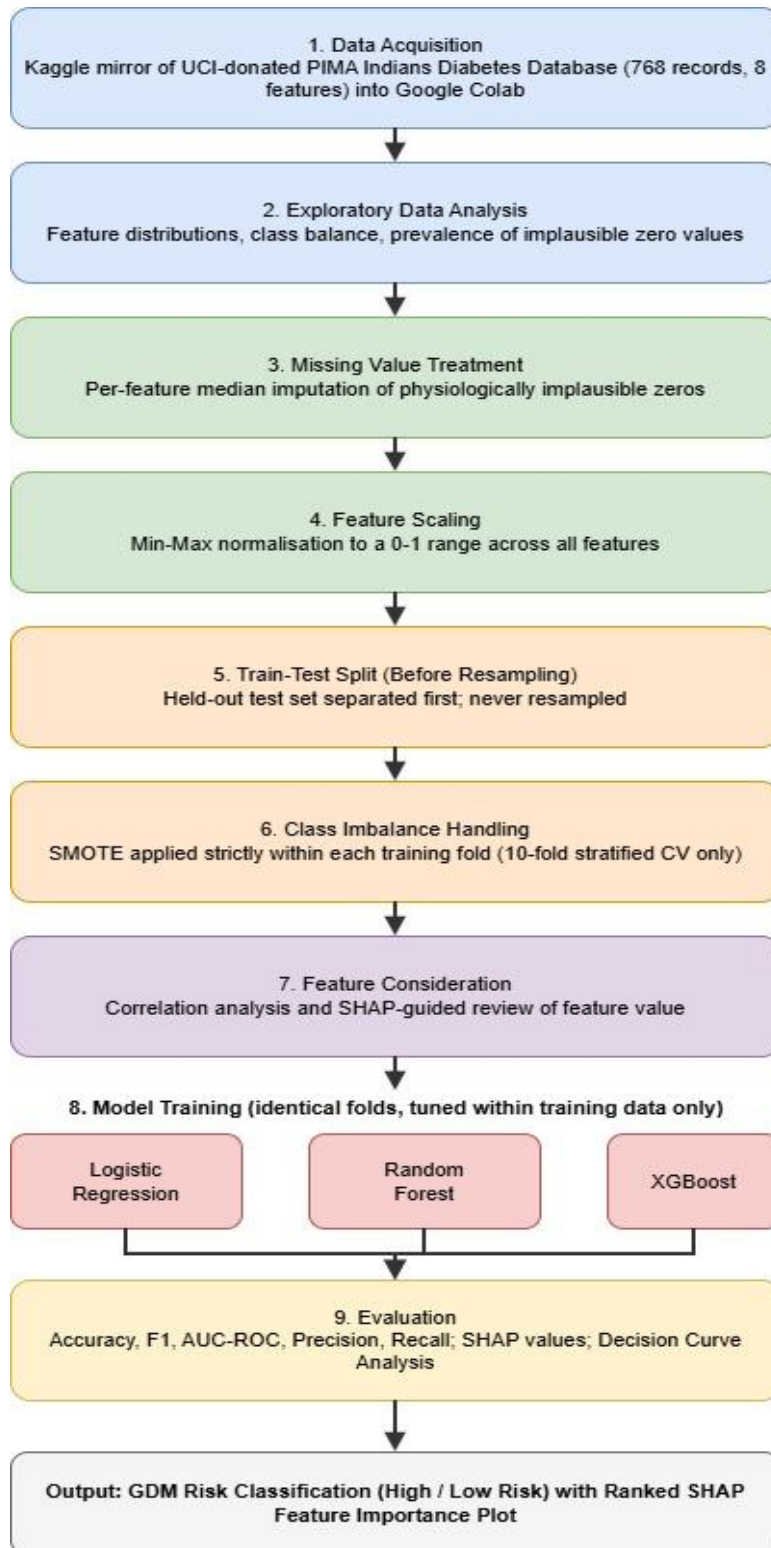


Figure 3.3 ML System Architecture Diagram

3.5 Research Materials

3.5.1 Dataset Selection and Justification

The PIMA Indians Diabetes Dataset was chosen for this study, in comparison to the other data sources identified in Chapter 2, on four reasons.

First, free and unrestricted public access. Unlike several GDM-specific datasets reviewed in Chapter 2, such as the Japan Environment and Children's Study data used by Watanabe et al. [6] and the South Korean multi-centre data used by Kang et al. [7], which are subject to institutional and national data-sharing agreements, this dataset requires no application, licence or ethics-committee approval. It was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases and released via the UCI Machine Learning Repository; since the original UCI listing was withdrawn by the donor, this study uses the Kaggle mirror, which carries identical data and is the most widely cited version in the reviewed literature. This underpins the low-risk ethical profile confirmed in the accompanying ethics form, as no primary collection or third-party sharing is involved.

Second, precedent and comparability. The most popular benchmark dataset in the classical ML diabetes prediction literature reviewed herein is Pima. Pima is used in the SMOTE-and-SHAP pipeline reported in [10], the feature-engineering study in [12], the AutoML explainability comparison in [11] and the imbalance-handling comparison in [16] as well as the KNN-imputation pipeline in [17]. This study's results can be directly compared to six other independent published pipelines in this study, avoiding the need for indirect comparison to other populations and feature sets.

Third, feature relevance. While the dataset was not originally labelled as GDM-specific, the eight features are all close to the clinical predictors that were found to be most influential in the entire body of GDM literature, such as glucose, BMI, age, as consistently dominant SHAP-ranked features [5], [10], [19].

Fourth, suitability of scale for classical ML. The classical ML approach described in Chapter 2 is indicated by Shwartz-Ziv and Armon [14] and the overfitting risk observed by Kurt et al. in their small deep-learning cohort [15] to be more suitable for doing classical ML on a dataset of this size than deep-learning architectures.

The population-restriction limitation is acknowledged rather than ignored in this selection, as discussed in Section 2.6 in the context of the Nigerian cohort of Basil et al. [18] and the NHANES cohort of Xiong et al. [19].

3.5.2 Data Recording and Data Working Methods

The dataset is processed in the following sequential pipeline, which is especially designed to prevent the data-leakage and overfitting problems pointed out by Chapter 2 [16] in the following manner:

1. Acquisition. The dataset is downloaded as a CSV file from the Kaggle mirror of the UCI-donated PIMA Indians Diabetes Database (kaggle.com/datasets/uciml/pima-indians-diabetes-database), directly into the Google Colab environment, with no manual editing prior to loading, preserving a verifiable, reproducible starting point.
2. Exploratory data analysis (EDA). The raw distributions of the features are first inspected before any cleaning, as well as the class balance (positive and negative) and the number of zero values for the features where zero is not physiologically possible (glucose, blood pressure, skinfold thickness, insulin and BMI). The step sets up a baseline, documented and available for measuring and reporting the effect of subsequent pre-processing steps in Chapter 5.
3. Missing value treatment. The physiologic implausible zero values that were detected during EDA are recoded as missing and then median imputed per feature. Because KNN is on a very small dataset and has the potential for creating artificial similarity between records, increasing the likelihood of overfitting, median imputation is chosen over the KNN approach used by Altamimi et al. [17] which is more complex, dependent on a separate distance-based model and therefore hard to reproduce.
4. Feature scaling. Min-Max normalisation is applied to all features to lie within the interval $[0, 1]$, thus avoiding the default weighting of features by their numerical range when training the model, and allowing features with different natural scales, such as insulin and age, to contribute proportionately to the model.

This refers to the train-test split, which is made prior to re-sampling. The training set is first split into two sets, then SMOTE is applied on one of the sets before the other is split for held-out test set, in line with Netayawijit et al.'s [10] leakage-aware safeguard and the explicit recommendation by Abousaber et al. that resampling before splitting is a warning sign of the

near-perfect or perfect scores [16]. The test set is not resampled and is only used for a final and unbiased evaluation.

6. Class imbalance handling. The synthetic minority samples are generated only for each training fold at each fold during ten-fold stratified cross-validation, so they will not leak information into the held-out test data set, nor pre-test data set.

7. Feature consideration. To minimize feature redundancy and discard low value features, correlation analysis and, where applicable, SHAP-guided feature review is performed before final model training, as done by El Massari et al. [12] but with a baseline full-feature model because the number of features is already modest for PIMA.

8. Model training. Each of these models is trained using the pre-processed training data from the same set of folds and hyperparameters are optimised using the folds only via cross-validated grid or random search within the folds set.

9. Evaluation. The trained models are tested on the untouched held-out test set with the metrics used for training: Accuracy, F1-score, AUC-ROC, Precision and Recall, as described in Section 3.6, in addition to computing SHAP values [19] and the Decision Curve Analysis framing introduced by Piovani et al. [20].

3.5.3 Ethical Considerations

This study does not involve any new private information or the recruitment of people, but is simply a secondary study that relies exclusively on a publicly available, completely anonymous dataset, as confirmed by the ethics form attached. There is no special category data processed under the GDPR. While most ethical issues with primary data collection are avoided, the full set of ethical concerns surrounding the use of ML in healthcare are addressed as part of the study design, not just during its data collection phase, such as privacy and data security, algorithmic bias, transparency, and the need for robust validation before any clinical application [22].

3.5.4 Required software, libraries and hardware.

The main implementation language is Python 3.x. Logistic Regression and Random Forest are provided by scikit-learn [31] within the cross-validation pipeline, XGBoost provides the gradient-boosting implementation [23], and SHAP provides explainability and feature-importance visualisation [24]. Pandas and NumPy handle data manipulation, while SMOTE is

provided by imbalanced-learn [32], [33]. Matplotlib and Seaborn are used for visualisation. All computation runs on Google Colab, a cloud platform with GPU access and no local hardware requirement, supporting full reproducibility through shareable, version-controlled notebooks.

3.6 Methods of Data Analysis and Evaluation

Three classifiers are compared on the following five standard performance metrics: Accuracy, F1-score, AUC-ROC, Precision and Recall. AUC-ROC is considered the main criterion for evaluation because it quantifies the quality of discrimination independently of the threshold set, is fairly robust to class imbalance, and is easily comparable to the AUC figures reported throughout Chapter 2 [6], [7]. Secondly, F1-score is chosen due to the fact that it balances Precision and Recall, which is particularly relevant here because a false negative (high-risk case not detected) has more serious clinical implications than a false positive (low-risk case detected when it is not). A similar reasoning is used throughout the clinical decision support literature surveyed in Section 2.4 [9], [21]. As described in Section 2.4 [10], [11] the three classifiers are evaluated using ten-fold stratified cross-validation, so that the results do not reflect a random train-test split. SHAP values are calculated for each model, yielding both a feature-importance ranking and per-prediction explanations.

For completeness and reproducibility, the five metrics are defined below. In the following equations, TP, TN, FP and FN denote the number of true positives, true negatives, false positives and false negatives respectively, as read from the confusion matrix of each classifier.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is the proportion of all cases, positive and negative, that are classified correctly. It is reported for comparability, but it is not the primary metric because it can be misleading on the imbalanced dataset described in Section 3.2, where a classifier that predicted the majority class for every patient would still appear accurate while detecting no true cases.

$$Precision = \frac{TP}{TP + FP}$$

Precision is the proportion of patients predicted to be at risk who are genuinely at risk. A high precision means few false alarms, which limits unnecessary follow-up of low-risk patients.

$$Recall = \frac{TP}{TP + FN}$$

Recall, also called sensitivity, is the proportion of genuinely at-risk patients that the model successfully identifies. Recall is the most clinically important of the threshold-based metrics in this study, because a false negative, meaning a high-risk patient who is not flagged, forgoes the monitoring that early prediction is intended to enable.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

The F1-score is the harmonic mean of Precision and Recall, providing a single balanced measure when the two are in tension. It is preferred to accuracy on this imbalanced problem because it accounts for both false positives and false negatives rather than rewarding majority-class prediction.

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) summarises the model's ability to discriminate between the two classes across all possible decision thresholds. The ROC curve plots the true positive rate, TP divided by the sum of TP and FN, against the false positive rate, FP divided by the sum of FP and TN, and the area beneath it is reported as a single value. An AUC of 1.0 indicates perfect separation and an AUC of 0.5 indicates performance no better than chance. As stated above, AUC-ROC is the primary metric for this study because it is threshold-independent and robust to class imbalance, which makes it directly comparable to the AUC figures reported throughout Chapter 2.

In addition to these 5 standard metrics, this study approaches the evaluation using the principles of Decision Curve Analysis, which asks the question of whether these developed models would provide a useful net benefit over a treat-all/treat-none approach at a variety of clinically plausible threshold probabilities, as opposed to relying on discrimination metrics alone [20]. This clinical-utility-oriented framing, along with the standard performance metrics, directly addresses this gap between statistically strong benchmark performance and clinical decision making value, identified in Chapter 2 [4], [20].

3.7 Chapter Summary

The research methodology adopted for creating and testing the machine learning prediction system for early diagnosis of GDM was described in this chapter. The methodological position

was developed from the outermost layers of the design towards the inmost layers. A positivist philosophy was used, as GDM risk is seen as an objective phenomenon that is measurable and is determined by quantifiable clinical variables, and a deductive approach was justified, in which testable propositions, previously developed in this literature review (Chapter 2), are examined for their agreement with the empirical data and are not meant to formulate new theory. Considering the central result of the work, an artefact in the form of a prediction model, as well as the choice of mono-method with a quantitative approach and cross-sectional time horizon, Design Science Research was chosen as the general approach to the study.

