

**ANL252**

**Python for Data Analytics**

# **End-of-Course Assessment - July Semester 2023**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **LIM KAIWEN** | **N2270422** |

**Tutorial Group: ­­­­­­­­­­­­ T03**

**Instructor’s Name: MUNISH KUMAR**

**Submission Date: 03/11/2023**

**Question 1**

**1) Remove missing values**

import pandas as pd

df = pd.read\_csv("ECA.csv")

#Find out how many missing values are in the dataset

df.isnull().sum(axis = 0)

PersonID 0

age 123

sex 0

bmi 0

children 0

smoker 0

region 0

charges 0

dtype: int64

#Remove the missing values

df.dropna(axis = 0 , how ="any", inplace = True)

It has been discovered that there are 123 missing values in the age column. Age is an important data in the data set as it allows us to identify the age of the person and determine if there is any correlation between age and medical cost as well as other variables.Therefore, to create an accurate analysis result, I will remove the missing data from the data set. After removing the missing values, the dataset contains 1217 rows and 8 columns of data.

**2) Reduce categorical variable**

df["sex"].replace({'M': 'male', 'F': 'female'}, inplace=True)

df.head(8)

PersonID age sex bmi children smoker region charges

0 1 19.0 female 27.900 0 yes southwest 16884.92400

1 2 18.0 male 33.770 1 no southeast 1725.55230

2 3 28.0 male 33.000 3 no southeast 4449.46200

3 4 33.0 male 22.705 0 no northwest 21984.47061

4 5 32.0 male 28.880 0 no northwest 3866.85520

5 6 31.0 female 25.740 0 no southeast 3756.62160

6 7 46.0 female 33.440 1 no southeast 8240.58960

7 8 37.0 female 27.740 3 no northwest 7281.50560

In the dataset, it can be observed there are four categorical data in the sex column, M, F, female, and male. It can be confusing and may lead to inaccurate results if left untreated as there can be only two categorical variables in the sex column which are female and male. It is logical to assume that the variable M refers to male and F refers to female, therefore, to reduce the number of categorical variables, I will replace all the M and F variables with male and female respectively. By removing the additional categorical variables, it improves the readability of the data and makes it more accessible to a broader audience. In addition, it also ensures that the data adheres to a consistent and well-understood convention. This consistency can help prevent errors and misunderstandings during data analysis.

3**) Remove duplicates data**

#Check if the dataset contains duplicates

duplicates = df.duplicated()

duplicate\_data = df[duplicates]

print(duplicate\_data)

PersonID age sex bmi children smoker region charges

100 100 38.0 male 19.3 0 yes southwest 15820.699

101 100 38.0 male 19.3 0 yes southwest 15820.699

#Remove the duplicates

final\_df = df.drop\_duplicates()

It is evident from the dataset that there are duplicates since each person's ID is unique, there can only be one unique ID per person. Hence, it is required to remove the duplicates as duplicated data can compromise the integrity of the dataset as these duplicates can skew the results and lead to inaccurate analysis results. By removing the duplicates, it also improves the efficiency of data processing. After removing the duplicates, the dataset contains 1215 columns and 8 rows of data.

[496 words]

**Question 2**

**Graph 1.1 Scatterplot of BMI vs Charges**

A graph of scatter plot of bmi versus medical cost

Description automatically generated

From graph 1.1, it can be observed there is a positive correlation between BMI and medical cost.

Generally, the data points are closely clustered together, and the trend line also shows an upward-sloping line which indicates that as BMI increases, the medical cost also increases. From the graph, it can be observed that people with higher BMIs (30 - 40) typically pay more for their health insurance since they are more likely to develop serious illnesses like high blood pressure and high cholesterol. Most people whose BMI are within the healthy range such as 18 – 25 usually pay less than 10K for their medical bills. In addition, some people even have to pay between 40k to 50K when their BMI is over 40.

import seaborn as sns

import matplotlib.pyplot as plt

import pandas as pd

df = pd.read\_csv("final\_df.csv")

#create the figure size

plt.figure(figsize=(10, 6)

)

sns.regplot(data=df, x='bmi', y='charges', scatter\_kws={"s": 10}, line\_kws={"color": "green"})

plt.title('Scatter Plot of BMI vs. Medical Cost')

plt.xlabel('BMI')

plt.ylabel('Medical Cost ($)')

plt.show()

