### 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 neighbour (KNN)

A. I will use the KNN and the following predictors to identify which customers are at high risk of churn?

• Children

• Income

• Tenure

• Bandwidth\_GB\_Year

• Age

1. 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.

A. The stakeholders can review the data provided by the analysis and create incentives to keep the customers that are likely to terminate their contracts with the company. This will lead to a lower churn rate.

### 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.
   1. K-Nearest Neighbours (KNN) is one of the simplest algorithms for classifying data items. It classifies new data points based on similarity measures i.e., distance function. The data items are classified by a majority vote to its neighbours.

Expected outcomes: The data will be classified based on their distance to predict their 'Churn' status 'Y' or 'N'.

1. Summarize one assumption of the chosen classification method.
   1. The classifier assumes that similar things are near to each other; that is, they are nearby. Chatterjee, M. (2021, April 19)
2. List the packages or libraries you have chosen for Python or R and justify how each item on the list supports the analysis.

A. I will utilize Python due to my previous interaction with it and its Pandas, matplotlib and Scipy modules. Additionally, I will be using Jupyter notebook as the IDE because it provides a user-friendly experience. Pandas is an excellent package for working with data set as it makes it easy to load and manipulate columns and/or rows to replace null values. Matplotlib plot is an easy way to create graphs for identifying outliers using histograms and/or boxplots.

### Part III: Data Preparation

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

1. Describe one data pre-processing goal relevant to the classification method from part A1.
   1. I will convert all the categorical data to binary to perform the classification analysis.
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.

##### Categorical Predictor:

• Churn

##### Continuous Predictor:

• Children • Income • Tenure • Bandwidth\_GB\_Year • Age

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

1. Import dataset to a DataFrame in Pandas.

2. Review the summary statistics of DataFrame, structure & data types.

3. Remove irrelevant columns from the DataFrame "Customer\_id", "zip code".

4. Split the training and test data

5. Load and fit the KNN classifier

6. Review the KNN training score

7. Find ideal testing and training accuracy

8. Provide AUC score

9. provide a copy of the prepared dataset "prepared\_dataset.csv" for use in the KNN model

### Part IV: Analysis

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

1. Text

   Description automatically generatedSplit the data into training and test data sets and provide the file(s).
2. Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.
   1. I used the KNN with n\_neighbors parameters set to 6 to perform the analysis, then I printed out the shape of the training data and then stored the prediction in the y\_pred. Lastly, print out the training accuracy of the model by using the score function.

Here is an example, the ‘green dot’ should be classified as either blue squares or red triangles. If k = 3 ‘solid line circle’, it is assigned to the red triangles because there are 2 triangles and only 1 square inside the inner circle. If k = 5 ‘dashed line circle’, it is assigned to the blue squares (3 squares vs. 2 triangles inside the outer circle).

A picture containing diagram

Description automatically generated

1. Graphical user interface, text, application, email, Teams

   Description automatically generatedProvide the code used to perform the classification analysis from part D2.

### Graphical user interface, application Description automatically generated

Text

Description automatically generated with medium confidence

### 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.

Graphical user interface, text, application, Word

Description automatically generated

Discuss the results and implications of your classification analysis.

1. The AUC score informs us how well is our model at predicting the correct label for 'Churn’. The value is scaled from 0 to 1 being the most accurate. This KNN model has a score of '0.8929' which means the model is adequate at predicting the Churn rate but could perform better if further analysis is performed on the independent variables.
2. Discuss one limitation of your data analysis.
3. Starting with an arbitrary choice of k = 6 nearest neighbours will yield dramatically different results. As shown above that a consideration must be observed and a systematic approach of trial and test must be tried to increase the accuracy of the prediction. I have also noticed that limiting the number of the independent variable to a select few can improve the performance considerably and reduce the memory and computational load which could lead to slow performance.
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.
5. The predictor variables create a low accuracy score of ‘0.89’. The stakeholders need to further invest in market analysis to identify competitive services and the marketing team needs to come up with products and services that will create an eco-system for the customer which will decrease the churn rate.

### Part VI: Demonstration

F. Panapto: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b83de77d-6f5b-4c82-afd3-ae1501858242>

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.

Pandas. (2021). Pandas DataFrames. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.dtypes.html

Get started with references. (2021). Jupyterbook. https://jupyterbook.org/tutorials/references.html#tutorials-references

Marques, A. M. (2020, March 11). How to show all columns / rows of a Pandas Dataframe? Towards Data Science. https://towardsdatascience.com/how-to-show-all-columns-rows-of-a-pandas-dataframe-c49d4507fcf

Starmer, J. (2018, March 5). StatQuest: Logistic Regression. YouTube. https://www.youtube.com/watch?v=yIYKR4sgzI8&t=121s

V. (2019, July 21). Pandas: Apply a function to single or selected columns or rows in Dataframe. ThisPointer. https://thispointer.com/pandas-apply-a-function-to-single-or-selected-columns-or-rows-in-dataframe/

Wijaya, C. Y. (2021, December 15). 5 Must-Know Dimensionality Reduction Techniques via Prince. Medium. https://towardsdatascience.com/5-must-know-dimensionality-reduction-techniques-via-prince-e6ffb27e55d1

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

Chantal D. Larose, & Daniel T. Larose. (2019). Data Science Using Python and R. Wiley.

Chatterjee, M. (2021, April 19). A Quick Introduction to KNN Algorithm. GreatLearning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/knn-algorithm-introduction/