

**ANL252 ECA**

**Python for Data Analytics**

**End of Course Assignment**

**ECA01**

**July 2023**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **ISAAC CHRISTOPHER ONG WEN XUAN** | **Q2180896** |

**SUBMITTED ON: 31 Mar. 23**

# **Question 1:**

**1. Addressing Missing Data:**

Upon examining the dataset, it was observed that the 'age' column had 123 missing entries (fig 1.1). Since 'age' is likely a significant variable when analyzing medical costs, it was crucial to address these gaps. Deleting these records could have led to a loss of valuable information, considering that they account for about 9% of the dataset. Therefore, I chose to fill in the missing 'age' values with the median age of the dataset. The median was used over the mean to mitigate the influence of any outliers or skewed distribution in the 'age' data.

**2. Statistical Analysis:**

A statistical summary of the dataset provided a snapshot of the central tendencies and dispersion of the numerical variables such as 'age', 'bmi', 'children', and 'charges'. This step is essential in understanding the data's distribution and in making informed decisions on further data processing needs. For instance, the average 'bmi' indicated that the dataset might include a significant number of overweight or obese individuals, which is an important factor in medical cost analysis.

**3. Encoding Categorical Variables:**

The dataset contained categorical variables ('sex', 'smoker', 'region') that machine learning algorithms cannot interpret in their raw form. To convert these into a machine-readable format, one-hot encoding was applied. This method created new columns for each category with binary values (0s and 1s), ensuring that no artificial ordering is implied among categories. For example, instead of 'sex' being a single column with 'male' or 'female', it was split into 'sex\_male' and 'sex\_female', with '1' indicating the presence of the attribute and '0' its absence.

import pandas as pd

# Load the dataset

file\_path = '/content/ECA.csv'

eca\_data = pd.read\_csv(file\_path)

# Display the first few rows of the dataframe

eca\_data.head()

# Check for missing values

missing\_values = eca\_data.isnull().sum()

# Get a statistical summary of the numerical features

statistical\_summary = eca\_data.describe()

missing\_values, statistical\_summary

# Impute missing values in 'age' with the median value

median\_age = eca\_data['age'].median()

eca\_data['age'].fillna(median\_age, inplace=True)

# Verify if missing values are filled

missing\_values\_after\_imputation = eca\_data.isnull().sum()

# Perform encoding for the categorical variables

# We use one-hot encoding for 'sex', 'smoker', and 'region' columns

eca\_data\_encoded = pd.get\_dummies(eca\_data, columns=['sex', 'smoker', 'region'], drop\_first=True)

# Display the first few rows of the encoded dataframe and the missing values check

missing\_values\_after\_imputation, eca\_data\_encoded.head()

Word Count: 376

# **Question 2:**

To provide meaningful insights from the ECA dataset on medical costs, I plotted three types of Graphs:

1. Distribution of Charges: A histogram to visualize the distribution of medical charges in the dataset.
2. Relationship between BMI and Charges: A scatter plot to observe the potential correlation between BMI and medical charges, which could suggest that higher BMI may be associated with higher medical costs.
3. Average Charges by Region: A bar chart to compare the average medical charges across different regions.

**Insights for Graph 1:**

A graph with blue lines

Description automatically generated

The histogram of medical charges reveals a right-skewed distribution, indicating that a larger number of individuals have medical costs on the lower end of the spectrum, while fewer individuals incur very high medical expenses. This skewness could be indicative of the presence of outliers or extreme values representing expensive medical procedures or treatments. The long tail to the right suggests that while most people may have relatively moderate healthcare costs, there's a significant variation with some facing much higher expenses, potentially due to chronic conditions, emergencies, or less common diseases that are more expensive to treat.

import matplotlib.pyplot as plt

import seaborn as sns

# Set the aesthetic style of the plots

sns.set\_style("whitegrid")

# Figure 1: Distribution of Charges

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

sns.histplot(eca\_data['charges'], kde=True, bins=30, color='skyblue')

plt.title('Figure 1: Distribution of Medical Charges')

plt.xlabel('Charges')

plt.ylabel('Frequency')

plt.show()

Next, we will create a scatter plot to explore the relationship between BMI and medical charges.

**Insights for Graph 2:**

A graph showing a diagram of a scatter plot

Description automatically generated

The scatter plot of BMI against medical charges does not show a clear linear relationship, but there appears to be a trend where individuals with higher BMI tend to have higher medical charges. This pattern could suggest that as BMI increases—potentially moving from a healthy range to overweight and then to obesity—medical costs may rise, possibly due to the increased risk of health issues associated with higher BMI. The cloud of points is denser at lower BMI values, which aligns with the general population distribution of BMI. Some points at higher BMIs also indicate significantly higher charges, which could correspond to obesity-related health conditions that are more costly to manage.

# Figure 2: Relationship between BMI and Charges

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

sns.scatterplot(x=eca\_data['bmi'], y=eca\_data['charges'], alpha=0.6, color='green')

plt.title('Figure 2: Scatter Plot of BMI vs. Medical Charges')

plt.xlabel('Body Mass Index (BMI)')

plt.ylabel('Charges')

plt.show()

**Insights for Graph 3:**

A graph of different colored squares

Description automatically generated

The bar chart showing the average medical charges by region indicates that there is regional variation in medical costs. Some regions have higher average charges than others, which could be due to a variety of factors, including cost of living, prevalence of certain health conditions, availability and use of medical services, and regional health care policies. This variation suggests that location can be a significant factor in the cost of medical care, which is an important consideration for healthcare providers and insurers when assessing risk and setting premiums.

