Diabetes Prediction with Machine Learning

Abstract

This project leverages a dataset of medical and demographic patient data to predict diabetes status (positive or negative) using machine learning techniques. The dataset, comprising 100,000 entries with features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level, is analyzed to build robust predictive models. Through preprocessing steps, such as label encoding and handling class imbalance, multiple models—including Random Forest Classifier, Logistic Regression, SVC, and XGB Classifier—are trained and evaluated using metrics like accuracy, precision, recall, F1 Score, and ROC AUC. The study aims to identify the most effective model for early diabetes detection, supporting healthcare professionals in risk assessment and personalized treatment planning, while also enabling researchers to explore relationships between risk factors and diabetes prevalence. The results highlight the superior performance of gradient boosting techniques, with recommendations for addressing data quality issues and enhancing model interpretability for clinical applications.

Problem Statement

The objective of this project is to develop a machine learning model to **predict whether a patient has diabetes** (binary classification: 0 for no diabetes, 1 for diabetes) based on medical and demographic features. The dataset includes key indicators such as HbA1c level (>6.5% typically indicates diabetes), blood glucose level, BMI, age, gender, hypertension, heart disease, and smoking history, which are known to influence diabetes risk. This problem is critical for:

- **Healthcare Applications**: Enabling early identification of at-risk patients to facilitate timely interventions and tailored treatment plans.
- **Research Purposes**: Investigating the relationships between medical/demographic factors and diabetes likelihood to inform preventive strategies. The challenge lies in handling class imbalance (only ~8.5% of patients have diabetes), ensuring model generalizability, and addressing data quality issues like duplicates (3,854 detected) to build a reliable and interpretable predictive system.

Data Information

The **Diabetes Prediction Dataset** is a collection of medical and demographic data from patients, designed to predict diabetes status (positive or negative). It contains **100,000 entries** with **9 columns**, comprising numerical and categorical features. The dataset is valuable for building machine learning models to identify at-risk patients and for researching the relationships between various factors and diabetes likelihood. Below is a detailed description of the features, combining source information and analysis from the Jupyter notebook:

Features

- **gender**:
  - **Type**: Categorical (object), encoded numerically using LabelEncoder (e.g., 0: Female, 1: Male, 2: Other).
  - **Description**: Represents the biological sex of the individual, which influences diabetes susceptibility. Categories: Male, Female, Other.
- **age**:
  - **Type**: Numerical (float64).
  - **Description**: Patient's age in years, ranging from 0.08 to 80 (mean: 41.89, std: 22.52). Diabetes is more prevalent in older adults.
- **hypertension**:
  - **Type**: Binary (int64).
  - **Description**: Indicates persistently elevated blood pressure, a risk factor for diabetes. Values: 0 (No hypertension), 1 (Hypertension; prevalence: ~7.5%).
- **heart_disease**:
  - **Type**: Binary (int64).
  - **Description**: Indicates the presence of heart disease, associated with increased diabetes risk. Values: 0 (No heart disease), 1 (Heart disease; prevalence: ~3.9%).
- **smoking_history**:
  - **Type**: Categorical (object), encoded numerically (e.g., 0: No Info, 1: Current, 4: Never).
  - **Description**: Represents smoking status, which can exacerbate diabetes complications. Categories: Not Current, Former, No Info, Current, Never, Ever.
- **bmi**:
  - **Type**: Numerical (float64).
  - **Description**: Body Mass Index, measuring body fat based on weight and height, ranging from 10.01 to 95.69 (mean: 27.32, std: 6.64). Categories:

<18.5 (Underweight), 18.5–24.9 (Normal), 25–29.9 (Overweight), ≥30 (Obese). Higher BMI is linked to increased diabetes risk.

- **HbA1c_level**:
  - **Type**: Numerical (float64).
  - **Description**: Hemoglobin A1c level, reflecting average blood sugar over 2–3 months, ranging from 3.5 to 9.0% (mean: 5.53, std: 1.07). Levels >6.5% typically indicate diabetes.
- **blood_glucose_level**:
  - **Type**: Numerical (int64).
  - **Description**: Blood glucose level at a given time, ranging from 80 to 300 (mean: 138.06, std: 40.71). High levels are a key diabetes indicator.
- **diabetes**:
  - **Type**: Binary target variable (int64).
  - **Description**: Indicates diabetes status. Values: 0 (No diabetes), 1 (Diabetes; prevalence: ~8.5%).

