DIABETES OUTCOME ANALYSIS

BY RAHIM NJAGI

Overview

Objective

The goal of this project is to develop model that precisely identifies the outcome of a patient having diabetes or not.

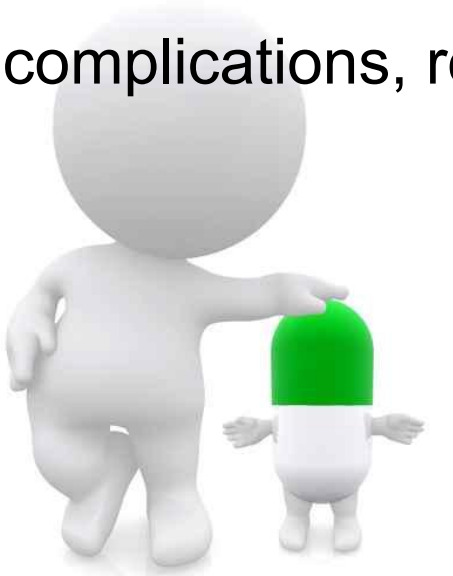
This can be significant in health institutions as it will help identify accurately patients diagnosed with diabetes.

BUSINESS AND DATA UNDERSTANDING

Business problem

-The primary business problem addressed in this project is the identification of high-risk patients for diabetes within a healthcare setting.

-Early detection is crucial for preventing serious complications, reducing healthcare costs, and improving patient outcomes.

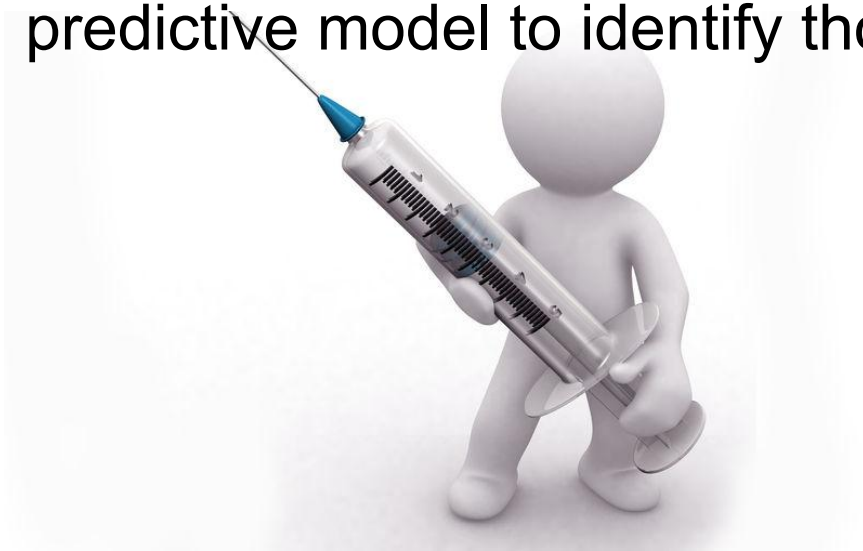


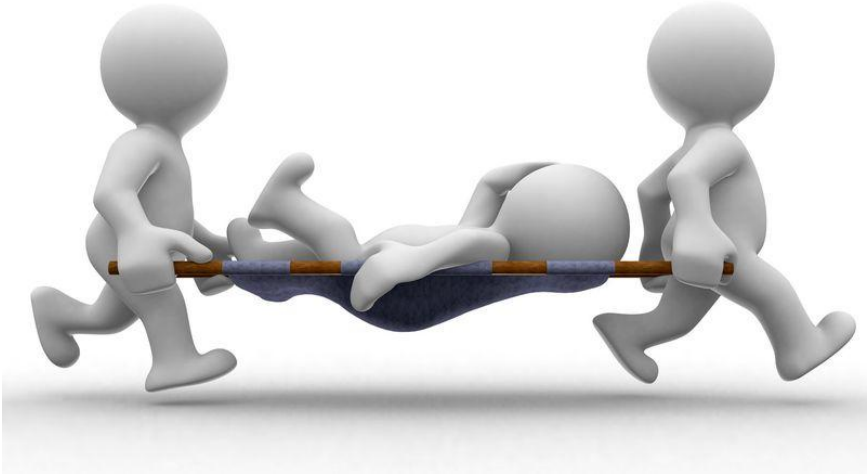
Data Understanding

-The dataset used in this project comes from a collection of patient records related to diabetes.

