

## Assignment Title Page

|  |  |
| --- | --- |
| **STUDENT NAME:** | Hilmi Bin Ishak |
| **P.I. NO.:** | N2181344 |
| **TITLE:** | ECA |
| **COURSE CODE:** | ANL252 |
| **TG:** | T03 |
| **SUBMISSION DATE:** | 01/11/2023 |

**Plagiarism and Collusion**

**Plagiarism** is the act of using or passing as one’s own, the ideas or writings of another without acknowledging or crediting the source from which the ideas are taken from.

**Collusion** is the act of submitting any academic work (including assignment, project or report) that was completed by another person and pass these work off as one’s own.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
| **Declaration** | | | | |
|  | | | | |
| I declare that this assignment is my own work, 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: | Hilmi Ishak |  | Date: | 01/11/2023 |
|  | | | | |

**Question 1**

Removal of missing data

# Importing of dataset

import pandas as pd

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

meddata

# Identifying missing values from dataset

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

# Creation of new dataset for data prep

prepmed = meddata.copy()

# Checking of unique data in age column

prepmed['age'].unique()

# Removal of nan in age column and verification

dprepmed = prepmed.dropna(axis = 0, how = 'any')

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

Prior to the start of data prep, we firstly would need to import the dataset using Pandas. Upon importing the dataset, we can then check for missing values in the dataset using the isnull() function. Before proceeding on to remove the missing values, we can create a new copy of the dataset so that we do not affect the original data itself and is able to make another copy if we are needed to restart our work. After a copy of the dataset is made, we then execute the dropna function on the Age column to remove missing values and do our verification by running the isnull() function again. We are using the dropna function to remove our missing values as averaging the age would give us a non-realistic age that would prove to be not useful for analysis. Removing missing values in this case would then allow the dataset to be more useful and allow us to have a better quality analysis of the data.

Rounding up values

# Rounding up charges data

roundprepmed = dprepmed.round({"charges":2})

The next data preparation that we can do to the dataset would be to the charges column. We would be prepping the data by converting the values to have 2 decimal places. The reason for such is because the values in that column beforehand had values with varying decimal points, some ranging from none to 5 decimal places. Therefore by setting each value to have 2 datasets, it would introduce uniformity in the values for said column. There is also no need to have such a varying amount of decimals and it also depends on the measurement. (Lydersen & Skovlund, 2023)

Converting of values

# Converting values from sex column

roundprepmed['sex'] = roundprepmed['sex'].replace(['F', 'M'], ['female', 'male'])

roundprepmed

The final data preparation that we are doing is to convert the values found in the ‘sex’ column of the dataset. In this column, the values are set to either male, female, F or M. To again introduce uniformity and also to allow for a more meaningful analysis, we then convert the F and M values in the column into female and male accordingly. For this, we would be utilizing the replace function in order to convert said data.

(438 words including in-text citations)

**Question 2**

BMI based on gender

#importing of matplotlib

import matplotlib.pyplot as plt

# filtering the data according to male and female

fig1 = roundprepmed['bmi'].loc[(roundprepmed['sex'] == 'female') & (roundprepmed['age'] > 18)]

fig2 = roundprepmed['bmi'].loc[(roundprepmed['sex'] == 'male') & (roundprepmed['age'] > 18)]

