Bachelor Thesis

Bachelor of Science (BSc.)

Department of Tech and Software

Major: Software Engineering

Application of Machine Learning Techniques for the Early Detection of Diabetes: A Comparative Study of Classification Models

Sebastian Russo

Matriculation Number: 79117092

First supervisor: Dr. Rand Kouatly

Second supervisor: Dr. Souad El Hassanie

Submitted on: 07.2025

Bibliographic Description and Presentation

- Bachelor in Science thesis 2025, 96 pages, 31 figures, 28 tables, 10 equations, 50 references.
- University of Europe for applied sciences, Department of Tech and Software

Statutory Declaration

I hereby declare that I have developed and written the enclosed Bachelor's Thesis completely by myself and have not used sources or means without declaration in the text. I clearly marked and separately listed all the literature and all the other sources which I employed when producing this academic work, either literally or in content. I am aware that the violation of this regulation will lead to the failure of the thesis.

Sebastian Russo

07.2025, Potsdam, Germany

Author’s signature

Abstract

Diabetes is a very common chronic disease and its detection in its early stages constitutes a determining factor for preventing severe health complications derived from it and the reduction of the strain it poses on both healthcare systems and the patients themselves. The present study explores and compares the performance of several different supervised machine learning model algorithms, specifically Logistic Regression, Decision Tree, Random Forest and Support Vector Machines. The models were trained and used in predicting diabetes from the Diabetes Binary Health Indicators BRFSS2015 dataset, comprised of over 250,000 entries with 21 health related predictor variables and 1 objective variable.

The selected standard classification metrics were accuracy, precision, recall, F1-score, ROC-AUC, K-fold cross validation and Confusion matrices to evaluate and compare the performance of each model. Among the implemented models, Random Forest achieved the highest overall performance, with an accuracy of 83.47%, a precision of 45.19% and a weighted F1-score of 0.83. On another note, Logistic Regression and Support Vector Machines showed the highest recall values (~75%), making them more effective at identifying actual diabetic or prediabetic patients.

The findings obtained from this study highlight a critical trade-off between the recall and precision from the trained models. While high recall reduces the missed diagnoses (false negatives), it increases incorrect diagnoses (false positives), which can burden healthcare systems and stress patients. It is suggested to consider a hybrid modelling strategy to combine the strengths of different models depending on the clinical context and priorities. The present study contributes to the growing field of intelligent healthcare systems and the importance of aligning model selection with medical priorities such as interpretability, sensitivity and robustness.

Table of Contents

List of Abbreviations

• **AI**	Artificial Intelligence
• **ANN**	Artificial Neural Networks
• **AUC**	Area Under the (ROC) Curve
• **AVG**	Average
• **BMI**	Body Mass Index
• **BRFSS**	Behavioral Risk Factor Surveillance System
• **CDC**	Center for Disease Control and Prevention
• **CPU**	Central Processing Unit
• **DL**	Deep Learning
• **EDA**	Exploratory Data Analysis
• **EHR**	Electronic Health Records
• **FN**	False Negative (Incorrect negative cases)
• **FP**	False Positive (Incorrect positive cases)
• **GDPR**	General Data Protection Regulation of the European Union
• **HIPAA**	Health Insurance Portability and Accountability Act of 1996 of the US
• **LinearSVC**	Linear Support Vector Classification
• **ML**	Machine Learning
• **OS**	Operating System
• **RAM**	Random Access Memory
• **RBF**	Radial Basis Function
• **ROC**	Receiver Operating Characteristic
• **ROC-AUC**	Receiver Operating Characteristic Area Under the Curve

• **SML**	Supervised Machine Learning
• **SMOTE**	Synthetic Minority Over-Sampling
• **SVM**	Support Vector Machine
• **T1D**	Type 1 Diabetes
• **T2D**	Type 2 Diabetes
• **TN**	True Negative (Correct negative cases)
• **TP**	True Positive (Correct positive cases)
• **UML**	Unsupervised Machine Learning
• **WHO**	World Health Organization

Table 1. List of abbreviations

Glossary of meanings

- Accuracy: Metric that measures overall correctness of a ML model (right predictions).
- ANN: Computing models that are inspired by the human brain, typically used for tasks like classification and pattern recognition
- AUC: Metric that indicates a ML model's ability to differentiate between classes. It is derived from the ROC curve.
- BMI: Health metric derived from height and weight to classify the body weight status.
- DL: Subset of ML models that utilize multi-layered neural networks to perform complex data tasks.
- F1-Score: Harmonic mean of precision and recall, balancing both metrics in one score.
- ML: Algorithms capable of learning patterns from a given data to make predictions and/or decisions.
- Precision: Metric that indicates how many of the positive predictions were actually correct.
- Recall: Metric that shows how well the model identifies actual positive cases.

- ROC: Plot that shows the performance of a binary classifier derived from the comparison of true and false positive rates.
- ROC-AUC: Single score that summarizes the ROC curve, where a higher value indicates a better performance because it measures the model's ability to distinguish between classes.
- SML: ML models that are trained exclusively on labeled data to predict outcomes.
- SMOTE: Method to balance class distribution by generating synthetic minority class samples.
- SVM: Supervised ML model that finds the best boundary to separate classes.
- UML: ML models that are trained exclusively on unlabeled data and have to find patterns without predefined outputs.

List of Tables

List of Figures

List of Equations

Chapter 1: Introduction

1.1 Background

Chronic diseases like diabetes are becoming increasingly prevalent among all societies around the world both in developed and developing countries, where there is a growing need for intelligent systems that can support early diagnosis and intervention, where traditional diagnostic that, while clinically accepted, often fail to utilize in full the vast amount of structured health data currently available.

Machine Learning (ML), a branch of Artificial Intelligence (AI), offers a promising solution by enabling automated and data driven prediction that can assist clinicians in identifying individuals at risk more efficiently and effectively. In the context of Software Engineering, integrating ML models into diagnostic tools represents a key advancement in building smart healthcare applications.

However, selecting the appropriate ML model involves a series of considerations related to trade-offs between accuracy, interpretability and deployment in an effective manner. This study evaluates and compares multiple supervised ML algorithms on a public health dataset to determine which of those models are the most effective and feasible for use in clinical decision supported systems.

1.2 Problem Statement

Diabetes is a chronic disease and an increasingly prevalent health affliction worldwide with significant risks for patients and burdening healthcare infrastructure and medical body. The early detection of this disease is key for its effective intervention and management. Traditional diagnostic methods are time and resource consuming and may not even leverage the full potential of the available data of the patient. With the increasing availability of datasets related to healthcare and the advancement of Machine Learning models, it is possible to use ML models for a more promising approach to determine diabetes disease at its early stages and more accurately. However, several considerations that must be taken into account and one of them is the need to assess which ML model performs the best and under which circumstances, particularly in terms of accuracy, interpretability and real world applicability since there is a lack of consensus on the best performing models for clinical and seamless implementation, which this proposed study aims to address.

1.3 Research Questions

1. Which Machine Learning algorithms provide the most accurate results in predicting diabetes based on health datasets?
2. How do different features (like cholesterol, BMI, age, sex etc.) influence the predictive power of ML models?
3. Can ensemble models outperform traditional single classifiers in diabetes prediction?
4. What challenges arise in implementing ML-based diagnostic tools in real-world healthcare systems?
5. How can the interpretability of ML models influence their adoption in clinical decision making in diagnostics?

1.4 Thesis Objectives

1.4.1 General Objective

- To explore and compare the performance and effectiveness of various machine learning classification models in the early detection of diabetes using a real world health dataset.

1.4.2 Specific Objectives

- To identify and preprocess the relevant diabetes related dataset.
- To implement and evaluate multiple ML algorithms, specifically Logistic Regression, Decision Tree, Random Forest and SVM.
- To assess the performance of each model using standard classification metrics.
- To analyze the feature importance and model interpretability in the context of medical decision making.
- To discuss the potential integration of ML models into healthcare systems for diagnostic support (not replacement).

1.5 Rationale of the Study

The present study contributes to the growing field of healthcare informatics and software engineering by evaluating the role of existing machine learning models in the medical field. As software solutions increasingly incorporate artificial intelligence more often, it is necessary to understand which models best support

diabetes prediction and can assist in the development of intelligent diagnostics systems. This study supports data driven healthcare, that can lead to improved patient outcomes through earlier intervention and may contribute to technological innovation and policies in medical diagnostics.

1.6 Scope

- The study focused on supervised classification machine learning models applied to the Diabetes Binary Health Indicators BRFSS2015 dataset.
- The dataset contains pre-cleaned, structured data with 21 total predictor variables and one binary target label.
- Evaluation is based on the previously mentioned, publicly available dataset derived from the Behavioral Risk Factor Surveillance System (BRFSS) by the CDC and curated by the kaggle user Alex Teboul.
- Model training followed a generalized pipeline; however, model specific parameters will be individually tuned as needed to optimize the performance of each algorithm.
- Evaluation metrics include standard classification metrics like Accuracy, Precision, Recall, F1-score, ROC-AUC, K-fold cross validation, Classification report, Training time and Confusion Matrix.

1.7 Delimitation

- The study does not involve real-time or live clinical data.
- The study does only the aforementioned dataset; no additional datasets will be combined or explored.
- It does not address diabetes treatment or progression prediction.
- It focuses on binary classification (diabetic/prediabetic vs non-diabetic), not multi-class (diabetic vs prediabetic vs non-diabetic for instance) or regression based problems.

Chapter 2: Theoretical Background

2.1 Introduction

The theoretical background outlines, as its name implies, the theoretical foundation that supports the present study on the application of machine learning (ML) techniques for early diabetes detection. This research has been based on the principles of data science, particularly for supervised machine learning and classification theory behind them. Through the establishement of a clear understanding of the average data science pipeline, the analysis of data sets (EDA), the utilized classification algorithms and the different model evaluation methods (metrics), it is possible to provide a conceptual foundantions that are necessary to implement and analyze machine learning models in the context of healthcare and diagnostics.

2.2 Data Science and ML Theory

Data science consists in an interdisciplinary field that utilizes and combines scientific methods/approaches, statistics, math, analytics, specialized algorithms and systems to extract information and insights from structured and/or unstructured data sets (Zarbin, Lee, Keane, & Chiang, 2021). Within this context, machine learning (a subfield of artificial intelligence) focuses on enabling systems to learn certain patterns from the provided data without hardcoding explicitly the program to do these tasks or know these aforementioned patterns beforehand (Badillo et al., 2020).

The following data science lifecycle, which is adapted and expanded from Wing (2019), usually includes the following stages:

- Data collection: Gathering and obtaining structured and/or unstructured data from various sources. Ethical principles are critical, especially when it comes to common people's private data.
- Pre-processing and data cleaning: Consists on removing inconsistencies, handling missing values and transform data into usable formats that the algorithms expect and that might negatively impact the training of the ML model.
- Feature engineering: From all the features that may be obtained from the dataset, selecting and transforming variables accordingly can greatly improve the model's performance.

- Machine learning model training: Different types of algorithms have been used, so the implemented ML models are capable to learn from a training set derived from the dataset and then tested with a smaller set (testing set). Each model implements a different set of parameters for tuning.
- Machine learning model evaluation: Subsequently to the training and testing of the ML model, it is necessary to evaluate and ensure that the model's performance is satisfactory by utilizing a specific set of metrics to validate its effectiveness and performance.
- Interpretation and deployment: Making the output of the model understandable and applicable to real world decision making and real case scenarios.

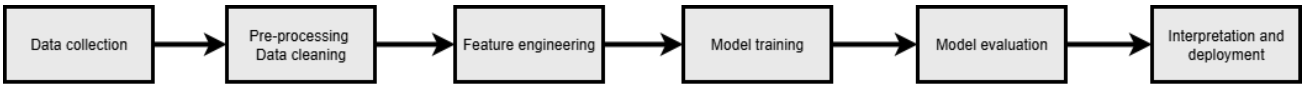

Figure 1. Data science lifecycle

Focusing on healthcare applications, this life cycle is crucial to develop accurate and interpretable predictive models that are appropriate to support clinical decisions, especially in the early diagnosis of diseases like diabetes.

2.3 Supervised Classification Algorithms

Supervised learning is a category of machine learning algorithms in which the model is trained on a dataset that is organized with input and output pairs. The ML algorithm model then starts learning how to map the input features (or X) to a target label (or Y), so, after being trained, the ML model is capable to predict a traget (or Y) for new, unseen, uncategorized data (Igual and Seguí, 2024). For the present study, the focus is completely on binary classification, where the goal is to predict if the patient is diabetic or not based on the given features of the selected dataset. The following consists in an overview of Supervised classification algorithms (from subsections 2.3.1, 2.3.2, 2.3.3 and 2.3.4).

2.3.1 Logistic Regression

According to Richards (2022), Logistic regression models specifically calculates the probability that a given input belongs to a class using sigmoid function:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)}}$$

Equation 1. Logistic Regression Formula

This formula provides the probability that the outcome y is 1 (as opposed to 0), given a set of features (the Xs), so the result always oscillates between 0 and 1. In this formula, "y" is the binary outcome, the "X" are the features and "β" are the coefficients or weights of each feature. The results of this model are interpretable and the model itself performs well with linearly separable data, but its performance tends to worsen as the dataset has too many features.

2.3.2 Decision Tree

According to Zhou (2022), Decision Trees split the given data based on feature values to form a structure similar to the form of a tree, where each of the nodes represent a decision and each "leaf" constitutes a prediction. They are fairly easy to interpret, handling both classification and regression tasks and support numerical and categorical data alike, however, an important note is that this ML model is very prone to overfitting the data, especially if the given dataset is too small or too noisy (presence of many outliers). The formulas used in this model are the following:

$$Gini\ impurity\ =\ \sum_{i=1}^{c} p(i)\ *\ (1-p(i))$$

Equation 2. Gini Impurity formula

Where "c" is the total classes and "p(i)" is the probability of picking a datapoint with class "I", this formula measures the probability of incorrectly classifying a randomly chosen element if it was labeled at random according to the distribution of the labels in the subset.

$$Entropy\ =\ -\sum_{i}^{c} p_i\ log_2\ p_i$$

Equation 3. Entropy formula

Where "pi" is the probability of picking an element of a class at random, this formula represents the amount of variance the data has or how heterogeneous the labels are in the subset.

$$Information\ Gain\ =\ Entropy(parent)\ -\ \sum_{i}(\frac{|D_i|}{|D|}\ *\ Entropy(D_i))$$

Equation 4. Information Gain formula

Information gain calculates the reduction in uncertainty it is gained regarding the target variable by splitting the dataset based on a particular attribute (feature), consisting in the difference between the original entropy of the "parent" set and the weighted entropy of the "child" after the split.

2.3.3 Random Forest

Random Forest is an ensemble machine learning method that constructs multiple decision trees during training to then output the majority of class (classification) as the final result. This ML model is capable of handling non-linear relationships well and provides feature importance. Since this model is an ensemble of Decision Tree, the formulas and math that support this model are the same as Decision Tree (Rahman, Md.A. et al, 2023).

It is also worth noting that, in comparison to Decision Trees, this ML is able to reduce overfitting and variance by leveraging several shallow trees. According to Liu, B. and Mazumder, R. (2024) Random forests implements:

- Bootstrap Aggregating (Bagging): Each tree is trained on a different random subset of the training data with replacements.
- Feature randomness: After each split in a tree, the Random Forest considers a random subset of features instead of all features.
- Model averaging: The final prediction, in this case for classification, is done by majority voting. In regression cases, it's done by averaging.

2.3.4 Support Vector Machines (SVM)

As described in the Scikit-learn official documentation, SVM realizes the optimal hyperplane that separates data into classes by maximizing the margin between support vectors. On the topic of non-linear data, certain types of SVM models can use the hardware's kernel in order to make it work as a Radial Basis Function (also known as RBF), which is used to project the given data into higher dimensions if necessary. Other types of SVM models, like LinearSVC, use a linear kernel that is optimized for linearly separable data, directly learning the linear decision boundary:

$$f(x) = w^T x + b$$

Equation 5. Decision Boundary equation

Where w is the weight of vector, x is input feature vector and b is the bias (intercept), this formula is the base of LinearSVC and is implemented in the following Hinge Loss formula.

$$Hinge\ Loss\ = \sum_{i=1}^{n} max(0,1\ y_i(w^T\ x_i\ +\ n))\ +\ \frac{\lambda}{2}\ ||w||^2$$

Equation 6. Hinge Loss formula

Where "yi" is the true label of the i^{th} example, "w^T *xi + n" is the implementation of the Decision Boundary equation, "max()" is the hinge loss that penalizes 0 loss when prediction is correct and outside the margin or misclassification, "$\lambda/2\ \|w\|^2$ " is the L2 regularization term that prevents overfitting. This formula is designed to enforce the idea of margin maximization, not only correct classification, but confident classification with a margin of separation.

2.4 Machine Learning Models comparison

Algorithm	Type	Interpretability	Training time	Handles Non-Linear	Robust to Outliers
Logsitic Regression	Linear	High	Fast	No	No
Decision Tree	Non-Linear	Medium-High	Fast	Yes	No
Random forest	Ensemble (Bagging)	Medium	Moderate	Yes	Yes
SVM	Linear (LinearSVC) or Non-linear (SVC with kernel)	Low-Medium	Fast (LinearSVC), Slow on large data (SVC)	Yes (SVC only)	Yes (SVC), No (LinearSVC without calibration)

Table 2. ML algorithms comparative table

2.5 Model Evaluation Metrics

The evaluation of ML models in the healthcare industry must go beyond just simple accuracy and satisfactory results. Misclassification, especially false negatives, can have severe consequences for the patients in medical contexts, therefore several metrics were used to evaluate each of the implemented models. The following overview of model evaluation metrics is adapted from Rainio, Teuho and Klén (2024):

- Accuracy: Proportion of total correct predictions (True positives and True negatives).

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

Equation 7. Accuracy formula

- Precision: Proportion of positive identifications (Positive predictions) that were actually correct (True).

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

Equation 8. Precision formula

- Recall (Sensitivity): Proportion of actual positives correctly identified (True positives).

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

Equation 9. Recall formula

- F1-Score: Consist in the harmonic mean of precision and recall times 2.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Equation 10. F1-score formula

- ROC-AUC: also known as Receiver Operating Characteristic Area Under the Curve, this metric measures the trade-off observed between true positive rate and the false positive rate, meaning that this metric measures the model's ability to distinguish between classes, where a 0.5 score constitutes random while 1.0 is a perfect separation of classes. The area under the curve provides an extra/aggregated performance measure for analysis.

The selected metrics offer a set of comprehensive insights into any model's performance, particularly working with imbalanced datasets that are very common in medical researches.

2.6 Feature Selection and its Importance in Machine Learning

Within the supervised machine learning category (moreover in medical diagnostics), feature selection constitutes a critical step that directly impacts the model performance, interpretability and generalization of unseen data. According to the review by Mahadeo, Dhanalakshmi and Balakrishnan (2022), Feature selection consist to the process of identifying and separating the most significant input variables (features or X) from the dataset because they are the input variables that contribute the most meaningfully to the prediction capabilities of the model when "deciding" the target outcome (predictions or Y), for the particular case of the present study, diabetes disease.

Feature selection constitutes a critical step in the data science pipeline and dataset preparation for the subsequent machine learning training (Adapted from Naheed et al., 2020):

- Improve model accuracy: Consists in the elimination of the irrelevant or redundant features that may just introduce noise and tamper with the accuracy.
- Reduce overfitting: By simplifying the model and ensures the focus is towards the most meaningful features only.
- Training Speed up: Simply lowering the dimensions of the data can increase the speed it takes to train any ML model.
- Enhance interpretability: It allows to better understand the influence of features and how they correlate to the output, especially in a healthcare context (indicators like BMI, age, glucose etc.), it is essential for its acceptance among professionals in the field.

2.7 Real-World Challenges for Deployment of ML models in Healthcare

Even though Machine Learning models often present an elevated accuracy in controlled environments and experiments, their deployment into a real world setting and specifically in a clinical environment presents a great deal of challenges that

must be addresssed not only to ensure the model's effectiveness and ethical use, but also when it comes to legal implications. Some of them include:

- Data privacy and Security: Healthcare data is highly sensitive information and subject to a great deal of regulations such as HIPAA in the United States or GDPR in the European Union, ensuring that data is securely stored, used responsibly and anonymous, this last point being specially non-negotiable for this and future researches (Ali, S et al., 2024).
- Integration with Clinical workflows: ML models must be integrated in a seamless matter into already existing hospital systems like Electronic Health Records (EHR) because tool's used by the medical staff can not disrupt their workflow and/or add complexity because it risks being ignored or underused (Sandhu, S et al. 202).
- Possible generalization across populations: ML models trained on specific datasets (for example, the BRFSS used in this study) do not mean that the same model will perform as correctly in other populations since datasets like BRFSS are based on the American population. This is because of inevitable biases or differences in demographics related to the lifestyle of the patients or the available healthcare infrastructure, making external validation essential for validation (Petersen et al., 2022).
- Interpretability and trust: The professionals working at a hospital or research institutton must be able to easily understand and completely trust the predictions of the ML models, especially because it involves life altering decisions like a disease diagnosis. The so called "Black-box" models are commonly avoided unless paired with interpretability tools for this very reason (Petersen et al., 2022).
- Legal and ethical concerns: Given that AI is still a very recent topic and a largely unregulated one to this date, a lot of questions that concern liability and fairness must be considered. Unfair models can reinforce health disparities and misdiagnoses (Petersen et al., 2022).
- ML model's maintenance: Healthcare data evolves over time and at an accelerated pace due to changes in clinical practices, diagnostic criteria or population health trends. Any ML model must be periodically retrained and

monitored to remain at an acceptable level of accuracy and relevance (Schinkel, M. et al, 2023).

