
# DIABETES ANALYSIS

PRESENTATION - 2024

# OVERVIEW

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Other factors such as gender, age, and bmi can also be contributors to the disease. This project will be focused in building models that can help predict whether a patient has the disease or not. This can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans.

# PROBLEM STATEMENT

The head of medical Board has tasked me to create a predictive model that can predict whether a patient has diabetes or not.

# OBJECTIVES

My top objectives for modelling will be:

1. Determine important factors that affect diabetes.
2. Determine the best model to predict diabetes.
3. Create a model that predicts diabetes with an accuracy of 96% or above.

# DATA SOURCES

The data for this modelling was obtained from kaggle.

The dataset contains the following columns:

Age, gender, bmi, Blood Glucose Level, HbA1c Level, Smoking history, Heart disease, Hypertension

# Analysis and Modelling

The analysis includes several statistical analysis and visualization techniques to uncover insights:

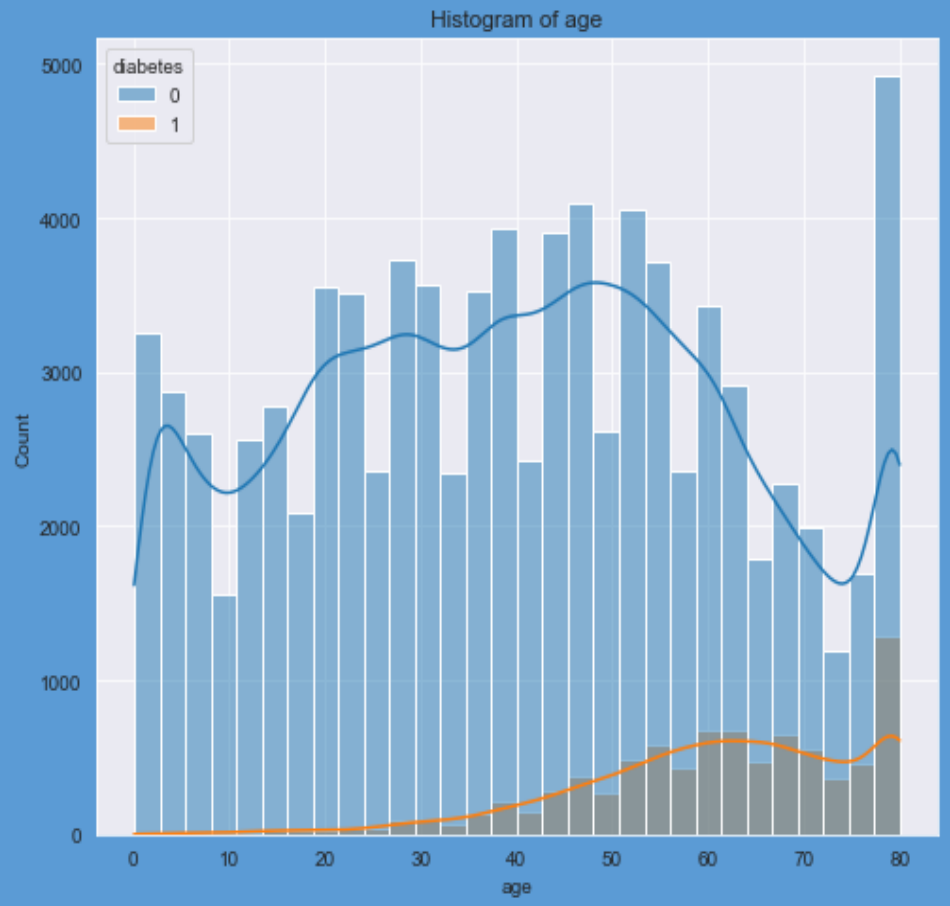
- Univariate Analysis: Examining the distribution of single variables, such as age and bmi, through histograms and count plots for hypertension, heart disease, smoking history and the diabetes column.
- Multivariate Analysis: A scatterplot of age vs bmi with hue as the diabetes column.
- Correlation Analysis: Exploring relationships between numeric variables to understand associations and check for multicollinearity

