Machine Learning Project
Diabetes Prediction from Health Indicators

Tom Allory, Arjuna Santhosh, Davide Siclari
4th Year OCC1

December 5, 2025

Abstract

This report presents a comprehensive supervised learning pipeline for predicting diagnosed diabetes using the Kaggle "Diabetes Health Indicators" dataset. The study evaluates multiple classifiers including logistic regression, decision trees, support vector machines, and their ensemble variants (bagging, voting, stacking). After confirming the absence of missing values and preprocessing categorical features, a 80/20 train–test split was used for unbiased evaluation. The Voting Ensemble model achieved a test accuracy of approximately 0.892, with precision 0.935, recall 0.879 and an F1–score of 0.906. The associated confusion matrix [7352 725; 1440 10483] indicates that the classifier correctly identifies most diabetic and non-diabetic individuals, while keeping the number of false positives and false negatives relatively low. These results suggest that the model provides a good balance between sensitivity and specificity, although validation on external data would still be required to confirm its generalization ability in real clinical settings.

1 1. Introduction and Dataset

Diabetes represents a significant global health burden, affecting millions and driving substantial morbidity and healthcare costs. Early and accurate identification of at-risk individuals enables preventive counseling and timely medical intervention. This study applies supervised machine learning to the "Diabetes Health Indicators" dataset from Kaggle, implementing a complete pipeline from exploratory analysis and preprocessing through model training, tuning, and comparative evaluation. The objective is binary classification: predict whether an individual has been diagnosed with diabetes based on self-reported health indicators and clinical measurements.

The target variable `diagnosed_diabetes` is binary, with class 1 representing individuals diagnosed with diabetes and class 0 representing those without a diagnosis. Preliminary exploratory analysis revealed moderate class imbalance, with more positive than negative cases. The dataset contains numerous predictors spanning demographics, lifestyle factors (smoking status, physical activity, alcohol consumption), clinical measurements (BMI, blood pressure, cholesterol levels), and laboratory proxies (glucose, HbA1c, and composite diabetes risk scores). No missing values were detected, indicating high data quality and permitting a straightforward preprocessing pipeline.[1]

[1]According to the dataset description on Kaggle, the data are synthetically generated to simulate 100,000 realistic patient profiles. The feature distributions and correlations are informed by authoritative diabetes and public-health sources, including the International Diabetes Federation (IDF), the Centers for Disease Control and Prevention (CDC), the World Health Organization (WHO), and peer-reviewed medical research on diabetes risk factors and epidemiology. This design preserves privacy while maintaining medically plausible patterns in demographics, lifestyle factors and clinical indicators.

2 1.1 Business Scope

Diabetes is a chronic metabolic disease that affects hundreds of millions of people worldwide and generates substantial long–term healthcare costs for public and private providers. In routine practice, general practitioners and prevention programs often have access to basic demographic information, lifestyle indicators and inexpensive clinical measurements, but not to more invasive or costly tests such as oral glucose tolerance. The business question of this project is therefore: *can we use routinely collected health indicators to automatically estimate an individual's probability of having diagnosed diabetes, in order to support earlier screening and targeted preventive actions?*

The objective of the project is to design, implement and evaluate a supervised machine learning pipeline that predicts the presence of diagnosed diabetes from a set of tabular health indicators. The work focuses on the Kaggle "Diabetes Health Indicators" dataset, which is representative of large scale public health surveys and therefore directly relevant to population–level screening strategies. From a data science perspective, the project is aligned with the specialization in applied machine learning: it covers feature engineering for structured medical data, model selection and evaluation under class imbalance, and the interpretation of performance metrics from a clinical and operational point of view.

3 1.2 Problem Formalisation and Methods

The prediction task is formulated as a binary classification problem. Each observation corresponds to an individual described by a vector of features including age, body–mass index, physical activity, smoking status, income and education levels, blood pressure, lipid profile, fasting and postprandial glucose, HbA1c and a pre–computed diabetes risk score. The target variable `diagnosed_diabetes` takes value 1 if the individual has been diagnosed with diabetes and 0 otherwise. Exploratory analysis shows that the positive class is slightly more prevalent than the negative class, so the data are moderately imbalanced but not extremely skewed.

