Proposal for a Diabetes Prediction System

Group conceptual project

Group 26 - Rhea Nair & Samiksha Sarda

Table of Contents

Business Understanding

As a healthcare startup, our primary goal is to leverage predictive analytics to enhance decision-making for healthcare systems, particularly in diabetes management. Diabetes is a critical public health challenge, and our predictive model aims to assist healthcare providers in identifying individuals at high risk of diabetes. By enabling early detection and intervention, we aim to reduce healthcare costs and improve patient outcomes.

The business objective is to develop a classification model to predict the likelihood of diabetes based on a variety of demographic, lifestyle, and medical variables. Predictor variables include factors like age, BMI, blood glucose levels, and smoking history. The target variable, diabetes, is binary, classifying individuals as either diabetic or non-diabetic. Success metrics include achieving an accuracy of 90% or higher, improving early diagnosis rates by 15%, and integrating predictive insights into healthcare management platforms seamlessly.

Key stakeholders include data scientists, project managers, and healthcare providers. This report reflects the preliminary development phase based on an initial dataset and serves as the foundation for a scalable solution.

Data Understanding

The dataset used in this project contains a variety of attributes relevant to diabetes prediction. These include demographic variables such as age and gender, behavioral variables like smoking history, and

clinical measures such as BMI and blood glucose levels. The data serves as a valuable resource for identifying patterns and trends associated with diabetes risk.

However, like most real-world datasets, this one is not without its challenges. Missing values are present in critical fields such as BMI, which, if left unaddressed, could compromise the reliability of the model. Additionally, there are anomalies in the data, including extreme values in blood glucose levels, which need to be carefully examined and handled to prevent skewing the model's predictions. Another significant challenge is the imbalance in the dataset, with far more non-diabetic cases than diabetic ones. This imbalance could bias the model towards predicting non-diabetic cases unless specific measures are taken to address it.

Despite these limitations, the dataset's strengths lie in its variety and relevance. Variables like HbA1c levels and blood glucose levels are strongly correlated with diabetes and are expected to play a crucial role in the model's accuracy. The dataset offers a comprehensive framework for analyzing the risk factors associated with diabetes, making it a suitable foundation for predictive modeling.

Data Preparation

The data preparation process was divided into four distinct phases: data integration, data cleaning, data transformation, and data reduction. Each phase was critical to ensure the dataset was reliable, streamlined, and ready for modeling.

Phase 1: Data Integration

The first phase involved consolidating data from various sources into a unified dataset. The initial dataset consisted of 100,000 rows and 9 columns, containing a mix of demographic, behavioral, and clinical

variables such as age, BMI, smoking history, and blood glucose levels. These attributes were loaded into a Pandas DataFrame, ensuring compatibility with Python-based data analysis and machine-learning libraries.

Redundant and irrelevant columns that could complicate the analysis, such as uninformative IDs or non-predictive metadata, were excluded during this step. The integration process ensured that the data was centralized and that only the most relevant attributes were retained for analysis.

Phase 2: Data Cleaning

The second phase focused on improving the quality and reliability of the data. Missing values, identified in columns such as BMI and smoking history, were imputed using a mean imputation method. This approach preserved the overall distribution of the dataset without introducing significant bias.

Anomalies and outliers in key variables like blood glucose levels and BMI were detected using box plots and statistical techniques. While these anomalies were identified, no rows were removed from the dataset. Instead, missing values in variables like BMI and blood glucose levels were filled by calculating and imputing the mean, ensuring the dataset remained complete and consistent. This approach preserved the dataset's size of 100,000 rows while improving its overall quality and reliability.

Phase 3: Data Transformation

This phase focused on preparing the dataset for machine-learning algorithms by encoding categorical variables like smoking history into numerical formats and standardizing numerical variables to ensure all predictors were on a comparable scale. These transformations enhanced interpretability and ensured effective model training without altering the dataset's original structure or adding new variables.

Outlier capping was applied to handle extreme values in key variables like BMI and blood glucose levels, preserving consistency and reliability. Using the Interquartile Range (IQR) method, outliers were defined as values falling below Q1−1.5×IQR or above Q3+1.5×IQR. These values were capped to the nearest acceptable boundary using the clip function, ensuring no data points were removed. This approach preserved the dataset's size while minimizing the potential influence of outliers, avoiding distortions in the model's learning process.

Phase 4: Data Reduction

The data was skewed, meaning there were significantly more non-diabetic cases than diabetic cases, which could bias the model toward predicting non-diabetic outcomes. To address this, we performed under-sampling, a technique that balances the dataset by reducing the number of samples from the majority class (non-diabetic cases) to match the number of samples in the minority class (diabetic cases).

Under-sampling helps create a balanced dataset by keeping all diabetic cases and selecting a smaller, equal number of non-diabetic cases. This ensures that the model has an equal opportunity to learn from both classes, improving its ability to correctly identify diabetic cases without being overly biased by the larger number of non-diabetic examples.

