12/2/2024

Predictive Modeling for Diabetes Risk Assessment

Project Report

Guided by: Zhe Zhang

BUAN 6356.002

*Group-15*

Dwaipayan Roy

Anuj Garg

Sucharita Naha

Karthik Raj Sarpangala

Table of Contents

[1. Executive Summary 3](#_Toc184239753)

[2. Background 3](#_Toc184239754)

[3. Data Description 4](#_Toc184239755)

[4. Data Understanding and Cleaning 4](#_Toc184239756)

[5. Methodology 7](#_Toc184239757)

[6. Results 13](#_Toc184239758)

[7. Conclusion 15](#_Toc184239759)

[8. References 15](#_Toc184239760)

## Executive Summary

This project aims to use classification models, correlation analysis, and trend analysis to determine whether an individual is diabetic or non-diabetic and identify any underlying trend factors.

The primary objective of our data mining approach is to develop a predictive model that can accurately identify individuals at risk of developing diabetes based on various health indicators. By analyzing the relationships between demographic and health-related variables, we aim to create a model that can be utilized for early intervention and personalized healthcare strategies. Our focus will be on determining the influence of factors such as age, hypertension, heart disease, body mass index (BMI), HbA1c levels, blood glucose levels, and smoking history on diabetes diagnosis.

## Background

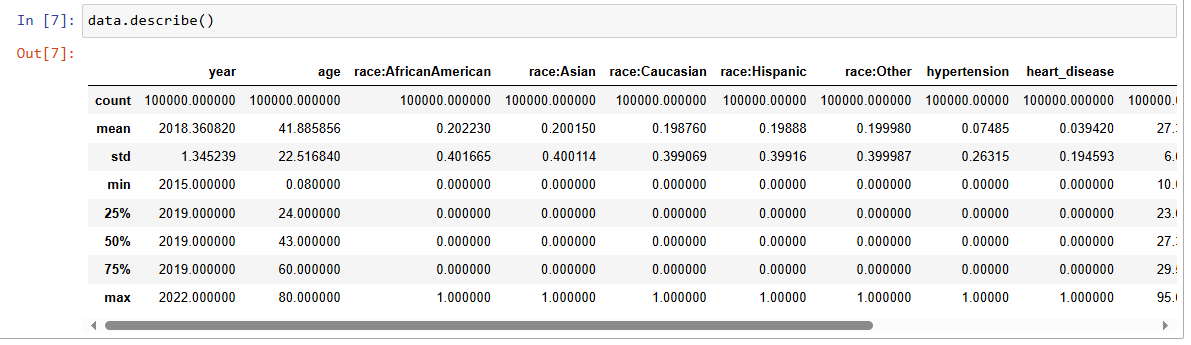
Diabetes is a growing global health concern, affecting millions of individuals worldwide. If left untreated, it can lead to a number of other health problems including blindness, kidney failure, heart attacks, and lower limb amputation. Symptoms of diabetes may occur suddenly. In type 2 diabetes, the symptoms can be mild and may take many years to be noticed and hence leads to delay in confrontation. Therefore, the ability to predict the likelihood of developing diabetes can significantly aid individual healthcare planning as well as public health interventions and clinical research. Given the sheer number of potential factors and the large volume of data, determining a trend from such data manually becomes impossible. This project aims to leverage a detailed dataset containing health and demographic information of 100,000 individuals to build predictive models and utilize various BI methodologies to perform an in-depth data analysis of the relationships between various health metrics and diabetes prevalence.

## Data Description

The dataset used for our analysis has 100,000 records of individuals, including various demographic and health-related features. The key attributes in the dataset are as follows:

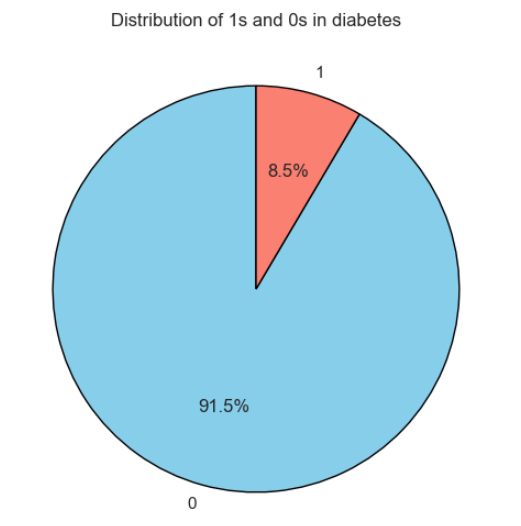
* year: Year of data collection
* age: Age of the individual (in years)
* race: Categorical variables representing racial demographics (African American, Asian, Caucasian, Hispanic, Other)
* hypertension: Binary variable indicating the presence of hypertension (0 = No, 1 = Yes)
* heart\_disease: Binary variable indicating the presence of heart disease (0 = No, 1 = Yes)
* bmi: Body Mass Index
* hbA1c\_level: HbA1c levels (indicates blood sugar control)
* blood\_glucose\_level: Blood glucose levels
* diabetes: Binary variable indicating diabetes diagnosis (0 = No, 1 = Yes)