2.8 Summary

The theoretical background chapter establishes the conceptual foundation for applying a supervised machine learning model (SML) and its techniques in the early diagnosis of diabetes. It outlines the common data science pipeline, which includes data collection, data preprocessing, feature engineering, model training, model evaluation and model deployment. Within this framework, the present study focuses on the binary classification to predict diabetes presence, where the results can be either a non-diabetic diagnosis or diabetic/prediabetic diagnosis.

The key ML models implemented in this study are the following:

- Logistic Regression: which is effective for linear problems, interpretable but limited with complex datasets.
- Decision Tree: intuitive model that features a tree-like shape but it is also prone to overfitting.
- Random Forest: an esemble approach that reduces overfitting while handling non-linear relationships.
- Support Vector Machines (SVMs): which is effective for high-dimensional and non-linear cases, certain types can use kernel functions at the cost of a slow training.

Feature selection also constitutes a crucial step for getting the best performance out of the implemented ML models that results in the improvement of accuracy, reduction of overfitting, speed up training and improving overall interpretability.

Lastly, real world challenges such as data privacy, seamless clinical workflow integration, population generalization, model interpretability, ethical and legal issues and ongoing ML model maintenance are discussed as important topics, highlighting the gap between the experimental success of research studies and the actual practical deployment in professional healthcare environments.

Chapter 3: Literature Review

3.1 Purpose of the Literature Review

The purpose of the literature review is to critically review pre-existing research related to the application of Machine Learning (ML) techniques for the early detection of diabetes. This chapter identifies the strengths, weaknesses and discoveries of previous related studies, with a focus on the usage of supervised ML algorithms for binary classification tasks in healthcare, specifically in the pre-emptive detection of diabetes disease. This review aims to highlight gaps in the current research, particularly in the comparative analyses of classification models, model interpretability and real world integration of ML tools in clinical settings, providing insights that acted as the foundation for the current study and justify its methodology and research questions.

3.2 Structure of the Literature Review

This chapter is organized in a thematical way to provide a coherent analysis of the field. Beginning with an overview of diabetes and the importance of early detection. Then exploring the role of Machine Learning to provide a coherent analysis of the field of medical diagnostics, followed by a detailed examination of the various ML models used for diabetes prediction. Then discuss the model performance comparison, interpretability in healthcare ML applications and challenges related to real world deployment of ML tools in live clinical environments. Lastly, a summary of the gaps identified and positioning the current research within the broader context of the existing literature.

3.3 Key Terms and Concepts

- Machine Learning (ML): Based on Baloglu, O., Latifi, S.Q. and Nazha, A. (2021), it is a field in the study of artificial intelligence (AI). It is concerned on developing and statistical studying of algorithms capable of learning from data in order to make predictions and/or decisions based on unseen data without requiring giving explicit instructions nor hard coding for those specific tasks.
- Supervised Machine Learning (SML): It is a specific type of machine learning model that is trained exclusively on a labeled datasets, meaning that, each data point within the dataset is always associated with an output label (pairs); in contrast, Unsupervised Mahcine Learning models (UML) are trained with

unlabeled datasets, or in another words there are no associated output labels (Shruthi H, 2022). Therefore, supervised ML models learn how to map the inputs to the correct outputs and make accurate generalizations with unseen data.

- Classification Machine Learning Models: A category within the Supervised Machine Learning models is the Classification ML models, which are used for predicting categorical outcomes. There are 3 possible classification ML models which consist in Binary, Multi-class and Multi-label (article in DataScienceTribe 2023), for this particular research study the proposed supervised ML models fall into the Binary classification, in order to distinguish between diabetic/prediabetic and non-diabetic individuals.
- Interpretability: According to Erasmus, A., Brunet, T.D.P. and Fisher, E. (2021), in the topic of Machine Learning, interpretability consists of the extent to which a human can understand the decisions and/or predictions performed by a trained model. This aspect is critical, especially when it comes to healthcare to ensure transparency, clinical adoption and trust.
- Early detections: This concept consists in the identification of any disease or chronic conditions at an initial stage before significant symptoms appear, which allows for timely intervention and improved outcomes for the patient (Setyati, R. et al, 2024).

3.4 Diabetes and the Need for Early Detection

The disease known as "Diabetes mellitus" is a chronic metabolic disorder characterized by the elevated levels of glucose in blood, resulting from defects in insulin secretion, insulin action or both. The most usual forms include Type 1 diabetes (T1D), Type 2 diabetes (T2D) and gestational diabetes. According to the World Health Organization (WHO, 2024), diabetes is found among the leading causes of death worldwide and often associated with long-term complications such as cardiovascular diseases, kidney failure, blindness and even lower limb amputation.

Type 2 diabetes accounts for about 90% of all cases, often developing gradually and remaining undiagnosed for years due to the fact is largely asymptomatic in nature (especially in early stages). Unnoticed diagnoses constitute a major concern, given that an early intervention to any disease before clinical

symptoms start showing can greatly reduce the risk of future complications and improve the quality of life of the patient, often times involving simple lifestyle changes or medications (American Diabetes Association, 2022).

Traditional screening methods, that are still used to this day, rely on periodic blood testing on glucose levels, HbA1c levels and patient reported symptoms. However, these methods are very much reactive and may miss a patient in its early stage or high-risk individuals, especially in populations with limited access to healthcare and deficient healthcare infrastructure. Moreover, traditional diagnostic processes are very time consuming and may not even leverage the full potential of patient medical data, which should include behavioral, demographic and lifestyle information.

Thus, the early detection is essential for initial timely treatment and also enabling preventative measures and strategies. Studies have shown that individuals diagnosed at early stages are more likely to respond positively to medical interventions, producing a better management of the disease and reducing healthcare costs (Zimmet et al., 2016). However, the ever increasing volume and complexity of health data has outpaced the ability of conventional diagnostic methods to process the aforementioned data efficiently.

Within the context of this research study, the integration of machine learning models in clinical settings offers a promising alternative for the early detection of diabetes (or other diseases). According to a study by Varma & Soni (2020), machine learning algorithms are capable of analyzing large datasets with, often times, a higher accuracy and in a faster manner than traditional statistical methods. Focusing on variables (features) like BMI, age, gender and physical activity among others, ML models were capable of identifying individuals with a high risk of diabetes before clinical symptoms arise and become apparent.

While diabetes as a disease continues to increase, especially in low and middle income countries, there is a clear necessity for a scalable, data-driven diagnostic tools and software solutions of the same nature that become increasingly urgent as time goes on. The early detection not only mitigates the adverse effects of the disease and improves the patient's overall health, but it also alleviates the broader burden put on public health systems. Therefore, research that studies and

advances on ML based solutions for diabetes disease prediction contributes significantly to both clinical and global health priorities.

3.5 Role of Machine Learning in Healthcare

There are several roles Machine Learning models can perform within the healthcare industry, some of them consist on:

- Enhance diagnostic accuracy: ML algorithms excel in analyzing complex data, in this case, medical data such as imaging and electronic health records to identify patterns indicative of diseases. The capabilities of ML models for detecting early signs of diseases like diabetic retinopathy, cardiovascular diseases and various types of cancer before critical symptoms start showing and with an accuracy comparable to traditional methods and even higher than them (Barth, S. and Flam, S, 2025).
- Personalize treatment plans: With the analysis of a patient's medical history and lifestyle factors, a ML model can assist in developing personalized treatment strategies tailored to each patient individually, improving the outcomes and minimizing adverse effects of the patient (Sarkar, K. et al., 2020).
- Predictive analytics for disease prevention: Being capable of processing vast amounts of data to predict the likelihood of disease development on an individual, meaning that these predictive capabilities enable healthcare providers to implement the previous point and perform early interventions (Kelley, 2024).
- Increase operational efficiency and reduce costs: Through the automation and acceleration of processes that traditional methods for diagnostics would occupy more resources and take a longer time, it is possible to reduce the operational costs, minimize human error and allocate resources more efficiently for healthcare institutions and patients alike (Kelley, 2024).
- Advancements in drugs: With the theoretically superior analysis capabilities, machine learning models can potentially also accelerate the discovery of drugs or the improvement of already existing ones by analyzing biological data to identify potential therapeutic targets and predict the efficacy of drug compounds (Kelley, 2024).

3.6 Comparative Studies of ML Models for Diabetes Prediction

As previously stated, the application of ML models in the prediction of diseases has gained a significant attention in recent years. Numerous studies have compared the performance of different ML algorithms on datasets, including those related to diabetes aiming to identify the most accurate and reliable model. These studies usually focus on binary classification tasks, where the objective is to distinguish between diabetic and non-diabetic based on various features related to the patient's health features.

3.6.1 Commonly Used Machine Learning Models

A variety of machine learning algorithms have been explored for diabetes prediction, including but not limited to:

- Logistic regression (LR): Simple yet widely used algorithm for binary classification tasks. It works exceptionally well with linearly separable data and provides interpretable results, making it potentially suitable for clinical settings where transparency is critical (Zhang, 2025).
- Decision tree (DT): A tree based algorithm model that splits the data into numerous branches based on the feature values, leading to a decision at each "leaf" or node, which is the prediction. Its intuitive, visual structure makes it one of the easiest to interpret, which is beneficial in clinical applications, but its prone to overfitting with noisy data (Abedini, 2020).
- Random forest (RF): An ensemble learning method that is composed of multiple decision trees and then aggregates their results. This model often performs well thanks to its ability to handle large datasets gracefully and also because of its robustness to overfitting, making it an appealing choice for diabetes prediction (Zhang, 2025).
- Support vector machines (SVM): Known for its capability to handle complex, high-dimensional data, this algorithm has been shown to perform well in diabetes prediction tasks, particularly when certain types (like SVC) use the kernels to transform non-linearly separable data into higher-dimensional spaces (Zhang, 2025).
- Neural networks (ANN): A more complex model than traditional classifiers since this one is a Deep learning model. Despite its complexity, it has been

applied to diabetes prediction tasks and yielding high performance, but its "black-box" nature poses a great challenge in terms of interpretability and understandabiltiy (Abedini, 2020).

- Ensemble methods (various): Techniques like Gradient Boosting and Voting Classifiers combine the outputs of several models to enhance the prediction accuracy, often outperforming single classifiers by reducing biases and variances (Mushtaq et al., 2025).

3.6.2 Performance Comparison Across Studies

Many academical studies have compared and implemented the models located in subsection 3.6.1 for the diabetes prediction and even other diseases, finding interesting variations in performance based on different aspects like dataset characteristics, feature selection and evaluation metrics. Therefore, the following consists in different key findings from comparative studies:

Zhang (2025) evaluated the models of Logistic regression, SVM, Random forest and XGBoost on the Prima Indians diabetes dataset (dataset initially considered for this thesis). Among the aforementioned models, XGBoost achieved the highest accuracy (85%) and ROC-AUC (91%), having BMI, glucose and age as key features. It is worth noting that the Random forest model also performed well but due to its lower interpretability compared to simpler models (albeit less accurate) like the Logistic regression, the Random Forest approach might limit a seamless clinical implementation.

Li et al. (2021) performed a similar study using the BRFSS 2015 dataset (selected for this study). Their results showed that Neuronal networks provided a superior accuracy performance but they were more prone to overfitting. Meanwhile, SVM models achieved a somewhat healthy balance between accuracy and interpretability.

On the other hand, Husain and Khan (2018) applied several ML models to the NHANES 2013-2014 dataset and developed an ensemble model using a majority voting technique. Utilizing a dataset with 10,172 samples and 54 features, the ensemble model proved to be the overall best predictor in terms of accuracy performance by achieving an ROC-AUC of 0.75. These findings highlight the potential of ensemble learning to enhance diabetes prediction at early stages.

3.6.3 Key Insights from Comparative Studies

- Accuracy vs Interpretability: While more advanced AI models, like DL neural networks or Random forest, are able to achieve high predictive accuracy in a healthcare context, their lack of transparency usually limits a practical use in a clinical setting. According to a study by Tonekaboni et al. (2019), a review and a survey of clinical staff, the authors highlighted that the interpretability of a model played a highly important role in building trust and implementation, where medical staff showed a preference for less complex models like Logistic regression and support for the SVM model due to their more explainable and clear nature for the outputs and results, even if the trade off meant a slight reduction in accuracy. This proves the need for explainable AI techniques that individuals with no technical education on the matter can still understand the results.
- Class imbalance handling: Many studies like Salmi et al. (2024) focus on addressing class imbalance since in medicine and real life scenarios, it is impossible to get balanced data and datasets. Therefore, techniques like SMOTE (Synthetic Minority Over-Sampling) or class weighting can mitigate the effects of any imbalances and avoid biased predictions towards the majority class
- Feature selection: The importance of selecting the most significant relevant features from the dataset is very important, given that unnecessary features will only clutter the model's training and performance.

3.7 Interpretability and Model Transparency in Healthcare

The transparency and interpretability of machine learning models are paramount for fostering trust, ensure accountability and facilitate adoption by healthcare professionals in a medical setting like hospital or institutions. As previously stated, more complex models like deep learning neural networks are more accurate but their inherent complexity hinders their preference by clinical professionals.

A study published by Luo et al. (2024) discussed the different trade-offs between different of machine learning models, specifically in terms of interpretability and accuracy. The aforementioned study supports the claim that, while "white-box" models potentially output a lower accuracy, their transparency and understandability

garner more trust among users despite "black-box" models yielding a better performance in accuracy.

It becomes clear that, while accuracy is an indispensable aspect of ML models to predict the likeness of a disease like diabetes, it is also important to balance it with a clear interpretability to ensure that AI systems are both effective and trustworthy in a healthcare environment.

3.8 Identified Gaps in the Literature

Despite the considerable amount of research dedicated to diabetes and other diseases prediction using machine learning several key gaps remain in the literature review. These gaps highlight areas where current studies may be insufficient and pinpoint where future studies are needed to further enhance the effectiveness of machine learning models in healthcare applications.

First and foremost, there is a clear lack of consensus on the "best" model for Diabetes prediction (or other diseases for that matter). A very recurring challenge is the lack of a definitive answer in this regard. While promising models reange from simpler models like Logistic regression all the way to more complex like a Deep learning models, the variation in results across studies suggest that the most optimal approach may depend on various different factors (dataset, feature selection and evaluation criteria) but the absence of a universal standard consensus makes it almost impossible to adopt a specific machine learning system in the clinical field, not to mention that certain models excel in specific scenarios, but a comprehensive, context-aware framework approach when selecting a model is still lacking a strong foundation to support the claim it is "the best".

Following the previous point, a lot of research between machine learning and the medical field seems to prioritize predictive accuracy over interpretability or usability in real-life scenarios, which are not optional for clinical adoption since doctors need the most reliable but also understandable results. A model can not just perform well, but also be clear and easy to understand in their context and even to patients that most of the time have no real education in the medical field nor in software engineering field. Unfortunately, it seems like a clear trade-off is that high-performing models (ANN or Ensemble methods) suffer from a "black-box" nature that makes them difficult for practitioners to intrepret and trust, meanwhile "white-box"

models (Logistic regression or Decision tree) are easy to interpret but lack the accuracy of their "black-box" counterpoarts. Studies that combine predictive performance and model transparency are limited but essential to close this gap.

Another aspect, particularly inevitable due to the nature of the medical field, is the class imbalance of data and datasets. In real-world healthcare datasets, the distribution of data is most of the time skewed, where, for example, there are more individuals diagnosed with diabetes compared to those who are not and as previously revised, class imbalance can lead to biased predictions where the model overestimates the likelihood of the majority class. There are techniques to mitigate this effect (SMOTE or Class weighting), but they are not capable of consistently apply them across all studies, leading to less reliable results in some cases.

Another significant gap is related to the limited number of reproductible and real-world valid studies. Numerous studies for different ML models have proposed, tested and have come to conclusions by using well known datasets of the likes of Pima Indians diabetes or NHANES, but this comes at the cost of the lack of validation of using more diverse, real-world data. Furthermore, many studies did not provide enough details on the model training, testing or hyperparameter tuning that their models undergo, difficulting the capacity to replicate or validate their results. In the healthcare field, this is crucial not only because it is necessary to ensure that models generalize well outside controlled environments and can be deployed effectively in clinical practice, but also because countless patient's can be negatively affected by incorrect predictions (False Positives and False Negatives) that can even lead to their deaths.

3.9 Summary

This chapter has reviewed the existing related literature on the use of supervised machine learning models for the early detection of diabetes chronic disease, emphasizing their relevance, strengths and limitations in real-world applications in the medical field. This review highlights the growing significance of ML in medical diagnostics, especially in chronic diseases like diabetes where early detection can drastically improve a patient's outcomes. Key concepts such as interpretability, model performance and the importance of an early intervention are defined and discussed.

A variety of commonly used ML models, ranging from interpretable "White-box" models to complex ensemble and neural network models "Black-box" have been explored, along with their comparative performance in recent studies. The consulted researches underscore a recurring situation in which where the more complex the model is, it achieves better metrics but constitutes challenges for clinical implementation due to their inherent complexity.

Several gaps have been identified, like the lack of consensus on the most suitable ML model for diabetes predictions, insufficient attention to model transparency and real-world usability, challenges associated with imbalanced datasets and limited generalization and reproducibility of studies in actual clinical settings. These gaps point to the need for future research that balances predictive power with interpretability and supports model deployment in real-world healthcare environments.

The insights derived from this literature review constitute the foundational justification for the methodology and model selection in the current research study, guiding the comparative evaluation of the different selected ML classifier models for diabetes prediction using a real-world dataset.

Chapter 4: Methodology

4.1 Research Design

This study employs a quantitative, experimental research design in nature to evaluate and compare the effectiveness of several supervised machine learning classification models in predicting diabetes. The methodology of the study is structured around the typical data science lifecycle, encompassing data acquisition, preprocessing, model development, model testing, model evaluation and results interpretation. A comparative approach is adopted in the analysis to determine the relative performance of selected algorithms using a consistent dataset with all of them and a consistent evaluation framework for all (with a set of specified evaluation metrics such as Accuracy, Precision, F1-score among others).

4.2 Dataset and Data Collection

The chosen data set for the present study, comes from the Diabetes Binary Health Indicators BRFSS2015 dataset, which is publicly available on Kaggle and originally prepared by the user Alex Teboul (2021). The aforementioned dataset is actually derived from the Behavioral Risk Factor Surveillance System (BFRSS) made in 2015, an annual health-related telephone survey conducted by the Center for Disease Control and Prevention (CDC). In total, the dataset contains the responses of 253,680 individuals, with 21 predictor variables and a binary target variable (Diabetes_binary). In the description of the dataset, it is stated that the dataset is already pre-cleaned, however, this dataset still has undergone additional preprocessing steps to ensure suitability for ML tasks. The 21 predictor variables and the binary target variable are as follows:

Variables	Description	Values / Encoding
Diabetes_binary (Target)	Diabetes status	0 = No diabetes, 1 = Diabetes or Prediabetes
HighBP	High blood pressure	0 = No, 1 = Yes
HighChol	High cholesterol	0 = No, 1 = Yes
CholCheck	Cholesterol checked in past 5 years	0 = No, 1 = Yes

BMI	Body Mass Index	Continuous numerical value
Smoker	Smoked at least 100 cigarettes in their lifetime (or 5 packs)	0 = No, 1 = Yes
Stroke	Ever had a stroke?	0 = No, 1 = Yes
HeartDiseaseorAttack	History of heart disease or attack (CHD or MI)	0 = No, 1 = Yes
PhysActivity	Physical activity in the past 30 days	0 = No, 1 = Yes
Fruits	Fruit consumption	0 = No, 1 = Yes
Veggies	Vegetable consumption	0 = No, 1 = Yes
HvyAlcoholConsump	Adult men >=14 drinks a week. Adult women>=7 drinks a week	0 = No, 1 = Yes
AnyHealthcare	Has any kind of healthcare coverage	0 = No, 1 = Yes
NoDocbcCost	Could not see a doctor due to cost in the last 12 years	0 = No, 1 = Yes
GenHlth	General health status	1 = Excellent to 5 = Poor
MentHlth	Days where mental health was not good	0–30 days (0 for no days with bad MentHlth and 30 for all days with bad MentHlth)
PhysHlth	Days where physical health was not good	0–30 days (0 for no days with bad PhysHlth and 30 for all days with bad PhysHlth)

DiffWalk	Serious difficulty walking or climbing stairs	0 = No, 1 = Yes
Sex	Gender	0 = Female, 1 = Male
Age	13-level age category (AGEG5YR see codebook)	Ordinal scale 1 to 13
Education	Education level (EDUCA see codebook)	Ordinal scale 1 to 6
Income	Income scale (INCOME2 see codebook)	Ordinal scale 1 to 8.

Table 3. Data set 22 columns

The selected dataset is very appropriate for predictive modelling due to its large size, diversified features, real-world relevance and origin. The survey responses capture key behavioral and demographic factors that contribute to diabetes risk. It is also worth noting that, the dataset is imbalanced, where the majority of responses are labeled as non-diabetic, which constitutes a problem for any ML model and their training because they may be biased towards the majority class (non-diabetic), neglect the minority class (diabetic/prediabetic) and leading to a poor generalization and deficient class separation capacity.