# Figure 3: Average Charges by Region

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

average\_charges\_by\_region = eca\_data.groupby('region')['charges'].mean().sort\_values()

sns.barplot(x=average\_charges\_by\_region.index, y=average\_charges\_by\_region.values, palette="viridis")

plt.title('Figure 3: Average Medical Charges by Region')

plt.xlabel('Region')

plt.ylabel('Average Charges')

plt.show()

word count: 499

# **Question 3:**

* Data Preparation: The features used for the decision tree model excluded any direct identifiers (like 'PersonID') and the target variable 'smoker' itself. One-hot encoded variables related to the target were also removed to prevent data leakage.
* Train-Test Split: The dataset was split into a training set (70%) and a test set (30%) to ensure that the model's performance is validated on unseen data.
* Model Training: A Decision Tree Classifier was trained on the training data. This non-parametric algorithm is ideal for classification tasks and can handle both numerical and categorical data, by learning simple decision rules inferred from the data features.
* Model Evaluation: After training, the model was tested on the test set to evaluate its performance. Metrics such as accuracy, precision, recall, and the F1 score were considered.
* Tree Visualization: The decision tree was visualized to interpret how the model makes its decisions. Due to space constraints, the depth was limited to 2 levels for a clearer view.
* The model achieved an accuracy of approximately 95.77%, with high precision and recall for both classes (fig 2.1). This performance indicates that the decision tree could effectively use the given features to differentiate between smokers and non-smokers. The visualization provided insights into the most influential factors, with the top nodes showing the most critical splits based on the input features.

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

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy\_score, classification\_report

from sklearn import tree

# Prepare data

X = eca\_data\_encoded.drop(['smoker\_yes', 'PersonID'], axis=1) # Features

y = eca\_data\_encoded['smoker\_yes'] # Target

# Split data into training and test sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# Train the Decision Tree Classifier

dtc = DecisionTreeClassifier(random\_state=42)

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

# Predict test set

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

# Evaluate the model

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

report = classification\_report(y\_test, y\_pred)

# Visualize Decision Tree

plt.figure(figsize=(20,20))

tree.plot\_tree(dtc, max\_depth=2, filled=True, feature\_names=X.columns, class\_names=['Non-Smoker', 'Smoker'])

plt.show()

# Output accuracy and classification report

print("Accuracy:", accuracy)

print("Classification Report:")

print(report)

Word Count: 321

# **Question 4:**

A diagram of a tree

Description automatically generated

The full decision tree provides a detailed map of how the model determines whether an individual is likely to be a smoker based on features in the dataset. The tree repeatedly splits from the root into smaller subsets based on certain thresholds, aiming to separate smokers from non-smokers as cleanly as possible.

**Key insights:**

* The first few splits at the top of the tree are typically the most informative. They represent the features that best divide smokers from non-smokers according to the model's training.
* The features at the top splits can be considered the most important predictors for the target variable. In this case, those might be factors like 'age', 'bmi', 'charges', which are intuitively linked to smoking habits.
* Each path from the root to a leaf represents a set of conditions that lead to a prediction. By following these paths, we can interpret the model's logic, which may align with known risk factors for smoking or reveal unexpected patterns.
* The leaves of the tree indicate the final decision (smoker or non-smoker) and the distribution of the classes at that node, giving an idea of the model's certainty.

# Re-train Decision Tree Classifier

# This might happen due to the stateful environment reset

dt\_classifier\_full = DecisionTreeClassifier(random\_state=42)

dt\_classifier\_full.fit(X\_train, y\_train)

# plot decision tree

plt.figure(figsize=(20,20))

tree.plot\_tree(dt\_classifier\_full, filled=True, feature\_names=X.columns, class\_names=['Non-Smoker', 'Smoker'], fontsize=10)

plt.title('Full Decision Tree Visualization')

plt.show()

Word Count: 224

# **Question 5:**

Yes, decision trees can be effectively used for exploratory data analysis (EDA), extending beyond their conventional role in prediction. In EDA, decision trees serve as a valuable tool for uncovering patterns, relationships, and structure within the data that may not be immediately apparent (Loh, 2011).

Their visual nature allows for an intuitive representation of the data-generating process, making it easier to communicate findings to non-technical stakeholders (Kuhn & Johnson, 2013). By examining how variables are split at each node, analysts can infer the relative importance of different features and detect interactions between variables (James et al., 2013).

Moreover, decision trees can help identify outliers and anomalies, as these will often result in leaves with very few instances or very different target values compared to the rest of the data (Hastie et al., 2009). This aspect of decision trees is particularly useful in domains like fraud detection where such insights are critical.

However, one should note that decision trees might introduce bias if the dataset is imbalanced or has missing values, which can lead to misleading insights if not properly accounted for (Molnar, 2020). Therefore, while decision trees are a powerful tool for EDA, their findings should be supplemented with other methods and domain knowledge to ensure robustness.

In summary, decision trees are not only predictive models but also exploratory tools that can reveal complex structures and guide further analysis. When used appropriately, they are a versatile addition to the data scientist's toolkit.

Word Count: 242

**References**

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

Loh, W.-Y. (2011). Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), 14-23.

Molnar, C. (2020). Interpretable Machine Learning. <https://christophm.github.io/interpretable-ml-book/>.

**Appendix**

A number of numbers on a white background

Description automatically generated

*Fig 1.1 summary of number of missing data*

A screenshot of a computer

Description automatically generated

*Fig 1.2 Seperation of data for greater accuracy*

A screenshot of a diagram

Description automatically generated

*Fig 2.1 Decision tree with accuracy report*