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)**

   How long will a customer stay with the company based on other factors?

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.**

   My goal is to build a random forest classification model that predicts the tenure of a customer based on other variables in an observation.

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.**

   A random forest algorithm classifier is a combination of the bagging method and feature randomness. It creates an unrelated 'forest' of decision trees. Feature randomness creates a random subset of features, ensuring the decision trees are not highly related. Decision trees look at all the feature splits possible, random forests select a sample of those splits (IMB Cloud Education, 2020). The random forest regressor is expected to output continuous predicted values of the target variable.

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

   A random forest is a non-parametric model, meaning that the data distribution can't be defined by finite parameters. Therefore, a random forest model assumes no specific distribution within the data set. It does however rely on the samples being reflective of the larger data set (Vishalmendekarhere, 2021).

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 to import/export and clean the data
   - sklearn.model_selection: train_test_split and GridSearchCV to split the data into training and testing data and to tune the model's hyperparameters.
   - sklearn.ensemble: RandomForestClassifierto create the regression model
   - sklearn.metrics: r2_score and mean_squared_error to evaluate the 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 prediction method from part A1.**

   Change all object type data to numeric data so that it can be passed into the random forest model.

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 continuous or categorical.**
   - CaseOrder: continuous
   - State: categorical
   - Zip: continuous
   - Lat: continuous
   - Lng: continuous
   - Population: continuous
   - Area: categorical
   - TimeZone: categorical
   - Children: continuous
   - Age: continuous
   - Income: continuous
   - Marital: categorical
   - Gender: categorical
   - Outage_sec_perweek: continuous
   - Email: continuous
   - Contacts: continuous
   - Yearly_equip_failure: continuous
   - Contract: categorical
   - InternetService: categorical
   - PaymentMethod: categorical
   - Tenure: continuous
   - MonthlyCharge: continuous
   - Bandwidth_GB_Year: continuous
   - Item1-Item8: categorical
   - Churn_Yes: categorical
   - Techie_yes: categorical
   - Port_modem_Yes: categorical
   - Tablet_Yes: categorical
   - Phone_Yes: categorical
   - Multiple_Yes: categorical
   - OnlineSecurity_Yes:categorical
   - OnlineBackup_Yes: categorical
   - DeviceProtection_Yes: categorical
   - TechSupport_Yes: categorical

- StreamingTV_Yes: categorical
- StreamingMovies_Yes: categorical
- PaperlessBilling_Yes: categorical

3. **Explain the steps used to prepare the data for the analysis. Identify the code segment for *each* step.**
   I will first import the original dataset and all necessary libraries. I will drop the variables Customer_id, Interaction, UID, and Job because they are not necessary from a business perspective. I will drop City and County because they are redundant with Zip. Then I will change object type categorical data to category codes and create dummy variables for object type categorical data that have only "Yes" and "No" values. Then I will drop all the "No" dummy variables to eliminate redundancy. See the Keim D209 Task Two Code document for the code segments.

4. **Provide a copy of the cleaned data set.**
   See Keim D209 Task Two Clean Data document.

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).**
   See Keim D209 Task Two Training Data and Keim D209 Task Two Test Data documents.

2. **Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.**
   I first created feature and target arrays from the cleaned data set, then split them into 80% training data and 20% testing data. I then created a random forest regression model and fit the model to the training data. Then I tuned my hyperparameters. I created a dictionary of parameters, used them in a grid search, then fit the grid search to the training data. After that, I extracted the best estimator, predicted the y labels based on the testing data, and printed the parameters of the best model. I then calculated the model's R^2 and MSE values. All code and calculations can be found in the Keim D209 Task Two Code document.

3. **Provide the code used to perform the prediction analysis from part D2.**
   See Keim D209 Task Two Code document.

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.**

The model has an R^2 score of 0.99788, meaning that the predictions fit the test data 99% of the time. The MSE of the model is 1.492, meaning that the average squared distance of the predicted values from the test values is 1.492. The RMSE of the model is 1.23, meaning that the model can predict a customer's tenure accurately within 1.23 months.

2. **Discuss the results and implications of your prediction analysis.**
   The results of the prediction analysis indicate that a random forest regressor is a good fit to predict a customer's tenure. It has a high R^2 score, a low MSE and a margin of error of less than two months.

3. **Discuss one limitation of your data analysis.**
   While the analysis can help us predict what types of customers will stay with the company for longer tenures, it cannot do this with 100% accuracy. It also cannot tell us why customers are choosing to stay or go.

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**
   The scores discussed in part E2 indicate that the model does a good job predicting a customer's tenure with the company. I recommend using the model to examine the features of customers predicted to stay longer to maximize customer retention.