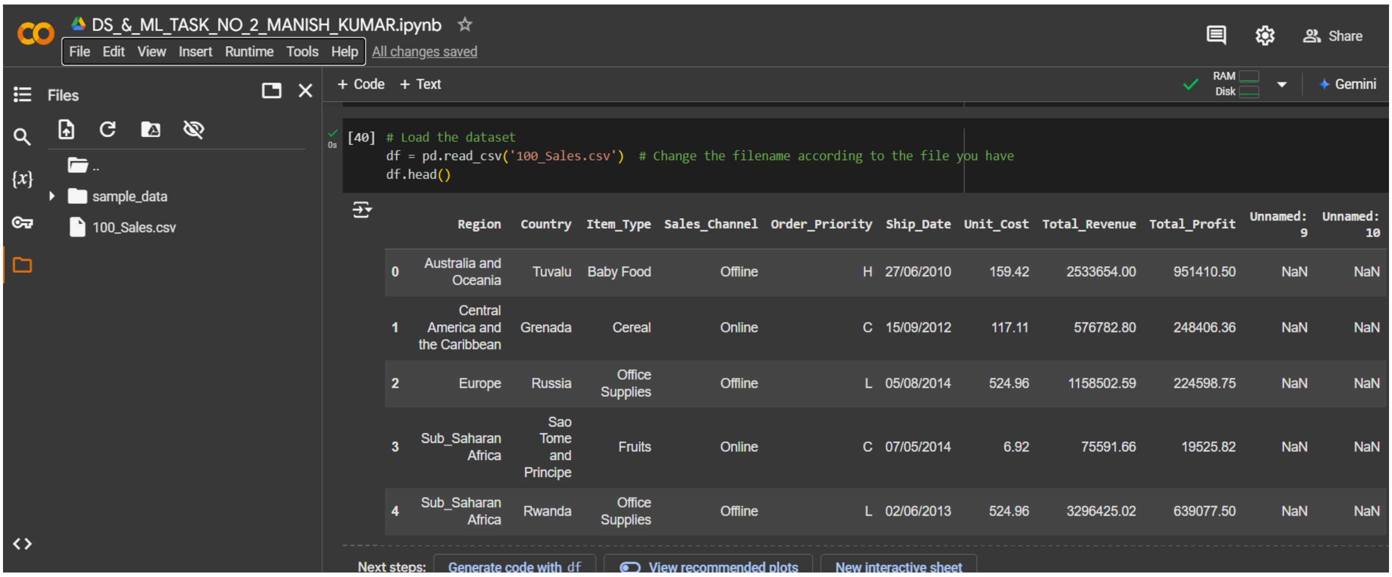
Task 2 Report
Classification with Decision Tree and Random Forest

1. Task Description

- The objective of this task is to apply two machine learning algorithms, Decision Tree and Random Forest, to predict the target variable Total_Revenue in the provided dataset.
- This dataset contains various features such as Region, Country, Item_Type, Unit_Cost, Total_Profit, and others that can influence sales revenue. The task involves several stages, including data preprocessing (handling missing values and encoding categorical variables), feature engineering, and model training.
- We will train both Decision Tree and Random Forest models to predict Total_Revenue and evaluate their performance using appropriate regression metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R²).
- Additionally, model optimization and evaluation will help determine the more accurate model for this prediction task.

