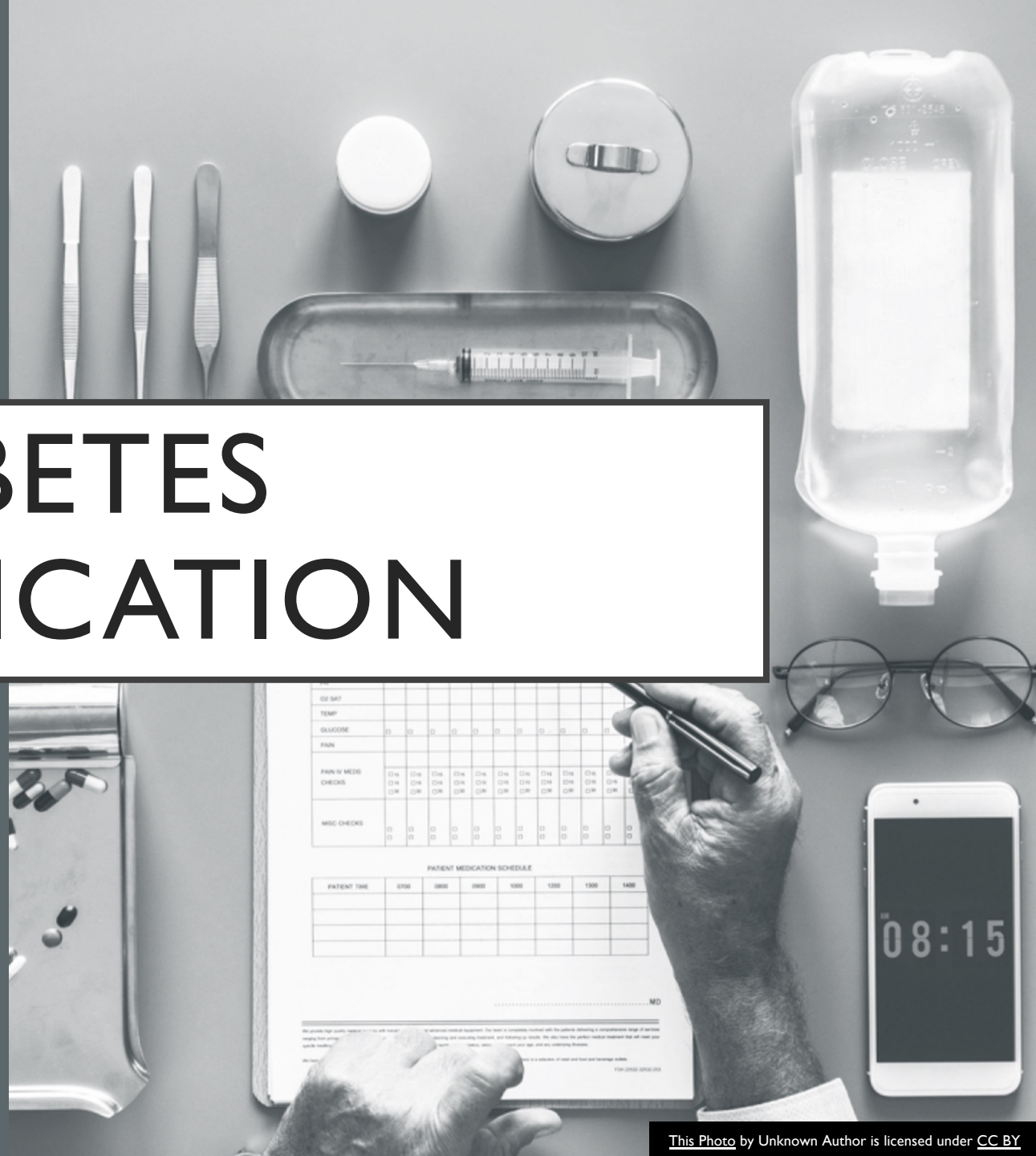
DIABETES CLASSIFICATION

Abrar Alsharidi

Morooj Aldeeb



CONTENTS

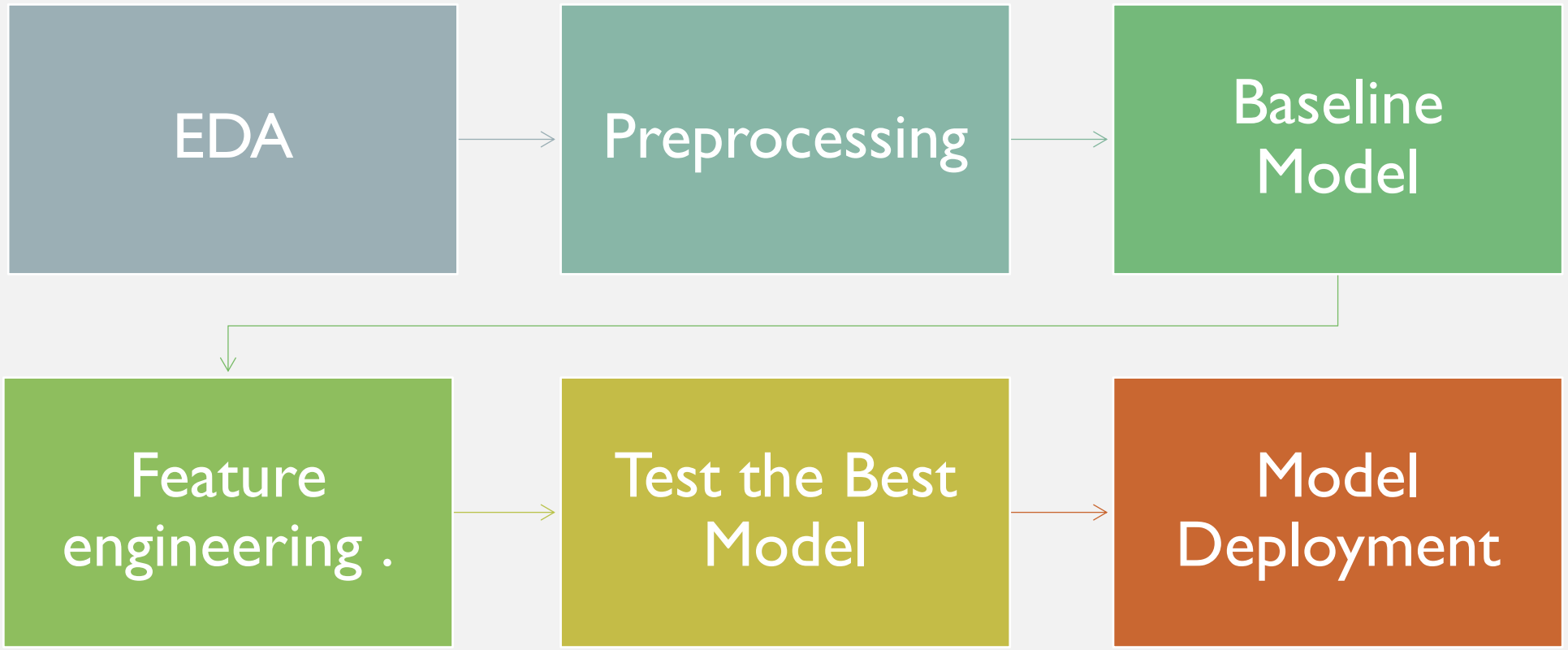
INTRODUCTION

- In this project and according to the data set, we will use supervised machine learning techniques. Particularly, binary classification to predict the outcome of a diabetic patient based on set of features which are thought to be very important, and which are likely able to explain the variability in disease outcome in different patients

OBJECTIVE

predict the diabetes degree based on the health condition of the patient and the frequency of patient visits to hospitals

METHODOLOGY



METADATA

- The data source used was the UCI website, it represents 10 years (1999-2008) of clinical care at 130 US hospitals.
- Data frame shape:-

Original data:101766 observations and 50 columns → After EDA: 99340 observations and 40 columns → After Feature engineering: 99340 observations and 37 columns

