CMPSC 445 (Spring 2024)
Machine Learning

Title: "Predictive Analytics for Early Detection of Diabetes"

Team Members: Sarim Kazmi, Syed Kazmi

TABLE OF CONTENTS

Introduction

Problem Statement

The World Health Organization reports that the prevalence of diabetes has significantly increased over the previous few decades, raising concerns about the disease's development on a global scale. If this chronic illness is not treated, it can result in major side effects such as kidney failure, cardiovascular disease, blindness, and amputation of a lower limb. We are addressing the issue of early diabetes detection, which can be difficult because the disease is sneaky and frequently shows no symptoms in its early stages. Our software uses machine learning techniques to forecast a person's risk of developing diabetes based on health data that can be obtained from regular doctor visits.

Objective

Our project's main goal is to create a predictive model that uses a machine learning algorithm to correctly identify people as either non-diabetic or diabetic. Through the examination of trends in a range of health-related variables, including age, BMI, blood glucose levels, and more, our goal is to determine the probability of diabetes before it becomes seriously apparent. The model will help medical professionals make well-informed choices regarding additional diagnostic procedures and treatments.

Motivation

This project has multiple motivations. Firstly, by lowering the need for complex and expensive treatments, early detection of diabetes can dramatically reduce the likelihood of complications and ease the strain on healthcare systems. Second, if diabetes is discovered early enough, those who are more susceptible to it may be able to stop it entirely from starting by making lifestyle changes. Additionally, the methodology can enable public health campaigns to more effectively target high-risk groups by facilitating large-scale screenings within populations. Furthermore, the predictive powers of our model might be incorporated into digital health platforms to improve the caliber of telemedicine services. We support the international effort to

stop the diabetes pandemic by providing patients and healthcare professionals with useful insights.

Related Work

Many statistical and machine learning techniques have been used in a large body of work in the field of diabetes prediction. Research has shown that algorithms like Support Vector Machines (SVM), Decision Trees, and Neural Networks are effective in predicting diabetes. In a noteworthy study published in the Journal of Medical Internet Research, SVM was used to leverage electronic health records and produce impressive predictive results. Another study that showed the promise of AI in managing chronic diseases was published in Diabetes Care. This study fed a neural network with clinical and demographic data. While our work is in line with existing studies, it goes beyond them by utilizing a Random Forest Classifier, which is a technique that is well-known for its high accuracy and resilience to overfitting.

Data

Data Source and Format

The dataset used in this work was acquired via Kaggle. Numerous health variables that are essential for anticipating the onset of diabetes are included in this dataset, which has been specially designed for diabetes research. The features cover both numerical and categorical data, including blood pressure readings, blood glucose and HbA1c levels, age, body mass index (BMI), and more. The goal variable for our predictive modeling is the patient information about diabetes diagnosis that is included in the dataset.

The information is organized into a Comma-Separated Values (CSV) file format, which is compatible with a wide range of programming environments and data processing tools. This format makes it simple to manipulate and handle the data, which enables a smooth integration into our pipeline for data analysis.

Data Example

A complete collection of 100,000 entries, each representing a person's health record, makes up the training and testing dataset. Based on these health variables, a predictive model can be created from the training dataset to determine the probability of diabetes. To convert into a format that can be used for machine

learning models, the inclusion of categorical data, such as {gender} and {smoking_history}, will require encoding techniques like one-hot encoding.

Installation Instructions

1. From github, either open in the github desktop and open in your IDE or download all the files individually.
2. When open in github desktop, copy all the csv files and the “model.joblib” to downloads or desktop so it's easier to call the file path. If downloaded individually, they will show up in downloads anyways.
3. In the “Project.ipynb” file and the “UI.py” file, update the file paths to wherever the csv and joblib files are stored.
4. Just update the user, downloads/desktop and the name of the file in the way below:
    - For Mac