2. Attach Screenshot of Output

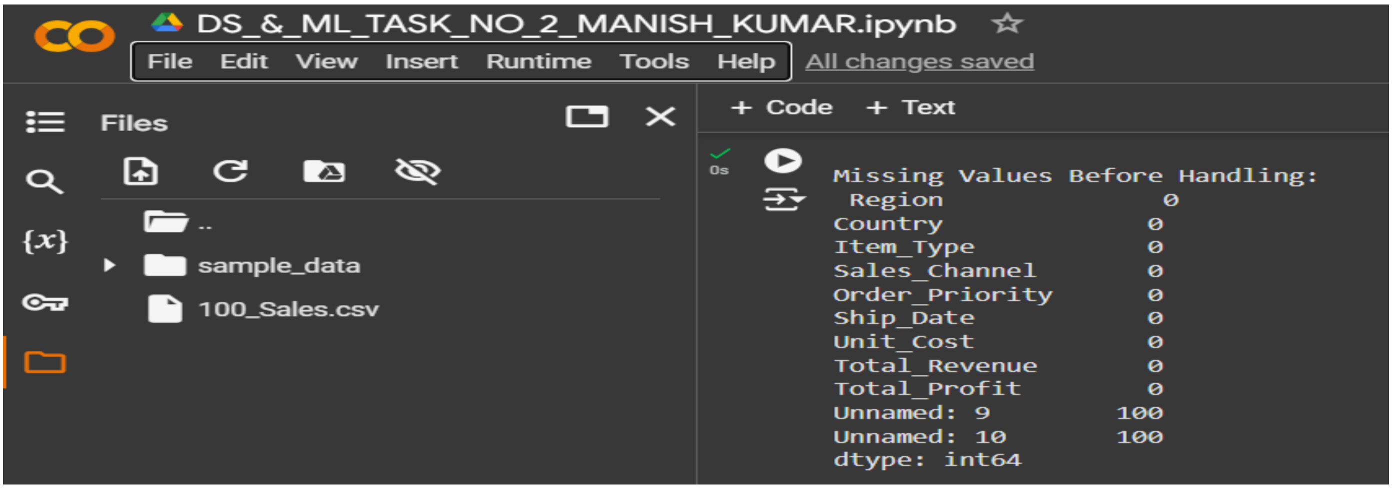
- **Dataset Preview:**


DS_&_ML_TASK_NO_2_MANISH_KUMAR.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Share
Files
+ Code + Text
RAM
Disk
Gemini
sample_data
100_Sales.csv
Next steps: Generate code with df View recommended plots New interactive sheet

```
[40] # Load the dataset
     df = pd.read_csv('100_Sales.csv')  # Change the filename according to the file you have
     df.head()
```

	Region	Country	Item_Type	Sales_Channel	Order_Priority	Ship_Date	Unit_Cost	Total_Revenue	Total_Profit	Unnamed: 9	Unnamed: 10
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	27/06/2010	159.42	2533654.00	951410.50	NaN	NaN
1	Central America and the Caribbean	Grenada	Cereal	Online	C	15/09/2012	117.11	576782.80	248406.36	NaN	NaN
2	Europe	Russia	Office Supplies	Offline	L	05/08/2014	524.96	1158502.59	224598.75	NaN	NaN
3	Sub_Saharan Africa	Sao Tome and Principe	Fruits	Online	C	07/05/2014	6.92	75591.66	19525.82	NaN	NaN
4	Sub_Saharan Africa	Rwanda	Office Supplies	Offline	L	02/06/2013	524.96	3296425.02	639077.50	NaN	NaN