Even thought the dataset is imbalanced, it is important to remember that this dataset comes from actual case, meaning that in real-life scenarios (clinical settings) the provided data to a ML model will most likely always be imbalanced due to the indiosincrasy of the medical field and nature itself, therefore a ML model must be able to tackle this issue efficiently rather than ignore it because it will be unavoidable, specially in developing countries.

According to the CDC (2024), 1 in 5 diabetics and almost 8 in 10 prediabetics are unaware of their condition until clinical symptoms become apparent, therefore, this dataset provides a vital foundation for developing appropriate predictive models, enabling early diagnosis and reducing overall healthcare burden.

4.3 Data Preprocessing

Despite the dataset being "cleaned", it is imperative to still undergo common preprocessing steps to optimize each model performance, given that the dataset consists of an imbalanced type with 21 features and a single binary target variable where all data entries seem to be floating point numbers:

1. Handle possible missing values: Inspect dataset to locate any null/invalid entries and apply imputation (median, mean, mode or regression imputation) or directly drop the said null/invalid entries as required for better results in the model's performance.
2. Remove duplicate values: Simply drop any value that is a duplicate of another given that their presence in the dataset may skew the model's training and lead to overfitting.
3. Encoding of ordinal features: Identify and encode any ordinal categorical variables, therefore, it is possible to ensure that the encoded values still preserve a correct logical order that the model can process. It is worth noting that for this research, this step was never implemented since the selected dataset has its data already encoded in all of its columns.
4. Data splitting: Split data into training and testing sets with stratification to tackle the class imbalance between both sets.
5. Check and handle class imbalance in the training set: Utilize SMOTE to oversmaple the minority class with synthetic samples or, alternatively, utilize Class weighting for models that support it.
6. Feature scaling: On the training set, standardize continuous features and Robust scaling (if necessary) for outliers, but ensure not to scale binary or already encoded categorical features. It is worth noting that some models do not support this step if SMOTE is already implemented.
7. Data shuffling: Shuffle the training data set to ensure it is as random as possible and leave test set as it is to simulate real-world scenarios as closely as possible.

4.4 Machine Learning Algorithms

Four Supervised classification algorithms have been selected for implementation and evaluation for this study, each of them is subject to further

refinement based on an exploratory analysis approach and their own particular idiosyncrasies:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machines (SVM)

Each of the listed models have been chosen for their interpretability, performance records and diversified algorithm approaches. They can be separated into linear, tree-based, ensemble and kernel-based (in some types) respectively.

4.5 Model Training

In order to train and validate each machine learning model and ensure a robust model performance, the following strategies have been followed:

- Train-test split: Utilize between 80/20 and 70/30 split ratio to divide the dataset into training and testing sets.
- Cross-validation: K-fold cross-validation (k=5 or 10) has been utilized to reduce the possibility of overfitting the models and ensure generalization across samples.

4.6 Evaluation Metrics for validation

To assess and evaluate each classification model comprehensively and fairly, the following metrics have been used:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-Score = 2 * (Precision * Recall) / (Precision + Recall)
- ROC-AUC (Area Under the Receiver Operating Characteristic Curve)
- Cross-Validation (Resampling technique for assessing model generalizability)
- Confusion Matrix (to analyze true positives, true negatives, false positives and false negatives).

The listed metrics have provided an accurate and rounded view on each of the model's effectiveness, especially in the context of medical diagnosis and the aforementioned class imbalance the dataset is stated to have and the clinical

context. These metrics also served as a clear comparison point between each of the ML models and their performance.

4.7 Feature Importance and Model Interpretability

To evaluate how the different features in the dataset have contributed to the prediction and facilitated clinical decision making:

- Feature importance scores have been analyzed (for the likes of tree-based models).

4.8 Tools and Technologies

The implementation of has been carried out using the programming language Python on the VSCode code editor (often incorrectly referred to as an IDE) using the following libraries:

- Scikit-learn: For machine Learning implementation and evaluation.
- Pandas and NumPy: For data manipulation and preprocessing from the dataset and analysis.
- Matplotlib and Seaborn: For data visualization (plotting and graphing) and results presentation in a comprehensible matter.
- Imblearn: For handling imbalanced datasets in machine learning, offering techniques like resampling and over/undersampling to improve model performance.
- Logging: For tracking and saving events, particularly the model's results when the script is run.
- Time: For returning the current time in seconds since the beginning of an epoch, to measure the time interval it took to train each model.
- Pathlib: For handling files and paths on the operating system, simply to facilitate the accessing of the CSV file of the dataset when scripts of the ML models are run.

4.9 Ethical Considerations

The present research study uses a publicly available data that contains no personal information that can lead to identifying any of the subjects that participated in the BFRSS dataset by the CDC. Therefore, the usage of this dataset poses

minimal ethical risk, however, ethical diligence is still maintained in the following ways:

- ML model fairness: The performance of each Supervised classification model across different demographic groups (for example, sex or age) will be analyzed to detect and minimize biases.
- Responsible use of AI technologies: Each insight found within this study was contextualized within the clinical realities and contexts, supporting the role of machine learning models as a tool to help in the diagnosis of diabetes disease and not a replacement of the expert judgement of the human professionals in the field.

4.10 Summary

The present chapter outlines the methodology that was adopted during this research study, from data acquisition to the ML model evaluation. By applying and comparing multiple Supervised machine learning classification models on a large scale health dataset, aiming to identify optimal strategies to detect diabetes and/or prediabetes in its early stages. There is a special emphasis on the interpretability, performance and clinical effectiveness of the models to ensure practical contributions to the healthcare domains.

Chapter 5: Results and Discussion

5.1 Hardware and Software specifications

The performance and efficiency of the training upon a ML model can be greatly increased by the hardware and software environment in which the script is executed and the research conducted. Topics like processing speed, available memory and system architecture can heavily influence how quickly and effectively a model can be trained (Scikit-learn Developers, 2024), especially on models that are very intensive like SVMs. During this research, all experiments were conducted on a machine with a system running Windows 11 OS and the following hardware specifications:

- CPU: Intel(R) Core(TM) i7-1255U 12th Gen processor (3.52 GHz)
- RAM: 16GB from 2 SK Hynix DDR4-3200 (8GB each)
- GPU: Intel(R) Iris(R) Xe integrated graphics
- Storage: 1TB SSD

Even though the system provides a somewhat adequate performance for datasets of moderate size, it is important to take into account that certain, more complex models and/or larger datasets might require an upgrade from these specifications because if not, it would result in longer training times or convergence issues due to lack of GPU acceleration or not even be able to meet the computational demands and overrunning the available RAM. Consequently, for this and future studies, these specifications must be considered when analyzing the results or attempts to replicate this work on different machines or environments.

5.2 Full script workflows

The following figure constitutes the general workflow that each script follows:

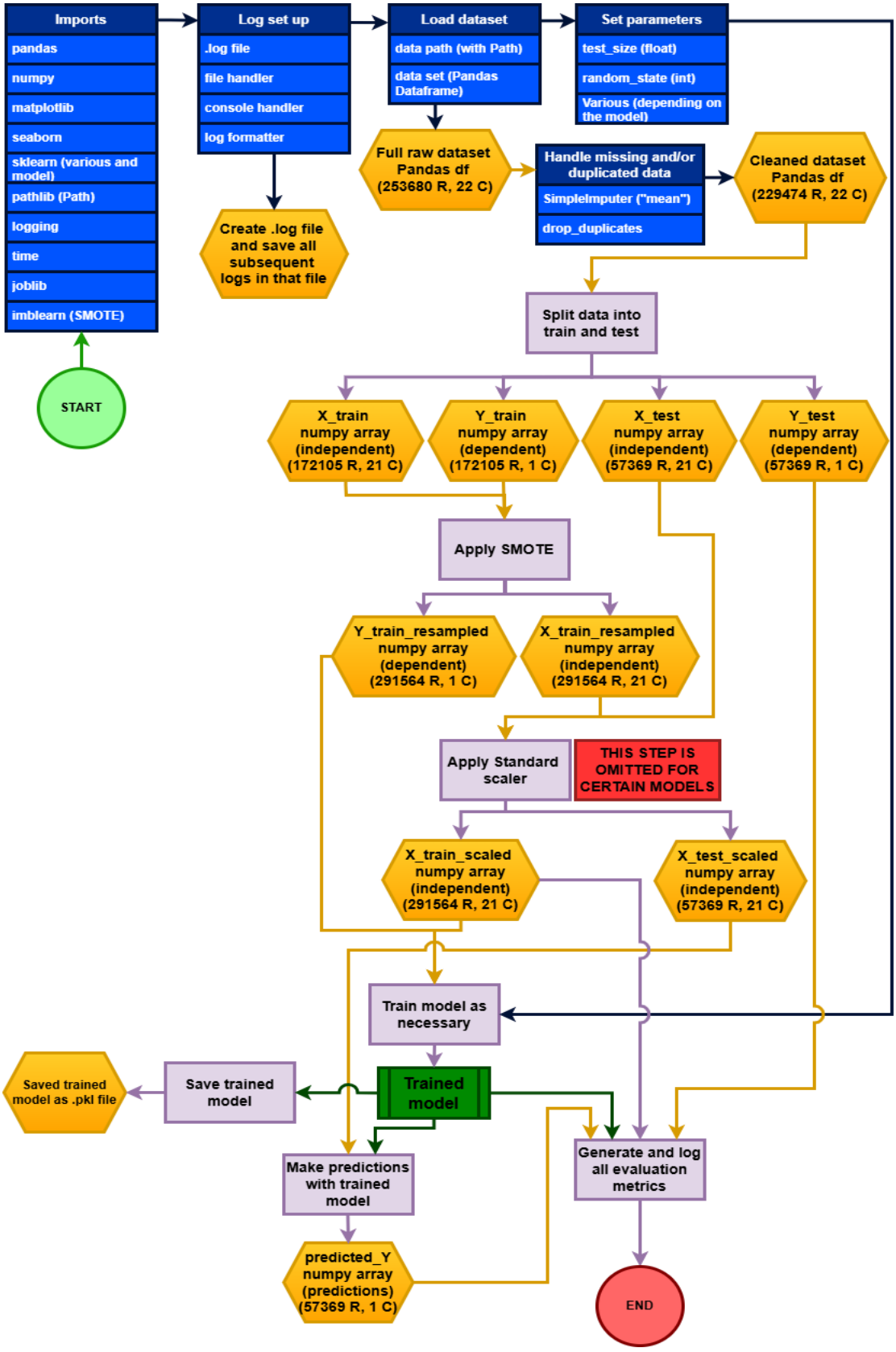

Figure 2. General workflow of ML model scripts

The 1st cell holds all the necessary Python libraries that are essential for the whole script. Libraries for data manipulation (pandas and numpy), visualization (matplotlib and seaborn), machine learning (sklearn and imblearn) and logging (logging, time, joblib and pathlib).

The subsequent 2nd cell sets up the logging configuration, generates the .log file and facilitates debugging and result tracking with a robust logging system is established. This includes a log file that captures execution details, console handler for real-time monitoring and structured formatter to maintain log consistency.

The 3rd cell loads the dataset as a Pandas dataframe. A dataframe comprised of 253,680 rows and 22 columns.

The 4th cell sets critical parameters for the pipeline, like test_size (test and train split ratio), random_state (Consistent results across runs) and model specific parameters to each model.

The 5th cell handles possible missing values with SimpleImputer with a mean strategy and removes possible duplicates to ensure unique and accurate training samples. This results in a cleaned dataset with 229,474 rows and 22 columns.

The 6th cell splits the data into Features (X) and target label (Y) and immediately also splits them into 4 numpy arrays:

- X_train and Y_train for training.
- X_test and Y_test for evaluation.

The 7th cell handles class imbalance by using SMOTE (Synthetic Minority Oversampling Technique) only on the training sets (X_train and Y_train). This results in 2 numpy arrays of 291,564 entries (X_train_resampled and Y_train_resampled)/

In the same 7th cell, standardization is applied by a standard scaler on X_train_resampled and X_test, which transforms the input features to a mean of 0 and unit variance. This results in 2 numpy arrays (X_train_scaled and X_test_scaled), however, an important note is that this step is only used with Logistic Regression and SVM (LinearSVC) given that it would not help Decision Tree nor Random Forest, in which cases they would use X_train_resampled and X_test given that X_train_scaled and X_test_scaled would not exist.

Then, the 8th cell is where the model is finally trained using the preprocessed training data. The modular structure of the Jupyter Notebook allows for experimentation of the different algorithms.

In the 9th cell is where the trained model is saved as a .pkl file in case the trained model is to be used or implemented in future softwares or for testing in a different study.

The 10th cell is where the already trained model is used to predict diabetes outcomes on the testing set, which yields the results (predicted_Y) that will be evaluated in the next cell.

The 11th and last cell is where the multiple evaluation metrics for the performance of each model, these metrics include:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC
- Classification Report
- K-fold Cross validation
- Confusion Matrix

These metrics are logged into the .log file to then be analyzed to ensure that the model performs effectively and with its best possible iteration, since we are interested in comparing them and determining which is the best model for the prediction of diabetes disease.

5.3 Data set analysis EDA (Exploratory Data Analysis)

The following plots, figures and tables represent the analysis of the raw data provided by the Control and Prevention (CDC) and their Behavrioal Risk Factor Surveillance System (BFRSS) made in 2015, which was prepared, cleaned to some extent and published by the user Alex Teboul (2021) in the Kaggle data science platform:

- **Data set overview**

Number of columns	Number of rows
22 (21 features and 1 objective variable)	253680 (total number of entries)

Table 4. Number of columns and rows of raw data set

- **Target variable distribution**

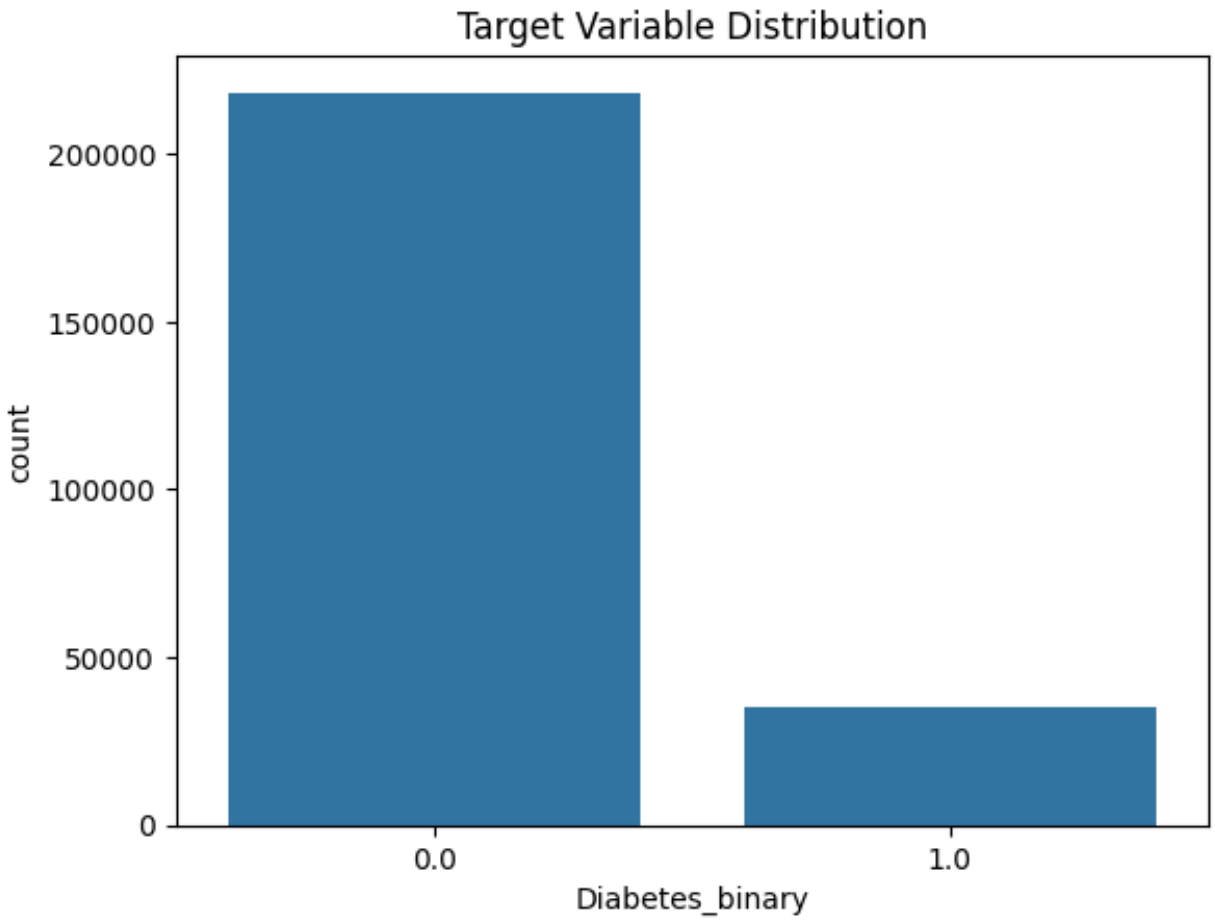

Figure 3. Target variable distribution

As seen in this plot, it is possible to visualize a clear class imbalance in which the majority classification is 0 or no diabetes (218334 out of 253680), whereas 1 or diabetes/prediabetes represents a clear minority in the dataset (35346 out of 253680). This ratio constitutes ≈86.07 of no diabetes, while ≈13.93% of diabetes and/or prediabetes.

- **Histograms of all independent variables**

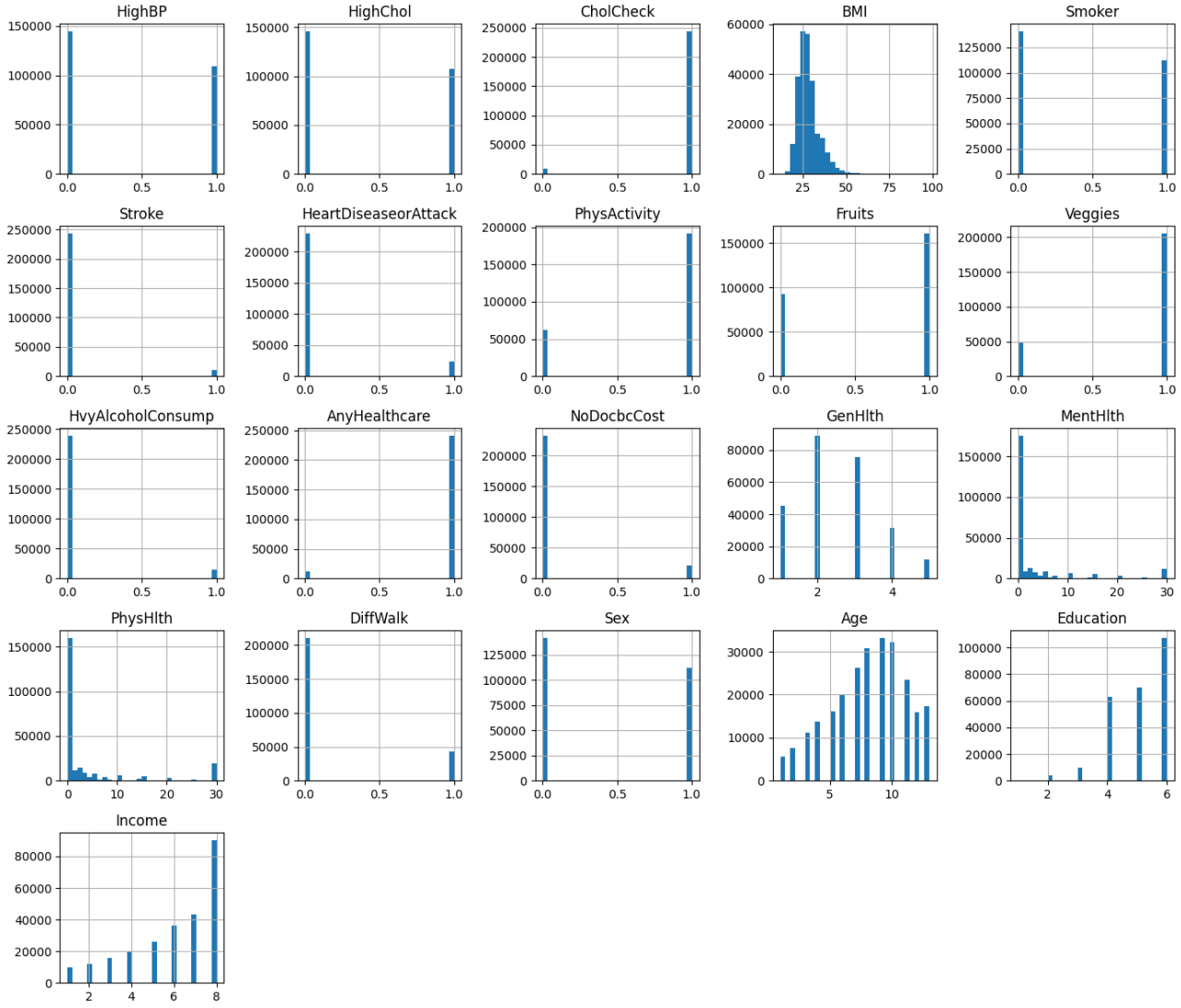

Figure 4. Histogram of all independent variables

There are several key takeaways from this series of histograms. Firstly, there is a clear imbalance for binary health conditions like stroke or heart diseases where the majority of responders claimed to not have these conditions (0 or no being the dominant values). Then, there is a high proportion of individuals that have undergone cholesterol checks and engaged in physical activities, which suggests a baseline level of health awareness and preventive behavior among the population. When it comes to self-reported health metrics, such ass mental and physical health, shows that while most reponders report low frequencies, there is a long tail of individuals that experience extended periods of health challenges. Lastly, on a social level, the sociodemographic distribution like education, income and age are skewed toward higher levels, indicating that the selected dataset may overrepresent older individuals

with a higher socio-economical status. These histograms showcase the relationship between independent variables and the target variable.

