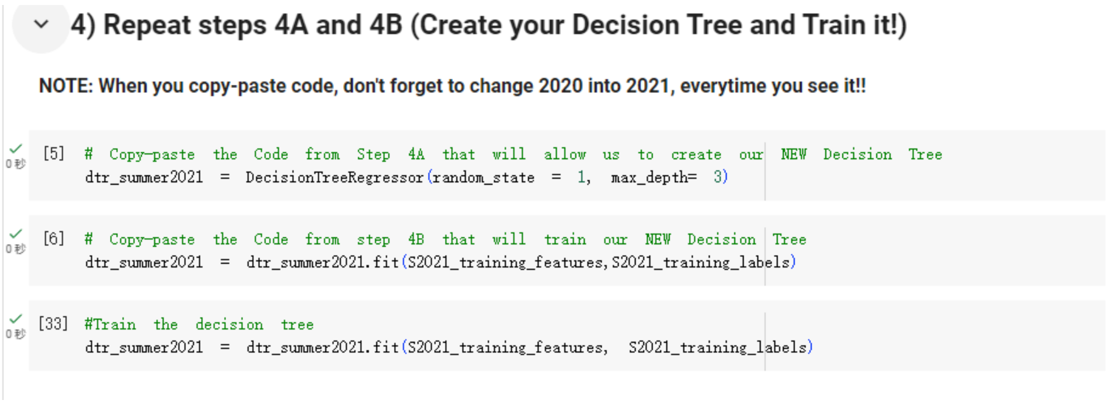
Modul 1 Coding Assignment Covid cases in California

Name: Po-Han Chen   Student ID:923446482

The code I added at Step 4 and Step 5(Practice part)



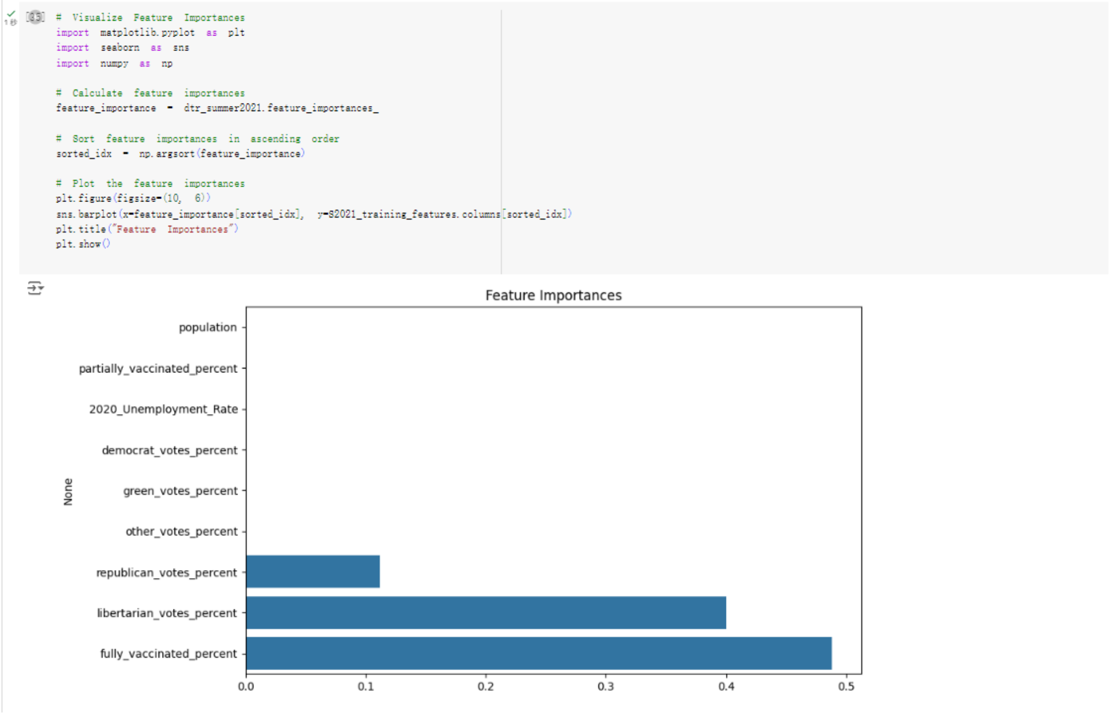
4) Repeat steps 4A and 4B (Create your Decision Tree and Train it!)

NOTE: When you copy-paste code, don't forget to change 2020 into 2021, everytime you see it!!

```
[5]  # Copy-paste the Code from Step 4A that will allow us to create our NEW Decision Tree
     dtr_summer2021 = DecisionTreeRegressor(random_state = 1, max_depth= 3)
```

```
[6]  # Copy-paste the Code from step 4B that will train our NEW Decision Tree
     dtr_summer2021 = dtr_summer2021.fit(S2021_training_features,S2021_training_labels)
```

```
[33] #Train the decision tree
     dtr_summer2021 = dtr_summer2021.fit(S2021_training_features, S2021_training_labels)
```



```
[35] # Visualize Feature Importances
     import matplotlib.pyplot as plt
     import seaborn as sns
     import numpy as np

     # Calculate feature importances
     feature_importance = dtr_summer2021.feature_importances_

     # Sort feature importances in ascending order
     sorted_idx = np.argsort(feature_importance)

     # Plot the feature importances
     plt.figure(figsize=(10, 6))
     sns.barplot(x=feature_importance[sorted_idx], y=S2021_training_features.columns[sorted_idx])
     plt.title("Feature Importances")
     plt.show()
```

Explain the code what I do:

After training the decision tree, you can visualize which features were the most important in making predictions. And use ‘ploty’ and ‘matplotlib’ to create a bar chart of feature importances.

A. Describe at least one main difference between the decision tree that was created with 2021 data and the 2020 decision tree.

The main difference is that the forecasts based on data from 2021 show significant differences in certain counties compared to the forecasts based on 2020 data. For example, counties like Alameda and Butte exhibit substantial discrepancies between the 2020 and 2021 data-driven predictions. The forecasts using 2021 data indicate a broader range and a more diverse distribution across the counties, unlike the more concentrated forecasts produced with 2020 data.

B. List at least three things you learned from the exercise and three things you'd like know more about.

Through this exercise, I learned three which are how to train a decision tree, how to separate data, and how to Visualize Feature Importance.

First of all, to create and train a decision tree model using Python, which involves feeding the data into the model to make predictions.

Second of all, I learned how to split the dataset into features which is inputs and labels which is outputs. This is important because the model needs to know what to learn from (features) and what to predict (labels).

Third of all, I learned the technique of creating a bar chart to visualize the most influential features or inputs for making predictions. This approach enables me to gain insight into what aspects the model is prioritizing.

The three things that I want to know more about that are improving model accuracy, different ways to measure model, and learning more visualization techniques.

About improving model accuracy, I'm wondering how to adjust its settings to improve that.

Besides I can just look at the results visually, I want to explore more methods to check how well the model is performing.

I want to learn more about visualization techniques, making data and model results visual much clearer.