D209 Classification Analysis Task 1

Part I

The purpose of this Data Mining Report is to be able to predict whether a person would be Readmitted into a Hospital. The model used is a K-Nearest Neighbor(knn) model. The goal is to identify key features that can help identify a patient having to be Readmitted. In identifying this, measures could be taken to ensure a patient is well educated in their health before they are released home.

Part II

The classification method looks at five variables to determine whether a person would eventually be readmitted into a hospital. The expected outcome was to have a high accuracy to be able to implement the model for future use. One assumption of the knn algorithm is that data is non-Parametric (Varghese, 2018). Distribution plots were created for all variables to determine whether the features were parametric or non-parametric. Once it was determined that the majority of the variables were nonparametric then knn was chosen.

Thirteen python packages or libraries were used for the model. These packages included pandas and NumPy to import the data and be able to manipulate the data as a Data frame and then convert it to a 2d array. Seaborn and Matplotlib were used for visualization purposes. Sklearn was the library where the majority of packages were imported. These consisted of preprocessing,

Neighbors, cross\_validate, MinMaxScaler, train\_test\_split, KNeighborsRegressor, mean\_squared\_error, GridSearchCV, and KNeighborsClassifier. These packages were used to prepare the data, instantiating the model, fitting the model, and evaluating the model. Lastly, sqrt was used imported from the math library to determine the Euclidean distance from the data points.

Part III

One data preprocessing goal was to scale the feature variables. This step was crucial in preparing the data set as two of the variables consisted of extremely high dollar amounts. Comparing these variables to the categorical variables which consisted of 1’s and 0’s would have greatly affected our model.

Of the 50 variables from the original model only 5 were used to determine Readmitted. These 5 variables were Initial days, Total Charge, Services CT Scan, Children, Initial\_admin\_Emergency Admission. Initial days is a continuous variable that tells us how many days the patient was initially admitted for. TotalCharge was a continuous variable that notified us of the total amount of the Hospital Stay. Services CT Scan was a categorical variable that informed us whether a person received a CT scan or not. Children was a continuous variable that notified us how many children a patient had. Initial admin Emergency Admission was a categorical variable that notified us if the patient was admitted through the Emergency Room as oppose to an observation checkup or Elective admission.

The steps taken to prepare the data for analysis were as follows. “df.isnull().sum()” Was used to identify any null values. “g = sns.pairplot(vals, hue = 'ReAdmis')” and “ax = sns.countplot(x=f, data = df, hue = 'Readmitted\_Yes')” were used to create plots to better understand the data and the distribution. A correlation matrix was then used to determine which variables would be most effective in creating the model. “correlation\_mat = df.corr()

corr\_pairs = correlation\_mat.unstack()

print(correlation\_matrix["Readmitted\_Yes"].sort\_values(kind="quicksort"))” was helpful in making sure the correlation amounts were showing in descending order.

Part IV:

The analysis technique best used to analyze the data was to calculate the confusion matrix for the model. Once the confusion matrix showed that the responses were being predicted correctly, a classification report was used to determine the models precision, recall and f1-score.

Attached are the outcomes for the model. The models shows a accuracy of 97%. The code used to generate the classification report was: print(classification\_report(y\_test,y\_pred))

|  |
| --- |
| precision recall f1-score support |
|  |
| 0 0.97 0.98 0.98 1224 |
| 1 0.97 0.96 0.96 776 |
|  |
| accuracy 0.97 2000 |
| macro avg 0.97 0.97 0.97 2000 |
| weighted avg 0.97 0.97 0.97 2000 |

Part V:

The accuracy of the model came out 97% which is substantially high. The AUC Score of the model came out to 98.89% which is also very high. These results are substantially high and could possibly lead to one of the drawbacks of the data. Since the dataset provided is considered “Toy” Data and is not from the real world. Then the high accuracy and high AUC Score can possibly be due to the fact that the data is intended to display these outcomes.

A recommended course of action would be to possibly reach out to local health institutions and advise that there is a possibility or opportunity to reduce Readmissions into the hospitals by investigating key variables.

# Bibliography

Varghese, D. (2018, 12 06). *Comparative Study on Classic Machine learning Algorithms*. Retrieved from towardsdatascience.com: https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222#:~:text=A%20general%20difference%20between%20KNN,whereas%20KNN%20is%20non%2Dparametric.