- **Summary statistics**

	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker
Count	253680					
Mean	0.139333	0.429001	0.424121	0.962670	28.382364	0.443169
STD	0.346294	0.494934	0.494210	0.189571	6.608694	0.496761
Min	0	0	0	0	12	0
25%	0	0	0	1	24	0
50%	0	0	0	1	27	0
75%	0	1	1	1	31	1
Max	1	1	1	1	98	1

Table 5. Summary statistics part 1

	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggie	DiffWalk
Count	253680					
Mean	0.040571	0.094186	0.756544	0.634256	0.811420	0.168224
STD	0.197294	0.292087	0.429169	0.481639	0.391175	0.374066
Min	0	0	0	0	0	0
25%	0	0	1	0	1	0
50%	0	0	1	1	1	0
75%	0	0	1	1	1	0
Max	1	1	1	1	1	1

Table 6. Summary of statistics part 2

	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth
Count	253680				
Mean	0.056197	0.951053	0.084177	2.511392	3.184772
STD	0.230302	0.215759	0.277654	1.068477	7.412847
Min	0	0	0	1	0
25%	0	1	0	2	0
50%	0	1	0	2	0
75%	0	1	0	3	2

Max	1	1	1	5	30

Table 7. Summary of statistics part 3

	Sex	Age	Education	Income
Count	253680			
Mean	0.440342	8.032119	5.050434	6.053875
STD	0.496429	3.054220	0.985774	2.071148
Min	0	1	1	1
25%	0	6	4	5
50%	0	8	5	7
75%	1	10	6	8
Max	1	13	6	8

Table 8. Summary of statistics part 4

These summaries provide a clearer insight into the distribution of each column and to which class they fall into more often, supporting the previous histograms plots and the key takeaways.

- **Missing values and duplicates**

No missing values were found, however, there were 24206 duplicated rows found that must be dealt with to avoid artificially inflating the importance of certain patterns or features.

- **Independent variables correlation with dependent variable**

Feature	Correlation with Diabetes_binary
GenHlth	0.293569
HighBP	0.263129
DiffWalk	0.218344
BMI	0.216843
HighChol	0.200276
Age	0.177442
HeartDiseaseorAttack	0.177282

PhysHlth	0.171337
Stroke	0.105816
MentHlth	0.069315
CholCheck	0.064761
Smoker	0.060789
NoDocbcCost	0.031433
Sex	0.031430
AnyHealthcare	0.016255
Fruits	-0.040779
Veggies	-0.056584
HvyAlcoholConsump	-0.057056
PhysActivity	-0.118133
Education	-0.124456
Income	-0.163919

Table 9. Independent variables correlation to dependent variable

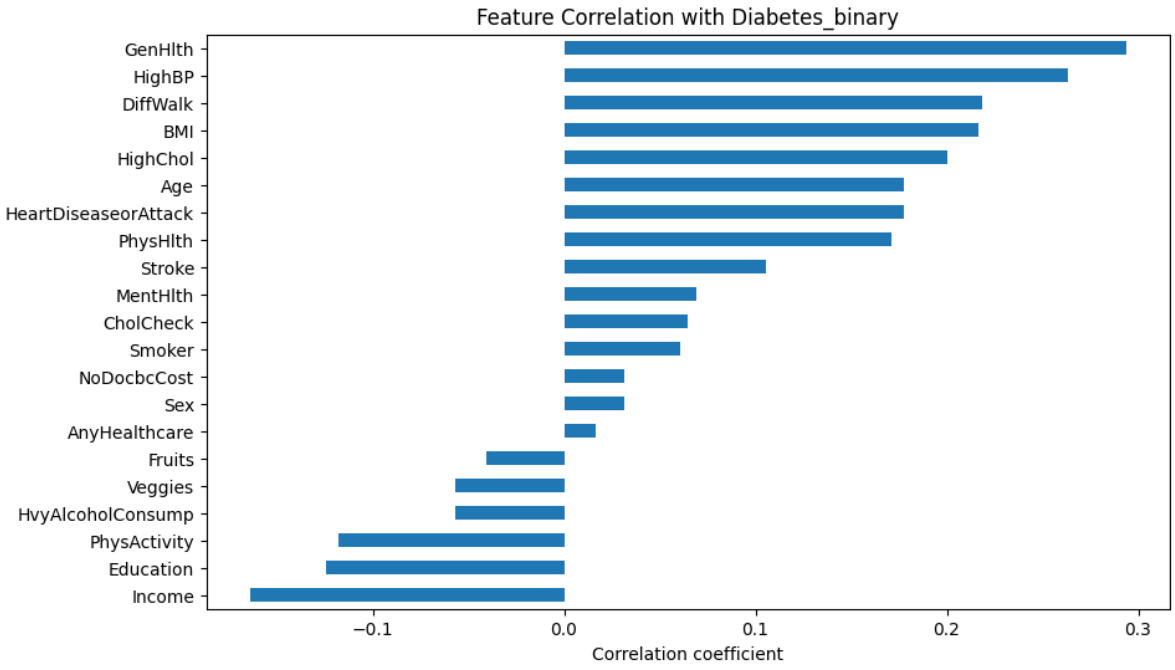

Figure 5. Independent variables correlation to Diabetes_binary

To identify the factors the most associated to diabetes disease in the selected dataset, Pearson correlation coefficients have been computed between all numerical and binary features of the dataset and the target variable Diabetes_binary.

The correlation between independent variables and the objective variable oscillates between -1 and 1, where -1 constitutes to the least correlated feature and 1 the most correlated feature. In this context, we can see that the General health is the most correlated feature to Diabetes_binary, while Income is the least correlated feature. However, it is important to note that these are correlations and not causal relationships, while they might align to existing medical knowledge, further modelling and/or controlled studies are required to establish a true causation.

- **Correlation matrix heatmap, all features**

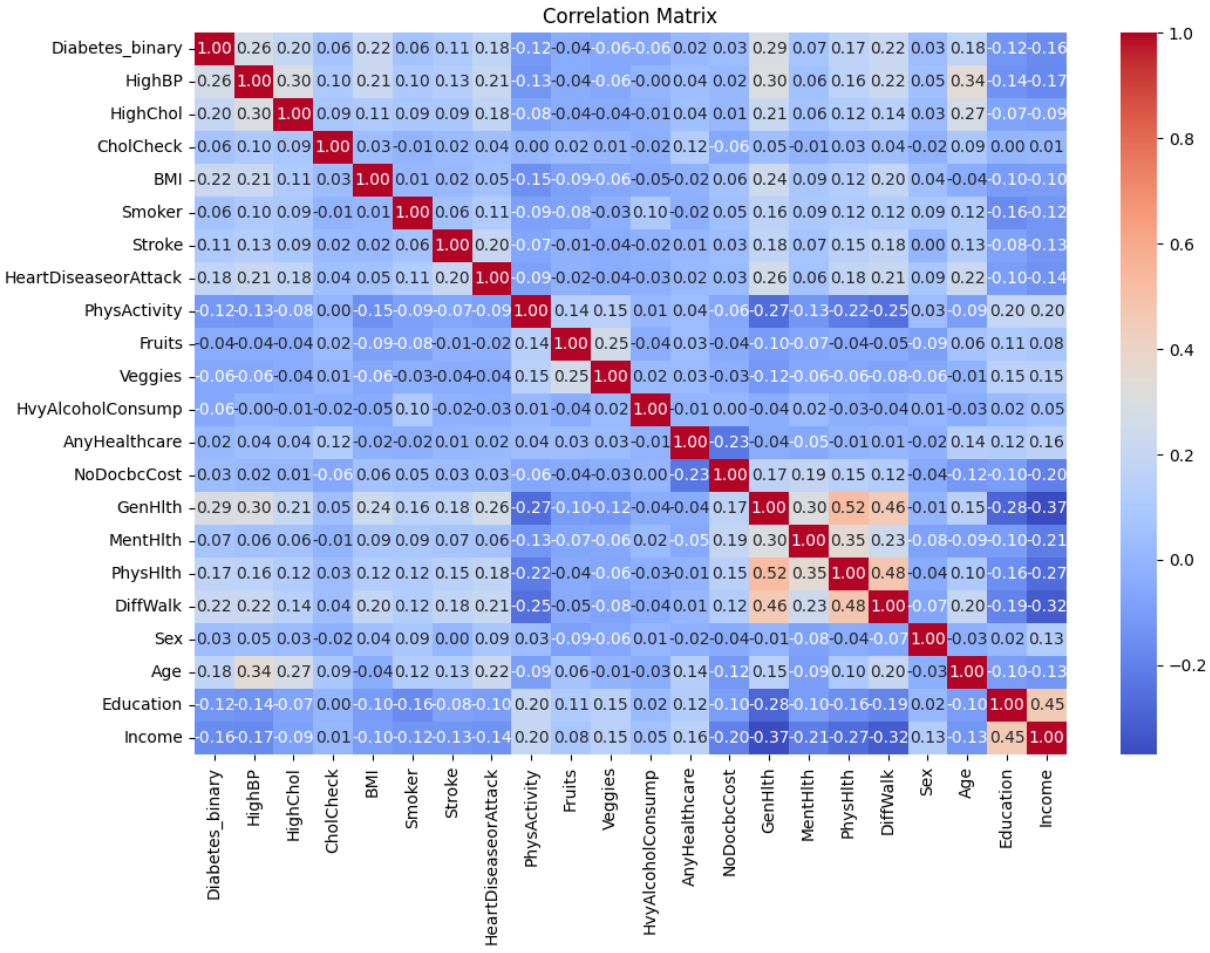

Figure 6. Correlation matrix between all columns of dataset

The present correlation matrix illustrates how each individual independent feature relates to diabetes and/or prediabetes presence. It is notable that, poor general health (GenHlth), high blood pressure (HighBP) and mobility issues

(DiffWalk) show the strongest positive correlations to diabetes or prediabetes that can confirm widely accepted medical risk factors. On a different note, features like higher income, better education and increased physical activity are weakly negatively correlated, suggesting that the aforementioned may be protective factors. Once again, these correlations are not causal and should be interpreted in the context of a broader multi-variable analysis.

- **Correlation matrix heatmap, top 8 features**

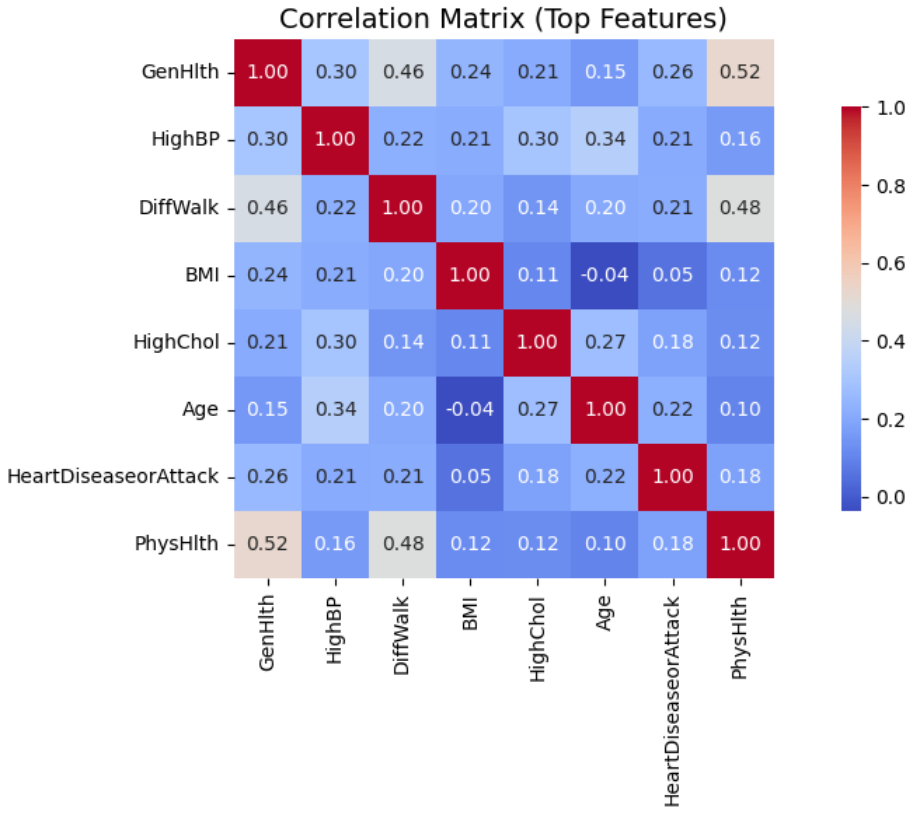

Figure 7. Correlation matrix between top 8 features of dataset

Based on the figure 6 and table 9, this correlation matrix contains only the top 8 features related to Diabetes_binary. Specifically, General health (GenHlth), High blood pressure (HighBP) and Difficult walking (DiffWalk) show the highest positive correlation among all 21 features, aligning to well-established medical risk factors for diabetes, not to mention that limited physical health (PhysHlth) has a moderate positive correlation that reinforces the relationship between physical well-being and diabetes risk.

Features like BMI (Body Mass Index), High cholesterol (Highchol) and age exhibit weaker correlations but still meaningful insights for the health status of each

individual patient, notably heart disease or heart attack history (HeartDiseaseorAttack) maintains a mild correlation, which could imply comorbidity.

Just like the previous correlation heatmap, these relationships do not constitute any causation, only correlation, they should always be interpreted within a broader analytical framework.

- **Countplot for categorical and binary features (Correlation, not causal)**

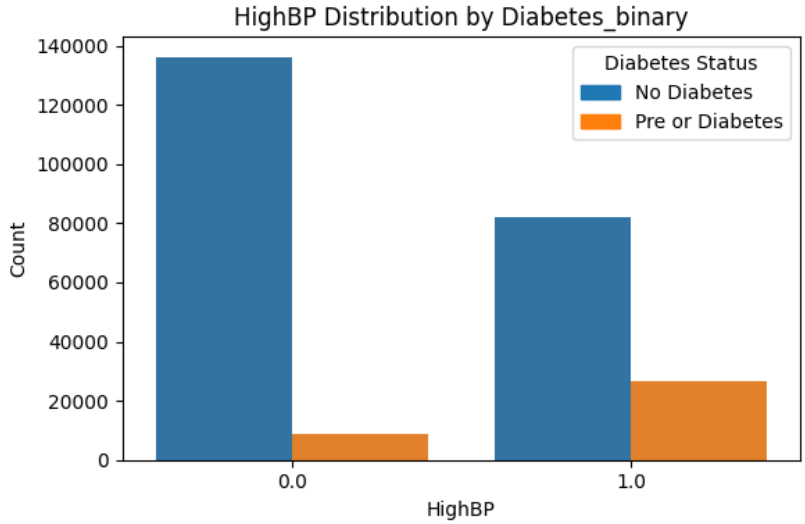

Figure 8. High blood pressure distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often suffer from high blood pressure, whereas people who do not suffer from diabetes often times do not suffer from high blood pressure either.

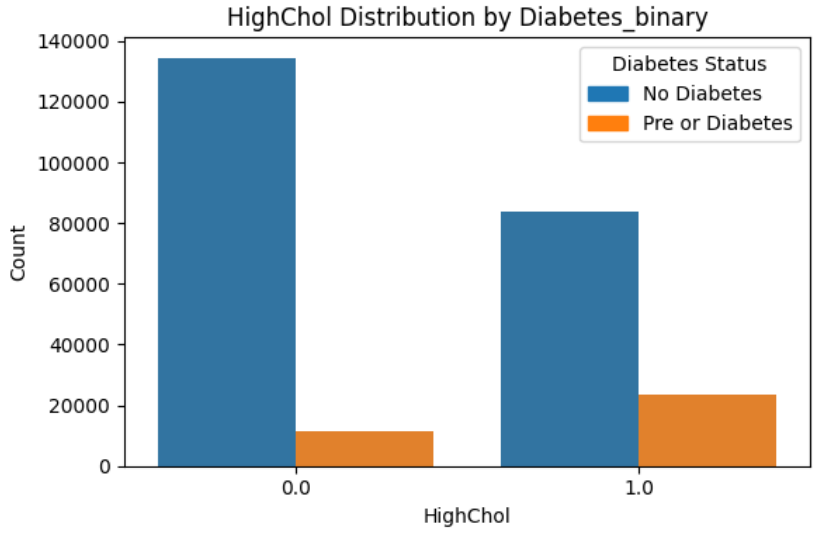

Figure 9. High cholesterol distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often suffer from high cholesterol, whereas people who do not suffer from diabetes often times do not suffer from high cholesterol either.

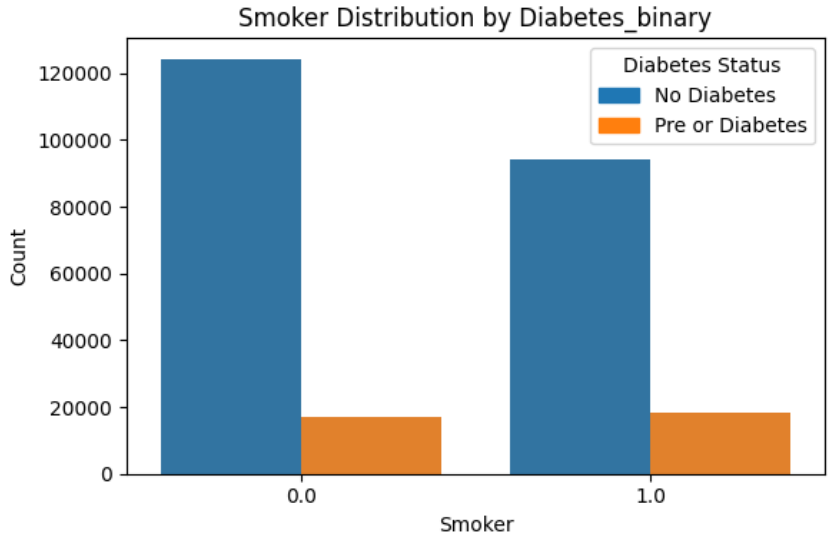

Figure 10. Smoker distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes (slightly) more often are smokers, whereas people who do not suffer from diabetes often times are not smokers, although in this case, the distribution is more even than with cholesterol and blood pressure.

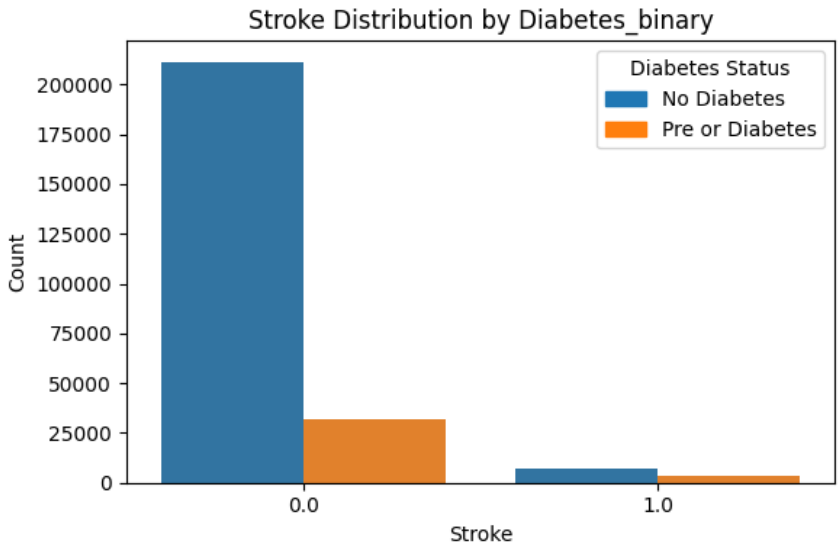

Figure 11. Stroke distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often have not suffered a stroke at least once in their lives, also, people who do not suffer from diabetes have not suffered from a stroke at least once in their lives either.

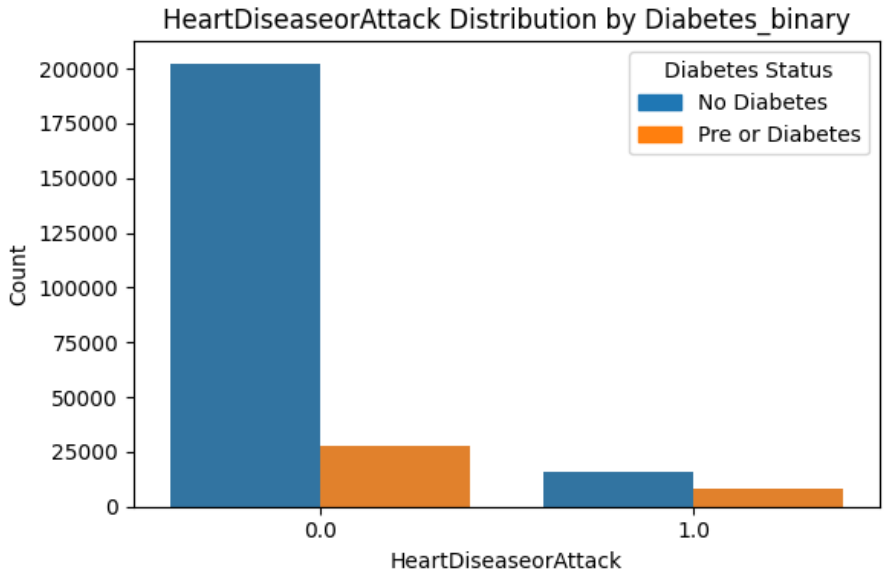

Figure 12. Heart disease or attack distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often have not suffered a heart attack nor do they possess a heart disease, also, people who do not suffer from diabetes have not suffered from a heart attack nor do they possess a heart disease either.

