**Course code: ANL252**

**Title: ECA**

**SUSS PI No.: Y2181412**

**Tutorial group: T03**

**Name: YIEW SHER KYM LYNETTE**

**Submission Date: 21 October 2023**

**Question 1 (321/500 words)**

The 3 data pre-processing tasks used to clean up and prepare the data sets are (1) removing rows with missing values, (2) removing duplicated rows, and (3) standardization of categories.

The following will be imported to ease usage of it throughout the assignment.

|  |
| --- |
| # to import numpy, pandas and matplotlib.pyplot  import numpy as np  import pandas as pd  import matplotlib.pyplot as plt  # to import from sklearn  from sklearn import tree  from sklearn import preprocessing  from sklearn import model\_selection  from sklearn import metrics |

The dataset will be imported and checked using the following code.

|  |
| --- |
| # to import the dataset  df\_ECA = pd.read\_csv('ECA.csv')  # to show the first 5 rows for checking of data  print(df\_ECA.head()) |

Removing rows with missing ‘age’ data provide valuable insight towards the demographics, behavior and risks when looked at with the data of ‘charges’, ‘region’, ‘’smoker’, ‘BMI’ and ‘sex’. Interpolating the age of participants will result in skewed outputs for analysis. Using the mean of the neighboring value, a mode, or others, will result in certain age groups having artificially more numbers; bias. Moreover, removing rows with missing ‘age’ data ensures data completeness since incomplete data introduces bias and inconsistency into the analysis. The following will check for missing rows.

|  |
| --- |
| # this checks for missing data  missing = df\_ECA.isnull() # checks each cell for missing data  # this count the missing values and outputs the total missing values for each column  null\_count = missing.sum()  print(null\_count) |

This will remove the rows with missing data.

|  |
| --- |
| # removing rows with missing values  df\_ECA = df\_ECA.dropna() |

Removing duplicate rows ensures the data of the row does not appear more significant than others. When certain data have duplicates, the interpretation of output will be inaccurate due to confusion that may arise from overrepresentation. When duplicated rows are removed, the accuracy and quality of any analysis taking place with it will be better. Bias is prevented and data integrity is restated. Data duplication could have also been present due to mistakes during data entry. By removing them, the unknown mistake is rectified. The following checks for duplicate rows.

|  |
| --- |
| # checking for duplicate rows  duplicated = df\_ECA.duplicated()  #show the count of unique (False) and duplicated (True) rows  print(duplicated.value\_counts()) |

The duplicate rows are removed with the following code.

|  |
| --- |
| # Removing rows with duplicates  df\_ECA = df\_ECA.drop\_duplicates() |

The standardization of categories enables data in the data set to be read easily for further processes such as creation of charts/diagrams for analyzing the data. Data that’s supposed to be under the same category will not be missed out and can be used and compared more efficiently. The following will be used to examine the current values for “sex”.

|  |
| --- |
| # to examine the values in "sex" column for the first 10 rows of the dataset  print(df\_ECA["sex"].head(10)) |

This is the code used to standardize the values for “sex”.

|  |
| --- |
| # creating a mapping to convert "female" to "F" and "F" to "F".  sex\_map = {  "female": "F",  "F": "F",  "male": "M",  "M": "M",  }  df\_ECA["sex"] = df\_ECA["sex"].replace(sex\_map)  # print(df\_ECA.head())  print(df\_ECA["sex"].head(10)) |

**Question 2 (498/500 words)**

**Figure 1**

A graph of blue bars

Description automatically generated

This bar graph displays the average charges the health insurance billed the primary beneficiary for the respective number of children that the primary beneficiary has covered under the insurance. The graph shows almost a bell curve where the peak would be for primary beneficiaries with 3 children. This means beneficiaries with 3 children are being billed the most as compared to the rest. Beneficiaries with less than or more than 3 children are billed a lower amount. Since the children are covered by the health insurance, this could mean more illness, health issues or sophistication of medical condition are being experienced and identified when there are 3 children. The same could be said for beneficiaries with 2 children but proportionately, it is reasonable that the average bill is slightly less than those with 3 children. Parents with more than 3 children are less stressed and fuss less over their children (Dube, 2018). This will explain the decline in charges billed although there are more children. There is better judgement of what and who requires medical care, so the occurrence of medical treatment is less frequent.

