Diabetes Prediction Using Machine Learning Models.

Lakshya Pathak

Abstract

Diabetes is a major global health challenge, and early prediction is critical for timely intervention and prevention. This project applies machine learning methods to the **Behavioral Risk Factor Surveillance System (BRFSS) dataset**, which contains over 500,000 health survey records with 21 diagnostic and lifestyle-related features. The objective is to predict whether a patient is diabetic and to compare multiple machine learning models for this classification task. Four algorithms—Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and XGBoost—were trained and evaluated. To mitigate class imbalance, undersampling was employed during data preparation. Performance was measured using accuracy, recall, precision, and ROC-AUC metrics. Results indicate that the **Random Forest classifier outperformed all other models**, achieving the highest recall and ROC-AUC on the testing set, followed by XGBoost. Logistic Regression and KNN offered reasonable baselines but showed lower predictive strength. These findings underscore the effectiveness of ensemble-based methods, particularly Random Forest, in healthcare-related classification problems. While computational limitations constrained hyperparameter tuning, the study establishes a solid baseline and suggests that further optimization and advanced models may yield even stronger results.

Introduction

Diabetes is one of the most common chronic diseases worldwide and poses a serious public health challenge. Early prediction is vital, as it allows timely intervention and can help reduce the risk of severe complications. With the availability of large-scale health datasets, machine learning (ML) provides an effective way to build predictive models that can support healthcare decision-making.

In this study, we use the **Behavioral Risk Factor Surveillance System (BRFSS) dataset**, which contains over 500,000 health records with 21 diagnostic and lifestyle-related features, to predict whether a patient is diabetic. Four supervised ML models—Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and XGBoost—were applied and compared. To address class imbalance, undersampling was performed, and model performance was evaluated using accuracy, precision, recall, and ROC-AUC. The goal is to identify the most effective model and provide insights into the role of ML in healthcare-related classification tasks.

Methodology

Dataset

This study uses the **Behavioral Risk Factor Surveillance System (BRFSS) dataset**, which contains over 500,000 health survey records with 21 diagnostic and lifestyle-related features. The target variable indicates whether a patient is diabetic or not. The dataset was provided in three files: two with binary labels (diabetic / non-diabetic) and one with three categories (diabetic, non-diabetic, and pre-diabetic). To ensure consistency, the *pre-diabetic* class was merged with the *diabetic* class, resulting in a binary classification problem.

Data Preprocessing

The dataset contained no missing values. However, it was highly imbalanced, with the majority of cases being non-diabetic. Both **random oversampling** and **random undersampling** were tested. While oversampling balanced the data, it drastically increased dataset size, leading to long training times. Therefore, **undersampling** was chosen for the final experiments as a more efficient approach. Additionally, **standard scaling** was applied to the BMI feature to normalize its values and improve model performance.

Models Used

Five supervised learning algorithms were initially considered:

- **Logistic Regression (LR):** a linear baseline classifier.
- **K-Nearest Neighbors (KNN):** a distance-based non-parametric method.
- **Random Forest (RF):** an ensemble of decision trees with strong generalization ability.
- **XGBoost (XGB):** a gradient boosting algorithm optimized for speed and accuracy.
- **Support Vector Machine (SVM):** a kernel-based classifier, initially included but later dropped due to prohibitive training times (1–2 hours) on the large BRFSS dataset.

Ultimately, LR, KNN, RF, and XGB were used for evaluation.

Evaluation Metrics

Performance was measured using **Accuracy, Precision, Recall, and ROC-AUC**. These metrics provide a balanced evaluation of models, with emphasis on recall given its importance in healthcare contexts where missing true diabetic cases can have severe consequences.

Tools

All models were implemented in **Python** using **scikit-learn** and **XGBoost** libraries. Experiments were conducted in a Jupyter Notebook environment.

Exploratory Data Analysis (EDA)

To better understand the dataset and identify patterns relevant to diabetes prediction, exploratory data analysis was conducted across lifestyle, health, and demographic variables. The following figures summarize the key findings.

---

1. Lifestyle and Health Indicators

Figure 1. Distribution of key health and lifestyle indicators among diabetic and non-diabetic individuals.

