STAT 515: Final Project - Progress Report

Due before May 1st , 2023, in-person during office hours.

Section Number:	STAT-515-005
Group Number:	Group 15
Names of the Members: **(As stated on Blackboard)**	Chimmula Akhil Reddy Kurella Bhavesh Yelagandula Shashank

Title of the Project:	Classifying Income class
Describe the data set(s) that were considered: **(Provide Background information, including the source and the variables)**	The extraction was done by Barry Becker from the 1994 Census database. Prediction task is to determine whether a person makes over 50K a year. age, work class, fnlwgt, education,education-num,marital-status,occupation: Relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country Source: Donor: Ronny Kohavi and Barry Becker Data Mining and Visualization Silicon Graphics. e-mail: ronnyk '@' live.com for questions.
Explain the reasons why the above data set(s) were selected:	I have selected this dataset because in general now a days prediction of Income class it falls under. And it as region column which help me to do more part of visualization which we can't grab that insight form a table and it also as multivariate dataset both with independent and dependent variables of both having quantitative and qualitative variables which is more strong reason to choose this dataset. **And after slight research I have found only classification as only been done using dataset where also we have continuous variables, I want to take advantage of it** to perform more than 2 Machine Learning Models on this dataset.

Research Questions:	1. What will be the income class of new instance (i.e. new adult) when he is going to graduate. 2. What will be the gain profit of new adult who as idea on his capability and his demographics want to predict it future gain profit. 3. I also want to find decision tree analysis on this data and want to classify by general statements passed after understanding the dataset.
Explain the rational behind each question and how questions relate to the data set:	"I want to find whether the particular instance falls under which category income class." The given dataset contains information about individuals' demographic and employment characteristics, and the prediction task is to determine whether an individual's income is above or below 50K. Therefore, this question is directly related to the prediction task and involves using a classification model to determine the income class of a particular instance based on its demographic and employment characteristics. "I also want to find continuous gain value using demographics of the instance." This question is related to the prediction task but involves a different type of analysis. Rather than classifying an instance into a binary category (i.e., above or below 50K), this question seeks to determine the continuous value of an individual's income based on their demographic and employment characteristics. This could be accomplished using regression analysis to predict an individual's income based on their demographic and employment features. "I also want to find decision tree analysis on this data and want to classify by general statements passed after understanding the dataset." This question involves exploring the dataset using decision tree analysis, which is a popular machine learning algorithm for classification and regression tasks. Decision tree analysis involves building a tree-like model of decisions and their possible consequences, which can be used to classify or predict the outcome of a particular instance based on its features. The goal of this question is to develop a set of general statements or rules based on the decision tree analysis, which can be used to interpret and understand the dataset in a more intuitive way.

Provide an overview of the statistical methods that will be used to answer the research questions:	I want to Binary classification technique where I can find a particular class belongs to and do test analysis for first question. I want to perform regression on gain column form my second research question.
Explain the rationale behind the methods used:	Binary classification models such as logistic regression, decision tree classifier, or SVM are appropriate for the first research question because they are designed to predict the outcome of a binary variable based on one or more predictor variables. In this case, the binary variable is income class (above or below 50K), and the predictor variables are the demographic and employment characteristics of the individuals in the dataset. These models are commonly used in classification tasks, and they can be trained on the dataset to learn the relationships between the predictor variables and the outcome variable. Regression models such as linear regression, multiple regression, or polynomial regression are appropriate for the second research question because they are designed to predict the continuous value of an outcome variable based on one or more predictor variables. In this case, the outcome variable is the income of the individuals, and the predictor variables are their demographic and employment characteristics. Regression models are commonly used in prediction tasks, and they can be trained on the dataset to learn the relationships between the predictor variables and the outcome variable. Decision tree algorithms such as ID3, C4.5, CART, or Random Forest are appropriate for the third research question because they are designed to build a tree-like model of decisions and their possible consequences based on the predictor variables in the dataset. The decision tree can be used to classify or predict the outcome of a particular instance based on its features. Additionally, statistical methods such as chi-square tests, t-tests, or ANOVA can be used to analyze the relationships between different variables

	in the dataset and identify which variables are most important for predicting income class. These methods are commonly used in exploratory data analysis and can help to identify patterns and relationships in the dataset.
Any Challenges or issues faced/facing:	Missing data: The dataset may contain missing values or incomplete records for some individuals, which could affect the accuracy of the models or analysis. Model selection: Choosing the appropriate statistical model for the given research question and dataset can be challenging. There are many different types of models available, each with its strengths and weaknesses, and selecting the appropriate one can be tricky. Implementation of model: I may face issues while building the application which in backend uses ML Model and takes user input and generate results. Interpreting results: Even if the statistical models are accurate, interpreting the results and drawing meaningful conclusions can be challenging. It is essential to understand the limitations of the statistical methods used and to contextualize the results within the broader research question.
Comments (If any):	I Have both dependent and independent variables multivariant and dependent variables have quantitative and qualitative data. I have a region class in my dataset where I want to take advantage in visualization aspect after building the model I want to implement it as an application which can be helpful to people who don't know how to run a R code can also.