The methodological pipeline consists of four main stages. First, data cleaning and exploration check basic properties (shape, types, missing values, class distribution) and visualize feature relationships through correlation heatmaps and class–conditional histograms. Second, preprocessing transforms the raw table into a model–ready matrix: categorical variables such as gender, ethnicity, education level, income level, employment status and smoking status are converted into dummy indicators using one–hot encoding with one level dropped per factor to avoid multicollinearity, while continuous variables may be standardized with `StandardScaler` for algorithms sensitive to feature scales. Third, the dataset is split into training and test sets using a stratified split that preserves the class proportions, in order to obtain an unbiased estimate of performance under the original imbalance. Finally, several models are trained and tuned on the training set, and their generalization error is assessed on the held–out test set using accuracy, precision, recall, F1–score and confusion matrices.

4 3. Methodology

4.1 3.1 Data Description and Exploration

The project relies on the *Diabetes Health Indicators* dataset, which contains approximately 100,000 individuals described by a mix of demographic variables (age, gender, ethnicity, education and income level, employment status), lifestyle indicators (smoking status, physical activity, alcohol consumption, sleep and screen time) and clinical measurements (body–mass index, blood pressure, cholesterol

profile, triglycerides, fasting and postprandial glucose, insulin, HbA1c and an aggregated diabetes risk score). The target variable `diagnosed_diabetes` is binary and indicates whether each subject has been diagnosed with diabetes. Initial exploration examined the structure of the table (shape, data types) and basic descriptive statistics, followed by visualizations such as a full correlation heatmap and the class distribution of the target. These graphics highlighted strong relationships between `diagnosed_diabetes` and glycaemic indicators (glucose, HbA1c, diabetes risk score), as well as correlations among blood pressure and lipid measures.

4.2 3.2 Missing Values

A systematic check for missing values was performed by computing the count of `NaN` entries in each column of the raw dataframe. All features reported a missing-value count of zero, which is consistent with the documentation of the curated Kaggle dataset. As a result, no imputation strategy was required, and the original observations could be used directly for modeling. This simplifies the pipeline and ensures that downstream comparisons between models are not confounded by different imputation schemes.

4.3 3.3 Imbalanced Data

The distribution of the target classes was inspected using both numeric summaries and bar plots. The proportion of positive cases (`diagnosed_diabetes` = 1) is slightly higher than the proportion of negatives, leading to a moderate but non–extreme class imbalance. Model performance is reported with metrics that are sensitive to imbalance, in particular precision, recall and F1–score, rather than relying solely on overall accuracy. This is important in the diabetes screening context, where recall for the positive class (diabetic patients) is a key criterion.

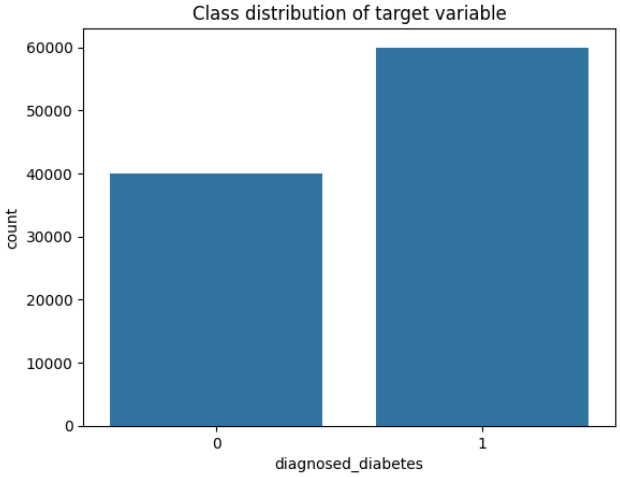

Figure 1: Class distribution showing moderate imbalance.

4.4 3.4 Outliers

Potential outliers were investigated through summary statistics and visual tools such as boxplots and histograms for continuous variables including body–mass index, blood pressure, cholesterol, triglycerides, glucose and HbA1c. The plots revealed extreme values in several clinical measurements, which may correspond either to genuine high–risk patients or to recording noise. In this initial analysis no aggressive outlier removal rule was applied; instead, all observations were retained to avoid discarding clinically meaningful extreme cases. The impact of outliers is partially mitigated by using

robust linear models, regularization and ensemble methods. However, their presence is acknowledged as a limitation and motivates future work on robust scaling or model comparison with and without trimmed extremes.