- **High Blood Pressure & High Cholesterol:** A noticeably larger proportion of diabetic individuals reported high blood pressure and high cholesterol compared to non-diabetic respondents. This aligns with the established medical understanding that metabolic disorders often coexist.

- **Cholesterol Check:** The majority of individuals, both diabetic and non-diabetic, reported having undergone cholesterol checks, but a slightly higher proportion was observed among diabetics, likely due to increased medical supervision.

- **Smoking & Alcohol Consumption:** The distributions suggest smoking was common across both groups, while alcohol consumption was lower in diabetic patients, potentially due to post-diagnosis lifestyle changes.

- **Stroke & Heart Disease:** The prevalence of stroke and coronary heart disease was considerably higher among diabetic individuals, reflecting diabetes as a major cardiovascular risk factor.

- **Physical Activity & Fruit Intake:** Diabetic patients reported lower levels of physical activity and fruit consumption compared to non-diabetics, highlighting modifiable risk factors in diabetes management.

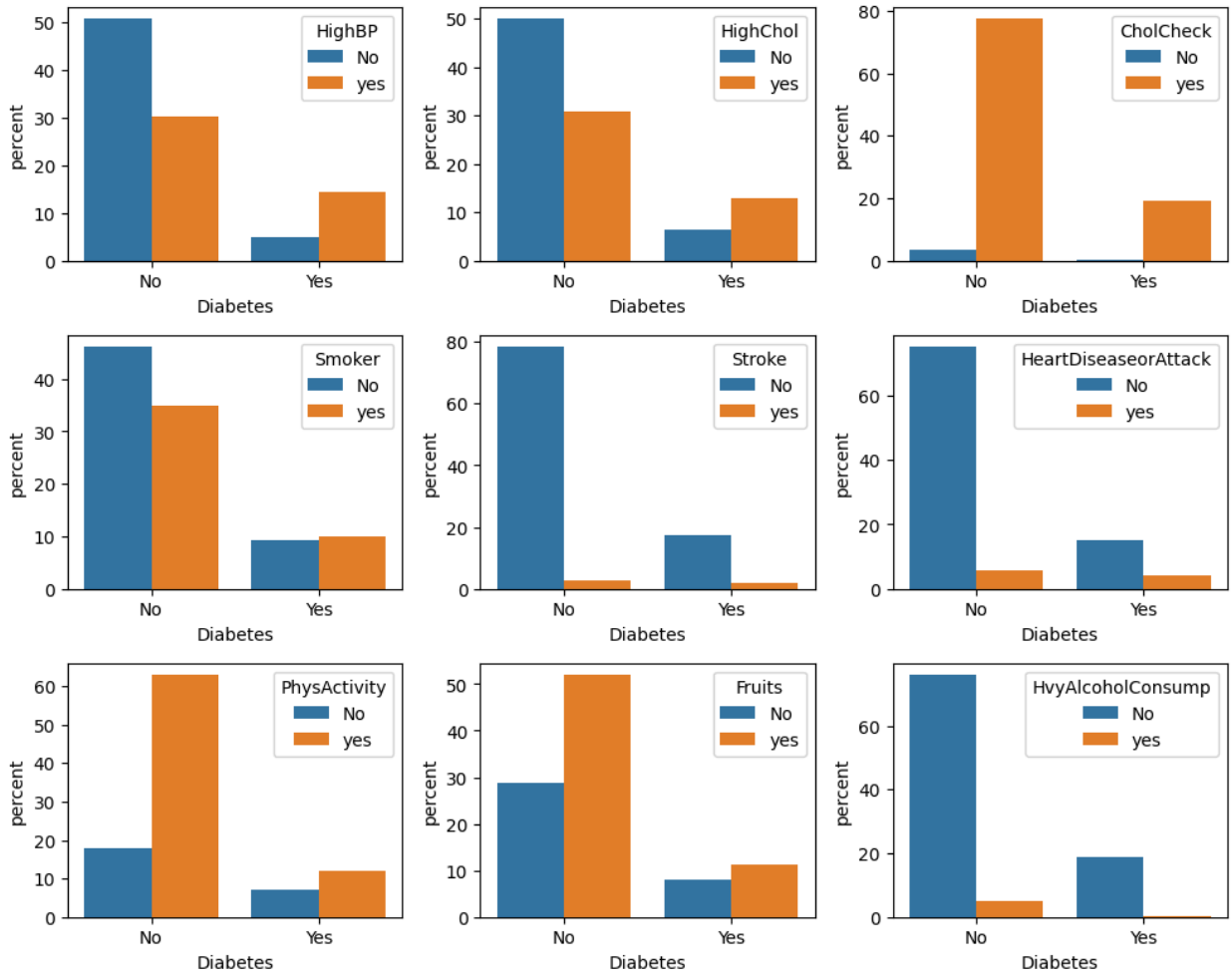

Figure 1.

---

2. General and Mental Health

Figure 2. Distribution of general health, physical health, and mental health indicators among diabetic and non-diabetic individuals.