## Data Understanding and Cleaning

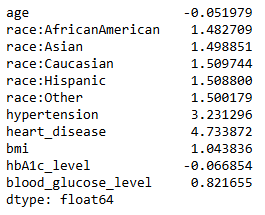


Describing the data helped us understand the below information:

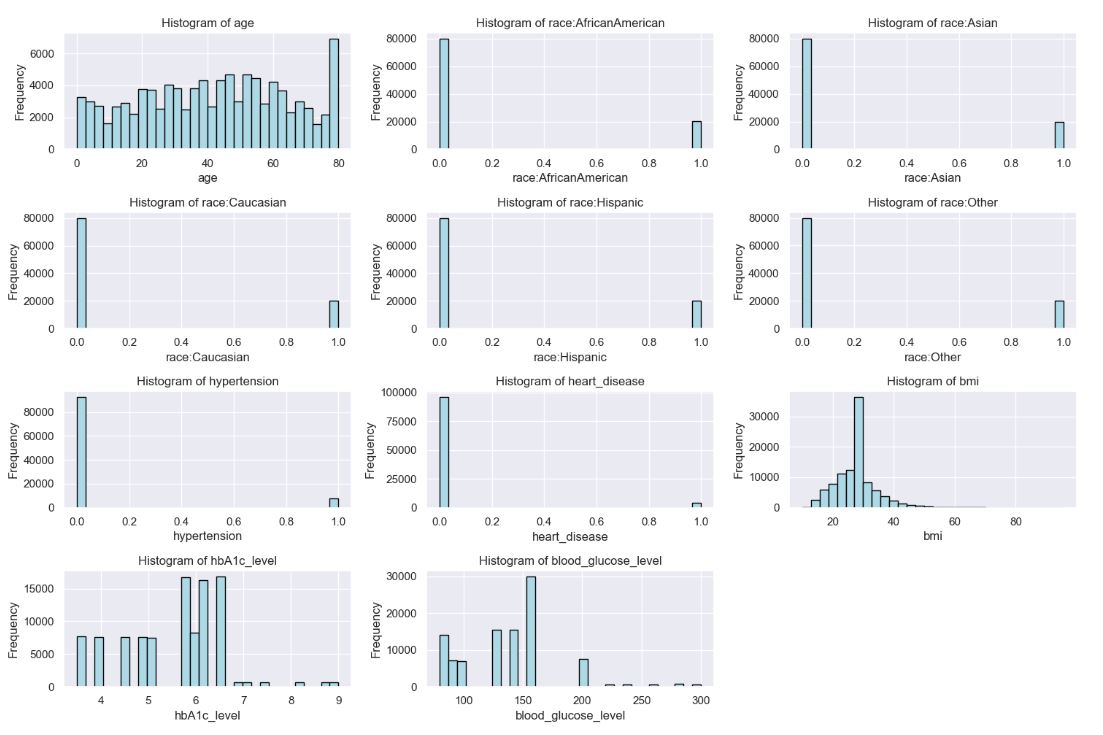
* In year, mean is 2018.36. The data consists the data from year 2015 to 2022.
* The average age is around 41.89 years, with a minimum of 0.08 years (infants). There’s a wide spread, with a standard deviation of 22.5 years, ranging from 0.08 to 80 years.
* About 7.5% of the people have hypertension, and 3.9% have heart disease, based on their mean values.
* Average BMI is about 27.32, with a standard deviation of 6.64, indicating variability in body mass index across the dataset.
* There’s a wide range, from 10.01 to 95.69, with most values between 23.63 (25th percentile) and 29.58 (75th percentile).
* The majority of values of HbA1c lie between 4.8 and 6.2, which aligns with standard thresholds for HbA1c in diabetes diagnosis.
* The average blood glucose level is 138.06, with a broad range (from 80 to 300). The 25th percentile (100) to 75th percentile (159) shows the typical range for most patients in the dataset.
* Around 8.5% of the people in the dataset have diabetes. This means that 91.5% of the entries are non-diabetic. Below is the depiction of this data in the form of a pie-chart.



