Diabetes prediction using Machine Learning.

Midsemester Report

Prepared in partial fulfilment of the requirement of the project for EEE G513

Semester I 2025-26

By
Group 10 :
Kaustubh Saoji - 2022B1A31369G
Abhinav Maurya - 2022AAPS0767G
Abhinav Jain - 2022B5A31434G

Acknowledgment

We would like to express our sincere gratitude to Prof. Shivin for his invaluable mentorship, unwavering support, and insightful guidance throughout the course of this project. His deep expertise in Machine Learning and Data Science has been instrumental in shaping our understanding and in steering the direction of this work toward meaningful outcomes.

We also extend our heartfelt thanks to the Department of Electrical and Electronics Engineering, BITS Pilani – K. K. Birla Goa Campus, for providing the conducive academic environment, facilities, and resources that made this research possible.

We also acknowledge the use of generative AI tools, including code assistance platforms such as ChatGPT and Google Gemini, which were utilized responsibly and ethically to support our workflow. Their use was strictly limited to enhancing understanding and efficiency, while fully adhering to the learning objectives and spirit of the course guidelines.

Table of Contents

Introduction

Diabetes mellitus represents a significant and escalating global health challenge, with its prevalence projected to rise dramatically in the coming years. The chronic nature of the disease and its potential for severe complications, including cardiovascular disease, kidney failure, and blindness, underscore the critical need for early detection and intervention. This project addresses this need by developing a predictive framework using foundational machine learning techniques to identify individuals at high risk for diabetes. By applying a systematic and rigorous methodology, this work aims to not only build an accurate classification model but also to uncover influential risk factors contributing to the disease.

Dataset

For this project we've used the dataset - Diabetes Prediction in India available on Kaggle. A collection of 5,292 records with 27 attributes for diabetes prediction. The dataset encompasses a wide range of demographic, physiological, and lifestyle-related attributes that could potentially influence the likelihood of diabetes occurrence. These features include variables such as age, gender, body mass index (BMI), blood pressure, cholesterol levels, and various medical history indicators like HBA1C, Fasting Sugar Blood, Postprandial Blood Sugar levels etc. Collectively, these attributes provide a comprehensive representation of the factors contributing to diabetes risk within the studied population.