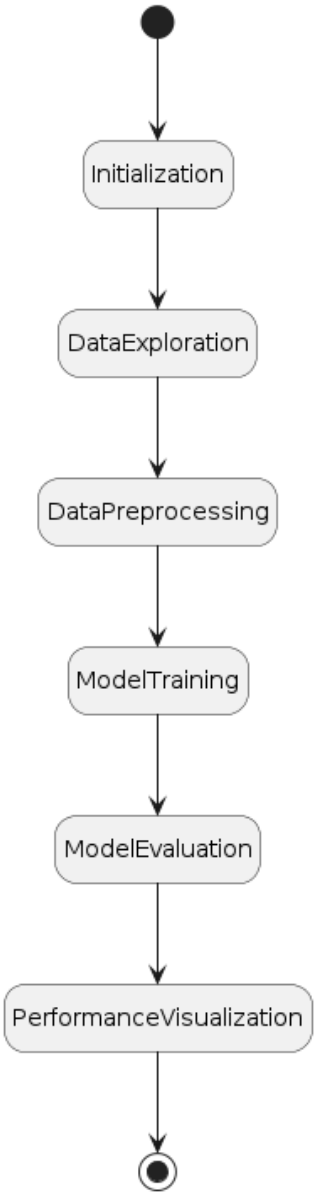
```
a. file_path = '/Users/sarimkazmi/Downloads/training_dataset.csv'
```

    - For Windows

```
- file_path = 'C:/Users/sarimkazmi/Downloads/training_dataset.csv'
```

5. Run all the “Project.ipynb” file
6. Run the “UI.py” file and then in the terminal enter:
   streamlit run file path of the UI.py file.
7. If you encounter any error, check if the python libraries are installed or not. Most likely, you might need to install the seaborn and joblib libraries.

Methodology



Schematic Diagram/Framework

The process flow depicted in the schematic diagram consists of the following sequential steps:

1. Initialization: Configuring parameters, loading required libraries and modules, and initializing variables are all part of setting up the environment.

2. Data Exploration: In this step, the dataset is carefully examined in order to comprehend the distributions of the features, spot outliers and missing values, and obtain some first understanding of the data structure.

3. Data Preprocessing: At this point, unprocessed data is cleaned up and put into a format that may be used to model. Missing value handling, categorical variable encoding, feature scaling, and normalization are examples of preprocessing procedures.

4. Model Training: During this stage, a machine learning algorithm is fed the preprocessed data in order to train the model. With the given features, the model learns to predict the target variable.

5. Model Evaluation: A different testing set is used to evaluate the model's performance following training. To determine the model's efficacy, evaluation measures like accuracy, precision, recall, and F1 score are computed.

6. Performance Visualization: Lastly, charts and graphs like ROC curves, confusion matrices, and performance bar graphs are used to display the findings of the model evaluation. This makes it easier to comprehend the model's output and pinpoint areas in need of development.

Data Visualization and Preprocessing

We use a variety of visualization tools to examine the distribution of various variables, such as age, BMI, and blood glucose levels, during the data exploration phase. These graphics, which are frequently density plots or histograms, aid in our comprehension of the data's central tendencies and dispersions. Additionally, we search for trends and abnormalities that might affect our model's ability to forecast the future.

One important step in preparing the dataset for modeling is preprocessing it. If there are random missing values in the data, we manage them by imputation techniques; otherwise, we remove the entries. We use one-hot encoding for categorical variables,

such as gender or smoking status, which turns these categories into a binary matrix—critical for models that need numerical input. For algorithms that are sensitive to the amount of input, such as support vector machines (SVMs), it is important to make sure that variables with greater scales do not disproportionately influence the model. To this end, numerical features might be standardized or normalized.

Procedures and Features

We carefully preprocess and engineer features to enhance model performance, and this divided our dataset into an 80% training and 20% testing split to ensure robust validation on unseen data. After loading, the data undergoes preprocessing to normalize distributions and handle missing values. Feature engineering plays a critical role, where new features are derived from existing ones and categorical variables are transformed using one-hot encoding, all contributing to the refined input dataset `X_new_train_encoded`. This meticulous preparation is important in optimizing the model's accuracy; this helps the model produce optimal results.