- **Missing values before handling:**

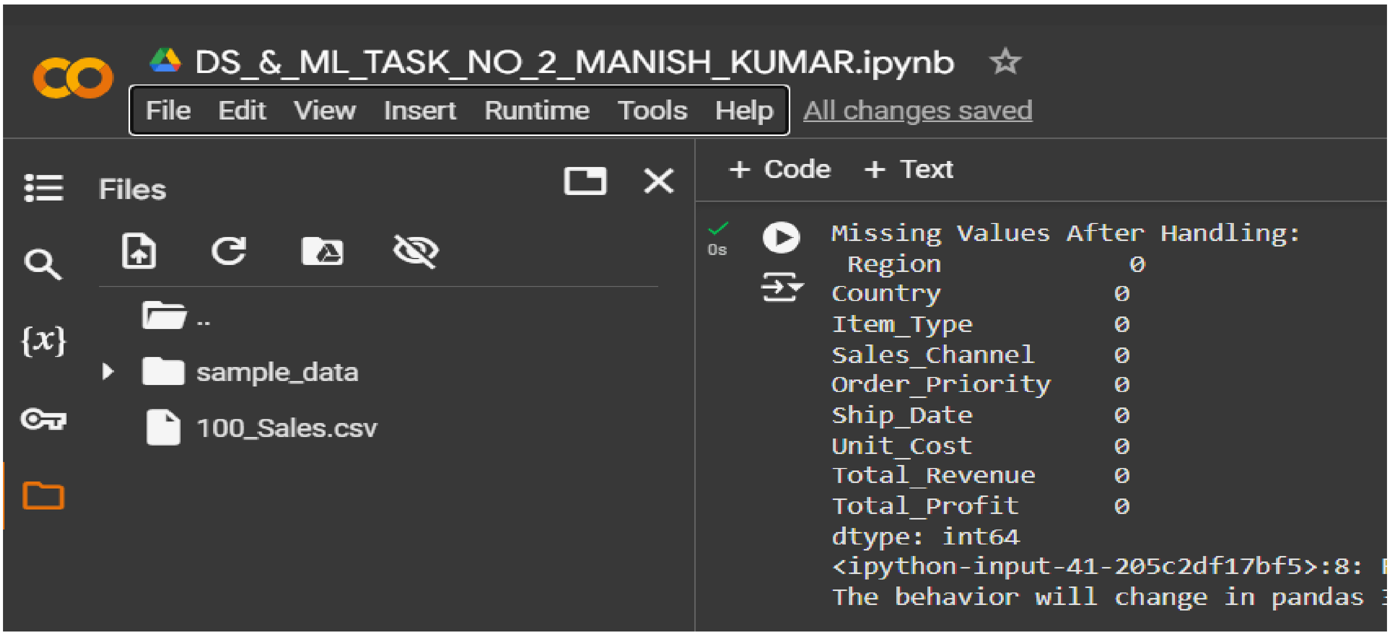

DS_&_ML_TASK_NO_2_MANISH_KUMAR.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Files
+ Code + Text
sample_data
100_Sales.csv

```
Missing Values Before Handling:
 Region             0
Country            0
Item_Type          0
Sales_Channel      0
Order_Priority     0
Ship_Date          0
Unit_Cost          0
Total_Revenue      0
Total_Profit       0
Unnamed: 9       100
Unnamed: 10      100
dtype: int64
```

Task 2 Report

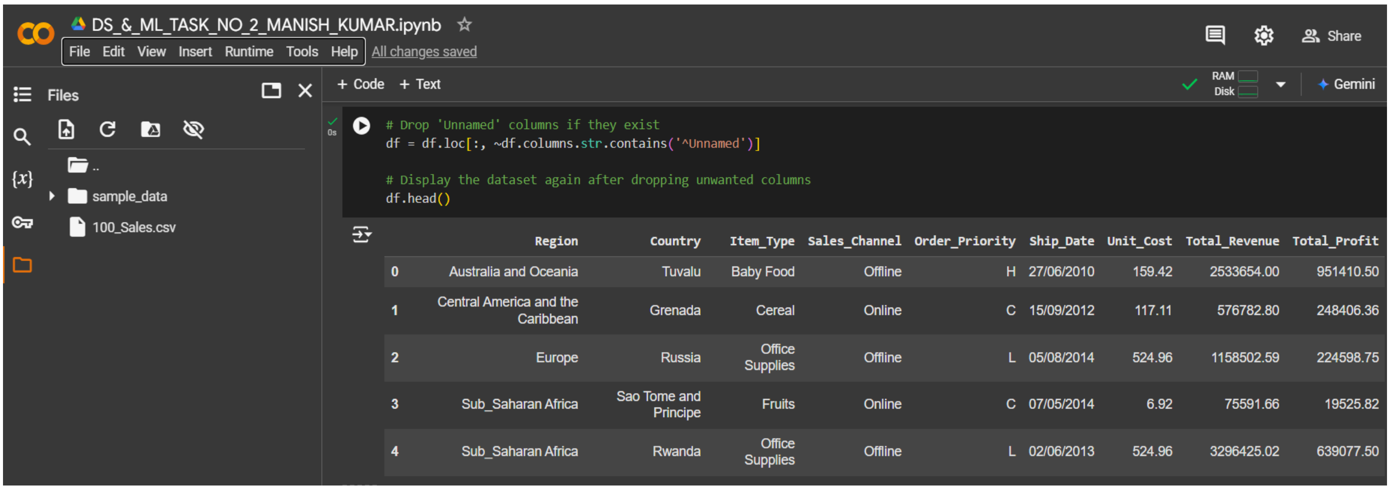
Classification with Decision Tree and Random Forest