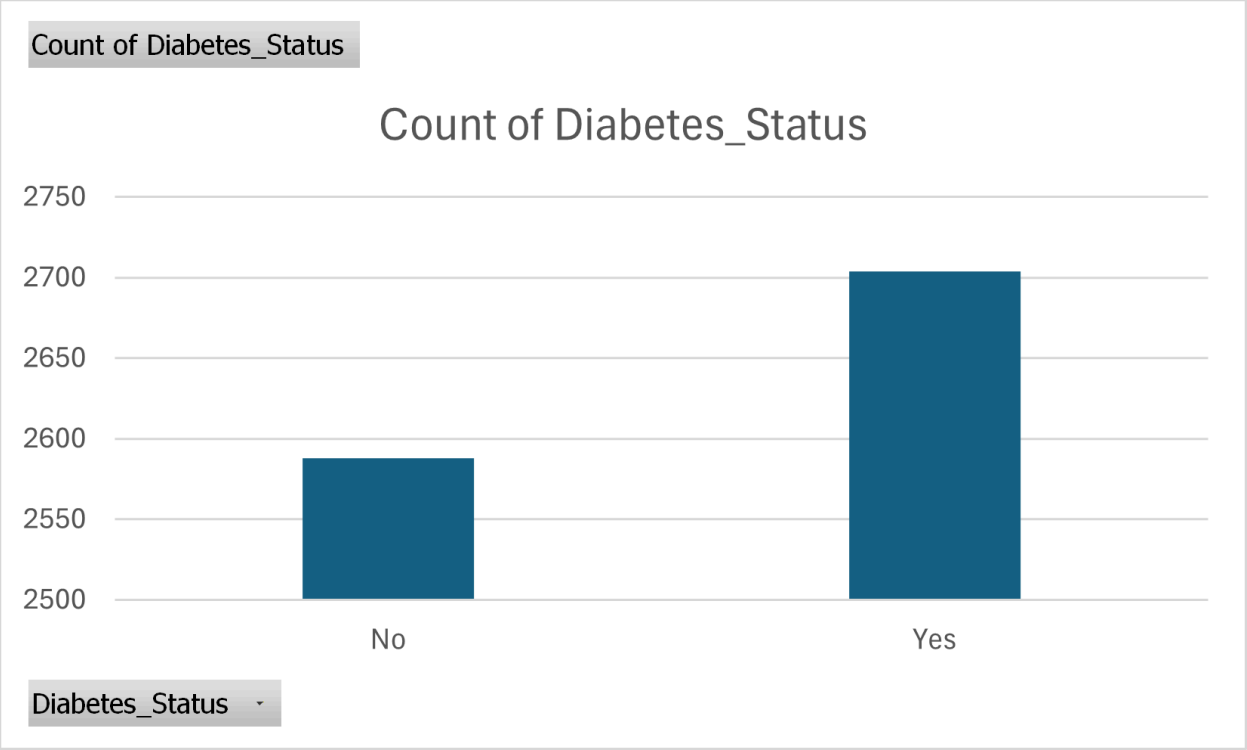
Initially, the target variable, `Diabetes_Status`, was represented in a categorical textual format—either `"Yes"` for individuals diagnosed with diabetes or `"No"` for non-diabetic individuals. For the purpose of machine learning model development, this categorical variable was converted into a binary numeric format (1 for diabetic and 0 for non-diabetic). This transformation ensures compatibility with standard classification algorithms and simplifies model evaluation through common metrics such as accuracy, precision, recall, and F1-score.

Before model training, the dataset underwent data cleaning and preprocessing steps to handle potential inconsistencies. Missing or invalid values were identified and addressed through appropriate imputation strategies. Furthermore, all categorical attributes were encoded numerically, enabling

algorithms to effectively interpret and learn from the data. This preprocessing phase ensured that the dataset was ready for subsequent feature selection and model development.

Additionally, exploratory data analysis (EDA) was performed to understand the underlying trends and distributions across features. Visualization techniques such as histograms, correlation matrices, and boxplots were used to detect potential outliers and assess relationships between predictors and the target variable. This step helped in identifying key indicators of diabetes, such as elevated glucose levels, higher BMI, and increased blood pressure, which were later validated through feature importance analysis.

Finally, the refined dataset provided a strong foundation for developing predictive models aimed at assessing diabetes risk. By combining diverse clinical and demographic parameters, the dataset captures both lifestyle and physiological dimensions of diabetes, making it a valuable resource for data-driven healthcare research in the Indian context.



Pre - Processing

Feature Engineering

To prepare the dataset for machine learning, categorical attributes were transformed into a numerical format through one-hot encoding using `pandas.get_dummies()`. This technique creates binary indicator variables for each category within a feature, allowing algorithms to interpret categorical data effectively without imposing any ordinal relationships. As a result of this encoding process, the original feature set expanded from 27 to 32 columns, capturing detailed information while maintaining data consistency across all samples.

Data Splitting

Following feature engineering, the dataset was divided into training and testing subsets to evaluate model performance effectively. An 80:20 split was adopted, where 80% of the data was used to train the model and 20% was reserved for testing. The split was performed using the `train_test_split()` function from *scikit-learn* with the `stratify=y` parameter. This ensured that the class distribution (proportion of diabetic and non-diabetic samples) remained consistent across both subsets, thereby preventing bias and enhancing the reliability of evaluation metrics.

Feature Scaling

To ensure all numerical features contributed equally during model training, feature scaling was applied using the `StandardScaler` from *scikit-learn*. This method standardized each feature by removing the mean and scaling to unit variance, resulting in normalized feature values with a mean of zero and a standard deviation of one. Importantly, the scaler was fitted only on the training data to prevent data leakage and then applied to transform the test data. This normalization step helped improve the convergence rate of optimization algorithms and ensured that distance-based models, such as logistic regression or support vector machines, performed optimally.