-It includes various health indicators that are crucial for predicting the likelihood of a patient being at high risk for developing diabetes.

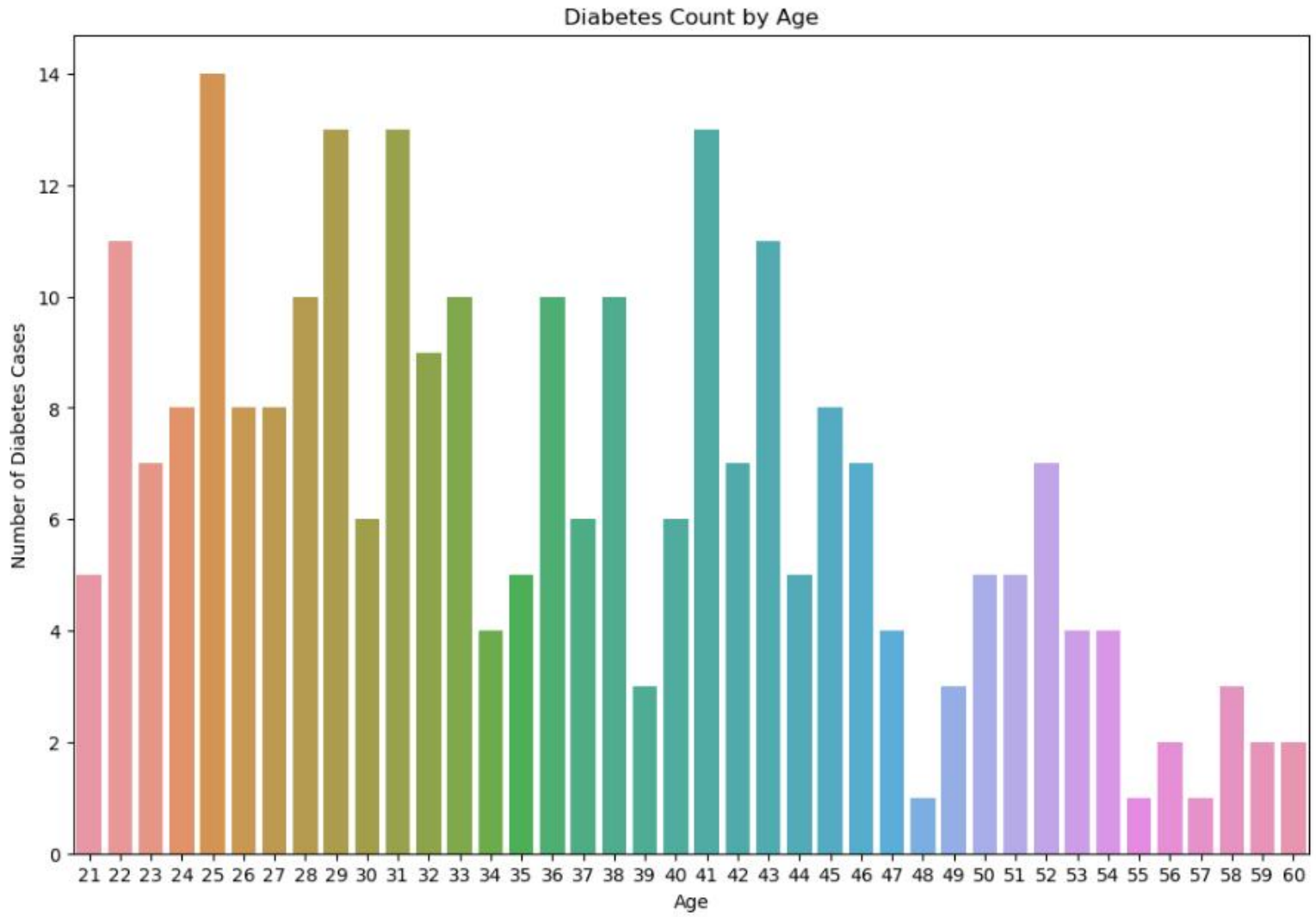
- The dataset provides a comprehensive view of patient health, allowing us to build a predictive model to identify those at risk.



