**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 prediction methods:

• **decision trees**

• random forests

• advanced regression (i.e., lasso or ridge regression)  
I will use the decision tree prediction method to help predict which customers are more likely to churn away and answer the question, what variables influence customer churn the greatest?

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 goals of the data analysis are to accurately predict which customers are most likely to churn away and determine the commonalities between them. This will allow us to focus on the variables that influence customer churn the greatest so that we can try to reduce it.

**Part II: Method Justification**

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

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

Decision trees work by splitting data into branches based on various feature values. This ends up leading to a tree-like structure where each branch represents a decision based on a feature. What I expect to happen is that the decision tree will group people who will churn away based on common features like monthly charge and tenure.

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

One assumption of decision trees are that each feature independently contributes to the final decision as far as what class something falls into. Interactions or dependencies between features are not accounted for.

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

Pandas: used for data manipulation and analysis

Numpy: used for numerical computations and handling large arrays or matrices

Sklearn: used for data preprocessing, model training, evaluation, standardization of numerical variables (StandardScaler), and is where our DecisionTreeClassifier is located

Seaborn: used to create various, static, informative, etc, visualizations

Scipy: used for advanced mathematical and numerical operations

Pylab: graphs and plots

Statsmodels: for more statistical based models

**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 prediction method from part A1.

Splitting the groups into training and test sets. The training set is used to teach the machine learning model and the test set is kept separate to evaluate the models performance.

2. Identify the initial data set variables that you will use to perform the analysis for the prediction question from part A1 and group *each* variable as numeric or categorical.

DummyChurn - Categorical

Children - Numeric

Bandwidth\_GB\_Year - Numeric

Age - Numeric

Tenure - Numeric

MonthlyCharge - Numeric

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

\*The code is shown in the attached notebook\*

First I performed some data cleaning processes (renaming columns, converting categorical variables, checking for null values, dropping unnecessary columns) to ensure that the data is free from errors, inconsistencies, and missing values. I also had to convert my Churn variable into numbers (1 = yes and 2 = no) since the words yes and no arent going to work in a model. Throughout this process I used the head function to make sure things were looking correct.

4. Provide a copy of the cleaned data set.  
\*Code shown in attached notebook\*

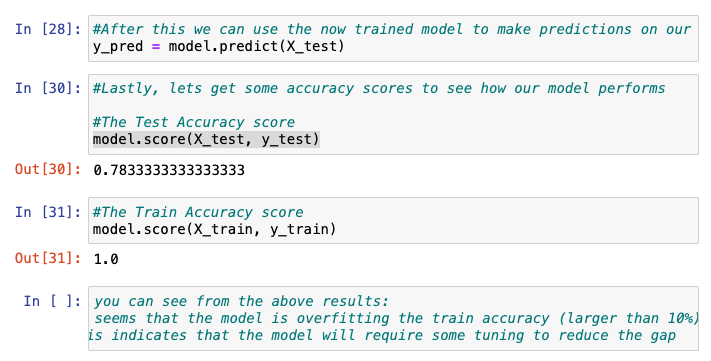
**Part IV: Analysis**

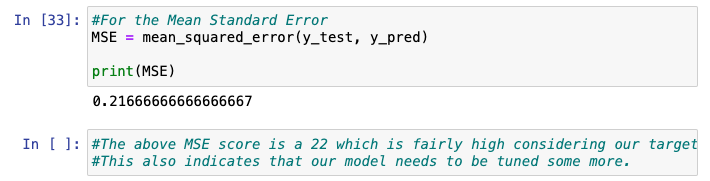
D. Perform the data analysis and report on the results by doing the following:

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

\*Code shown in attached notebook\*

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





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

\*Code shown in attached notebook\*

**Part V: Data Summary and Implications**

E. Summarize your data analysis by doing the following:

1. Explain the accuracy and the mean squared error (MSE) of your prediction model.

I calculated a test accuracy score of .79 and a train accuracy score of 1. This indicates that our model is overfitting and needs to undergo some tuning to reduce that gap. The Mean Standard Error that I got is .22 which is also fairly high considering our target variable only goes from 0 to 1. This further supports that our model needs to undergo some tuning.

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

Based on the accuracy scores discussed above, the model must be tuned some in order to give accurate predictions that we can be confident in.

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

One limitation of my data analysis using decision trees is the tendency to overfit. Apparently this is something that is very common when using decision trees and has to be remedied by tuning certain parameters like maximum depth and minimum numbers of samples required to cause a branch to split.

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 results, I would advise a company depending on these results to wait while we improve our model so that when we do use it we can be sure we are making a decision we can be confident in, as mistakes can be costly in the business world.

**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 - NVMx | D209 (student creators) [assignments*](https://wgu.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx#folderQuery=%22d209%22&folderID=%22dcfdf6dd-9e05-4665-b1b6-ac72018a20ab%22)*]." 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. Acknowledge web sources, using in-text citations and references, for segments of third-party code or data used to support the analysis. Be sure the web sources are reliable.

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

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