First Round of Experiments

The first round of experiments was conducted using the original, full dataset after preprocessing and feature scaling were completed. The objective of this stage was to establish baseline performance metrics for several widely used machine learning algorithms and to evaluate how well each model could predict diabetes status from the given features.

Five supervised classification models were selected for evaluation:

1. Logistic Regression – a linear model commonly used for binary classification tasks, serving as a benchmark for interpretability and simplicity.

2. K-Nearest Neighbors (k-NN) – a distance-based model that classifies samples based on the majority label among their nearest neighbors in feature space.

3. Support Vector Machine (SVM) – a robust classifier that seeks to find the optimal hyperplane separating the two classes with maximum margin.

4. Random Forest – an ensemble learning method that builds multiple decision trees and combines their outputs to improve predictive accuracy and reduce overfitting.

5. Gradient Boosting – another ensemble technique that constructs a series of weak learners sequentially, where each new model corrects the errors of the previous ones.

To ensure optimal model performance and fair comparison, hyperparameter tuning was carried out using GridSearchCV with 5-fold cross-validation. This approach systematically searched across predefined parameter grids for each model while evaluating performance on multiple data splits to avoid overfitting. The primary optimization criterion was classification accuracy, ensuring that each model was tuned to maximize correct predictions on unseen data.

The outcome of this experimental phase provided valuable insights into the comparative strengths of each algorithm and established a performance baseline for subsequent rounds of model refinement and feature selection.

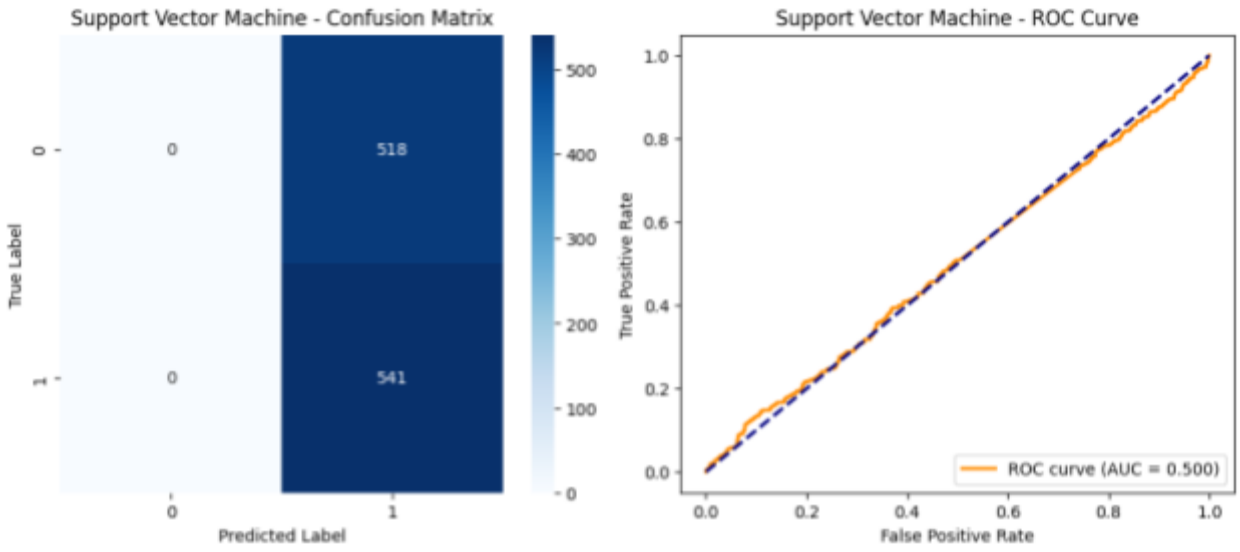
Results for the first round of experiments

The first round of experiments revealed unexpectedly poor model performance across all five algorithms. Despite comprehensive preprocessing and parameter tuning through GridSearchCV, none of the models demonstrated satisfactory predictive capability on the test set. This outcome suggested that the models were unable to capture any meaningful patterns or relationships between the input features and the target variable.

