Tasmia Kayenat

Capstone 2 - Final Report

Predicting Hospital Readmissions Within 30 Days Using Patient and Clinical Data.

Problem Statement:

Everyday we see thousands of people getting admitted to the hospitals. Even after providing the same care, some of the patients need to get readmitted to the hospitals. Is age a factor here? Does race make a difference when it comes to hospital readmissions? Can we predict the pattern of readmissions based on the variables like age, race, gender?

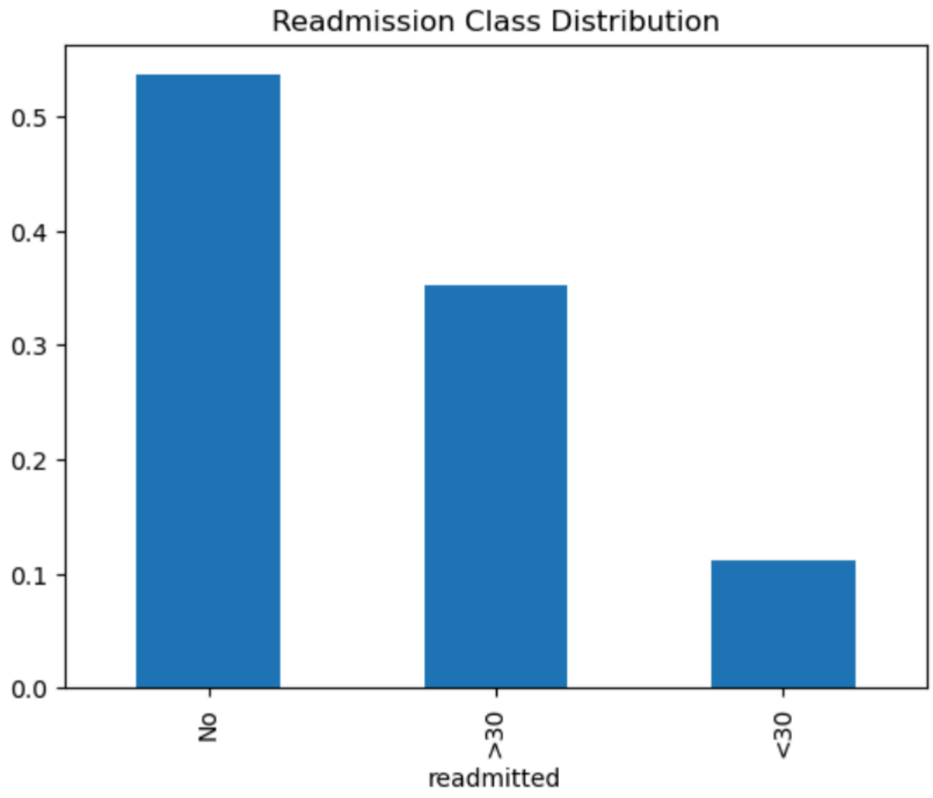
My aim is to develop a machine learning model that predicts the likelihood of a patient being readmitted to the hospital within 30 days after discharge. We know that unplanned readmissions are not only costly but also reflect the quality of patient care. By identifying high-risk patients using factors such as age, primary diagnosis, length of stay, discharge disposition, and comorbidities, hospitals can implement targeted interventions to reduce readmission rates.

Data Wrangling:

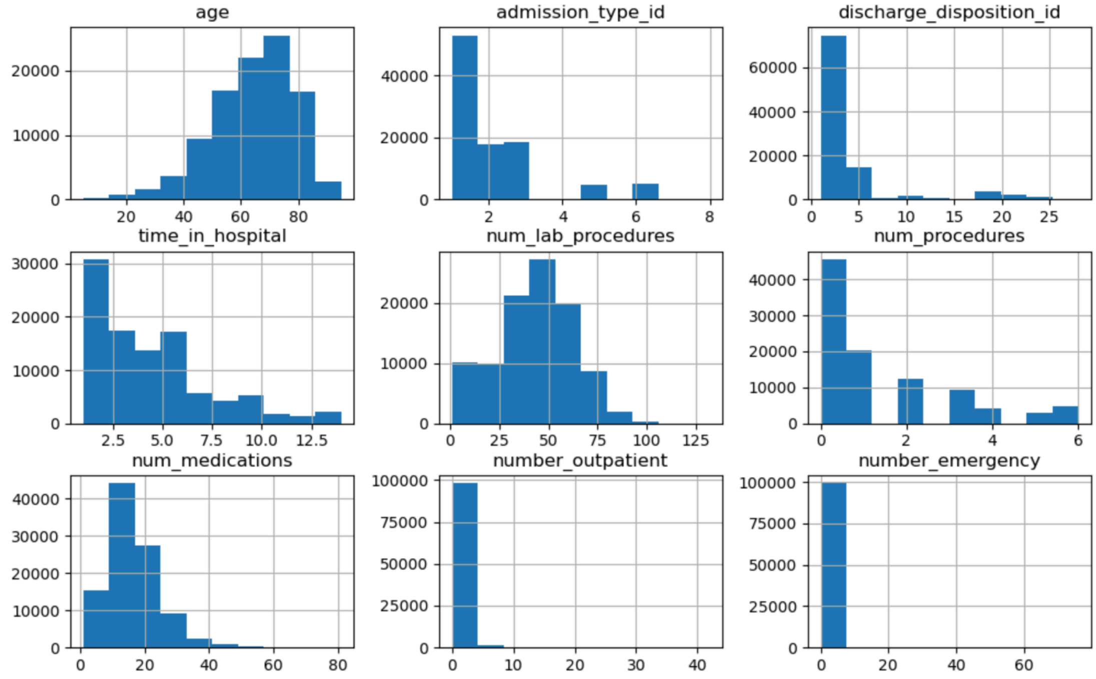
The raw data was taken from Kaggle, which is the data of Diabetic patients being admitted to 130 US hospitals between the years 1999 - 2008. It primarily had 101766 rows and 50 columns. I started by loading the libraries and the data. Then I checked the dataset, its shape, the columns and how the data looked. After looking at the dataset, I noticed there were lots of null values or placeholders. Columns with '?' were replaced with NaN and columns that were unnecessary were dropped. The next step was to look for inconsistency and some columns had duplicate values. Duplicated values were removed and inconsistent values were converted. Then I checked for outliers and finally saved the clean data that had 101766 rows and 43 columns.

Exploratory Data Analysis:

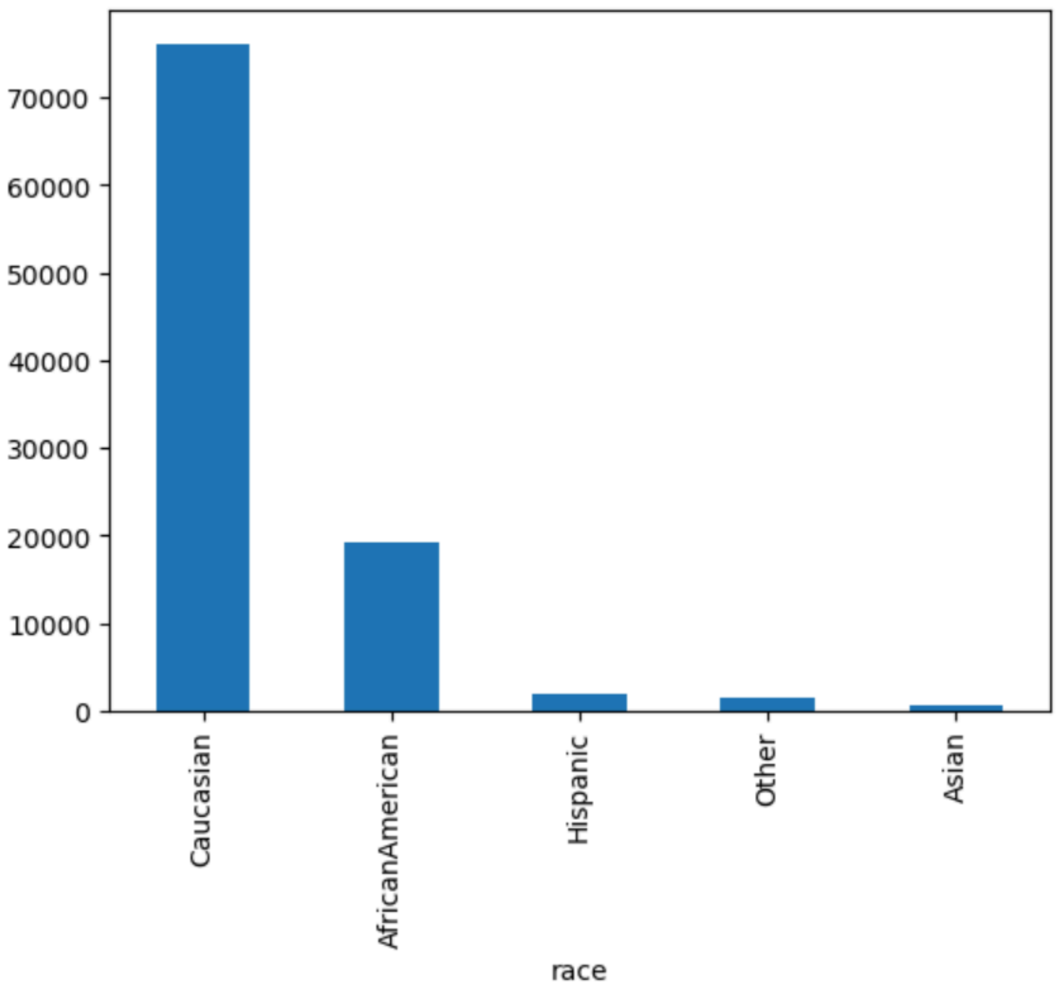
For this part of the project, I checked the cleaned data, did some data wrangling, uncovered meaningful patterns and relationships among variables. From the first visualization, we can see that the rate of people getting readmitted within 30 days was the lowest.