Schematic Diagram Analysis

From the first step of setting up the environment to the last step of visualizing the model's performance, the accompanying image describes a straightforward and linear approach. This framework makes sure that the data science lifecycle proceeds logically, which makes it possible to design and assess the prediction model in an organized manner. Every stage is designed to reinforce the one before it, resulting in a model that is comprehensible and strong.

Experiments

Data Division (Training/Testing)

We divided the dataset into two separate sets: one for training (80%) and another for testing (20%) in order to ensure that our model learned efficiently and was appropriately evaluated. This separation is a common machine learning procedure that guards against overfitting and guarantees that the model's functionality is evaluated on data that hasn't been seen before.

Parameter Tuning

We experimented with various values for parameters like `n_estimators} (the number of trees in the forest) and `max_depth} (the maximum depth of the trees) in order to do hyperparameter tuning on the Random Forest classifier. For example, adding more estimators could increase the accuracy of the model, but at the expense of computational efficiency. In a similar vein, choosing a suitable {max_depth} can improve the model's ability to generalize and keep it from fitting to noise in the training set.

Evaluation Metrics

We used a number of indicators to assess our model's performance:

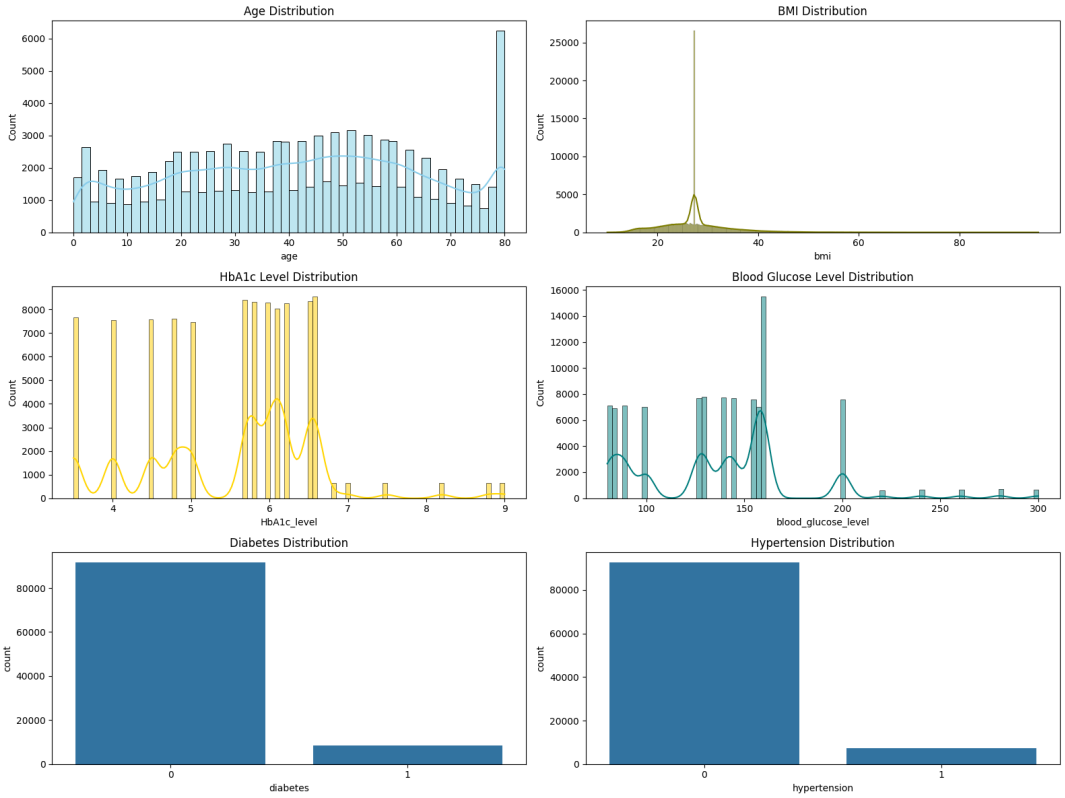
- Accuracy: Determines how accurate the model is overall by dividing the number of accurately predicted instances by the total number of instances.