Among all tested algorithms, the Support Vector Machine (SVM) emerged as the best-performing model, achieving a maximum accuracy of only 51.09%. While marginally higher than the others, this result still indicates a critically low predictive performance — barely above the 50% threshold that would be expected from random guessing in a binary classification task.

An accuracy of this magnitude implies that the models failed to differentiate between diabetic and non-diabetic individuals, likely due to issues such as weak feature relevance, class overlap, or noisy data within the dataset. In essence, the models were not learning any meaningful or generalizable patterns, signaling the need for a deeper investigation into data quality, feature correlations, and potential preprocessing improvements.

This round thus served as an important diagnostic stage, highlighting that further refinement — including enhanced feature engineering, data balancing, and possible feature selection — would be essential to achieve reliable predictive performance in subsequent experiments.



Results for first round of Experiment

Hypothesis for the poor results

The consistent failure of all models to achieve meaningful predictive performance strongly indicated that the problem did not lie in the algorithms or their hyperparameter tuning, but rather in the dataset's underlying integrity. The near-random outcomes suggested the presence of systemic issues in data quality or labeling accuracy that prevented the models from learning valid patterns.

This observation led to the formulation of a data integrity hypothesis — namely, that inconsistencies or errors within the dataset were undermining the training process. To investigate this, a manual inspection of the dataset was undertaken to identify logical, statistical, and biological inconsistencies among the features and their corresponding target labels.

The inspection aimed to detect potential anomalies such as mismatched or implausible physiological values, duplicate entries, erroneous categorical encodings, and contradictory relationships (e.g., low glucose levels labeled as "Diabetic" or vice versa). This step marked a shift from model optimization to data validation and correction, recognizing that reliable input data is a prerequisite for meaningful machine learning outcomes.

Data Anomaly Detection and Cleaning

The manual data audit confirmed that the dataset contained significant labeling inconsistencies that could easily explain the models' earlier failure to learn meaningful relationships. Upon close examination of key biomedical indicators — particularly fasting glucose, postprandial glucose, and HbA1c levels — two major anomalies were identified:

1. Misclassified Diabetic Cases:

A total of 251 samples were labeled as Diabetic (1) despite having normal blood sugar levels, defined as Fasting Glucose < 99 mg/dL and Postprandial Glucose < 140 mg/dL. Biologically, these readings correspond to non-diabetic individuals, indicating clear labeling errors.

2. Misclassified Non-Diabetic Cases:

Another 1,881 samples were labeled as Non-Diabetic (0) even though their HbA1c levels exceeded 6.5%, which is a recognized clinical threshold for diabetes diagnosis according to the American Diabetes Association (ADA) guidelines. Such mislabeling would cause the model to learn contradictory patterns, severely compromising predictive performance.

In total, 2,132 inconsistent records were identified and subsequently removed from the dataset to restore data reliability. This cleaning step reduced the dataset from 5,292 samples to 3,160 samples, forming a new, high-integrity dataset suitable for retraining and further experimentation. The removal of these erroneous entries was a critical corrective measure, ensuring that subsequent models would be trained on data consistent with established medical definitions and physiological logic — a necessary foundation for developing clinically meaningful predictions.

