BAN 501

Business Analytics Project 3 (100 points possible)

Project Name	Predicting income above 50k
Project Due Date	Sunday by 11:59pm
Requirements	

This is an individual project. You are not to share code, insights, documents, or ideas. If you have questions, please direct them to me. Any code or projects that are duplicated or are very similar will receive 0 points.

Project Overview: Predict whether income exceeds $50,000 per year based on census data. This will be a challenging project where you will have to make decisions on how to proceed. Analyze the attached dataset according to supervised learning (Module 10) and the other tools and techniques we have learned thus far in the course. This dataset includes historical census data with a binary outcome of whether the individual earned an income above $50,000 (50k). Please see the data description file for a description of the data and variables. This project consists of two parts: the analysis in Python and a written report.

Instructions: Predict whether an individual's income is above $50,000 per year using the census_income.csv dataset. Use supervised learning (and versions of the sklearn algorithms we covered in Module 10) to analyze and determine whether an individual will earn above $50k/yr.

Generate insightful plots and charts to visualize key aspects of the data, such as income distribution across different demographic groups. You may create all the plots you wish, but you must only include a few key plots in your written report. Remember, you should always include descriptive statistics regardless of other approaches you utilize.

This dataset includes several categorical variables and some missing data. If you use any categorical variables, you will need to user your own learning and research to determine how to handle them and any missing data.

This dataset also requires that you set up your x and y variables a little differently than we covered in class. Rather than setting up lists and arrays, you will do something like the following. In this example, I am using age and education_level as predictor (independent) variables. This is not the solution, it is just an example:

```
k = 15
newage = 35
newhour = 60

# Create the x and y variables and the target values for the model.
x = income[['age', 'hours_per_week']]  # Independent variables
y = income['income_above_50k']  # Target variable
target_values = [[newage, newhour]]
```

This example does not use lists and sets the x and y variables directly. Also, notice that the target_array we used in the class lesson is different for this dataset. We should call it target_values and set the values as a collection without reshaping the items.

Next, because the outcome in this project is binary (1 or 0), you will need to use slightly modified versions of the supervised learning algorithms. Rather than using a regressor for each one, you will need to use a classifier. The only differences are what you import and how you apply the classifier. Here is an example of the changes the nearest neighbors algorithm:

```
from sklearn.neighbors import KNeighborsClassifier

knnclassifier = KNeighborsClassifier(n_neighbors=k)
```

Notice the difference in these two code snippets. Everything is the same except for using a classifier rather than a regressor. You will need to do this for each type of algorithm. Do not use linear regression for this project, but experiment with different supervised learning algorithms provided in the sklearn modules, including nearest neighbor, decision tree, random forest, and neural network classifiers. Compare and contrast the performance of each algorithm/model to determine the most suitable one for predicting income levels above $50,000.

Create a 75%/25% train-test split of the data. Train the models then test them using the test data. Calculate the error (prediction) scores for each algorithm to determine which is best. Discuss this in your report.

As part of your analysis and report, include examples of test values to illustrate how the best model predicts income levels for different individuals. Begin by describing the demographic characteristics of the selected individuals, such as age, education level, occupation, or whatever variables you included in your model. Next, present the predictions generated by the model, indicating whether each individual is predicted to earn above or below $50,000 per year. Finally, interpret the results to offer insights into the model's performance, discussing any observed patterns or trends and their implications.

In a separate Word document, write a brief report on what you found. Share your key insights and conclusions. Insert data, output, or plots to support your insights and conclusions. In the text, reference any data, output, or plots that you used and explain how those led you to your insights and conclusions. Do not include less helpful or unnecessary output or plots. Include an introductory paragraph and a strong conclusion in your report. Depending on the size and number of plots you include, you should aim for a report that is between 2-4 pages in length. You are free to use any format or style that you prefer.

Submit your report as either a Word document or a PDF. The deliverables for this project include your Python project code and your summary report.

Files for this project:
Project dataset: census_income.csv
Data description: census_income description.txt

Evaluation
This project will be evaluated out of a possible of 100 points. You are welcome to add additional functionality and to utilize your creativity in making the analysis even better. You will receive highest marks for writing a strong report that focusses on a few key insights and conclusions. You will also be evaluated in how you approach the project and how you handle the categorical variables and missing data. Your analysis should be strong and thorough.
Deliverables
Submit your written report and your project code on Canvas by 11:59pm on Sunday.