Data Characteristics

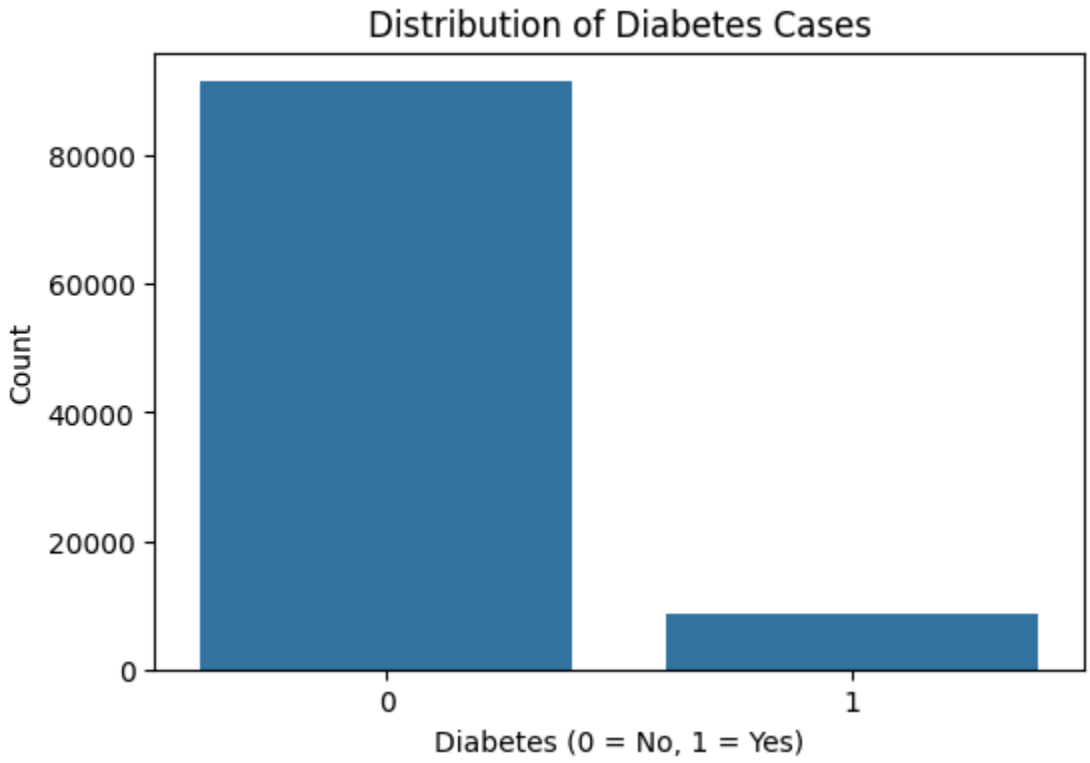
- **Data Types**:
  - **Categorical**: gender, smoking_history (converted to numerical via label encoding).
  - **Numerical**: age, BMI, HbA1c_level, blood_glucose_level.
  - **Binary**: hypertension, heart_disease, diabetes.
- **Missing Values**: No missing values (isnull().sum() returns 0 for all columns).
- **Duplicates**: 3,854 duplicated rows, which may inflate model performance if not addressed.
- **Shape**: (100,000 rows, 9 columns).
- **Class Imbalance**: The target variable, diabetes, has a mean of 0.085, indicating only 8.5% of patients have diabetes, necessitating techniques like SMOTE for balanced modeling.
- **Statistical Summary** (from data.describe()):
  - **Age**: Mean 41.89, median 43, indicating a slightly right-skewed distribution.
  - **BMI**: Mean 27.32, median 27.32, with a wide range suggesting diverse body compositions.
  - **HbA1c_level**: Mean 5.53, with 75% of values ≤6.2, aligning with non-diabetic ranges.
  - **Blood_glucose_level**: Mean 138.06, with a broad range indicating variability in glucose control.
  - **Hypertension and Heart Disease**: Low prevalence (7.5% and 3.9%, respectively), suggesting sparse positive cases.