Descriptive Statistics

- Age: Mean: 41.89, Standard Deviation: 22.47
- BMI: Mean: 27, Standard Deviation: 5.65
- Blood glucose: Mean: 137.50, Standard Deviation: 38.64

Modeling

Model Selection

To predict diabetes, three modeling techniques were evaluated to determine the best fit for the dataset's characteristics: Random Forest, XGBoost, and Decision Tree. After a thorough comparison, XGBoost emerged as the most effective model, particularly for handling the dataset's skewed distribution and identifying underrepresented diabetic cases.

Why This Model?

XGBoost, or Extreme Gradient Boosting, is an advanced machine-learning algorithm known for its ability to build an ensemble of decision trees sequentially. In this process, each new tree corrects errors made by the previous ones, continuously improving the model's accuracy. This iterative refinement, combined with its application of both classification and regression trees, makes XGBoost exceptionally capable of capturing complex patterns in imbalanced datasets like ours.

Reasons for Choosing XGBoost:

1. **Handling Imbalanced Data:** XGBoost incorporates a weighted loss function that prioritizes minority classes. This feature enabled the model to give greater importance to diabetic cases, effectively reducing the likelihood of false negatives. In contrast, Random Forest aggregates results without explicitly addressing class imbalances, and Decision Trees often overfit, focusing excessively on dominant classes.
2. **Regularization for Robustness:** To prevent overfitting, XGBoost employs L1 (Lasso) and L2 (Ridge) regularization techniques, ensuring the model generalizes well to unseen data. This approach

enhances performance stability, which is critical in healthcare applications where reliability is paramount.

3. **Efficient and Versatile Design:** XGBoost integrates tree pruning and handles missing values efficiently, which streamlined the model training process. These capabilities allowed it to accommodate the dataset's imperfections, such as missing entries, while maintaining high accuracy.

Model Configuration

The XGBoost model was carefully designed to optimize its performance in predicting diabetes by configuring key hyperparameters and leveraging its advanced capabilities. The model was built using the XGBClassifier, a powerful implementation of the gradient boosting algorithm specifically tailored for classification tasks. The configuration included setting the number of estimators to 500, allowing the model to construct up to 500 decision trees. Each tree contributed to the overall prediction by addressing errors from previous iterations, enabling the model to improve sequentially. The learning rate was set to 0.1, striking a balance between learning speed and precision. This ensured the model captured patterns effectively without overreacting to noise in the data.

The parameter gamma, which controls the minimum reduction in loss required to make a split at a node, was set to 0, enabling the model to explore all possible splits and search comprehensively for optimal tree structures. Additionally, a random state of 16 was specified to ensure reproducibility, maintaining consistency in the model's behavior and results across different runs. The dataset was divided into training and testing subsets, with 80% used for training the model and 20% reserved for testing. This split ensured the model was thoroughly evaluated on unseen cases, reflecting its real-world applicability.

The XGBoost model's boosting framework sequentially constructed trees, each correcting the errors made by its predecessors. This iterative learning process allowed the model to effectively capture complex

relationships between variables such as age, BMI, and blood glucose levels. Its tree-based structure prioritized key predictors like HbA1c levels and blood glucose levels, which are highly correlated with diabetes risk. Though gamma was set to 0 in this configuration, XGBoost inherently employs L1 and L2 regularization, penalizing overly complex models to prevent overfitting and ensuring the model generalizes well to new data.

XGBoost also handled missing values efficiently, automatically determining the best way to split data points with incomplete entries. This feature streamlined preprocessing and improved robustness. With the combination of 500 estimators and a moderate learning rate, the model had sufficient capacity to learn intricate patterns in the data while avoiding overfitting. This design ensured the XGBoost model accurately classified diabetic and non-diabetic cases, effectively handled imbalanced data, and delivered reliable predictions aligned with the project's goals.

Ultimately, XGBoost outperformed both Random Forest and Decision Tree across key metrics, including accuracy and recall, especially for the diabetic class. Its superior ability to manage imbalanced data, coupled with its advanced features and robustness, made it the ideal choice for this project.

Evaluation

The XGBoost model demonstrated strong performance in predicting diabetes, achieving an overall accuracy of 97%, indicating its reliability in classifying most cases correctly. For non-diabetic cases (class 0), the model achieved a precision of 0.97 and a recall of 1.00, with an F1-Score of 0.98, reflecting excellent performance. For diabetic cases (class 1), the model achieved a precision of 0.95, a recall of 0.69, and an F1-Score of 0.80, showing solid performance but with room for improvement in recall.

The macro average metrics, providing the unweighted mean across classes, showed a precision of 0.96, recall of 0.84, and F1-Score of 0.89. Weighted averages, accounting for class imbalance, were consistent at 0.97, aligning with the high overall accuracy. The confusion matrix revealed 18,249 correctly classified non-diabetic cases and 1,164 diabetic cases, alongside 55 false positives and 532 false negatives. An AUC-ROC score of 0.84 highlighted the model's strong ability to distinguish between the two classes, even with an imbalanced dataset.