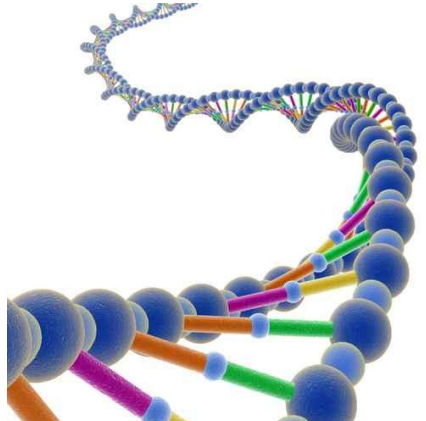


Perfoming EDA

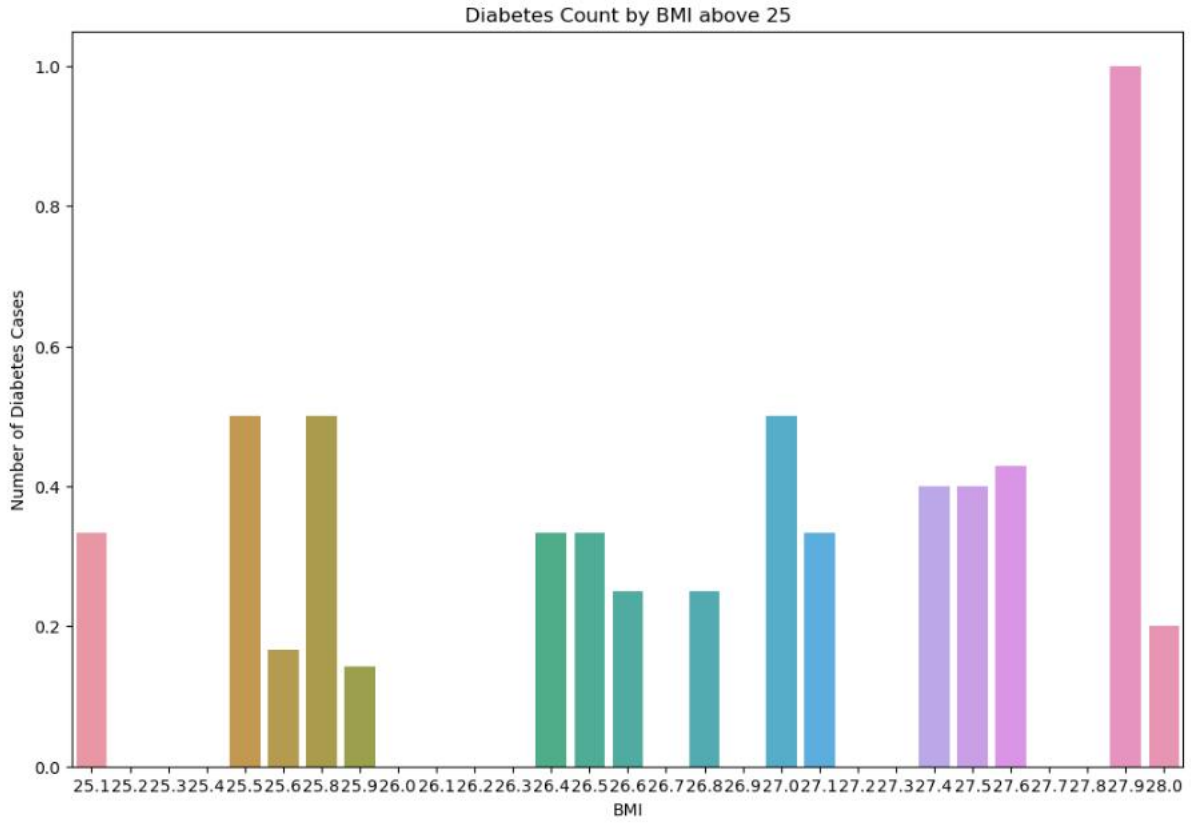
Analysis of Diabetes by Age



Age and Diabetes Cases: The number of diabetes cases increases with age. This is evident as older age groups tend to have higher counts of diabetes cases compared to younger age groups.