❖ Missing values after handling:


DS_&_ML_TASK_NO_2_MANISH_KUMAR.ipynb
File Edit View Insert Runtime Tools Help
All changes saved
Files
+ Code + Text
sample_data
100_Sales.csv
0s

```
Missing Values After Handling:
 Region           0
Country          0
Item_Type        0
Sales_Channel    0
Order_Priority   0
Ship_Date        0
Unit_Cost        0
Total_Revenue    0
Total_Profit     0
dtype: int64
<ipython-input-41-205c2df17bf5>:8: F
The behavior will change in pandas 3
```

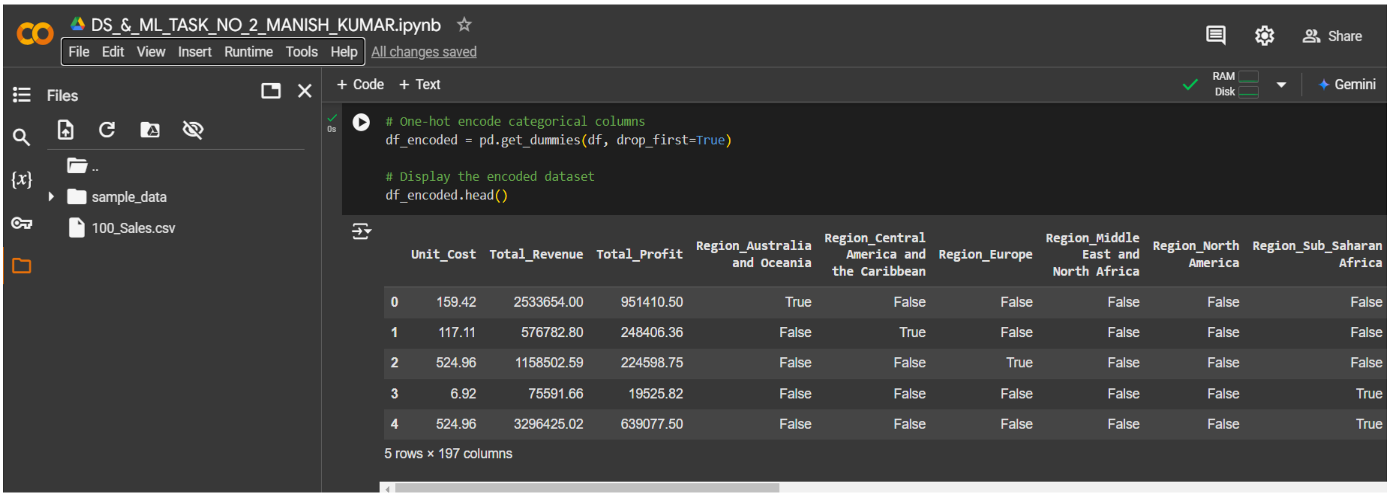
❖ Drop unnamed columns:



```python
# Drop 'Unnamed' columns if they exist
df = df.loc[:, ~df.columns.str.contains('^Unnamed')]

# Display the dataset again after dropping unwanted columns
df.head()
```

