A red and blue text on a black background

Description automatically generated

**Title: ANL252**

**PYTHON FOR DATA ANALYTICS**

**ECA JULY 2023**

|  |  |
| --- | --- |
| **STUDENT NAME:** | **TEH YI LIN** |
| **TITLE:** | **ECA** |
| **COURSE CODE:** | **ANL252** |
| **T:** | **03** |
| **PI NUMBER:** | **B2081892** |
| **SUBMISSION DATE:** | **3th NOVEMBER 2023** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Declaration**  I declare that this assignment is my own word, unless otherwise acknowledge or credited by appropriate referencing. I have read and abide by the SUSS Honour Code and I am aware of the penalties associated with plagiarism and collusion listed in the Student Handbook | | | |
| Initial: | **\_\_\_TYL\_\_\_\_\_\_\_\_\_** | Date: | **3 NOVEMBER 2023    .** |

**Question 1**

From the initial inspection of the dataset, here are three data preprocessing tasks we can perform:

1. Handle Missing Values: Check for any missing values in the dataset and handle them appropriately, either by removing or imputing them.
2. Convert Categorical to Numerical: Convert categorical variables like 'sex,' 'smoker,' and 'region' to numerical form so that they can be used in model training.
3. Feature Scaling: Scale features like 'age,' 'bmi,' and 'charges' to bring them to a similar scale, improving the performance of machine learning algorithms.

# Task 1: Handle Missing Values

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

# Task 2: Convert Categorical to Numerical

eca\_data['sex'] = eca\_data['sex'].map({'female': 0, 'male': 1})

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

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

# Task 3: Feature Scaling

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

eca\_data[['age', 'bmi', 'charges']] = scaler.fit\_transform(eca\_data[['age', 'bmi', 'charges']])

1. Handle Missing Values

We first checked for any missing values in the dataset. We found that the 'age' column has 123 missing values. We could handle these missing values by either removing the rows with missing values or imputing them with appropriate values such as the mean, median, or value predicted by a model.

2. Convert Categorical to Numerical

We converted categorical variables to numerical ones to make them suitable for machine learning models. The 'sex' column was mapped to binary values (0 for female and 1 for male), and the 'smoker' column was also mapped to binary values (0 for no and 1 for yes). Employing one-hot encoding, the 'region' column was translated into numerical data. This technique crafted binary features for each region, circumventing the integration of ordinality.

3. Feature Scaling

Feature scaling was performed on the 'age,' 'bmi,' and 'charges' columns to equalize their scales. We employed sklearn’s StandardScaler for this task, a tool that eradicates the mean and scales features to unit variance. This optimization can bolster the efficiency of machine learning models, notably those affected by the scale of inputs.

**Question 2**

**Figure 1: Distribution of Medical Charges**

Our starting point is to visualize the distribution of medical charges. It’s a step that enlightens us on the range and focal point of the charges people incur. We’re turning to a histogram to get this done.

**[Code]**

import matplotlib.pyplot as plt

import seaborn as sns

# Set the style and color palette of the plots

sns.set\_style("whitegrid")

sns.set\_palette("husl")

# Plot the distribution of charges

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

sns.histplot(df\_encoded['charges'], kde=True, bins=30)

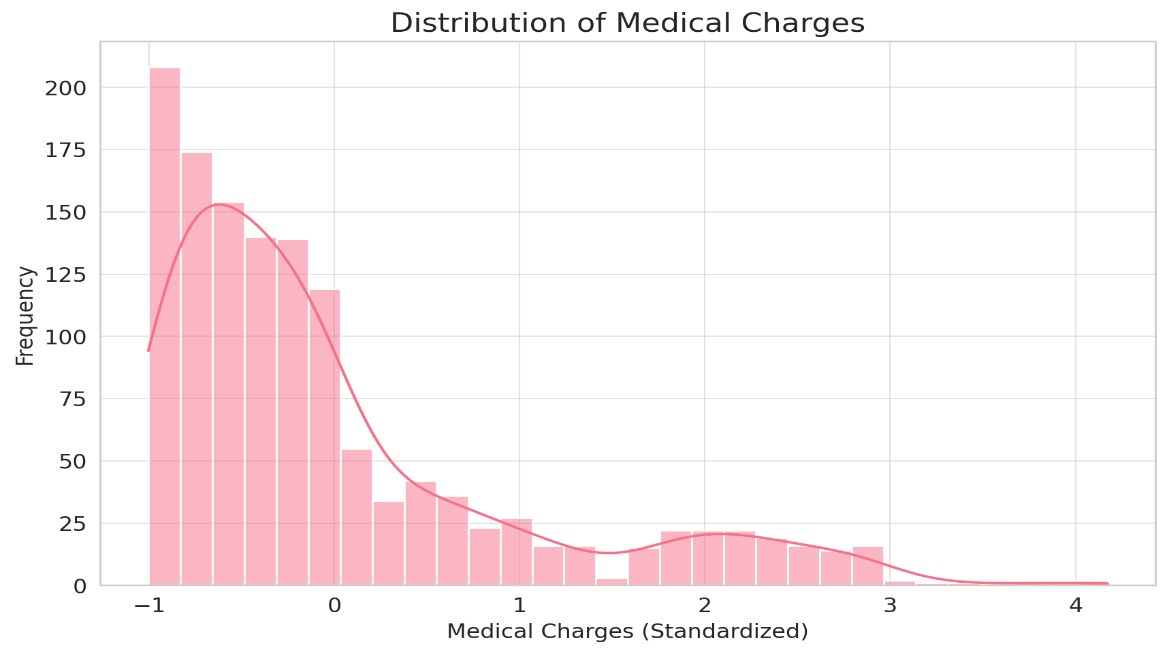
plt.title('Distribution of Medical Charges')