* We checked the distribution and skewness of the continuous variables to identify potential data transformations.



* BMI and blood\_glucose\_level are continuous variables and we performed the transformation on theose features.

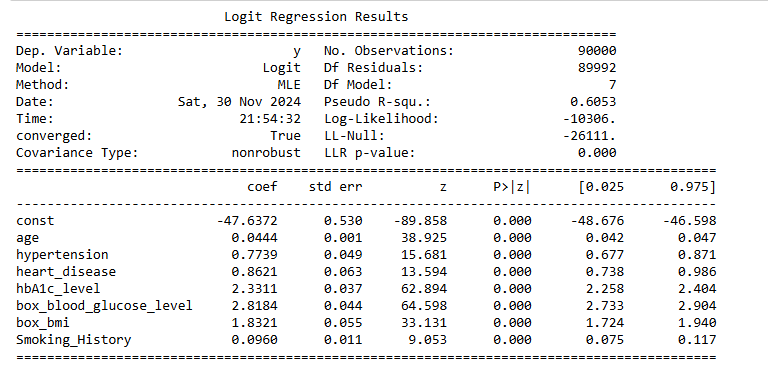
****

* After transformation, the skewness of bmi and blood\_glucose\_level was significantly reduced using boxcox. The new values are around 0.01 and -0.003 respectively.
* Further, we transformed the categorical features like ‘Smoking\_History’ into numerical values to aid the analysis using label encoder.

## Methodology

* 1. **Logistic Regression**

To analyze the data and predict diabetes occurrence, logistic regression was applied as a classification technique. Logistic regression is a statistical method used to model the probability of a binary outcome based on one or more predictor variables. The model successfully converged after 10 iterations.



The Pseudo R-squared value of 0.6053 suggests that the model explains about 60.53% of the variance in the outcome variable.The results indicate that all predictor variables are statistically significant (P-value < 0.05), with HbA1c level and blood glucose level being the most influential predictors.

Below is the confusion matrix without SMOTEENN:

**A black and white text

Description automatically generated**

From the confusion matrix, we observe that the model correctly predicted 9041 negative cases (no diabetes) and 527 positive cases (diabetes). There were 349 false negatives and 83 false positives.

The accuracy of the model on the test set is 95.68% and the precision score is 86.4%. 0.864 indicates that when the model predicts diabetes, it is correct approximately 86.4% of the time. However, the recall score is only 60%. Hence, we will use SMOTEENN to oversample the minority class and undersample the majority class.

Below is the confusion matrix with SMOTEENN:

A black numbers and symbols on a white background

Description automatically generated

The recall score in this scenario is almost 93%, which is quite high.

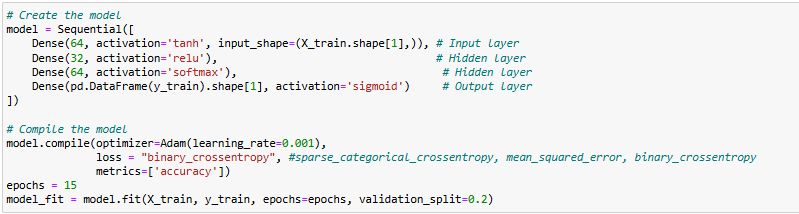
The logistic regression model has demonstrated excellent predictive performance for diabetes prediction, with high accuracy and precision scores. The model’s significant features, such as age, hypertension, HbA1c level, and blood glucose levels, which indicate the strong relationship between these factors and the risk of diabetes.

* 1. **Neural Network**

We applied a Neural Network (NN) model to predict diabetes. They are particularly powerful for capturing non-linear relationships in data. Here, we built and trained a feedforward neural network to classify patients based on their likelihood of having diabetes.

To address class imbalance in the dataset, we applied SMOTEENN to create a balanced training dataset. The resampled data was then split into training and test sets using stratified sampling to preserve class proportions.

Below is the neural network model we used.



The neural network consisted of the following layers:

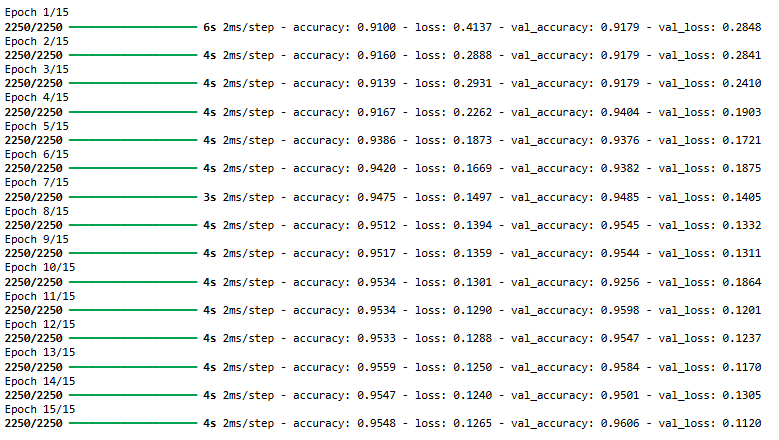
Input Layer: 64 neurons.

