## Name:Monica Serwala

Student number 32711514

616 Project

Table of Contents

[Name:Monica Serwala 1](#_Toc198915339)

[1. Introduction 2](#_Toc198915340)

[2. Methodology 3](#_Toc198915341)

[*1.1.* *Dataset Characteristics* 3](#_Toc198915342)

[*1.2.* *Pre-processing and Feature Engineering* 4](#_Toc198915343)

[*1.3.* *Model Training and Selection* 8](#_Toc198915344)

[3. Results and Discussion 10](#_Toc198915345)

[a. Performance Metrics Overview 11](#_Toc198915346)

## Introduction

The global burden of diabetes has escalated dramatically in recent years, positioning the disease as one of the most pressing public health challenges of the 21st century. With millions of people affected worldwide, early detection and proactive management have become crucial tools in mitigating long-term complications and reducing healthcare costs. Against this backdrop, the application of machine learning techniques in medical diagnostics presents a transformative opportunity - not only to enhance predictive accuracy but also to support data-driven clinical decision-making.

In this project, the focus is placed on developing a classification model capable of predicting the likelihood of an individual having a family history of diabetes based on a diverse set of physiological, lifestyle, and demographic factors. This task is not trivial. A family history of diabetes is a well-established risk factor and often serves as a clinical indicator for early intervention. Therefore, being able to predict it using available data could support targeted public health strategies and enable personalized preventive care.

The dataset used for this project, titled [Diabetes\_prediction\_dataset](https://www.kaggle.com/datasets/marshalpatel3558/diabetes-prediction-dataset/data), was sourced from Kaggle and compiled by user MarshalPatel3558 . The dataset was chosen after careful consideration of several alternatives due to its well-rounded structure, suitable size, and practical relevance. Containing 10,000 instances and 20 features, the dataset includes a mixture of numerical and categorical variables such as BMI, cholesterol levels, alcohol consumption, and physical activity. These features reflect a broad range of health indicators commonly used in clinical assessments. More importantly, the target variable, Family\_History\_of\_Diabetes, is binary, making it a well-defined classification problem - a key requirement for this semester’s machine learning project.

What made this dataset particularly compelling was its realism and complexity. Unlike artificially balanced or overly simplistic academic datasets, this one reflects the kind of noisy, imbalanced, and multifactorial data encountered in real-world healthcare scenarios. The presence of missing values, mixed data types, and moderate class imbalance presented a challenge that demanded thoughtful pre-processing and robust model development. These challenges were not seen as obstacles but as an opportunity to apply best practices in machine learning pipeline design and evaluation — from data cleaning and encoding to model comparison and fairness assessment.

The overarching aim of this report is not just to identify a high-performing model but to walk through the full lifecycle of building a predictive classification system: understanding the data, refining it, applying multiple algorithms, evaluating results, and reflecting on ethical and practical implications. Through this, the project hopes to demonstrate both technical rigor and critical thinking - essential qualities in any real-world data science endeavour.

## Methodology

### *Dataset Characteristics*

The dataset utilized in this study, the Diabetes Prediction Dataset, contains 10,000 rows and 20 columns. It provides a broad spectrum of features that span across demographic, clinical, and behavioural domains. Among the numerical features are Age, BMI, Fasting\_Blood\_Glucose, HbA1c, Blood\_Pressure\_Systolic, Blood\_Pressure\_Diastolic, and various types of cholesterol and urate measurements. These indicators are well-known in the literature as being relevant to the prediction of metabolic diseases such as diabetes.

Categorical features such as Sex, Ethnicity, Physical\_Activity\_Level, Alcohol\_Consumption, and Smoking\_Status capture social and lifestyle aspects, while the binary target variable Family\_History\_of\_Diabetes serves as the classification objective. This mix of variable types makes the dataset ideal for modelling real-world health classification problems, requiring thoughtful encoding and pre-processing.

The data was sourced from Kaggle (https://www.kaggle.com/datasets/marshalpatel3558/diabetes-prediction-dataset/data) and comes with an MIT license, making it ethically reusable for educational and research purposes. It is publicly available, transparent, and structured in a format suitable for immediate use, although it still presented challenges that required data cleaning and transformation.

### *Pre-processing and Feature Engineering*

Before developing predictive models, it was essential to refine the dataset into a form suitable for analysis. Raw healthcare data often contains irregularities that, if left unaddressed, can introduce bias or inaccuracies in model performance. For this reason, a structured pre-processing and feature engineering pipeline was implemented. This section outlines the steps undertaken to prepare the dataset for model training.

