**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

Using K-nearest neighbor algorithm can we find which customers are at high risk of Churn from our dataset?

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.

Stakeholders will benefit from this analysis by being able to understand and identify customers that are churning. Furthermore, stakeholders will then be able to come up with a plan, marketing, or changing initial service that will better benefit the customer and reduce churn rates.

**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.

K-nearest neighbor algorithm analyzes the data by working on the distance between data points. Also, it is based on the premise that similar data points are close to each other. Basically, the neighbors are categorized in a certain way, and will be placed in the same category based on the virtue of them being close to each other. Furthermore, the further away neighbors are from each other will most likely fall in a different classification category.

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

K-nearest neighbor doesn’t make assumption since it is not parametric algorithm. KNN doesn’t care about the assumption for data distribution, but a general assumption that can be made is that data points that exist close to each other are highly similar.

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 and libraries used for Python:

1. Numpy
2. Pandas
3. Seaborn
4. Matplotlib
5. Sklearn

Numpy, Pandas, Matplotlib are the most basic imports used in data science projects and provide methods that allow for visualizations, reading, and saves for the dataset. Also, Seaborn was used for more enhanced visualization in graphs, plots and matrices. Finally, Sklearn helped split, train, predict, handle different algorithms “KNeighborClassifier”, reports, and metrics for the data.

**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.

One data preprocessing goal was transforming the binary variables from “Yes, No” to “1’s, 0’s” to have all numerical values to work with.

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.

Continuous Variables:

* Children
* Income
* Outage\_sec\_perweek
* Email
* Contacts
* Yearly\_equip\_failure
* Tenure
* MonthlyCharge
* Bandwidth\_GB\_Year
* Age

Categorical Variables Nominal and Ordinal:

* Contract
* Techie
* Phone
* Tablet
* Port\_modem
* InternetService
* Multiple
* OnlineSecurity
* OnlineBackup
* DeviceProtection
* TechSupport
* StreamingTV
* StreamingMovies
* PaperlessBilling
* PaymentMethod
* Churn
* Timely response
* Timely fixes
* Reliability
* Options
* Respectful response
* Courteous exchange
* Evidence of active listening

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

1. Import packages and libraries.
2. Read dataset into a data frame.
3. View columns, drop meaningless columns, and update survey questions to appropriate names.
4. Check if any null or empty columns.
5. Check for any misspellings in the data.
6. Check if all previous modifications to the data took effect ex: survey questions have proper names.
7. Get statistics on the data.
8. Get unique values for columns.
9. Save a clean copy of the data set.
10. Backup to google drive.

Text

Description automatically generatedText, letter

Description automatically generatedText

Description automatically generatedText, letter

Description automatically generatedTable

Description automatically generatedGraphical user interface, application

Description automatically generatedTable

Description automatically generated

Chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedDiagram

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedTable

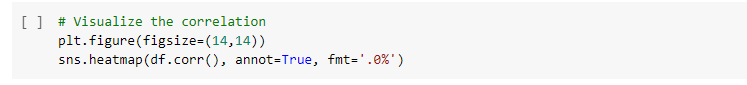
Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidenceGraphical user interface, text, application

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidenceTable

Description automatically generated with medium confidence

Chart

Description automatically generated

4.  Provide a copy of the cleaned data set.

Copy provided in folder-“ final\_prepared\_data\_csv”.

Text

Description automatically generated with medium confidence

**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).

Graphical user interface, application, Word

Description automatically generatedText

Description automatically generatedGraphical user interface, text, application

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidenceTable

Description automatically generated with medium confidenceA picture containing calendar

Description automatically generatedGraphical user interface, text, application, chat or text message

Description automatically generated

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

To begin my analysis, I wanted to see if we were to guess “no” all the time in a customer not to churn, how accurate would our model be. I achieved this by using a basic formula of:

Customer count no – customer count yes / customer count no. This gave a basis of a prediction outcome, where we want our model to do better than this initial guess.

Text

Description automatically generated

Below are screen shots that shows accuracy scores for using KNN model to increase the accuracy and give better predictions.

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

Table

Description automatically generated with low confidenceGraphical user interface

Description automatically generated with medium confidenceChart, line chart

Description automatically generatedText

Description automatically generatedTable

Description automatically generated

**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 scaling improved the model accuracy from the original formula used shown above of around 64% “Calculation if we guess”, to the initial model of an accuracy of 72.5%, and finally to final model accuracy of 74.5%. Furthermore, the AUC also increased from an initial model of 0.744 to final model of 0.778.

Pictures of scores shown above slides.

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

The results of the classification analysis were an increase in accuracy, AUC score, precision and f1 score. The accuracy of the initial model of 0.725, precision of “0” 0.80, precision of “1“0.48, f1-score “0” 0.82, f1-score “1” 0.45, and ACU 0.744 to accuracy of the tuned parameters model of 0.745333333…, precision of “0” 0.81, precision of “1“0.53, f1-score “0” 0.86, f1-score “1” 0.47, and ACU 0.778. Furthermore, this was all done by implementing GridSearchCV for parameter tuning, which runs through all the different parameters and gives the best combination of parameters.

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

Limitations occurred with the KNN algorithm, when tuning the parameters, we must decrease the size of the range because it takes a lot of time to run and heavy on memory. Furthermore, KNN is delicate to the scale of data, so for higher scale, the calculations for distances that are very high can produce a mediocre/poor result.

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.

Our prediction is not very high at 74.5% but was increased from original human guess of 64%. Also, the use of KNN makes it hard to figure feature importance variables, so we might want to use a different algorithm that can help determine which features are more important than others. So, the course of action should start with looking for features that have a high customer churn rate. For example, port\_modem and device\_protection have high churn rates, improving those services can help with customer churn.

**Part VI: Demonstration**

F.  Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.

Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.

Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.

To submit your recording, upload it to the Panopto drop box titled “Data Mining I – NVM2.” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.

G.  Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

None used

H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

None used

I.  Demonstrate professional communication in the content and presentation of your submission.