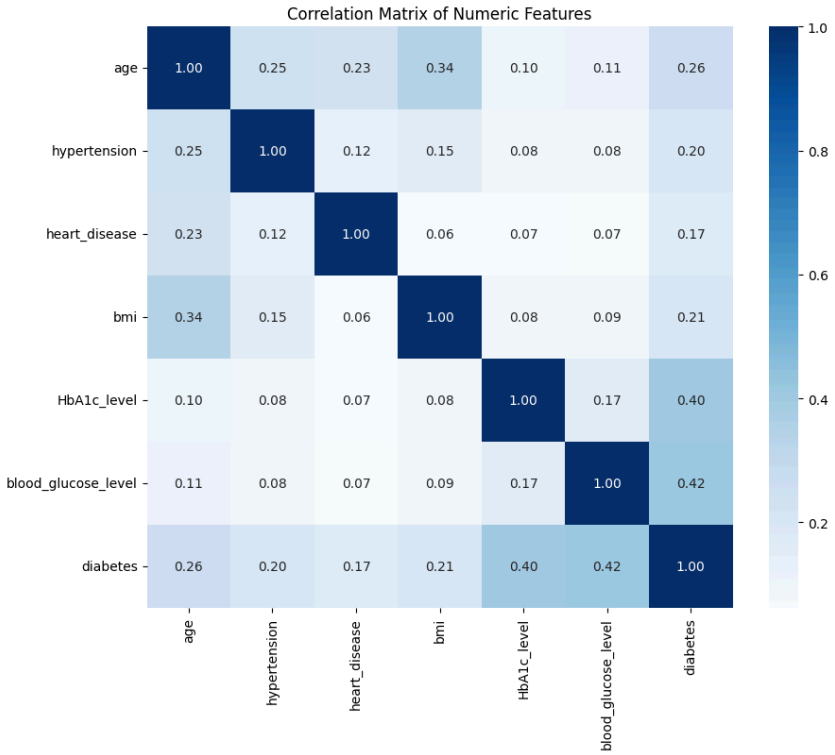
- Precision: This important indicator is important when the costs of false positives are large since it represents the percentage of positive identifications that were genuinely true.

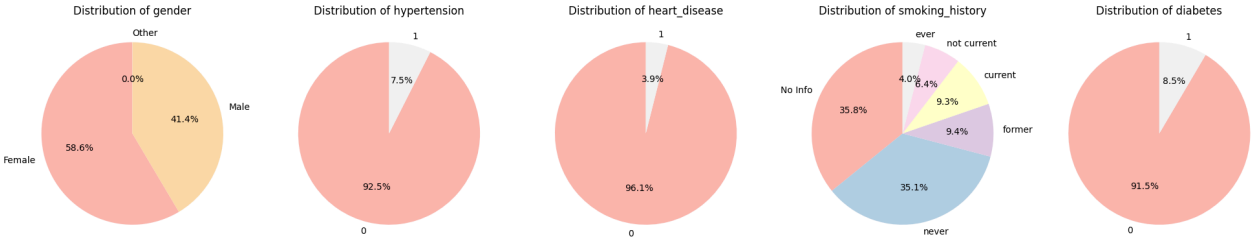
- Recall (Sensitivity): Indicates the percentage of true positives that were accurately recognized; this is important for medical diagnosis since it can be expensive to miss a condition.