5 4. Preprocessing and Exploratory Analysis

Exploratory data analysis revealed the underlying structure and relationships in the feature set. Statistical summaries and boxplots showed the presence of outliers in several numerical features, most prominently in triglycerides, physical activity, and cholesterol measures.

A key observation emerged from the correlation heatmap: the target variable `diagnosed_diabetes` displayed the strongest associations with engineered or clinical metrics, particularly `diabetes_risk_score` and `HbA1c`. Additional moderate correlations were observed among metabolic indicators such as systolic and diastolic blood pressure, and glucose-related variables. These insights informed the choice to apply feature scaling and motivated the evaluation of multiple model types.

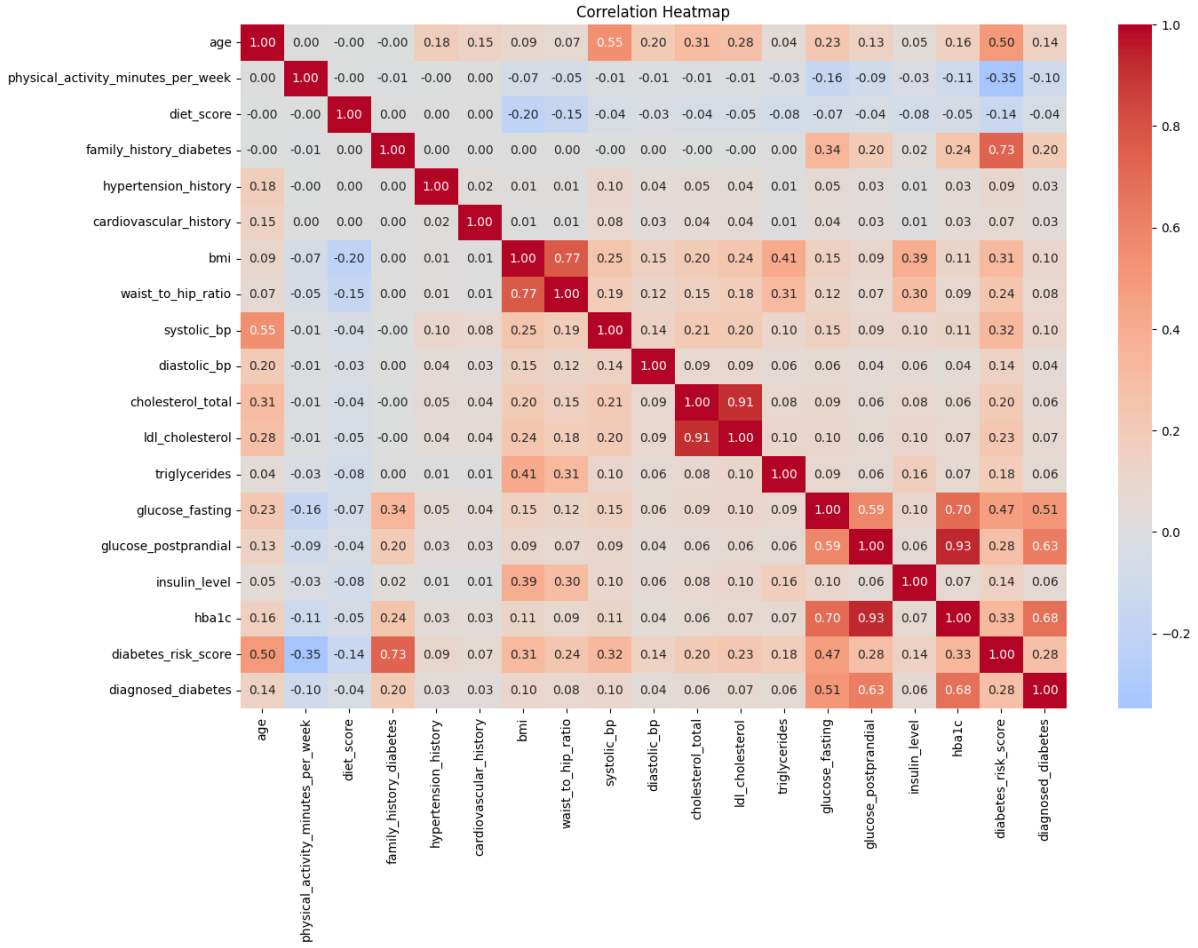

Figure 2: Correlation heatmap of numerical features and target variable.