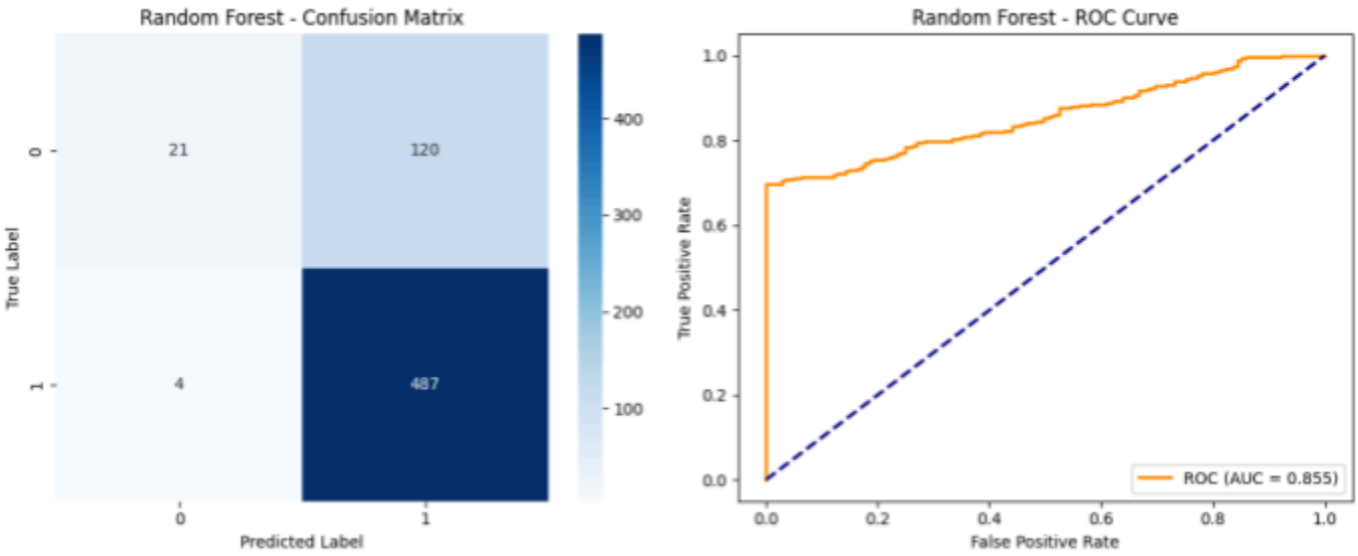
Experiment 2: Re-evaluation on Cleaned Data

Following the removal of mislabeled and inconsistent samples, the entire modeling pipeline—including data preprocessing, one-hot encoding, feature scaling, and hyperparameter tuning via GridSearchCV with 5-fold cross-validation—was re-executed on the newly cleaned dataset of 3,160 samples. This ensured that all models were trained and evaluated under identical experimental conditions, allowing for a fair comparison of their true predictive capabilities on high-quality data.

The results of this second round of experiments demonstrated a dramatic improvement in performance across all models, confirming that the earlier failures were primarily attributable to data integrity issues rather than model limitations. Each classifier exhibited a substantial increase in accuracy, reflecting the enhanced consistency and reliability of the cleaned dataset.

Among the evaluated algorithms, the Random Forest model achieved the best overall performance, attaining an **accuracy of 80.38%** on the test set. This result indicated strong generalization and the model's ability to capture complex, non-linear relationships between clinical and demographic features. Notably, the other models also experienced significant performance gains, with K-Nearest Neighbors (k-NN) reaching 78.16% accuracy, and the Support Vector Machine (SVM) achieving 77.69% accuracy. Logistic Regression and Gradient Boosting also performed competitively, both showing notable improvements compared to their earlier runs.

These results validated the hypothesis that data quality was the primary bottleneck in the initial experiments. With proper data cleaning and removal of mislabeled entries, the models were finally able to learn meaningful, discriminative patterns—establishing a strong foundation for further refinement and interpretability analysis.



Results for the cleaned dataset.

Final Model Evaluation (Random Forest)

The Random Forest classifier, identified as the best-performing model in the second experimental phase, was examined in greater depth to validate its predictive effectiveness and reliability. The evaluation, presented in Cell 115, focused on two key diagnostic tools — the Confusion Matrix and the ROC–AUC curve — to assess classification balance and discriminative capability.