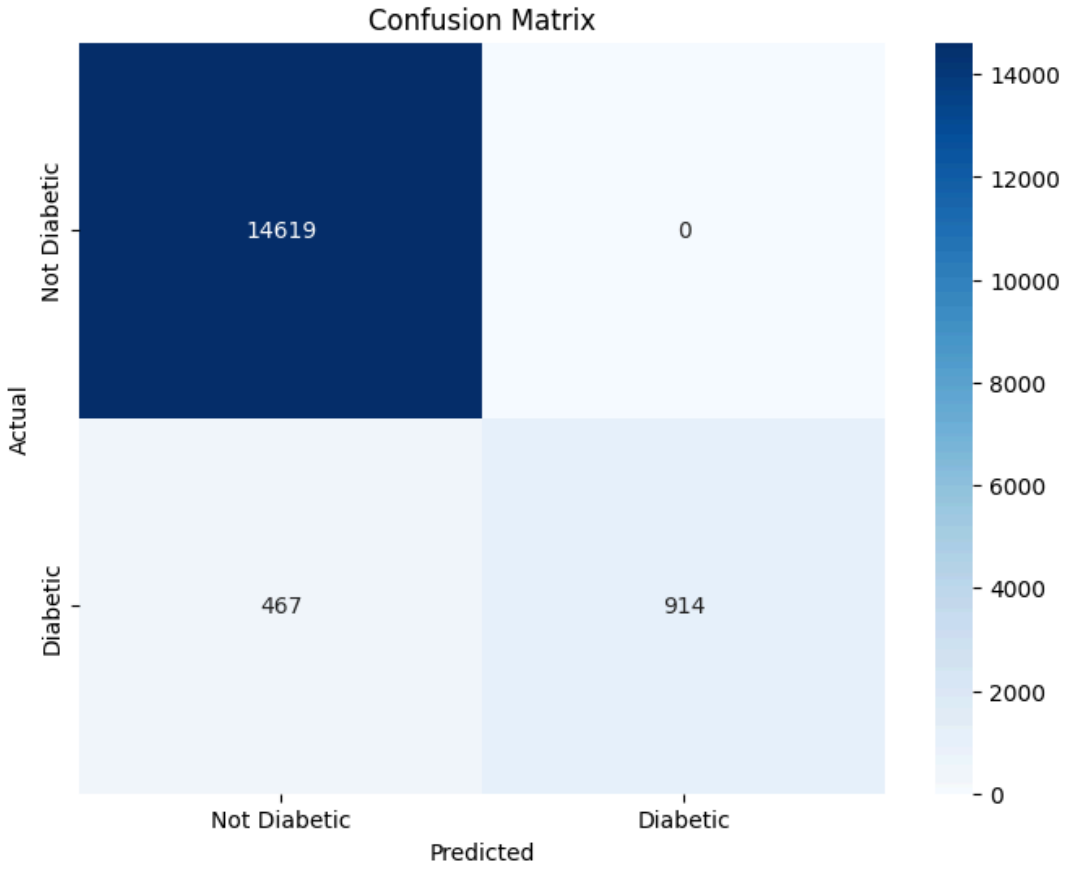
- F1 Score: The harmonic mean of recall and precision, which yields a single score that takes into account both factors. This is especially helpful in cases where the class distribution is uneven, as it is in our dataset.

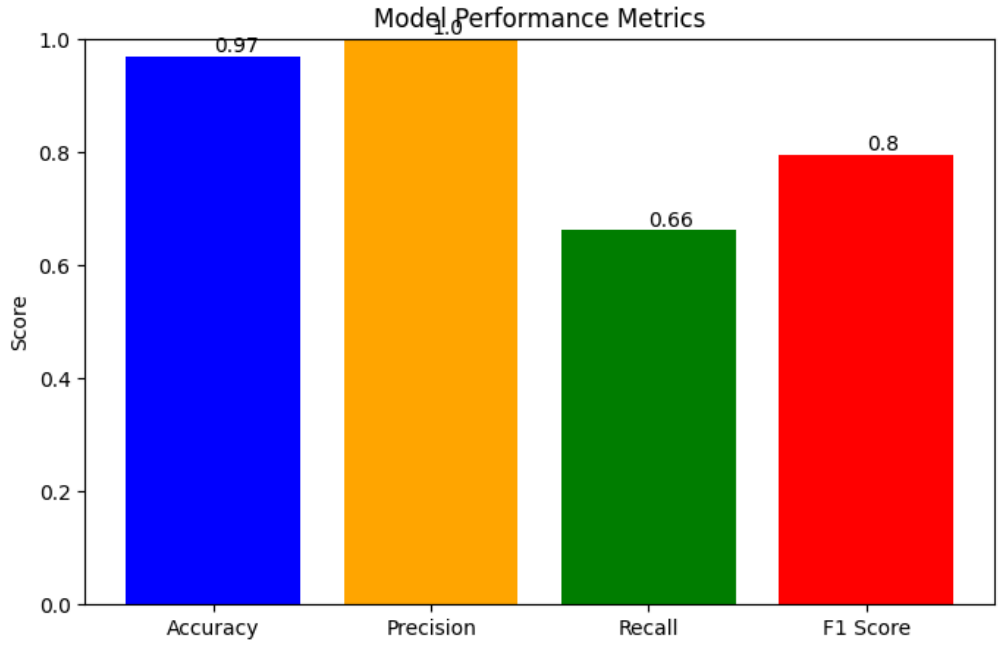
Results (Tables/Graphs)