Tools and Libraries

Tools: Jupyter notebook, Python

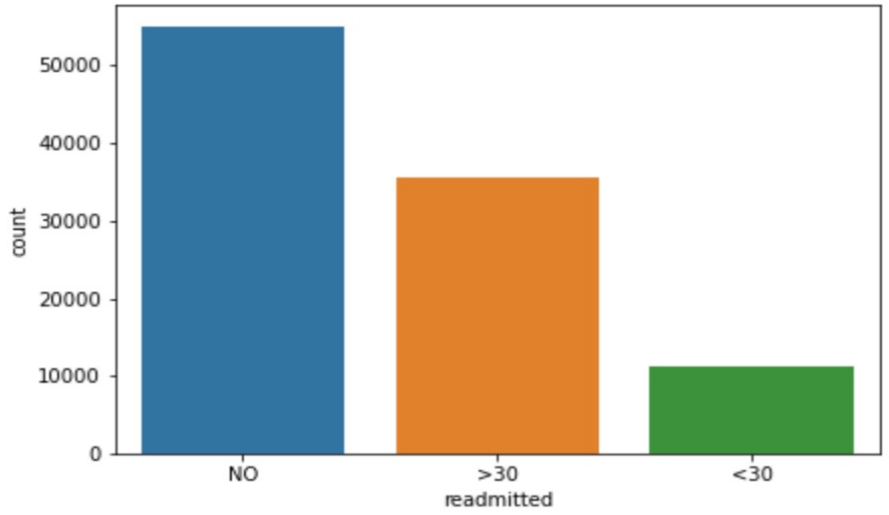
libraries: Pandas ,NumPy ,Matplotlib. ,Seaborn, Scikit-learn

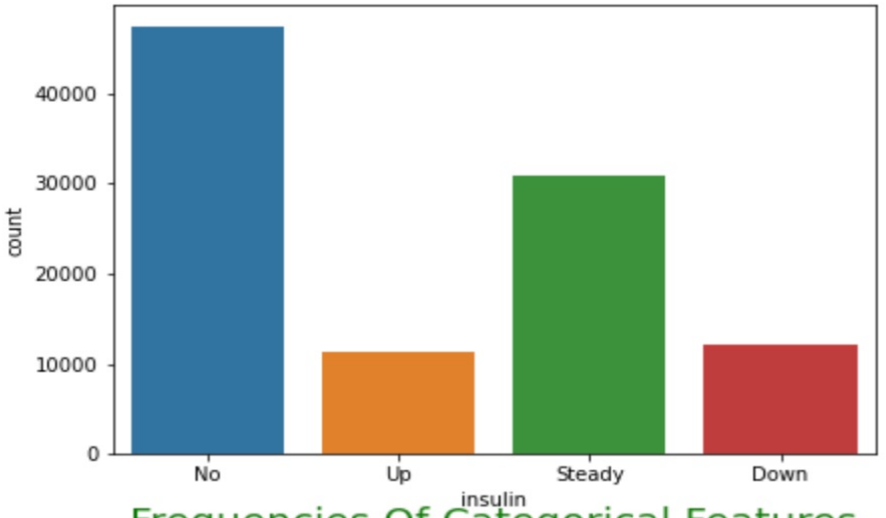
EDA

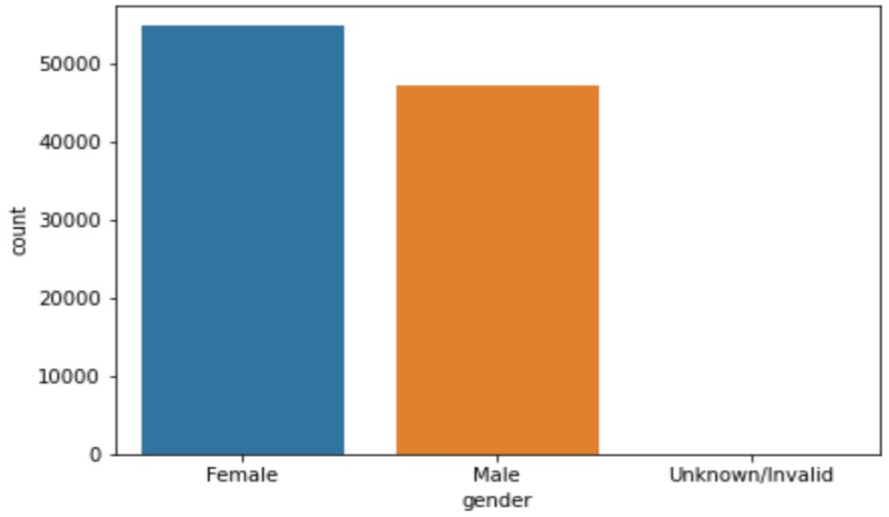

MISSING VALUES
-Dropping null columns
Categorical Values
- get dummies
- Label Encoder
(metformin)
Dropping Unneeded Column
Creating New Columns
- readmit & diabetes