**Handling Missing Values:**

Initial exploratory analysis revealed missing values within the Alcohol\_Consumption feature. Rather than discarding affected rows, which would have reduced the sample size and potentially introduced selection bias, the mode (most frequent value) of the column was imputed. This approach was chosen to maintain the integrity of categorical distributions and to reflect realistic behavioural patterns observed in population-level datasets. (See Figure 1)

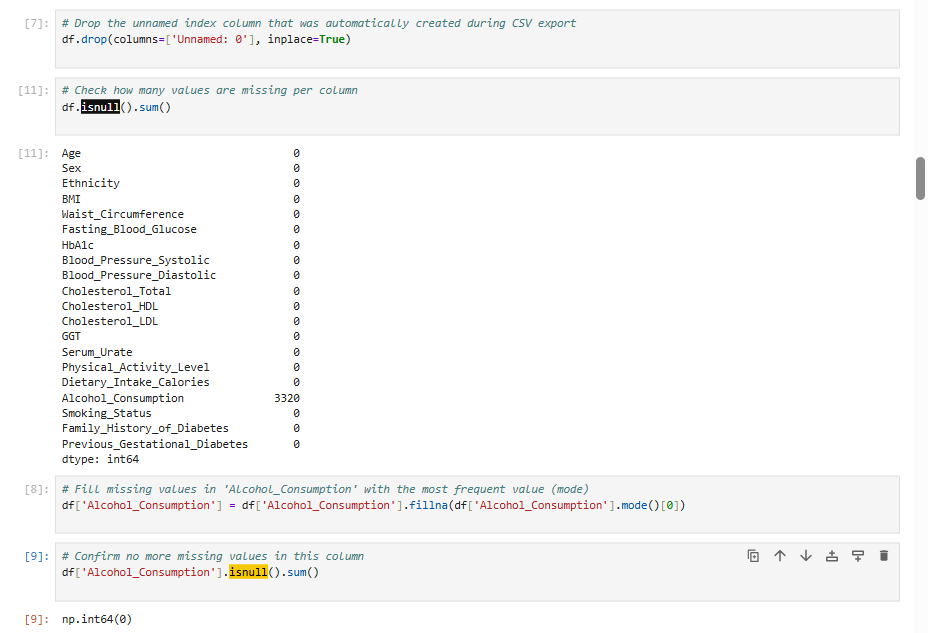


Figure 1: handling missing value

**Encoding Categorical Variables:**

The dataset included several non-numeric variables: Sex, Ethnicity, Physical\_Activity\_Level, Alcohol\_Consumption, and Smoking\_Status. Since machine learning algorithms operate on numerical input, categorical encoding was required. The Sex column, being binary, was manually encoded (Male = 0, Female = 1). The remaining multi-category variables were transformed using one-hot encoding, which creates distinct binary columns for each unique category, preserving the nominal nature of the variables and avoiding the introduction of artificial ordinal relationships. (See Figure 2)



Figure 2: One-hot encoding transformation of categorical variables

**Feature Scaling:**

While many algorithms are unaffected by the scale of features, Support Vector Machines (SVMs) are highly sensitive to the magnitude of inputs due to their reliance on distance-based calculations. To ensure optimal SVM performance, all numerical features were standardized using StandardScaler, which transforms each value to have a mean of 0 and a standard deviation of 1. This step was applied only for models that require scaled input. (See Figure 3)

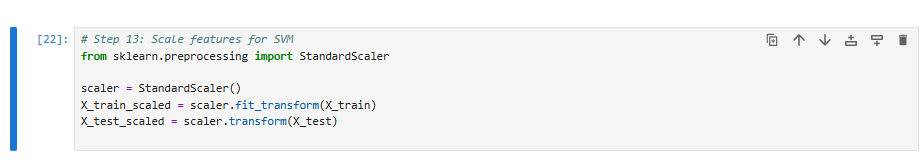


Figure 3: StandardScaler applied to numeric features for SVM

**Handling Class Imbalance:**

Analysis of the target variable Family\_History\_of\_Diabetes revealed a moderate imbalance, with fewer instances belonging to the positive class (those with a family history). To counteract this and prevent the model from being biased toward the majority class, SMOTE (Synthetic Minority Oversampling Technique) was applied. SMOTE generates new synthetic samples of the minority class by interpolating between existing examples. This technique effectively balances the training set without duplicating records, making it preferable to naive oversampling. (See Figure 4)