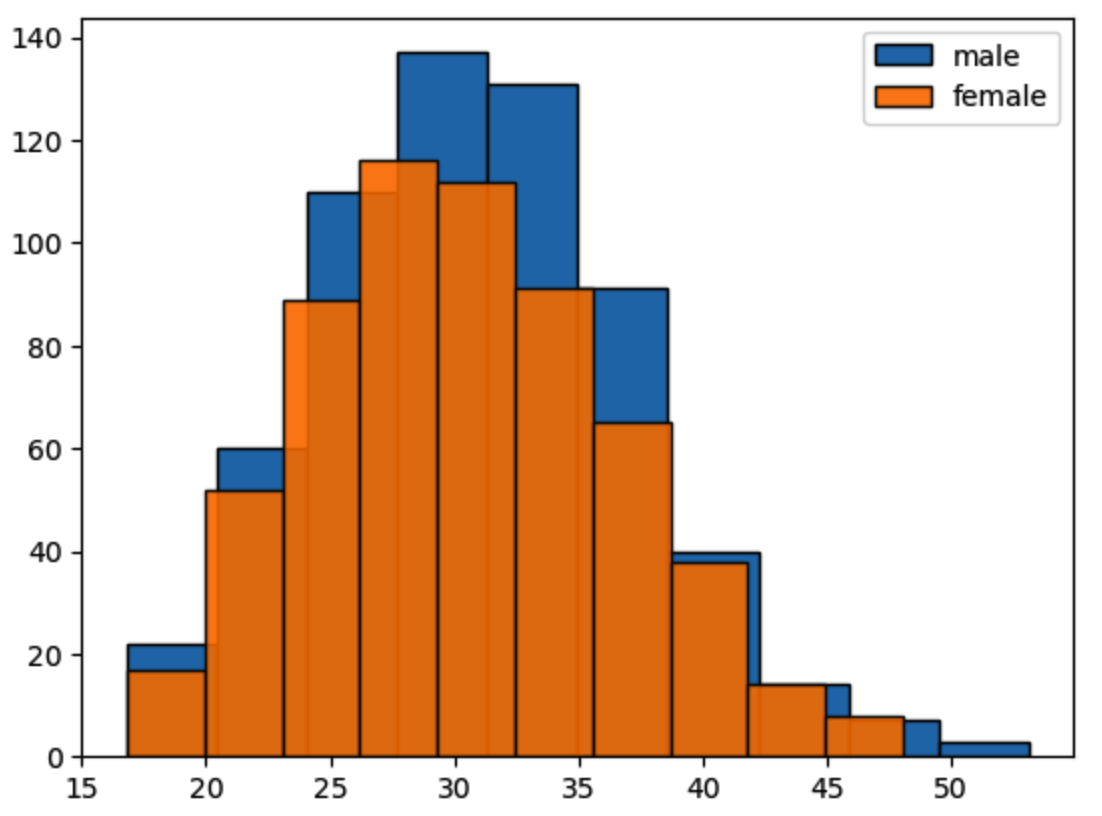
# creating histogram with formatting

plt.hist(fig2, label = 'male', edgecolor = 'black')

plt.hist(fig1, label = 'female', alpha = 0.9, edgecolor = 'black')

plt.legend()

plt.show()

****

From the output shown above, we can see that the males have a higher BMI than the female. The healthy BMI range would be between 18.5 to 24.9 (Centers for Disease Control and Prevention, n.d.), and as seen from the figure above, there are more males than females that are in the healthy BMI range. The BMI ranges that falls under the 25 and above ranges are then considered overweight with those falling under BMI 30 and above are then considered obese. We can then see that the males have a much higher number that are obese and overweight as compared to females in the US.

Box and whiskers for BMI based on Region

# import matplotlib

import matplotlib.pyplot as plt

# setting of data

data\_1 = roundprepmed['bmi'].loc[roundprepmed['region'] == 'southwest']

data\_2 = roundprepmed['bmi'].loc[roundprepmed['region'] == 'southeast']

data\_3 = roundprepmed['bmi'].loc[roundprepmed['region'] == 'northwest']

data\_4 = roundprepmed['bmi'].loc[roundprepmed['region'] == 'northeast']

alldata = [data\_1, data\_2, data\_3, data\_4]

fig = plt.figure(figsize =(10, 7))