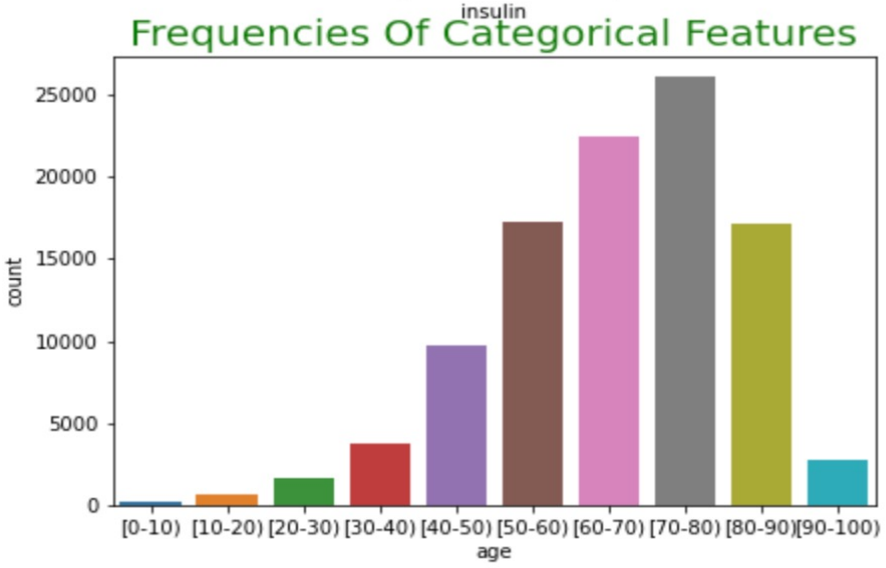
DATA UNDERSTANDING

Viewing the variables in the dataset to understand the dataset better



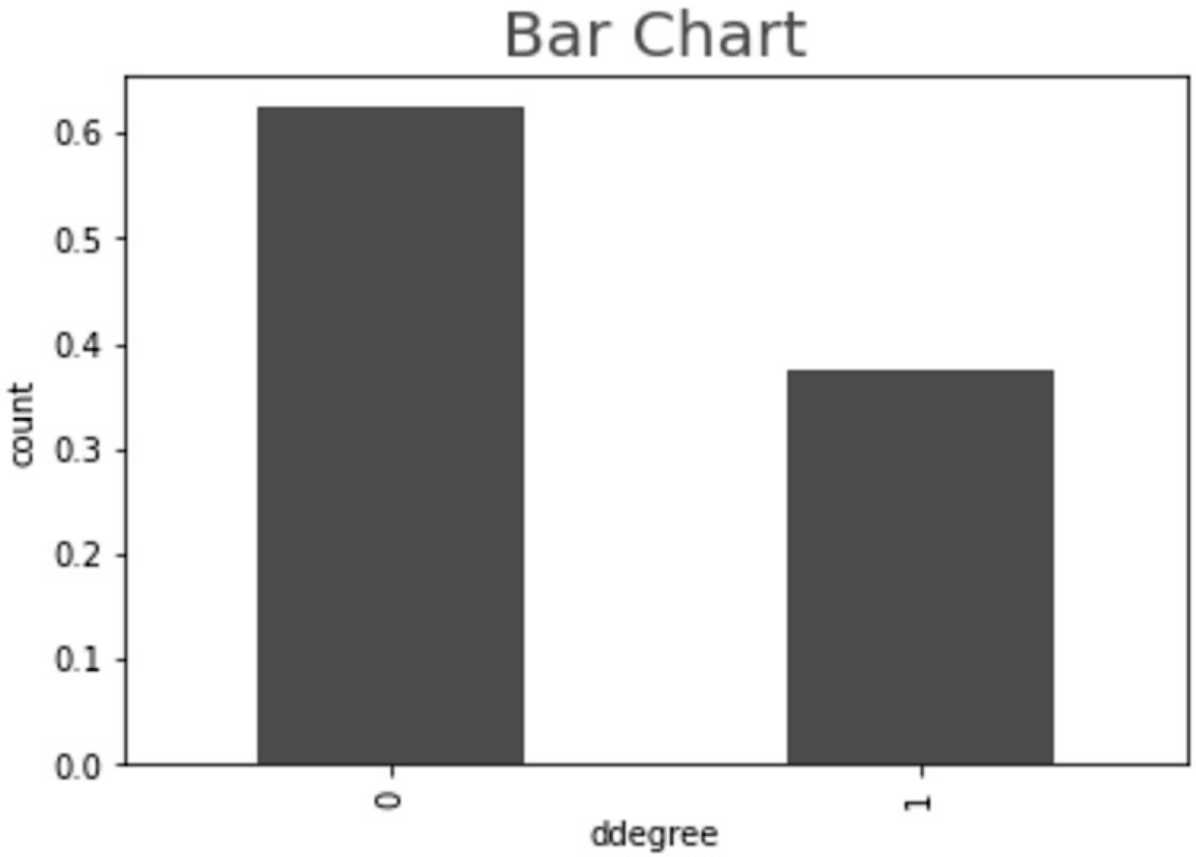






DIABETES DEGREE

Here it is clear that the data is nearly balanced



CLASSIFICATION MODELSLASSIFICATION MODELLING