This is the code to generate the figure.

|  |
| --- |
| # getting the data  children = [0, 1, 2, 3, 4, 5]  average\_charges = []  for c in children:  data = df\_ECA[df\_ECA["children"] == c]  average\_charge = data["charges"].mean() # this calculates the average charges billed by the health insurance to the primary beneficiary  average\_charges.append(average\_charge)  # plot settings  plt.figure(figsize=(7, 4))  plt.xlabel("Number of Children")  plt.ylabel("Average charges health insurance billed")  plt.title("The average medical bill of primary beneficiary to the number of children")  # plot the graph  plt.bar(children, average\_charges, width=0.6)  plt.show() |

**Figure 2**

A graph of smoking and smoking

Description automatically generated

This scatter plot shows the relationship between the payments that have been made by the primary beneficiary and their age, as well as the correlation of whether the primary beneficiary is a smoker or a non-smoker. At a glance, the smoker beneficiaries are paying more than their counterparts who do not smoke. Smoker primary beneficiary is likely paying more due to the negative health connotations associated and experienced with smoking and secondhand smoke. There is also an upward linear trend observable for both smoker and non-smoker whereby payments required to be made are higher as they age. This could be due to lower coverage by the health insurance or higher medical bills due to aging.

This is the code to generate the figure.

|  |
| --- |
| # label Encoding for "smoker"  smoker\_map = {  "yes": 1,  "no": 0  }  df\_ECA["smoker"] = df\_ECA["smoker"].replace(smoker\_map)  # getting the data  age\_s = df\_ECA[(df\_ECA["smoker"] == 1)]["age"] # smoker  payment\_s = df\_ECA[(df\_ECA["smoker"] == 1)]["charges"]  age\_ns = df\_ECA[(df\_ECA["smoker"] == 0)]["age"] # non-smoker  payment\_ns = df\_ECA[(df\_ECA["smoker"] == 0)]["charges"]  # plot settings  plt.figure(figsize=(6, 4), dpi=120)  plt.xlabel("Age")  plt.ylabel("Payment")  plt.title("Payment against Age")  # plot the graph  plt.scatter(age\_s, payment\_s, c="orange")  plt.scatter(age\_ns, payment\_ns, c="blue")  plt.legend(["smoker", "non-smoker"])  plt.plot() |

**Figure 3**

A graph of blue and orange bars

Description automatically generated

This multiple bar graph shows the total payments made by each region by smokers and non-smokers. On the most part, the smokers are being billed less than the non-smokers except for the southeast region of the US residential area. Making a connection to **Figure 2**, this may indicate that most of the smokers in the dataset live in the southeast region. It may also be an indication that a higher proportion of smoker aging population resides in the southeast region. Thus, causing the total payment to be much higher than each of the other total payments respectively. The total payment for southwest non-smoker is the least among non-smokers and second least among the smokers. Making a connection to **Figure 1**, where primary beneficiaries with more than or less than 3 children are billed lesser than those with 3 children, the proportion of primary beneficiaries in southwest with less than or more than 3 children is likely to be higher.