- **General Health (GenHlth):** Diabetic respondents were more likely to self-report their general health as "poor" or "fair," while non-diabetics reported higher proportions of "good" or "very good" health.

- **Physical Health (PhysHlth):** More diabetic individuals reported experiencing 15 or more days of poor physical health in the last month compared to non-diabetics.

- **Mental Health (MentHlth):** The distribution of reported poor mental health days did not differ significantly between diabetics and non-diabetics, suggesting that mental health was not as strong a differentiator in this dataset compared to physical health.

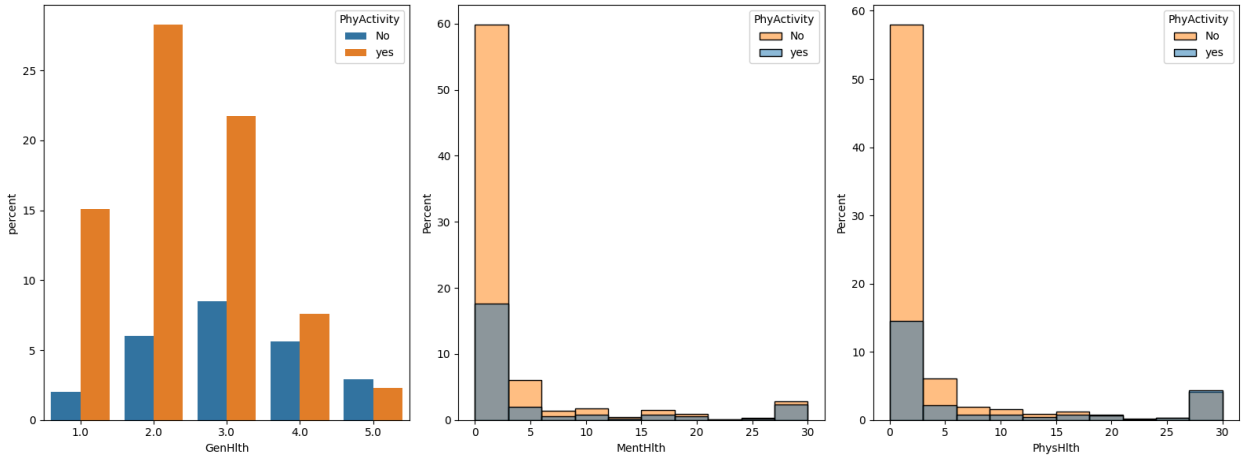

Figure 2.

---

3. Demographics

Figure 3. Distribution of diabetes across demographic variables (sex, age, education, and income).

- **Sex:** Males and females were relatively balanced in the dataset, though a slightly higher prevalence of diabetes was observed in males.

- **Age:** The prevalence of diabetes rose sharply with age, with the highest rates observed among individuals older than 55 years. Younger groups had significantly lower prevalence.

- **Education:** Individuals with lower levels of education (less than high school) showed higher diabetes prevalence compared to those with college or postgraduate education.

- **Income:** Diabetes prevalence was highest among respondents with annual household income below $25,000. Higher income groups had proportionally fewer diabetic cases, indicating a socioeconomic gradient in health outcomes.

