Data1030 Final Project: Diabetes Prediction

Xuanyao Qian
Brown University, Data Science Institute

GitHub:
https://github.com/Pafuuuu/Brown_Data1030_Final_Project.git

December 2025

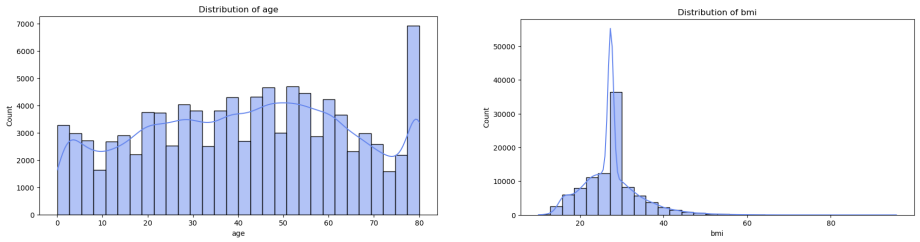
1 Introduction

Diabetes is a growing global health concern, and early identification of individuals at risk is critical for improving clinical outcomes and guiding intervention strategies. Motivated by the need for accurate and accessible diagnostic tools, this analysis examines a range of medical and demographic factors to build predictive models capable of classifying diabetes status effectively. Such tools can support healthcare professionals in identifying high-risk patients, enabling personalized treatment planning and preventive care. Previous work has been done by Tasin et al. where they performed a range of machine learning techniques for diabetes prediction and achieved a best F1 score of 0.81 using XGBoost[2].

The dataset used in this study is sourced from Kaggle's Diabetes Prediction Dataset, which aggregates electronic health records from multiple healthcare providers. The dataset contains 100,000 patient records, with diverse patient information across 8 features, including 4 numerical features: age, body mass index (BMI), H1bAc level, blood glucose level and 4 categorical features: gender, smoking history, hypertension, heart disease. As a binary classification problem, the task is to determine whether an individual has diabetes (1) or does not (0) based on the available attributes[1].

2 Exploratory Data Analysis (EDA)

An initial exploration of the dataset reveals several important characteristics that directly influence model development and evaluation. The numerical features, particularly age and BMI, display distinctly non-normal distributions. Age reflects a large proportion of older patients in the sample. While BMI shows an abnormal pattern: a sharp spike concentrated at a single value, followed by a long right tail.



(a) Distribution of Age (b) Distribution of BMI

Figure 1: Distribution of Age(left) and BMI(right)

Further investigation confirms that this spike results from **pre-existing mean imputation** in the original dataset, where all missing BMI values were replaced using the global mean. This artificially inflates the frequency of the mean BMI value and distorts the natural distribution, potentially misleading algorithms that depend on distributional structure. To correct this issue, all previously imputed BMI values were reverted back to missing (NA). After this restoration, BMI exhibits a more realistic slightly right-skewed distribution, with **25,495 missing values** requiring appropriate imputation later in the analysis.