Age Distribution
Count
age
BMI Distribution
Count
bmi
HbA1c Level Distribution
Count
HbA1c_level
Blood Glucose Level Distribution
Count
blood_glucose_level
Diabetes Distribution
count
diabetes
Hypertension Distribution
count
hypertension


Correlation Matrix of Numeric Features
age 1.00 0.25 0.23 0.34 0.10 0.11 0.26
hypertension 0.25 1.00 0.12 0.15 0.08 0.08 0.20
heart_disease 0.23 0.12 1.00 0.06 0.07 0.07 0.17
bmi 0.34 0.15 0.06 1.00 0.08 0.09 0.21
HbA1c_level 0.10 0.08 0.07 0.08 1.00 0.17 0.40
blood_glucose_level 0.11 0.08 0.07 0.09 0.17 1.00 0.42
diabetes 0.26 0.20 0.17 0.21 0.40 0.42 1.00
age
hypertension
heart_disease
bmi
HbA1c_level
blood_glucose_level
diabetes


Distribution of gender
Other
0.0%
Male
41.4%
Female
58.6%
Distribution of hypertension
1
7.5%
92.5%
0
Distribution of heart_disease
1
3.9%
96.1%
0
Distribution of smoking_history
ever
not current
current
No Info
4.0%
6.4%
9.3%
35.8%
9.4%
former
35.1%
never
Distribution of diabetes
1
8.5%
91.5%
0


Confusion Matrix
Not Diabetic
Diabetic
Actual
14619
0
467
914
Not Diabetic
Diabetic
Predicted
14000
12000
10000
8000
6000
4000
2000
0



- Histograms and Distributions: These are graphs that display the frequency distribution of specific attributes, like age, BMI, blood sugar, and HbA1c. For instance, the Age Distribution histogram points to a distribution that is right-skewed, suggesting a higher proportion of young people. The BMI distribution shows a peak at the normal range, with tails indicating the prevalence of underweight and overweight people.

- Correlation Matrix: The pairwise correlation between characteristics is displayed in this heatmap. For instance, a strong positive correlation has been shown between diabetes and BMI, indicating that a higher BMI is linked to a higher chance of developing diabetes.

- Pie Charts: While the majority of participants do not have heart disease or hypertension, the gender distribution pie chart indicates a higher proportion of female participants. The smoking history chart, which shows that most people are not smokers, offers insights about the smoking behaviors of the general public.

- Confusion Matrix: The true positives, true negatives, false positives, and false negatives are displayed in the matrix for our model. It shows that while our model reflects individuals who are not diagnosed, it also includes a significant proportion of false negatives, which is problematic in the medical industry.

Analysis of the Results

The algorithm appears to be quite good at classifying the non-diabetic condition, based on its high accuracy of 97%. It is troubling, though, because the recall of 66% suggests that the model is not as good at recognizing all real instances of diabetes. Although the F1 score of 0.8 achieves a compromise between precision and recall, there is clearly potential for improvement given the previously indicated reduced recall, particularly in terms of lowering false negatives—a critical component of any medical diagnostic tool.