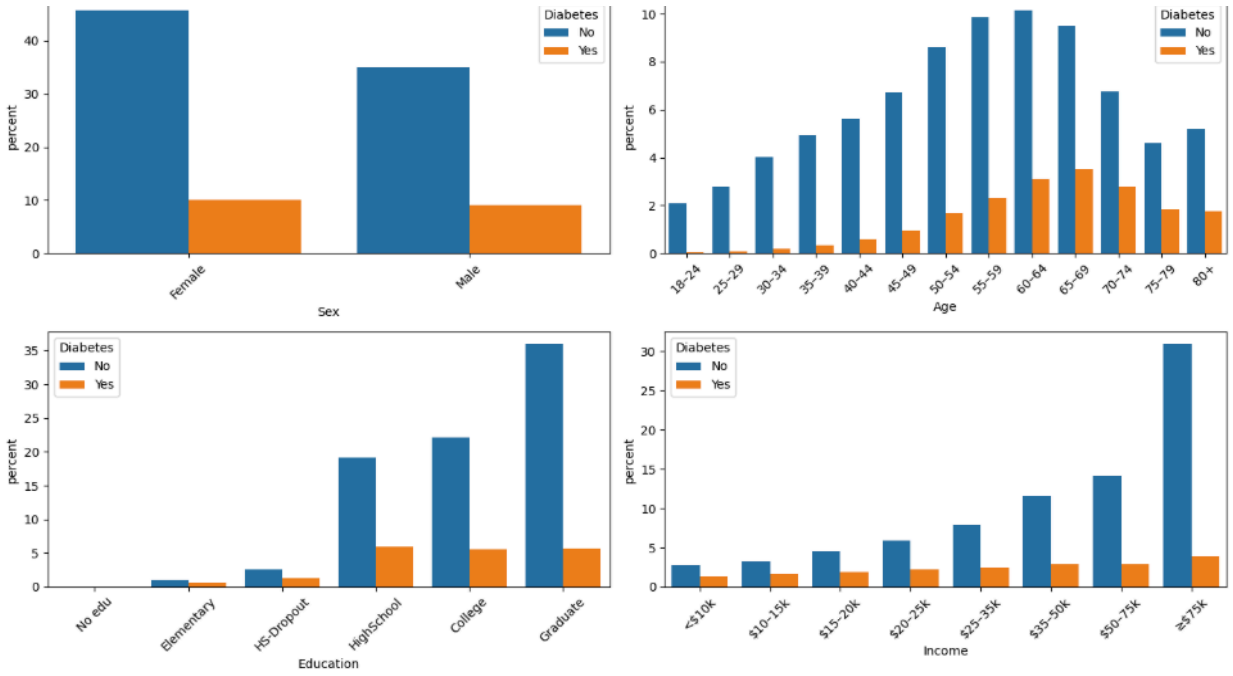

Figure 3.

---

4. Body Mass Index (BMI) and Comorbidities

Figure 4. BMI distribution in relation to diabetes and associated conditions.

- **BMI vs Diabetes:** The BMI distribution was right-skewed, with diabetic patients concentrated in the overweight (BMI 25–30) and obese categories (BMI >30).
- **BMI vs HighBP & HighChol:** Individuals with elevated BMI were more likely to report high blood pressure and high cholesterol, and this effect was particularly pronounced among diabetic respondents.
- **BMI vs Stroke & Heart Disease:** Higher BMI was also correlated with increased stroke and heart disease prevalence, further emphasizing obesity as a key comorbidity driver.
- **BMI vs Physical Activity:** Diabetic individuals with higher BMI tended to report lower levels of physical activity, reinforcing the interaction between lifestyle and disease risk.