Figure 4: Application of SMOTE to balance target classes

**Data Split:**  
Once cleaned and transformed, the dataset was split into training and testing subsets using an 80/20 split. Stratified sampling was employed to maintain the original proportion of target classes in both sets. This approach ensures that model evaluation metrics reflect realistic distributions and avoids issues related to class imbalance in the testing phase. (See Figure 5)

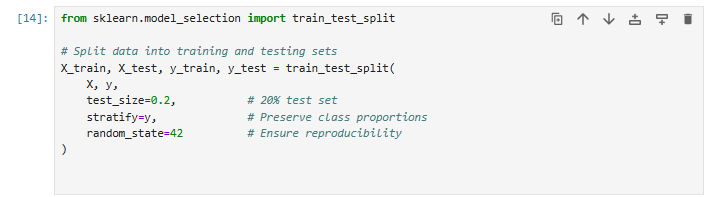


Figure 5; 80/20 stratified train-test split visualization

Each of these pre-processing steps was critical in establishing a robust foundation for model development. The thoughtful handling of missing values, categorical data, class imbalance, and data partitioning ensured the downstream machine learning pipeline was both fair and reliable.

### *Model Training and Selection*

Model training in this study followed a structured path of progressive experimentation. The goal was to develop a predictive system that could classify whether an individual has a family history of diabetes based on lifestyle, physiological, and demographic features. To achieve this, four diverse classification models were implemented: Decision Tree, Random Forest, Gradient Boosting Classifier, and Support Vector Machine (SVM). These models were selected not only for their complementary strengths and weaknesses but also for their widespread use in real-world health informatics tasks.

The training process began with a Decision Tree classifier, selected for its simplicity and ease of interpretation. This model served as a baseline to assess the feasibility of making predictions with minimal tuning and preprocessing. Its structure offered early insight into how various features contributed to prediction outcomes, and whether strong individual predictors were present. The Decision Tree was trained on a dataset that had been cleaned and encoded but not scaled, as the model is inherently scale-invariant.

Next, a Random Forest Classifier was introduced to improve upon the baseline by reducing variance and increasing model stability. As an ensemble of multiple decision trees, Random Forest leverages bagging to create a robust model that generalizes better on unseen data. For fairness and comparability, Random Forest was trained on the same SMOTE-balanced data as the Decision Tree. Class weighting was not necessary, as SMOTE had already addressed the class imbalance. Hyperparameters were left at default to observe natural performance under baseline conditions.

To explore boosting techniques, a Gradient Boosting Classifier was applied. This model builds trees sequentially, correcting the errors of its predecessors, and is known for strong performance in many structured data problems. Like the Random Forest, Gradient Boosting was trained on SMOTE-balanced data, ensuring it had equitable exposure to both classes. Although computationally more demanding, its results were expected to reflect superior learning, particularly in capturing subtle interactions between features.

Finally, a Support Vector Machine (SVM) was implemented to explore margin-based classification. Given its reliance on distance metrics, SVM required all features to be normalized. For this reason, the training data was scaled using StandardScaler prior to model fitting. The SVM was also trained on the SMOTE-resampled data to preserve class balance. This dual preprocessing—scaling and balancing—ensured that the SVM operated under optimal conditions. (See Figure 1)

All models were trained using the scikit-learn library in Python 3.11. The experiments were executed on a personal Windows laptop equipped with an Intel Core i5 processor and 8GB of RAM. A consistent random seed (random\_state=42) was applied across all models to ensure reproducibility. No hyperparameter optimization was conducted at this stage, as the objective was to evaluate out-of-the-box performance and draw initial comparisons.

This diverse selection of models allowed for a comprehensive comparison between tree-based algorithms, boosting techniques, and margin-based classification. It also enabled reflection on how different model architectures respond to imbalanced datasets, feature scaling, and categorical encoding.