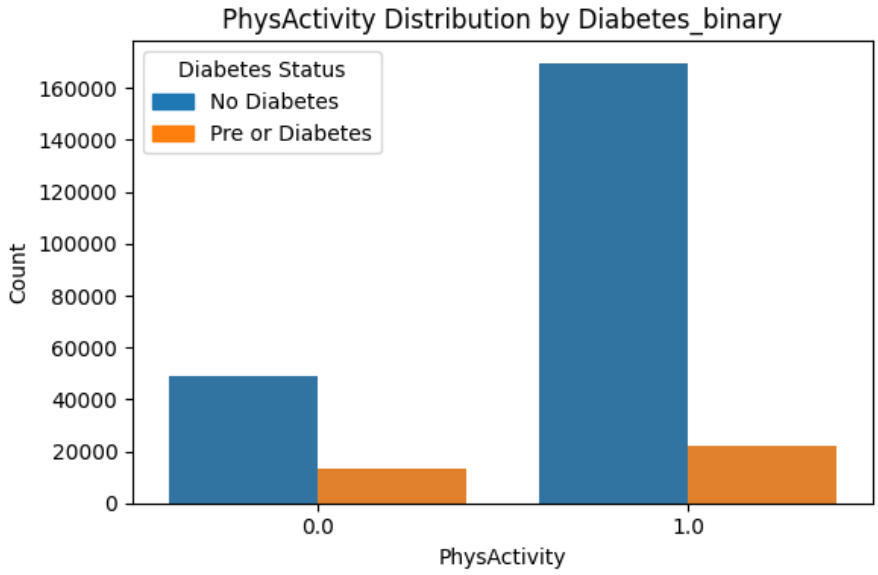

Figure 13. Physical activity distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often practice some sort of physical activity, also, people who do not suffer from diabetes practice some sort of physical activity as well.

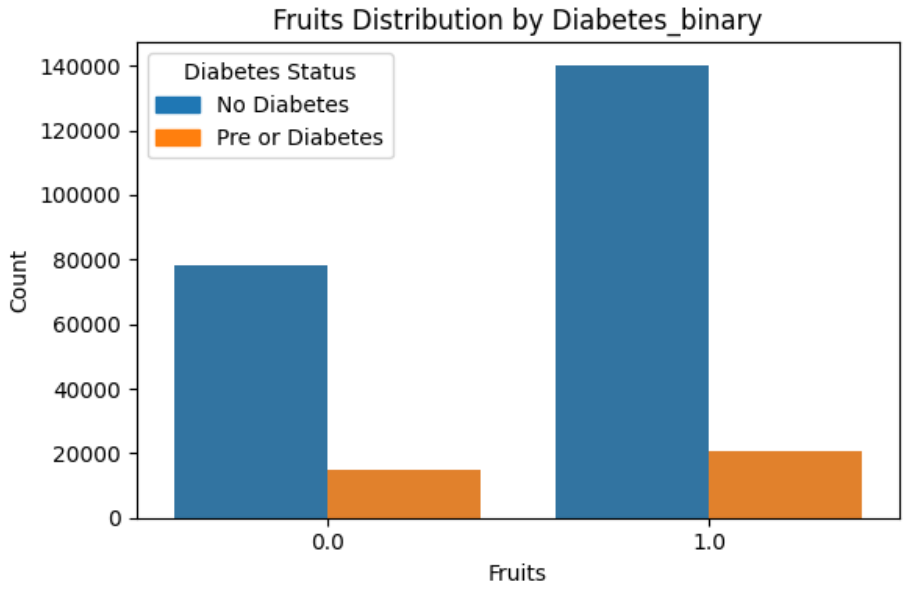

Figure 14. Fruit consumption distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often also consume fruits on a regular basis, also, people who do not suffer from diabetes consume fruits on a regular basis as well.

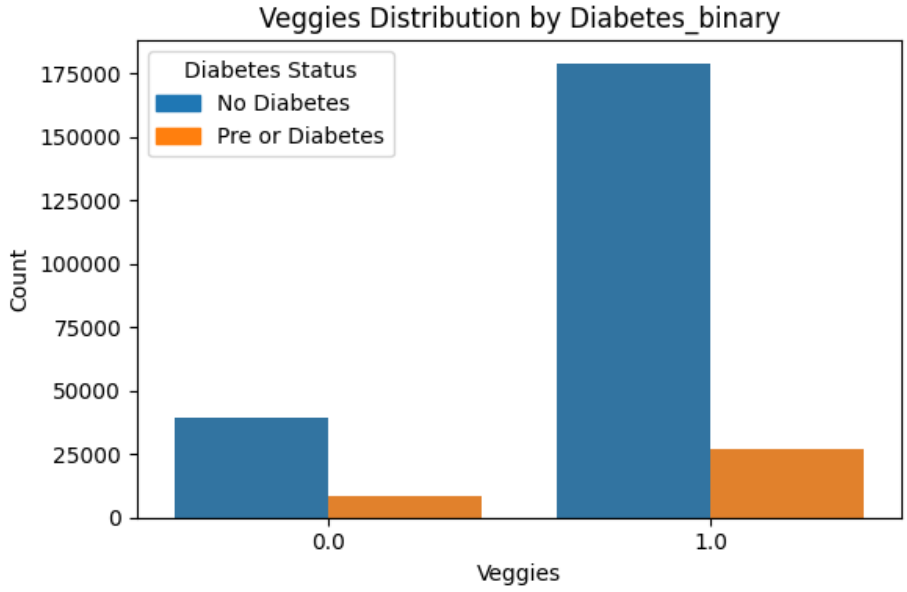

Figure 15. Vegetable consumption by Diabetes presence

This plot implies that people with diabetes or prediabetes more often also consume vegetables on a regular basis, also, people who do not suffer from diabetes consume vegetables on a regular basis. In this case, the implications have a higher disparity compared to fruit consumption as well.

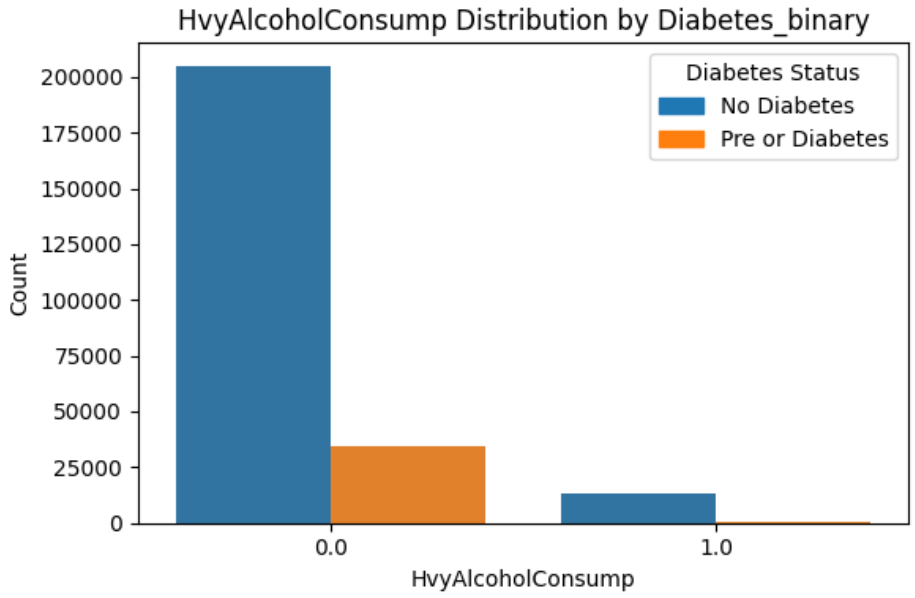

Figure 16. Heavy alcohol consumption distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often also do not heavily consume alcohol, also, people who do not suffer from diabetes do not heavily consume alcohol either.

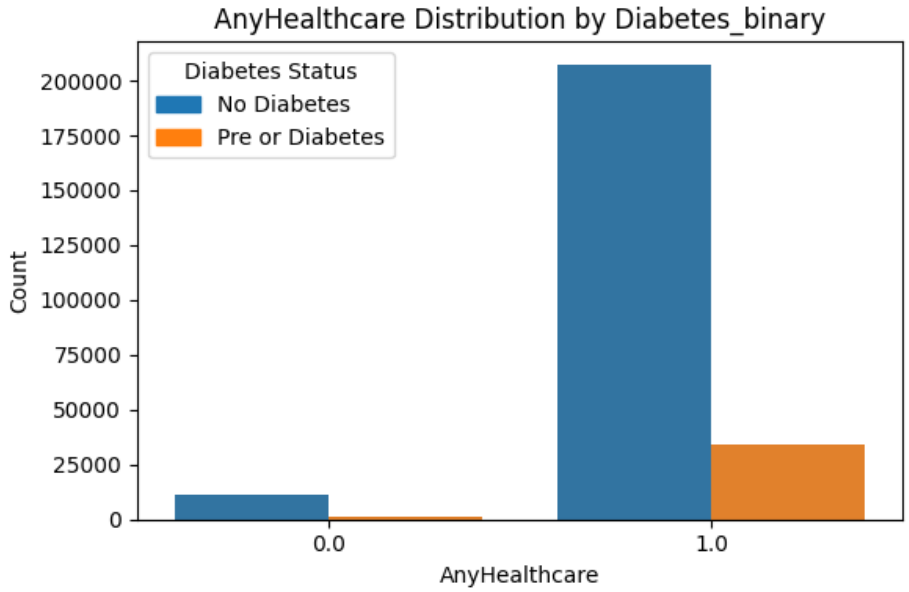

Figure 17. Any healthcare distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often also have some sort of healthcare in their lives, also, people who do not suffer from diabetes still have some sort of healthcare in their lives as well.

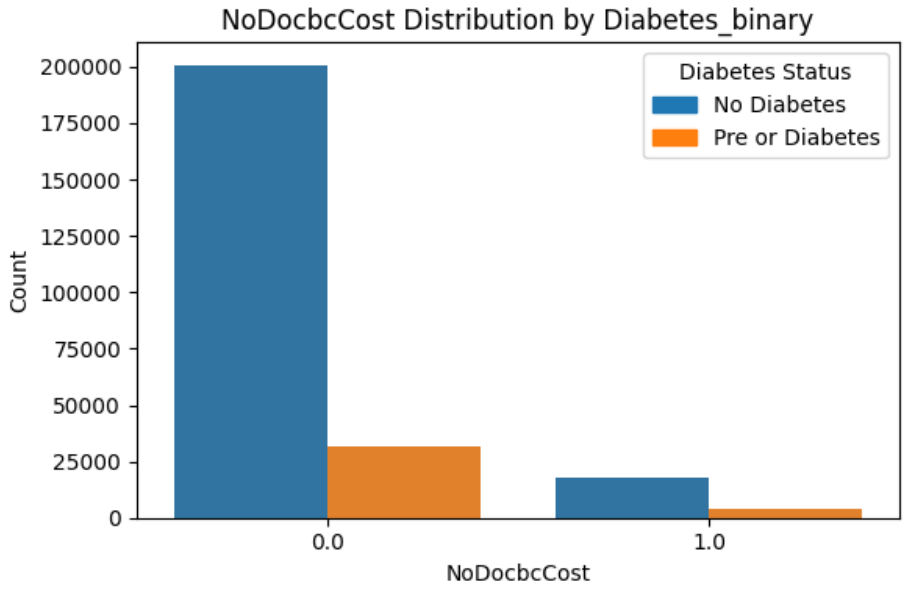

Figure 18. No visit to the Doctor due to cost distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes were more often unable to go for a medical visit due to its cost, also, people who do not suffer from diabetes were unable to go for a medical visit due to its cost as well.

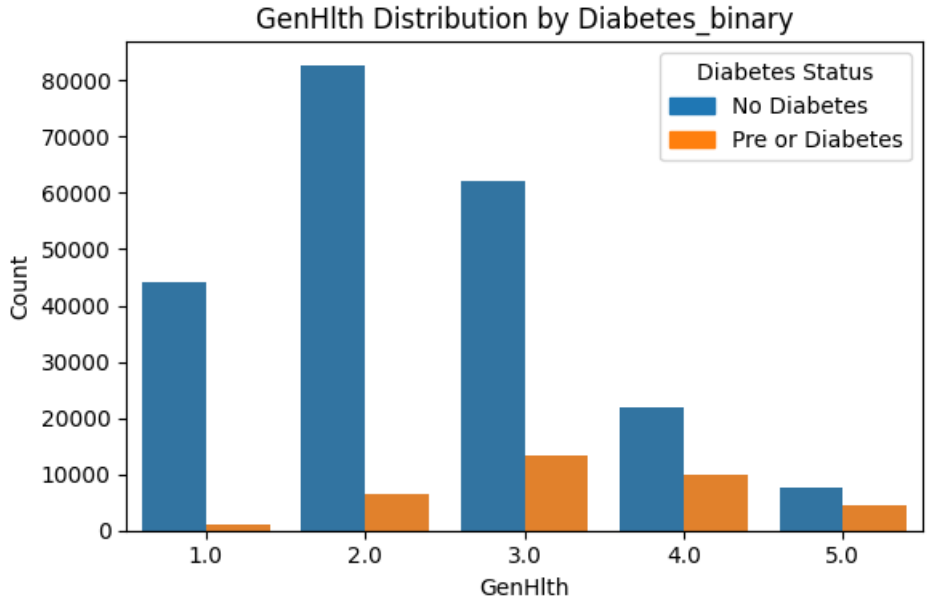

Figure 19. General health distribution by Diabetes presence

This plot ranges from 1=Excellent general health to 5=Poor health. It implies that people who suffer the most from diabetes or prediabetes have a poorer general health, also, people who do not suffer the most from diabetes have a general excellent general health.

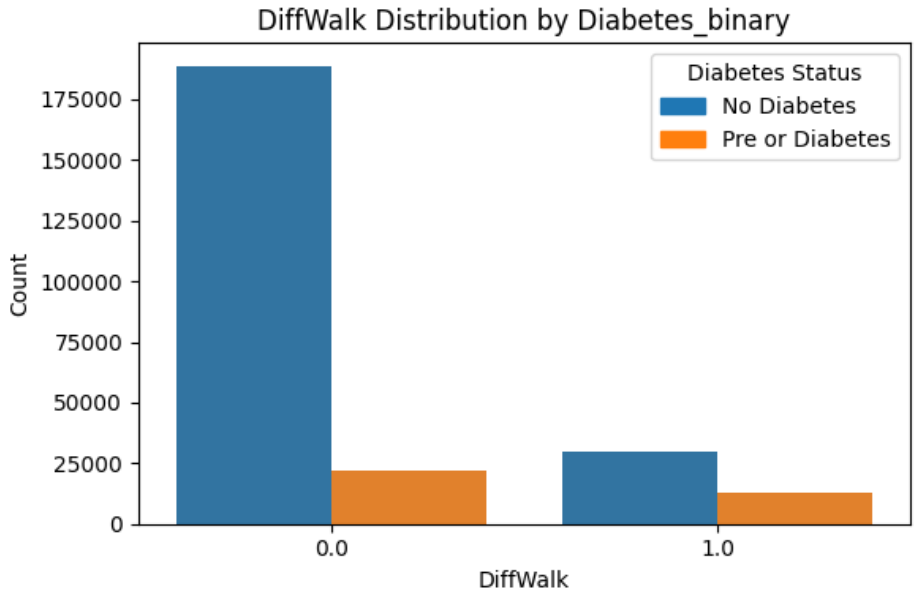

Figure 20. Difficulty walking distribution by Diabetes presence

This plot implies that people with diabetes or prediabetes more often do not have difficulties walking or going up stairs, also, people who do not suffer from diabetes do not have difficulties walking or going up stairs as well.

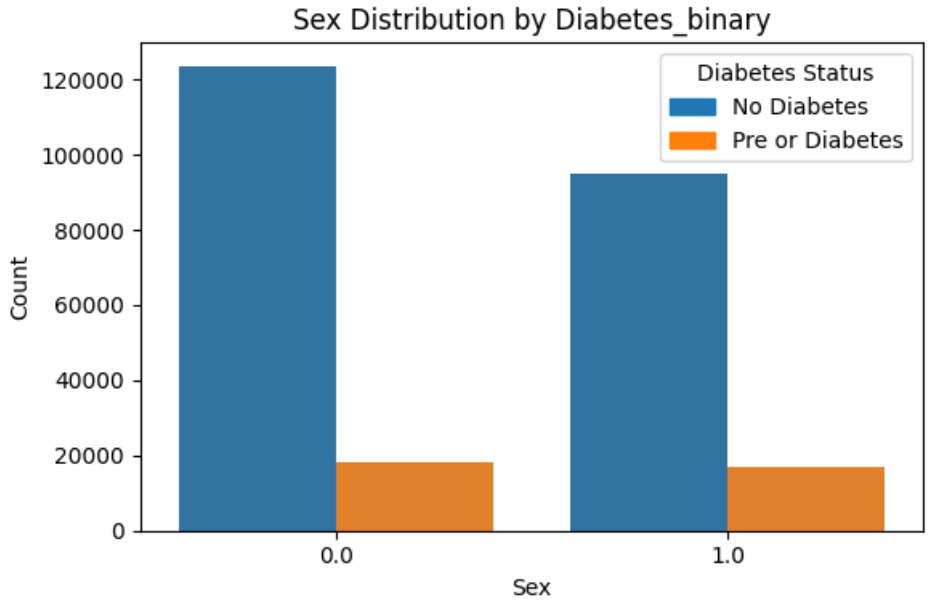

Figure 21. Sex distribution by Diabetes presence

This plot maps 0 = Female and 1 = Male. It implies that people who suffer from diabetes or prediabetes are more often Females, also, people who do not suffer from diabetes are Females as well. Although when it comes to suffering from diabetes or prediabetes, the distribution is very even in comparison to the non-diabetes distribution.

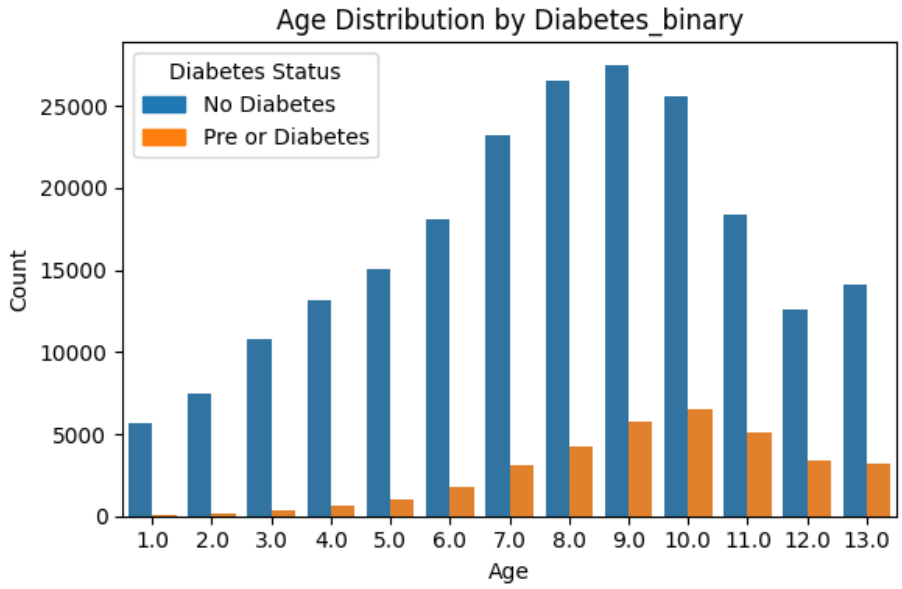

Figure 22. Age distribution by Diabetes presence

This plot has 13 levels of age ranges in which 1 = 18–24y, 2 = 25–29y, 3 = 30–34y, 4 = 35–39y, 5 = 40–44y, 6 = 45–49y, 7 = 50–54y, 8 = 55–59y, 9 = 60–64y, 10 = 65–69y, 11 = 70–74y, 12 = 75–79y and 13 = 80 or older. It implies that people who suffer from diabetes or prediabetes are more often between 55 to 70 years old, also, people who do not suffer from diabetes are between 45 and 70 years old.

