NVM2 — NVM2 TASK 1: CLASSIFICATION ANALYSIS

Jason Smith  
Student ID: 293963  
[jsmi268@wgu.edu](mailto:jsmi268@wgu.edu)  
WGU – MSDA  
D209- Data Mining Task 1 - Hospital Data  
Program Mentor: Christiana Okhipo

**Part I: Research Question**

A1. 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) is a classification algorithm that will be used in this study. KNN makes predictions based on similar groupings of data points. This study will attempt to answer the following question: Can hospital readmissions be predicted with an acceptable degree of accuracy? K-nearest neighbor (KNN) is a classification algorithm that will be used for this objective.

A2.  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 project stakeholders require the ability to predict which patients are most likely to be readmitted into the hospital and any correlations that may offer clues that lead to patient readmittance.

These patients can then be targeted for additional information and more attention to potentially avoid costly readmission.

**Part II: Method Justification**

B1. Explain how the classification method you chose analyzes the selected data set. Include expected outcomes.

The K-nearest neighbor (KNN) algorithm classifies data points based on similarities of multiple variables to enable its predictive ability. Data points that are in a close proximity of one another are considered most similar. KNN accepts an input parameter, called “K”, which represents the classification size or the neighboring data point to be compared against. I expect the higher “K” values to yield better results with less noise.

B2. Summarize **one** assumption of the chosen classification method.

KNN models assumes that points in close proximity to one another are quite similar while points that are not close to one another are dissimilar (Nelson, P.1). The distance between points can be measured multiple ways. For example, the Euclidean method measures the distance between two points by using a straight line. Manhattan and Minkowski are two additional measurement techniques that are not described in this study.

B3. List the packages or libraries you have chosen for Python and justify how *each* item on the list supports the analysis.

Graphical user interface, text, application, email

Description automatically generated

**Part III: Data Preparation**

C1. Describe **one** data preprocessing goal relevant to the classification method from part A1. Describe **one** data preprocessing goal relevant to the classification method from part A1.

KNN requires text variables to contain numeric data. The target variable and some of the predictor variables contain text data. These variables will need to be encoded to numeric values prior to using this data classification technique.

C2.  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

|  |  |
| --- | --- |
| **Variable** | **Classification** |
| Initial\_days | Continuous |
| Population | Continuous |
| Treatment | Categorical |
| Doc\_visits | Categorical |
| VitD\_levels | Continuous |
| ReAdmis | Categorical |

C3.  Explain *each* of the steps used to prepare the data for the analysis. Identify the code segment for *each* step.

* Identify number of rows and columns in the data set

Graphical user interface, application

Description automatically generated

* Describe univariate statistics

A picture containing table

Description automatically generated

* Rename mis-spelled column



* Create subset of the data

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

* Identify data types and whether Null values exist

Table

Description automatically generated

* Encode Categorical data values

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

* Create a cross tabulation of discrete target and features variables

Graphical user interface, text, application, email

Description automatically generated

* Create data frame containing discrete variables to be used with univariate visualizations

A picture containing text

Description automatically generated

* Create a countplot of discrete target and feature variables

A picture containing shape

Description automatically generated

A picture containing bar chart

Description automatically generated

AxesSubplot(0.125,0.125;0.775x0.755) None

--------------------------------

Chart, bar chart

Description automatically generated

AxesSubplot(0.125,0.125;0.775x0.755) None

--------------------------------

Chart, bar chart

Description automatically generated

AxesSubplot(0.125,0.125;0.775x0.755) None

--------------------------------

Chart, bar chart

Description automatically generated

AxesSubplot(0.125,0.125;0.775x0.755) None

* Create boxplots of discrete target and features

A picture containing graphical user interface

Description automatically generated

* Create a correlation table of target and feature variables

Table

Description automatically generated

* Create a heatmap of discrete target and features

Graphical user interface

Description automatically generated with medium confidence

* Variables to be used with Datatypes

Graphical user interface, text

Description automatically generated

* Check for null values



Soft\_drink 0

HighBlood 0

Stroke 0

Arthritis 0

Diabetes 0

Hyperlipidemia 0

BackPain 0

Allergic\_rhinitis 0

Reflux\_esophagitis 0

Asthma 0

Initial\_admin 0

Complication\_risk 0

Services 0

Courteous 0

Initial\_days 0

Doc\_visits 0

Population 0

Treatment 0

Hours 0

Admissions 0

VitD\_levels 0

ReAdmis 0

C4.  Provide a copy of the cleaned data set.