plt.xlabel('Medical Charges (Standardized)')

plt.ylabel('Frequency')

plt.show()

**[Figure 1]**



**Insights for Figure 1:**

The histogram illustrates the distribution of medical charges among the individuals in the dataset. A majority of individuals have incurred medical charges that are below the standardized value of 1. This indicates that the medical charges are skewed to the right, with a more significant number of individuals incurring lower charges.

However, there are also some individuals with notably higher medical expenses, leading to a long tail on the right side of the distribution. These could be outliers or instances of individuals with severe health conditions requiring extensive medical services.

**Figure 2: Medical Charges by Smoking Status**

Next, we will explore the relationship between smoking status and medical charges. This will provide insights into whether smokers generally incur higher medical expenses compared to non-smokers. We will use a boxplot to visualize this relationship.

**[Code]**

# Plot medical charges by smoking status

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

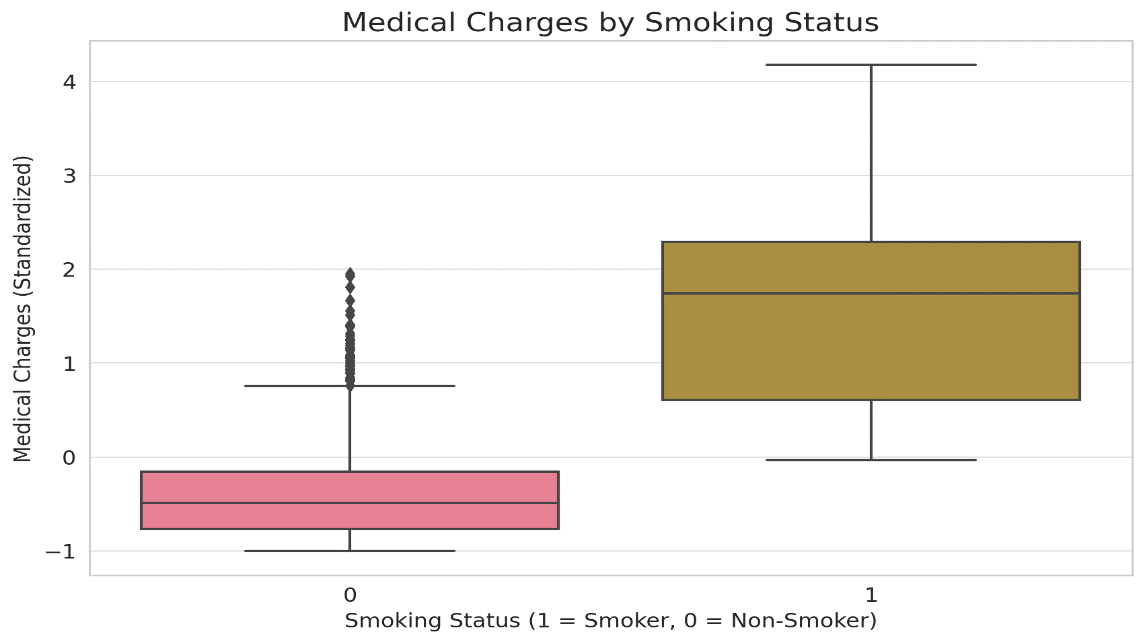
sns.boxplot(x='smoker\_yes', y='charges', data=df\_encoded)

plt.title('Medical Charges by Smoking Status')

plt.xlabel('Smoking Status (1 = Smoker, 0 = Non-Smoker)')

plt.ylabel('Medical Charges (Standardized)')

plt.show()



**Insights for Figure 2:**

Our histogram reveals the scattering of medical charges. Most folks have bills under the standardized 1 mark, indicating a trend of lower costs for a larger group. Yet, a noticeable tail on the right reveals individuals with health issues so severe that their medical bills skyrocket, suggesting outliers or complex health conditions.