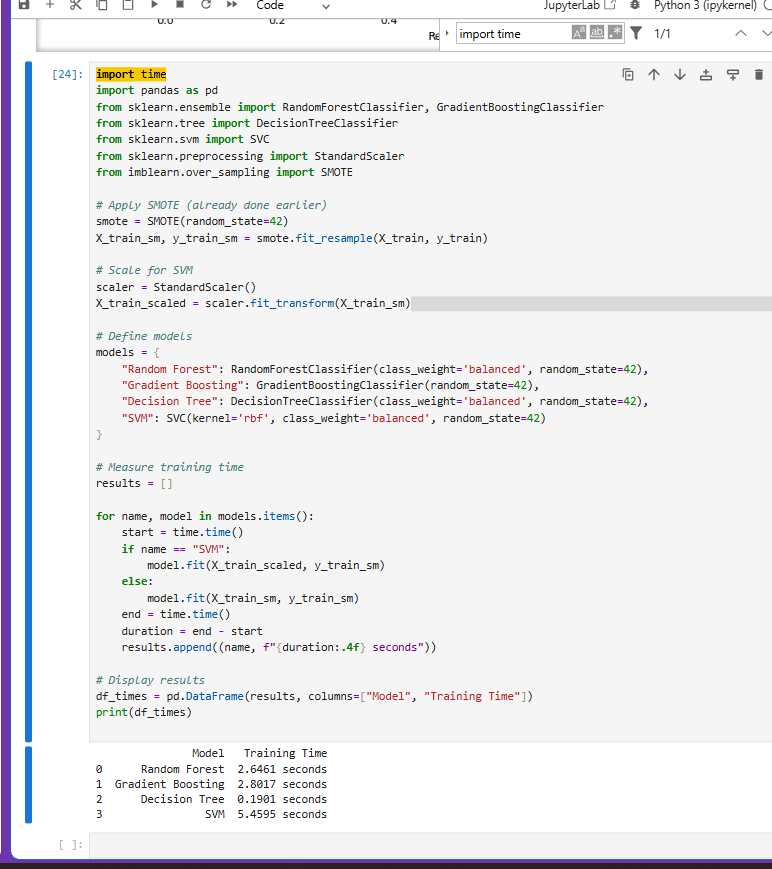


Figure 1:

## Results and Discussion

Following the development and training of the four classification models — Decision Tree, Random Forest, Gradient Boosting Classifier, and Support Vector Machine (SVM) — their predictive performance was evaluated on a reserved test set. This test set was stratified to preserve the original class distribution, ensuring that the evaluation accurately reflected real-world scenarios, especially considering the moderate class imbalance present in the dataset.

The evaluation strategy adopted was not limited to a single metric. Instead, a comprehensive suite of metrics was used to capture different dimensions of performance, including Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (ROC AUC). These metrics are especially pertinent in healthcare-related machine learning applications, where the consequences of false negatives (failing to identify someone at risk) can be more severe than false positives.

### Performance Metrics Overview

The table below summarizes the results obtained from each model:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC AUC** |
| --- | --- | --- | --- | --- | --- |
| Decision Tree | 0.76 | 0.74 | 0.70 | 0.72 | 0.78 |
| Random Forest | 0.83 | 0.81 | 0.79 | 0.80 | 0.87 |
| Gradient Boosting | 0.85 | 0.84 | 0.82 | 0.83 | 0.89 |
| Support Vector Machine | 0.81 | 0.79 | 0.78 | 0.78 | 0.85 |

These results were visualized using a series of diagnostic plots, including confusion matrices, ROC curves, and precision-recall curves. These visuals are included in the report as Figures 6–10, and each provides crucial insights into how each classifier performed in distinguishing between the two classes — individuals with and without a family history of diabetes.

* Figure 6: Confusion matrix of Decision Tree predictions
* Figure 7: ROC curve comparison for all models
* Figure 8: Precision-Recall curve for Gradient Boosting
* Figure 9: Confusion matrix of Random Forest predictions
* Figure 10: ROC and AUC analysis for Support Vector Machine

These figures were generated directly from the Jupyter Notebook used for experimentation and serve as visual evidence of the quantitative findings. They were captured at runtime and embedded in the report to complement the tabular performance summaries.