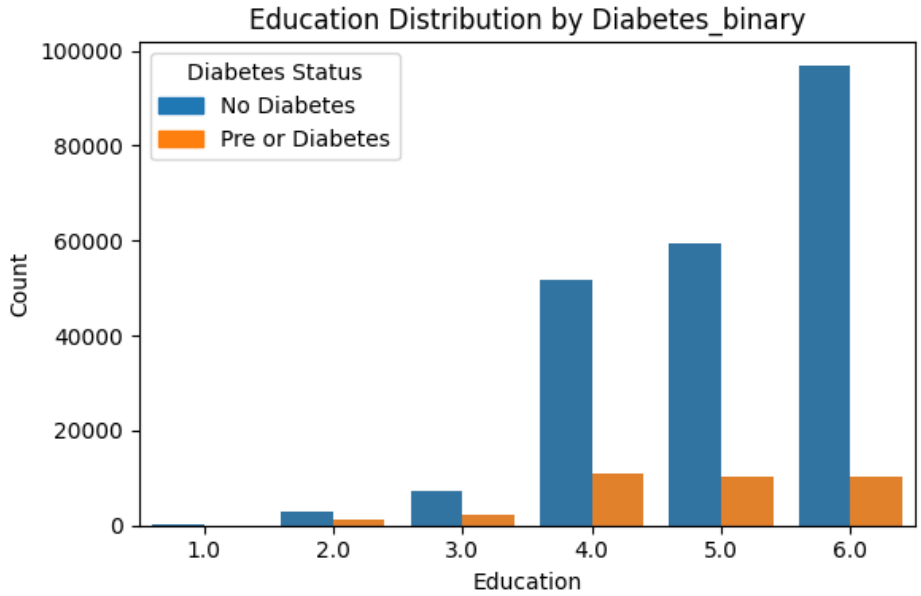

Figure 23. Education distribution by Diabetes presence

This plot has 6 levels of education and they represent 1 = Never attended, 2 = Grades 1 to 8, 3 = Grades 9 to 11, 4 = Grade 12 or GED, 5 = College or Technical school (no degree) and 6 = College graduate (4-year degree or higher). It implies that

people who suffer from diabetes or prediabetes also have at least graduated from High School, also, people who do not suffer from diabetes have at least graduated from High School as well.

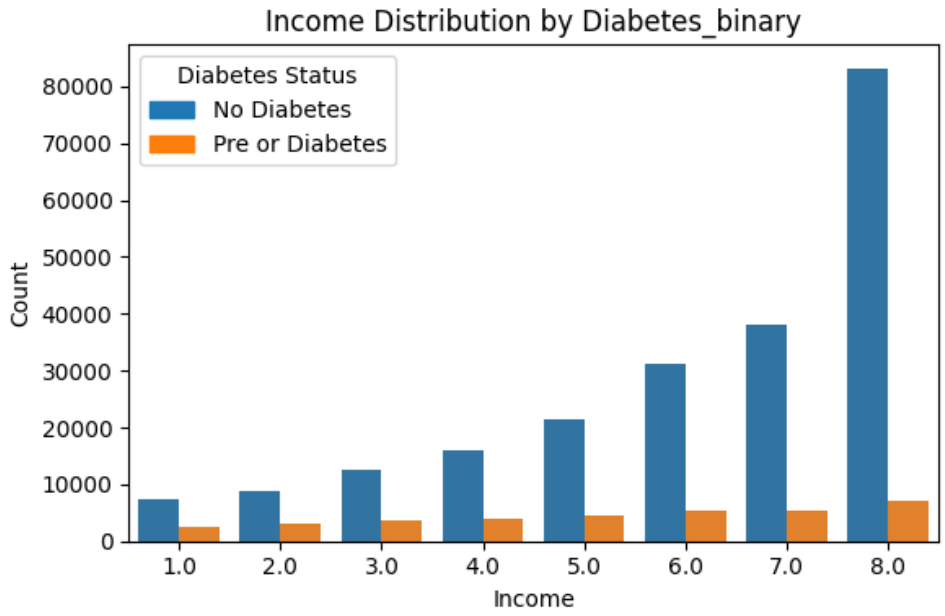

Figure 24. Income distribution by Diabetes presence

This plot has 8 levels of educations and they represent 1 = less than 10k $, 2 = 10k to 15k $, 3 = 15k to 20k $, 4 = 20k to 25k $, 5 = 25k to 35k $, 6 = 35k to 50k $, 7 = 50k to 75k $ and 8 = 75k $ or more. It implies that people who suffer from diabetes or prediabetes more often have a higher income, also, people who do not suffer from diabetes have a higher income as well, in a more notorious way.

- **KDE for continuous features (Correlation, not causal)**

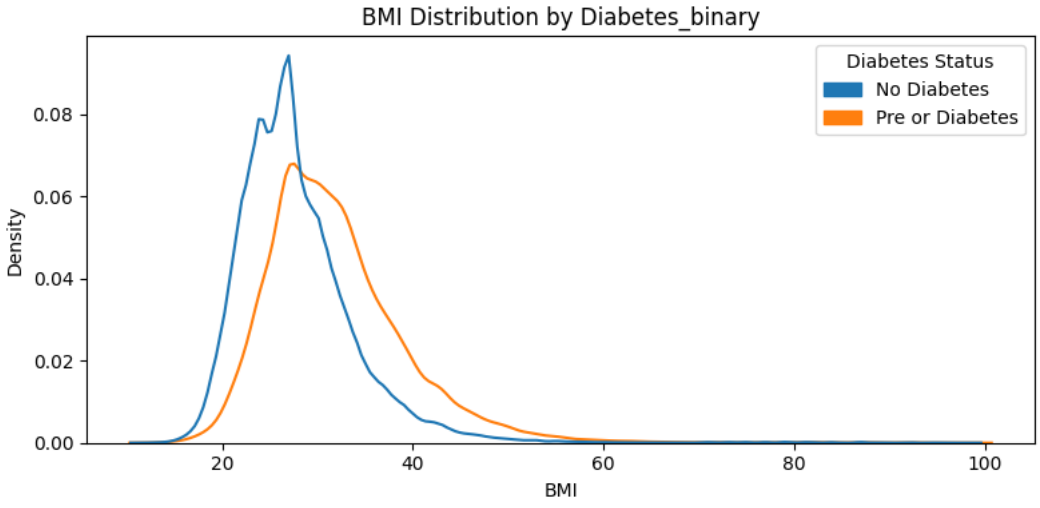

Figure 25. BMI distribution by Diabetes presence

In this plot, the values usually oscillate between 10 and 60, derived from the patient's weight in kg divided by the patient's height in meters to the power of 2, where the lower the value, the more underweight and the bigger the value, the more overweight. For reference, anything below 16 is considered severely underweight (malnourished) and anything equal to or above 40 is considered morbidly obese. It implies that the people who suffer the most from diabetes or prediabetes range between a BMI of 20 and 40, the same when it comes to people who do not suffer from diabetes.

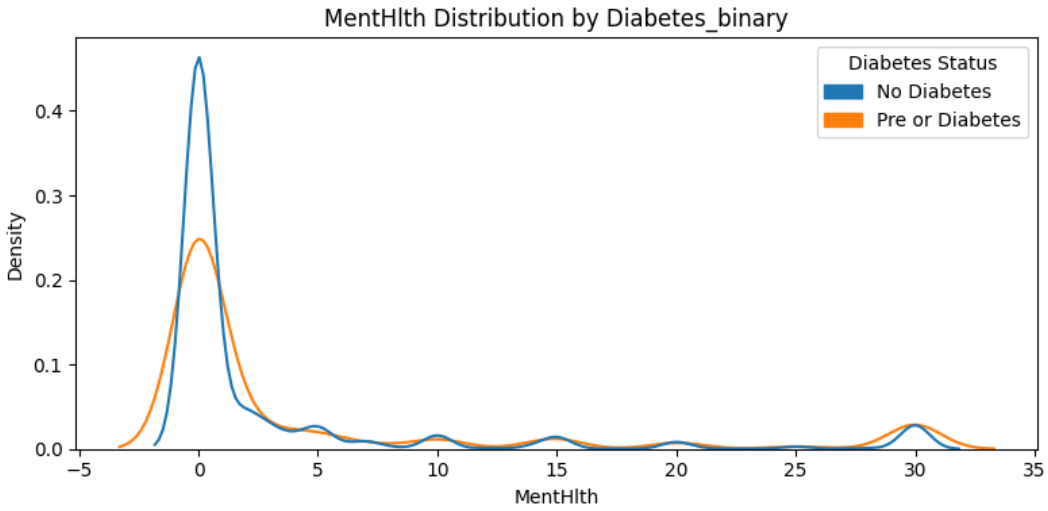

Figure 26. Days with bad Mental health by Diabetes presence

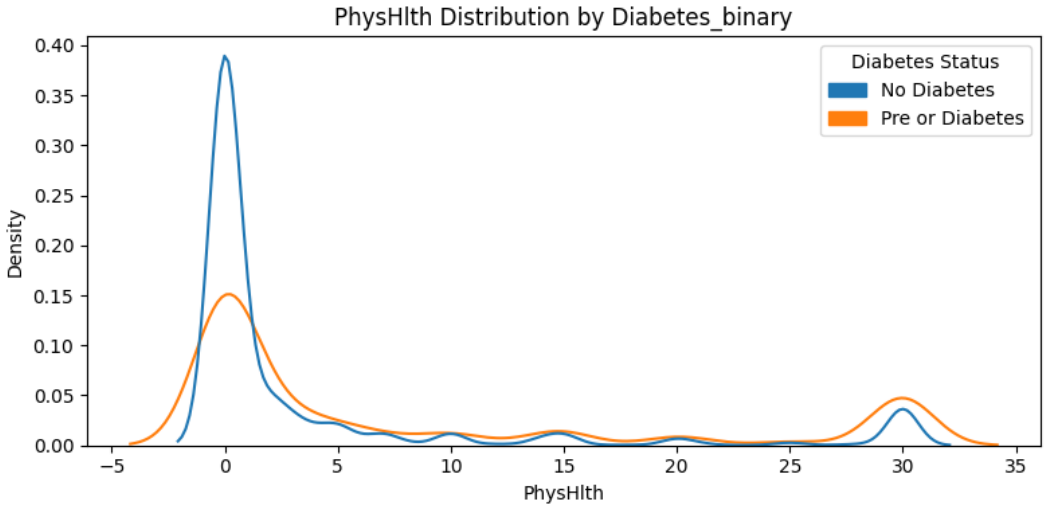

Figure 27. Days with bad Physical health by Diabetes presence

In the 2 present plots, they range from 0 to 30, representing the number of days in which the responder considers their mental and/or physical health has not been good, where 0 means that there were no days with negative health and 30 to

represents all days with negative health. Both plots imply the people who suffer the most from diabetes or prediabetes did not have a single day of bad mental or physical health, same with the people who do not suffer from diabetes. However, it is worth noting that those who claimed to have 30 days of bad mental or physical health present a surge in which the people suffer from diabetes or prediabetes more often than not.

5.4 Logistic Regression Model Results

The Logistic Regression Model constitutes the simplest among the selected ML models for this research (Not to say it is simple as a concept). The tuning of the aforementioned model is as follows:

Parameter	Value	Description
class_weight	"balanced"	Adjust the weights inversely to the class frequencies to handle imbalance
test_size	0.25	When splitting the independent and dependent variables, 25% is testing and 75% is for training
max_iter	1000	Maximum number of optimization iterations before forcing a stop
random_state	42	Selected seed for random number generation to ensure reproductibility
solver	lbfgs	Optimization algorithm (default)
penalty	l2	Regularization technique to reduce overfitting (default)
c	1	Inverse of regularization strength (default)

Table 10. Logistic Regression model parameters for tuning

Regarding the data preparation before training the Logistic Regression model is as follows:

Step (In order)	Description	Method/Tool used	Result/Notes
Missing Value Handling	Checked any null values	SimpleImputer (strategy='mean')	No missing values detected

Duplicate Removal	Checked for any duplicated rows	drop_duplicates()	24206 duplicates removed
Train/Test Splitting	Split dataset into training and testing subsets	train_test_split (test_size=0.25)	75% training, 25% testing; stratified
Class Imbalance Handling	Address imbalance in target variable	SMOTE (random_state=42)	Class ratio balanced. 0 and 1: 145782
Feature Scaling	Standardized features for uniform scale	StandardScaler()	Applied to both training and test sets using .fit_transform() and .transform()
Data shuffling	Randomly shuffled training data	shuffle() (random_state=42)	Shuffled training set after resampling and scaling

Table 11. Data preparation for Logistic Regression model training

These parameters and preparation steps for the data before training the Logistic Regression wielded the best results and performance for the model. The results of the model are as follows (tables and figures):

Metric	Value
Training time (s)	0.2 to 0.3s on average
Accuracy (0-1)	0.7154
Precision (0-1)	0.3186
Recall (0-1)	0.7560
F1-Score (0-1)	0.4483
ROC-AUC (0-1)	0.8076

AVG Cross-Validation F1 (0-1)	0.7504 ± 0.0017 (std deviation)
5-Fold Cross-Validation F1 (0-1)	[0.7478, 0.7497, 0.7504, 0.7521, 0.7519]

Table 12. Logistic Regression model evaluation Metric results

Classification report				
	Precision	**Recall**	**F1-score**	**Support**
0 (No diabetes)	0.94	0.71	0.81	48595
1 (Diabetes or Prediabetes)	0.32	0.76	0.45	8774
Accuracy			0.72	57369
Macro avg	0.63	0.73	0.63	57369
Weighted avg	0.85	0.72	0.75	57369

Table 13. Logistic Regression model Classification Report

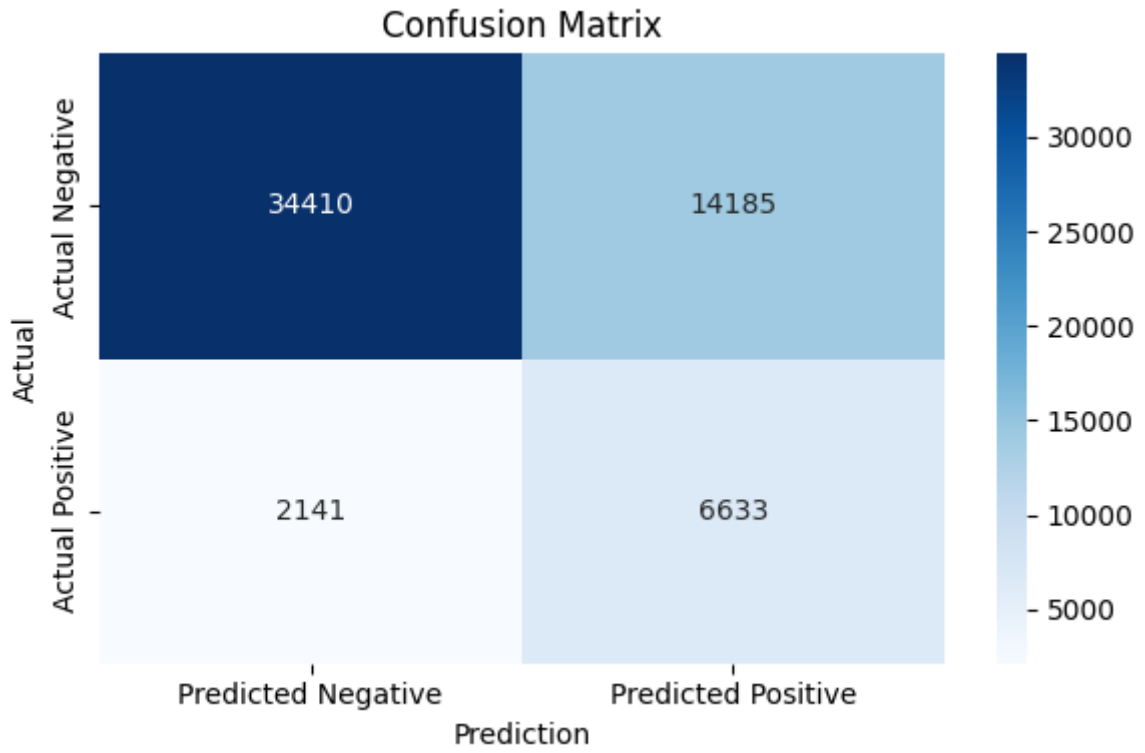

Figure 28. Logistic Regression model Confusion Matrix

The evaluation metrics prove that the Logistic Regression ML model performs well when it comes to separating the classes, especially considering the class

imbalance confirmed in the EDA. The ROC-AUC of 0.81 implies that the model is good at distinguishing between the 2 classes (remembering that 0.5 constitutes randomness and 1.0 is a perfect separation). However, the precision is 0.32, which is low but common when recall is prioritized via techniques like SMOTE and class weight balancing. The precision is high for class 0 (no diabetes) but low for class 1 (diabetes/prediabetes), recall is strong for class 1 (0.76) so the model detects most diabetes cases, the macro average gives equal weight to both classes so performance is balanced and the weighted average favors the majority class (0), inflating overall scores.

The confusion matrix reinforces the idea of a high recall (few missed positives) and low precision (many false alarms). Therefore, this model is recall-oriented (ideal when the idea is not to miss potential diabetes cases) but with the expense of precision that can raise false alarms that might be acceptable in a medical screening context and lastly, the ROC-ACU and CV values prove the model's stability and generalized learning.

5.5 Decision Tree Model Results

The Decision Tree Model constitutes a strong model that generally delivers a good performance while not increasing its complexity. The tuning of the aforementioned model is as follows:

Parameter	Value	Description
class_weight	"balanced"	Adjust the weights inversely to the class frequencies to handle imbalance
test_size	0.25	When splitting the independent and dependent variables, 25% is testing and 75% is for training
max_depth	16	Maximum permitted depth of the decision tree to prevent overfitting
random_state	42	Selected seed for random number generation to ensure reproductibility

Table 14. Decision Tree model parameters for tuning

Regarding the data preparation before training the Decision Tree model is as follows:

Step (In order)	Description	Method/Tool used	Result/Notes
Missing Value Handling	Checked any null values	SimpleImputer (strategy='mean')	No missing values detected
Duplicate Removal	Checked any duplicated rows	drop_duplicates()	24206 duplicates removed
Train/Test Splitting	Split dataset into training and testing subsets	train_test_split (test_size=0.25)	75% training, 25% testing; stratified
Class Imbalance Handling	Address imbalance in target variable	SMOTE (random_state=42)	Class ratio balanced. 0 and 1: 145782
Feature Scaling	SKIPPED	SKIPPED	SKIPPED
Data shuffling	Randomly shuffled training data	shuffle() (random_state=42)	Shuffled training set after resampling and scaling

Table 15. Data preparation for Decision Tree model training

These parameters and preparation steps for the data before training the Decision Tree wielded the best results and performance for the model. The results of the model are as follows (tables and figures):

Metric	Value
Training time (s)	2.3 to 2.6 on average
Accuracy (0-1)	0.8164
Precision (0-1)	0.3966
Recall (0-1)	0.3848
F1-Score (0-1)	0.3906
ROC-AUC (0-1)	0.7663

AVG Cross-Validation F1 (0-1)	0.8580 ± 0.0020 (std deviation)
5-Fold Cross-Validation F1 (0-1)	[0.8549, 0.8575, 0.8607, 0.8599, 0.8568]

Table 16. Decision Tree model evaluation Metric results

Classification report				
	Precision	**Recall**	**F1-score**	**Support**
0 (No diabetes)	0.89	0.89	0.89	48595
1 (Diabetes or Prediabetes)	0.40	0.38	0.39	8774
Accuracy			0.82	57369
Macro average	0.64	0.64	0.64	57369
Weighted average	0.81	0.82	0.82	57369

Table 17. Decision Tree model Classification Report

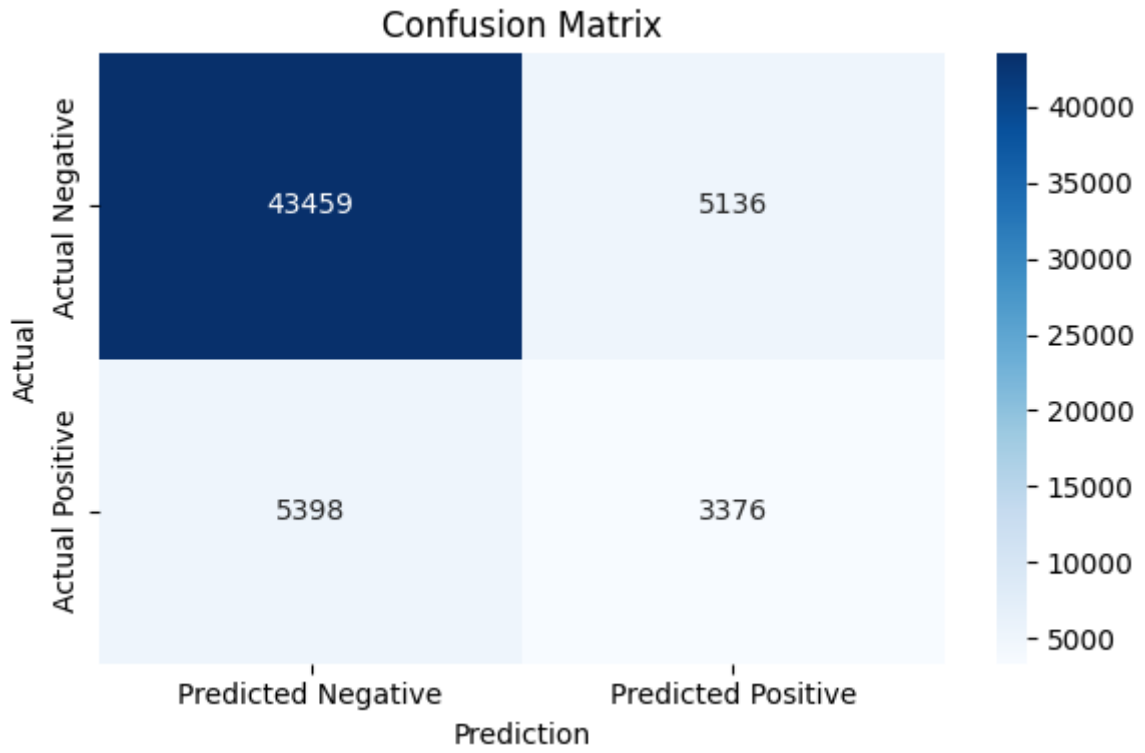

Figure 29. Decision Tree model Confusion Matrix

The evaluation metrics of the Decision Tree model indicate a higher overall accuracy (0.81) compared to that achieved by the Logistic Regression model.