Hidden Layer1: 32 neurons.

Hidden Layer2: 64 neurons.

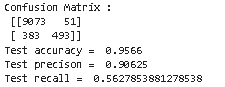
Output Layer: Sigmoid activation to predict binary outcomes (diabetes or no diabetes).

The neural network was trained over 15 epochs using an 80-20 validation split. The model achieved the following accuracy and loss values during training:



We got following readings in the confusion matrix of neural network.

Without SMOTEENN:



In the above scenario, we got good accuracy and precision values. However, the recall score is only around 56% which proves that the false negatives are high.

With SMOTEENN:

A black numbers and symbols

Description automatically generated

In this scenario, we got promising results with regards to accuracy, precision and recall value.

The precision score indicates that when the model predicts diabetes, it is correct approximately 90% of the time.

* 1. **Decision Tree**

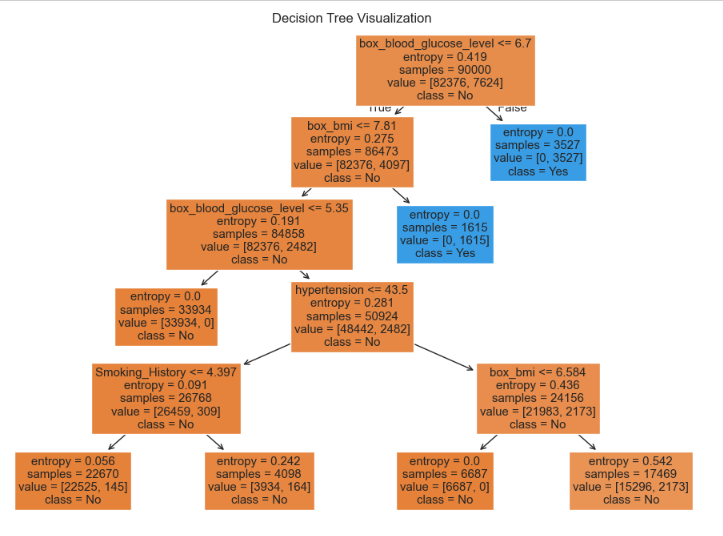
We used below parameters for our Decision Tree model.

Criterion: 'entropy' is used to measure the quality of splits based on information gain.

Max Depth: 5 is the limit of the tree's depth in order to prevent overfitting.

Random State: 42, which is to ensure reproducibility of results.

The Decision Tree algorithm provides the advantage of interpretability. The model learned decision rules from the features to classify patients as diabetic or non-diabetic. By setting a maximum depth of 5, we ensured the model was simple enough to avoid overfitting while still capturing meaningful patterns in the data.



The tree starts with a root node testing if box\_blood\_glucose\_level <= 6.7 and branches into various decision paths.

Second Level, left branch tests box\_bmi <= 7.81 and right branch leads to a leaf node with class "Yes". At lower levels, splits were based on:

Blood glucose level (<=5.35)

Hypertension (<=43.5)

Smoking history (<=4.397)

BMI (<=6.584).

We got following readings in the confusion matrix of Decision Tree.

Without SMOTEENN:

A screenshot of a computer

Description automatically generated

The accuracy of 96.69% indicates that the model correctly classified a significant majority of the test data. The precision score of 100% suggests that all instances predicted as positive (diabetes) by the model were correctly classified, meaning there were no false positives. However, the recall score is only around 62%.

With SMOTEENN:

A black numbers and symbols on a white background

Description automatically generated

In this scenario, we got promising results with regards to accuracy, precision and recall value.

The precision score indicates that when the model predicts diabetes, it is correct approximately 97.5% of the time.

The Decision Tree classifier delivered excellent predictive performance, achieving both high accuracy and perfect precision. This makes it a strong candidate for diabetes prediction, especially in scenarios where interpretability is important.

* 1. **Random Forest**

The Random Forest model was created using following hyperparameters.

Number of Estimators is 100 which specifies the number of decision trees in the ensemble.

Max Depth is 10 which limits the depth of each tree to prevent overfitting.

Random State is 42 which ensures reproducibility of results.