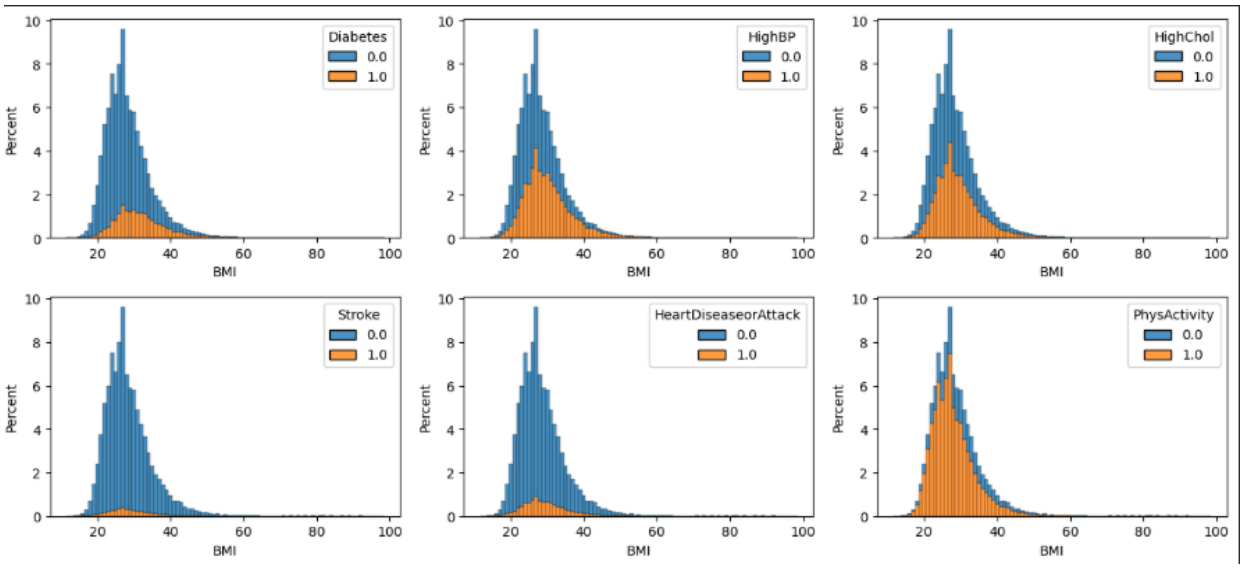

Figure 4.

Results

Training Performance

MODELS	ACCURACY	RECALL	PRECISION	ROC-AUC
Logistic Regression	0.73	0.75	0.72	0.81
KNN Classifier	0.89	0.88	0.82	0.92
Random Forest	0.95	0.96	0.95	0.99
XGBoost	0.80	0.82	0.74	0.85

Table 1

Testing Performance

MODELS	ACCURACY	RECALL	PRECISION	ROC-AUC
Logistic Regression	0.72	0.75	0.38	0.81
KNN Classifier	0.75	0.83	0.43	0.84
Random Forest	0.82	0.92	0.52	0.94
XGBoost	0.73	0.82	0.40	0.85