Exploratory Data Analysis (EDA)

The EDA section examines the dataset's characteristics, distributions, and relationships between features to inform preprocessing and modeling decisions. The analysis includes class distribution, cross-tabulations, a correlation matrix, and specific insights into age and smoking history.

1. Class Distribution

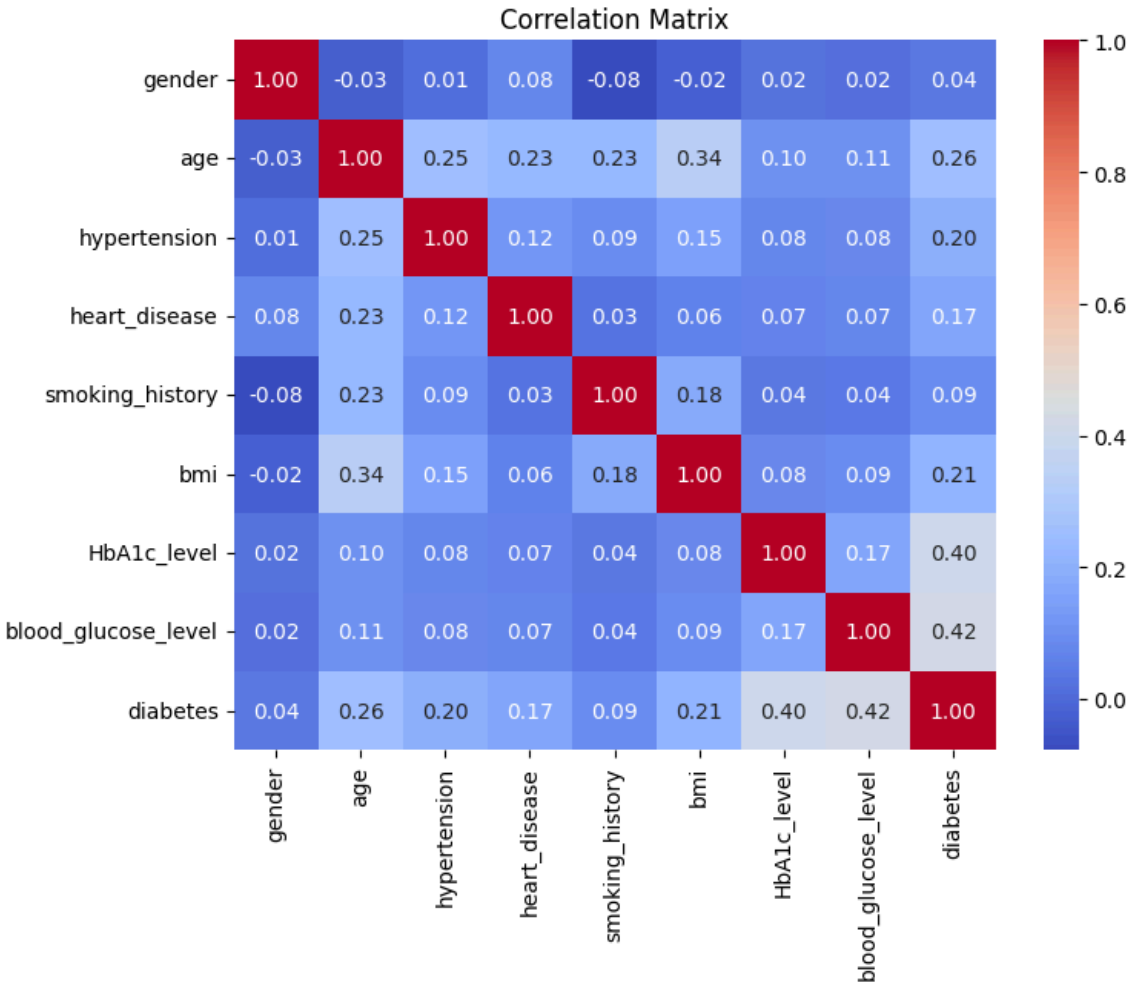
- **Diabetes Class Distribution**:
  - 0 (No diabetes): 91.5%
  - 1 (Diabetes): 8.5%
- **Insight**: The dataset is highly imbalanced, with only 8.5% of patients having diabetes. This necessitates techniques like SMOTE to oversample the minority class, ensuring models do not overly favor the majority class and miss diabetes cases.