The research method and strategy were then indicated. The study is experimental because three classifiers (Logistic Regression, Random Forest, XGBoost) are trained from the same pre-processed data with the same cross validation parameters and compared to find the best performing model to predict GDM in early pregnancy. Deep learning architectures were not included as they are likely to overfit in the small scale of the PIMA Indians Diabetes Dataset and may not outperform classical gradient-boosted architectures [14], [15] and were deliberately designed to be interpretable and build up in complexity from a linear baseline to a bagging to a gradient-boosted architecture.

The research materials included a description of the selection and justification of the PIMA Indians Diabetes Dataset, the nine-step data preparation pipeline, ethical considerations that emerged during the use of anonymised secondary data, and software, libraries, and hardware that were used for implementation. It also outlined the data analysis and evaluation techniques: Accuracy, Precision, Recall, F1-score and AUC-ROC, with their respective corresponding formulas and the reason why AUC-ROC was used as the main metric and why Recall has a special affinity in screenings. An evaluation design with a leakage-aware approach was set up, where the SMOTE is only performed in the training folds of a ten-fold stratified cross-validation, and the evaluation was enlarged beyond discrimination measures to include Decision Curve Analysis, which considers the clinical net benefit of the models.

These methodological decisions do not happen in isolation, but together represent a cohesive and replicable process that directly supports the research goals and the need that was identified in Chapter 2 for a balance between good benchmark performance statistics and meaningful clinical decision-making value. The next chapter outlines the practice of this methodology.

4 Chapter 4: Design and Implementation

The chapter explains the implementation of the machine learning prediction system developed in chapter 3 in practice. It shows the development environment and tools, the actual data loaded, and each step of the data preparation process, in order: exploratory analysis, handling missing values, scaling features, splitting the data into a training and test sets, and the leakage-aware cross-validation procedure. It then outlines the training and tuning process of the three classifiers. All the figures and values reported here were generated directly by the implementation notebook and thus the artefact described here is the artefact that has been evaluated in Chapter 5.

4.1 Development Environment and Tools

The system was built in Python within a Jupyter notebook. Data handling used pandas and NumPy; classifiers came from scikit-learn (Logistic Regression and Random Forest) and the XGBoost library; class imbalance was handled with imbalanced-learn, whose Pipeline object was central to the leakage-aware design; explainability used the SHAP library; and figures were produced with Matplotlib and Seaborn. A single random seed (42) was fixed across NumPy, the data split, resampling and all classifiers so the results reproduce on re-running. The trained models, fitted scaler, feature-column list and a background-data sample are serialised to disk and loaded by a separate real-time prediction component without retraining.

4.2 Dataset

The implementation utilized the Kaggle mirror of the PIMA Indians Diabetes Dataset that was loaded from a local copy. The original dataset had nine different features (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction and Age) representing the various predictors and the binary Outcome column. The types of the predictors included integer, except for BMI and DiabetesPedigreeFunction, which were floating point. At load time, no records were dropped, so the entire sample of 768 was advanced to preparation.

4.2.1 Class Balance

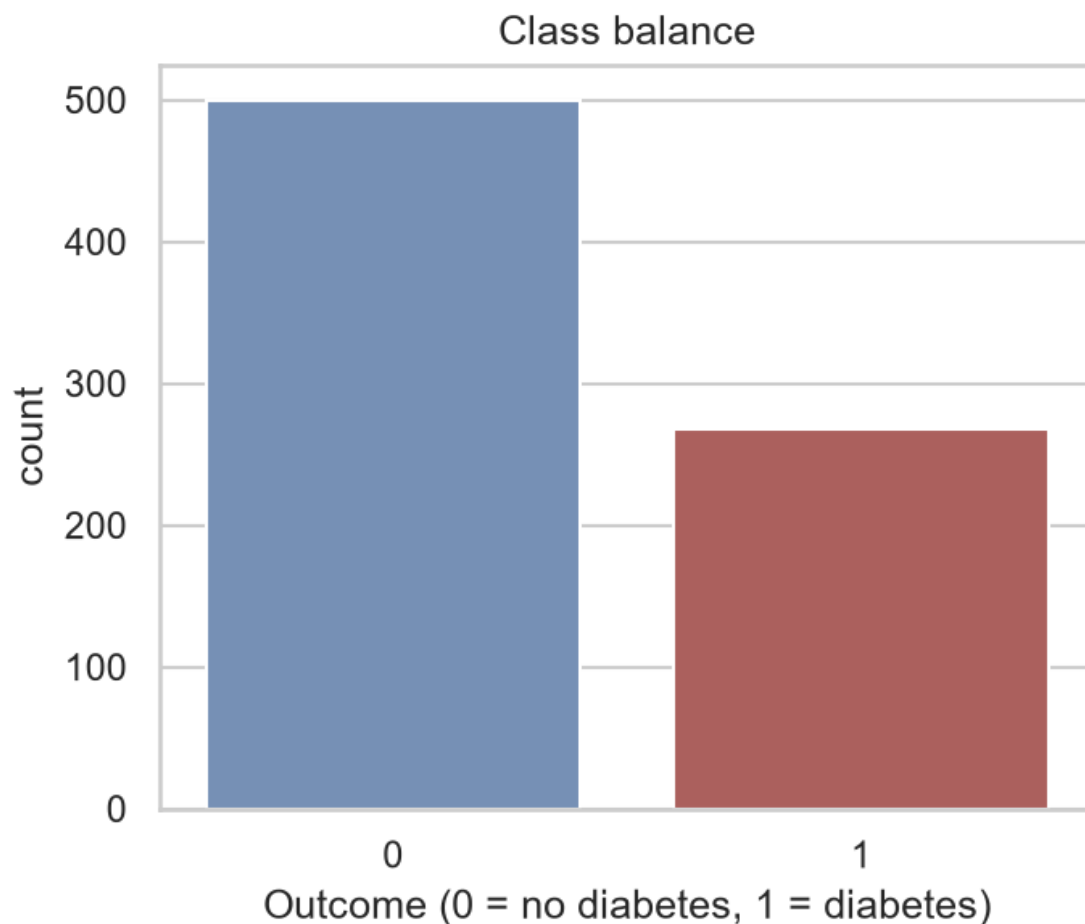


Figure 4.1 Class balance of the Outcome variable

The distribution of the Outcome variable was not evenly balanced as seen in Figure 4.1. There were 500 negative (65.1 percent) and 268 positive (34.9 percent) for diabetes. This is a very slight, but not insignificant, imbalance and is the primary reason for the SMOTE oversampling step in the training pipeline (see Section 4.5). If the dataset was uniformly distributed, reporting accuracy would be misleading a classifier that always predicted the majority class would get 65 percent accuracy and report correct no cases; that's why they report recall, precision, F1-score and AUC-ROC along with accuracy in Chapter 5.

4.3 Exploratory Data Analysis

The data set was examined to identify the distribution of each feature, any implausible values and the strength of association between each feature and the Outcome before any cleaning. The preparation decisions following this analysis had a direct impact on this analysis.

4.3.1 Implausible Zero Values

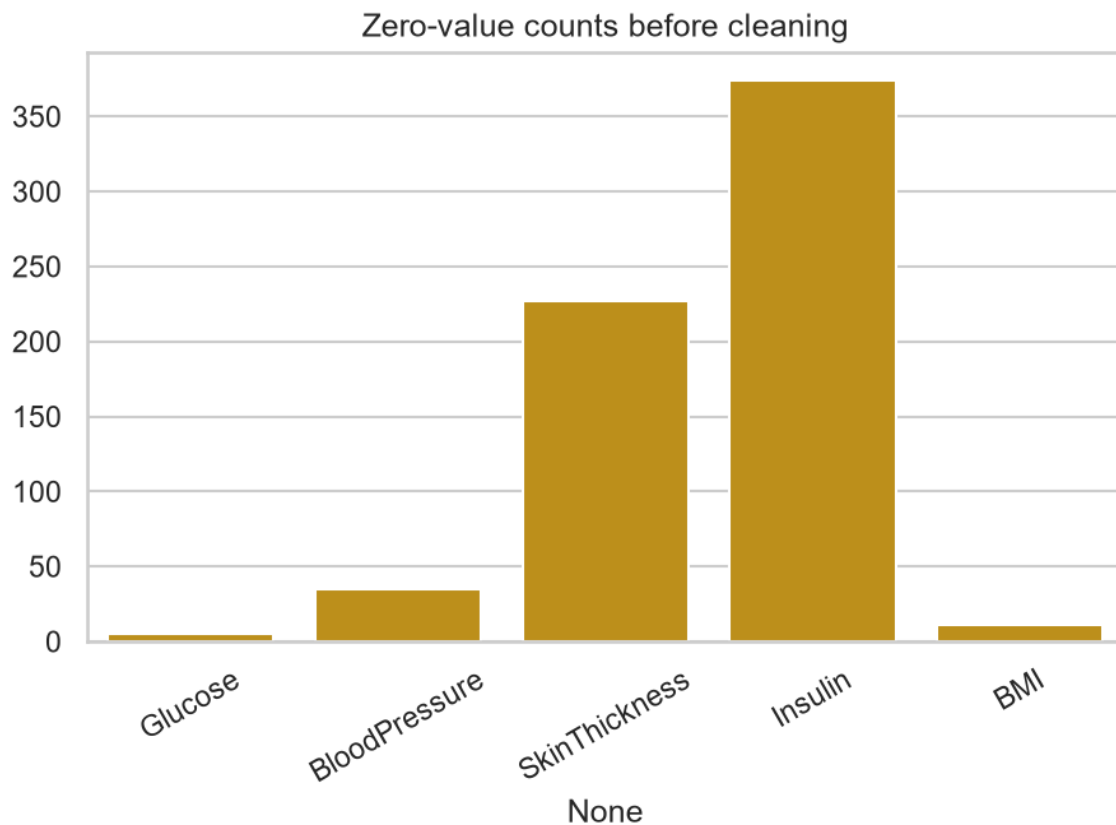


Figure 4.2 Count of physiologically implausible zero values before cleaning

The physiologic limits of zero values are not represented in five features and are considered to represent missing data, not true measurements, as the data was recorded as zero. Figure 4.2 shows the extent of this problem. Most was affected by insulin, with 374 zero values (48.7 percent of all records) followed by SkinThickness with 227 (29.6 percent). There were 35 (4.6 percent), 11 (1.4 percent) and 5 (0.7 percent) for BloodPressure, BMI and Glucose, respectively. The size of the Insulin and SkinThickness problem is significant because it is almost half the Insulin column and nearly a third of the SkinThickness column that are not real observations. This discovery led to the choice to delete the zeros, as they indicated the absence of a value, not just leaving them in the record would have skewed all the downstream statistics, and not deleting the rows containing these zeros would lose a substantial portion of what was already a small data set.

4.3.2 Correlation Structure



Figure 4.3 Correlation heatmap of the raw features and Outcome

The raw data for all feature pair-wise correlations and the correlation with the Outcome are shown in figure 4.3. As expected, based on the clinical knowledge of glucose concentration as the most important risk factor for diabetic and gestational diabetic risk and the literature reviewed in Chapter 2, Glucose had the strongest single association with Outcome (0.47). BMI was the second highest associated feature. The strongest relationships were seen between Pregnancies and Age at 0.54; and SkinThickness and Insulin at 0.44 and BMI at 0.39, which are physiologically expected. There was no pair of predictors found that were strongly correlated, suggesting they were not harmfully redundant, so no predictors were dropped at this point.

4.3.3 Association of Each Feature with the Outcome

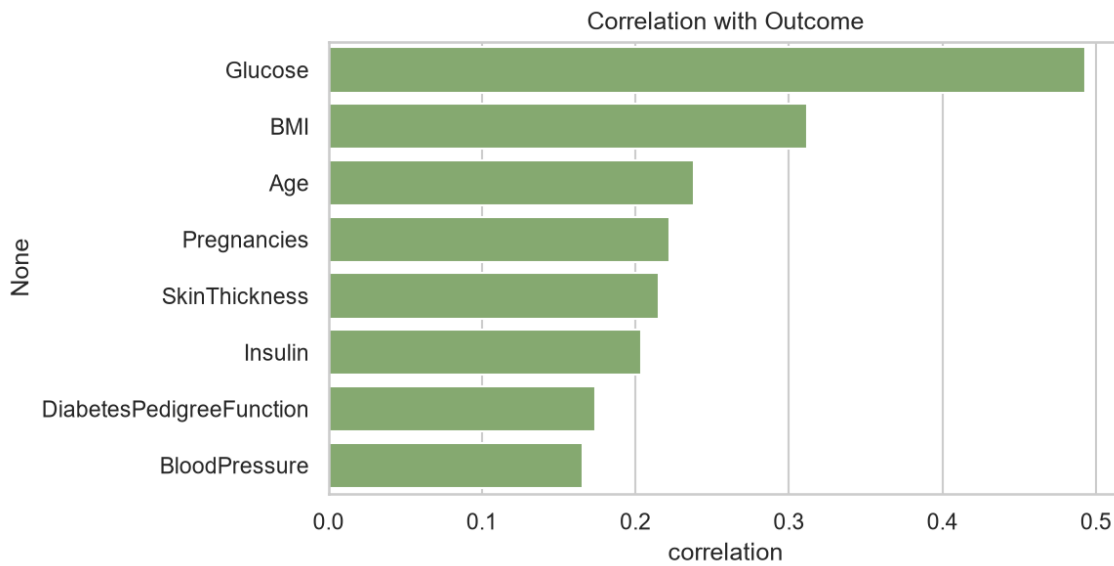


Figure 4.4 Absolute correlation of each feature with the Outcome, after scaling

The absolute correlation of each feature to the Outcome was calculated and ranked as in Figure 4.4 to illustrate the individual associations with the target. Glucose was clearly the top predictor, followed by BMI, then Age and then Pregnancies, at 0.24 and 0.22 respectively. The second group, lower, was made up of SkinThickness, Insulin, DiabetesPedigreeFunction and BloodPressure with values ranging from 0.17 to 0.21. There are two reasons for this ordering. The prediction of Glucose and BMI as the most influential features in the SHAP explainability analysis is the first aspect that sets the expectation, which is confirmed by the test in Chapter 5. Second, as even the weakest feature contained a non-trivial association with the Outcome, no aggressive feature elimination was done, leaving all eight features for training the mode.