However, the model's other metrics like precision and recall for the positive class (diabetes/prediabetes), are relatively low at 0.4 and 0.38 respectively. This implies that the model is good at correctly identifying the negative class (non-diabetes) with both precision and recall at 0.89, but struggles more with the positive class (diabetes/prediabetes).

The ROC-AUC score of 0.77 implies that there is a moderate discrimination capacity between both classes, but is lower than the Logistic regression's 0.81, a slightly less robust ranking, predictions and class separation. The confusion matrix further confirms this, the model procedures a moderate number of false negatives (FN, missed diabetes cases) and false positives (FP), a balanced but modes performance on detecting positive cases.

Despite this, the model proves to be very stable and with a generalizable performance during training with a 5-fold Cross-validation F1 score (0.86). Overall, the Decision Tree model shows a strong overall accuracy and stable learning, but may require further complementary techniques to improve its recall to diabetes cases, which is a key factor in diagnostics because missing positives are catastrophic.

5.6 Random Forest Model Results

The Random Forest Model constitutes an ensemble model which basically implements main Decision Tree models. The tuning of the aforementioned model is as follows:

Parameter	Value	Description
class_weight	"balanced"	Adjust the weights inversely to the class frequencies to handle imbalance
test_size	0.25	When splitting the independent and dependent variables, 25% is testing and 75% is for training
max_depth	16	Maximum permitted depth of the decision tree to prevent overfitting
random_state	42	Selected seed for random number generation to ensure reproductibility

n_estimators	100	Number of trees in the forest to build for the ensemble.
min_samples_split	15	Minimum samples required to split a node to control tree growth.
min_sample_leaf	5	Minimum samples required at a leaf node to prevent overfitting.
max_features	"sqrt"	Limit number of features considered for each split; "sqrt" (default)
n_jobs	-1	Uses all available CPU cores to parallelize training and improve performance.

Table 18. Random Forest model parameters for tuning

Regarding the data preparation before training the Random Forest model is as follows:

Step (In order)	Description	Method/Tool used	Result/Notes
Missing Value Handling	Checked any null values	SimpleImputer (strategy='mean')	No missing values detected
Duplicate Removal	Checked any duplicated rows	drop_duplicates()	24206 duplicates removed
Train/Test Splitting	Split dataset into training and testing subsets	train_test_split (test_size=0.25)	75% training, 25% testing; stratified
Class Imbalance Handling	Address imbalance in target variable	SMOTE (random_state=42)	Class ratio balanced. 0 and 1: 145782
Feature Scaling	SKIPPED	SKIPPED	SKIPPED
Data shuffling	Randomly shuffled training data	shuffle() (random_state=42)	Shuffled training set resampling and scaling

Table 19. Data preparation for Random Forest model training

These parameters and preparation steps for the data before training the Random Forest wielded the best results and performance for the model. The results of the model are as follows (tables and figures):

Metric	Value
Training time (s)	9 to 10 on average
Acurracy (0-1)	0.8347
Precision (0-1)	0.4519
Recall (0-1)	0.3798
F1-Score (0-1)	0.4127
ROC-AUC (0-1)	0.8077
AVG Cross-Validation F1 (0-1)	0.8889 ± 0.0014 (std deviation)
5-Fold Cross-Validation F1 (0-1)	[0.8874, 0.8873, 0.8896, 0.8916, 0.8887]

Table 20. Random Forest model evaluation Metric results

Classification report				
	Precision	Recall	F1-score	Support
0 (No diabetes)	0.89	0.92	0.90	48595
1 (Diabetes or Prediabetes)	0.45	0.38	0.41	8774
Accuracy			0.83	57369
Macro average	0.67	0.65	0.66	57369
Weighted average	0.82	0.83	0.83	57369

Table 21. Random Forest model Classification Report

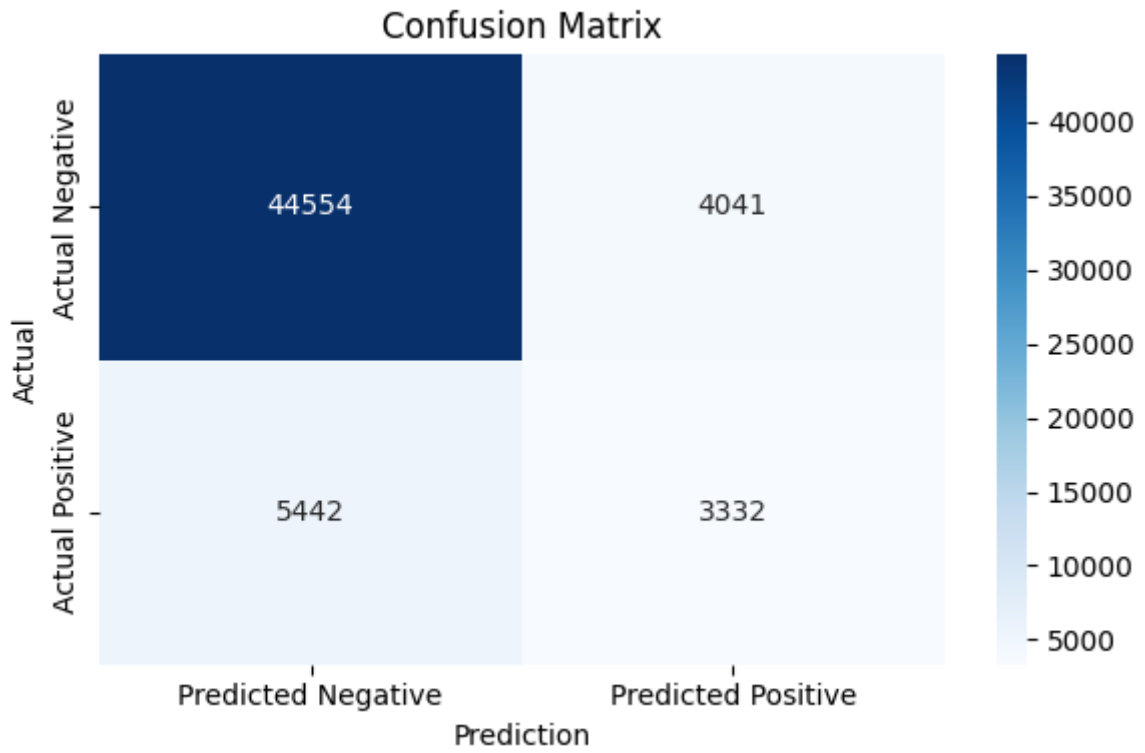

Figure 30. Random Forest model Confusion Matrix

The evaluation metrics of the Random Forest model demonstrate a strong overall performance in general, thanks to its ensemble nature, showing an accuracy of 0.83 (slightly higher than Decision Tree and Logistic Regression models). Precision and recall for the positive class (diabetes/prediabetes) are 0.45 and 0.38 respectively, which improves over Decision Tree precision, but recall remains modest at best. This indicates the ensemble model is better at identifying positive cases (TP) without significantly increasing false positives (FP).

It also performs very well on the negative class (non-diabetes), achieving a precision of 0.89 and recall of 0.92, a highly reliable model in ruling out diabetes when the prediction comes as negative and the ROC-AUC socre of 0.81 reflects a solid ability to distinguish the 2 classes, matching that of Logistic Regression model and surpassing Decision Tree. Furthermore, the Confusion Matrix supports this, suggesting the model is more conservative in flagging diabetes (likely due to imbalanced class distributions), but it's still excellent generalization and stability during training, evidenced by the 5-fold Cross-Validation F1 scores averaging 0.8889, making it robust and less prone to overfitting. While recall on positive class could still be improved, this model strikes a strong balance between accuracy, generalization and interpretability.

5.7 SVM Model Results

The Support Vector Machine (SVM) model is a supervised learning algorithm capable of finding the optimal hyperplane to separate classes. There are certain types of SVM models and for this research, it was decided to use the LinearSVC given the size of the dataset. The tuning of the aforementioned model is as follows

Parameter	Value	Description
class_weight	"balanced"	Adjust the weights inversely to the class frequencies to handle imbalance
test_size	0.25	When splitting the independent and dependent variables, 25% is testing and 75% is for training
random_state	42	Selected seed for random number generation to ensure reproductibility
max_iter	2000	Maximum number of optimization iterations before forcing a stop
c_reg_strength	0.01	Inverse of regularization strength (smaller values = stronger regularization)
cross_val	5	Number of folds in cross-validation to assess model performance

Table 22. SVM (LinearSVC) model parameters for tuning

Regarding the data preparation before training the SVM (LinearSVC) model is as follows:

Step (In order)	Description	Method/Tool used	Result/Notes
Missing Value Handling	Checked any null values	SimpleImputer (strategy='mean')	No missing values detected
Duplicate Removal	Checked any duplicated rows	drop_duplicates()	24206 duplicates removed
Train/Test Splitting	Split dataset into training and testing subsets	train_test_split (test_size=0.25)	75% training, 25% testing; stratified

Class Imbalance Handling	Address imbalance in target variable	SMOTE (random_state=42)	Class ratio balanced. 0 and 1: 145782
Feature Scaling	Standardized features for uniform scale	StandardScaler()	Applied to both training and test sets using .fit_transform() and .transform()
Data shuffling	Randomly shuffled training data	shuffle() (random_state=42)	Shuffled training set after resampling and scaling

Table 23. Data preparation for SVM (LinearSVC) model training

These parameters and preparation steps for the data before training the SVM (LinearSVC) wielded the best results and performance for the model. The results of the model are as follows (tables and figures):

Metric	Value
Training time (s)	3.1 to 3.5 on average
Acurracy (0-1)	0.7153
Precision (0-1)	0.3180
Recall (0-1)	0.7525
F1-Score (0-1)	0.4470
ROC-AUC (0-1)	0.8072
AVG Cross-Validation F1 (0-1)	0.7502 ± 0.0014 (std deviation)
5-Fold Cross-Validation F1 (0-1)	[0.74813051 0.74953025 0.74983792 0.75188195 0.75137278]

Table 24. SVM (LinearSVC) model evaluation Metric results

Classification report				
	Precision	Recall	F1-score	Support
0 (No diabetes)	0.94	0.71	0.81	48595
1 (Diabetes or Prediabetes)	0.32	0.75	0.45	8774
Accuracy			0.72	57369
Macro average	0.63	0.73	0.63	57369
Weighted average	0.85	0.72	0.75	57369

Table 25. SVM (LinearSVC) model Classification Report

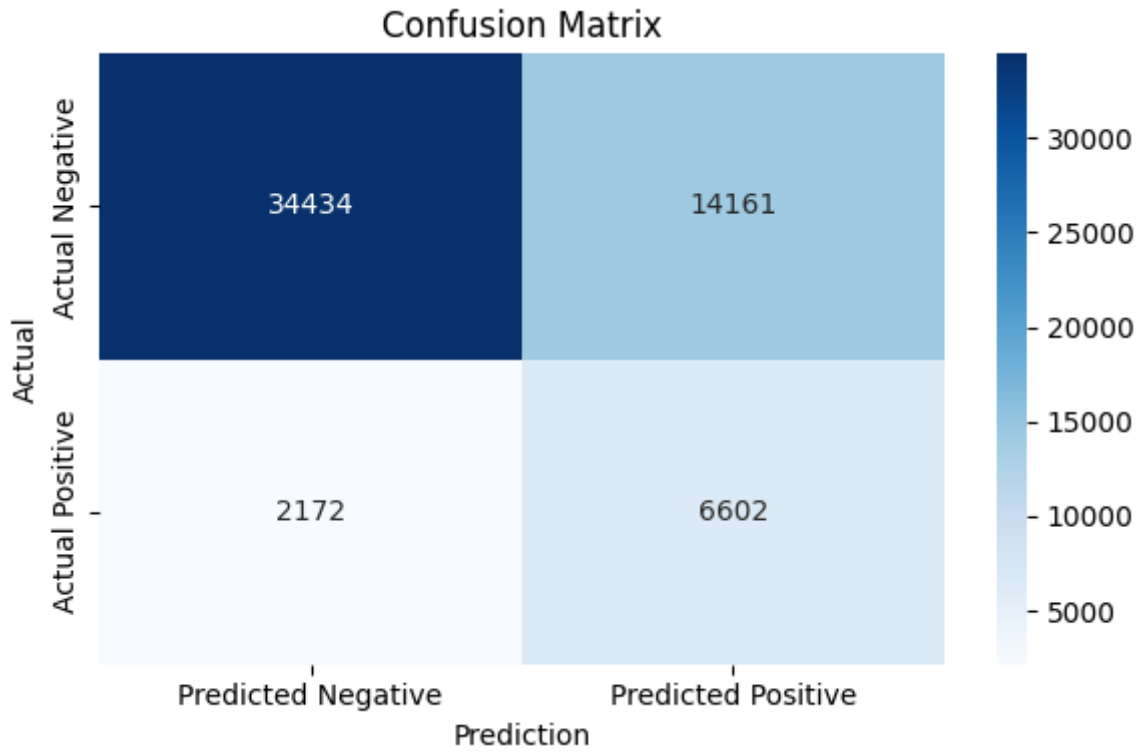

Figure 31. SVM (LinearSVC) model Confusion Matrix

The results of the evaluation metrics of the SVM (Linear SVC) model indicate a notable difference in performance model compared to the previous ensemble model. With an accuracy of 0.7153, the model performs reasonably well overall when it comes to classification, but is still lower than the ensemble based approaches. That said, the LinearSVC stands out in its recall ability for positive cases (diabetes/prediabetes) with a recall of 0.7525 (higher than Random Forest and Logistic Regression), but it comes at the cost of precision (0.3180). This trade-off

implies an aggressive approach in flagging positive cases, resulting in more true positives (TPs) but also increases the false positives (FPs).

Regarding the negative class (non-diabetes), the model also performs strongly with a precision of 0.94 and a recall of 0.71, quite reliable to identify non-diabetic individuals, but less conservative than Random Forest. Regarding the ROC-AUC score of 0.8072, it demonstrates that despite the class imbalance and relatively low precision on the positive class (diabetes/prediabetes), the model remains stable compared to Random Forest and Logistic Regression when it comes to class separation.

The confusion matrix of this model further highlights the model's behavior, with a large number of non-diabetic patients correctly classified, but also a big portion are misclassified as diabetic due to high sensitivity. Still, it is possible that this is preferable in medical screenings, given that a missed diabetic diagnosis case is more critical than a false alert. When it comes to its generalization performance, it shows an excellent 5-fold Cross-Validation F1 score by averaging a 0.7502 ± 0.0014, a well regularized model that is less prone to overfitting despite the data imbalance.

5.8 Direct comparison between all models

Within this section, there are side-by-side comparisons between the performance metrics for all the implemented machine learning models evaluated in this study. A concise overview of each model's strengths and weaknesses and also facilitating an easier interpretation, helping to identify the most effective model for the prediction of diabetes disease.

5.8.1 Metrics comparison between all models

//////////////////////////	Machine Learning Models			
Metrics	**Logistic Regression**	**Decision Tree**	**Random Forest**	**SVM (LinearSVC)**
Training time (s)	0.2 - 0.3s avg	2.3 - 2.6 avg	9 - 10 avg	3.1 - 3.5 avg
Acurracy (0-1)	0.7154	0.8164	0.8347	0.7153
Precision (0-1)	0.3186	0.3966	0.4519	0.3180

Recall (0-1)	0.7560	0.3848	0.3798	0.7525
F1-Score (0-1)	0.4483	0.3906	0.4127	0.4470
ROC-AUC (0-1)	0.8076	0.7663	0.8077	0.8072
AVG Cross-Validation F1 (0-1)	0.7504 ± 0.0017 (std dev)	0.8580 ± 0.0020 (std dev)	0.8889 ± 0.0014 (std dev)	0.7502 ± 0.0014 (std dev)

Table 26. Evaluation Metrics comparison between all models

From the present table, there are several key takeaways from this direct comparison of the listed evaluation metrics:

- Overall performance: Random Forest got the best results across the listed key metrics, specifically accuracy (0.8347), precision (0.4519), and cross-validation F1-score (0.8889), a strong generalization. While Logistic Regression and SVM have lower accuracy and precision, they present a higher recall (>0.75), making both useful for minimizing FN.
- Precision vs Recall trade-off: Logistic Regression and SVM favor a higher sensitivity (recall), which is beneficial to detect possible diabetic cases, meanwhile, Random Forest has a better balance with a higher precision without totally sacrificing recall.
- Training time: Logistic Regression trained the fastest by a large factor (0.2 - 0.3s), suitable for a quick deployment. Random Forest was the slowest (9 - 10s) but with a better performance.
- Stability: Random forest and SVM (LinearSVC) had both the lowest standard deviation (±0.0014) in their cross-validation, indicating a reliable performance in comparison to Random Forest and Logistic Regression.
- ROC-AUC: All models achieved a high ROC-AUC (~0.807) except for Decision Tree (0.7663), but overall, a good class separation capacity.

Random Forest emerged as the most effective of all for early diabetes prediction in terms of accuracy, precision and stability, but, SVM and Logsitic Regression model's high recall rates attest their strength in sensitive detection scenarios (prioritized to avoid missing TP). These results support the idea of a Hybrid strategy in which ensemble models (Random forest) can be used for general

prediction and simpler models (Logistic Regression or SVM) when transparency and high recall are necessary for clinical adoption

5.8.2 Confusion matrices comparison between all models

//////////////////////////	Machine Learning Models			
Confusion Matrix sections	**Logistic Regression**	**Decision Tree**	**Random Forest**	**SVM (LinearSVC)**
TP (Real diabetes)	6633	3376	3332	6602
FP (False diabetes)	14185	5136	4041	14161
TN (Real non-diabetes)	34410	43459	44554	34434
FN (False non-diabetes)	2141	5398	5442	2172

Table 27. Confusion matrices comparison between all models

From the direct comparison of the Confusion matrices of all models, certain takeaways can be observed:

- TP: Logistic Regression and SVM predicted the highest number of true diabetic cases (6633 and 6602), a strong sensitivity that is crucial in a clinical setting.
- FP: Logistic Regression and SVM had the highest FP counts (around 14000), indicating they misclassified non-diabetic individuals as diabetic more often than the other 2 models, the aforementioned trade-off for higher recall.
- TN: Random Forest and Decision Tree had the highest number of correctly identified non-diabetic cases (44554 and 43459), strong when it comes to distinguishing negative cases.
- FN: Random Forest and Decision Tree had more FN (around 5400), making them more likely to miss actual diabetic cases (potential critical limitation).

These matrices reinforce the idea that Logistic Regression and SVM prioritize sensitivity (catch more diabetes cases at the cost of higher FP). While Random

Forest has the best overall balance in its metrics, it still misses more actual diabetic cases than the 2 previously mentioned (Decision Tree is similar but with lower reliability). Clearly, there must be a practical consideration between minimizing FN and overall classification precision, a critical decision point depending on healthcare priorities.

5.8.3 Classification reports comparison between all models

Class/Metric	Logistic Regression	Decision Tree	Random Forest	SVM (LinearSVC)
Class 0 (Non-diabetes)	//			
Precision	0.94	0.89	0.89	0.94
Recall	0.71	0.89	0.92	0.71
F1-score	0.81	0.89	0.90	0.81
Support	48595			
Class 1 (diabetes/ prediabetes)	//			
Precision	0.32	0.40	0.45	0.32
Recall	0.76	0.38	0.38	0.75
F1-score	0.45	0.39	0.41	0.45
Support	8774			
Macro average	//			
Precision	0.63	0.64	0.67	0.63
Recall	0.73	0.64	0.65	0.73

F1-score	0.63	0.64	0.66	0.63
Weighted average	//			
Precision	0.86	0.81	0.82	0.85
Recall	0.72	0.82	0.83	0.72
F1-score	0.75	0.82	0.83	0.75

Table 28. Classification report comparison between all models

The classification report, while a little more complex, shows key differences between all models:

- Class 0 (Non-diabetes):
  - All models performed well on predicting non-diabetics with a high precision (0.89 to 0.94) and F1-scores (0.81 to 0.90).
- Class 1 (Diabetes/Prediabetes):
  - Logistic Regression and SVM stood out in terms of recall (0.75 and 0/76), correctly identifying most of the actual diabetes cases (TP), but their precision remained low (0.32 for FP).
  - Random Forest balanced with the highest precision (0.45), decent recall (0.38) and better F1-score (0.41).
- Macro and Weighted averages:
  - Random Forest had the best macro-averaged performance, once again indicating a balanced effectiveness across both classes.
  - In weighted averages (that accounts for class imbalance), Random Forest leads with an F1-score of 0.83, followed by Decision Tree 0.82

The class-wise metrics reveal a key distinction between the models, which is that Logistic Regression and SVM are better and identifying actual diabetic cases (recall), while Random Forest offers the best overall balance across both classes. This reinforces the need to weigh sensitivity (recall) vs precision and class balance when choosing a ML model for early diabetes detection.