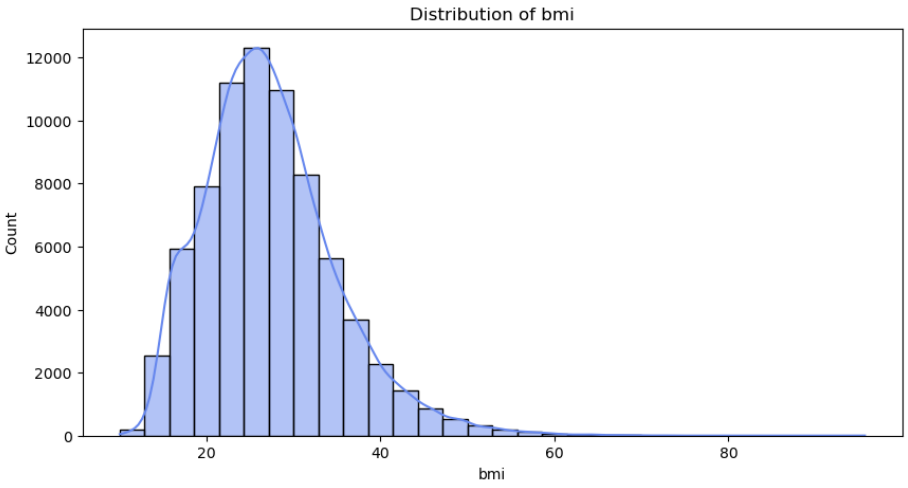


Figure 2: Distribution of BMI after removing imputed values

The categorical variables also reveal notable imbalances that must be accounted for during modeling. Medical conditions such as heart disease and hypertension are heavily skewed toward the negative class, indicating that only a small fraction of patients report these diagnoses. The target variable, diabetes, displays a similar pattern, with significantly fewer positive than negative cases. These class imbalances introduce a risk of biased model learning, where

prediction algorithms may default to the majority class. To mitigate this, both the train–test split and all cross-validation procedures will be stratified, ensuring that the distribution of key categorical variables—most importantly the diabetes outcome—remains proportionally consistent across evaluation folds.

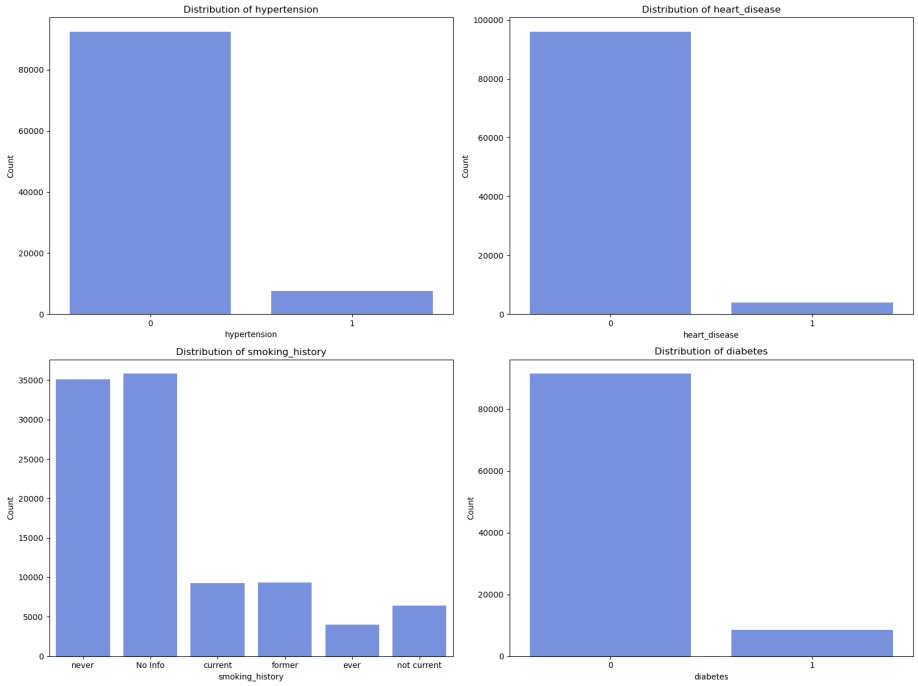


Figure 3: CountPlot of hypertension(top left), heart disease(top right), smoking history(bottom left), diabetes(bottom right)

A deeper investigation into the smoking history variable uncovers additional complexity related to ambiguous labeling and extensive missingness. In total, 35,816 observations—over one third of the dataset—are labeled as "No Info." Among the non-missing entries, smoking history includes the categories not current, former, never, ever, and current. The categories not current, former, and ever exhibit overlapping meaning, but their precise distinctions are not documented. For model simplicity and interpretability, these three similar labels were consolidated into a single category, former. Interestingly, analysis of the diabetes rate across smoking groups reveals that individuals with missing smoking information exhibit a substantially lower prevalence of diabetes compared to all clearly labeled categories. This indicates that missingness is likely not random and may encode certain hidden patterns. Instead of discarding these observations or attempting to force-fit them into existing smoking categories, the missing values will be preserved as their own category, "missing," to retain this potentially informative signal in the modeling stage.