# creation of plot

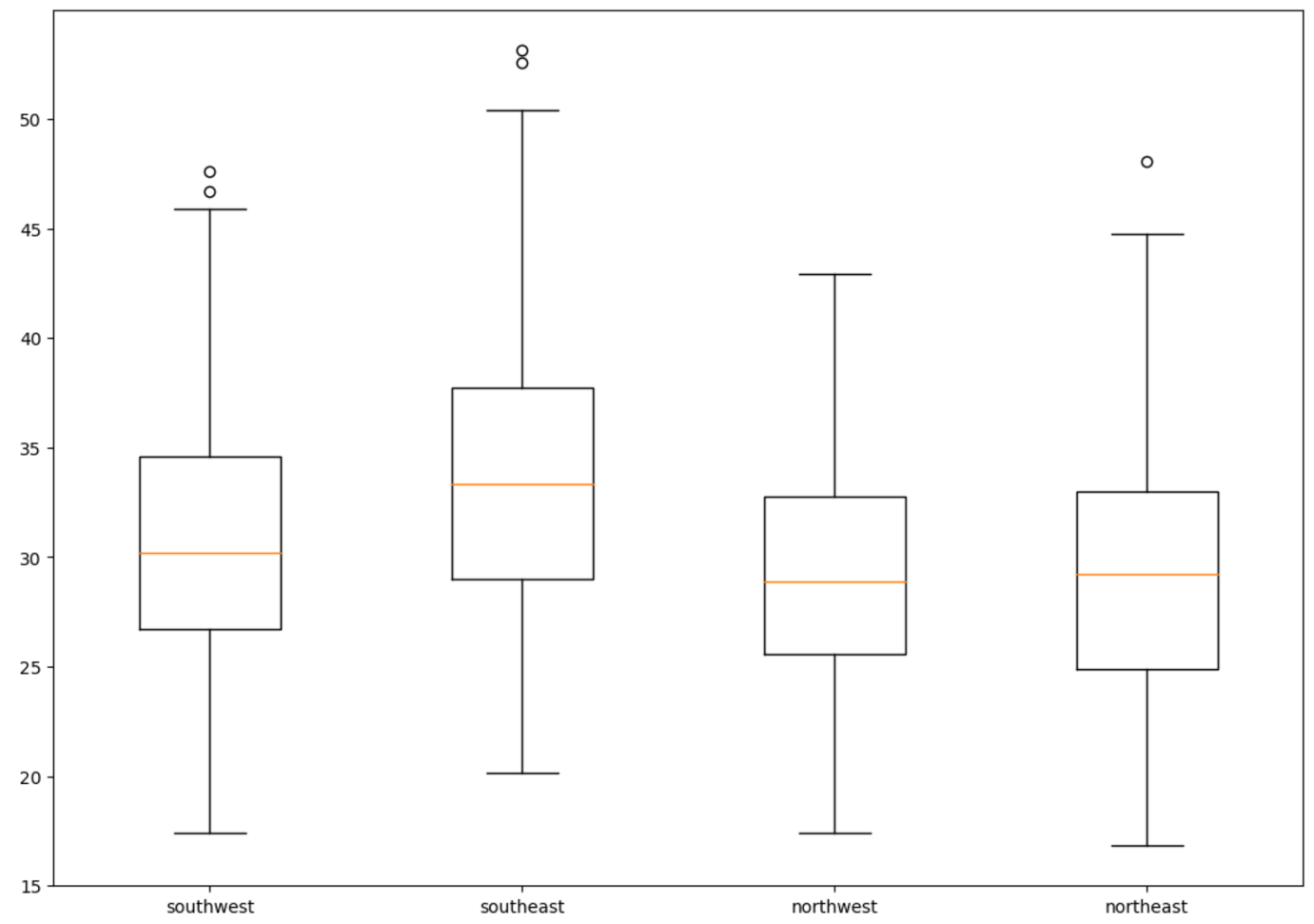
ax = fig.add\_axes([1, 1, 1, 1])

bp = ax.boxplot(alldata)

# show plot with renaming

plt.xticks([1, 2, 3, 4], ['southwest', 'southeast', 'northwest', 'northeast'])

plt.show()



The Southeast region has the highest median than the other regions, which means that on average, there are a higher number of people with high BMI, or are obese, in that region as compared to the rest. The ranges of each region is roughly similar however Northwest has a smaller box plot which mean that its data is less dispersed than the others. There are also outliers present for all the regions except for Northwest, which indicates that the distributions of the regions with said outliers could have a skewed distribution and this is clearly seen in the Southwest region. This then affects the mean of the data for said region, indicating that said region also has a high number of obese people.