2. Correlation Matrix

The correlation matrix (using Pearson correlation) reveals relationships between numerical features after encoding categorical variables (gender, smoking_history):

- **Strong Positive Correlations**:
    - **diabetes and HbA1c_level**: 0.40. Higher HbA1c levels (>6.5% indicates diabetes) are strongly associated with diabetes, aligning with medical knowledge.
    - **diabetes and blood_glucose_level**: 0.42. Elevated blood glucose levels are a key indicator of diabetes.
    - **age and BMI**: 0.34. Older patients tend to have higher BMI, both of which are risk factors for diabetes.
- **Moderate Positive Correlations**:
    - **diabetes and age**: 0.26. Diabetes prevalence increases with age.
    - **diabetes and hypertension**: 0.20. Hypertension is a known risk factor for diabetes.
    - **diabetes and heart_disease**: 0.17. Heart disease is associated with higher diabetes risk.
    - **diabetes and BMI**: 0.21. Higher BMI increases diabetes likelihood.
- **Weak Correlations**:
    - **gender, smoking_history**: Show negligible correlations with diabetes (0.04 and 0.09, respectively), suggesting limited direct influence.
- **Insight**: HbA1c_level and blood_glucose_level are the strongest predictors of diabetes, followed by age, BMI, hypertension, and heart_disease. The weak correlations for gender and smoking history suggest they may have less direct impact, though cross-tabulations provide further context.

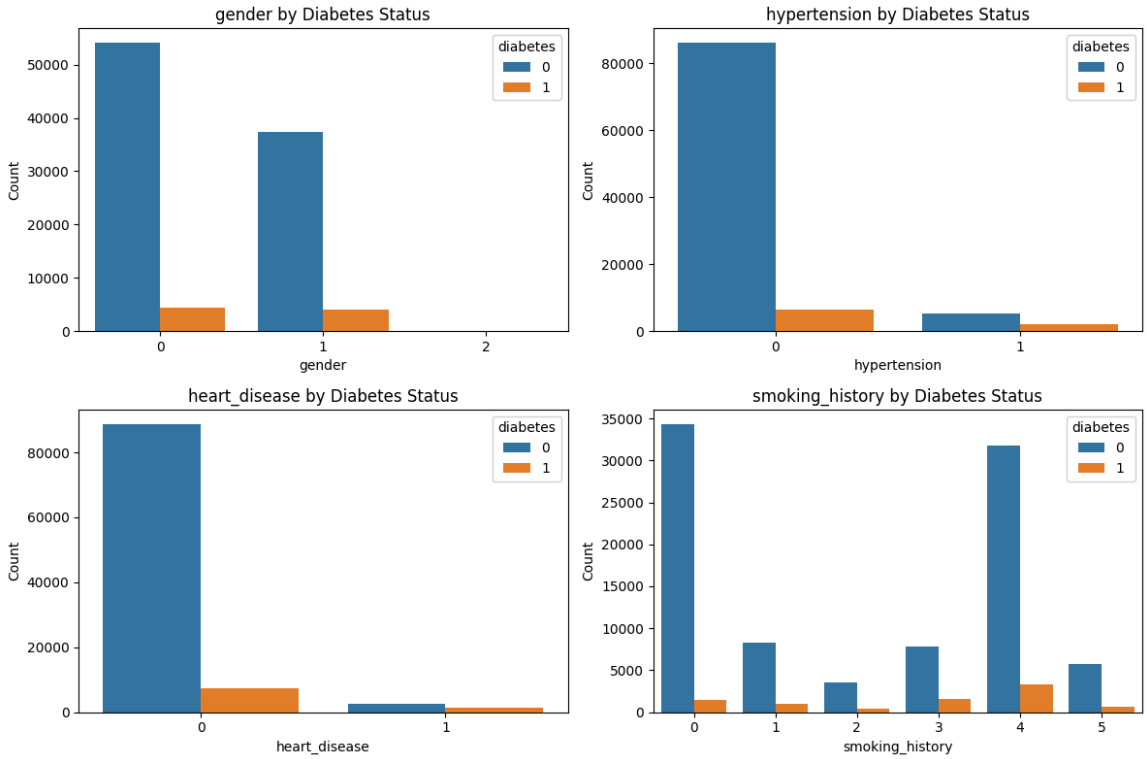


3. Cross-Tabulations

Cross-tabulations examine the proportion of diabetes cases across categorical and binary features:

- **Gender and Diabetes**:
  - Female (0): 7.62% diabetic.
  - Male (1): 9.75% diabetic.
  - Other (2): 0% diabetic.
  - **Insight**: Males have a slightly higher diabetes prevalence than females. The "Other" category has no diabetes cases, possibly due to its small sample size.
- **Hypertension and Diabetes**:
  - No hypertension (0): 6.93% diabetic.
  - Hypertension (1): 27.90% diabetic.