We split our data into: training set: 80%, Validation 10% and Testing 10%

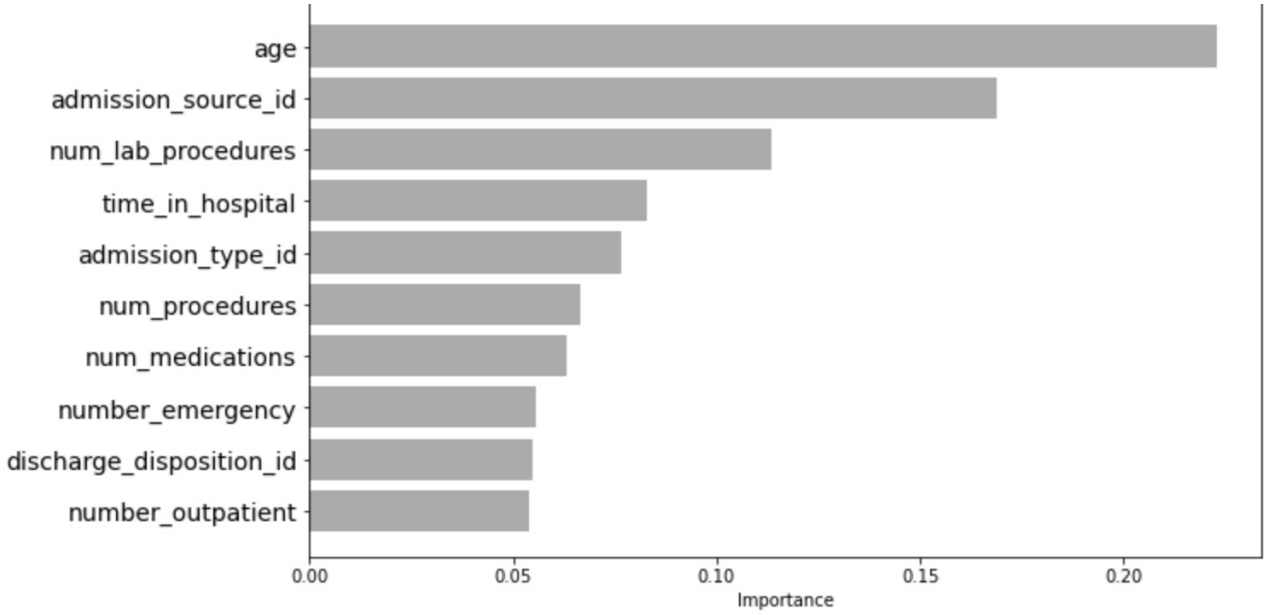
Model Name	Training Score	Validation Score
K-nearest Neighbors	0.82	0.66
Logistic Regression	0.66	0.43
Decision Tree	0.66	0.53
Random Forest	0.99	0.69
XGBoost Classifier	0.71	0.70
Bernoulli	0.66	0.66
AdaBoost	0.67	0.67