Average Charges by Region

import matplotlib.pyplot as plt

d1 = roundprepmed['charges'].loc[roundprepmed['region'] == 'southwest'].mean()

d2 = roundprepmed['charges'].loc[roundprepmed['region'] == 'southeast'].mean()

d3 = roundprepmed['charges'].loc[roundprepmed['region'] == 'northwest'].mean()

d4 = roundprepmed['charges'].loc[roundprepmed['region'] == 'northeast'].mean()

region = roundprepmed['region'].unique()

values = [d1, d2, d3, d4]

fig = plt.figure(figsize = (10, 5))

# creating the bar plot

plt.bar(region, values, color ='maroon',

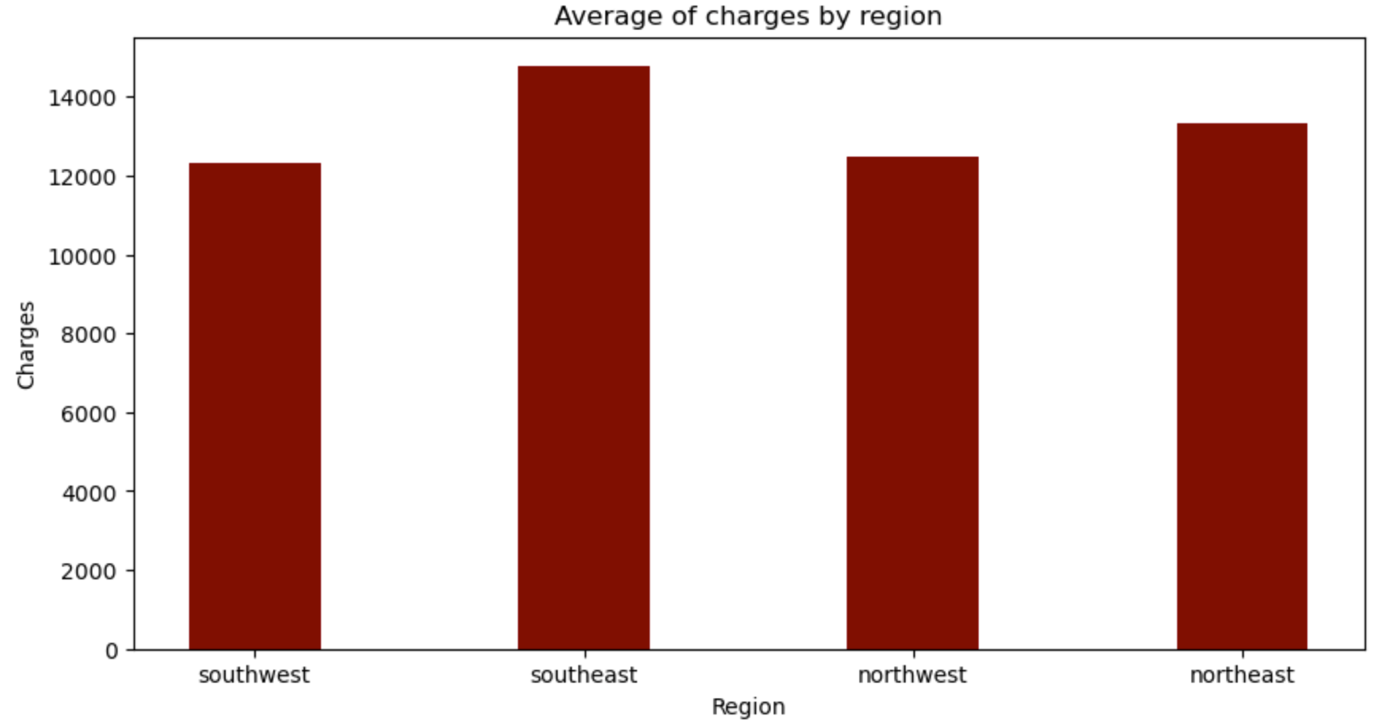
width = 0.4)

plt.xlabel("Region")

plt.ylabel("Charges")

plt.title("Average of charges by region")

plt.show()



From this bar chart, it is clear that the Southeast region has the highest average of charges as compared to the rest. From the box plot earlier, it shows that there are more obese people in said region as compared to the rest. This could mean that there could be higher charges for those people in that region that could have health complications due to their obesity.