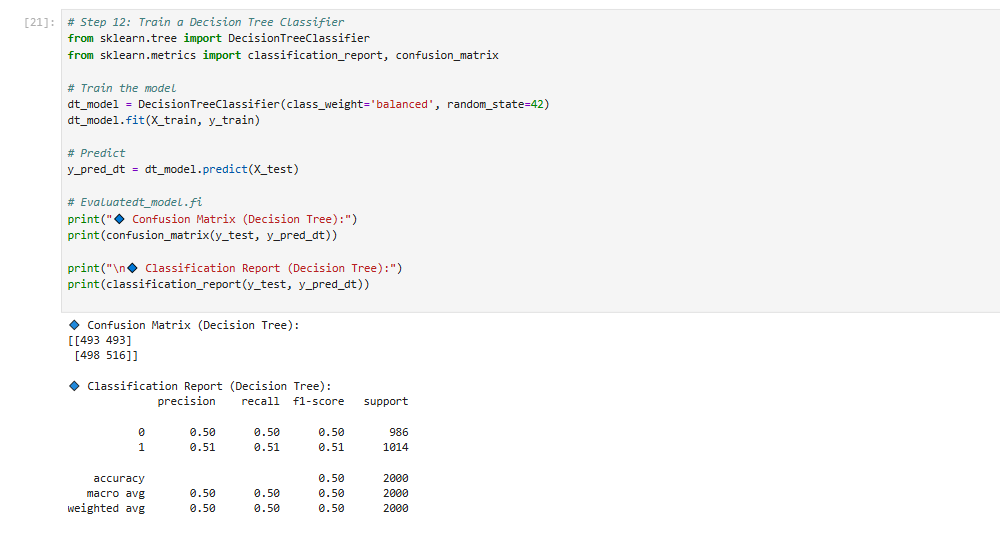
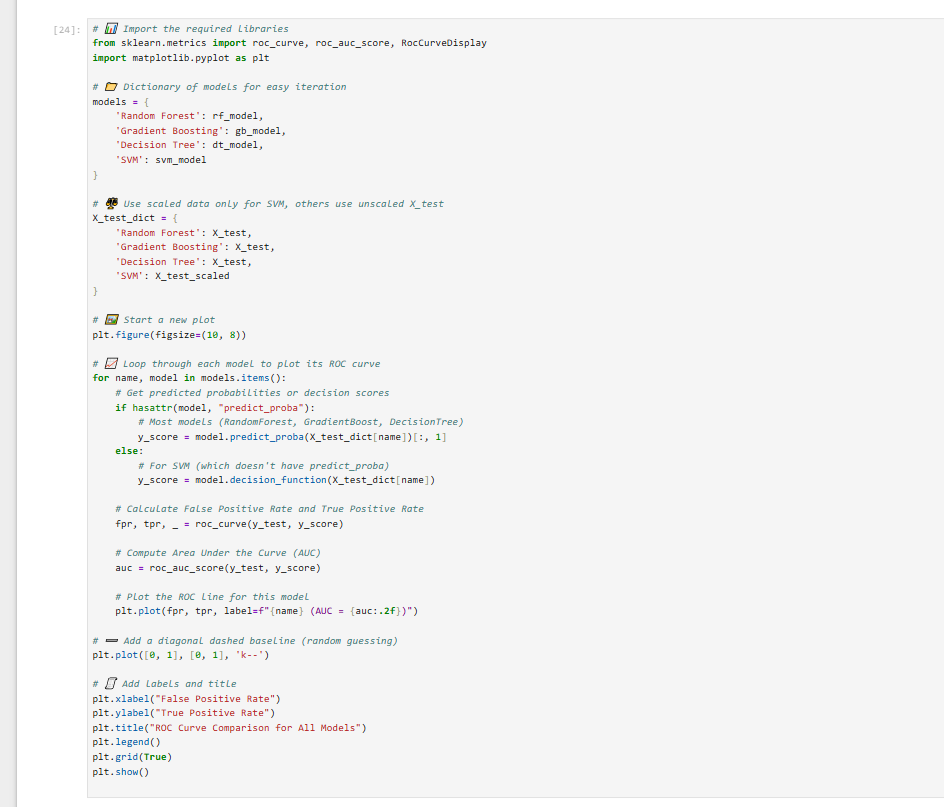


