Project Report

Summary

In this project, we have worked on the 'diabetic_data.csv' dataset. The given dataset contains records of diabetic patients admitted to US hospitals from 1999 to 2008.

We have performed **Data Cleaning**, **Data Transformation**, **EDA (Exploratory Data Analysis)**, and the **Model Prediction** using **Random Forest Classifier** and **Model Evaluation** using **Cross-Validation Procedure**.

Part 1: Building up a basic Predictive Model

1. Data Cleaning and Transformation:

- This Data Frame consists of **101766 Rows** and **50 Columns.** With the **missing values** in the form of **"?"** in the dataset.
- Only column **"weight"** has **missing values** of **more than 50%** compared to other columns.
- Transformed the **"age"** column in the middle value according to the given range.
- Replaced the possible missing values in columns **diag_1**, **diag_2**, and **diag_3** by the **number 0**.
- Converted the values of the column "readmitted" into **0** for **"NO"** and **1** for **"<30"** and **">30**" for binary classification.
- Dropped **unnecessary columns** for **easy data exploration**.

Correlation of other numeric columns with respect to "readmitted":

```
readmitted                1.000000
number_inpatient          0.217194
number_diagnoses          0.112564
number_emergency          0.103011
number_outpatient         0.082142
patient_nbr               0.074093
time_in_hospital          0.051289
num_medications           0.046772
admission_source_id       0.039986
num_lab_procedures        0.039253
admission_type_id        -0.004923
discharge_disposition_id -0.014852
encounter_id             -0.038267
num_procedures           -0.044748
```

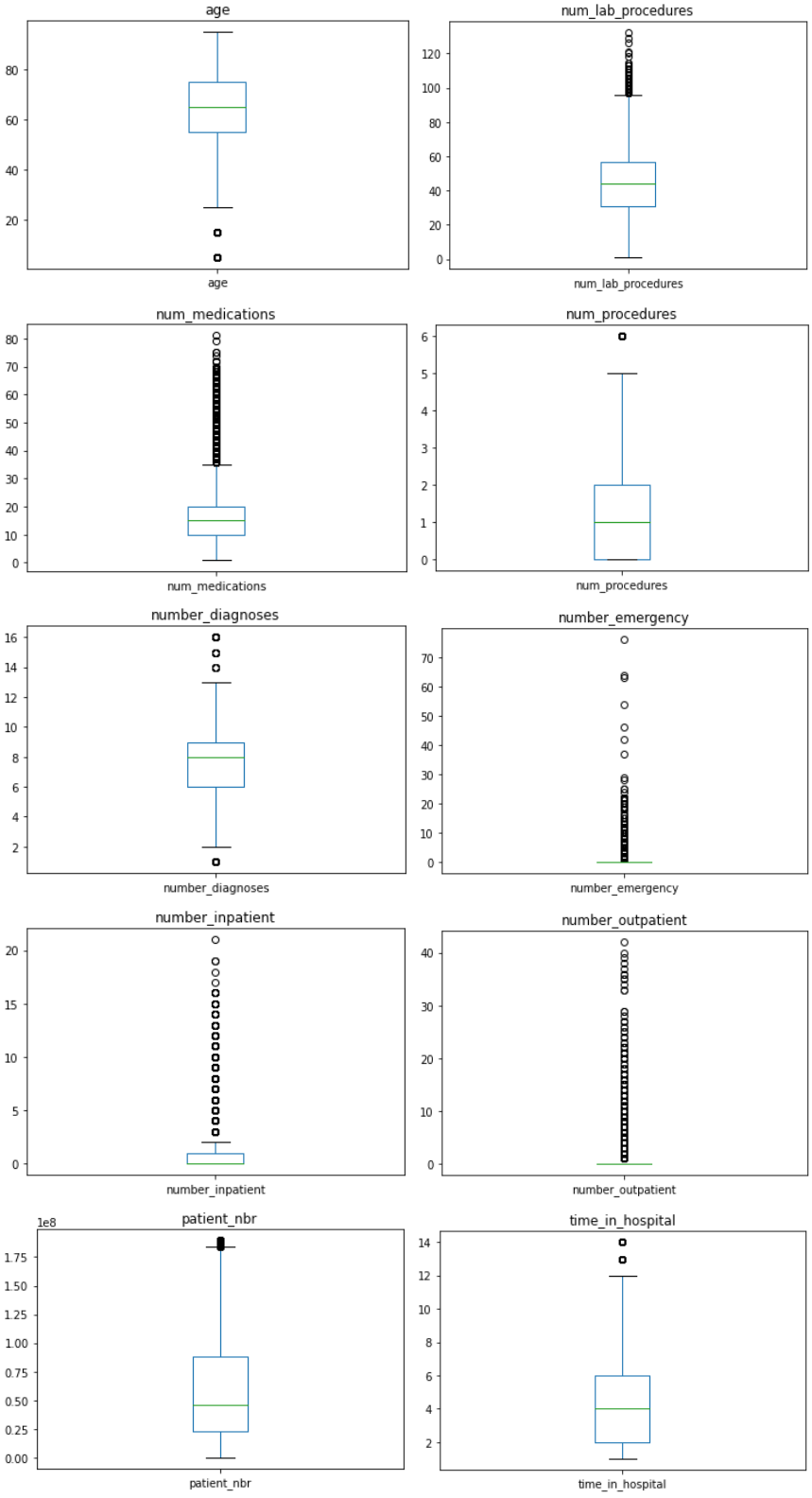
Categorical Features:

- The categorical features after the removal of unnecessary features are:- **'race', 'gender', 'diag_1', 'diag_2', 'diag_3', 'change'**

Numerical Features:

- The numerical features after the removal of unnecessary features are:- **'encounter_id', 'patient_nbr', 'age', 'time_in_hospital', 'num_lab_procedures', 'num_procedures', 'num_medications', 'number_outpatient', 'number_emergency', 'number_inpatient', 'number_diagnoses', 'diabetesMed', 'readmitted'**

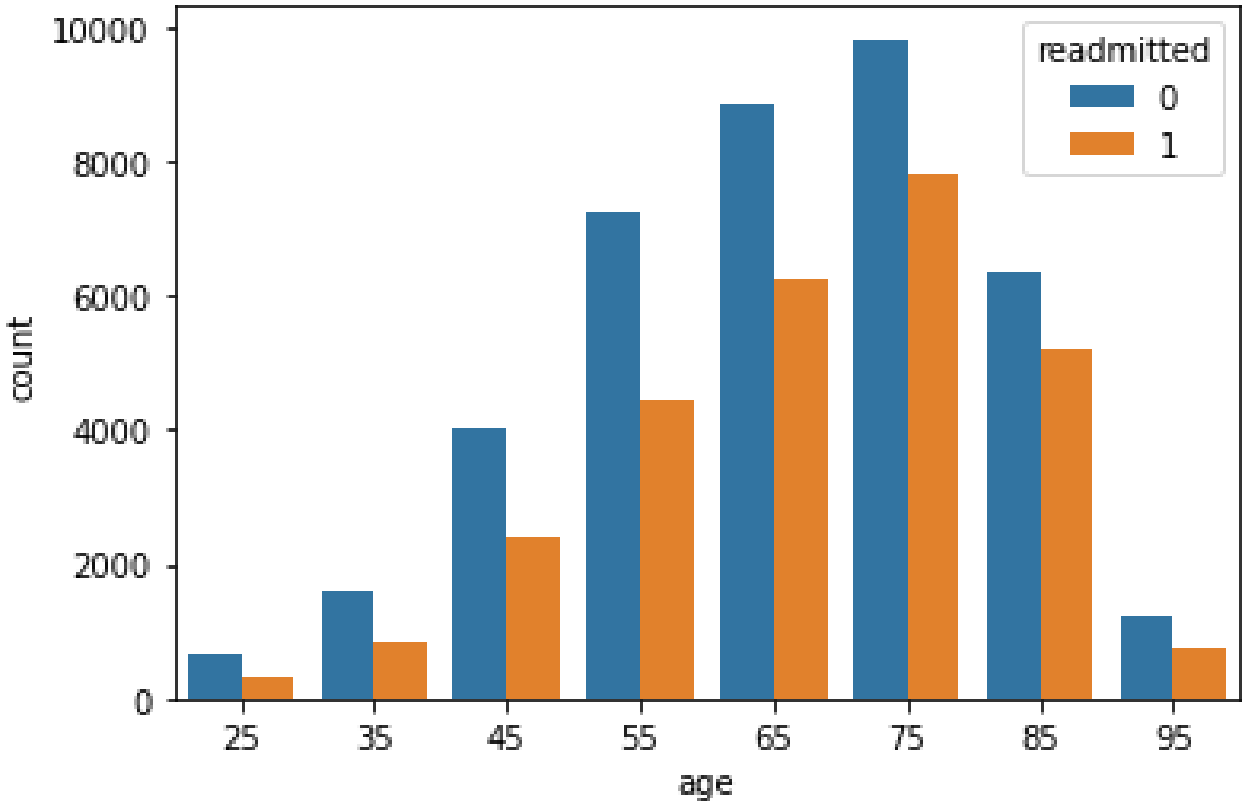
Outliers in the Dataset:



- After removing outliers, the resulting data frame has **67799 Rows** and **13 Columns**.

2. Data Exploration:

➢ **Impact of "age" on "readmission" using CountPlot:**

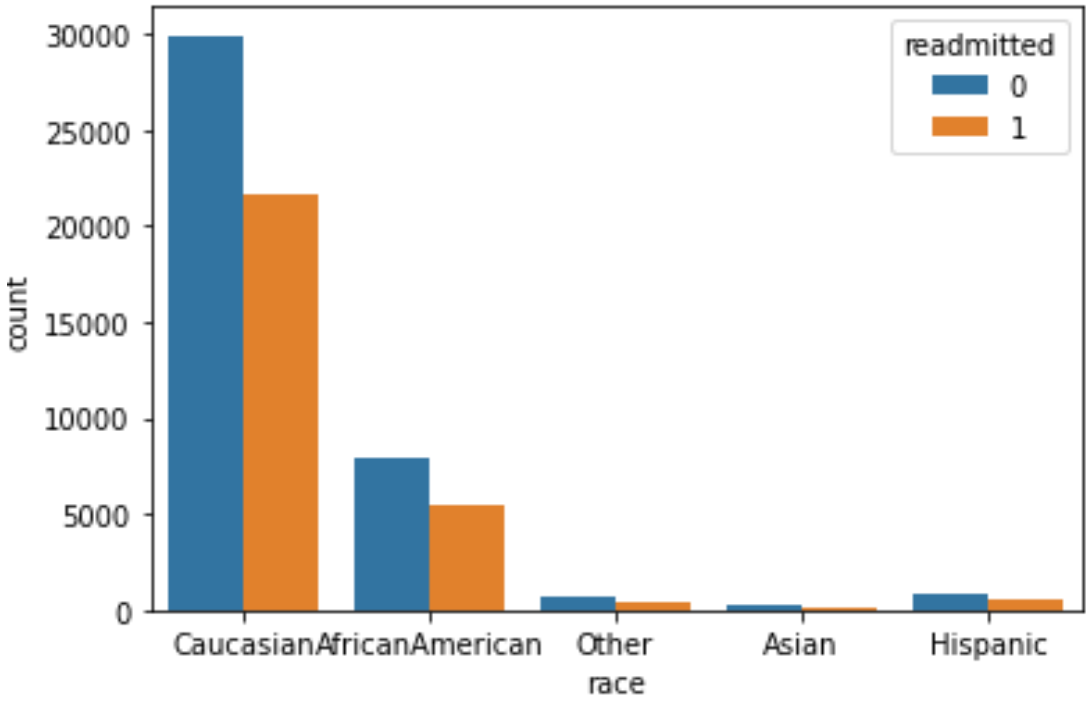


Observation:

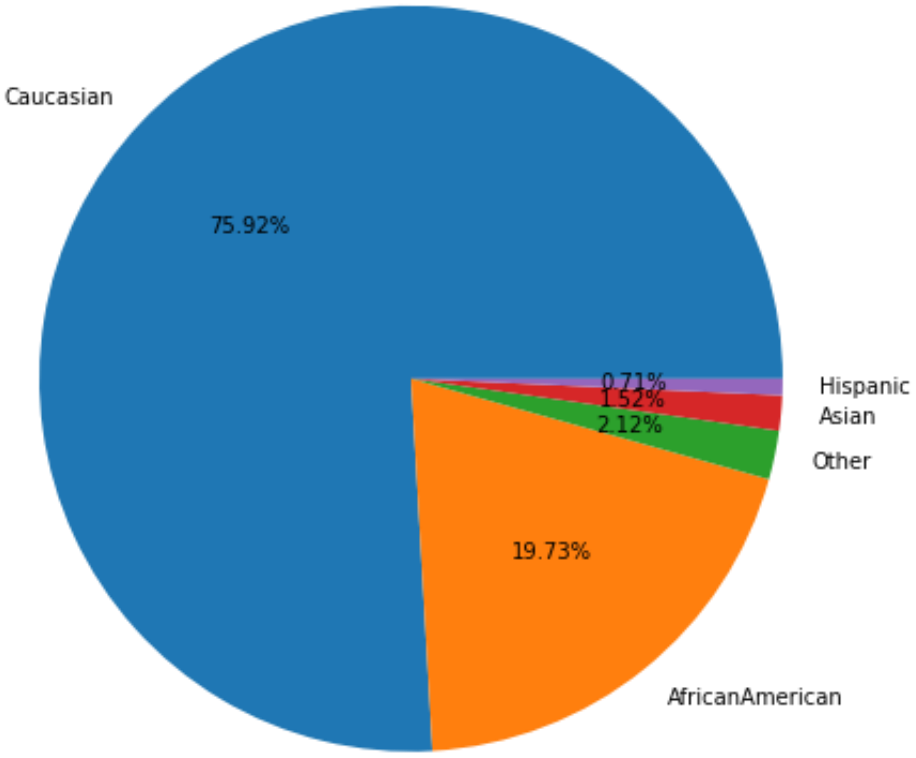
- As the age increases, the chances of getting readmitted increases

➢ **"readmission" according to the "race" using CountPlot:**

- Caucasian 51472
- AfricanAmerican 13379
- Hispanic 1435
- Other 1031
- Asian 482



Race availability using PieChart:

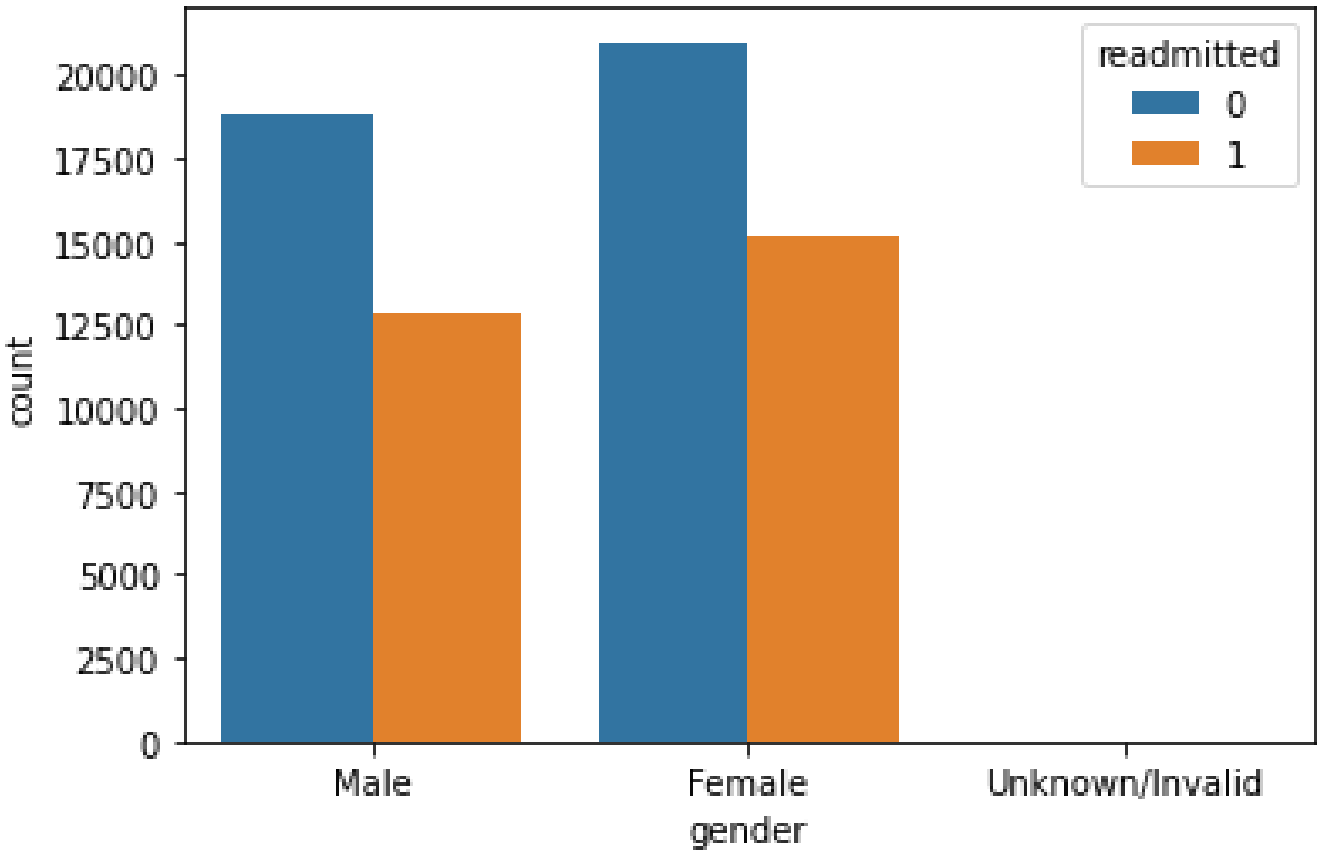


Observation:

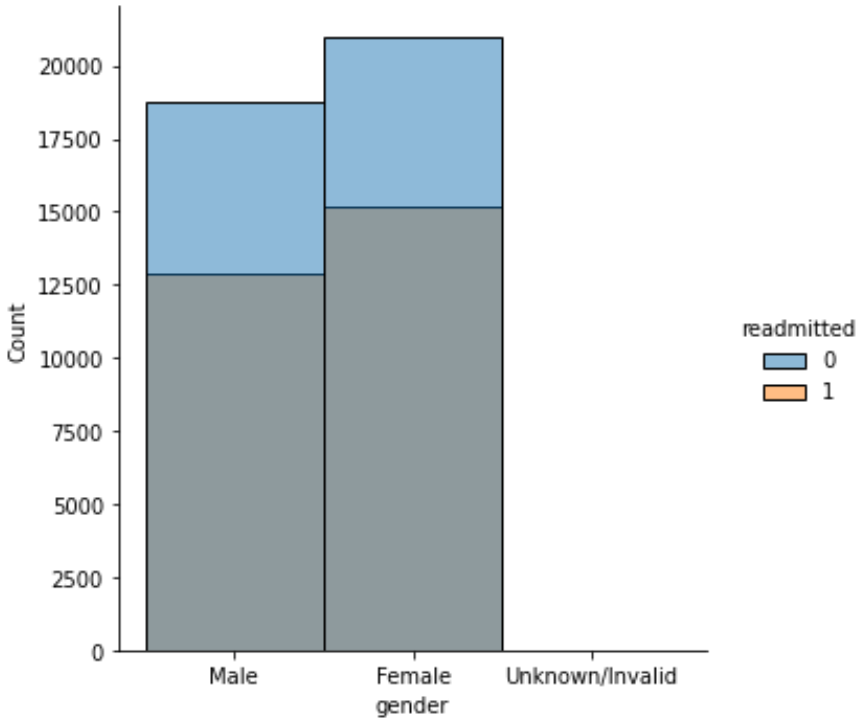
- Availability of **Caucasian** is **75.92%**, **AfricanAmerican** is **19.73%**, **Hispanic** is **0.71%**, **Asian** is **1.52%**, and **Other** is **2.12%**.