(498 words including in-text citation)

**Question 3**

# importing of libraries

import sys

import matplotlib

%matplotlib inline

import pandas

from sklearn import tree

from sklearn.tree import DecisionTreeClassifier

import matplotlib.pyplot as plt

# creation of copy of dataset

dtprepmed = roundprepmed.copy()

# converting values to numerical and selecting columns

p1 = {'no': 0, 'yes': 1}

p2 = {'female': 0, 'male': 1}

dtprepmed['smoker'] = dtprepmed['smoker'].map(p1)

dtprepmed['sex'] = dtprepmed['sex'].map(p2)

columns = ['sex', 'children']

x = dtprepmed[columns]

y = dtprepmed['smoker']

# fitting the decision tree

decisiont = DecisionTreeClassifier()

decisiont = decisiont.fit(x, y)

# creating the decision tree

plt.figure(figsize=(35, 25))

tree.plot\_tree(decisiont, feature\_names = columns, fontsize = 12)

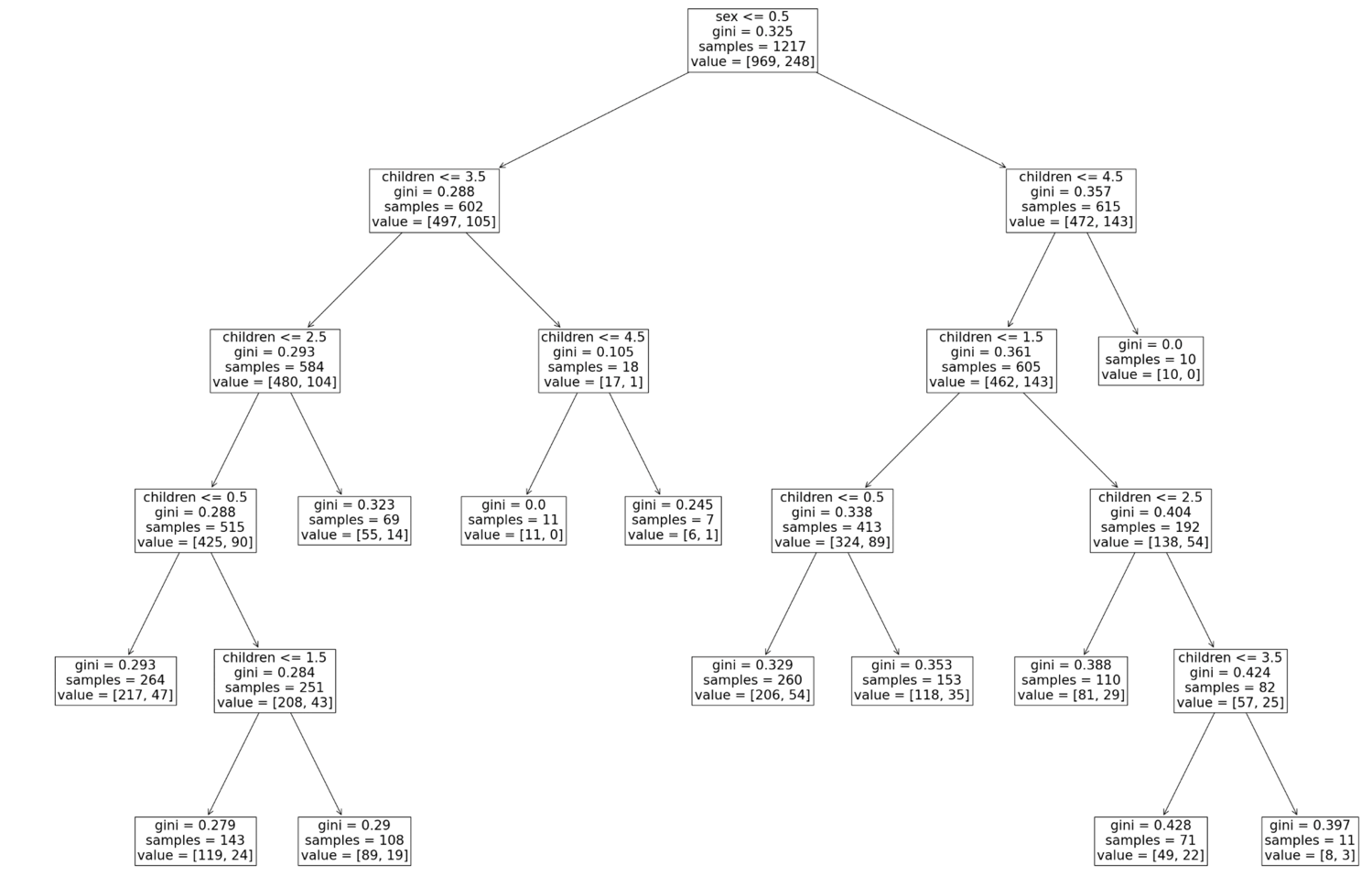
plt.show()