5.9 Discussion

The performance from all four ML models showcases critical insight into their applicability for early diabetes prediction. Each model demonstrated their own strengths and limitations, supporting the argument that there is no single universally superior model but rather a more context dependent issue based on the priorities of the medical body.

5.9.1 Performance vs. Clinical Priorities

Related to raw performance, Random Forest consistently outperformed other models in terms of accuracy (83.47%0), precision (45.19%), F1-score (0.83 weighted) and cross-validation stability (±0.0014). These metrics support that Random Forest is a highly reliable and generalizes well across data splits, making it an excellent candidate for a robust and scalable deployment.

On the other hand, Logistic Regression and SVM (LinearSVC) stood out for their high recall results (~75%), both models capable of correctly identifying the majority of true diabetic or prediabetic cases (TPs). This trait in a medical diagnostics context is critical when it comes to minimizing the false non-diabetic cases (FNs) to ensure patients with the disease are not overlooked.

Meanwhile, Decision Tree showed an overall intermediate performance between all of these approaches, but suffering a great deal of higher false negatives (FNs) and lower stability, making this model less than ideal as a standalone choice.

An essential question from the clinical part arises based on these results. Knowing the ML model limitations, should a model prioritize minimizing the false positives (misdiagnosis of healthy individuals) or the false negatives (missed actual diabetic cases). In many cases, especially for chronic disease management like diabetes, missing a diagnosis can be more harmful than a false alarm, especially knowing that, at early stages, simple lifestyle changes can help the patient without the need for invasive procedures or the need for drug consumption. From this perspective, a model with a high recall (Logistic Regression and SVM) may be more valuable despite their lower precision, but the aforementioned is not trivial, since an excessive number of false positives (FPs) can burden healthcare systems and patients alike since the disease will progress unchecked until it is too late.

Given these considerations, a hybrid modeling approach should be considered. By combining the strengths of multiple models to achieve a more balanced and adaptable diagnostic pipelines can be created. This strategy may offer a better compromise between robustness, sensitivity and interpretability, aligning with real world medical requirements.

5.9.2 Confusion Matrices and Trade-offs

The confusion matrices revealed that, by a large marging, Logistic Regression and SVM models hold the highest number of TPs but also the highest number of FPs, highlighting a common (and expected) precision-recall trade-off, where a higher sensitivity lead to more false alarms but also catch correct diabetes or prediabetes cases more often, underlying a better class separation but at the cost of more FNs.

These trade-offs must be carefully considered in clinical practice. FNs (missed cases) could lead to severe consequences if the diabetic condition goes untreated and the patient continues to worsen, on the other hand, FPs (false alarms) could lead to unnecesary wast of medical resources (like testing) but may still be preferable in early screening scenarios in the diagnosis of diabetes or prediabetes.

5.9.3 Class-Wise Metrics and Interpretability

Metrics related to classes further support subsections 5.9.1 and 5.9.2. Random Forest maintained a balanced performance across both classes with the highest F1-score for non-diabetics (0.90) and the best precision for diabetics (0.45). Meanwhile, Logistic Regression and SVM remained strong in diabetic recall with a lower precision.

Regarding the transparency and interpretability of a ML model within medical contexts, like clinical facing tools, Logistic Regression may be preferable despite its lower overall performance due to its interpretability and fast training time, this last point may also make Logistic Regression a very maintainable model among the implemented models.

5.9.4 Hybrid Strategy Perspective

As previously mentioned, the potential for a hybrid strategy approach becomes increasingly attractive to get the best out of all the implemented models for better results overall. For example:

- Random Forest can be utilized for general population screening or bulk prediction, where performance and accuracy are the most important aspects.
- Logistic Regression or SVM can be deployed in sensitive detection pipelines instead, where high recall and explainability are critical. For example, follow up triaging or clinical review.

A hybrid perspective opens the possibility of leveraging the best traits of each model based on the healthcare task at hand.

5.9.5 Training time and computational costs

A very important aspect of any implementation of AI models into existing fields of study and their practical counterparts is the training time and the computational cost. Especially considering that the field is the medical/clinical field, it is imperative to deploy models that bring the best results while dealing efficiently with large datasets and real-time systems with the best management of available computing resources. Focusing purely on training time and computational cost, the 4 selected models during this study can be ordered from least computationally taxing and quickest to train to the most computationally taxing and longest to train:

1. Logistic Regression: Trained the fastest by a large margin compared to the other models with an approximate of 0.2 to 0.3 seconds each training time. These results make this model ideal for rapid depolyment, rapid prototyping, real-time interference and/or applications with frequent retraining requirements.
2. Decision Trees: While not exceeding in any of the evaluation metrics, this model is the second fastest to train with a moderate time of 2.3 to 2.6 seconds, offering a good balance between speed and interpretability. They might be suited for scenarios where transparent decision making and low latency interference are necessary, although it can not be overlooked the fact that this comes at the cost of possible overfitting risks if not pruned as necessary.
3. LinearSVC (Support Vector Machine): Compared to Decision Trees, it took around 3.1 to 3.5 seconds, showing the extra computational cost required in optimizing the margin based objective function and the second most computationally demanding of the 4 selected models. While slower than simpler linear models, SVMs can yield a higher accuracy in some linearly

separable datasets with metrics comparable to Logistic Regression models, but their scalability to large datasets is limited.

4. Random Forest: By far the most intensive computationally speaking with around 9 to 10 seconds for training. These results are expected since, as an ensemble model, it builds several trees ranging from tens to hundreds in parallel, each requiring computation over bootstrapped datasets and random feature subsets. Of course, this cost is compensated by the superior overall predictive performance and robustness to overfitting compared to Decision Trees.

Overall, the most limiting model in the training time and computational cost aspect is Random Forest, especially within low resource environments. Meanwhile, Logistic Regression and Decision Trees offer much faster training times and require lower computational resources, making them simpler and more interpretable use cases. Finally, LinearSVC proves to be somewhat high in this regard but with results similar to Logistic regression metrics wise, making it a special case.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

During this study, the performance of various Machine Learning classification algorithms have been evaluated and compared, specifically Logistic Regression, Decision Tree, Random Forest and SVM for the early detection of diabetes chronic disease using a large, real-world health dataset. The findings obtained from this research reaffirm the growing potential of machine learning implementation in clinical decision support while highlighting the important trade-offs between model complexity, accuracy and interpretability.

The key findings within this study can be summarised to the following bullet points:

- "Black box" models (like Random Forest) are more accurate but more complex, take longer to train and demand more computational resources.
- "White box" models (like LinearSVC, Logistic Regression and Decision Tree) are less accurate but simpler, take less time to train and demand less computational resources.
- Logistic Regression is the best model in terms of training time and computational cost, followed by Decision Tree, SVM (LinearSVC) and lastly Random Forest as the most demanding in both regards.
- Random Forest consistently outperformed the other 3 models in terms of overall evaluation metrics (accuracy, precision and F1-score).
- Random Forest is the most suited for general population screening and scalable deployment.
- Random Forest inherent "Black box" nature limits and presents important barriers for direct and seamless implementation in clinical settings.
- Logistic Regression and SVM, while weaker in overall evaluation metrics, achieved the higher recall values (identified more TPs and minimized FNs.

There is an emphasis on the argument that there is no universally "best" model given the strengths and weaknesses each model presents, making it more related to the optimal choice depending on the priorities of medical diagnostics. For instance, if minimizing the number of missed cases (FNs) is critical, then models with a high-recall may be better, but if reducing false alarms (FPs) and maximization of

general performance is more important, ensemble models are a better option. This supports the idea that individual models, also referred to as "White box", might be less accurate, but they are quicker to train, more interpretable and less complex without the drawback of overfitting towards the majority class.

Given these insights, it is proposed that a hybrid modelling approach might offer a more balanced solution that leverages the different models in different stages of the diagnostic pipeline and uses them in the areas where they are the strongest.

6.2 Future Work

Several routes can be taken to further enhance the applicability and impact of this research and its purpose:

- Clinical consultation: Future studies should involve direct input from medical professionals to understand clinical preferences. Would practitioners prefer models with higher recall and interpretability, even if it leads to more false positives? Or do they prefer more precise, stable, but complex models that risk missing some true cases?.
- Model calibration and optimization: Exploring the tuning threshold or cost-sensitive learning could help balance the false positives and false negatives more effectively based on a specific healthcare context.
- Broader dataset exploration: Expanding the analysis to include additional or more recent datasets could test model robustness and generalization across the different populations and time periods.
- Explainability tools: Since explainability is a big issue with more complex models, integrating explainable AI techniques like SHAP or LIME could bridge this particular gap between models like ensemble ones and their great, balanced performance with the interpretability required by healthcare standards.
- System integration testing: Future studies and researches should also explore deploying trained models into simulated or even real-world healthcare settings to fully evaluate their usability, response time and acceptance among medical staff.

References

1) **Teboul, A**. (2021). *Diabetes Health Indicators Dataset*. Kaggle. Available at: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset (Accessed and downloaded: 12th May 2025).
2) **Centers for Disease Control and Prevention (CDC)** (2022). Behavioral Risk Factor Surveillance System. Kaggle. Available at: https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system (Accessed: 12th May 2025).
3) **Zarbin, M.A, et al.** (2021). *Data science*, in: Translational Vision Science & Technology. ARVO Journal, 10 (8), p. 20. Available at: https://tvst.arvojournals.org/article.aspx?articleid=2776501 (Accessed: 20th May 2025).
4) **Badillo, S, et al.** (2020). *An introduction to machine learning*, in: Clinical Pharmacology & Therapeutics. Wiley-Blackwell, 107(4), pp. 871–885. Available at: https://ascpt.onlinelibrary.wiley.com/doi/10.1002/cpt.1796 (Accessed: 20th May 2025).
5) **Igual, L. and Seguí, S.** (2024). *Supervised learning*, in: Introduction to Data Science. Springer eBooks, pp. 67–97. Available at: https://link.springer.com/chapter/10.1007/978-3-031-48956-3_5 (Accessed: 20th May 2025).
6) **Wing, J.M**, (2019). *The Data Life Cycle*, in: Harvard Data Science Review, 1(1). Available at: https://hdsr.mitpress.mit.edu/pub/577rq08d/release/4 (Accessed: 20th May 2025).
7) **Özsu, M.T**, (2024). *Foundations and scoping of data science*. arXiv preprint. Available at: https://arxiv.org/abs/2301.13761 (Accessed: 20th May 2025).
8) **Richards, J.A**, (2022). *Supervised classification techniques*, in: Remote Sensing Digital Image Analysis. Springer eBooks, 6, pp. 263–367. Available at: https://link.springer.com/chapter/10.1007/978-3-030-82327-6_8 (Accessed: 21st May 2025).
9) **Rainio, O., Teuho, J. & Klén, R**, (2024). *Evaluation metrics and statistical tests for machine learning*. Scientific Reports 14(1). Available at: https://www.nature.com/articles/s41598-024-56706-x (Accessed: 21st May 2025).

10) **Petersen, E. *et al*,** (2022). *Responsible and Regulatory conform Machine Learning for Medicine: A Survey of Challenges and solutions*. IEEE Access, 10, pp. 58375–58418. Available at: https://ieeexplore.ieee.org/document/9783196 (Accessed: 21st May 2025).

11) **Ali, M.S. et al**, (2024). *Federated Learning in Healthcare: model misconducts, security, challenges, applications, and Future Research Directions - A Systematic review*. Available at: https://arxiv.org/abs/2405.13832 (Accessed: 21st May 2025).

12) **Khattak, F.K. et al**, (2023). *MLHOPs: Machine Learning for Healthcare Operations*. Available at: https://arxiv.org/abs/2305.02474v1 (Accessed: 23rd May 2025).

13) **Schinkel, M. et al**, (2023). *Detecting changes in the performance of a clinical machine learning tool over time*, in: EBioMedicine, 97. Available at: https://pubmed.ncbi.nlm.nih.gov/37793210/ (Accessed: 23rd May 2025).

14) **Chandrashekar, G. and Sahin, F**, (2022). *Stability of feature selection algorithm: A review*, in: Journal of King Saud University - Computer and Information Sciences. Springer eBooks 34(4), pp. 1060-1073. Available at: https://www.sciencedirect.com/science/article/pii/S1319157819304379 (Accessed: 25th May 2025).

15) **Naheed, N. et al**, (2020). *Importance of features selection, attributes selection, challenges and future directions for medical imaging data: a review*, in: Computer Modeling in Engineering & Sciences. Tech Science Press, 125(1), pp. 314–344(31). Available at: https://www.ingentaconnect.com/contentone/tsp/cmes/2020/00000125/00000001/art00017#Refs (Accessed: 25th May 2025).

16) **World Health Organization: WHO**, (2024). *Diabetes*. Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes (Accessed: 26th May 2025).

17) **American Diabetes Association**, (2022). *American Diabetes Association Releases 2023 Standards of Care in Diabetes to Guide Prevention, Diagnosis, and Treatment for People Living with Diabetes*, Press Release. Available at: https://diabetes.org/newsroom/american-diabetes-association-2023-standards-care-diabetes-guide-for-prevention-diagnosis-treatment-people-living-with-diabetes (Accessed: 26th May 2025).

18) **Zimmet, P. et al**, (2016). *Diabetes mellitus statistics on prevalence and mortality: facts and fallacies*, in: Nature Reviews Endocrinology, 12(10), pp. 616–622. Available at: https://www.nature.com/articles/nrendo.2016.105 (Accessed: 26th May 2025).

19) **Varela, R. et al**, (2021). *Hyperglycemia and hyperlipidemia can induce morphophysiological changes in rat cardiac cell line*, in Biochemistry and Biophysics Reports, 26. Available at: https://pubmed.ncbi.nlm.nih.gov/33912691/ (Accessed: 27th May 2025).

20) **Soni, M**, (2020). *Diabetes Prediction using Machine Learning Techniques*, IJERT, 9(9). Available at: https://doi.org/10.17577/IJERTV9IS090496 (Accessed: 28th May 2025).

21) **Barth, S. and Flam, S**, (2025). *Machine Learning in Healthcare: Guide to Applications & benefits*. Available at: https://www.foreseemed.com/blog/machine-learning-in-healthcare (Accessed: 28th May 2025).

22) **Sarkar, K. et al**, (2020). *Machine Learning for Health (ML4H) 2020: Advancing Healthcare for All*. ML4H. Available at: https://proceedings.mlr.press/v136/sarkar20a/sarkar20a.pdf (Accessed: 28th May 2025).

23) **Kelley, K**, (2024). *Machine learning in healthcare: applications, use cases, and careers*. Caltech. Available at: https://pg-p.ctme.caltech.edu/blog/ai-ml/machine-learning-in-healthcare-applications-use-cases-careers (Accessed: 28th May 2025).

24) **Scikit-learn developers**, (no date). *LogisticRegression*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (Accessed: 29th May 2025).

25) **Scikit-learn developers**, (no date). *1.10. Decision Trees*. Available at: https://scikit-learn.org/stable/modules/tree.html (Accessed: 29th May 2025).

26) **Scikit-learn developers**, (no date). *RandomForestClassifier*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (Accessed: 29th May 2025).

27) **Scikit-learn developers**, (no date). *SVC*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC (Accessed: 30th May 2025).

28) **Scikit-learn developers**, (no date). *LinearSVC*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html (Accessed: 30th May 2025).

29) **Scikit-learn developers**, (no date). *sklearn.svm*. Available at: https://scikit-learn.org/stable/api/sklearn.svm.html (Accessed: 30th May 2025).

30) **Zhang, Z**, (2025). *Comparison of machine learning models for predicting Type 2 diabetes risk using the PIMA Indians Diabetes Dataset*, in: Journal of Innovations in medical research. Paradign ACAD Press, 4(1). Available at: https://www.paradigmpress.org/jimr/article/view/1532 (Accessed: 31st May 2025).

31) **Abedini, M., Bijarr, A. and Banirostam T**, (2020). *Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network*. IJARCE, 9(7). Available at: https://www.academia.edu/43751015/Classification_of_Pima_Indian_Diabetes_Dataset_using_Ensemble_of_Decision_Tree_Logistic_Regression_and_Neural_Network (Accessed: 31st May 2025).

32) **Sandhu, S. et al**, (2020). *Integrating a machine learning system into clinical workflows: Qualitative study*, in: Journal of Medical Internet Research, 22(11), p. e22421. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7714645/ (Accessed: 31st May 2025).

33) **Mushtaq, Z. et al**, (2022). *Early Detection of Diabetes Using a Hybrid Approach Based on the Voting Classifier*, in: Open Journal of Applied Sciences, 15(4), pp. 784-797. Available at: https://www.scirp.org/journal/paperinformation?paperid=141602 (Accessed: 1st June 2025).

34) **Rahman, Md.A. et al**, (2023). *Machine Learning-Based Approach for Predicting Diabetes employing Socio-Demographic Characteristics*, in Algorithms, 16(11), pp. 503. Available at: https://www.mdpi.com/1999-4893/16/11/503 (Accessed: 1st June 2025).

35) **Balakrishnan, K., P, R. and Mahadeo, U**, (2021). *Analysing stable feature selection through an augmented marine predator algorithm based on*

opposition-based learning, in: Expert Systems, 39(1). Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12816 (Accessed: 1st June 2025).

36) **Zhou, V**, (2022). *Random forests for complete beginners*, in: victorzhou.com. Available at: https://victorzhou.com/blog/intro-to-random-forests/#23-training-a-decision-tree-when-to-stop (Accessed: 2nd June 2025).

37) **Zhou, V**, (2022). *A Simple Explanation of Information Gain and Entropy*, in: victorzhou.com. Available at: https://victorzhou.com/blog/information-gain/ (Accessed: 2nd June 2025).

38) **Zhou, V**, (2022). *A Simple Explanation of Gini Impurity*, in: victorzhou.com. Available at: https://victorzhou.com/blog/gini-impurity/#example-3-an-imperfect-split (Accessed: 2nd June 2025).

39) **Husain, A., Khan, M.H,** (2018). *Early Diabetes Prediction Using Voting Based Ensemble Learning*, Advances in Computing and Data Sciences, ICACDS. Springer eBooks, 905(1), pp. 95-103. Available at: https://link.springer.com/chapter/10.1007/978-981-13-1810-8_10 (Accessed: 2nd June 2025).

40) **Wu, Y. et al**,(2021). *LIME: Learning Inductive Bias for primitives of Mathematical Reasoning*, in: Proceedings of the 38th International Conference on Machine Learning. PMLR, pp. 11251–11262. Available at: https://proceedings.mlr.press/v139/wu21c.html (Accessed: 3rd June 2025).

41) **Tonekaboni, S. et al**, (2019). *What clinicians want: Contextualizing explainable machine learning for clinical end use*. Available at: https://arxiv.org/abs/1905.05134v2 (Accessed: 3rd June 2025).

42) **Salmi, M. et al**, (2024). *Handling imbalanced medical datasets: review of a decade of research*, in: Artificial Intelligence Review, 57(273). Available at: https://link.springer.com/article/10.1007/s10462-024-10884-2 (Accessed: 3rd June 2025).

43) **Nasarian, E. et al**, (2024). *Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework*, in: Information Fusion, 108. Available at: https://www.sciencedirect.com/science/article/abs/pii/S1566253524001908 (Accessed: 3rd June 2025).

44) **Centers for Disease Control and Prevention**, (2024). *Diabetes Basics*. Available at: https://www.cdc.gov/diabetes/about/index.html (Accessed: 4th June 2025).

45) **Shetty, S.H. et al**, (2022). *Supervised Machine Learning: Algorithms and Applications*, in: Fundamentals and Methods of Machine and Deep Learning: Algorithms, Tools and Application. Wiley, pp. 1-16. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119821908.ch1 (Accessed: 5th June 2025).

46) **Baloglu, O., Latifi, S.Q. and Nazha, A**, (2021). *What is machine learning?*, in: Archives of Disease in Childhood Education & Practice, 107(5), pp. 386–388. Available at: https://ep.bmj.com/content/107/5/386.abstract (Accessed: 5th June 2025).

47) **robot learner**, (2023). *A comprehensive guide to binary, Multi-Class, and Multi-Label* classification, in: DataScienceTribe. Available at: https://www.datasciencebyexample.com/2023/06/09/binary-classification-vs-multi-class-classification-vs-multi-label-classification (Accessed: 5th June 2025).

48) **Erasmus, A., Brunet, T.D.P. and Fisher, E**, (2021). *What is Interpretability?*, in: Philosophy & Technology, 34(4), pp. 833–862. Available at: https://doi.org/10.1007/s13347-020-00435-2 (Accessed: 6th June 2025).

49) **Setyati, R. et al**, (2024). *The Importance of Early Detection in Disease Management*, in: Journal of World Future Medicine Health and Nursing, 2, pp. 51-63. Available at: https://www.scirp.org/reference/referencespapers?referenceid=3882714 (Accessed: 6th June 2025).

50) **Liu, B. and Mazumder, R**, (2024). *Randomization can reduce both bias and variance: a case study in random forests*. Available at: https://arxiv.org/abs/2402.12668 (Accessed: 19th June 2025).