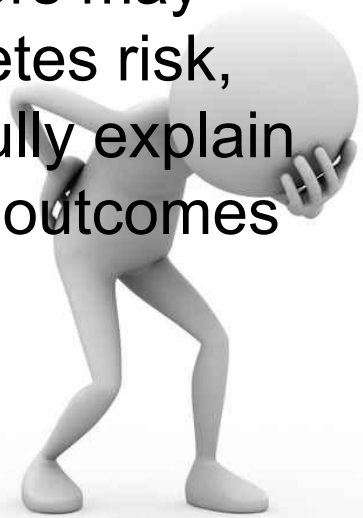


Analysis of diabetes count by BMI above 25



Increased Risk at Higher BMI Levels: As BMI increases, there seems to be a general trend of increasing proportions of individuals with diabetes:

Irregular Patterns: This irregularity suggests that other factors may also be influencing diabetes risk, and BMI alone doesn't fully explain the variance in diabetes outcomes



MODELING

Why modeling?

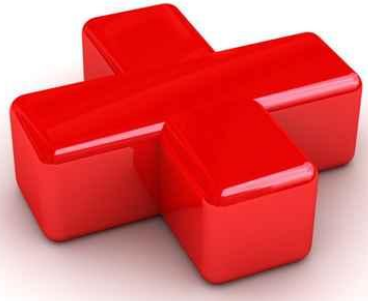
-Machine learning models tend to offer more
accuracy for tasks like classification or regression where patterns are complex for basical statistical methods.

-For example, predicting diabetes based on multiple features like age, BMI, blood pressure, etc., often requires ML to handle the complexity.

Data Preprocessing

Before building the model we first preprocess the data

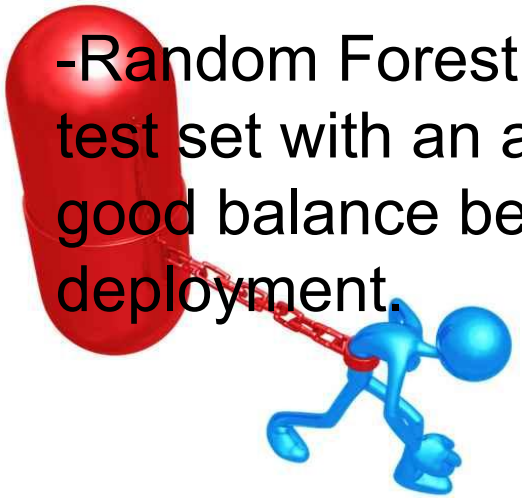
This involves:
 -standardadizing features through scaling
 - handling class imbalance through smote
-performing a train-test split



Model Performance.

-Logistic Regression: The logistic regression model showed solid performance, with an ROC-AUC of 0.80 on the test set, indicating good discrimination between classes. It is interpretable but might struggle with more complex relationships in the data.

-Decision Tree: The decision tree model exhibited high accuracy on the training set (suggesting overfitting) but lower performance on the test set, with a test set ROC-AUC of 0.73. This suggests that while the model is interpretable, it might not generalize well to unseen data.



-Random Forest: The Random Forest model achieved the highest performance on the test set with an accuracy of 0.82 and an ROC-AUC of 0.85. This model provides a good balance between accuracy and generalization, making it the most suitable for deployment.

Recommendation

Adopt the Random Forest Model for Deployment:

-Why: Given its superior performance in terms of accuracy and ROC-AUC, the Random Forest model is recommended for deployment in identifying high-risk diabetes patients. It can provide reliable predictions even in complex, non-linear data scenarios.

-Implementation: This model should be integrated into the healthcare provider's system to flag high-risk patients, enabling timely interventions.

Use Logistic Regression for Interpretability:

-Why: If interpretability is a critical requirement, particularly for explaining the model's predictions to non-technical stakeholders, the logistic regression model should be used. This model can serve as a secondary tool to help explain individual predictions or to provide a simpler, more interpretable option.

-Implementation: Utilize the logistic regression model to provide insights to healthcare professionals, helping them understand key factors influencing diabetes risk

Continuous Monitoring and Model Updates:

-Why: The healthcare environment and patient data can change over time, potentially affecting model performance. Continuous monitoring is necessary to ensure that the model remains accurate and reliable.

-Implementation: Establish a process for regular model evaluation and retraining using new data to maintain and improve predictive accuracy over time.

Explore Additional Features:

-Why: Adding more relevant features (e.g., lifestyle factors, family medical history) could enhance the model's predictive power.

-Implementation: Work with domain experts to identify and integrate additional features that could provide more insight into diabetes risk factors, potentially improving model performance.

THANK YOU