Upon creating a copy of the dataset for the creation of a decision tree, we then convert the values of the smoker column as well as sex column into numerical values. This is due to their values being either Yes or No as well as Male or Female. We then fit the data accordingly to which we use the columns sex and children for our feature columns. The children column already has numerical values thus there is no need to convert its values.

The reason on why we are using children is due to the fact that smoking has harmful effects to others around them. Children affected by second hand smoke can have harmful side effects such as infections to their ear or to their respiratory systems. (Centers for Disease Control and Prevention, n.d.) If the females are the ones who smoke, it may also affect child birth with there being a risk of birth defects. (SingHealth, n.d.) Thus, from the decision tree, we are seeing on how likely it is for someone to smoke, starting with their gender and the number of children they have.

(282 words with in-text citations)

**Question 4**



Starting from the top, out of the entire population of the data, only 248 would be smokers. However, once split into male and female, we can see that for the male side, majority of the population would still consider smoking and is evident if they have less than 4 children. It can be seen that the bulk of the males that would consider smoking would be those that do not have children but despite so, there are still some that would still smoke when they have children.

As for the females, there are lesser that would consider smoking again with the bulk of them considering to be smokers if they do not have children. The number that would consider smoking if they do not have children are similar to those of the males.

(133 words)

**Question 5**

Decision trees can be effectively used for exploratory data analysis. They give us the visuals needed to see the relationship between variables in our dataset. This makes handling a large dataset easier as a decision tree allows us to see which variables are significant and also allows us to see the significance of the relationships between different numbers of variables. (Master's in Data Science, n.d.)

A decision tree is also easy to understand (Master's in Data Science, n.d.) and thus, it can be used effectively as people who have not much experience in analytics could also be able to understand a decision tree and derive insights on their own. Decision trees also requires less data preparation (Hillier, 2023) which means that insights can be derived from decisions trees quicker as compared to other data modelling techniques.

However, decision trees is prone to overfitting, whereby it can’t handle too many variables thrown at it. (Hillier, 2023) This requires specific selections of the variables wanted in order not to complicate the design so that we can achieve a meaningful analysis from said tree. Decision trees are also unable to handle all types of data, in which it is not suited for data that are continuous. (Hillier, 2023)

(205 words with in-text citations)

# Bibliography

Lydersen, S., & Skovlund, E. (2023, May 5). *Do we need all the decimals?* Retrieved from Tidsskrifet: https://tidsskriftet.no/en/2023/05/do-we-need-all-decimals

Centers for Disease Control and Prevention. (n.d.). *Health Problems Caused by Secondhand Smoke*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/tobacco/secondhand-smoke/health.html#:~:text=In%20children%2C%20secondhand%20smoke%20exposure,infant%20death%20syndrome%20(SIDS).&text=Since%201964%2C%20about%202%2C500%2C000%20people,caused%20by%20secondhand%20smoke%20exposure.

Centers for Disease Control and Prevention. (n.d.). *About Adult BMI*. Retrieved from Centres for Disease Control and Prevention: https://www.cdc.gov/healthyweight/assessing/bmi/adult\_bmi/index.html

SingHealth. (n.d.). *Pregnancy and Smoking*. Retrieved from SingHealth: https://www.singhealth.com.sg/patient-care/conditions-treatments/pregnancy-and-smoking

Master's in Data Science. (n.d.). *What is a Decision Tree?* Retrieved from Master's in Data Science: https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/

Hillier, W. (2023, April 17). *What is a Decision Tree and How Is It Used?* Retrieved from CareerFoundry: https://careerfoundry.com/en/blog/data-analytics/what-is-a-decision-tree/