This is the code to generate first.

|  |
| --- |
| # Dummy encoding  categorical\_features = ["region"]  df\_ECA = pd.get\_dummies(df\_ECA, columns=categorical\_features)  print(df\_ECA.head()) |

After the above is generated, the following code will be able to generate the figure.

|  |
| --- |
| # Label Encoding for "smoker"  smoker\_map = {  "yes": 1,  "no": 0  }  df\_ECA["smoker"] = df\_ECA["smoker"].replace(smoker\_map)  # Get the data  regions = ["northeast", "northwest", "southeast", "southwest"]  total\_payments\_s = []  total\_payments\_ns = []  for r in regions:  column = f"region\_{r}" # Format to get the correct column  data\_s = df\_ECA[(df\_ECA[column] == 1) & (df\_ECA["smoker"] == 1)] # Rows where region is r and smoker is "yes"  data\_ns = df\_ECA[(df\_ECA[column] == 1) & (df\_ECA["smoker"] == 0)] # Rows where region is r and smoker is "no"    total\_payment\_s = data\_s["charges"].sum()  total\_payment\_ns = data\_ns["charges"].sum()    total\_payments\_s.append(total\_payment\_s)  total\_payments\_ns.append(total\_payment\_ns)    # Plot settings  plt.figure(figsize=(7, 4))  plt.xlabel("Region")  plt.ylabel("Total Payment")  plt.title("Total Payment by Region")  # Plot the graph  bar\_width = 0.35  bar\_positions = range(len(regions))  plt.bar(bar\_positions, total\_payments\_s, bar\_width, label="Smokers", color="orange")  plt.bar([pos + bar\_width for pos in bar\_positions], total\_payments\_ns, bar\_width, label="Non-Smokers", color="blue")  plt.xticks([pos + bar\_width / 2 for pos in bar\_positions], regions)  plt.legend()  plt.show() |

**Question 3 (300/300 words)**

The precursor to building decision tree will be discussed. The following code will be needed first.

|  |
| --- |
| # Dummy encoding  categorical\_features = ["sex"]  df\_ECA = pd.get\_dummies(df\_ECA, columns=categorical\_features)  print(df\_ECA.head()) |

This will be dummy encoding the categorial feature of “sex” into the dataframe ‘df\_ECA’ using the pandas library ‘pd.get\_dummied()’. ‘sex’ column will be converted into two binary columns “sex\_F” and “sex\_M” with 0 or 1 to represent whether there is the category. Next, the code below prepares the data for supervised machine learning using decision tree classifier to classify smoker and non-smoker, per the question to make dependable variable ‘smoker’. The features (x) and target (y) will be separated whereby features (x), the "age", "sex\_F", "sex\_M", "bmi", "children", "region\_northeast", "region\_northwest", "region\_southeast", "region\_southwest", "charges" will be used to make predictions and target (y) “smoker” will be the variable the decision tree will be aiming to predict.

|  |
| --- |
| # separating the features (x) and target (y) variables to pass them as inputs to the decision tree  features = ["age", "sex\_F", "sex\_M", "bmi", "children", "region\_northeast", "region\_northwest", "region\_southeast", "region\_southwest", "charges"]  target = ["smoker"]  x = df\_ECA[features]  y = df\_ECA[target] |

Next, this code below will split the dataset into training and testing sets for supervised machine learning. The training set ‘x\_train’ and ‘y\_train’ will train the model for machine learning and the test set ‘x\_test’ and ‘y\_test’ will be used to for performance evaluation of the decision model.

|  |
| --- |
| # split the data into train and test  x\_train, x\_test, y\_train, y\_test = model\_selection.train\_test\_split(  x, y, test\_size = 0.25, random\_state = 42  ) |

The next step will be to fit the decision tree. In the code below, the maximum depth is limited so the tree output will not be huge. The boxes will also be easier to see.

|  |
| --- |
| # fitting the decision tree  # classifier will be used because the classifications are smoker or non-smoker than predicting a continuous value (regression)  tree\_model = tree.DecisionTreeClassifier(max\_depth=3) # this limits the max depth  tree\_model.fit(x\_train, y\_train, sample\_weight=None) |

Lastly (optional), the accuracy of the decision tree model can be checked to have a gauge of how well it models the data. This provides insight on whether the decision tree can be used as a predictor and the relevance of insights gained from the output (Decision tree). The following code can be used.

|  |
| --- |
| # getting a prediction of the decision tree's accuracy  y\_pred = tree\_model. predict(x\_test)  # Get accuracy score  acc = metrics.accuracy\_score(y\_test, y\_pred)  print(acc) |

The high accuracy score of about 96% indicates that though the max-depth is 3, the decision tree is a good model for the data and any insights gained later are probably relevant.