From a business perspective, the model's high precision ensures reliable identification of diabetic cases, reducing unnecessary follow-ups for non-diabetic individuals. However, the recall for diabetic cases indicates 31% of actual cases were missed, potentially delaying diagnosis and treatment. Addressing this limitation through hyperparameter tuning, class weighting, or sampling methods like SMOTE could improve recall. Despite this, the XGBoost model delivers actionable insights and demonstrates high reliability, meeting the project's goals.

Deployment

Deploying the XGBoost model involves integrating it into healthcare systems for real-time diabetes risk prediction. The trained model will be exported using tools like pickle or joblib and integrated into web-based applications or APIs, enabling healthcare providers to input patient data and receive instant predictions. Containerization tools such as Docker will ensure consistent performance across systems.

Post-deployment, automated monitoring will track performance metrics like accuracy and recall. If performance drift occurs due to changing demographics or trends, the model will be retrained using updated data. An automated pipeline with tools like Apache Airflow will streamline retraining schedules, while feedback loops with healthcare stakeholders will validate predictions and refine functionality.

These deployment strategies ensure the XGBoost model provides accurate predictions, integrates seamlessly into workflows, and delivers measurable benefits to healthcare providers and patients.

Bibliography

1. E. Selvin, "Declining Diabetes Incidence Rates in the U.S.," *JAMA Internal Medicine*, vol. 181, no. 2, pp. 223-224, 2021. [Online]. Available: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2775594.
2. S. Smith et al., "Advances in Diabetes Prediction Models," *PubMed Central*, vol. 18, pp. 1123-1145, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10762598/.
3. Y. Li et al., "Performance Comparison of Machine Learning Models for Diabetes Prediction," *PubMed Central*, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9690067/.
4. Centers for Disease Control and Prevention, "Health, United States, 2022 – Data Finder," *CDC National Center for Health Statistics*. [Online]. Available: https://www.cdc.gov/nchs/hus/data-finder.htm.

Appendix

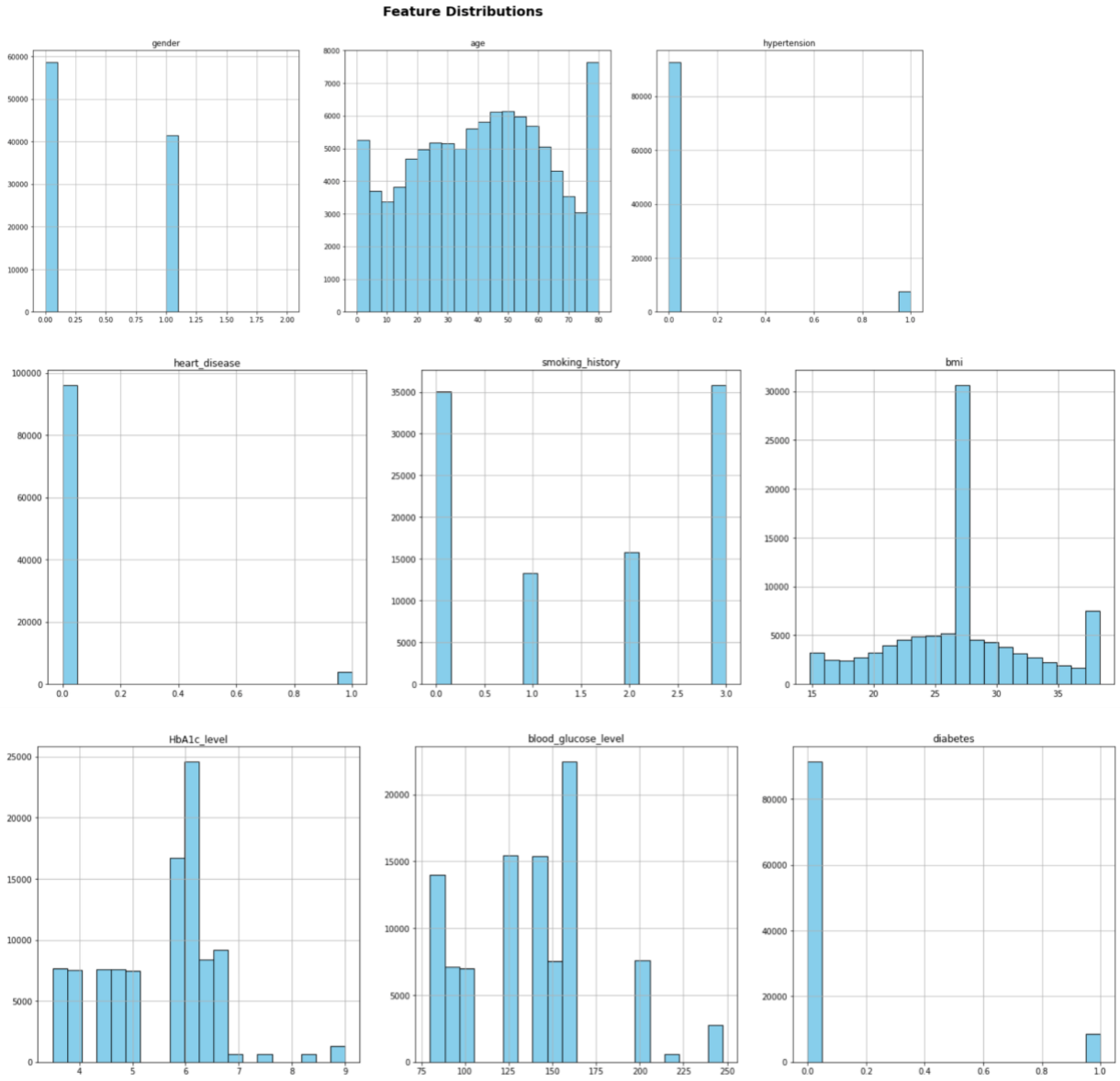
Diabetes Prediction System:

https://colab.research.google.com/drive/1XdtCJ9EgmFukGQmIFfSn_X9oBqQ6U5OS?usp=sharing

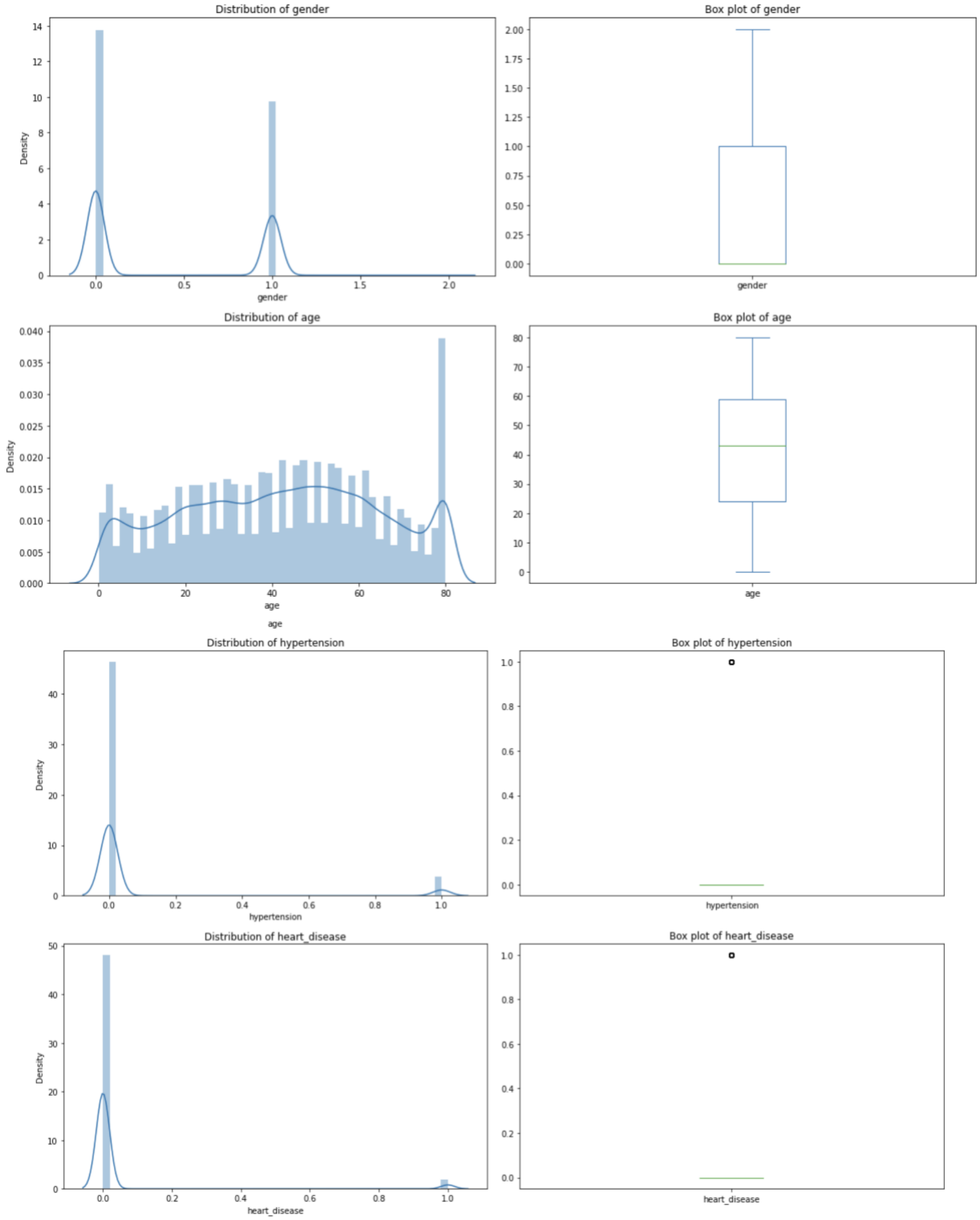
Dataset can be accessed from:

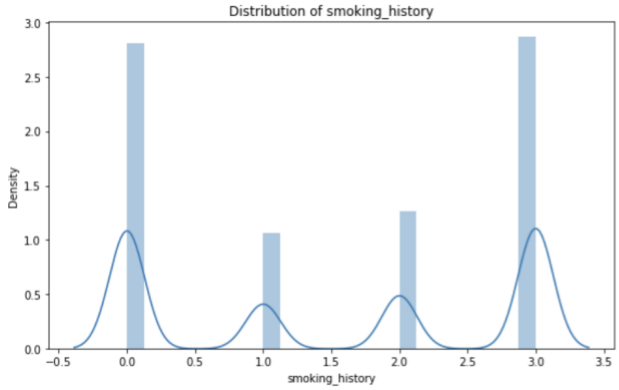
https://www.kaggle.com/search?q=diabetes+prediction

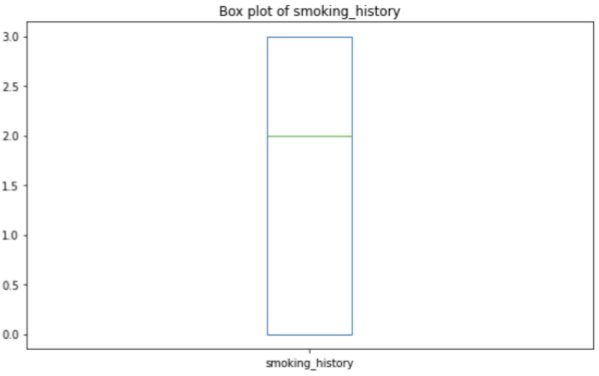
Feature Distributions

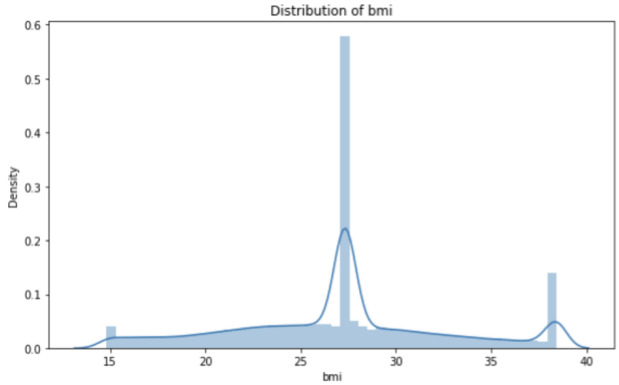


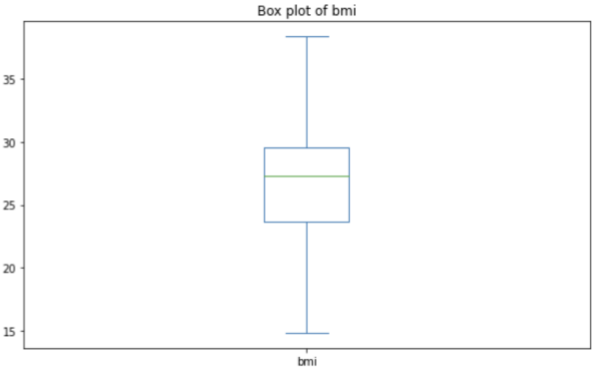
Feature distribution and box plot

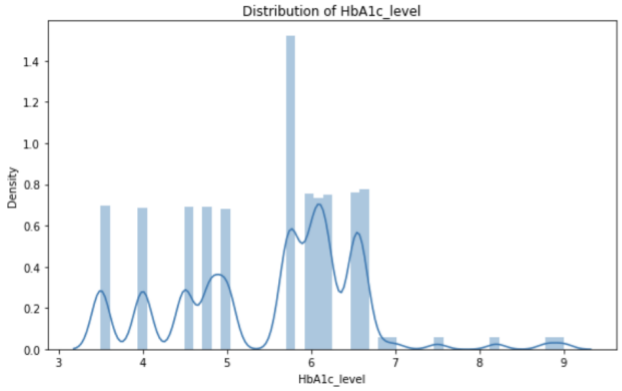


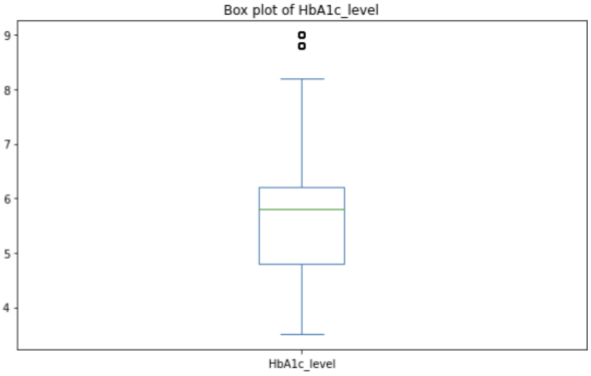

Distribution of smoking_history
Density
smoking_history

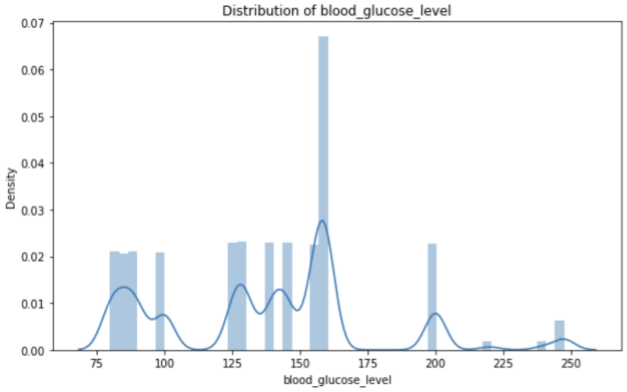

Box plot of smoking_history
smoking_history


Distribution of bmi