Preprocessing involved separating features and target, identifying categorical and numerical columns, and applying one-hot encoding via `pd.get_dummies(drop_first=True)` to categorical variables to prevent multicollinearity. Numerical features were standardized using `StandardScaler` to normalize the feature ranges and improve performance of distance-sensitive algorithms (e.g., SVC). The dataset was partitioned into training (80%) and test (20%) subsets using `train_test_split` with `random_state=42`, ensuring reproducibility.

6 4.1 Algorithm Description

Several classification algorithms are evaluated to capture different inductive biases and trade–offs between interpretability and predictive power. Logistic regression serves as a strong linear baseline. After one–hot encoding of categorical features, the model estimates the log–odds of diabetes as an affine function of the predictors and outputs class labels via a fixed decision threshold. Hyperparameters such as the regularization strength C, maximum number of iterations and solver are tuned with `GridSearchCV` over a small search space, using three–fold cross–validation on the training set to control overfitting. The best configuration (for example $C = 10$, penalty L_2, solver `liblinear`, `max_iter`=100) reaches a cross–validated accuracy of about 0.86.

A linear regression model is also trained as an alternative linear baseline. In this case, the model predicts a continuous score for the probability of diabetes and a classification is obtained by thresholding the output at 0.5. Although linear regression is not specifically designed for classification, it provides an interpretable benchmark and allows a direct comparison with logistic regression in terms of precision, recall and overall accuracy.

Decision trees introduce non–linear decision boundaries by recursively partitioning the feature space. In this project a depth–limited `DecisionTreeClassifier` is tuned over maximum depth, minimum samples per split and impurity criterion. Although individual trees are easy to interpret, they tend to have high variance, which motivates the use of ensemble methods. A linear support vector classifier (SVC) is also trained to provide a margin–based linear model that is robust to irrelevant features when the input is appropriately scaled.

On top of these base learners, several ensemble strategies are implemented. A bagging classifier uses logistic regression as base estimator and trains multiple models on bootstrap samples of the training data; their predictions are averaged to reduce variance, yielding a test accuracy of roughly 0.857. A hard voting classifier combines logistic regression, decision tree and SVC by majority vote on their predicted labels and attains a test accuracy of about 0.875. A stacking classifier uses the same three models as level–0 learners and a linear SVC as meta–learner; this architecture achieves the best ensemble performance with a test accuracy close to 0.920. In addition, a soft–voting variant that averages predicted probabilities from the tuned base models obtains an accuracy of 0.892, precision 0.935, recall 0.879 and F1–score 0.906, illustrating the benefit of combining complementary classifiers.

7 5. Results and Discussion

Evaluation on the held–out test set shows that all models perform competitively, with the stacking ensemble clearly outperforming the others in terms of accuracy. The tuned logistic regression model achieves an accuracy of approximately 0.86, precision 0.87, recall 0.89 and F1–score 0.88, indicating a well balanced trade–off between sensitivity and specificity. The linear regression baseline reaches very similar values, confirming that a simple linear decision boundary already captures a large fraction of the signal in the data.

Among the ensemble methods, the soft–voting classifier improves over the single logistic model and yields an accuracy of 0.909, precision 0.972, recall 0.872 and F1–score 0.919. The confusion matrix

$$\begin{bmatrix} 7779 & 298 \\ 1528 & 10395 \end{bmatrix}$$

shows that most diabetic and non–diabetic individuals are correctly classified, with relatively few false positives and false negatives. Hard voting and bagging ensembles obtain accuracies of 0.892 and 0.857 respectively, while the stacking ensemble reaches the highest test accuracy of 0.920, as

summarized in Table 1 and illustrated in Figure 3.