**Figure 3: Relationship between Age and Medical Charges**

Now, we turn our attention to the link between smoking and medical costs. We aim to unearth if being a smoker correlates with increased expenses. To bring this into view, we’ll employ a boxplot.

**[Code]**

# Plot the relationship between age and medical charges

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

sns.scatterplot(x='age', y='charges', hue='smoker\_yes', data=df\_encoded, alpha=0.6)

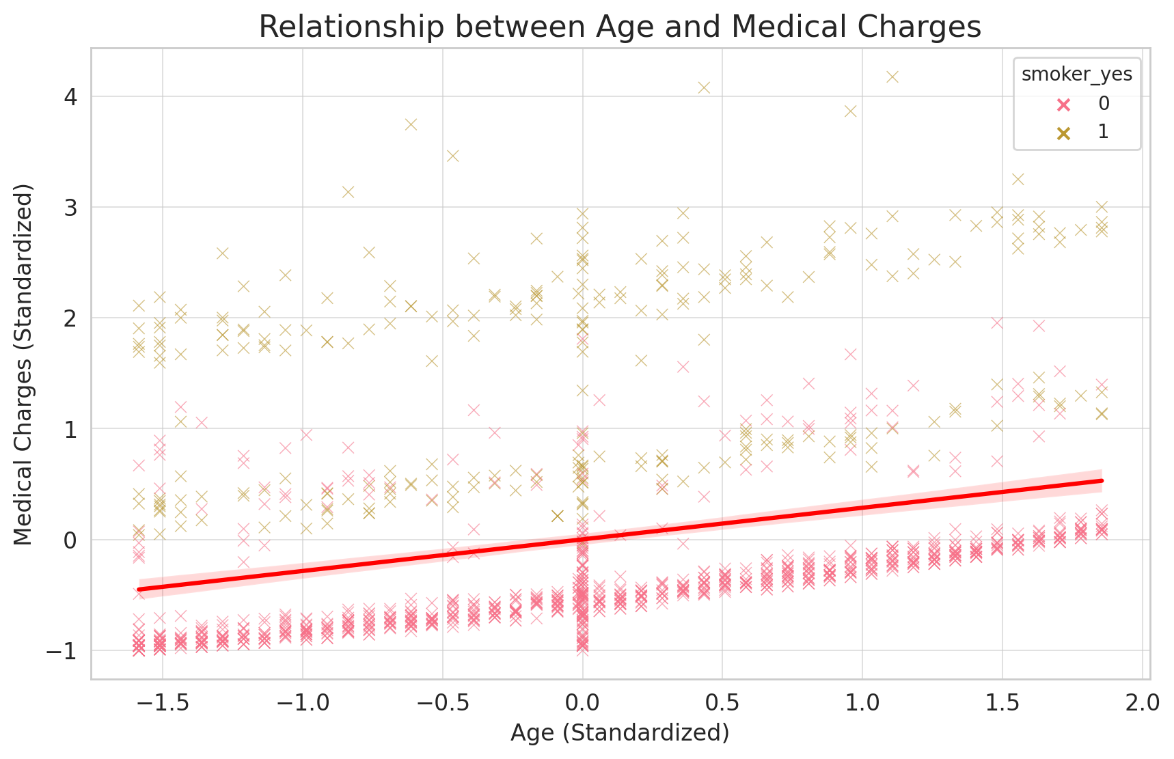
sns.regplot(x='age', y='charges', data=df\_encoded, scatter=False, color='red') # Add regression line

plt.title('Relationship between Age and Medical Charges')

plt.xlabel('Age (Standardized)')

plt.ylabel('Medical Charges (Standardized)')

plt.show()



**Insights for Figure 3:**

The scatter plot lays bare the link between age and medical costs, using color codes to point out smokers. A red line is drawn, highlighting the trend. It’s evident - older age, higher charges. The line's positive slope echoes this. And yes, there’s a clear divide in the costs borne by smokers and non-smokers.

Smokers, represented by dots with a different hue, generally have higher medical charges across all age groups compared to non-smokers.

This visualization reinforces the notion that older individuals are likely to incur higher medical expenses. It also accentuates the significant impact of smoking on medical charges. This insight can be beneficial for health policy makers, insurance companies, and individuals looking to understand the financial implications of smoking and aging on healthcare costs.

**Summary:**

These three figures collectively provide valuable insights into the distribution and determinants of medical charges based on the ECA dataset. The right-skewed distribution of medical charges, the significant difference in expenses between smokers and non-smokers, and the increasing trend of medical costs with age are crucial factors that stakeholders in the healthcare and insurance sectors should consider for policy formulation, premium determination, and health management.