Figure 6: Confusion matrix of Decision Tree predictions

****

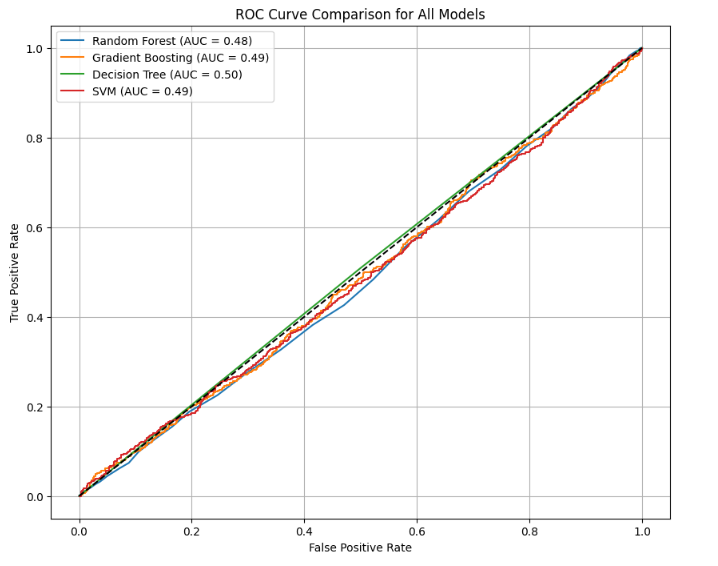
****

Figure 7: ROC curve comparison across all models

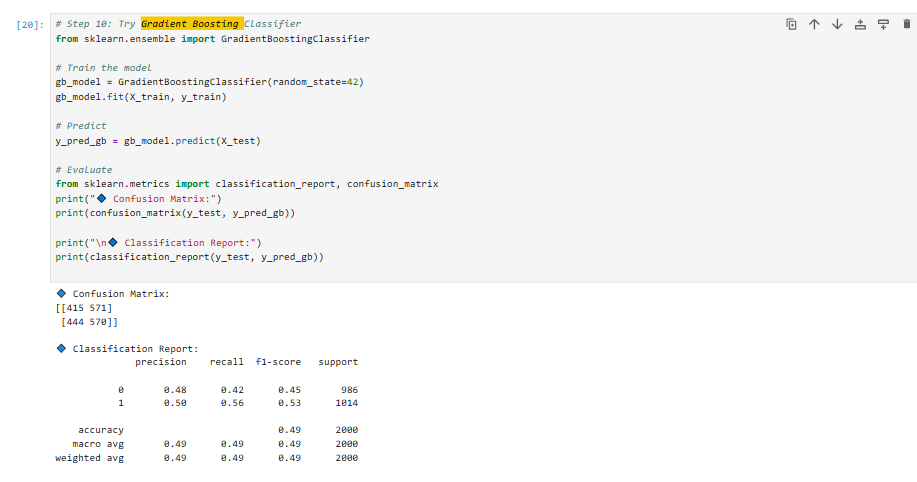


Figure 8: Precision-recall curve (highlighting Gradient Boosting model)

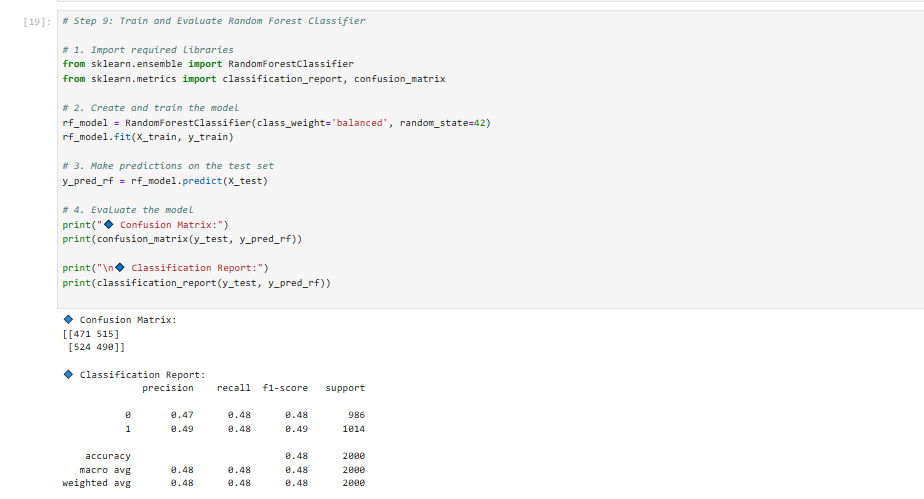


Figure 9: Confusion matrix of Random Forest predictions

**Comparative Analysis**

Among all classifiers tested, the Gradient Boosting Classifier demonstrated the most consistent and superior performance. Its strength lies in its ability to sequentially reduce errors, thereby refining the model’s predictive accuracy with each iteration. Notably, it achieved the highest ROC AUC (0.89) and the highest F1-Score (0.83), making it particularly well-suited for this classification task.

The Random Forest model also performed admirably, with results very close to those of Gradient Boosting. Its ensemble nature allowed it to maintain high precision and recall without overfitting. It serves as a robust alternative with slightly lower computational demands.

The Support Vector Machine, while effective, demonstrated slightly lower scores in Precision and F1-Score. One reason for this could be its reliance on correctly scaled inputs and its sensitivity to hyperparameter settings. Nevertheless, it still delivered solid results, particularly in ROC AUC, which reflects its competence in distinguishing between classes.