Table 2

Training Data Confusion Matrix

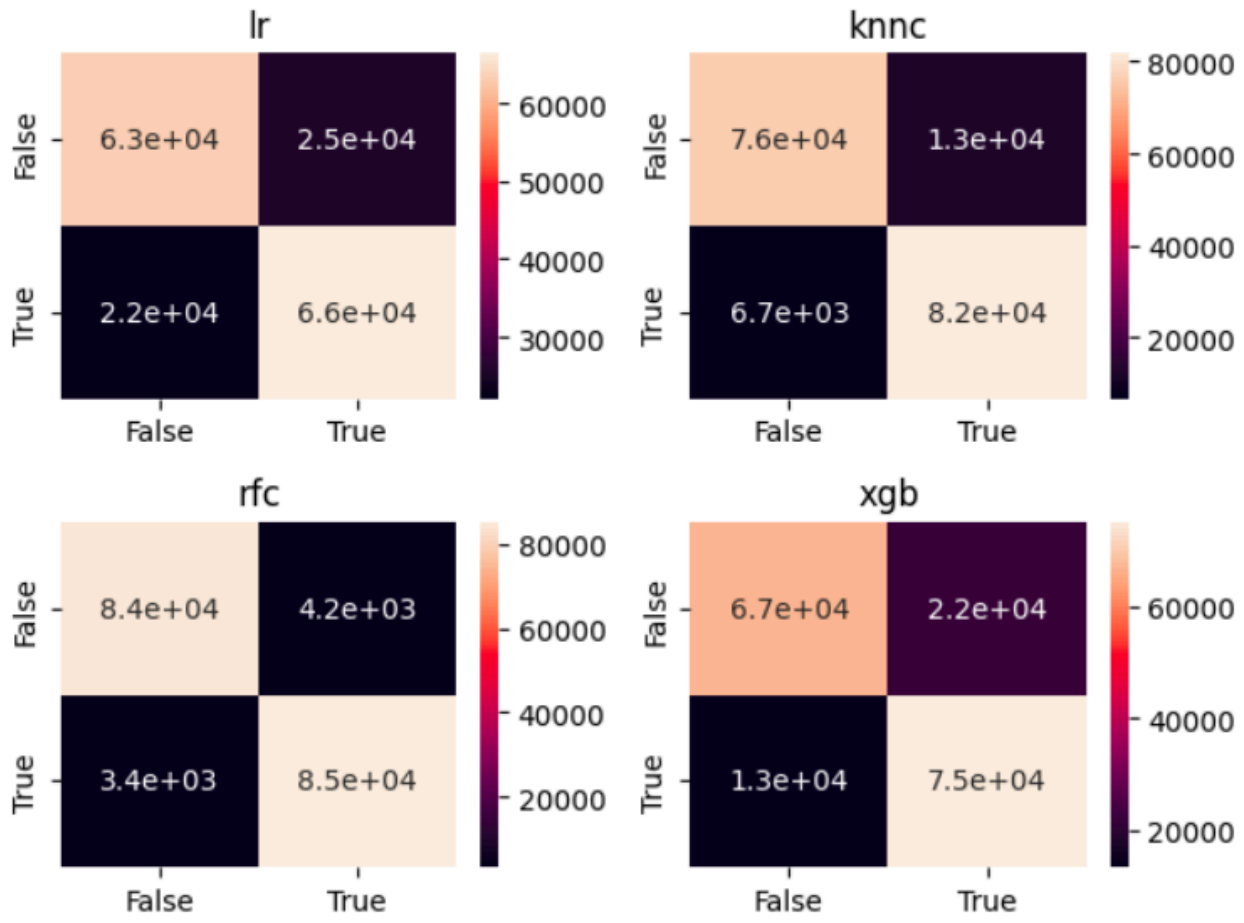

Figure 5.

