**Logo

Description automatically generated**

Course code : **ANL252**

Course title : **Python for Data Analytics**

Assessment title : **End-of-Course Assessment (ECA)**

**July Semester 2023**

Student name **: Lee Jay Hoon Jenny**

Student PI number : **Y2070655**

Submission date **: 3 November 2023**

**Question 1**

Three (3) data pre-processing tasks to clean and prepare the dataset:

1. **Treating Categorical Values**

This task involves the encoding of categorical values in the columns ‘sex’, ‘region’, and ‘smoker’. The categorical values were converted into numerical representations, ensuring consistency among the data and their usability in the subsequent analyses.

Initially, there were discrepancies in the ‘sex’ column, with entries being labeled both 'M' and 'F' as well as 'male' and 'female.'  This task harmonised the ‘M’ and ‘male’ entries with the integer value 1, and ‘F’ and ‘female’ entries with the integer value 2.

The ‘region’ and ‘smoker’ columns were treated with the LabelEncoder.

The values ‘northeast’, ‘northwest’ , ‘southeast’, and ‘southwest’ were converted to numerical values 0, 1, 2, and 3, respectively. Additionally, the ‘smoker’ column type was changed to boolean, holding either a 1 or 0 value.

1. **Treating Numerical Values**

The column ‘bmi’ was converted from a numerical variable to a categorical one.

The wide range of BMI values in the dataset was categorised into the following 4 categories:

1. Underweight,
2. Healthy,
3. Overweight
4. Obese.

This categorisation will allow the easier drawing of insights in the subsequent analyses and enable a clearer understanding of an individual’s health status when compared to the other variables.

1. **Treating Missing Values**

There were 123 missing values in the ‘age’ column, which was treated by imputing them with the mean age of the dataset.

This decision was made as completely removing the rows with the missing values would have resulted in the loss of relevant data. Furthermore, it is essential to recognize that there are other columns holding crucial data related to medical costs, not just the age column.

Although imputation does introduce some degree of inaccuracy in the ‘age’ column, it is a considered trade-off that preserves other crucial data elements essential for subsequent analysis.

Prior to the above three steps, the ‘PersonID’ column was removed, as it only serves as a unique identifier and does not provide meaningful information for analysing the dataset.

Codes for the above treatments:

|  |
| --- |
| *#Import Library* import numpy as np import pandas as pd from sklearn import preprocessing dataset = pd.read\_csv("ECA.csv")  *# Remove personID* dataset = dataset.drop('PersonID', axis=1)  *### Task 1 ###* *# Convert the sex Field* dataset['sex'] = dataset['sex'].replace(['M','male'], 1) dataset['sex'] = dataset['sex'].replace(['F','female'], 2) *#Encoding Categorical Variables into numerical format* label\_encoder = preprocessing.LabelEncoder() *#Region Field* dataset['region']= label\_encoder.fit\_transform(dataset['region']) *#Smoker Field* dataset['smoker']= label\_encoder.fit\_transform(dataset['smoker']) dataset['smoker'] = dataset['smoker'].astype('bool')  *### Task 2 ###* *#Categorise BMI to Underweight, Healthy, Overweight, Obese* dataset['bmi']=np.where(((dataset['bmi']<18.5)), 0, dataset['bmi']) *#Underweight (0)* dataset['bmi']=np.where(((dataset['bmi']>=18.5) & (dataset['bmi']<=24.9)), 1, dataset['bmi']) *#Healthy (1)* dataset['bmi']=np.where(((dataset['bmi']>24.9) & (dataset['bmi']<=29.9)), 2, dataset['bmi']) *#Overweight (2)* dataset['bmi']=np.where(((dataset['bmi']>29.9)), 3, dataset['bmi']) *#Underweight (3)* dataset['bmi'] = dataset['bmi'].astype('int64')  *### Task 3 ###* *#Check which columns have missing values* dataset.info() *#Fill the missing values for the age columns with mean --> Dataset* dataset['age'].fillna(dataset['age'].mean(), inplace=True) *# Convert the 'age' column to integer* dataset['age'] = dataset['age'].astype(int) |

[342 words]

**Question 2**

The three (3) charts based on the medical costs dataset:

**Chart 1**: Grouped Bar Chart - Distribution of Smokers by BMI and Gender

* Among the 276 smokers, 82% are considered unhealthy (Underweight/Overweight/Obese) based on their BMI. The remaining 18% of smokers are categorised as healthy. This suggests that a majority of smokers have higher chances of facing health concerns.
* In total, there are 161 male smokers and 106 female smokers. With 85% of male smokers and 82% of female smokers, being categorised as unhealthy, the distribution reveals that there is no gender-related disparity here.
* Within the subset of unhealthy smokers, there are significantly more overweight and obese smokers compared to underweight smokers. Among male smokers, 84% are found to be overweight and obese. The female smokers mirror a similar trend, with 78% falling into the overweight and obese categories.

This shows a trend of overweight and obesity amongst unhealthy smokers.

To summarise this graph, the analysis showed that a sizable percentage of smokers are unhealthy and run the risk of developing health issues, regardless of their gender.

A graph of smokers and bmi

Description automatically generated

Chart 1 Codes:

|  |
| --- |
| *# Imports* import numpy as np import pandas as pd import matplotlib.pyplot as plt  *# Prepare dataset for smokers sorted by BMI and Sex* bmi\_sex\_smoker = dataset.groupby(['bmi','sex'])['smoker'].sum().reset\_index() bmi\_sex\_smoker['sex'] = bmi\_sex\_smoker['sex'].replace([1], 'Male') bmi\_sex\_smoker['sex'] = bmi\_sex\_smoker['sex'].replace([2], 'Female')  *# Change sex to categorical data* female\_smoker = bmi\_sex\_smoker[bmi\_sex\_smoker['sex'] == 'Female'] male\_smoker = bmi\_sex\_smoker[bmi\_sex\_smoker['sex'] == 'Male']  *# Plot graph* plt.figure(figsize=(6,6)) plt.bar(male\_smoker['bmi'] - 0.2, male\_smoker['smoker'], width=0.4, label='Male') plt.bar(female\_smoker['bmi'] + 0.2, female\_smoker['smoker'], width=0.4, label='Female')  *# Add data labels to the bars* for i, value in enumerate(male\_smoker['smoker']):     plt.text(i - 0.2, value + 0.5, str(value), ha='center', va='bottom')  for i, value in enumerate(female\_smoker['smoker']):     plt.text(i + 0.2, value + 0.5, str(value), ha='center', va='bottom')  plt.xlabel("BMI Range") plt.ylabel("Number of smokers") plt.title("Distribution of Smokers by BMI and Gender") plt.xticks([0,1,2,3],['underweight','healthy','overweight','obese']) plt.legend() plt.show() |

**Chart 2**: BoxPlot - Distribution of Insurance Charges: Non-Smokers vs Smokers

* The boxplots below have been labelled with the key statistical values rounded to two decimal places: minimum, maximum, median, 25th percentile, and 75th percentile.
* At first glance, there is a stark contrast between the insurance charges for smokers and non-smokers. Non-smokers pay significantly less for insurance, with a median charge of $7345.41. This is 78.6% lower than smokers' median charge of $34371.51.
* Despite the 46 outliers among the boxplot for non-smokers, the highest outlier remains below the 75th percentile value for smokers. Ignoring outliers, the maximum insurance charge of the non-smoker is $22395.74, while the smoker’s is $63770.43.

This further proves the disparity between the insurance charges for smokers and non-smokers.

In summary, there is a clear disparity in insurance charges between non-smokers and smokers. Based on the key statistical values, the charges incurred by non-smokers are consistently lower, reaffirming that a non-smoker pays less insurance premiums than a smoker.