**Question 4 (200/200 words)**

**Figure 4**

A diagram of a company

Description automatically generated

The most important feature for separating classes would be hospital charges as anything $14,511.863 or below (first box) is sufficient to separate most non-smokers. As seen from gini 0.006, an almost perfect purity where almost all data belongs to this class, smoking is very likely to increase hospital charges above said threshold. Furthermore, $13,833.654 or below further separates smokers from non-smokers using a lower charge. A purity of 0.0 was achieved for 652 samples whereby the non-smokers being charged by the health insurance $13,833.654 or below. This is slightly more than two-thirds of the initial sample of 911. Concluding based on charges, smoker beneficiaries are billed higher by health insurance. This is consistent with the findings from **Figure 2**.

For 3.5 or less children, which in this case rounding down to 3 children is sensible, 96 0.0 purity smoker beneficiaries had 3 or less children, while only 2 non-smoker beneficiaries had 3 or less children. This high number is consistent with **Figure 1** where average charges for health insurance billed is highest for beneficiaries with 3 children and second highest for beneficiaries with 2 children due to a higher proportion of smoker beneficiaries with higher charges.

The following codes were used.

|  |
| --- |
| # plotting the tree  plt.figure(figsize=(7, 4), dpi=300)  tree.plot\_tree(tree\_model,  feature\_names=features,  class\_names=["non-smoker", "smoker"],  filled=True) # colour  plt.show() |

**Question 5 (288/300 words)**

Decision trees can be effectively used for exploratory data analysis. Exploratory data analysis is a method used for dataset analysis and investigation (IBM, n.d.). Thereafter, the main characteristics of the dataset will be summed up by data scientists using this method (IBM, n.d.). Decision trees do not require a large dataset in order to make the rules. They are ideal for showing which feature of the dataset is able to give the most information, allowing for the identified features to be used first when making analysis. These features are also the most important for the purpose of making predictions (Scikit-learn, n.d.). The most influential datasets can be found high in the tree, and this will allow the data scientists to decide on which feature they should select or explore further.

Moreover, the layout of a decision tree will allow data scientists to be able to see how and where data has been divided into their classes or groups and understand the relationship of the features better. By identifying the interaction between features, data scientists will be able to identify possible interest patterns that they would like to explore further for the dataset.

Lastly, decision trees are able to detect outliers and reflect them, clearly show class distributions for understanding potential balances or imbalances and will segment the dataset into groups or clusters that can be useful. It also frees up work for the data scientists, so they are able to focus on carrying out quality work with the given dataset (IBM, n.d.).

Hence, the prioritization of influential datasets and possible meaningful interactions of features, as well as the preparatory nature of work carried out in decision trees to enable sense-making of the data, makes decision trees effective models for exploratory data analysis.

**References**

Dube, R. (2018, April 24). *Mom survey says: Three is the most stressful number of kids.* Today. <https://www.today.com/parents/mom-survey-says-three-most-stressful-number-kids-t127551>

IBM. (n.d.). *What is a Decision Tree?* <https://www.ibm.com/topics/decision-trees>

IBM. (n.d.). *What is exploratory data analysis?* <https://www.ibm.com/topics/exploratory-data-analysis#:~:text=Exploratory%20data%20analysis%20(EDA)%20is,often%20employing%20data%20visualization%20methods>.

Scikit-learn. (n.d.). *1.10. Decision Trees.* <https://scikit-learn.org/stable/modules/tree.html>