	Region	Country	Item_Type	Sales_Channel	Order_Priority	Ship_Date	Unit_Cost	Total_Revenue	Total_Profit
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	27/06/2010	159.42	2533654.00	951410.50
1	Central America and the Caribbean	Grenada	Cereal	Online	C	15/09/2012	117.11	576782.80	248406.36
2	Europe	Russia	Office Supplies	Offline	L	05/08/2014	524.96	1158502.59	224598.75
3	Sub_Saharan Africa	Sao Tome and Principe	Fruits	Online	C	07/05/2014	6.92	75591.66	19525.82
4	Sub_Saharan Africa	Rwanda	Office Supplies	Offline	L	02/06/2013	524.96	3296425.02	639077.50

❖ Encode categorical columns:



```python
# One-hot encode categorical columns
df_encoded = pd.get_dummies(df, drop_first=True)

# Display the encoded dataset
df_encoded.head()
```

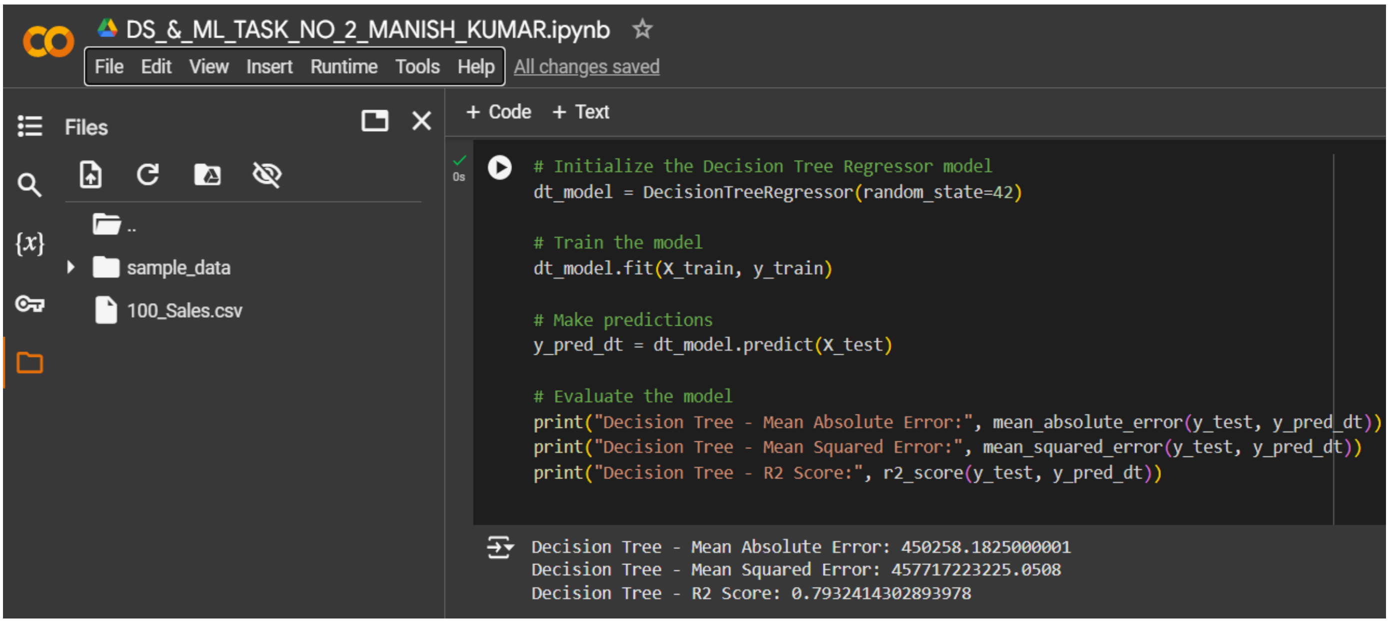
	Unit_Cost	Total_Revenue	Total_Profit	Region_Australia and Oceania	Region_Central America and the Caribbean	Region_Europe	Region_Middle East and North Africa	Region_North America	Region_Sub_Saharan Africa
0	159.42	2533654.00	951410.50	True	False	False	False	False	False
1	117.11	576782.80	248406.36	False	True	False	False	False	False
2	524.96	1158502.59	224598.75	False	False	True	False	False	False
3	6.92	75591.66	19525.82	False	False	False	False	False	True
4	524.96	3296425.02	639077.50	False	False	False	False	False	True