# Fitting of classification models

Models such as logistic regression, decision trees and random forests are fit

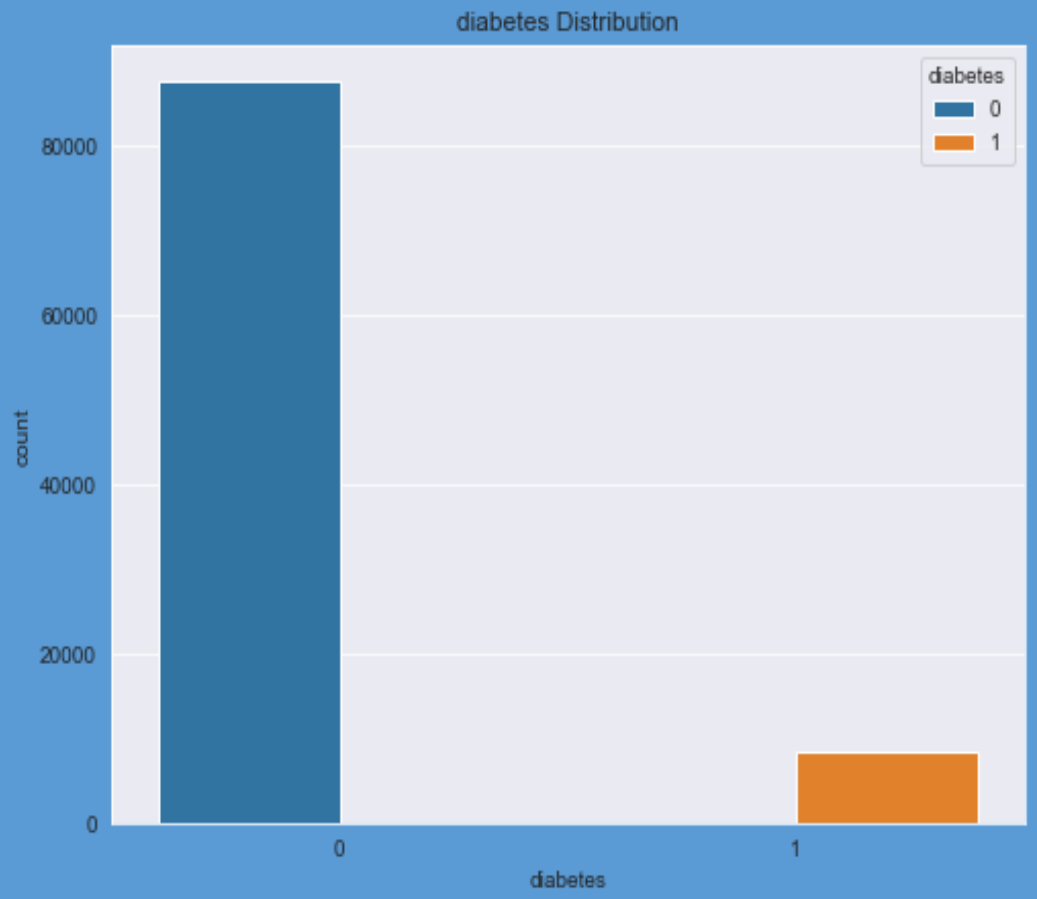
Model Comparison: The performance of the models is compared based on evaluation metrics and ROC curves to identify the best-performing model.

# HISTOGRAM OF AGE



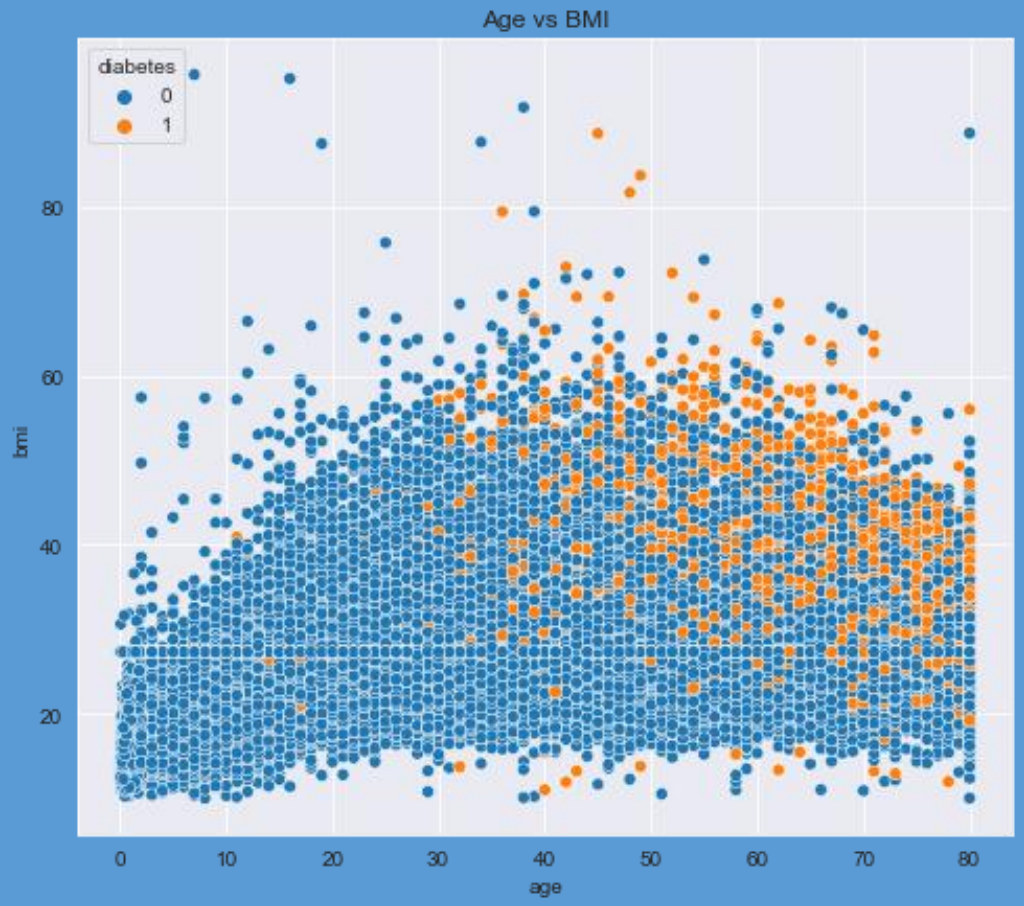

The histogram displays age with hue as the target. It can be seen that patients above 40 are at high risk of having diabetes.