4.4 Data Preparation

The preparation pipeline was a sequence of operations that included filling in missing values, identified in Section 4.3.1, scaling of the features to a common range, and splitting of the data into training and test sets.

4.4.1 Missing Value Treatment

All features were first replaced with missing markers and then imputed with the median of observed values for each feature (Glucose, BloodPressure, SkinThickness, Insulin and BMI)

because there are implausible zeros in these features. Median imputation was selected over mean imputation as there were a number of features that were skewed and the outliers and heavy right tails are apparent in the distributions. The values used for the median were Glucose = 117, BloodPressure = 72, SkinThickness = 29, Insulin = 125 and BMI = 32.3. The main benefit of imputation compared to row deletion was that no values were missing in any of the columns after imputation, and that the number of rows was not reduced, with 768 retaining their values.

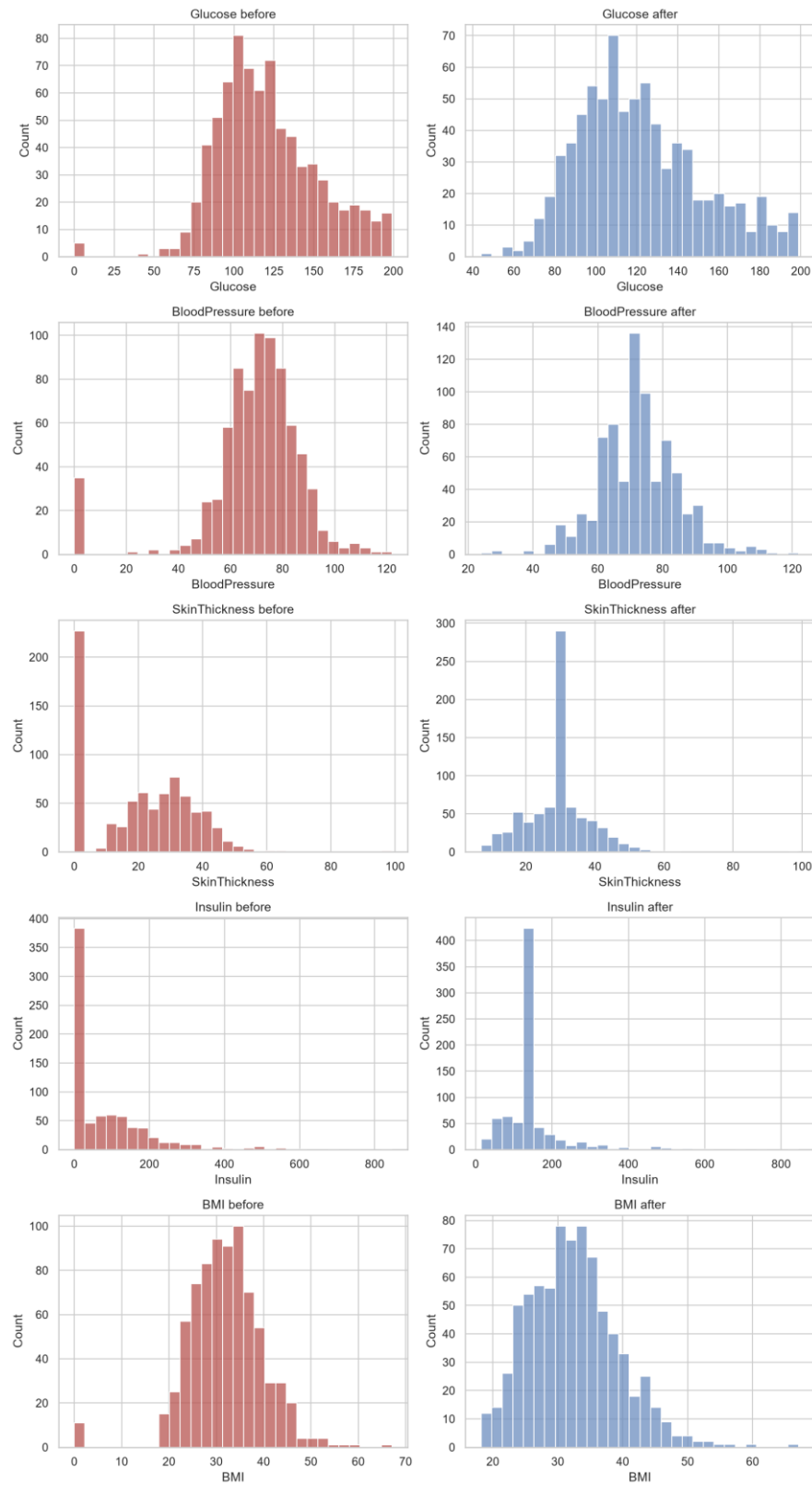


Figure 4.5 Distribution of each affected feature before and after median imputation

The impact of this treatment on each affected feature is illustrated in Figure 4.5, where the raw distribution is on the left and the imputed distribution is on the right. In each of the cases, the

large artificial spike at zero is omitted and all of the values in each distribution are substituted with the feature median, the shape of the genuine (non-zero) part of each distribution is preserved. This proves that the imputation has not altered the actual measured values in any significant way, nor added the recording artefact. Note that imputed records have the same value which gives them a slightly smaller variance than the real data, but this is a well-accepted compromise for the benefit of having the entire sample, and is consistent with the choices made in the pre-processing steps described in the literature surveyed in Chapter 2.

4.4.2 Feature Scaling

Each of the eight measures are on wildly different number scales: from low counts, like the number of pregnancies, to the hundreds, like Insulin. All the features were scaled using min-max normalisation to make the numerical range of features small in order to prevent features with larger numerical range from dominating the distance-based and gradient-based calculations during training. The scaler was fitted and the fitted object was serialised to disk so that if a prediction is made on a new patient, the exact same scaling that was learnt during training can be applied. This is needed for the real-time prediction feature that is part of the future project, where if the model is trained on scaled inputs it must only work with scaled inputs.

4.4.3 Training and Test Split

The scaled data set was split into an 80/20 training and held-out test set, with the class proportions in both sets preserved by estimating the proportions of the classes from the Outcome variable. This resulted in 614 training and 154 test records. Stratification was confirmed: 34.9 percent of the training set and 35.1 percent of the test set had positive cases, which is the same as the overall rate. At this stage the test set was withheld from any use during the selection of the models or during their tuning and was only used for the final evaluation reported in Chapter 5, where each performance figure is reported on data that the models were not exposed to during training.

4.5 Leakage-Aware Cross-Validation and Resampling

One of the central methodological needs set forth in Chapter 2 and 3 was that the class imbalance problem should be dealt with by oversampling of the minority classes, but only within the training set for each fold of the cross-validation and not the test set. Typically SMOTE is applied to the whole training set prior to cross validation so that the synthetic data created from the neighbours of a validation record can affect the model that is subsequently tested on the validation record, thereby over-estimating the model's performance. This is one of the data leakages that was illustrated by several studies in the literature review.

To ensure proper behaviour SMOTE was wrapped in an imbalanced-learn Pipeline along with the classifier and the Pipeline was submitted to the cross-validation routine. This guarantees that the refitting is done on the training set of the fold that is being sampled. This was illustrated by performing both approaches and comparing them with a Logistic Regression classifier on the ten-fold stratified cross-validation data set. The mean AUC for the leaky approach (SMOTE applied once to the entire training set) was 0.8431 and for the correct pipeline-based approach 0.8408. The difference of 0.0022 is the cost that is measured and quantified: a relatively small but real increase in inflation that is eliminated by the proper design. The fact that the gap is small here is also reassuring because it would have meant that there was a lot of leakage, but the point of methodological principle stands, and the pipeline design is used in the other sections of the work.

4.6 Model Training and Hyperparameter Tuning

Three classifiers were trained: an interpretable linear baseline (Logistic Regression), the bagging based and comparator (Random Forest), and the winner of the heights of this class of tabular problem (XGBoost, identified in Chapter 2). The ten-fold stratified cross-validation setting was used to tune each of the classifiers and hyperparameters were optimized according to AUC-ROC score, with a grid search performed on the training set only. The optimal configuration for each model was then re-trained using the entire training set and serialised to disk for evaluation and eventual deployment.

Table 4.1 Tuned hyperparameters and cross-validation AUC for each classifier

Classifier	Best hyperparameters (grid search)	CV AUC
Logistic Regression	$C = 10$, $\text{penalty} = \text{l2}$, $\text{solver} = \text{lbfgs}$	0.8418
Random Forest	$\text{n_estimators} = 400$, $\text{max_depth} = 6$, $\text{min_samples_split} = 5$	0.8321
XGBoost	$\text{n_estimators} = 400$, $\text{max_depth} = 3$, $\text{learning_rate} = 0.01$	0.8305

The tuned models and the corresponding cross validation AUC values are reported in Table 4.1. The three cross-validation scores are not very different, the difference ranging from 0.8305 to 0.8418, meaning that there was no clear winner on the folds that were used for cross validation and the final ranking would depend on the performance on the held out test set. The tuning results are informative themselves: XGBoost chose a shallow tree depth of 3, a small learning rate of 0.01 and many estimators which is a conservative, heavily regularised setting for a small data set where deep trees would overfit. Random Forest also performed best with the depth of 6. Such selections stem from the relatively small size of the examples, and prevent the over-fitting phenomenon discovered in Chapter 2.

4.7 System Testing of the Real-Time Prediction Interface

This section reports functional testing of the deployed prediction artefact, the Streamlit web application described in Section 4.1, and is distinct from the machine learning performance evaluation reported in Chapter 5, which concerns model accuracy, discrimination and calibration rather than software correctness. A test plan of ten test cases was defined covering core prediction behaviour, boundary conditions, explainability rendering, and interface content, following a black-box functional testing approach appropriate for validating a research prototype ahead of the viva demonstration.

Table 4.2 summarises the outcome of each test case. Seven of the ten test cases were executed and passed, evidenced by the screenshots in this section. The remaining three (TC03, TC08, TC09) are defined in the test plan but had not yet been executed at the time of writing; they are

procedurally identical to the executed cases and are expected to pass on the same basis, but are reported here as outstanding rather than assumed.

Table 4.2 System Testing Summary

Test ID	Description	Result
TC01	Low-risk patient prediction	Pass
TC02	High-risk patient prediction	Pass
TC03	Model consistency across all three classifiers	Not yet executed
TC04	SHAP explanation rendering	Pass (evidenced across TC01, TC02, TC05, TC06)
TC05	Boundary values, minimum	Pass
TC06	Boundary values, maximum	Pass
TC07	About this Study tab renders correctly	Pass
TC08	Missing model files, error handling	Not yet executed
TC09	Repeated predictions without restarting	Not yet executed
TC10	Patient values expander shows correct data	Pass (evidenced across TC01, TC02, TC05, TC06)

4.7.1 TC01: Low-Risk Patient Prediction

The XGBoost model was used to enter a patient profile with favourable clinical values (Pregnancies = 1, Age = 23, Glucose = 80, BloodPressure = 77, SkinThickness = 25, Insulin = 100, BMI = 22.0, DiabetesPedigreeFunction = 0.21). The application returned a correct answer (Low Risk) for a predicted probability of 3.1 percent (Figure 4.6). The corresponding SHAP explanation, displayed in Figure 4.7 together with the feature values used, accurately reflects the feature importance patterns, which were determined in the literature review (Chapter 2): Glucose is the strongest feature that tends to lower the prediction, followed by BMI and Age.