5 rows × 197 columns

Task 2 Report

Classification with Decision Tree and Random Forest

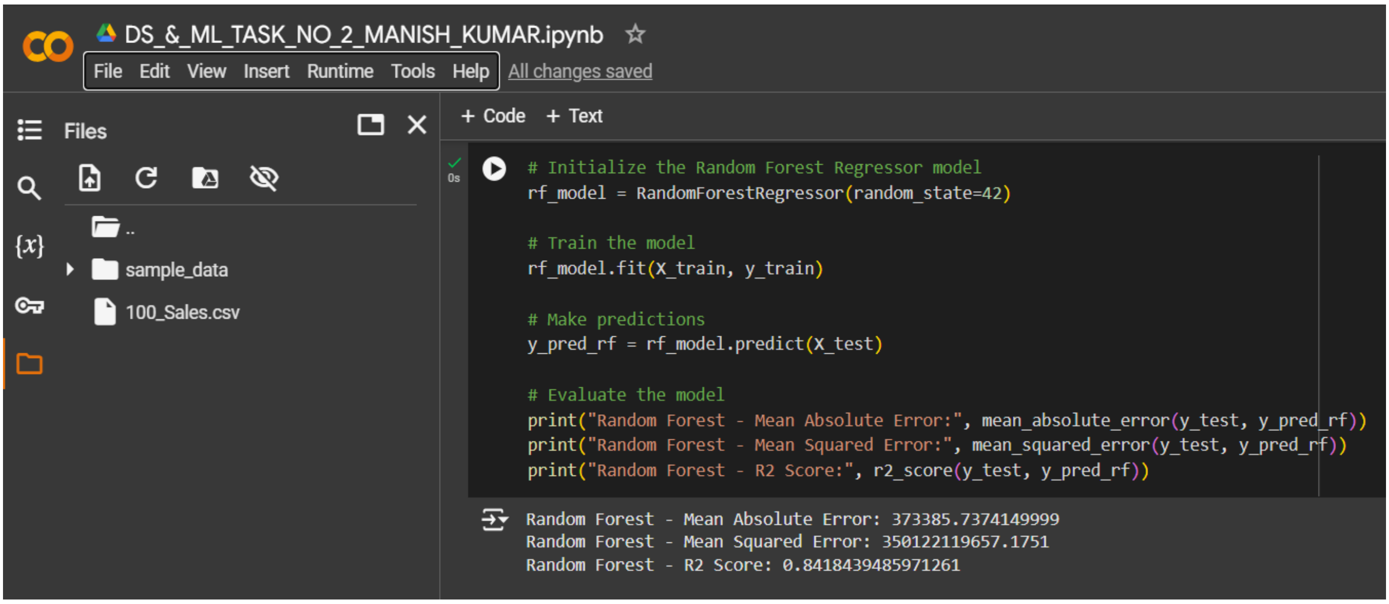
❖ Decision Tree Model:


DS_&_ML_TASK_NO_2_MANISH_KUMAR.ipynb
File Edit View Insert Runtime Tools Help
All changes saved
Files
..
sample_data
100_Sales.csv
+ Code + Text
0s

```
# Initialize the Decision Tree Regressor model
dt_model = DecisionTreeRegressor(random_state=42)

# Train the model
dt_model.fit(X_train, y_train)

# Make predictions
y_pred_dt = dt_model.predict(X_test)

# Evaluate the model
print("Decision Tree - Mean Absolute Error:", mean_absolute_error(y_test, y_pred_dt))
print("Decision Tree - Mean Squared Error:", mean_squared_error(y_test, y_pred_dt))
print("Decision Tree - R2 Score:", r2_score(y_test, y_pred_dt))
```

```
Decision Tree - Mean Absolute Error: 450258.1825000001
Decision Tree - Mean Squared Error: 457717223225.0508
Decision Tree - R2 Score: 0.7932414302893978
```