    - **Insight**: Patients with hypertension are significantly more likely to have diabetes, supporting its role as a risk factor.
- **Heart Disease and Diabetes**:
    - No heart disease (0): 7.53% diabetic.
    - Heart disease (1): 32.14% diabetic.
    - **Insight**: Heart disease substantially increases diabetes likelihood, consistent with medical literature.
- **Smoking History and Diabetes**:
    - No Info (0): 4.06% diabetic.
    - Current (1): 10.21% diabetic.
    - Former (2): 11.79% diabetic.
    - Ever (3): 17.00% diabetic.
    - Never (4): 9.53% diabetic.
    - Not Current (5): 10.70% diabetic.
    - **Insight**: The "Ever" smoking category has the highest diabetes prevalence (17%), followed by "Former" (11.79%), suggesting past smoking significantly increases risk. "No Info" has the lowest rate, likely due to missing data.



4. Smoking History Distribution

- **Distribution**:
    - No Info (0): 35.82%
    - Never (4): 35.10%
    - Ever (3): 9.35%
    - Current (1): 9.29%
    - Not Current (5): 6.45%
    - Former (2): 4.00%
- **Insight**: A significant portion of the dataset (35.82%) has "No Info" for smoking history, which may introduce bias if not handled properly. "Never" smokers are the largest known category, while "Former" smokers are the smallest, despite their higher diabetes prevalence.

5. Patients with Age < 1 Year

- **Summary Statistics** (911 patients):
    - Age: Mean 0.53, std 0.24, range 0.08–0.88.
    - Diabetes: Mean 0.0 (no diabetes cases).
- **Insight**: None of the 911 patients under 1 year have diabetes, which aligns with the rarity of diabetes in infants. This subgroup may not contribute to predictive modeling for diabetes and could be considered for exclusion to focus on more relevant age groups.

6. Key EDA Insights

- **Class Imbalance**: The 91.5% vs. 8.5% split in diabetes status confirms the need for imbalance handling (e.g., SMOTE) to improve model performance on the minority class.
- **Feature Importance**: HbA1c_level and blood_glucose_level show the strongest correlations with diabetes, followed by age, bmi, hypertension, and heart_disease. These features should be prioritized in modeling.
- **Categorical Features**: hypertension, heart_disease, and certain smoking_history categories (e.g., Ever, Former) significantly increase diabetes risk, while gender shows a weaker effect.
- **Data Quality**: The presence of 3,854 duplicates and 35.82% "No Info" in smoking_history highlights the need for deduplication and handling of missing/unknown data.
- **Age Consideration**: Patients under 1 year are not relevant for diabetes prediction and may be excluded to improve model focus.

Data Preprocessing

Data preprocessing ensures the dataset is clean, consistent, and suitable for machine learning models. The following steps were applied to address categorical encoding, missing data, class imbalance, and feature scaling, based on insights from EDA.

1. Label Encoding

- **Features**: gender (Male, Female, Other) and smoking_history (Not Current, Former, No Info, Current, Never, Ever).
- **Method**: Used LabelEncoder to convert categorical variables into numerical format (e.g., gender: 0, 1, 2; smoking_history: 0 to 5).
- **Reason**: Most machine learning algorithms require numerical inputs, and label encoding ensures compatibility.