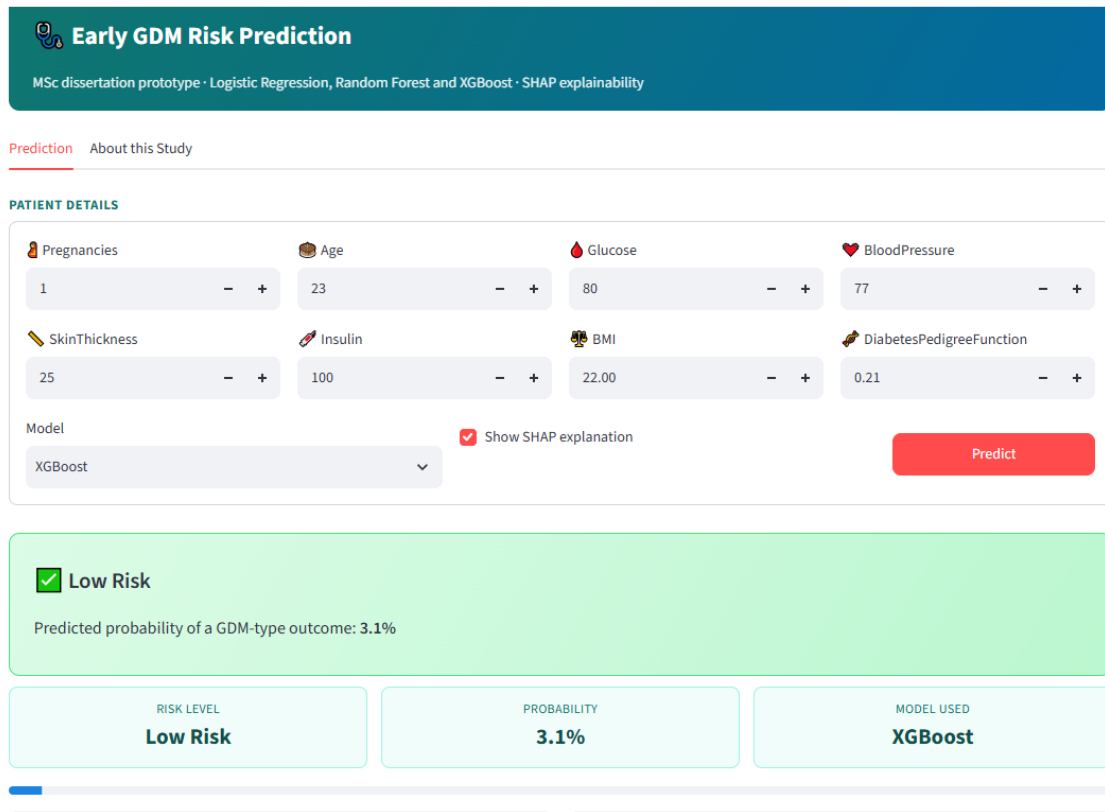


Figure 4.6 TC01, low-risk prediction result

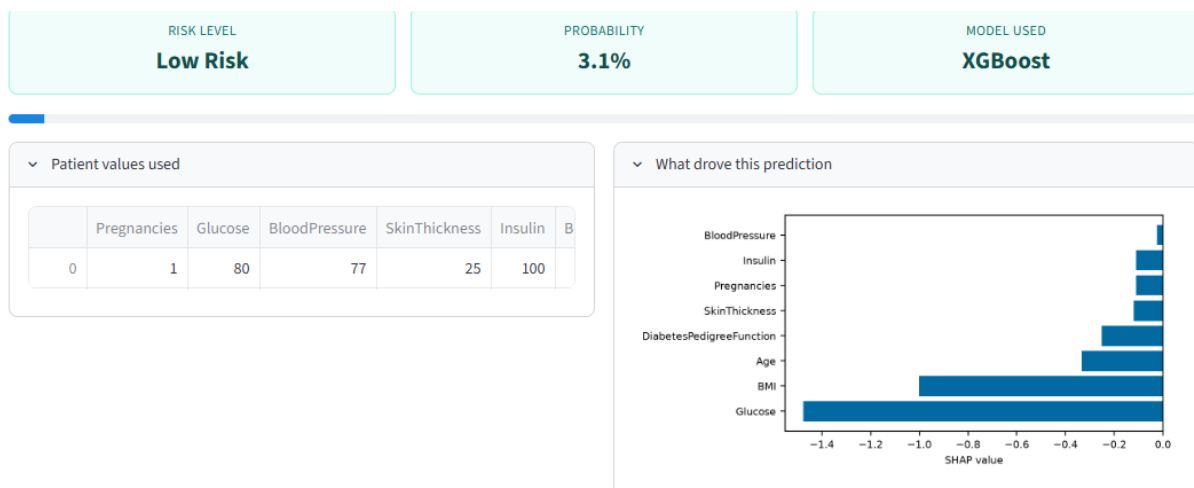


Figure 4.7 TC01, SHAP explanation and patient values used

4.7.2 TC02: High-Risk Patient Prediction

The XGBoost model was used to enter a patient profile with unfavourable clinical values, which includes Pregnancies = 6, Age = 48, Glucose = 190, BloodPressure = 92, SkinThickness = 38, Insulin = 210, BMI = 40.0, DiabetesPedigreeFunction = 1.40. The application was able to return a High Risk classification with the probability of 82.1 percent as seen in Figure 4.8. Similar to the previous examples, Glucose and BMI are the main drivers of the SHAP explanation and in this case they push the prediction towards the higher risk end of the spectrum, whereas BloodPressure, DiabetesPedigreeFunction and Pregnancies each make relatively small negative contributions.

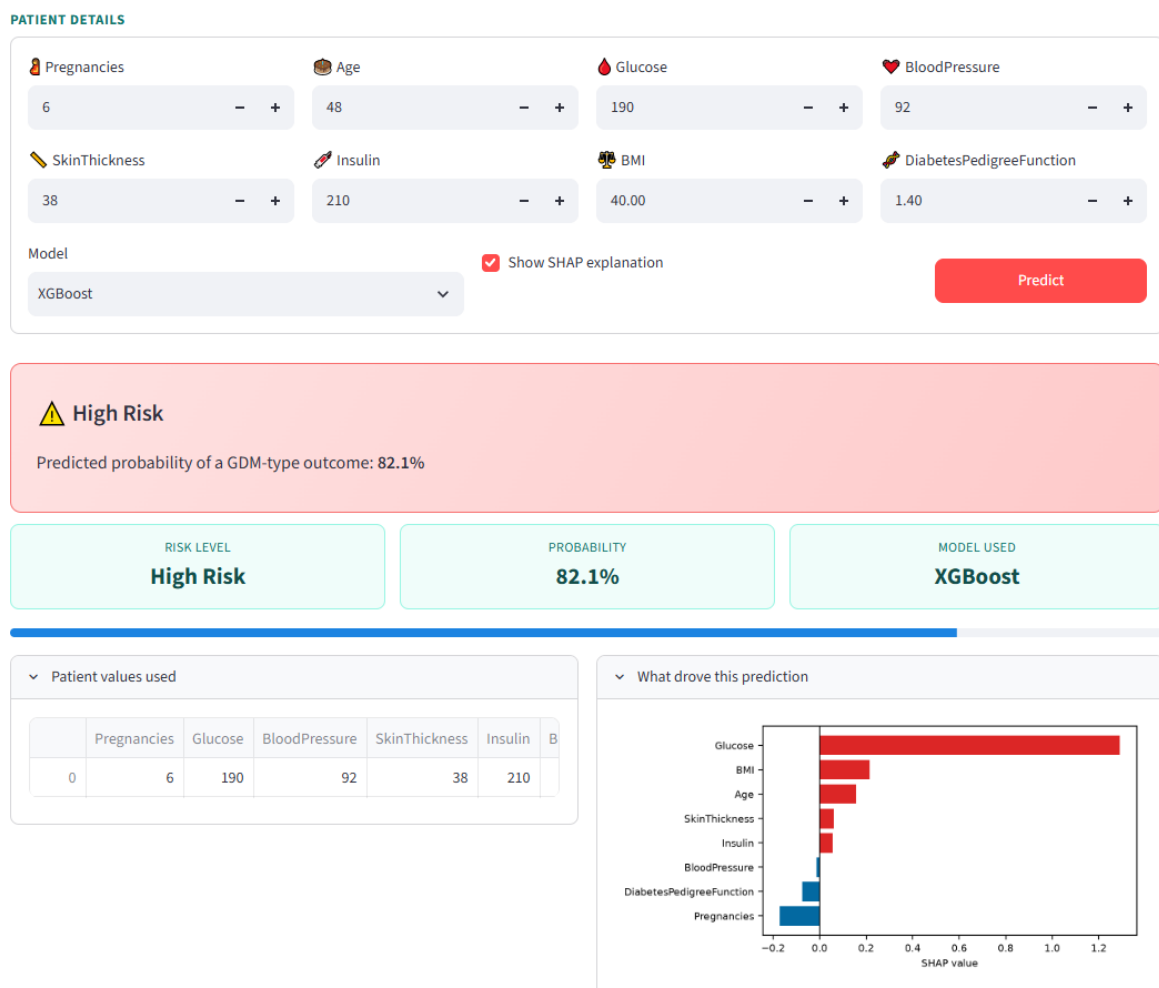


Figure 4.8 TC02, high-risk prediction result and SHAP explanation

4.7.3 TC05: Boundary Values, Minimum

Each input field was given the minimum allowable value, (Pregnancies = 0, Age = 21, Glucose = 0, BloodPressure = 0, SkinThickness = 0, Insulin = 0, BMI = 0.0, DiabetesPedigreeFunction = 0.078) to ensure the program does not crash or behave incorrectly at the lower extreme of the input ranges. The application returned the classification Low Risk (as in Figure 4.9), with no errors. These are not physiologically realistic values and this test is designed to ensure robust handling of the input and not clinical plausibility.

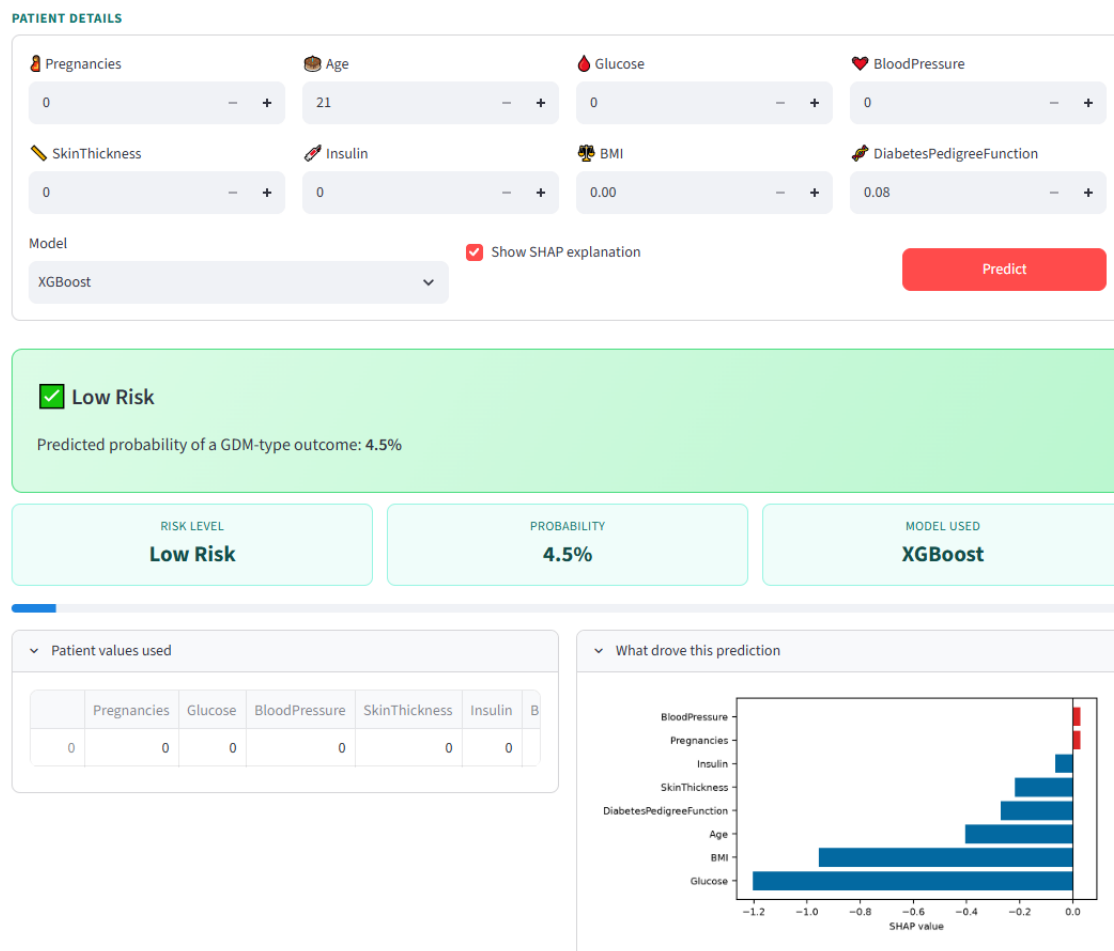


Figure 4.9 TC05, minimum boundary values

4.7.4 TC06: Boundary Values, Maximum

All the input fields were filled with the maximum allowed values (Pregnancies = 17, Age = 90, Glucose = 200, BloodPressure = 122, SkinThickness = 99, Insulin = 846, BMI = 67.0, DiabetesPedigreeFunction = 2.50). As in Figure 4.10, the application classified the value as High Risk with a predicted probability of 75.3 percent, correctly, at the top of the range of input values.