**A graph of a number of smokers

Description automatically generated**

Chart 2 Codes:

|  |
| --- |
| *# Imports* import numpy as np import pandas as pd import matplotlib.pyplot as plt  *# Insurance charges sorted by smoker vs non-smoker* smoker\_charges = dataset['charges'].loc[dataset['smoker'] == 1] nonsmoker\_charges = dataset['charges'].loc[dataset['smoker'] == 0] *# Plot graph* plt.figure(figsize=(9,8)) boxplot = plt.boxplot([nonsmoker\_charges, smoker\_charges], flierprops={'marker': 'o', 'markersize': 4, 'markerfacecolor': 'cyan'}, widths=0.3) plt.xlabel("Smoking Status") plt.ylabel("Insurance Charges") plt.title("Distribution of Insurance Charges: Non-Smokers vs Smokers") plt.xticks([1,2],['Non-Smoker', 'Smoker']) plt.yticks(np.arange(0, 75000, 5000)) *# Annotate the box plot with min, max, median, Q1, and Q3 values* ax = plt.gca() ax.set\_xlim(1-1, 2+1) *# Draw median line* for line in boxplot['medians']:     x, y = line.get\_xydata()[1]     plt.text(x-0.15, y+500, '%.2f' % y,          ha='center', fontsize=9) *# Draw Q1 and Q3 lines* for line in boxplot['boxes']:     x, y = line.get\_xydata()[0]     plt.text(x-0.25,y+500, '%.2f' % y,          ha='left',          va='top', fontsize=9)     x, y = line.get\_xydata()[3]     plt.text(x-0.25,y+500, '%.2f' % y,          ha='left',          va='top', fontsize=9)  *# Draw min/max lines* for line in boxplot['caps']:     x, y = line.get\_xydata()[1]     plt.text(x-0.18,y, '%.2f' % y,          ha='right', fontsize=9)  *# Find number of outliers* outliers = [flier.get\_ydata() for flier in boxplot['fliers']] non\_smoker\_outliers = len(outliers[0]) smoker\_outliers = len(outliers[1])  plt.show() |

**Chart 3**: Scatterplot - Distribution of Insurance Charges by the Number of Children

* The scatterplot has been annotated with the average insurance charges for each group, rounded to two decimal places.
* From preliminary observation, it appears that individuals with four or five children have a lower average charge than those with zero to three children. However, there are significantly fewer individuals with more than three children, rendering this observation inaccurate.
* Further analysis of average charges showed minimal disparities between individuals with zero and one child ($12377.97 and $12731.17). Similarly, those with two and three children also showed minimal variation in average charges ($15073.56 and $15355.32).

This suggests a consistent premium threshold for individuals with zero to one child and an elevated premium for those with two to three children.

While there appears to be an upward trend in insurance charges when having more children, there is insufficient data regarding individuals having four to five children to make a definitive conclusion on this.

**A graph showing the amount of insurance charges

Description automatically generated**

Chart 3 Codes:

|  |
| --- |
| *# Imports* import numpy as np import pandas as pd import matplotlib.pyplot as plt  *# Get the average insurance charge per number of children* mean\_charges = dataset.groupby('children')['charges'].mean().reset\_index() *# Get the number of people per number of children* no\_of\_children = dataset.groupby('children')['charges'].count()  *# Plot graph* plt.figure(figsize=(9,8)) plt.scatter(dataset['children'], dataset['charges'], label='Insurance Charges', alpha=0.5) plt.xlabel('Number of Children') plt.ylabel('Insurance Charges') plt.title('Distribution of Insurance Charges by Number of Children') plt.yticks(np.arange(0, 75000, 5000)) ax = plt.gca() ax.set\_xlim(-0.5, 5.7)  *# Annotate and label data with average charges and number of people* for children, charges in enumerate(mean\_charges['charges']):     text = "%.2f\nCount: %d" % (charges, no\_of\_children[children])     plt.annotate(text, (children+0.1, mean\_charges['charges'].where(mean\_charges["children"]==children)[children]))  plt.legend() plt.show() |

[500 words]

**Question 3**

The decision tree was achieved using the DecisionTreeClassifier object provided by the scikit-learn library.

In the first step, I determine the dependent variable and the independent variables. As given in the question, the dependent variable is ‘smoker’. The independent variables are the other variables in the dataset which are 'sex ','bmi','children','region' and 'charges'.

In the next few steps, the DecisionTreeClassifier model will be trained, and then used to predict the dependent ‘smoker’ variable,

To do this, I first divide the dataset into training and test sets using the ‘train\_test\_split’ function provided by scikit-learn. A 70-30 split is used on the data such that 70% of the data is allocated for training (X\_train and Y\_train), and 30% is reserved for testing (X\_test and Y\_test).