# Distribution of diabetes



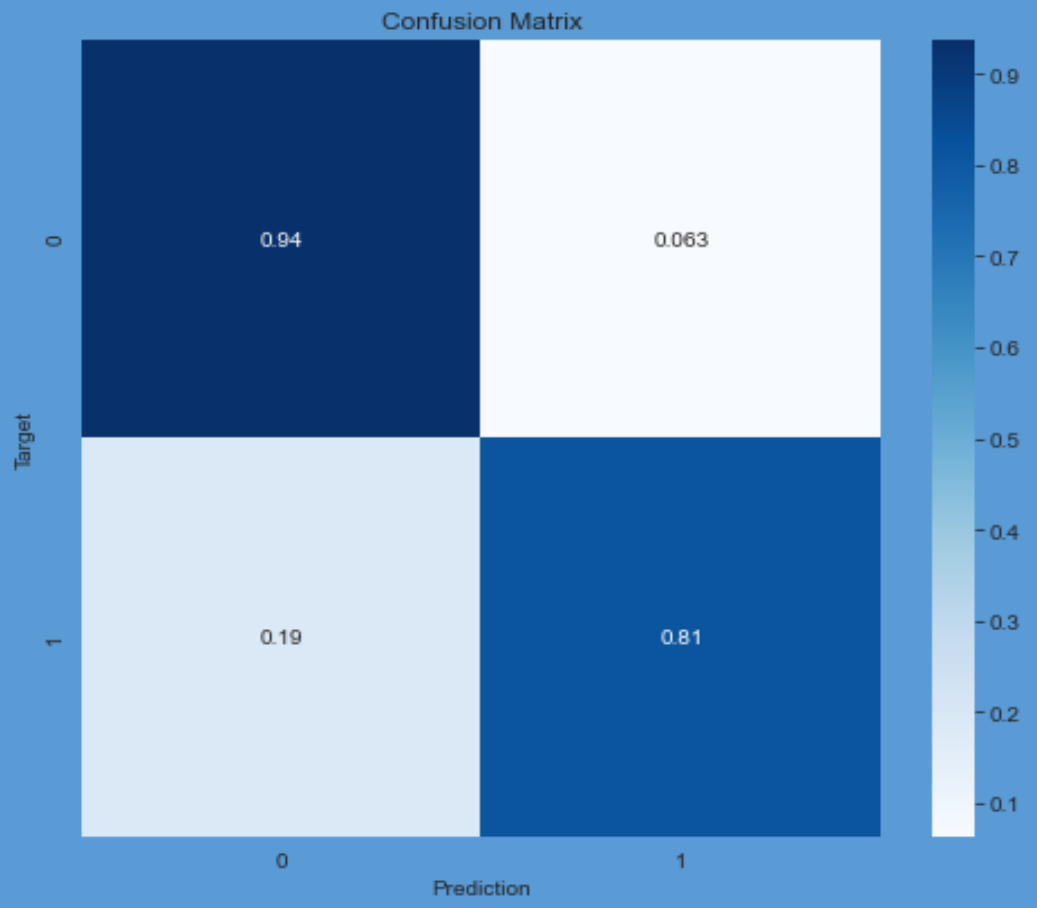

The distribution of the target column. Patients with no diabetes are more than those with diabetes. This shows there is an imbalance in the dataset.

# Scatterplot of age versus bmi




The distribution of age versus bmi with hue as the target. Younger patients with bmi have a low risk of having diabetes compared to patients above 40 with a high bmi.

# Confusion matrix




Normalised confusion matrix of the final chosen model. 0.94 shows where the patient was predicted not to have diabetes when in fact they did not have diabetes, while 0.81 shows where the patient was predicted to have the disease when in fact they had the disease.

# Recommendations

- Despite the model achieving a good accuracy of 92%, more data should be collected for patients with diabetes.

- Addition of more features which are used to create the models. More features can help improve recall, especially for cases where the patient has diabetes.

# Suggestions for future improvement

- Collection of more data on things such as family history, physical activity level and diet information could help improve the model's predictive accuracy.

- Training could be done on other machine learning models. This can help in improving the accuracy of predictions.

- Feature engineering such as interaction terms and polynomial features could help improve the model's performance.

THANK

YOU!