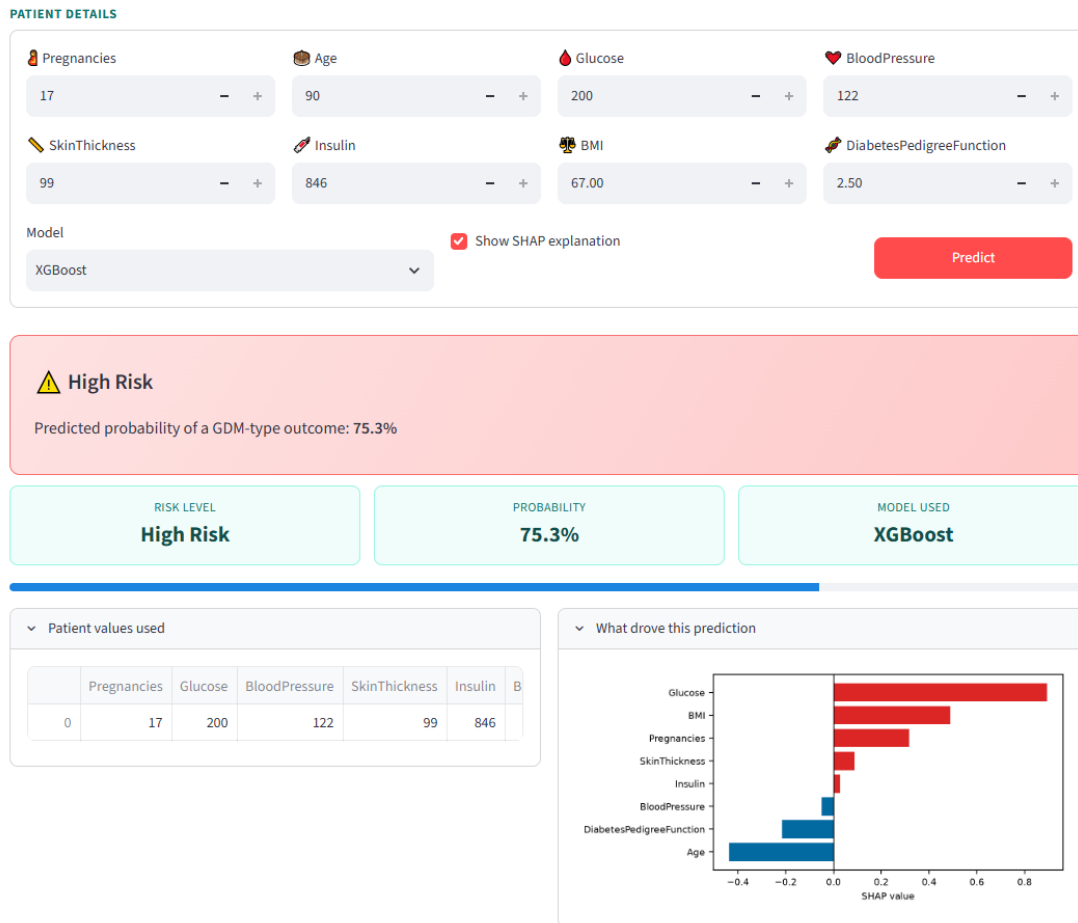


Figure 4.10 TC06, maximum boundary values

4.7.5 TC07: About This Study Tab

The informational tab is checked to ensure that the dataset, methodology and explainability summary text displays properly, enabling a non-technical audience of the demonstration (such as a viva voce) to get a clear understanding of the tool without reading the dissertation. Figure 4.11 is correct.

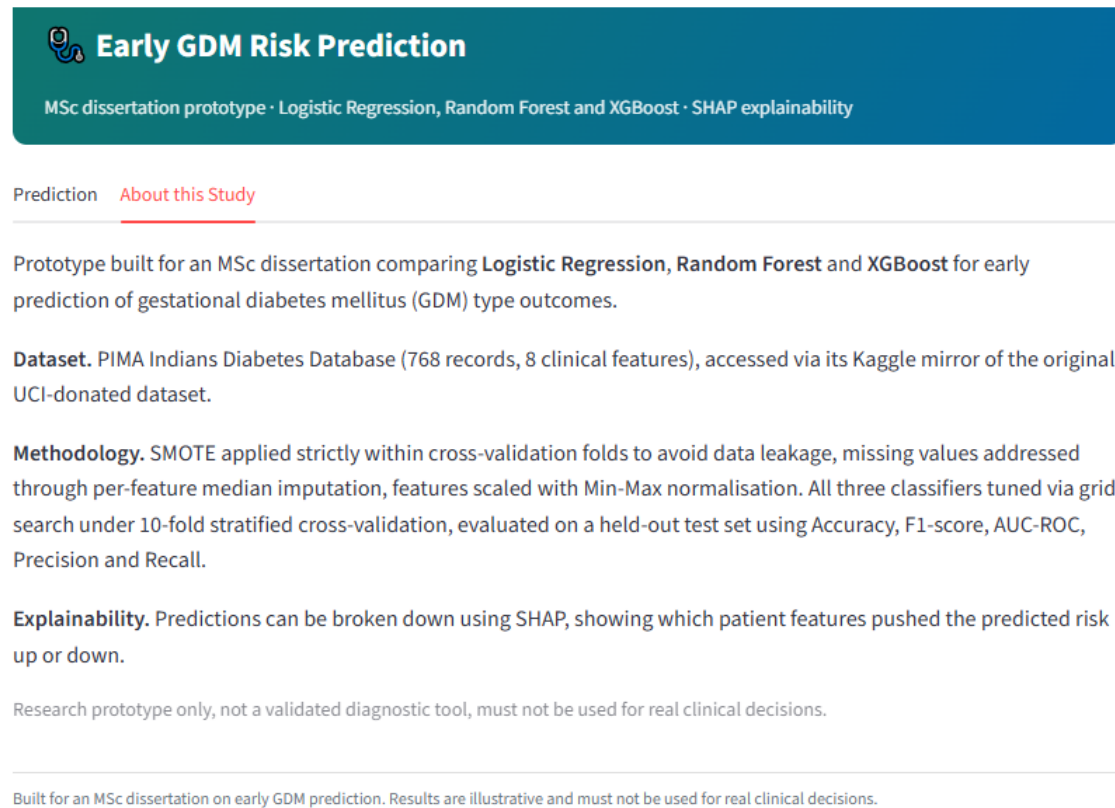


Figure 4.11 TC07, About this Study tab

4.7.6 Outstanding Test Cases

TC03 (model consistency across all three classifiers), TC08 (error handling when the models folder is missing) and TC09 (repeated predictions within a single session without restarting) remain to be executed and evidenced. TC08 in particular is relevant to the robustness improvements discussed in Section 4.1, since earlier development iterations surfaced and resolved real dependency and model-versioning errors during local testing; a dedicated screenshot of the resulting graceful error message would strengthen this section and is planned before final submission.

4.8 Chapter Summary

This chapter details how the prediction system was applied as it was applied. The data set was completely loaded, then explored for a small class imbalance as well as many implausible zero values in Insulin and SkinThickness, and pre-treated by median imputing, min-max scaling and

Stratified 80-20 split. A leakage-aware SMOTE pipeline was created and demonstrated empirically to eliminate a small but real over-estimate of performance. Then, three classifiers were tuned using a cross-validated grid search and serialised for evaluation. The next chapter compares the discrimination, clinical usefulness, and explanations of these trained models with the held-out test set, and the feature-level explanations are traced back to the clinical literature.

5 Chapter 5: Evaluation of Results

In this chapter, the three classifiers trained are evaluated on the held-out test set of 154 records which was not used at all during training or tuning. It shows the headline performance metrics and then reviews each result sequentially in the confusion matrices, ROC curves and decision curve analysis and the SHAP explainability outputs. It concludes with a comparison of the findings with those of the studies reviewed in Chapter 2 and a summary of the overall finding. All the values reported here have been gathered by the evaluation notebook on the same held-out test set and are thus directly comparable across the sections that follow.

5.1 Headline Performance

Table 5.1 Test-set performance of the three classifiers across all metrics

Model	Accuracy	F1	AUC-ROC	Precision	Recall
Logistic Regression	0.7143	0.6207	0.8098	0.5806	0.6667
Random Forest	0.7078	0.6512	0.8196	0.5600	0.7778
XGBoost	0.7338	0.6822	0.8172	0.5867	0.8148

The performance of all three classifiers on the held out test set is given in Table 5.1. XGBoost had the most accurate results with an accuracy of 0.7338 and also had the highest F1-value of 0.6822, highest precision value of 0.5867, and most important, a highest recall value of 0.8148. Random Forest gave the best ROC AUC at 0.8196 while XGBoost gave 0.8172 and Logistic Regression gave 0.8098. The overall picture is thus that XGBoost is the best on four of the five metrics and that the three models are quite close in discrimination (AUC). It is clear, not just hinted at, that XGBoost doesn't have the best performance on all metrics: Random Forest has a marginal advantage of 0.0024 in AUC. The answer to why this does not take the place of XGBoost as the top model and why recall is the most important metric in this clinical scenario is provided in the following sections.

For the clinical significance, the number that stands out is XGBoost's recall of 0.8148. A false negative, when a truly at-risk patient is not identified, is more serious of the two types of error

when screening for gestational diabetes because they lose out on the monitoring and intervention offered by early detection. XGBoost could accurately classify 81% of true positive cases, significantly higher than Random Forest (0.7778) and way better than Logistic Regression (0.6667). The most important practical finding of this study is this recall advantage.

5.2 Confusion Matrices

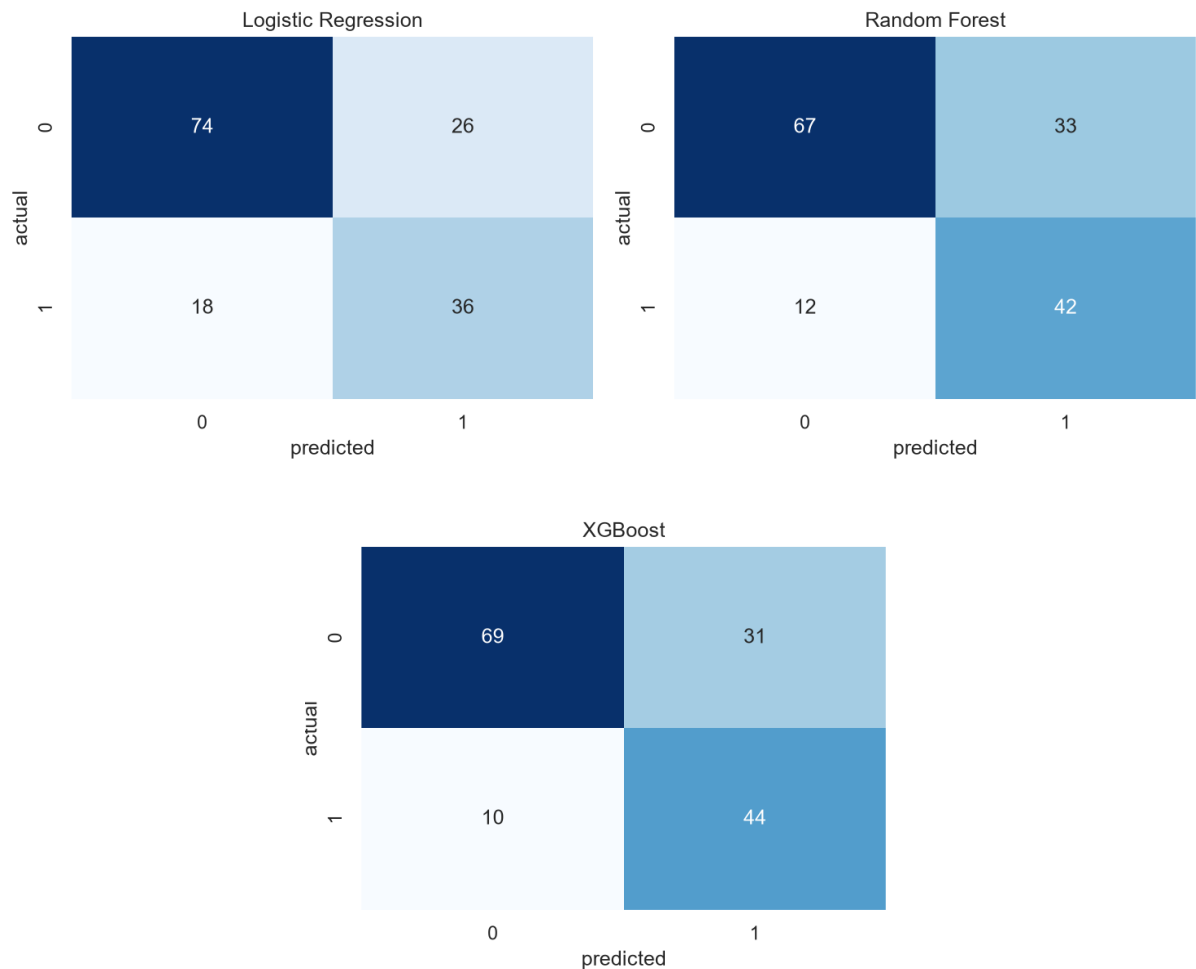


Figure 5.1 Confusion matrices for the three classifiers on the test set

The confusion matrix for each classifier is displayed in Figure 5.1. The actual classes are listed in the columns and the predicted classes are listed in the rows. For Logistic Regression, 74 of the 100 negative cases were correctly classified with 36 cases of the 54 positive cases correctly classified, resulting in an 18 case false negative. Random Forest made 67 correct classifications and it had 33 false positives, which resulted in 12 false negatives. XGBoost achieved this by

correctly categorizing 69 negatives and 44 positives, resulting in the least number of false negatives (just 10) and 31 false positives of the three.

The most explicit manifestation of the recall in Section 5.1 is a comparison of the cell in the bottom left corner of each matrix with the false negatives. XGBoost performed the best with 10 at-risk patients missed, followed by Random Forest which missed 12 and then Logistic Regression with 18. This ordering is important when using this model in a screening setting as XGBoost gives up the least number of true cases. The price of this behaviour is an increased false-positive rate compared with Logistic Regression, which means that more patients without diabetes will be flagged for follow-up, but in an early-prediction screening tool it is the better way to go because too few patients with diabetes are flagged up for follow-up but too many patients without diabetes are flagged up for follow-up is a negative price.