After this, I trained the model using the training data (X\_train and Y\_train) via the fit method. This allows the classifier to learn and map the independent variables to the dependent variable based on the training data.

Once the training has completed, the classifier can then be used to predict the dependent variable 'smoker' for the test dataset (X\_test).

To understand the accuracy of the classifier, the metrics.accuracy\_score function is used. The function compares the predicted values (y\_pred) to the actual values in the test set (Y\_test) and computes the accuracy score.

|  |
| --- |
| *# Imports* from sklearn import preprocessing from sklearn import tree from sklearn import metrics from sklearn.tree import DecisionTreeClassifier from sklearn.model\_selection import train\_test\_split  independant\_variables = ['age','sex','bmi','children','region','charges'] dependant\_variables= ['smoker']  X =dataset[independant\_variables] Y =dataset[dependant\_variables]  *# Split dataset into training set and test set with 70% training and 30% test* X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.3, random\_state=1)  clf = DecisionTreeClassifier()  *# Train Decision Tree Classifer* clf = clf.fit(X\_train,Y\_train)  *#Predict the response for test dataset* y\_pred = clf.predict(X\_test)  print("Accuracy:",metrics.accuracy\_score(Y\_test, y\_pred)) |

[215 words]

**Question 4**

The decision tree is coloured orange and blue, representing the ‘non-smoker’ and ‘smoker’ categories respectively.

* From the decision tree, it is evident that the people who were charged a lesser amount for their insurance costs are non-smokers. The left-side of the tree represents individuals with lower insurance charges, while the right-side represents individuals with higher insurance charges. There are more orange nodes clustered on the left.
* The blue nodes are scattered around but grow in colour intensity on the right-side of the tree, which reveal that a large number of smokers have charges within the range of $14525.11 to $33473.89.
* The blue nodes on the tree’s left-side are there as the ‘age’ is below 29.5, suggesting that the individuals who smoke have lower insurance charges because they are younger.

* In contrast, the orange nodes on the tree’s middle-right-side reveal that non-smokers have higher insurance charges as they ‘age’ within the range of 43 to 53, suggesting that older individuals have higher insurance charges.
* Finally, the orange nodes on the tree’s right-side reveal that non-smokers with high insurance charges have a high BMI and are considered obese (according to the categorical sorting in Question 1).

A diagram of a flowchart

Description automatically generated

|  |
| --- |
| feature\_names=['age','sex','bmi','children','region','charges']  plt.figure(figsize=(32,20)) tree.plot\_tree(clf, fontsize=8, filled=True, feature\_names=feature\_names) plt.show() |

[194 words]

**Question 5**

The decision trees cannot be effectively used for exploratory data analysis. According to (Hillier, 2023), decision trees are associated with classification and prediction tasks for their ability to segment data into smaller parts. This aligns with Question 4 observations, where the dataset was broken down into various nodes, allowing better data visualisation. For instance, it was easy to see how many samples had the feature ‘age <= 21.0’. For data analytics, the breaking down of data may be useful, allowing one to discover patterns, detect outliers, and see the interactions between features.

In a specific case study, (Automated Fulfillment, 2018) used decision trees for exploratory data analysis and data visualisation using a retail store dataset. Their objective was to assess classification accuracy, feature importance, and visualise target segmentation. However, they obtained mixed results when attempting to determine the important features. The decision tree deemed one of the variables, ‘average property value’ as the most important, despite conflicting evidence, stressed a need for further investigation to validate the decision tree’s conclusion.

To conclude, while decision trees may offer and facilitate the visual exploration of data, they might be unsuitable for conducting a comprehensive analysis independently or covering the entire spectrum of analytical requirements.

**References**

* Automated Fulfillment. (2018, March 12). *Exploratory Analysis with Decision Trees*. Exploratory Analysis with Decision Trees. Retrieved October 22, 2023, from https://automated-fulfillment.com/experiment-1/
* Hillier, W. (2023, April 17). *What Is a Decision Tree and How Is It Used?* What Is a Decision Tree and How Is It Used? Retrieved October 22, 2023, from https://careerfoundry.com/en/blog/data-analytics/what-is-a-decision-tree/

[202 words]