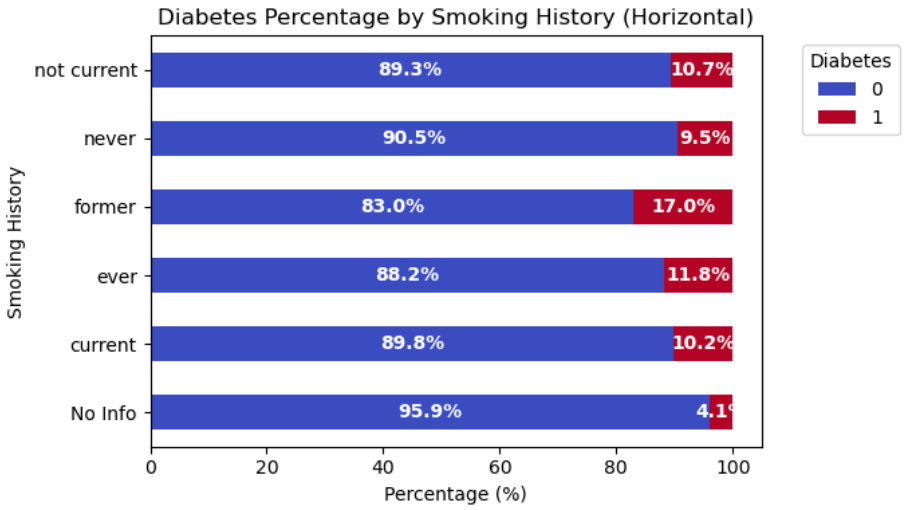


Figure 4: Diabetes Percentage by Smoking hIstory

3 Methods

The dataset was split into a stratified 60/20/20 train–validation–test partition to ensure that each subset preserved the original imbalance between diabetic and non-diabetic cases. Stratification prevented unstable estimates that could arise from folds with too few minority-class examples. All model development was performed exclusively on the training data, while the test set remained untouched for final evaluation. Hyperparameter tuning used five-fold stratified cross-validation, maintaining consistent class ratios within each fold and providing a reliable, leakage-free estimate of model performance..

The preprocessing pipeline addressed missing values, scaling, encoding, and class imbalance in a unified and reproducible framework. Missing BMI values were imputed with an Iterative Imputer, which captures multivariate relationships more effectively than simple imputation while remaining computationally lighter than KNN. Numerical features were standardized using StandardScaler, and categorical variables were encoded with OneHotEncoder. To mitigate class imbalance, SMOTE oversampling was applied within each training fold, preventing synthetic samples from leaking into validation or test sets and reducing bias toward the majority class. However, since oversampling distort the original distribution of the dataset and may cause unforeseen results, both models with and without SMOTE will be trained for further evaluation.

Four supervised learning algorithms were implemented—logistic regression, support vector machines, random forests, and XGBoost—covering both linear and non-linear modeling approaches. Each model was embedded in the pre-

processing pipeline and tuned using GridSearchCV. Logistic regression tuning focused on regularization strength, penalty type, and solver to balance flexibility and convergence stability. Random forest hyperparameters such as the number of trees, maximum depth, and minimum sample thresholds were adjusted to control model complexity and variance. For SVMs, the margin penalty parameter and choice of linear versus non-linear kernels were tuned to identify the most effective decision boundary. XGBoost optimization involved tree depth, learning rate, and subsampling rate, which together influence model complexity, learning stability, and generalization.

Model	**Hyperparameters Tuned**
Logistic Regression	C: [0.01, 0.1, 1, 10] penalty: L2 solver: lbfgs
Random Forest	n_estimators: [200, 500] max_depth: [5, 10, None] min_samples_split: [2, 5] min_samples_leaf: [1, 2]
SVM	C: [0.1, 1, 10] kernel: [rbf, linear]
XGBoost	max_depth: [3, 5, 7] learning_rate: [0.01, 0.1, 0.2] subsample: [0.7, 1]

Table 1: Hyperparameters Tuned for Each Model

Model performance was evaluated using the F1 score. This metric was selected because accuracy is inappropriate for imbalanced datasets: a classifier predicting the majority class at all times would achieve high accuracy despite failing to identify any diabetic patients. The F1 score provides a balance between precision and recall and therefore offers a more meaningful assessment of clinical prediction quality.