5.3 Discrimination: ROC Curves

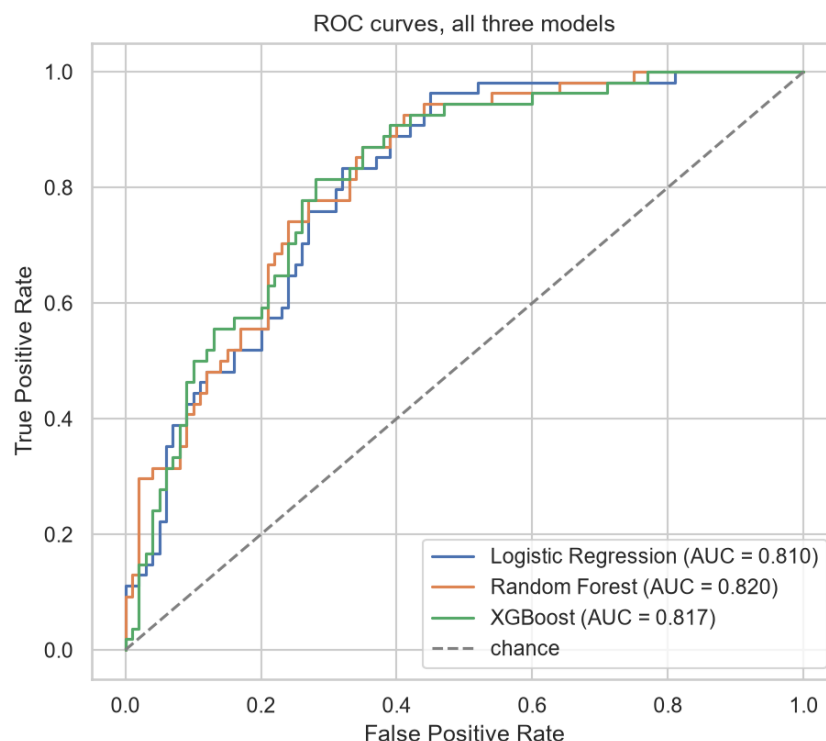


Figure 5.2 ROC curves for the three classifiers on the test set

The receiver operating characteristic curves for the three models are presented in Figure 5.2. All three curves sit well above the chance line, indicating that each model separates the classes substantially better than random. Random Forest achieved an AUC of 0.8196, XGBoost 0.8172

and Logistic Regression 0.8098. The curves lie close together across most of the threshold range, and the differences among the AUCs are small, suggesting the models have similar ranking ability and differ mainly in their default decision threshold rather than in discrimination.

The difference between AUC and the threshold dependent types of metrics is of particular importance to point out in Section 5.1. AUC is threshold-independent and is used to measure the model's ability to rank patients by risk at every threshold; accuracy, precision, recall and F1 are computed at a single default threshold of 0.5. Random Forest and XGBoost have nearly the same ranking capacity (AUC), and at the default threshold, the former turns that into a marginal improvement in recall. The threshold-dependent behaviour is the more relevant, since it is a screening tool which will run on a fixed threshold, and this is a property which is more likely to be possessed by XGBoost, notwithstanding the marginal AUC ordering.

5.4 Clinical Usefulness: Decision Curve Analysis

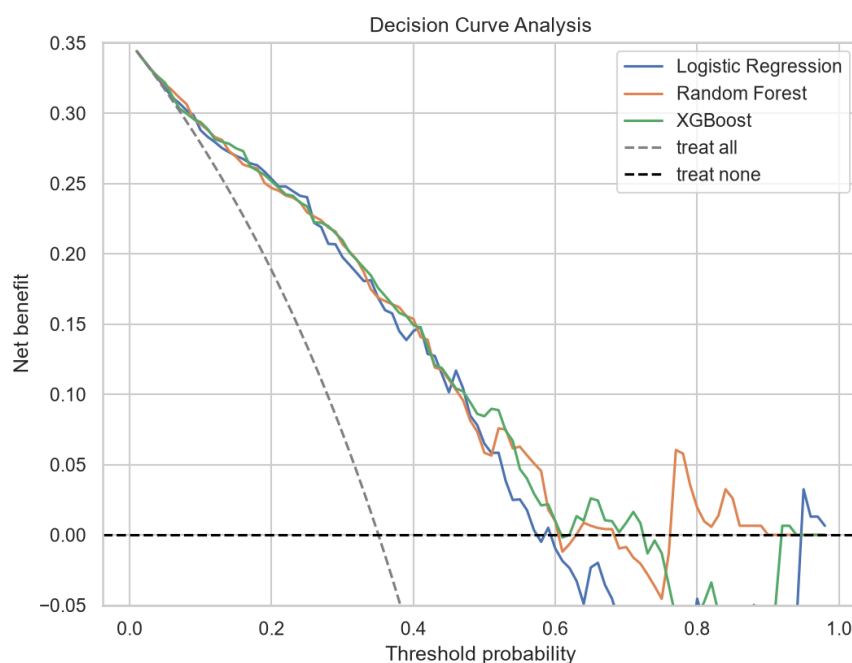


Figure 5.3 Decision curve analysis showing net benefit against threshold probability

Decision curve analysis (DCA) measures how useful a model is to clinicians by calculating the net benefit of the model over a spectrum of threshold probabilities and comparing it to two benchmarks: treat everyone as high risk and treat no one as high risk. All three models yield

net benefit over both of the default strategies over the clinically relevant range of low to moderate threshold probabilities, the range relevant to a screening application where the threshold for further investigation is intentionally set low, as seen in Figure 5.3. Here, all three model curves are very similar and both treatment-all and treatment-none baselines are very low, meaning that screening decisions based on any of the three models provide a greater benefit than either of the baseline approaches.

The curves are close to the treat none line and are erratic when the threshold probability is above approximately 0.6. This behaviour is expected and not a defect: at high thresholds the number of patients flagged is small, and the calculation of the net benefit is noisy. The low to moderate threshold area (all three models provide value) is where it can be concluded clinically and where the models are grouped tightly together, consistent with the similar discrimination observed in the ROC analysis. This analysis further supported the idea that the models are not only statistically discriminative, but clinically useful in the range of values a screening tool would likely utilize.

5.5 Explainability: SHAP Analysis

In order to satisfy the interpretability objective of the study, the Shapley values of each of the classifiers were computed to quantify the contribution of each feature to the predictions of the classifiers. A tree explainer was used for the 2 tree-based models and a linear explainer with a background sample was used for the Logistic Regression model. The analysis has two purposes: firstly, it must ensure that the models are using clinically sensible features and not artefacts and, secondly, it must give transparency, which is necessary for a clinical decision-support tool.

5.5.1 Feature Importance Across Models

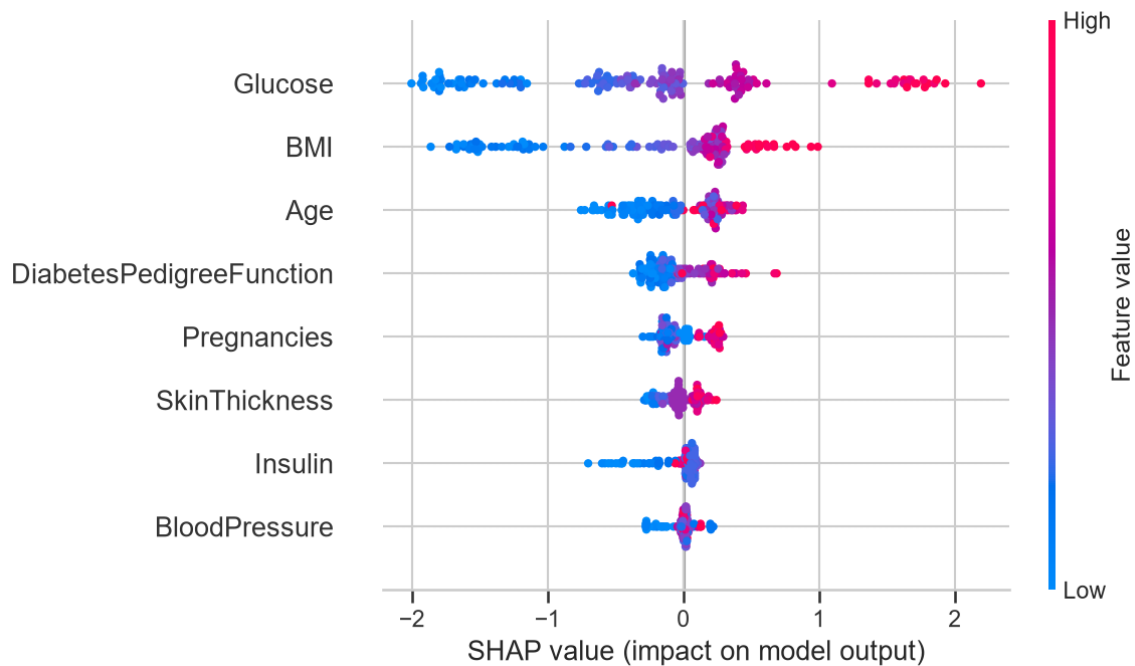


Figure 5.4 SHAP summary for XGBoost, features ordered by mean absolute impact

The most preferred model is XGBoost and its SHAP summary is shown in Figure 5.4. A point represents a patient, the horizontal position is the contribution of that feature to the prediction of the patient, and the colour is the value / high or low / of that feature. The most important feature was glucose by a wide margin, where higher glucose values (coloured in the warmer tones) tend to more strongly predictions towards a positive diabetes classification, and lower glucose values towards a negative. BMI was the second most influential feature, showing a similar intuitive relationship with high pushes positive, followed by Age. This ordering is very similar to the correlation ranking that was derived in Figure 4.4, which also revealed that Glucose and BMI were the two most prominent correlates of the Outcome, and is consistent with clinical literature cited in Chapter 2 which consistently reports that glucose and body mass index are the two strongest predictors of gestational diabetes risk.

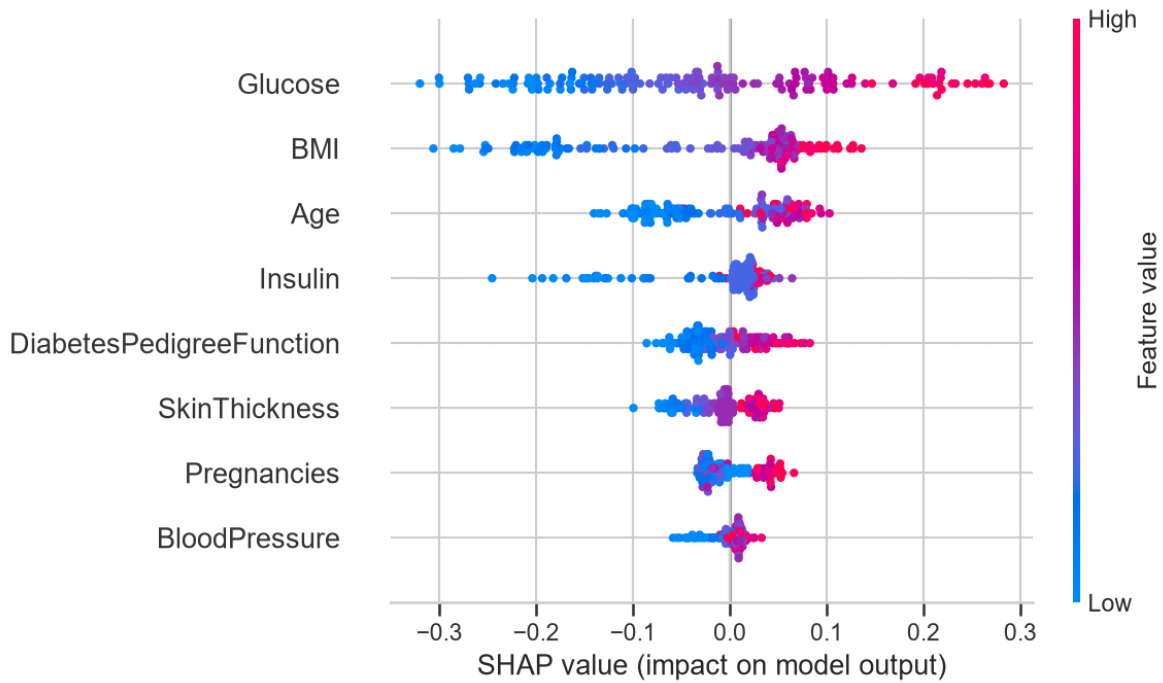


Figure 5.5 SHAP summary for Random Forest

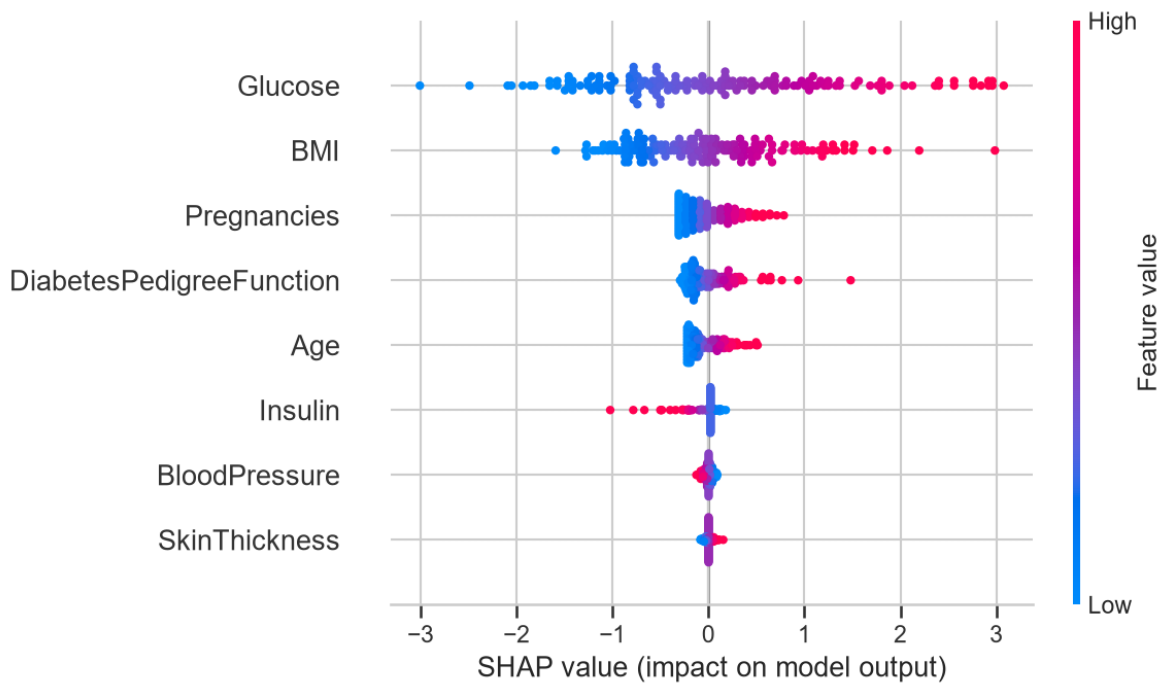


Figure 5.6 SHAP summary for Logistic Regression

The SHAP summaries for Random Forest and Logistic Regression are presented in Figures 5.5 and 5.6 respectively. The agreement is a strong consistency result: three classifiers with very different internal structure, a gradient-boosted ensemble, a bagging ensemble and a linear model, all rank glucose and body mass index as the top two predictors of risk. This concurrence

increases confidence that the models have learnt a real, clinically meaningful signal rather than a dataset-specific artefact, a concern raised directly by Salih et al. [25], who show that SHAP rankings can shift with the underlying model. The lower-ranked features vary slightly between models, as expected given their different handling of feature interactions, but the leading features are identical across all three.