**Graph 1.2 Bar plot medical cost by region**

A graph of different colored bars

Description automatically generated

Graph 1.2 shows the distribution of medical costs by region. According to graph 1.2, the residents of the US Southeast region have the highest medical bills ($4,832815) that are covered by their health insurance. This could indicate that a greater proportion of smokers or people with higher BMI live in the southeast. Hence, mayors in respective regions of the US might wish to create more campaigns to promote healthy lifestyle choices and work to deter people from smoking. The northeast region has the second highest medical bills, coming in at $3,941420, followed by the northwest region with $3,708439, and the southwest region with $3,632342.

import matplotlib.ticker as ticker

colors = sns.color\_palette("Set2")

sns.set(style="whitegrid")

plt.figure(figsize=(10, 6))

ax = sns.barplot(data=df, x='region', y='charges', estimator=sum, ci=None, palette=colors)

plt.xlabel('Region', fontsize=14)

plt.ylabel('Total Medical Cost', fontsize=14, labelpad=14) # Adjust the labelpad

plt.title('Total Medical Costs by Region', fontsize=14)

ax.set\_yscale("linear")

#shows the figures

def format\_func(value, tick\_number):

return f"${int(value):,}"

ax.yaxis.set\_major\_formatter(ticker.FuncFormatter(format\_func))

for p in ax.patches:

ax.annotate(f"${p.get\_height():.0f}", (p.get\_x() + p.get\_width() / 2., p.get\_height()), ha='center', va='center', fontsize=12, color='black', xytext=(0, 10), textcoords='offset points')

plt.xticks(rotation=0)

plt.tight\_layout()

plt.show()

**Graph 1.3: Boxplot of smoker vs medical cost**

A graph showing a number of patients

Description automatically generated

It can be observed in graph 1.3 that the median medical expenditures paid by smokers and non-smokers are $34456.35 and $7418.52 respectively. As smokers are more likely than non-smokers to suffer from catastrophic illnesses like lung cancer, it makes sense that their medical costs would be higher when it comes to their health insurance. In addition, people who smoke can even be charged more than $60k for their medical charges as compared to a non-smoker who only pays less than $40k for their medical costs.

# Encode the 'smoker' column into numerical values

df['smoker\_encoded'] = df['smoker'].map({'yes': 1, 'no': 0})

sns.set(style="whitegrid")

ax = sns.boxplot(data=df, x='smoker\_encoded', y='charges', palette='Set3')

ax.set\_xticklabels(['No', 'Yes'])

#add legends

plt.xlabel('Smoker')

plt.ylabel('Medical Cost')

plt.title('Distribution of Medical Costs by Smoker')

#Add the median labels

medians = df.groupby(['smoker'])['charges'].median()

for i, median in enumerate(medians):

ax.text(i, median, f'Median: {median:.2f}', verticalalignment='center', size='medium', color='black', weight='semibold')

plt.show()

[498 words]

**Question 3**

The dataset is loaded into a Pandas DataFrame, which contains both numerical and categorical data. To prepare the decision tree, one-hot encoding is implemented to convert categorical variables such as ‘sex’ and ‘region’. This process converts these categorical variables into binary columns, which allows the model to work effectively with categorical data. To prevent multicollinearity, the ‘drop\_first = True’ is also implemented to eliminate one category from each one-hot encoded column.

To form the basis for the model’s prediction, the feature selection is performed by specifying the variables such as ‘age’, ‘charges’ , ‘bmi’ ‘sex\_male’, and ‘region’ which are the columns to be extracted from and ensuring ‘smoker’ is the target variable that I want to predict using the decision tree. Using the if statement also ensures that every categorical variables are included in the model.

To assess the model’s performance, the dataset is divided into training and testing sets using the train\_test\_split from SciKit learn. By setting the ‘random\_’state’ parameter, it ensures

reproducibility in the model runs.

To prevent overfitting, the decision tree classifier is created with a maximum depth of 3 so that it can maintain the complexity of the decision tree. Based on the training data, the model will be trained and learned to make predictions based on the identified relationships within the dataset.

Feature names are transformed into a list to make feature visualisation in the decision tree easier. Making predictions with the test data is how the model's performance is assessed, and the accuracy of these predictions is computed and reported.

Based on the feature selections, the model's rules are visually represented by the decision tree diagram. By interpreting the particular feature values that affect these predictions, the interpretation of the decision tree provides insights into how the model categorises people as smokers or non-smokers.