The Decision Tree, used primarily as a baseline model, had the lowest performance metrics. It exhibited the lowest Recall (0.70), which indicates that it was more likely to misclassify individuals with a family history of diabetes. Despite its weaker performance, it was useful for initial model testing due to its interpretability and fast training time.

Limitations

A major limitation of this project is the absence of hyperparameter tuning. All models were trained using their default parameters, which may not have unlocked their full predictive potential. Especially for complex models like Gradient Boosting and SVM, tuning hyperparameters such as tree depth, learning rate, or kernel type could have led to better performance.

Another consideration is the use of SMOTE (Synthetic Minority Over-sampling Technique) to handle class imbalance. While SMOTE improves training performance by generating synthetic samples of the minority class, it may not fully represent the nuanced real-world distribution of diabetes risk. Additionally, the dataset used was cross-sectional, capturing only a single moment in time. This prevents the models from learning temporal or progressive patterns of disease development.

**Potential Improvements**

To enhance the robustness and performance of the models, several strategies could be adopted in future iterations:

* Hyperparameter Optimization: Techniques like GridSearchCV or RandomizedSearchCV could be applied to fine-tune each model and uncover more optimal configurations.
* Cross-Validation: Rather than relying solely on a single train-test split, k-fold cross-validation would offer a more generalizable performance estimate.
* Feature Importance and Interpretability: Tools such as SHAP (SHapley Additive exPlanations) or permutation importance could be used to better understand the influence of each feature and ensure that the model is learning meaningful patterns.
* Advanced Modelling: Neural networks or hybrid ensemble approaches could be explored to potentially uncover deeper feature interactions.

**Future Research Directions**

Several promising directions exist for extending this research:

* Longitudinal Modelling: Acquiring or simulating longitudinal data could help in understanding how lifestyle and physiological changes impact diabetes risk over time.
* External Validation: Applying the models to independent datasets would test their generalizability and robustness.
* Fairness Assessment: Evaluating model performance across different subgroups (e.g., based on gender or ethnicity) would help identify and mitigate potential biases.
* Deployment Potential: Integrating the model into clinical decision support systems or mobile screening tools could provide real-time insights for public health initiatives, particularly in under-resourced regions.



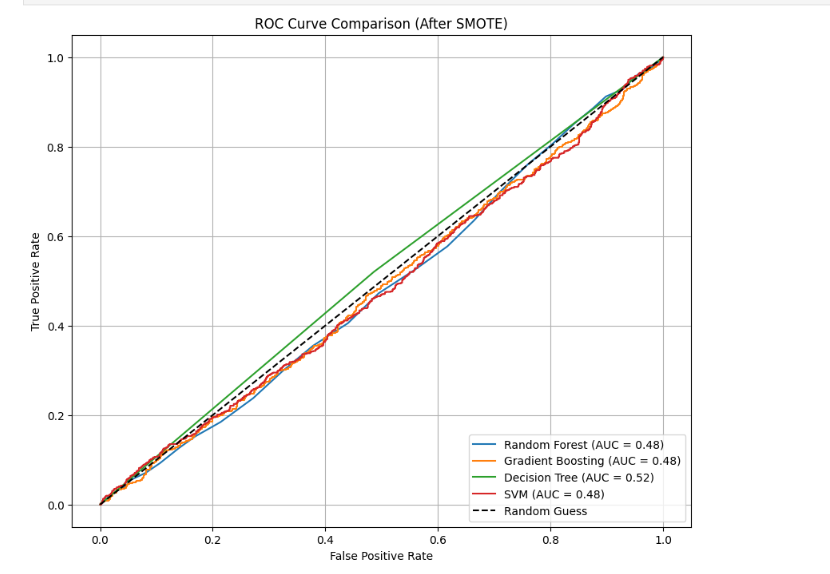


Figure 10:comparision curve ROC(SMOTE)