2. Impute "No Info" in Smoking History Using KNN

- **Step 1: Convert 'No Info' to NaN**:
    - Replaced smoking_history value 0 ("No Info") with np.nan to treat it as missing data.
    - **Reason**: "No Info" (35.82% of the data) introduces bias if treated as a valid category; imputing it provides a more meaningful representation.
- **Step 2: Scale Numerical Features for KNN**:
    - Applied StandardScaler to numerical features (age, bmi, HbA1c_level, blood_glucose_level) to standardize them (mean 0, std 1).
    - **Reason**: KNN imputation relies on distance metrics, which are sensitive to feature scales. Scaling ensures equal weighting of features.
- **Step 3: Apply KNN Imputation**:
    - Used KNNImputer (n_neighbors=5, weights='uniform') on features gender, age, hypertension, heart_disease, bmi, HbA1c_level, blood_glucose_level, and smoking_history to impute missing values in smoking_history.
    - **Reason**: KNN imputation predicts missing values based on the nearest neighbors, leveraging relationships between features (e.g., age and smoking_history correlation of 0.23).
- **Step 4: Convert Imputed Values Back to Integers**:
    - Rounded smoking_history values (KNN outputs floats) and converted them to integers.
    - **Reason**: smoking_history is a categorical variable, and integer values align with its encoded format.
- **Result**: Post-imputation distribution of smoking_history:

    - Never (4): 56.62%
    - Ever (3): 21.31%
    - Current (1): 9.40%
    - Not Current (5): 6.86%
    - Former (2): 5.81%
    - **Insight**: Imputation significantly reduced "No Info" entries, with "Never" becoming the dominant category, aligning with its pre-imputation prevalence (35.10%). The distribution now reflects more realistic smoking patterns.

3. Filter Out Rows Where Age < 1

- **Method**:
    - Reverted scaling of numerical features using scaler_knn.inverse_transform to interpret age in its original scale.
    - Filtered out rows where age < 1 (originally 911 rows, none with diabetes).
- **Reason**: EDA revealed that patients under 1 year have no diabetes cases and are not relevant for prediction, as diabetes is rare in infants. Removing them focuses the model on more relevant age groups.

4. Define Features and Target

- **Features (X)**: All columns except diabetes (gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level).
- **Target (y)**: diabetes (0 or 1).
- **Reason**: Separates the predictor variables from the target for supervised learning.

5. Split the Dataset into Training and Testing Sets

- **Method**: Used train_test_split with an 80-20 split (test_size=0.2, random_state=42).
- **Result**:
    - Training set shape: (79,272, 8) before SMOTE.
    - Testing set shape: (19,818, 8).
- **Reason**: Splitting ensures the model is evaluated on unseen data, preventing overfitting and assessing generalizability.

6. Handle Class Imbalance Using SMOTE

- **Method**: Applied SMOTE (random_state=42) to the training set to oversample the minority class (diabetes = 1).
- **Result**:

    - Training set shape after SMOTE: (144,936, 8).
    - Training label distribution after SMOTE: 50% (0), 50% (1).
- **Reason**: The original dataset is imbalanced (91.5% no diabetes, 8.5% diabetes). SMOTE balances the training set, improving the model's ability to learn from the minority class and reducing bias toward the majority class. SMOTE is applied only to the training set to avoid data leakage.

7. Scale Numerical Features for Modeling

- **Method**:
    - Applied a new StandardScaler (scaler_model) to numerical features (age, bmi, HbA1c_level, blood_glucose_level) in the training set (fit_transform) and testing set (transform).
- **Reason**: Models like Logistic Regression and SVC are sensitive to feature scales. Scaling ensures all numerical features contribute equally, improving model performance and convergence.

8. Key Preprocessing Outcomes

- **Dataset Shape**:
    - After filtering age < 1: 99,089 rows (100,000 - 911).
    - Training set after SMOTE: 144,936 rows (oversampled to balance classes).
    - Testing set: 19,818 rows.
- **Class Balance**: The training set now has an equal distribution of diabetes and non-diabetes cases, addressing the imbalance issue.
- **Smoking History**: KNN imputation effectively handled "No Info" entries, resulting in a more representative distribution of smoking categories, with "Never" (56.62%) being the most common.
- **Data Quality**: While age filtering and imputation improved data relevance, duplicates (3,854) remain unaddressed and should be considered in future steps to avoid data leakage.

Model Training

Four machine learning models were trained to predict diabetes status, leveraging the preprocessed dataset. The models were selected based on their ability to handle binary classification tasks and their varying strengths in capturing different types of patterns in the data. Default hyperparameters were used for all models, as hyperparameter tuning was not applied.