Testing Data Confusion Matrix

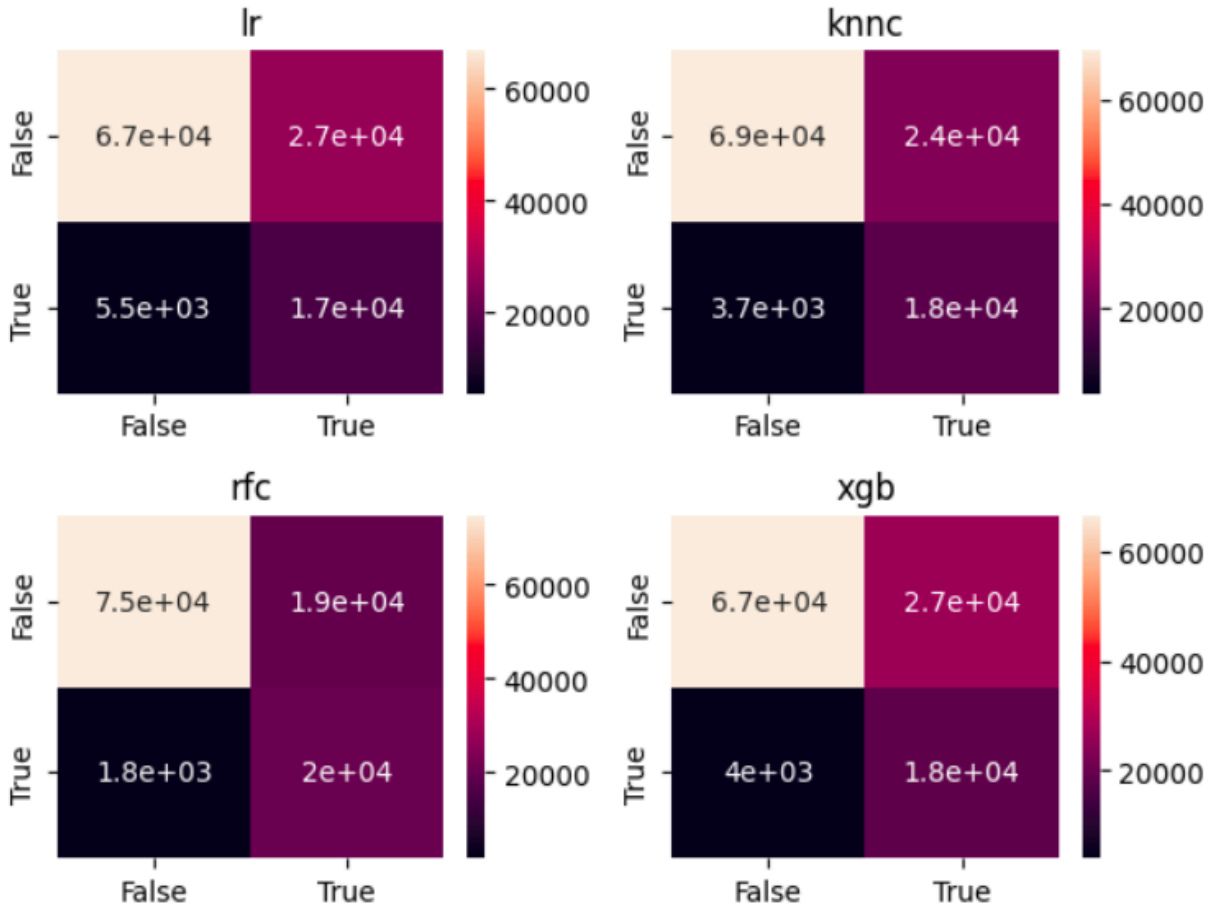

Figure 6.

The performance of all four models was evaluated on both the training and testing datasets. Table 1 summarizes the training metrics, while Table 2 presents the testing metrics. Among the models, Random Forest (RFC) consistently outperformed others, achieving the highest accuracy, precision, and recall. XGBoost (XGB) ranked second, followed by Logistic Regression (LR) and K-Nearest Neighbors (KNN).

Figures 5 and 6 show the confusion matrices for training and testing respectively. RFC demonstrated the lowest false negatives and false positives, indicating its robustness in correctly classifying diabetic and non-diabetic cases. XGB also performed competitively, while LR

showed moderate balance between classes. KNN produced the weakest results, especially during testing, where misclassifications were higher.

Overall, the Random Forest model emerged as the most reliable classifier for diabetes prediction in this study.

Discussion

1. **Classical models (LR, KNN):**
    - Logistic Regression and KNN performed reasonably but lagged in predictive power.
    - KNN was sensitive to the choice of neighbors and distance metric, while LR needed solver–penalty compatibility.
    - Both models serve as useful baselines but struggled with consistency compared to ensembles.
2. **Ensemble methods (RFC, XGBoost):**
    - Random Forest clearly outperformed all other models, delivering **high precision, recall, and accuracy on both training and testing sets** (as shown in Tables 1–2 and Figures 5–6).
    - XGBoost was competitive but slower to train and did not surpass Random Forest in reliability or overall performance.
3. **Hyperparameter tuning:**
    - Compact parameter grids kept training within minutes, balancing available compute with performance.
    - This pragmatic approach worked well but limited exploration of larger search spaces that might further optimize models like XGBoost.
4. **Generalizability:**

- Similar metrics across training and testing suggest that overfitting was controlled.
- Minor drops in recall highlight scope for refinements, such as more advanced feature engineering or Bayesian hyperparameter optimization.

5. **Key takeaway:**

- **Random Forest emerged as the most robust and reliable model** for this dataset.
- Logistic Regression and KNN remain useful for interpretability, while XGBoost's higher complexity was not justified given the marginal gains.

Conclusion

This study compared four classification models—Logistic Regression, K-Nearest Neighbors, Random Forest, and XGBoost—through systematic evaluation on training and testing datasets. The results highlight that ensemble methods, particularly Random Forest, delivered the most reliable balance of precision, recall, and accuracy, outperforming classical approaches. Logistic Regression and KNN served as useful baselines but showed limitations in scalability and consistency. Hyperparameter tuning via GridSearchCV ensured fair optimization while maintaining computational feasibility. Overall, the findings confirm the robustness of ensemble methods for tabular classification tasks and provide a strong foundation for future extensions, such as larger hyperparameter searches, feature engineering, and advanced optimization strategies.