The first section of this assignment covered data cleaning so it will not be discussed here, that is visible in the jupyter file. This report will review the data exploration from the diabetes dataset of patient admissions from 1999 to 2008 in US hospitals as well as the predictive model built, clusters and includes recommendations on the current model built.

# Part 1.2:

Chart, bar chart

Description automatically generated**Figure 1: Histogram of age Impact on readmissions**

A black screen with white text

Description automatically generated with low confidence**Figure 2: Percentages of all age groups and readmission**

The age column of the graph has been set to be the middle value of an age group for example an age group of [40-50] was set to 45. In figure 1 above, the graph represents the number of times a patient has or has not been readmitted back into the hospital, the orange column (1) represents a patient who has been readmitted while the blue column (0) represents a patient who has not been readmitted. On the graph, it can be seen as a patient’s age rises, the rate of readmissions rises with it. This rise stops at 75 and has a decline in the last two columns, this decline could be affected by other factors like the average life expectancy in the US at the time was around 80 years old. This graph can therefore prove the hypothesis of age having a higher impact on readmissions until a certain age; in our graph that age is 75 years old. A table (Figure 2) has also been made as it was noticed that more people in other age groups would have checked into a hospital for Diabetes, therefore percentages adding the total readmissions for each age group and dividing it by whether a patient has been readmitted or not has been done. This table further proves the hypothesis as age increases the chance of readmission is higher.

The next hypothesis explores the impact race has on readmissions.

Chart, bar chart, funnel chart

Description automatically generated**Figure 3: Histogram of race impact on readmissions**

A picture containing text, road, screenshot

Description automatically generated**Figure 4: Histogram of age Impact on readmissions**

In figure 3, the graph represents the impact race has on readmissions. Similar structure to Figure 1 with the orange column representing who has been readmitted and the blue column who has not been readmitted. Figure 4 presents the percentage of readmissions for each ethnic group, in which it shows that African Americans has the third highest readmission rate with 32.77% compared to Caucasians which has 38.89% and the group called other which has 35.01%. A relationship between race and readmissions cannot be seen therefore race does not have any impact on whether a patient would be readmitted again.

The relationship between gender and readmission will be explored next.

Bar chart

Description automatically generated**Figure 5: Histogram showing gender impact on readmissions**

Graphical user interface, text, application

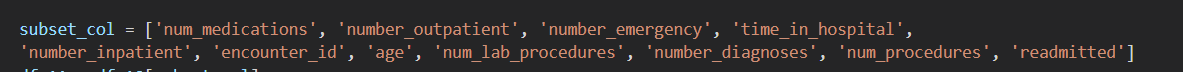
Description automatically generated **Figure 6: Histogram showing gender impact on readmissions**

In figure 6, the histogram represents gender impact on readmissions. Overall, there were a higher number of women patients being readmitted than men patients. This proves the hypothesis of women being more likely to be readmitted than men. The table has been added due to a higher number of female patients compared to males, a percentage shown in figure 6, which was calculated to find if gender has an impact on readmissions. Women do have a slightly higher readmission rate compared to and the very close values of a difference of 0.66% between male and female readmissions which could be seen as insignificant, but for the hypothesis of women being more likely to be readmitted than men it would be seen as true.

Graphical user interface, text, application

Description automatically generated**Diagnosis types impact on readmissions**

**Model Building**

The subset columns provided were used as the features while targeting the readmitted column.

After this step the dataframe was immediately split into training and test sets with a test size of 20% and a training size of 80%. We used the logistic regression model as this was more time efficient that other models, this was prioritised over accuracy. Once this was completed a confusion matrix was plotted.

A confusion matrix is used to evaluate and summarise he performance of the model.

True negative values on the confusion matrix below shows the number of negative values that are classified accurately, in this model the true negative was 2425 which means the negative values were classified accurately.

False negative shows the number of actual positive values that are classified as negative. In this model the false negative value was 1515 …….

False positive shows the number of actual negative values that are classified as positive, in this model the false positive was 0 which means …………

Chart, treemap chart

Description automatically generatedTrue positive shows the number of positive values that are classified accurately, in the model the true positive value was 0 which means there are no correctly classified values .

Following the confusion matrix, a cross validation evaluation was made to find the accuracy of the model which was 0.63.

Calendar

Description automatically generatedThe classification report helps in checking how accurate a model is, the report shown below returns precision, recall and f1-scores for each class, with the help of the classification report we can see the precision and recall of 1 is 0.00 so the f1-score for it would also be 0.00, this means the model is not as accurate as it only predicting Not readmitted values. This would need to be fixed to improve the model.

**Improved Model**