Density
bmi

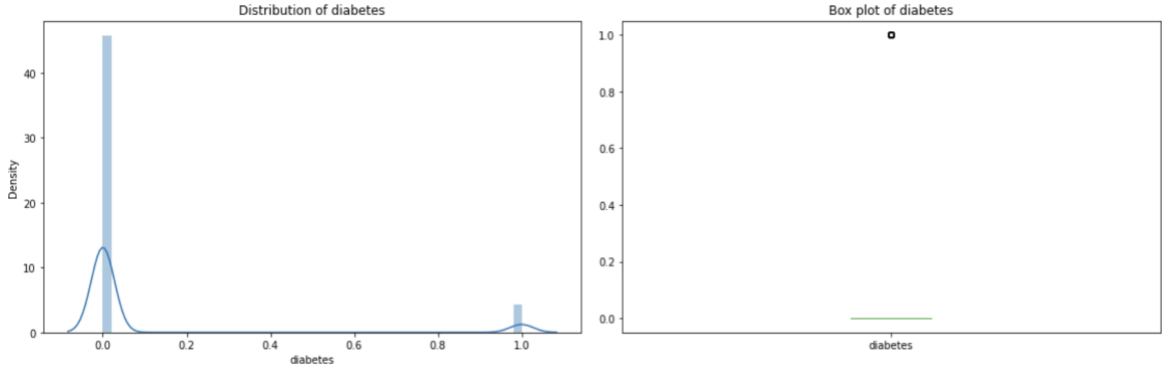

Box plot of bmi
bmi


Distribution of HbA1c_level
Density
HbA1c_level

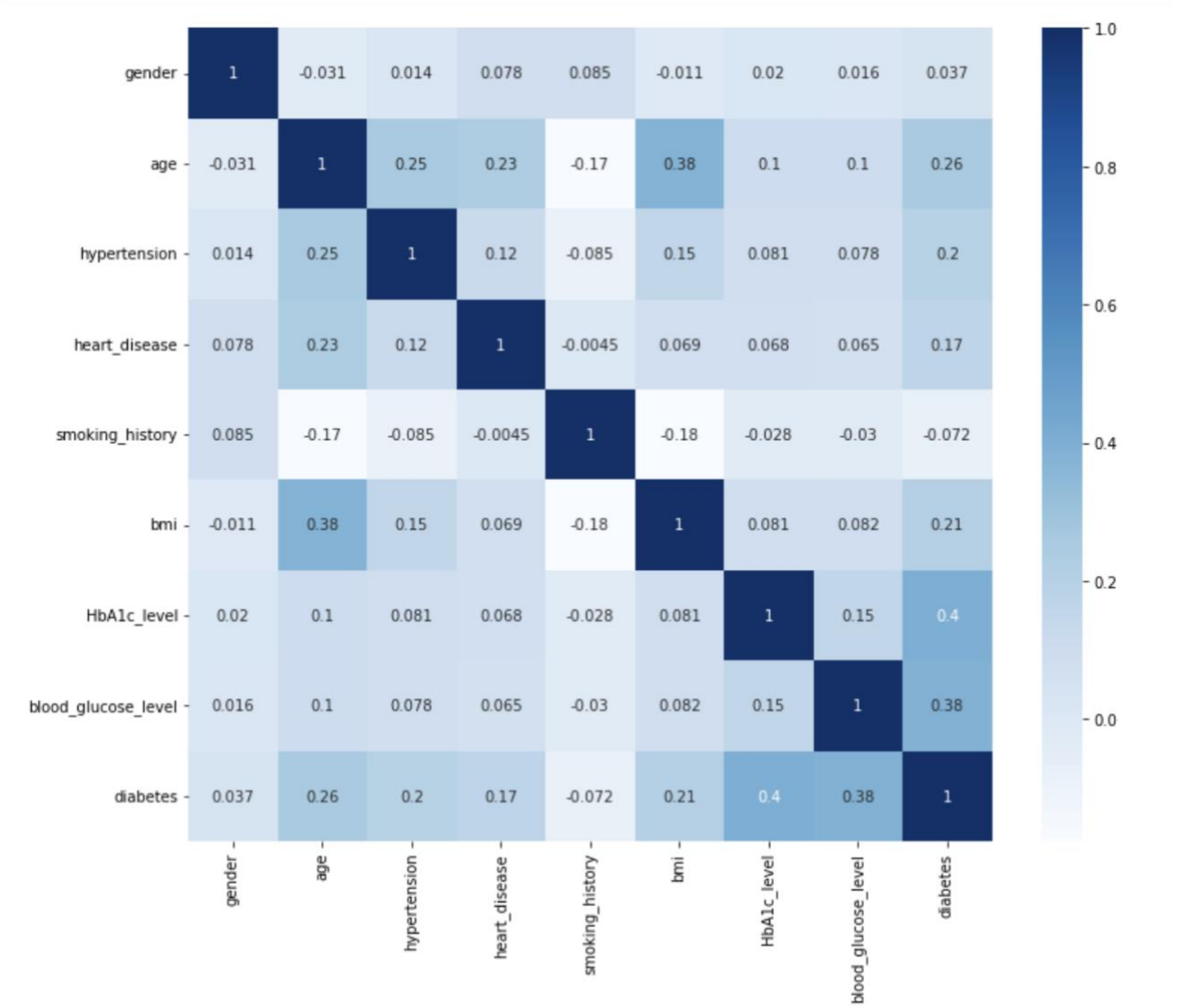

Box plot of HbA1c_level
HbA1c_level


Distribution of blood_glucose_level
Density
blood_glucose_level

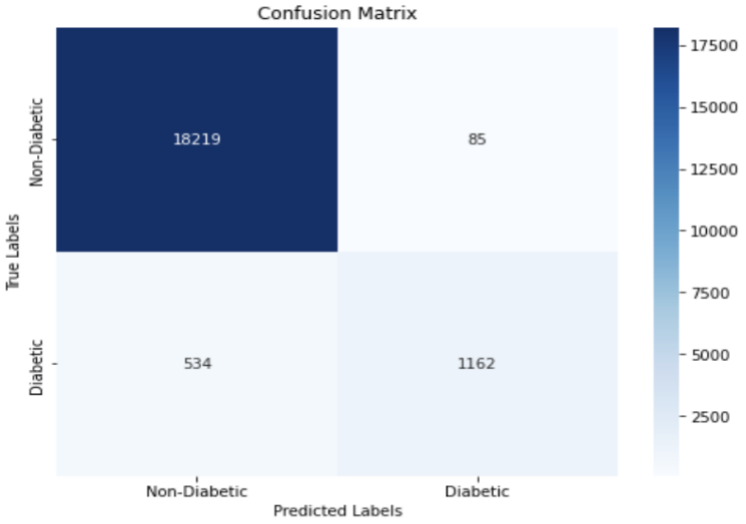

Box plot of blood_glucose_level
blood_glucose_level


Distribution of diabetes
Density
diabetes
Box plot of diabetes
diabetes

Correlation Matrix

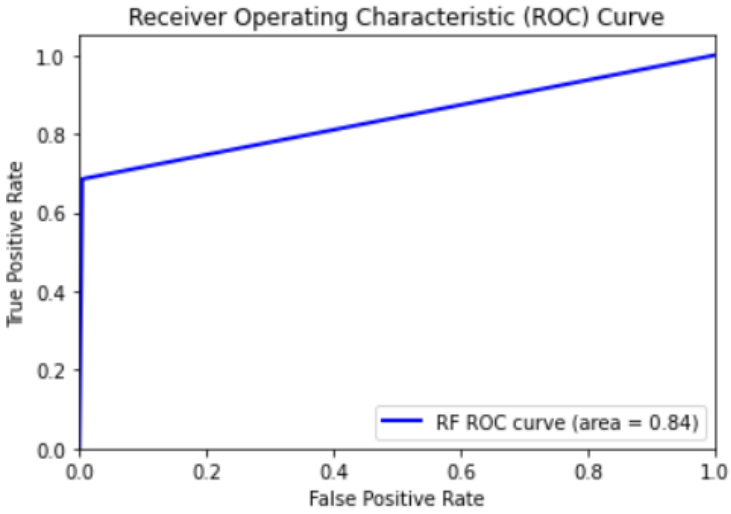


Random Forest: Confusion Matrix