The accuracy of 96.68% highlights the model's strong predictive capability. The precision score of 99% indicates that all instances classified as positive (diabetes) were correct, with no false positives.

Random Forest builds multiple decision trees and averages their predictions to improve accuracy and robustness. By using 100 estimators and a controlled maximum depth of 10, the model effectively captured important patterns in the data without overfitting.

We got following readings in the confusion matrix of Decision Tree.

Without SMOTEENN:

A screenshot of a computer

Description automatically generated

The readings are somewhat similar to the decision tree. In this scenario too we got 62% recall score.

With SMOTEENN:

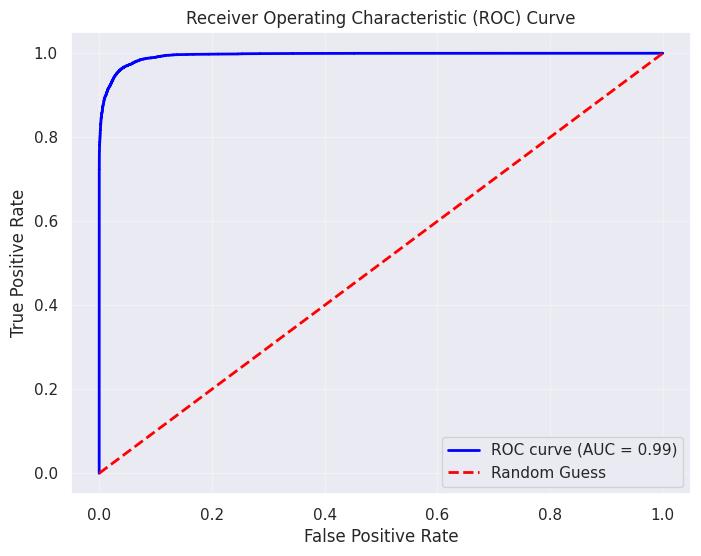
A black numbers and symbols

Description automatically generated with medium confidence

After SMOTEENN, the recall score improved to almost 96%. The above readings indicate that the accuracy is 96% and the precision is close to 96%. Hence, we got great results with Random Forest.

## Results

**Model Performance:** The diabetes prediction project evaluated multiple machine learning models to assess their effectiveness in identifying potential diabetes cases. Each model demonstrated strong predictive capabilities, with notable variations in performance metrics:

* **Logistic Regression** achieved an accuracy of 92.6% and a precision of 92.67%, showing robust performance in classifying diabetes cases.
* **Neural Network** presented an accuracy of 92.6% and a precision of 90.3%, indicating high reliability in positive diabetes predictions.
* **Decision Tree** emerged as the top-performing model with the highest accuracy of 92.8% and a perfect precision score of 89.3%, suggesting zero false positive predictions.
* **Random Forest** closely followed with an accuracy of 96% and a precision of 95.9%, demonstrating remarkable predictive power and minimal false positive rates.
* 

We have also plotted the ROC curve and the Area under the curve value is 0.99. As the recall value is high, we can say that there are less false negative.

During the analysis we got to know about the key predictors like HbA1c levels and blood glucose which were helpful in decision-making.

**Critical Predictors**: Across all models, key factors influencing diabetes prediction included:

* HbA1c levels
* Blood glucose levels
* Age
* Hypertension
* Body Mass Index (BMI)

**Dataset Insights**:

* Total records: 100,000.
* Diabetes prevalence: 8.5% of the dataset.
* Age range: 0.08 to 80 years (average 41.89 years).
* Hypertension: 7.5% of individuals.
* Heart disease: 3.9% of individuals.

**Methodological Strengths**:

* Addressed class imbalance through techniques like SMOTEENN.
* Implemented multiple models to ensure robust prediction.
* Used feature transformations to improve model performance.
* Maintained model interpretability, especially with Decision Tree and Random Forest.

## Conclusion

The **Random Forest model** particularly distinguishes itself with its exceptional performance, achieving the high accuracy, high precision and high recall value. This indicates that the Random Forest algorithm was most effective in correctly classifying diabetes cases in our dataset, with less instances of false positive predictions.

These results highlight the potential of machine learning techniques in early diabetes detection, providing promising tools for healthcare professionals to identify at-risk individuals more accurately. However, further validation with larger and more diverse datasets would be crucial to confirm these promising findings and ensure the models' generalization.

## References

Data source: [Diabetes Clinical Dataset](https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset/data)

Acknowledgments

The success of this project is a result of the collaborative efforts of the project team. We extend our gratitude to everyone who contributed to the development and implementation of the predictive model.