FEATURE ENGINEERING

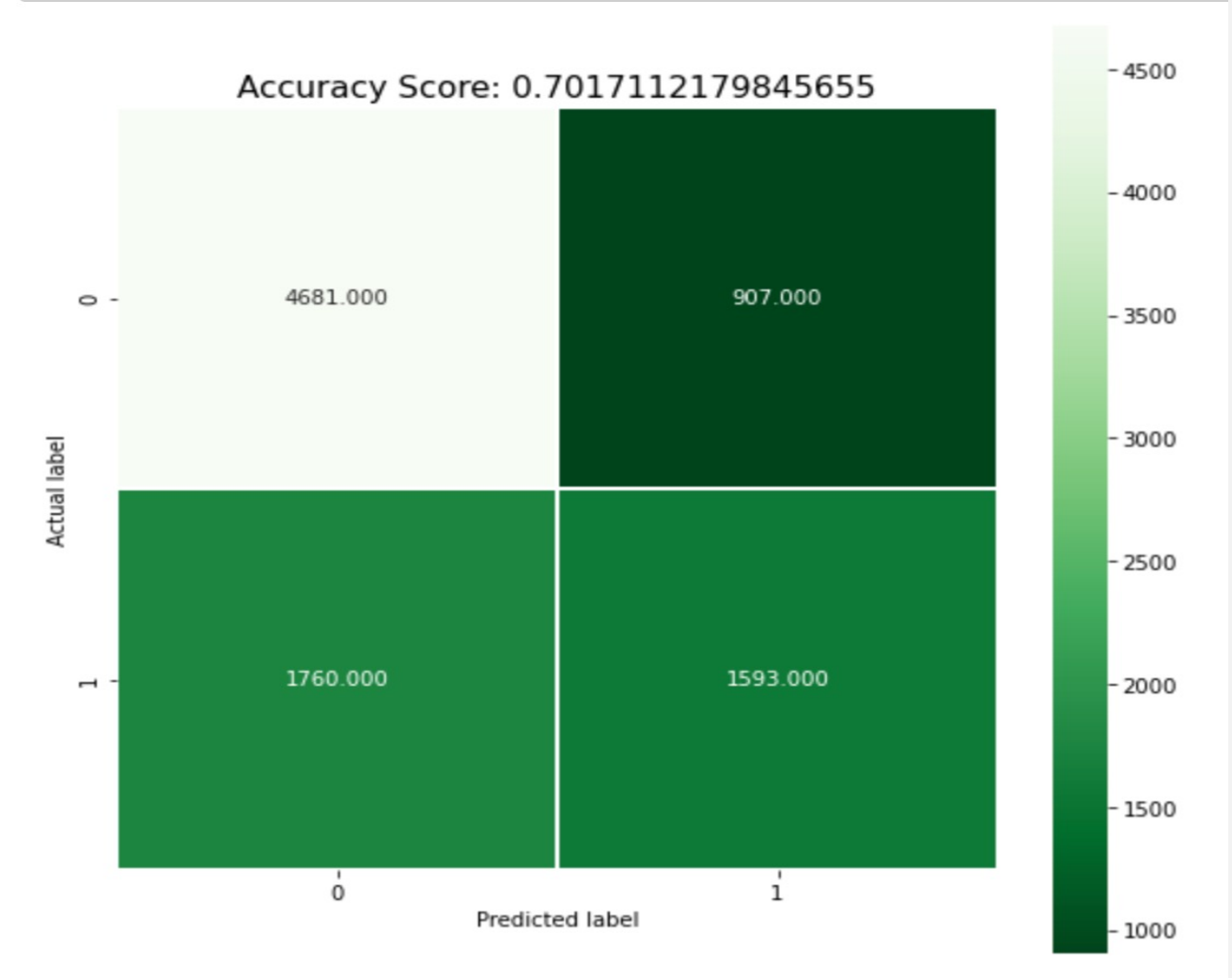
- We converted Readmitted from Nominal into binary by encoding No=0 and others=1. while DiabetesMed was encoded as yes=1 and no=0 and created a completely new column called ‘ddegree’ using Logic AND between Readmitted and DiabetesMed.
- We used Standard Scaler (Z-score).
- We used the top 10 most influential features.

BEST 10 FEATURES AFTURE ENGENEERING

- Find the most important columns with our target and remove the other.