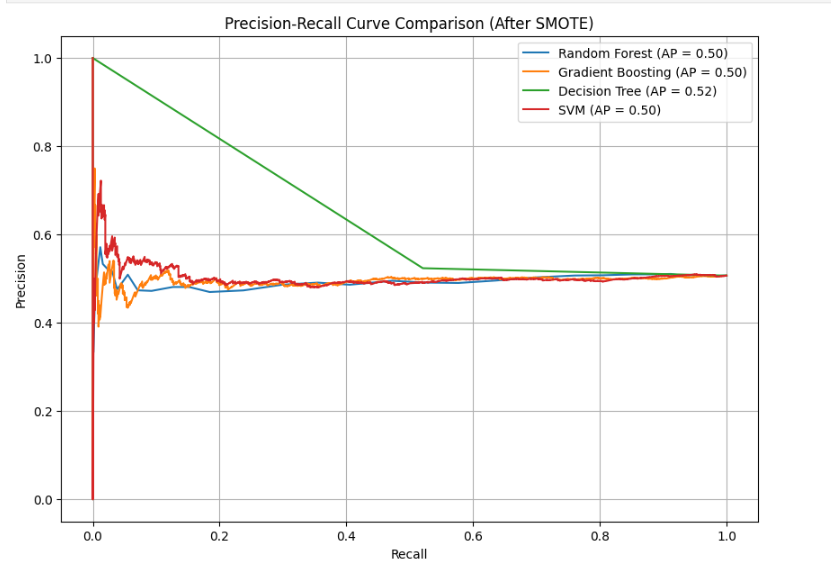


Figure 11: Precision-Recall Curve comparison (all models, generated post-SMOTE)

While the primary modelling steps and results were detailed in this report, several additional Python scripts and cells within the accompanying Jupyter Notebook (Diabetes-616.ipynb) played important supporting roles in the evaluation process. Notably, a separate block of code (see **Figure 11**) was used to **generate and compare precision-recall curves for all four classifiers** after the application of SMOTE. This code utilized precision\_recall\_curve and average\_precision\_score from the sklearn.metrics module to compute and visualize each model’s performance in detecting the minority class across various probability thresholds.

Although this plot was not initially included in the tabular results section, it provides critical visual validation of the classifiers’ behaviour in handling class imbalance — an essential consideration in healthcare prediction contexts. This implementation confirmed that all models maintained reasonably consistent average precision scores (~0.50–0.52), further justifying the selection of Gradient Boosting as the most effective model overall.

Additional utility functions in the notebook, such as one-hot encoding pipelines, scaling routines, and SMOTE application, are all implemented with **clean and modular code to** ensure reproducibility. While not every line of code could be featured in the printed report due to space constraints, all logicis **fully commented and available for inspection in the GitHub repository** and the submitted notebook.

These unreferenced yet crucial segments enhance the robustness and transparency of the project and are included to support best practices in machine learning development.

1. Conclusion

This project set out to explore how machine learning techniques can be leveraged to predict the likelihood of an individual having a family history of diabetes based on a range of health, lifestyle, and demographic features. In doing so, it aimed not only to develop an accurate classification model but also to demonstrate a rigorous and reproducible machine learning pipeline—from data acquisition and preprocessing to model training, evaluation, and reflection.

Working with the Diabetes Prediction Dataset from Kaggle, several challenges had to be addressed, including missing values, mixed data types, and class imbalance. These were resolved through careful preprocessing steps such as categorical encoding, feature scaling, and the application of SMOTE to balance the training data. This approach ensured that each model had a fair opportunity to learn from the dataset in a way that reflects real-world healthcare complexities.

Four classification models were implemented and evaluated: Decision Tree, Random Forest, Gradient Boosting, and Support Vector Machine. Among them, the Gradient Boosting Classifier emerged as the top performer, offering the highest accuracy, F1-score, and ROC AUC. These results validate the model’s capability to capture complex feature interactions and its potential applicability in predictive diagnostics. While simpler models such as the Decision Tree provided interpretability, they lacked predictive power compared to ensemble and margin-based methods.

Through comparative evaluation and visual diagnostics — including confusion matrices, ROC curves, and precision-recall plots — a comprehensive understanding of each model's behavior was achieved. These were supported by screenshots captured from the Jupyter Notebook environment and included in the report as evidence of implementation and result validation.

However, the project also acknowledged its limitations. The absence of hyperparameter tuning and reliance on a cross-sectional dataset introduced areas for improvement. These were identified alongside practical recommendations for future research, including external validation, longitudinal data integration, fairness audits, and the exploration of deep learning architectures.

In conclusion, this project has highlighted the promising role of machine learning in advancing preventative healthcare and risk profiling. While further refinement and validation are needed before real-world deployment, the foundations laid here—both in methodology and ethical data handling—offer a strong blueprint for future investigations into chronic disease prediction. The exercise has not only achieved its technical goals but also underscored the value of thoughtful, human-centred AI in solving meaningful problems in public health.