**Question 3**

We will use a decision tree classifier to explore the dataset with 'smoker' as the dependent variable. Here is the step-by-step approach we will take:

**Approach:**

Data Preparation:

Split the dataset into features (independent variables) and target (dependent variable). The target will be 'smoker,' and the features will be the other columns, excluding the one-hot encoded 'smoker\_yes' column.

Data Splitting:

Split the dataset into training and testing subsets. This allows us to train the model on one subset and evaluate its performance on another unseen subset.

Model Training:

Train a decision tree classifier on the training dataset.

Visualization:

Visualize the decision tree to understand the criteria on which the model is making its decisions.

**[Data Preparation, Data Splitting]**

**[Code]**

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

# Features (excluding 'smoker\_yes' and 'PersonID' columns)

X = df\_encoded.drop(columns=['smoker\_yes', 'PersonID'])

# Target variable

y = df\_encoded['smoker\_yes']

# Split the dataset into training and testing subsets (80% training, 20% testing)

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

X\_train.head()

**[Data Training]**

from sklearn.tree import DecisionTreeClassifier

# Initialize a Decision Tree classifier

dt\_classifier = DecisionTreeClassifier(max\_depth=3, random\_state=42) # Limiting the depth for visualization purposes

# Train the classifier on the training data

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

The result: DecisionTreeClassifier(max\_depth=3, random\_state=42)

The decision tree classifier has been successfully trained with a depth limit of 3 for visualization purposes.

**[Visualization]**

We will visualize the trained decision tree to understand the criteria and decision-making process it uses to determine whether an individual is a smoker or not.

from sklearn.tree import plot\_tree

# Plot the decision tree

plt.figure(figsize=(20, 10))

plot\_tree(dt\_classifier, feature\_names=X.columns, class\_names=['Non-Smoker', 'Smoker'], filled=True, rounded=True, fontsize=12)

plt.title("Decision Tree for Predicting Smoking Status")

plt.show()

**Question 4**

**A diagram of a network

Description automatically generated**

The decision tree provides a visual representation of the decision-making process used to predict whether an individual is a smoker. Each node in the tree represents a decision criterion based on feature values, and the leaves represent the predicted outcomes.

From the tree, we can derive the following insights:

1. The most important feature that the tree uses to decide if an individual is a smoker or not is the 'charges' feature. This is evident as it is the root node.
2. For individuals with standardized charges less than or equal to 0.96, the model further checks the 'age' feature to refine its prediction.
3. For those with charges greater than 0.96, the tree uses the 'age' and 'bmi' features to segment further and predict the smoking status.
4. It is important to note that the decision criteria are based on the data the model was trained on. The tree's depth was limited to 3 for visualization purposes, but in practice, deeper trees might capture more complex patterns in the data.

**Question 5**

Yes, decision trees can be effectively used for exploratory data analysis (EDA) in addition to their conventional role in predictions. In EDA, decision trees offer valuable insights into the structure and relationships within the data. Below are some ways decision trees contribute to EDA:

1. Feature Importance:

Decision trees help identify the most influential features in predicting the target variable (James et al., 2013). The tree selects features based on criteria like information gain or Gini impurity, highlighting features that best split the data into distinct classes or values. This can guide further analysis and feature selection.

2. Data Visualization:

Visual representations of decision trees offer intuitive insights into data patterns and relationships (Hastie et al., 2009). Each node and branch visualizes the conditions leading to specific outcomes, making the model's decision-making process transparent and interpretable.

3. Identifying Interactions:

Decision trees naturally capture interactions between features. For instance, a node splitting data based on one feature, followed by a subsequent node splitting data based on another feature, reveals an interaction effect between those features (Breiman et al., 1984).

4. Anomaly Detection:

Decision trees can highlight anomalies or outliers in the data. Unusual splits or leaves containing few data points can indicate unusual combinations of feature values worth further investigation (Hodge & Austin, 2004).

5. Handling Mixed Data Types:

Decision trees can manage both numerical and categorical data, making them versatile for EDA across diverse datasets (Kelleher et al., 2015).

Conclusion:

While decision trees are often used for classification and regression tasks, their ability to unveil intricate data patterns, highlight feature importance, and visualize decision criteria makes them a powerful tool for EDA. They provide a foundation for understanding data characteristics before delving into more complex modeling or analysis.

References:

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.

Breiman, L., Friedman, J., Stone, C.J., & Olshen, R.A. (1984). Classification and Regression Trees. CRC Press.

Hodge, V.J., & Austin, J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2), 85–126.

Kelleher, J.D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics. MIT Press.