(CONFUSION MATRIX)
XGBOOST CLASSIFIER



HYPER-PARAMETERS

- The best model tested so far is XG Boost followed by Random Forrest.

```
n_estimators=30000,
max_depth=4,
objective='binary:logistic',
learning_rate=.05,
subsample=.8,
min_child_weight=3,
colsample_bytree=.8
```

CLASSIFICATION MODELS

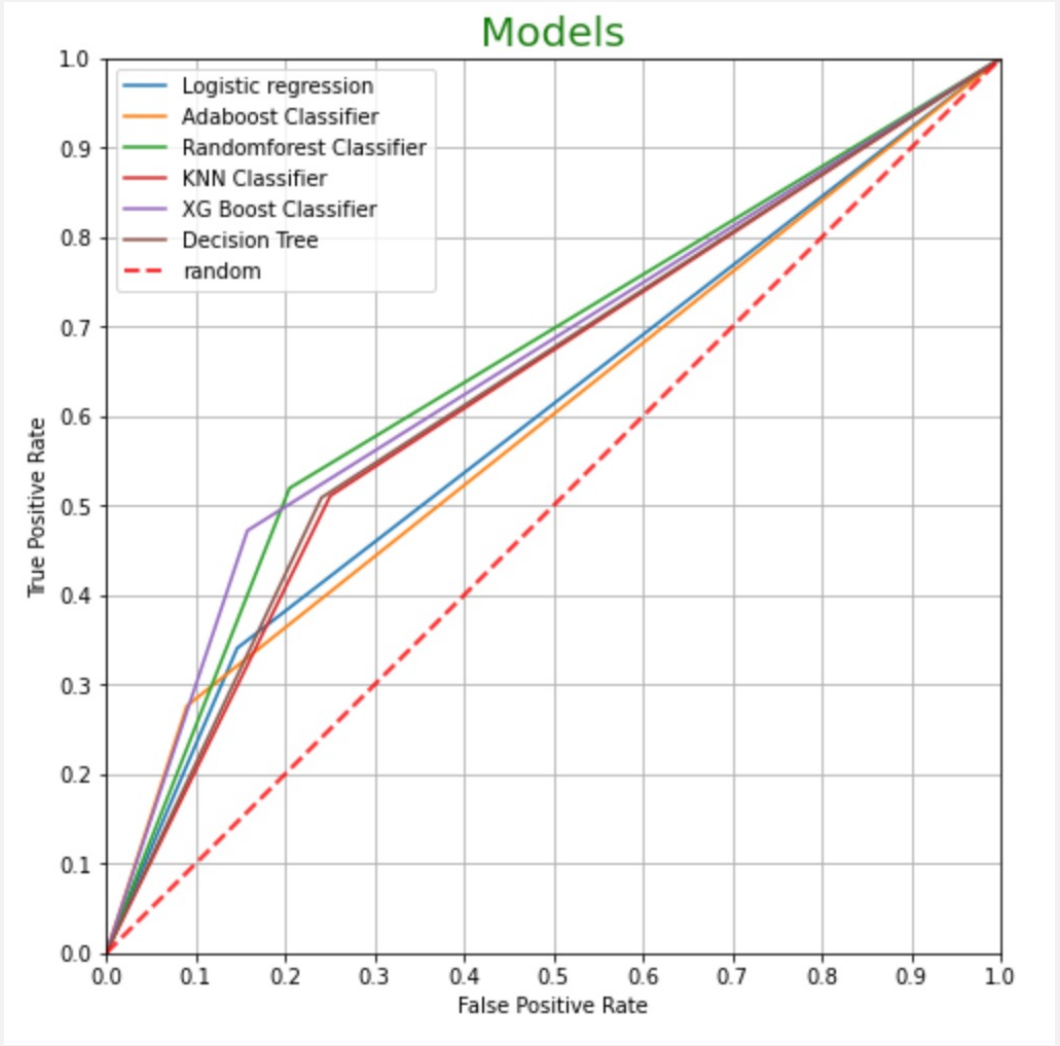
Tunned Models	Accuracy	Recall	Precision	F1 Score
K-nearest Neighbors	0.65	0.66	0.65	0.66
Logistic Regression	0.66	0.34	0.59	0.43
Decision Tree	0.66	0.50	0.56	0.53
Random Forest	0.69	0.53	0.59	0.56
XGBoost Classifier	0.70	0.47	0.63	0.54

XGBOOST CLASSIFIER

Dataset	Accuracy	F1 Score
Testing	0.70	54

MODELS(ROC CURVE)

Explains the best two models in terms of accuracy (XG boost) and Random Forrest Classifiers



FUTURE WORK

- Applying more complex models to classify diabetic patients perhaps with better results. i.e., Deep Learning (Neural networks) models.

THANK YOU FOR LISTEN ☺