Uncertainty in model performance was measured in two ways. Variation across cross-validation folds provided an estimate of uncertainty attributable to different data partitions and highlighted the stability of each algorithm. In addition, stochastic learning processes within ensemble methods contributed further variability; although random seeds were fixed to increase reproducibility, some randomness inherent to tree-based models remains. Reporting both the mean and standard deviation of the cross-validated F1 score allowed for a more nuanced comparison of models and facilitated the identification of performance

differences that exceeded sampling variability.

4 Results

A majority-class baseline model was included for comparison and, as expected, though accuracy was as high as 0.915, the model produces zero precision and recall, making it meaningless to calculate F1 score, which underscore the need for more sophisticated modeling approaches.

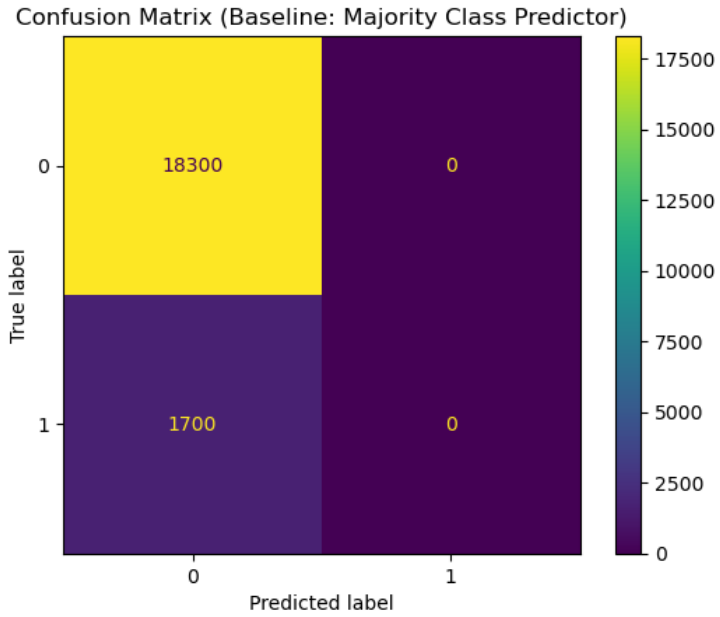
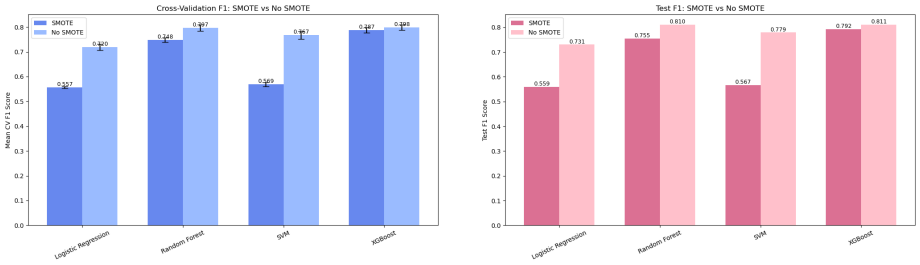


Figure 5: Confusion Matrix For Baseline Predictor

Figure 6 compares cross-validation and test F1 scores for four classifiers trained with and without SMOTE. Across all models, those trained on the original imbalanced data consistently outperform their SMOTE-augmented counterparts. Logistic regression shows the largest improvement, with mean CV F1 rising from about 0.56 to 0.72, and SVM exhibits a similar shift from roughly 0.57 to 0.77. Random forest and XGBoost are less affected by oversampling but still achieve higher F1 scores without SMOTE (approximately 0.75→0.80 and 0.79→0.80, respectively). The small error bars indicate that these performance gaps are stable across folds rather than due to variability in the splits.