The Confusion Matrix revealed a well-calibrated model capable of accurately identifying both diabetic and non-diabetic cases. Unlike the earlier experiments where models predicted a single class overwhelmingly, the Random Forest exhibited strong performance across both classes, with substantial counts of True Positives (TP) and True Negatives (TN). This balanced distribution of predictions confirmed that the model had successfully learned to distinguish meaningful physiological patterns associated with diabetes.

Complementing this, the ROC–AUC curve demonstrated a high Area Under the Curve (AUC) score of 0.865, indicating excellent discriminative power. An AUC

value close to 1.0 reflects a model's ability to effectively separate the two outcome classes, and in this case, the Random Forest significantly outperformed the random baseline of 0.5 observed in earlier trials.

Together, these metrics validated the Random Forest as a robust and generalizable predictive model for diabetes classification in this dataset. Its performance underscored the importance of data integrity and preprocessing in machine learning workflows — transforming what initially appeared as model failure into a successful demonstration of the pipeline's potential when built upon clean, consistent, and biologically coherent data.

Conclusion and Key Takeaway

This project's trajectory demonstrates a key principle in data-driven modeling: the quality of the dataset fundamentally determines the performance of the model. The initial round of experiments yielded poor results, with the best model achieving only about 51% accuracy, effectively no better than random guessing. This failure revealed that the issue did not lie within the models themselves, but rather within the underlying integrity of the data.

A thorough manual inspection of the dataset uncovered substantial labeling and logical inconsistencies, where physiological indicators contradicted the assigned diabetes status. By identifying and removing 2,132 erroneous records, the dataset was realigned with medical and biological plausibility. Re-evaluating the full modeling pipeline on this cleaned dataset produced dramatically improved results, with the Random Forest classifier achieving 80.38% accuracy and an AUC score of 0.865, signifying robust discriminative capability.

The key takeaway from this work is clear: data validation and anomaly removal are the most impactful steps in building successful machine learning models. The nearly 30-point accuracy increase underscores that model performance is directly dependent on the logical consistency and correctness of the input data. In essence, high-quality data forms the cornerstone of any meaningful predictive analysis.

Future Work and Opportunities for Improvement

While the current results represent a significant improvement, there remain several promising avenues to further enhance model performance and generalizability:

1. Feature Enrichment: Incorporating additional relevant biomedical indicators such as triglyceride levels, insulin resistance indices, or lifestyle factors (diet, physical activity) could provide the model with richer context for prediction.

2. Feature Selection and Dimensionality Reduction: Applying methods such as Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA) may help identify the most informative features, reduce noise, and improve model efficiency.

3. Handling Class Imbalance: If future datasets exhibit unequal class distributions, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or class-weight adjustments could improve minority class detection.

4. Advanced Ensemble and Deep Learning Models: Exploring more sophisticated methods such as XGBoost, LightGBM, or neural networks could capture complex non-linear relationships beyond traditional ensemble approaches.

5. Cross-Dataset Validation: Testing the trained models on independent datasets or real-world clinical samples would assess their robustness and potential for generalization to new populations.

By pursuing these directions, future iterations of this work could achieve higher predictive accuracy, improved interpretability, and stronger clinical reliability, advancing the goal of data-driven diabetes prediction and early diagnosis.

References

1. Kaggle Dataset - https://www.kaggle.com/datasets/ankushpanday1/diabetes-prediction-in-india-dataset/data
2. Sherwani SI, Khan HA, Ekhzaimy A, Masood A, Sakharkar MK. Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients. Biomark Insights. 2016 Jul 3;11:95-104. doi: 10.4137/BMI.S38440. PMID: 27398023; PMCID: PMC4933534.
3. Vani K, Renuka A. Correlation of glycated haemoglobin with fasting and post prandial blood glucose in Type 2 diabetes [Internet]. Int J Clin Biochem Res. 2020 [cited 2025 Nov 04];7(3):380-383. Available from: https://doi.org/10.18231/j.ijcbr.2020.081
4. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
5. Pandas: McKinney, W. (2010). *Data structures for statistical computing in Python*. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56).
6. Waskom, M. L. (2021). *Seaborn: Statistical data visualization. Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021