1. Model Selection

- **RandomForestClassifier**:
    - **Reason**: A tree-based ensemble model that handles non-linear relationships and provides feature importance. It is robust to unscaled data and effective for imbalanced datasets when paired with SMOTE.
- **LogisticRegression**:
    - **Reason**: A simple, interpretable model that serves as a baseline. It assumes linear relationships between features and the target, making it suitable for comparison with more complex models.
- **SVM (Support Vector Machine, SVC)**:
    - **Reason**: Excels in high-dimensional spaces and can capture non-linear patterns using a kernel (e.g., RBF). It benefits from scaled features, as applied in preprocessing.
- **XGBClassifier**:
    - **Reason**: A gradient boosting algorithm known for high accuracy, handling complex feature interactions, and robustness to imbalanced data when paired with SMOTE.

2. Training Process

- **Dataset**: Used the preprocessed training set (144,936 rows after SMOTE, 8 features) and testing set (19,818 rows).
- **Training**: Each model was trained on the SMOTE-balanced training set, ensuring equal representation of both classes (50% diabetes, 50% no diabetes).
- **Evaluation Metrics**:
    - **Accuracy**: Overall correctness of predictions.
    - **Precision**: Proportion of positive predictions that were correct (important to minimize false positives in medical contexts).
    - **Recall**: Proportion of actual positives correctly identified (critical to minimize false negatives, ensuring diabetes cases are not missed).
    - **F1-Score**: Harmonic mean of precision and recall, balancing both metrics for imbalanced data.
    - **ROC-AUC**: Measures the model's ability to distinguish between classes, robust to imbalance.

3. Implementation Notes

- Models were implemented using scikit-learn (RandomForestClassifier, LogisticRegression, SVC) and XGBoost (XGBClassifier).

- Training was performed on the scaled features, ensuring compatibility with models like Logistic Regression and SVM, which are sensitive to feature scales.

Model Comparison

The performance of the four models was evaluated on both the training and testing sets using the metrics defined above. The results are summarized in the table below, extracted from model_comparison_results.csv.

Model	Set	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	Train	0.9993	0.9997	0.9989	0.9993	1.0000
Random Forest	Test	0.9654	0.8588	0.7130	0.7791	0.9668
Logistic Regression	Train	0.8898	0.8898	0.8899	0.8898	0.9648
Logistic Regression	Test	0.8896	0.4282	0.8633	0.5725	0.9609
SVM	Train	0.8908	0.8884	0.8938	0.8911	0.9646
SVM	Test	0.8893	0.4278	0.8680	0.5732	0.9609
XGBoost	Train	0.9821	0.9981	0.9661	0.9818	0.9985
XGBoost	Test	0.9695	0.9131	0.7118	0.8000	0.9777

Observations

- **Training Performance**:
    - **Random Forest**: Near-perfect scores across all metrics (Accuracy: 0.9993, ROC-AUC: 1.0000), indicating potential overfitting.
    - **XGBoost**: High performance (Accuracy: 0.9821, F1-Score: 0.9818), but slightly lower than Random Forest, suggesting less overfitting.
    - **Logistic Regression and SVM**: Similar performance (Accuracy ~0.89, ROC-AUC ~0.964), but significantly lower than tree-based models, likely due to their inability to capture non-linear patterns effectively.
- **Testing Performance**:
    - **XGBoost**: Best overall test performance (Accuracy: 0.9695, F1-Score: 0.8000, ROC-AUC: 0.9777), with the highest precision (0.9131), indicating fewer false positives.
    - **Random Forest**: High test accuracy (0.9654) but lower recall (0.7130) and F1-Score (0.7791) compared to XGBoost, suggesting it misses more diabetes cases.
    - **Logistic Regression and SVM**: Similar test accuracy (~0.889), but very low precision (~0.428) despite high recall (~0.86), leading to poor F1-Scores (~0.573). This indicates many false positives, making them less reliable for medical applications.
- **Overfitting**:
    - Random Forest and XGBoost show significant drops in performance from training to testing (e.g., Random Forest F1-Score: 0.9993 to 0.7791), indicating overfitting, though XGBoost is less affected.
    - Logistic Regression and SVM show minimal overfitting, with stable performance between training and testing, but their overall performance is suboptimal.

Result Analysis

The results provide insights into the models' effectiveness for diabetes prediction, their strengths and weaknesses, and their suitability for clinical applications.