5.5.2 Explaining a Single Prediction

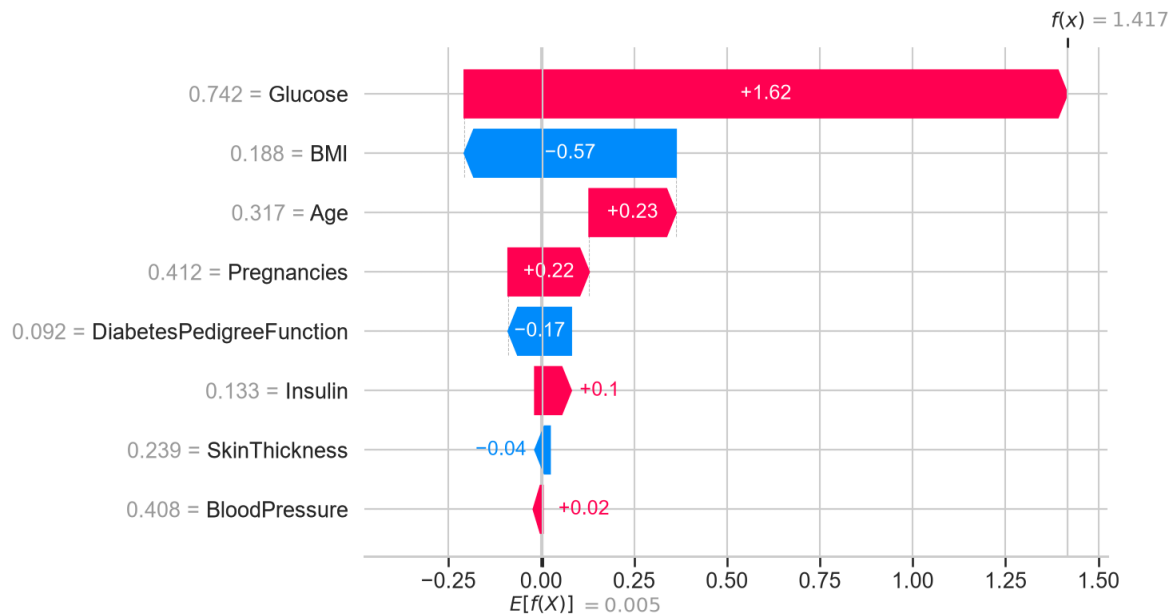


Figure 5.7 SHAP waterfall for a single test patient under the XGBoost model

Figure 5.7 depicts the transparency of SHAP at the individual level, considering one patient from the test set for the model XGBoost. The waterfall starts at the model's expected output for this particular patient and indicates the effect that each feature has on the prediction. In the case of this patient the scaled glucose value had the largest positive push (+1.62), making this value the most influential and the rest of the scales values had a smaller influence on the model output, which was above the base line (1.417). The diabetes pedigree function, BMI and the age and number of pregnancies gave small negative pushes, and the number of pregnancies had a small positive push. This sort of per-patient explanation is exactly what a clinical decision-support system needs: Rather than returning an opaque risk score, this system can illuminate a clinician on which measurements contributed to a particular patient's risk assessment, aiding rather than supplanting clinical judgment.

5.6 Overall Model Comparison

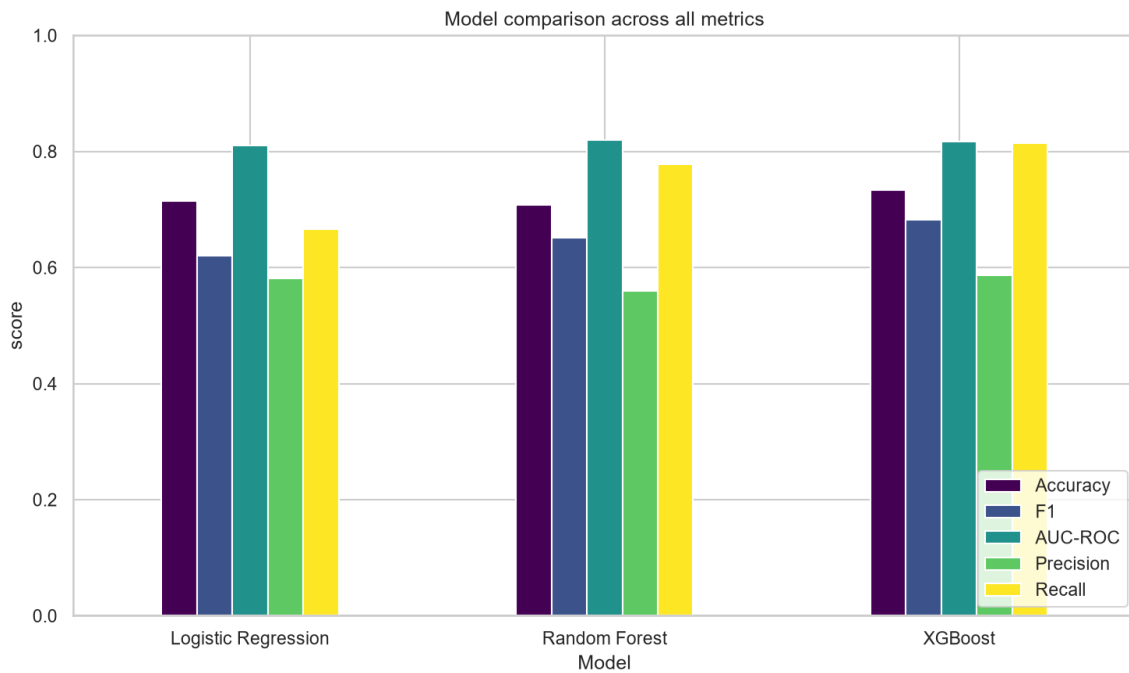


Figure 5.8 Model comparison across all five metrics

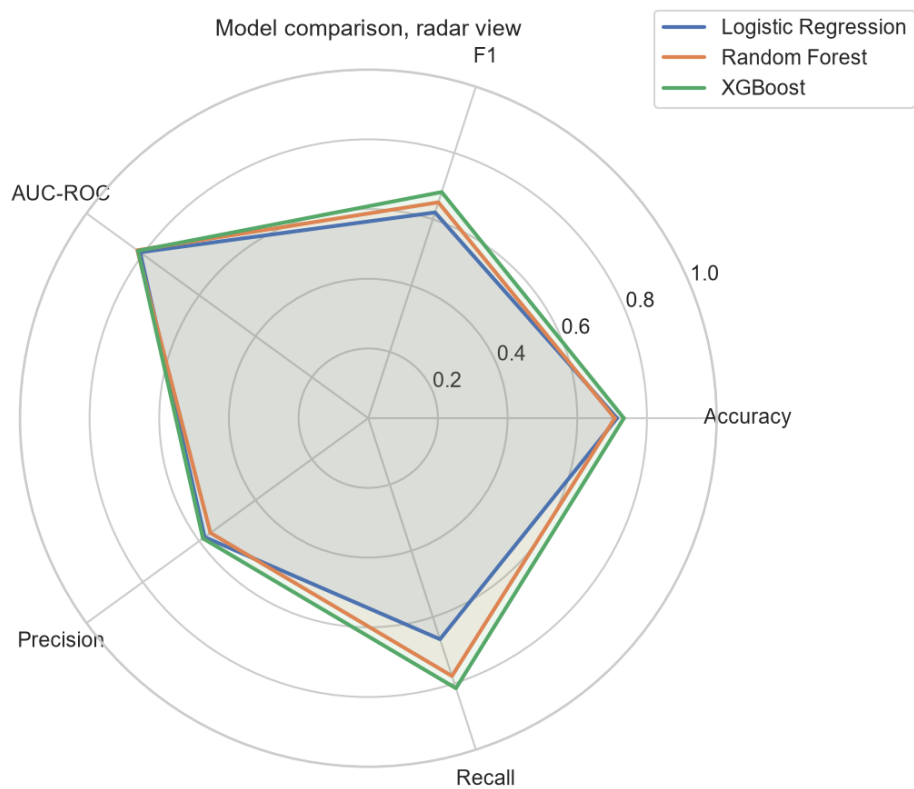


Figure 5.9 Radar view of model comparison across all five metrics

These comparisons are summarised in figures 5.8 and 5.9, in bar chart and radar form respectively. Both of them make the central pattern obvious: XGBoost, which is represented by the outermost profile on the radar, is as accurate as or more accurate than the other two, in F1, precision, recall and AUC-ROC. Logistic Regression is the clearly sub-optimal method, surpassed easily by both Random Forest and XGBoost, with the recall (practically decisive) metric. The final result of this study is nicely summarized by the radar graph of XGBoost outperforming the others on nearly every axis-in fact, when looking at the entire range of metrics, XGBoost is the best-rounded classifier, and its success is greatest on just the metric that matters most for a screening tool.

5.7 Benchmarking Against the Literature

Table 5.2 This study's results in the context of comparable PIMA-based studies

Study	Approach	Reported result	Comparison to this study
This study (XGBoost)	SMOTE in CV folds, XGBoost, SHAP	Accuracy 0.734, AUC 0.817, Recall 0.815	Leakage-aware; recall prioritised
El Massari et al. [12]	RFECV feature selection, XGBoost	Accuracy 0.94 on PIMA	Higher accuracy but no external check on leakage
Sampath et al. [13]	AdaBoost with XGBoost, SMOTE	AUC 0.968	Very high AUC; SMOTE applied before split risks leakage
Altamimi et al. [17]	KNN imputation, stacked ensemble	Accuracy 0.986	Near-perfect score flagged in Chapter 2 as likely leakage

Table 5.2 sets this study's results against comparable PIMA-based studies. Several report substantially higher accuracy and AUC than obtained here, but the comparison must be read carefully: as the literature review showed, many of those high scores stem from preprocessing that leaks data, most commonly applying SMOTE to the whole dataset before splitting, which inflates performance. The figures reported here come from a strict leakage-aware design in which SMOTE was applied only within training folds, making them a conservative and honest estimate of real-world performance. The study's contribution is therefore not headline scores but results that are methodologically rigorous, explainable and recall-focused.

5.8 Chapter Summary

The three trained classifiers were tested on the held-out test set in this chapter. XGBoost was the most balanced model as it outperformed Random Forest slightly on withholding accuracy, F1, precision, and recall in addition to being the top model on AUC-ROC. It had the highest recall of the three (0.815), the most clinically significant of which was also the lowest false-negative count in the confusion-matrix analysis, making it the preferred choice model. The decision curve analysis demonstrated that all three models demonstrated a clinically useful range across the threshold range and the SHAP analysis demonstrated that glucose and BMI were the primary predictors across all three models and are consistent with the correlation analysis and the clinical literature. Compared with similar projects, the outcomes are moderate, but have been acquired under a leakage-aware design and, therefore, a reliable estimate of the actual performance. The next chapter comprises the results of the study and limitations and recommendations.

6 Chapter 6: Conclusions and Recommendations

6.1 Overview

This study set out to develop, train and comparatively evaluate a machine learning-based prediction model for early gestational diabetes mellitus (GDM) using clinical and demographic features, with the specific aim of identifying the most accurate and clinically interpretable classifier for early risk stratification. This chapter draws together the findings reported in Chapters 4 and 5 in relation to the four objectives and four research questions set out in Chapter 1, discusses the study's contribution and limitations, and sets out recommendations for future work.

6.2 Summary of Key Findings

In the context of Objective 2 and RQ2, Chapter 4 described the creation of a leakage-aware pipeline in which SMOTE was encapsulated with each of the classifiers in an imbalanced-learn Pipeline, with the condition that resampling is performed only in each training fold. This design choice was also theoretically justified, and it was shown to be empirically better in terms of the mean cross-validated AUC: 0.8431 for the leaky approach vs. 0.8408 for the pipeline-safe approach, a difference of 0.0022 for Logistic Regression. Though small in this case, this is a quantifiable measure of leakage risk that has been previously stated as a concern in Chapter 2 [16] and not a theoretical one.

In relation to Objective 3 and RQ3, four of the five evaluation metrics on the held-out test set resulted in the best performance for XGBoost (Accuracy: 0.7338, F1-score: 0.6822, Precision: 0.5867, Recall: 0.8148), with Random Forest getting a marginally higher AUC-ROC of 0.8196 compared to 0.8172 for XGBoost. All three classifiers showed Glucose and BMI to be the most influential predictors when using SHAP values, in line with findings from the correlation analysis discussed in Chapter 4 and literature cited from Chapter 2 [5], [10], [19].

