**Part I: Research Question**

A. Describe the purpose of this data mining report by doing the following:

1. Propose **one** question relevant to a real-world organizational situation that you will answer using **one** of the following classification methods:

• *k*-nearest neighbor (KNN)

• Naive Bayes

For my research question, I plan on trying to predict a patient’s readmission status based on multiple independent variables. Using k-nearest neighbors, I plan to construct an analysis that can benefit stakeholders in determining if they can predict future patient status. This will help to highlight what factors may influence the readmission of patients at a later timeline. Using the predictions, stakeholders can begin to educate patients that fall into the categories that influence readmission status.

2. Define **one** goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

The goal of my analysis is to present to stakeholders the ability to take patient information and determine that patient’s readmission potential. Using the available information and python, I will construct a model that can synthesize the records and predict the outcome with high reliability. These scores will help stakeholders in the final acceptance of the project.

**Part II: Method Justification**

B. Explain the reasons for your chosen classification method from part A1 by doing the following:

1. Explain how the classification method you chose analyzes the selected data set. Include expected outcomes.

The classification method that I used will analyze the data and determine what the expected outcome of the next data point should be. I set my classifier to analyze the nearest 6 data points. By doing this I hope to be able to predict with high reliability that the next data point will be truthful. By taking the information found in the records and creating a test and training set, I can train the model with the data, and use the neighbor’s method to generate the next outcome in regards to a patient’s readmission potential.

2. Summarize **one** assumption of the chosen classification method.

Choosing k-nearest neighbors, the one main assumption that ties in with this method are that things close together have a lot in common (Grant, 2019). By using this method, we can pick out groups of data that surround unlabeled data points and attempt to label them based on nearby data. Using this method, we can assume that data points that surround our unlabeled data will share characteristics with the unlabeled data. Using the predict method, we can then attempt to label that data point based on its neighbors.

3. List the packages or libraries you have chosen for Python or R, and justify how *each* item on the list supports the analysis.

Packages that were used in the processing of the data included; pandas which allowed me to create data frames of the data and perform processing. Numpy, to deal with numbers and equations within the data. Matplotlib, to allow for graphics and visual representation of the data. K-neighbors classifier, allowed me to perform the labeling of the unlabeled data and was the backbone of the project. After the first prediction, I used test/train split to further process the data and perform a split that allowed me to train the model. I imported logistic regression to perform regression of the split data and used roc score to perform a score analysis of the final model and finally, I imported roc curve to graph the final model.

**Part III: Data Preparation**

C. Perform data preparation for the chosen data set by doing the following:

1. Describe **one** data preprocessing goal relevant to the classification method from part A1.

After importing the data, I will print out the variables and determine what columns can be dropped. By doing this I can easily drop variables that are not needed in the analysis, columns such as customer ID or interaction ID will not fit into the classification model and can be dropped. This saves processing time and allows for a slimmer model in the end.

2. Identify the initial data set variables that you will use to perform the analysis for the classification question from part A1, and classify *each* variable as continuous or categorical.

For my final analysis, the variables I selected were:

|  |  |
| --- | --- |
| Case Order | Continuous |
| Population | Continuous |
| Children | Continuous |
| Age | Continuous |
| Income | Continuous |
| ReAdmis | Categorical |
| VitD\_levels | Continuous |
| Doc\_visits | Continuous |
| Full\_meals\_eaten | Continuous |
| vitD\_supp | Continuous |
| Initial\_days | Continuous |
| TotalCharge | Continuous |
| Additional\_charges | Continuous |

3. Explain *each* of the steps used to prepare the data for the analysis. Identify the code segment for *each* step.

Included in video and file

4. Provide a copy of the cleaned data set.

File Attached

**Part IV: Analysis**

D. Perform the data analysis and report on the results by doing the following:

1. Split the data into training and test data sets and provide the file(s).

Attached to submission

2. Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.

To analyze the data after performing the predictions I used the scoring method to print out the score of the data that went into determining the prediction. First I performed a test of the data without any split and saved the score to compare it to after splitting the data. After splitting the data, I can then compare my new score to the old score. This will help to determine if the model is performing as needed, or if the research question needs to be changed.

3. Provide the code used to perform the classification analysis from part D2.

Attached to submission

**Part V: Data Summary and Implications**

E. Summarize your data analysis by doing the following:

1. Explain the accuracy and the area under the curve (AUC) of your classification model.

The accuracy of my final classification model was close to 1 at 0.99. This points to the model working as it should and predicting the outcome of patient readmissions. The graph of my score confirms this accuracy, as more of the model is placed under the curve than above it. This validates the choice and analysis that my model has done, and better presents a confidence level to stakeholders.

2. Discuss the results and implications of your classification analysis.

After using the k-nearest neighbor’s classification on my data, I was able to classify and predict with a certain confidence level the readmission potential for patients. By using this method, and fine-tuning it based on the research question, I can better assist stakeholders and others that have a vested interest in determining whether patients will be readmitted based on various factors in their lives.

3. Discuss **one** limitation of your data analysis.

One limitation that I can see with the data, is that predictions can always be wrong. If I am to predict the possibility of a patient being readmitted, the data and the analysis must be thorough and accurate. Therefore, there must be some implicit trust between those that are gathering the data, and those that are processing and modeling the data.

4. Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.

Based on my analysis of the provided data, I believe stakeholders must start reaching out to members of the population that fall into categories that are predisposed to be readmitted to the hospital. By providing early education and awareness, patients can prevent illnesses or maladies that will contribute negatively to their overall health. In states or income levels, patients that are highlighted in the analysis can be given a higher focus to prevent them from falling victim to health-related problems.

## References

Grant, P. (2019). *Introducing k-Nearest Neighbors*. Towards Data Science. Retrieved 18 May 2021, from https://towardsdatascience.com/introducing-k-nearest-neighbors-7bcd10f938c5#:~:text=The%20k%2Dnearest%20neighbors%20algorithm%20is%20a%20common%20classification%20model,between%20the%20two%20data%20points.