(a) Validation F1 Score

(b) Test F1 Score

Figure 6: Validation and Test F1 Scores of models with and without SMOTE

The underperformance using SMOTE might due to the mixed-feature nature of the dataset. Interpolating minority samples in a mixed feature space—especially with one-hot encoded categorical variables—creates synthetic points that do not correspond to realistic patient profiles and blur class boundaries. This added noise is particularly harmful for linear and margin-based models like logistic regression and SVM, which rely on clean separations.

Since XGBoost gives the highest test F1 Score of 0.811, it is chosen as the winner model of the four. Investigating the model further using a confusion matrix, the no-SMOTE XGBoost model reveals strong overall predictive performance, achieving an accuracy of 97.3%, which is about 29.4 standard deviations higher than the baseline model's accuracy of 91.5%. Precision for the positive (diabetes) class is exceptionally high at 0.98, indicating that when the model predicts a patient has diabetes, it is almost always correct. However, recall is noticeably lower at 0.69, meaning the model fails to identify roughly 31% of true diabetes cases. This imbalance between precision and recall suggests that while the classifier is highly conservative—avoiding false positives—it misses a substantial portion of positive cases, a trade-off driven by the underlying class imbalance and the model's preference for minimizing incorrect diabetes predictions.

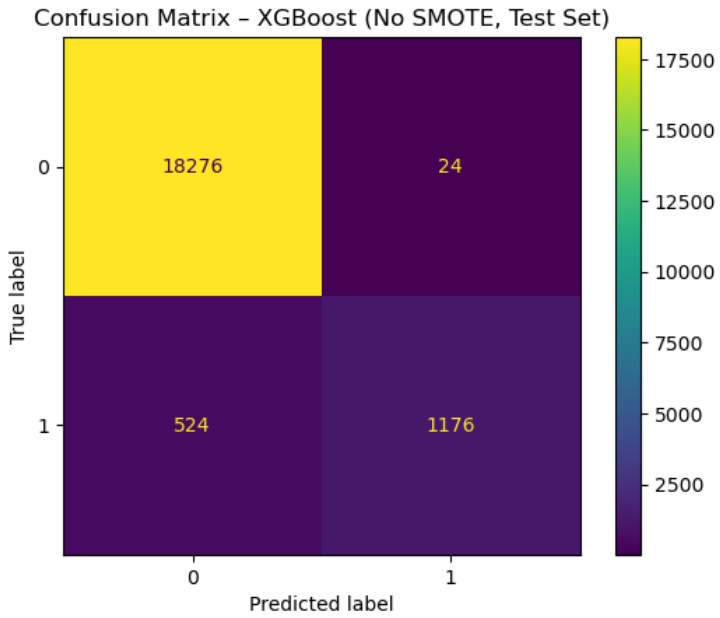


Figure 7: Confusion Matrix For No SMOTE XGBoost

When looking at global feature importance through permutation feature importance, gain, and cover, the model consistently identifies HbA1c level and blood glucose level as the dominant predictors of diabetes status.Permutation Importance, which directly measures performance degradation when features are shuffled, shows a dramatic drop in accuracy when these two biomarkers are perturbed, while all other features have negligible effect. This indicates that HbA1c and glucose carry the majority of the model's predictive signal and are essential for distinguishing diabetic and non-diabetic individuals.

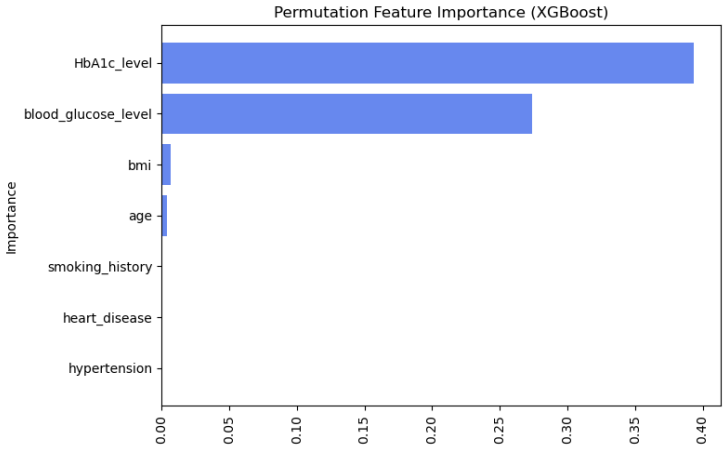
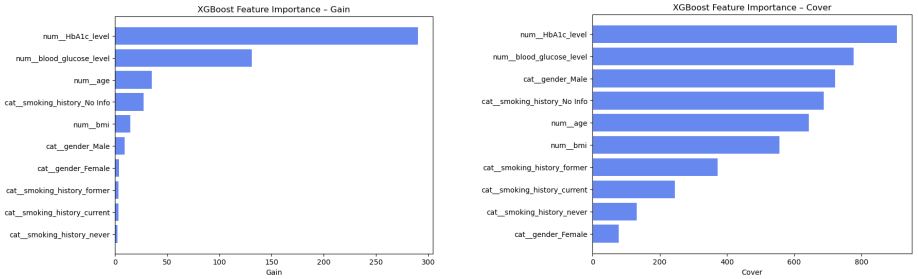