1. Model Performance Analysis

- **XGBoost**:
    - **Strengths**: Highest test accuracy (0.9695), precision (0.9131), and ROC-AUC (0.9777), making it the most reliable model for distinguishing between diabetic and non-diabetic patients. Its F1-Score (0.8000) balances precision and recall effectively.

    - **Weaknesses**: Recall (0.7118) is slightly lower than Logistic Regression and SVM, meaning it misses some diabetes cases. Some overfitting is evident (train F1-Score: 0.9818 vs. test: 0.8000).
    - **Suitability**: Best overall model for this task, especially in medical contexts where high precision (fewer false positives) is crucial to avoid unnecessary interventions.
- **Random Forest**:
    - **Strengths**: High test accuracy (0.9654) and ROC-AUC (0.9668), indicating good overall performance. Precision (0.8588) is strong, though lower than XGBoost.
    - **Weaknesses**: Lower recall (0.7130) and F1-Score (0.7791) compared to XGBoost, meaning it misses more diabetes cases. Significant overfitting (train Accuracy: 0.9993 vs. test: 0.9654).
    - **Suitability**: A strong contender but less effective than XGBoost due to lower recall and more pronounced overfitting.
- **Logistic Regression and SVM**:
    - **Strengths**: High recall (~0.86–0.87), ensuring most diabetes cases are identified. Minimal overfitting (stable performance between train and test).
    - **Weaknesses**: Very low precision (~0.428), leading to many false positives and poor F1-Scores (~0.573). This makes them unreliable for clinical use, as false positives can lead to unnecessary treatments.
    - **Suitability**: Not suitable for this task due to poor precision and overall performance, despite high recall.

2. Clinical Implications

- **Best Model**: XGBoost is recommended for diabetes prediction due to its balanced performance (F1-Score: 0.8000) and high precision (0.9131). In a medical context, high precision minimizes false positives, reducing unnecessary patient anxiety and interventions.
- **Recall Consideration**: While XGBoost's recall (0.7118) is lower than Logistic Regression and SVM, it is still reasonable, and its overall performance outweighs the alternatives. Missing some diabetes cases is a concern, but the high ROC-AUC (0.9777) indicates strong class separation.
- **Feature Importance**: From EDA, HbA1c_level (corr: 0.40) and blood_glucose_level (corr: 0.42) are key predictors, likely driving XGBoost's performance. These align with medical knowledge, as levels above 6.5% (HbA1c) and high glucose are diagnostic for diabetes.

3. Limitations

- **Overfitting**: Random Forest and XGBoost exhibit overfitting, which may inflate training performance and reduce generalizability. Further regularization (e.g., reducing tree depth) or cross-validation could mitigate this.
- **Duplicates**: The 3,854 duplicated rows remain unaddressed, potentially inflating performance and risking data leakage.
- **Interpretability**: XGBoost and Random Forest are less interpretable than Logistic Regression, which may be a concern in clinical settings where explainability is critical.
- **Class Imbalance**: While SMOTE balanced the training set, the test set remains imbalanced (8.5% diabetes), which may skew metrics like precision and recall. Techniques like weighted loss functions could further improve performance.

4. Future Improvements

- **Address Duplicates**: Remove or analyze duplicated rows to prevent data leakage and ensure realistic performance.
- **Enhance Recall**: Adjust model thresholds or use cost-sensitive learning to improve recall for XGBoost, ensuring fewer diabetes cases are missed.
- **Improve Interpretability**: Use SHAP or LIME to explain XGBoost predictions, making it more suitable for clinical use.
- **Cross-Validation**: Implement k-fold cross-validation to better assess model generalizability and reduce overfitting.
- **External Validation**: Test the model on an independent dataset to confirm its performance in real-world scenarios.

5. Conclusion

XGBoost is the best-performing model for diabetes prediction in this study, with a test accuracy of 0.9695, precision of 0.9131, and ROC-AUC of 0.9777. It effectively identifies at-risk patients, supporting early intervention and personalized treatment planning in healthcare settings. However, addressing overfitting, duplicates, and interpretability is crucial before deployment. The model also enables researchers to explore the impact of medical and demographic factors on diabetes, with HbA1c_level and blood_glucose_level confirmed as key predictors.