[297 Words]

**Question 4**

A diagram of a network

Description automatically generated

The decision tree is created using the code below.

from sklearn.tree import DecisionTreeClassifier

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score

import matplotlib.pyplot as plt

from sklearn.tree import plot\_tree

df = pd.read\_csv('final\_df.csv')

# One-hot encode 'sex' and 'region'

df = pd.get\_dummies(df, columns=['sex', 'region'], drop\_first=True)

# Verify that 'region\_northeast' is included in the one-hot encoding

if 'region\_northeast' not in df.columns:

df['region\_northeast'] = 0

# Select features and target variable

X = df[['age', 'charges', 'bmi', 'sex\_male', 'region\_northwest', 'region\_southeast', 'region\_southwest', 'region\_northeast']]

y = df['smoker']

# Split the data into a training set and a test set

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Create and train the decision tree classifier with a max depth

clf = DecisionTreeClassifier(max\_depth=5)

clf.fit(X\_train, y\_train)

# Convert the Index to a list for feature\_names

feature\_names = X.columns.tolist()

# Make predictions on the test set

y\_pred = clf.predict(X\_test)

# Evaluate the model's accuracy

accuracy = accuracy\_score(y\_test, y\_pred)

print(f"Accuracy: {accuracy:.2f}")

# Plot the decision tree with an enlarged figure size

plt.figure(figsize=(12, 6))

plot\_tree(clf, feature\_names=feature\_names, class\_names=['no', 'yes'], filled=True, rounded=True)

plt.show()

The decision tree analysis provides crucial information about the factors that determine a person's position as a "smoker". "Charges" turns out to be the key predictor: those who pay more for healthcare are more likely to smoke, whereas people who pay less are more likely to not smoke. This demonstrates the substantial financial correlation between smoking and healthcare expenses.

Secondary contributors such as "age" and "bmi" indicate that while older people are more likely to not smoke, younger people are generally more likely to be smokers. In addition, smoking may be linked to why people have higher BMI values.

‘Sex’ also has an adverse impact because the variable 'sex\_male' suggests that being male makes one more likely to smoke, which is consistent with previous gender-based smoking trends. It also shows that region is an important component as the decision tree takes into account regional differences in smoking prevalence. To comprehend the smoking habit, this spatial effect is crucial.

The main factors that determine whether a person is a "smoker" are, essentially, healthcare costs, age, BMI, gender, and region. These insights provide useful knowledge of the complex variables affecting smoking behaviour, with "charges" emerging as the most significant and important predictor.

[Word count:200]

**Question 5**

A decision tree is a non-parametric supervised learning algorithm for classification and regression tasks. (Gupta, 2017) Decision tree can be effectively used as it offers a clear and intuitive representation of the data which allows non-technical stakeholders to be able to comprehend the information easily. One of the benefits is that a decision tree can perform feature selection therefore it is able to rank the variables according to their significance, which aids in identifying the key variables that influence the target variable. This is useful especially in exploratory data analysis as users can prioritise more research into the most influential variable. (Gupta, 2017)

Secondly, a decision tree can handle both categorical and numerical data, which simplifies the process of handling diverse data types. In addition, it can also handle multi-output problems.

Thirdly, unlike other data analytics techniques which often require a lot of effort in terms of data preparation and data normalisation, a decision tree requires minimal data preparation from users. (Hillier et al., 2023)

Lastly, in terms of visualisation, a decision tree also serves as a visual aid as the tree structure provides makes it easier to spot patterns and connections in the data. It makes complex interactions easier to understand by making splits and decisions made at each node easily identifiable. (Gupta, 2017)

Although a decision tree can provide essential insights and make predictions for the users. There are still limitations in using a decision tree for exploratory data analysis. For example, decision trees are particularly vulnerable to overfitting. Overfit models are less reliable for making predictions as they capture noise in the data. Results of the decision tree can be affected even with small changes in the dataset. Hence, the sensitivity may be a disadvantage if the data sets are often being updated. (Gupta, 2017)

[Word Count: 298]

**References**

Hillier, W.(2023, April 17). *What is a decision tree and how is it used?*. CareerFoundry. https://careerfoundry.com/en/blog/data-analytics/what-is-a-decision-tree/

Gupta, P. (2017, November 12). *Decision trees in machine learning*. Medium. https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052