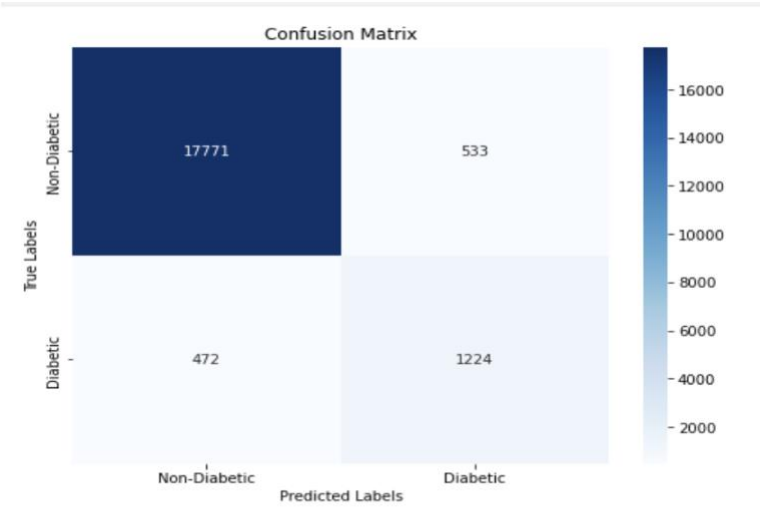



```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98     18304
           1       0.93      0.69      0.79      1696

    accuracy                           0.97     20000
   macro avg       0.95      0.84      0.89     20000
weighted avg       0.97      0.97      0.97     20000
```

Random Forest: ROC Curve

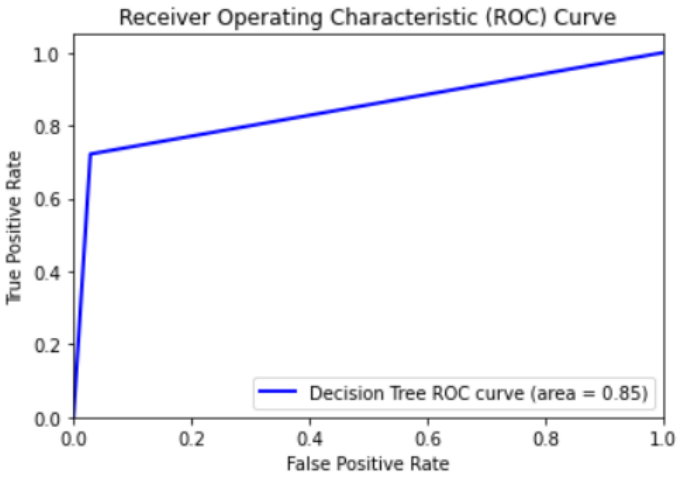


Decision Tree: Confusion Matrix

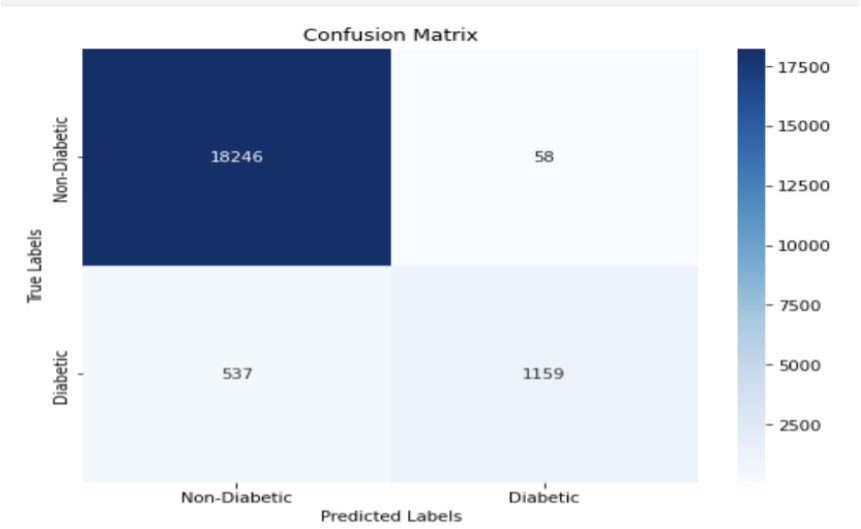



```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97     18304
           1       0.70      0.72      0.71      1696

    accuracy                           0.95     20000
   macro avg       0.84      0.85      0.84     20000
weighted avg       0.95      0.95      0.95     20000
```

Decision Tree: ROC Curve



XG Boosting: Confusion Matrix



Classification Report:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	18304
1	0.95	0.68	0.80	1696
accuracy			0.97	20000
macro avg	0.96	0.84	0.89	20000
weighted avg	0.97	0.97	0.97	20000

XG Boosting: ROC Curve