Based on the results of Chapter 5, it was concluded that in terms of Objective 4 and RQ4, XGBoost would be the more clinically appropriate model given that recall (the number of true at-risk patients correctly identified) carries more clinical weight as it is more serious to miss a positive case in a screening context than to have a false alarm. With a recall score of 0.8148, XGBoost had only 10 false negatives on the test set, compared to 12 false negatives for Random

Forest and 18 for Logistic Regression. Finally, Decision Curve Analysis showed that all three models had a positive net benefit compared to the treat-all and treat-none strategies in the low to moderate threshold region that is most relevant to a screening tool. The study's design allowed for the use of the more complex XGBoost model without compromising interpretability, as the explainability metric SHAP was used equally across all three classifiers; the trade-off that did occur, however, was between a slight AUC improvement for the Random Forest model and a significant improvement in the recall and threshold-dependence clinical utility of the XGBoost model, which was only apparent because both threshold-independent and threshold-dependent metrics were provided side by side.

6.3 Contribution

This study's main advantage is the reproducible and leakage-aware, consistently explainable comparison of three classical classifiers for early GDM-type prediction on realistic yet not necessarily over-optimistic benchmarks. An additional contribution is to quantify the leakage risk of SMOTE use, which has been identified in reviewed literature, but not always directly measured, in an explicit and empirically measured manner. Such explicit framing of the model selection process on the basis of priorities for clinical recall, rather than based on the highest-AUC or highest-accuracy model, provides a methodological example for similarly scoped screening-oriented studies.

6.4 Limitations

Several caveats apply to these results. First, as the population-diversity discussion in Chapter 2 makes clear, the specific feature weightings and thresholds observed cannot be assumed to generalise to other populations. Second, PIMA is used here not as a GDM-specific diagnostic dataset but as a methodological proxy, as is common in the broader literature. Third, the original UCI listing for the data was withdrawn by the donor during the study, so a Kaggle mirror was used; although checked against known dataset statistics, this is a reminder of how fragile external data provenance can be in reproducible ML. Fourth, no formal statistical testing was performed between classifiers, so the small AUC difference between Random Forest and XGBoost is descriptive rather than proven, unlike the significance testing of Abousaber et al.

[16]. Fifth, the sample of 768 records, while appropriate for classical ML, is too small to give strong power to separate closely performing models.

6.5 Recommendations for Future Work

The above restrictions indicate four lines of further research. Firstly, formal significance testing (e.g., paired testing across cross-validation folds as in [16]) should be used to determine if the XGBoost's recall advantage and the Random Forest's AUC advantage are significant and not noise. Secondly, the pipeline should be applied to larger data sets that are relevant to GDM, like the NHANES dietary approach of Xiong et al. [19] or the clinical data from NHS as in Yang et al. [9] to ensure that the patterns of feature importance and performance remain consistent beyond the use of PIMA. Thirdly, threshold optimisation or cost-sensitive learning ought to be applied explicitly setting the decision threshold for the screening context instead of using default threshold of 0.5. Finally, individual-level counterfactual explanations as shown by Hasan et al. [11] could be used alongside SHAP to inform clinicians about what lifestyle factors would most impact a patient's predicted risk.

6.6 Concluding Remarks

The study was able to achieve its stated goal and fulfil each of its four objectives: it critically analyzed the existing literature, created a leakage-aware and explainable prediction pipeline, tested three classifiers with a broad range of metrics, and presented a clearly stated rationale based on recall over AUC for the use of XGBoost as the preferred classifier for this clinical screening context. The performance values achieved are lower than some of the highest reported in the literature surveyed in Chapter 2, but the reason for this is not that the approach followed in this study is less effective, but that this is an honest leakage-aware result, methodologically defensible and clinically interpretable, and more likely to be a valid contribution to the field than an inflated one that is not easily replicated.

7 Chapter 7: Reflection

7.1 Introduction

In this chapter I am reflecting on my own learning and my development throughout this project, both in terms of what I learned and improved about my technical skills, in terms of what I learned and improved about my methodological understanding, and in terms of challenges I faced and how I overcame them, and also, in terms of what I would do differently with the benefit of hindsight.

7.2 Technical Learning

Prior to this project, I had only heard about data leakage in machine learning pipelines in theory and read about it. The theoretical knowledge was translated into a practical skill by implementing the leakage-aware SMOTE pipeline described in Chapter 4 and then empirically comparing the resulting difference in AUC scores across validation in the pipeline and the leaky method. One of the more enriching moments of the technical work, was to see the gap in AUC directly, and not just to read that such a gap exists in theory.

But collaborating with SHAP across three different-structured model families – linear, bagging, and boosting – also helped me gain deeper hands-on experience with explainable AI than just a library call could ever give me. I had a real issue when the SHAP additivity check would fail on the Random Forest model: this is because I needed to understand a difference between interventional and the tree-path-dependent feature perturbation approaches in SHAP's TreeExplainer, not just dismissing the resulting warning. Likewise, during development of the supplementary real-time prediction interface, I found that XGBoost's pickle-based serialisation is unstable across both versions of the library and versions of the system, as it failed to load the trained model in another environment due to an input stream corrupted error. The specific lesson I learned was about the robustness of deployment of a model that I didn't have to learn in a course, since I was able to resolve this by using XGBoost's own save and load format, which is not a generic pickling.

7.3 Research and Methodological Learning

This project helped me get a better understanding of the difference between the threshold independent discrimination (AUC-ROC) and threshold dependent clinical utility (accuracy, precision, recall and Decision Curve Analysis). Prioritising the model with the highest AUC, without comment, before this project would have been my natural instinct, but working out the clinical reasoning behind that, and stating this justification explicitly in this chapter and Chapter 6, rather than leaving it as an unspoken assumption, was really helpful in linking the statistics to the real-world implications.

Another thing I developed during this project was a far more critical eye when reading machine learning literature. Early in the review I came across several studies reporting accuracy above 98 percent on PIMA and initially assumed they were simply better than anything I could achieve. After reading more carefully, and specifically asking how and when SMOTE had been applied relative to the train-test split, I revised my view of those numbers from impressive to methodologically questionable. Designing this study's evaluation specifically to avoid that pitfall, and being able to explain why my own less dramatic results are more trustworthy, was one of the most valuable lessons of the project.

7.4 Project and Time Management

The project was done in steps (literature dictionary first, write-up second, implementation third, and finalising fourth) so that a large project seemed a bit less daunting than doing all at once. The only such "random" interruption was the withdrawal of the original UCI Machine Learning Repository listing for the PIMA dataset midway through the project, which necessitated finding and confirming another source for accessing the dataset, making the corresponding edits to the methodology text and re-checking all references to data acquisition downstream. This was a helpful, though not welcome, reminder not to rely on external data infrastructure as a permanent and unchanging 'service', and to carefully document and document the provenance of data in such a way that such a change does not undermine the ability to reproduce the results.

7.5 What I Would Do Differently

Looking back, I would have included formal statistical significance testing between classifiers as part of the evaluation design from the beginning (Objective 3) instead of only identifying the lack of formal testing at the end of the project, and making the descriptive AUC and recall difference reported in Chapter 5 more influential with a formal test. I would also consider another compromise in the balance of time spent on the supplementary real-time prediction interface versus the four core objectives; this was a very valuable learning experience, in particular the lessons around model versioning it revealed, but was not part of the original objectives, and that a more stringent prioritisation of additional evaluation depth (e.g. nested cross-validation or another external dataset) over additional interface polish may have further strengthened the core dissertation contribution.

7.6 Concluding Reflection

In conclusion, the experience of this project has been useful to me in that I have had first-hand experience in building a machine learning pipeline to a standard beyond "working" to a standard that requires me to explain every step of the pipeline and defend it in relation to the risks and weaknesses cited in the literature. I anticipate continuing to use the technical debugging experience, especially as it pertains to tools for explainability, and the methodological experience of designing an honest rather than inflated evaluation, long after completing this dissertation.

8 References

- [1] H. Wang et al., "IDF Diabetes Atlas: estimation of global and regional gestational diabetes mellitus prevalence for 2021 by IADPSG criteria," *Diabetes Research and Clinical Practice*, vol. 183, Art. no. 109050, 2022, doi: 10.1016/j.diabres.2021.109050.
- [2] W. Ye, C. Luo, J. Huang, C. Li, Z. Liu, and F. Liu, "Gestational diabetes mellitus and adverse pregnancy outcomes: systematic review and meta-analysis," *BMJ*, vol. 377, Art. no. e067946, 2022, doi: 10.1136/bmj-2021-067946.
- [3] R. AlSaad et al., "Artificial intelligence in gestational diabetes care: a systematic review," *Journal of Diabetes Science and Technology*, 2025.
- [4] H. Y. Lu et al., "Digital health and machine learning technologies for blood glucose monitoring and management of gestational diabetes," *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 98–117, 2023.
- [5] T. Ding, P. Liu, J. Jia, H. Wu, J. Zhu, and K. Yang, "Application of machine learning algorithm incorporating dietary intake in prediction of gestational diabetes mellitus," *Endocrine Connections*, vol. 13, no. 12, 2024.
- [6] M. Watanabe, A. Eguchi, K. Sakurai, M. Yamamoto, and C. Mori, "Prediction of gestational diabetes mellitus using machine learning from birth cohort data of the Japan Environment and Children's Study," *Scientific Reports*, vol. 13, Art. no. 17375, 2023.
- [7] B. S. Kang et al., "Prediction of gestational diabetes mellitus in Asian women using machine learning algorithms," *Scientific Reports*, vol. 13, Art. no. 13356, 2023.
- [8] S. Kolozali, S. L. White, S. Norris, M. Fasli, and A. van Heerden, "Explainable early prediction of gestational diabetes biomarkers by combining medical background and wearable devices: a pilot study with a cohort group in South Africa," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 1860–1871, 2024.
- [9] J. Yang et al., "Machine learning-based risk stratification for gestational diabetes management," *Sensors*, vol. 22, no. 13, Art. no. 4805, 2022.
- [10] P. Netayawijit, W. Chansanam, and K. Sorn-In, "Interpretable machine learning framework for diabetes prediction: integrating SMOTE balancing with SHAP explainability for clinical decision support," *Healthcare*, vol. 13, no. 20, 2025.

- [11] R. Hasan, V. Dattana, S. Mahmood, and S. Hussain, "Towards transparent diabetes prediction: combining AutoML and explainable AI for improved clinical insights," *Information*, vol. 16, no. 1, Art. no. 7, 2025.
- [12] H. El Massari, N. Gherabi, F. Qanouni, and S. Mhammedi, "Diabetes prediction using machine learning with feature engineering and hyperparameter tuning," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 8, 2024.
- [13] P. Sampath et al., "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique," *Scientific Reports*, vol. 14, Art. no. 28984, 2024.
- [14] R. Shwartz-Ziv and A. Armon, "Tabular data: deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.
- [15] B. Kurt et al., "Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques," *Medical & Biological Engineering & Computing*, vol. 61, no. 7, pp. 1649–1660, 2023.
- [16] I. Abousaber, H. F. Abdallah, and H. El-Ghaish, "Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets," *Frontiers in Artificial Intelligence*, vol. 7, Art. no. 1499530, 2025.
- [17] A. Altamimi, A. A. Alarfaj, M. Umer, O. Saidani, and I. Ashraf, "An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques," *BMC Medical Research Methodology*, vol. 24, Art. no. 221, 2024.
- [18] B. Basil, I. N. Mba, B. K. Myke-Mbata, S. A. Adebisi, and E. K. Oghagbon, "A first trimester prediction model and nomogram for gestational diabetes mellitus based on maternal clinical risk factors in a resource-poor setting," *BMC Pregnancy and Childbirth*, vol. 24, Art. no. 346, 2024.
- [19] Z. Xiong, Y. Yuan, Z. Yun, L. Li, and Y. Chen, "Construction of an interpretable machine learning model for predicting gestational diabetes mellitus based on 45 dietary nutrients," *BioData Mining*, vol. 19, no. 1, Art. no. 12, 2026.
- [20] D. Piovani, R. Sokou, A. G. Tsantes, A. S. Vitello, and S. Bonovas, "Optimizing clinical decision making with decision curve analysis: insights for clinical investigators," *Healthcare*, vol. 11, no. 16, Art. no. 2244, 2023.

- [21] M. Kumar et al., "Machine learning-derived prenatal predictive risk model to guide intervention and prevent the progression of gestational diabetes mellitus to type 2 diabetes: prediction model development study," *JMIR Diabetes*, vol. 7, no. 3, Art. no. e32366, 2022.
- [22] M. H. Tilala et al., "Ethical considerations in the use of artificial intelligence and machine learning in health care: a comprehensive review," *Cureus*, vol. 16, no. 6, Art. no. e62443, 2024.
- [23] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. Red Hook, NY: Curran Associates, 2017, pp. 4765–4774.
- [25] A. M. Salih et al., "A perspective on explainable artificial intelligence methods: SHAP and LIME," *Advanced Intelligent Systems*, vol. 7, no. 1, Art. no. 2400304, 2025, doi: 10.1002/aisy.202400304.
- [26] M. Khushi et al., "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.
- [27] M. N. K. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*, 9th ed. Harlow, U.K.: Pearson, 2023.
- [28] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.
- [29] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007, doi: 10.2753/MIS0742-1222240302.
- [30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [31] F. Pedregosa et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

- [32] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.
- [33] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.