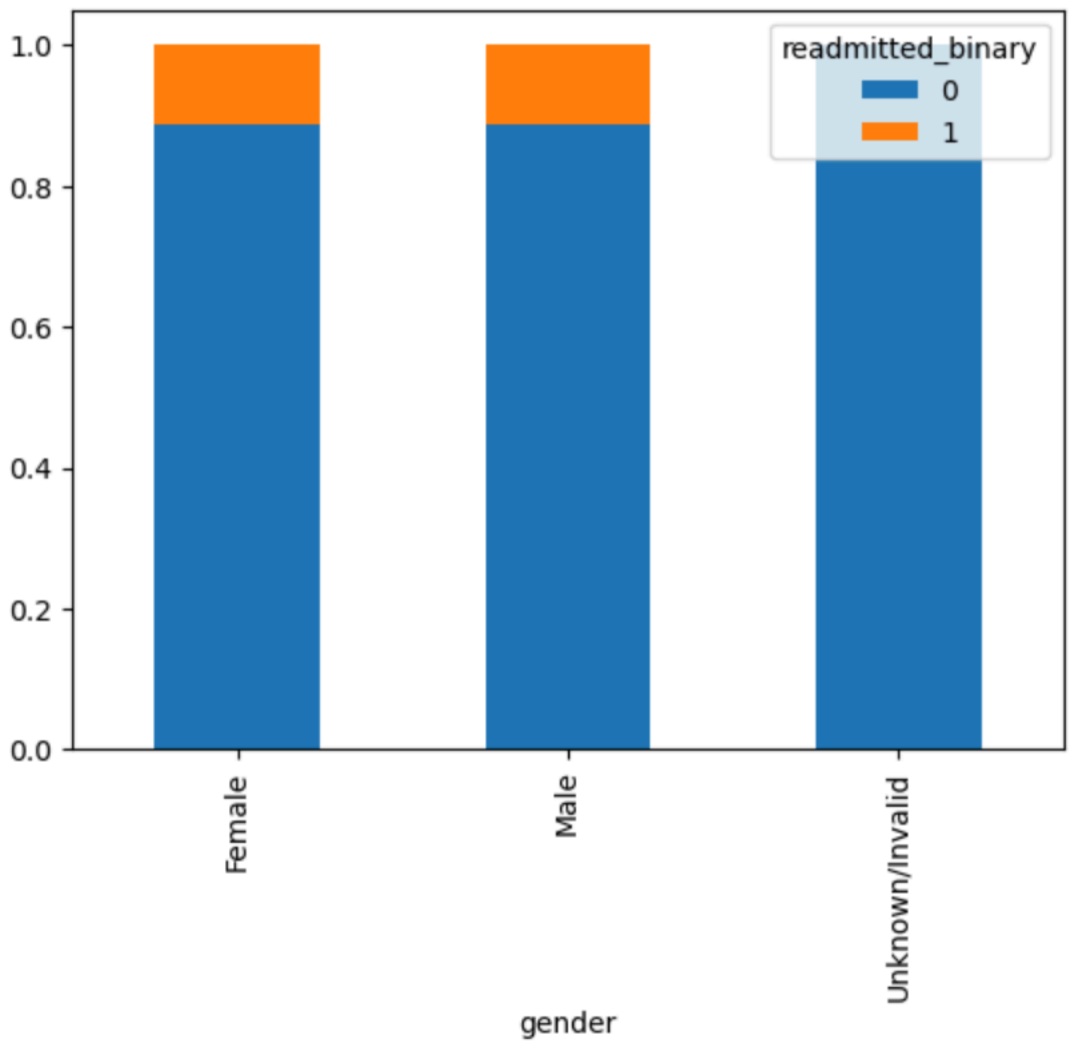


The dataset has two types of data; numeric variables and categorical variables. I analyzed the numeric variables and below is the visualization:



We can see that most patients fall between the age of 60 - 80. Race and gender distributions were a bit imbalanced with white and female being the majority of patients.


70000
60000
50000
40000
30000
20000
10000
0
Caucasian
AfricanAmerican
Hispanic
Other
Asian
race



The dataset was found to be imbalanced with fewer positive readmission cases, which influenced model selection and evaluation strategy.

Preprocessing, Training and Modeling:

To ensure the dataset was ready for machine learning, several preprocessing steps were performed including encoding categorical features and preparing for modeling. Label encoding was applied to binary categorical features such as gender. Then the train test split was performed. The cleaned and encoded dataset was split into training 80% and testing 20% subsets using stratified sampling to maintain class distribution.

Preprocessing Step	**Action Taken**
Missing values handling	Dropped high missing columns
Categorical encoding	One hot encoding
Train test split	Stratified 80/20 split

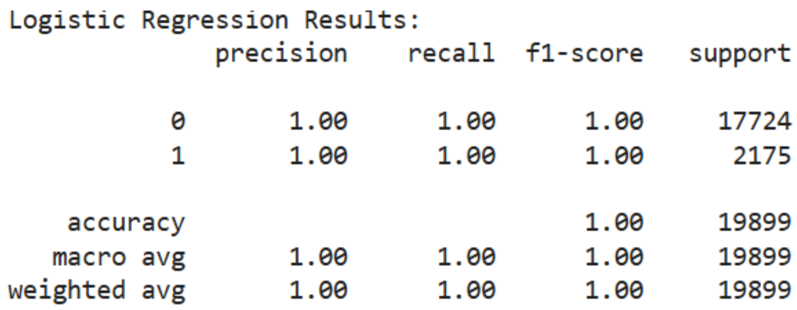
Several classification models were trained and evaluated to identify the best-performing algorithm based on various performance metrics. The models that were trained using the preprocessed data are:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- XGBoost Classifier

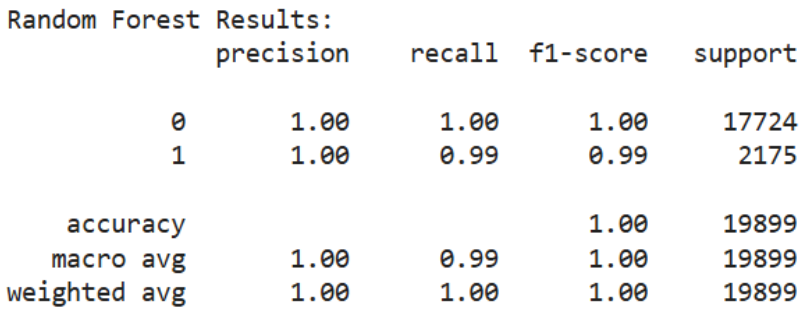
The metrics that were used are:

- Accuracy
- Precision
- Recall
- F1 Score
- ROC AUC Score

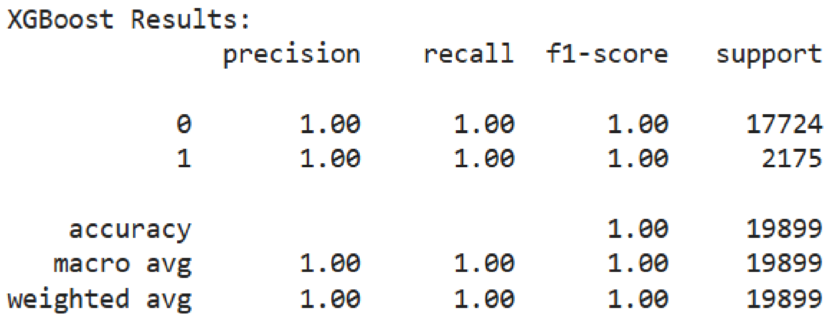
The result summary:



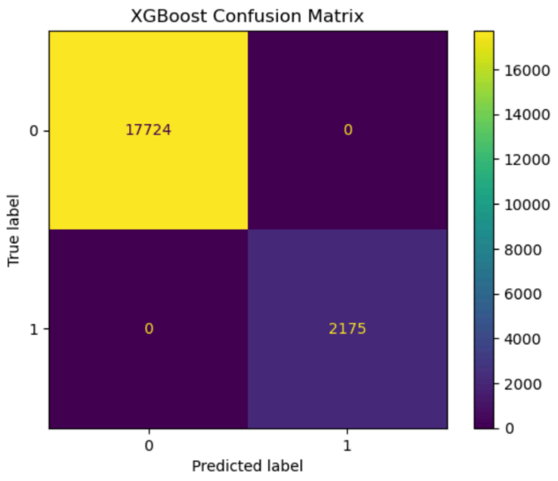
```
Logistic Regression Results:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     17724
           1       1.00      1.00      1.00      2175

    accuracy                           1.00     19899
   macro avg       1.00      1.00      1.00     19899
weighted avg       1.00      1.00      1.00     19899
```



```
Random Forest Results:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     17724
           1       1.00      0.99      0.99      2175

    accuracy                           1.00     19899
   macro avg       1.00      0.99      1.00     19899
weighted avg       1.00      1.00      1.00     19899
```



```
XGBoost Results:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     17724
           1       1.00      1.00      1.00      2175

    accuracy                           1.00     19899
   macro avg       1.00      1.00      1.00     19899
weighted avg       1.00      1.00      1.00     19899
```



Conclusion:

My project aimed to predict hospital readmissions using various machine learning models, including Logistic Regression, Random Forest and Gradient Boosting. Despite thorough preprocessing and model tuning, the predictive performance across these models had poor outcome.

AUC Scores: Ranged from approximately 0.47 to 0.53, indicating limited discriminative ability.

Accuracy: While some models achieved higher accuracy, this was misleading due to class imbalance, with models predominantly predicting the majority class.

Confusion Matrices: Revealed that models struggled to correctly identify readmitted patients, often misclassifying them as non-readmitted. The number of medications, patient age, and prior length of stay, indicating that patient history and treatment complexity play roles in readmission likelihood.

Several factors contributed to the models' limited performance: Insufficient Feature Set: The dataset lacked comprehensive clinical details such as specific procedures, medication types, and vital signs, which are crucial for accurate predictions

Class Imbalance: A disproportionate number of non-readmitted patients led to models biased towards predicting the majority class.

Model Limitations: Standard machine learning models may not capture the complex patterns associated with hospital readmissions without richer data.