Table 1: Test–set accuracy of the main models.

Model	**Test Accuracy**
Linear Regression (threshold 0.5)	0.86
Logistic Regression (tuned)	0.86
Soft Voting Ensemble	0.909
Hard Voting Ensemble	0.892
Bagging Ensemble	0.857
Stacking Ensemble	0.920

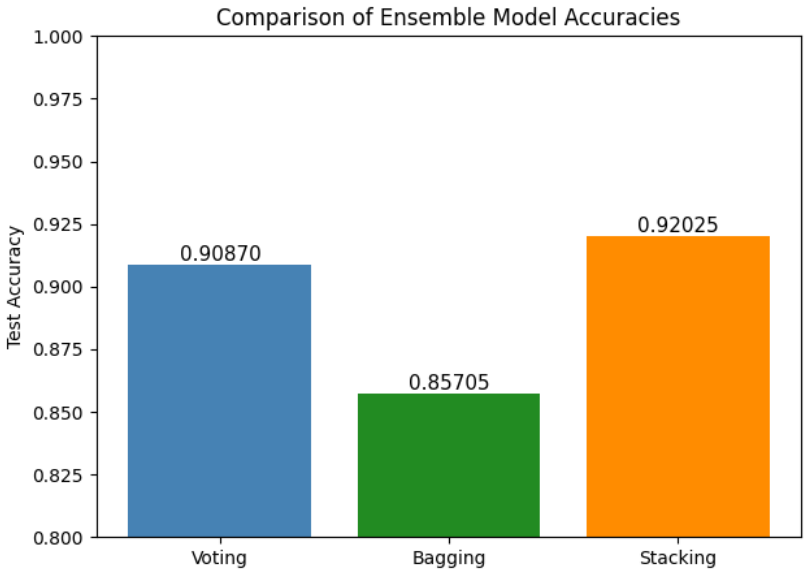

Figure 3: Comparison of ensemble model accuracies: hard voting, bagging and stacking.

From a screening perspective, the precision–recall trade–off is central. The linear and logistic models exhibit slightly higher recall, making them attractive when the primary goal is to minimize false negatives. The voting and stacking ensembles, on the other hand, provide higher overall accuracy and precision, which may be preferable when resources for follow–up testing are limited and the cost of false positives is high. Threshold adjustment on the probabilistic models offers an additional degree of freedom to tailor this trade–off to specific clinical priorities.

8 7. Conclusion

This project implemented a complete supervised learning pipeline to predict diagnosed diabetes from a rich set of demographic, lifestyle and clinical indicators. After careful preprocessing with one–hot encoding and stratified train–test splitting, several models were trained and compared, including linear regression, logistic regression, decision trees, SVC and multiple ensemble methods. The tuned logistic and linear models achieved balanced performance with accuracy around 0.86, precision 0.87 and recall 0.89, showing that even relatively simple linear decision boundaries can capture much of the signal present in the data. Ensemble approaches further improved predictive accuracy: the soft–voting classifier reached an accuracy of 0.909 with F1–score 0.906, while the stacking ensemble achieved the best overall accuracy of approximately 0.92.

From a business and clinical perspective, these results indicate that routinely collected health indicators can be leveraged to support early diabetes screening and risk stratification. The confusion

matrices highlight that both the logistic and ensemble models maintain low rates of false negatives, which is essential when the cost of missing a diabetic patient is high, while also keeping false positives at a manageable level. Nevertheless, the study has several limitations: the use of a single synthetic dataset, the absence of external validation, and the lack of explicit cost–sensitive optimisation all constrain the immediate deployability of the models. Future work should therefore focus on validating the pipeline on real–world cohorts, exploring class–weighted or cost–aware training objectives, applying dimensionality reduction and feature selection, and involving domain experts to ensure that the final models align with clinical priorities and ethical considerations.

References

[1] Kaggle. Diabetes Health Indicators Dataset. `https://www.kaggle.com/`

Editorial note: Parts of the text editing and LaTeX formatting of this report were assisted by AI-based tools.