Figure 8: Permutation Feature Importance

The Gain metric, which captures the average improvement in splitting purity contributed by each feature, reinforces this finding: HbA1c exhibits the highest gain, followed by glucose, suggesting that these biomarkers produce the most informative splits across the ensemble. Cover, which reflects the number of samples affected by splits on each feature, likewise places these two features at the top, indicating that they influence a large portion of the tree structure.



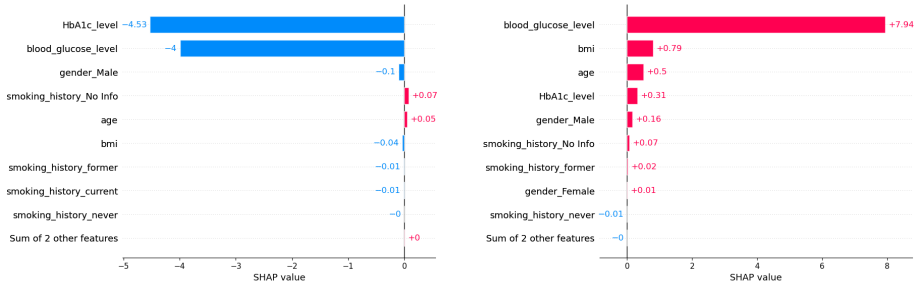
(a) Gain (b) cover

Figure 9: Feature Importance for gain(left) and cover(right)

Categorical variables—including gender and smoking history—show consistently low importance across all metrics.Overall, the combined evidence demonstrates that physiological metabolic indicators (HbA1c, blood glucose) dominate the predictive landscape, while demographic and behavioral features add limited value.

The global trend carries to local feature importance, inspecting individual predictions with class 0 and class 1, HbA1c and blood glucose level remains the

most importance feature for predicting diabetes when looking at local importance using SHAP.



(a) Class 0 (b) Class 1

Figure 10: SHAP Importance for Class 0(left) and Class 1(right) Prediction

5 Outlook

Although the current modeling pipeline performs well, several improvements could further enhance its effectiveness. The hyperparameter search was deliberately limited for computational efficiency; expanding the tuning space—particularly for tree depth, learning rates, and regularization terms—may yield models that better capture nonlinear patterns in the data. A second opportunity lies in refining the evaluation metric. Because clinical screening prioritizes minimizing missed diagnoses, shifting from the F1 score to a recall-weighted metric such as the F2 score would better align model selection with real-world medical priorities.

Further gains could come from incorporating additional clinical variables, such as family history or longitudinal glucose measurements, which would provide richer context and help the model differentiate more effectively between diabetic and non-diabetic cases. Finally, alternative imbalance-handling strategies, such as class-weighted losses or selective oversampling of borderline samples, could address class skew without introducing the synthetic noise observed with SMOTE. Together, these enhancements would strengthen both the predictive accuracy and practical applicability of the model in clinical settings.

References

[1] Mohammed Mustafa. Diabetes prediction dataset. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset, 2022. Kaggle Dataset.

[2] I. Tasin, T. U. Nabil, S. Islam, and R. Khan. Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, 10(1-2):1–10, 2022.