- African Americans are more likely to be readmitted than other ethnic groups as their readmission rate.

➢ **Readmission According to Gender using CountPlot:**



Displot:



Observation:

- Women patients are more likely to be readmitted than men.

- **Impact of Diagnose type on readmission:**

Observation:

- As Diagnose types i.e. diag_1, diag_2, diag_3 are 'object' data type columns and contains a huge numbers of values.
- Therefore, they can't be plotted against readmitted feature.

3. Model Building:

Procedure:

- Separating Independent **"x"** and Dependent **"y"** variable.
- Train Test Splitting the dataset in **x_train**, **x_test**, **y_train**, **y_test**.
- Building the **Random Forest Classifier** and training the model.

Model Evaluation:

Confusion Matrix:

[6097, 1902]

[3393, 2168]

Cross Validation Methods:

	precision	*recall*	*f1-score*	*support*
0	*0.64*	*0.76*	*0.70*	*7999*
1	*0.53*	*0.39*	*0.45*	*5561*
accuracy			*0.61*	*13560*
macro avg	*0.59*	*0.58*	*0.57*	*13560*
weighted avg	*0.60*	*0.61*	*0.60*	*13560*

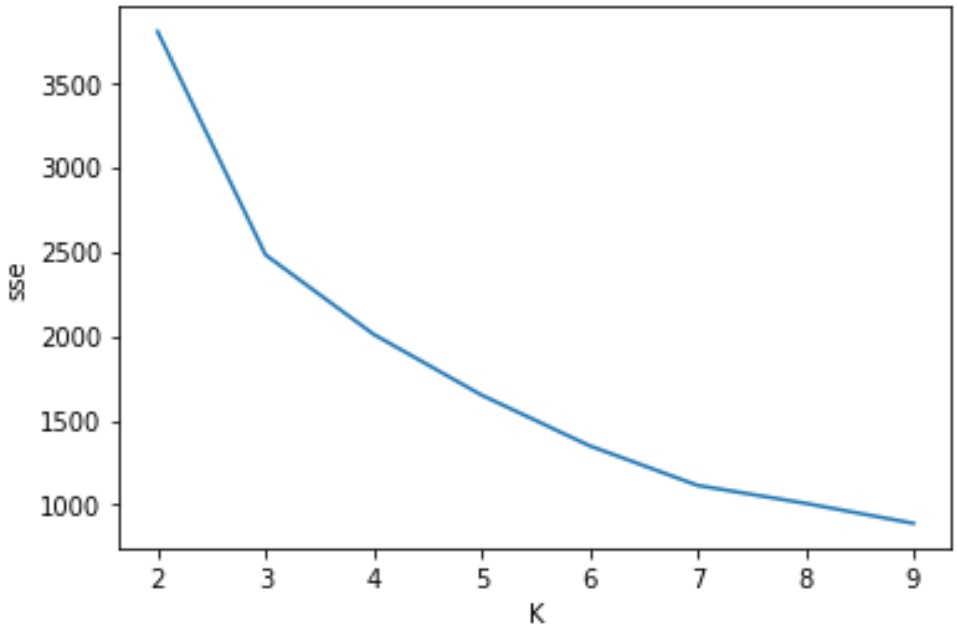
Observation:

- By Seeing this classification Report, we can conclude that the model did not predicted well with the testing data.
- By providing the accuracy of 61% approx.

Part 2: Improved Model

1. K-Means Clustering Approach:

- Taking 2 attributes **“age”** and **“num_lab_procedures”** for clustering.
- Scaling the attributes for better clustering using MinMaxScaler.
- Finding the better K value for the number of clusters in K-means clustering.
- Plotting the sum of squared error for K value.



- Training the clustering algorithm for clusters.
- Plotting the clusters.