D209 Decision Tree Task 1

The purpose of this Data Classification Report is to be able to predict whether a person would be Readmitted into a Hospital. The model used is a Decision Tree model. The model is ideal since the data provided is labeled data and the outcome that we are looking to identify is a classification outcome such as a yes or no. If the data been unlabeled then we would have had to use an alternative such as an unsupervised model. The way the model works is by going through each iteration of the algorithm, it iterates through every unused attribute of the set and calculates the entropy of that attribute. It then selects the attribute which has the smallest entropy value. The set is then split by the selected attribute to produce subsets of the data. The algorithm continues to recurse on each subset, considering only attributes never selected before. The goal is to identify key features that can help determine whether a patient is going to be Readmitted. Once those features are identified, measures can be taken to ensure a patient is well educated in their health before they are released from the hospital.

The classification method looks at 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 Decision Tree algorithm is that Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model (Rawale, 2018). Steps were taken to ensure that most of the feature variables were categorical. However, some variables are still discreet and steps were taken to ensure that they did not affect the model in a negative way. This was done by standardizing the variables.

Ten 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 tree, DecisionTreeClassifier, accuracy\_score, confusion\_matrix, r2\_score, plot\_tree, plot\_confusion\_matrix, confusion\_matrix. These packages were used to prepare the data, instantiating the model, fitting the model, and evaluating the model.

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. Scaling the variables consisted of determining the highest value in those columns and then dividing each value by the highest amount. This allows the ranges of the particular column to have a minimum value of 0 and the highest value to be 1. All other values are a decimal number below 1. This is extremely important in ensuring all features could be properly computed by the model in balanced way. Ensuring that the features that were scaled had values below 1 and the categorical features had values of 1’s and 0’s would ensure our Decision Tree model would be balanced.

Of the 50 variables from the original model only 9 were used to determine Readmitted. These 9 variables were Initial days, Total Charge, Age, Children, Initial\_admin\_Emergency Admission, Asthma\_No, Services\_CT Scan, State\_TX, State\_IL, State\_SC. Initial days is a continuous variable that tells us how many days the patient was admitted for. TotalCharge was a continuous variable that told us the total amount of the Hospital Stay. Age informed us the age of the patient. Children informed us of the number of children a patient had. Services CT Scan was a categorical variable that informed us whether a person received a CT scan or not. 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. Asthma\_No was a variable that told us if the person had asthma. State\_TX, State\_IL, and State\_SC is a variable that let us know if the patient was in one of those States.

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 charts to better understand the distribution of the data. 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.

The analysis technique used to evaluate the model was to calculate the confusion matrix for the model. Once the confusion matrix showed that the responses were being predicted correctly, a report was used to determine the models precision, recall and accuracy.

Attached are the outcomes for the model. The model shows it has an accuracy of 98%. The code used to generate the classification report was: cm = confusion\_matrix(y\_train, y\_pred).

|  |
| --- |
| The True Positive rate / Recall per class is: [0.980079 0.973379] |
| The Precision per class is: [0.984545 0.965797] |
| The False Alarm rate per class is: [0.026621 0.019921] |
| The Miss Rate rate per class is: [0.019921 0.026621] |
| The Classification error of each class is [0.022375 0.022375] |
| The Accuracy of each class is [0.977625 0.977625] |
|  |
| The average Recall is: 0.9767288675269773 |
| The average Precision is: 0.9751713842471355 |
| The average False Alarm is: 0.023271132473022732 |
| The average Miss Rate rate is: 0.023271132473022732 |
| The average Classification error is 0.022375 |
| The average Accuracy is 0.977625 |

This was the code used to generate that output:

Text

Description automatically generated

The accuracy of the model came out 98% which is very high. This means that from the predicted values, the model was able to identify the accurate outcome 98% of the time. The RMSE was not calculated for this model since the model used was a classification tree and not a regressor tree. Since we used a classification tree, the accuracy and precision are used to evaluate the model. Both the precision and accuracy were 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 real-world data, then the high accuracy and high precision can possibly be due to the fact that the data is intended to display these positive 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. If we are able to implement the model, then we would be able to reduce readmission for the hospitals which could lead to more or better quality of care for patients who are visiting for the first time. In doing so, the hospital would be able to distribute their resources and have a better opportunity for growth and expansion in different areas.