See attached file: nvm2.csv

**Part IV: Analysis**

D1.  Split the data into training and test data sets and provide the file(s).

The “train\_test\_split” package from SKLearn was used to split the data. Seventy-five percent of the data was utilized for train and the remainder (25%) went to test. Refer to attached files: *xtrain.csv, xtest.csv, ytrain.csv, ytest.csv*

D2.  Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.

I used the *accuracy\_score* and *classification\_report* functions of SKLearn.metrics to determine accuracy and precision after each run of KNN. See below:

Graphical user interface

Description automatically generated with medium confidence

I also used the *Confusion Matrix* function (included in SKLearn.metrics) (Tarek Atwan, p.1) to determine the true positives and true negatives. See below:

Chart, treemap chart

Description automatically generated

KNN requires an input variable called the “K” value. I made an initial best guess of 5 as to the “K” value and received an accuracy value of 81%. I then charted the error curves two different ways(Tarek Atwan, p.1). These charts provide clues to determine an ideal “K” value. Screen shots of both error charts follow:

Chart, line chart

Description automatically generated

Chart

Description automatically generated with medium confidence

Based on the two charts, I concluded that a “K-value” of 15 is the minimum “K-value” to yield a high accuracy value of .8364(84%).

D3.  Provide the code used to perform the classification analysis from part D2.

* Scaling the input data



* Scaling training and testing data set



* Run KNN with an initial guess at the K value (5) and an ideal K-value (15)

Text, letter

Description automatically generated

* Determine the KNN accuracy based on the current K-value

Table

Description automatically generated

* Use error curves to choose an ideal K-value (Tarek Atwan, P.1)

Text

Description automatically generated

Chart, line chart

Description automatically generated

* Determine True Positive, True Negative, False Positive, and False Negative using Confusion Matrix (Atwan, p.1).

Text

Description automatically generated

Chart, treemap chart

Description automatically generated

**Part V: Data Summary and Implications**

E.  Summarize your data analysis by doing the following:

E1.  Explain the accuracy and the area under the curve (AUC) of your classification model.

The model accuracy is 84% (0.836). The precision of predicting a patient being readmitted is 93%. The precision of predicting a patient not being readmitted in 72%. The accuracy and precision score values indicate the model’s strong ability to prediction abilities.

E2.  Discuss the results and implications of your classification analysis.

K Nearest Neighbors was used to classify and predict which patients were likely to be readmitted. Features having the strongest correlation were selected. These features include: 'Initial\_days', 'Population','Treatment','Doc\_visits', 'VitD\_levels'. An optimal K-valuewas chosen based on charting out the error rate vs. K-values. The relatively high accuracy and precision of this model means hospital executives can predict which patients are likely to be readmitted into the hospital based on the above features.

E3.  Discuss **one** limitation of your data analysis.

KNN is quite sensitive to the scale of a dataset and can be thrown off quite easily by irrelevant features (Daniel Nelson, p. 1). This model can be used as a focal point to target patients that possibly should be further interviewed and monitored for hospital readmittance. There is no guarantee that patients that are identified as likely to be readmitted will be readmitted.

E4.  Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.

Hospital executives and personnel can be trained to identify and assist those patients likely to be readmitted and / or identify those patients that are not likely to fully recover from their ailments. Hospital personnel can start to look for common patterns that lead to readmittance. Reducing readmittance, not only leads to hospital cost reduction, but also leads fewer patients checking in to the hospital, reduced suffering of the patients, and lower workloads and stress levels of the hospital staff.

**Part VI: Demonstration**

F.  Panopto video link:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=ea35f010-fbf9-47d5-b06a-ae5900033a7d

G.  **Sources for Third Party Code**

1. Tarek Atwan(2017, P.1). How To Plot A Confusion Matrix In Python.  
   How To Plot A Confusion Matrix In Python – Tarek Atwan – Notes on Artificial Intelligence. <https://tatwan.github.io/How-To-Plot-A-Confusion-Matrix-In-Python>
2. Bashir Alam(2022, P.1). Implementing KNN Algorithm using Python. <https://hands-on.cloud/implementing-knn-algorithm-using-python>

H.  **Sources**

1. Daniel Nelson. (2020, (P. 1)). What is a KNN (K-Nearest Neighbors)?  
   <https://www.unite.ai/what-is-k-nearest-neighbors/#:~:text=The%20main%20limitation%20when%20using%20KNN%20is%20that,algorithm%2C%20the%20proper%20value%20for%20K%20is%20chosen>.