Conclusion

Discuss any Limitation

Numerous obstacles came up during the course of the endeavor. Predictions could be skewed if the dataset was not representative of the total population. For example, the data can be biased toward specific age groups or show little variation in ethnicity, which could affect how well the model applies to a larger population. Moreover, not all pertinent variables that influence diabetes may be covered by the elements listed, such as specific lifestyle choices or genetic predispositions. Model performance shows a tendency to overlook diabetic patients, even while accuracy is high and recall is relatively low. A mismatch in class or characteristics that fail to fully portray the subtleties of the development of diabetes could be the cause of this.

Discuss any Issue not Resolved

The possibility of underdiagnosis in minority groups in the event that the dataset is not diverse is one of the problems left unresolved in this iteration. Furthermore, it's possible that the feature set is too basic to identify intricate relationships in the data that raise the risk of diabetes. Furthermore, the model's predictions might be impacted by unidentified biases that occurred during the data collection process and were not taken into consideration.

Future Direction

There are various ways to enhance the present model in future development. One immediate suggestion is to gather a larger and more varied dataset with a greater range of characteristics that may influence the risk of diabetes. To capture more complicated correlations in the data, further feature engineering approaches and the addition of interaction terms could be investigated. Recall performance could be

enhanced by using deeper learning models or ensemble learning techniques, which are more complex machine learning models. Moreover, methods like SMOTE (Synthetic Minority Over-sampling Technique) or modified class weights during model training could be used to address the problem of class imbalance. Additionally, longitudinal studies could be used to monitor changes over time and offer deeper insights into the development of diabetes.

Appendix

A. Data Overview
- **Number of Records:** 100,000 (training and testing dataset)
- **Number of Features:** 9
- **Features:**
  - **gender:** Categorical (Female, Male, Other)
  - **age:** Numerical (Continuous)
  - **hypertension:** Binary (0: No, 1: Yes)
  - **heart_disease:** Binary (0: No, 1: Yes)
  - **smoking_history:** Categorical (never, current, former, No Info)
  - **bmi:** Numerical (Continuous)
  - **HbA1c_level:** Numerical (Continuous)
  - **blood_glucose_level:** Numerical (Discrete)
  - **diabetes:** Binary (Target Variable; 0: No, 1: Yes)

B. Visualization Samples

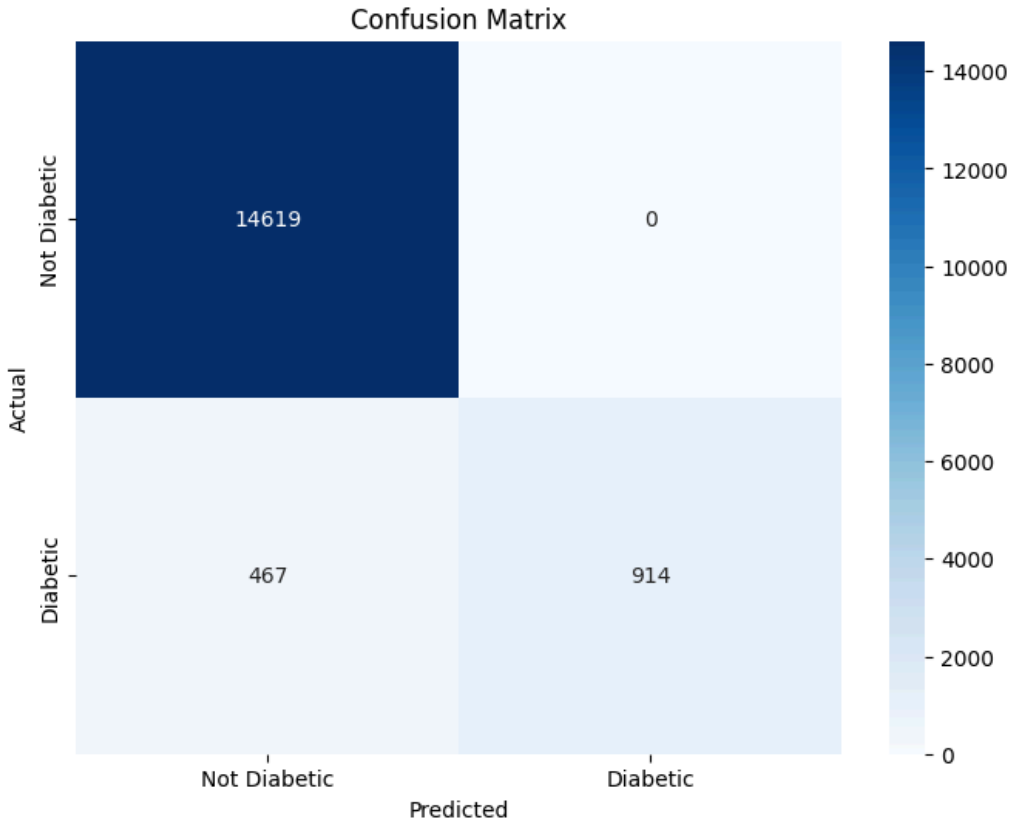
- Histograms for `age`, `bmi`, `HbA1c_level`, `blood_glucose_level`

- Pie charts representing the distribution of `gender`, `hypertension`, `heart_disease`, `smoking_history`, and `diabetes`

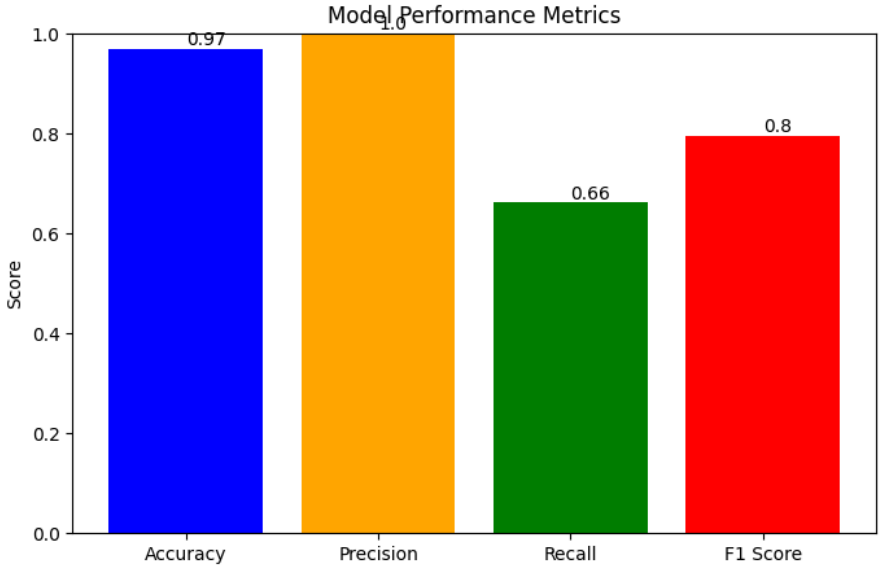
- Correlation matrix heatmap showing relationships between features

C. Model Evaluation Output

- Confusion matrix showcasing the model’s performance in classifying diabetic and non-diabetic instances



- Bar chart displaying model performance metrics including Accuracy, Precision, Recall, and F1 Score



- The model was integrated into an application, and the application is capable of identifying, based on certain inputs, whether a patient is diabetic or non-diabetic; this means a software application capable of identifying whether a patient is diabetic or not is developed, at the end.

References

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

https://arxiv.org/ftp/arxiv/papers/1205/1205.5921.pdf

https://www.researchgate.net/publication/236952762_Random_Forests

https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

https://machinelearningmastery.com/bagging-and-random-forest-for-imbalanced-classification/

https://scikit-learn.org/stable/modules/preprocessing.html

https://scikit-learn.org/stable/modules/model_evaluation.html

https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_features.html

https://seaborn.pydata.org/generated/seaborn.histplot.html

https://www.geeksforgeeks.org/plotting-correlation-matrix-using-python/