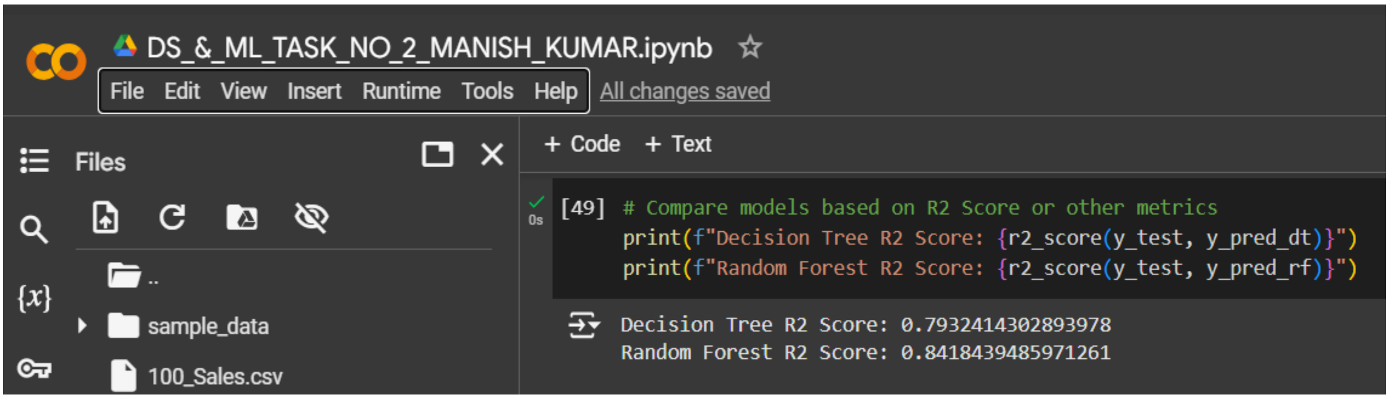
❖ Random Forest Model:


DS_&_ML_TASK_NO_2_MANISH_KUMAR.ipynb
File Edit View Insert Runtime Tools Help
All changes saved
Files
..
sample_data
100_Sales.csv
+ Code + Text
0s

```
# Initialize the Random Forest Regressor model
rf_model = RandomForestRegressor(random_state=42)

# Train the model
rf_model.fit(X_train, y_train)

# Make predictions
y_pred_rf = rf_model.predict(X_test)

# Evaluate the model
print("Random Forest - Mean Absolute Error:", mean_absolute_error(y_test, y_pred_rf))
print("Random Forest - Mean Squared Error:", mean_squared_error(y_test, y_pred_rf))
print("Random Forest - R2 Score:", r2_score(y_test, y_pred_rf))
```

```
Random Forest - Mean Absolute Error: 373385.7374149999
Random Forest - Mean Squared Error: 350122119657.1751
Random Forest - R2 Score: 0.8418439485971261
```

❖ Model comparison:


DS_&_ML_TASK_NO_2_MANISH_KUMAR.ipynb
File Edit View Insert Runtime Tools Help
All changes saved
Files
..
sample_data
100_Sales.csv
+ Code + Text
0s
[49]

```
# Compare models based on R2 Score or other metrics
print(f"Decision Tree R2 Score: {r2_score(y_test, y_pred_dt)}")
print(f"Random Forest R2 Score: {r2_score(y_test, y_pred_rf)}")
```

```
Decision Tree R2 Score: 0.7932414302893978
Random Forest R2 Score: 0.8418439485971261
```

Task 2 Report

Classification with Decision Tree and Random Forest

3. Describe Widget/Algorithm Used in Task

Algorithms Used:

- **Decision Tree Regressor**: A non-linear model used for regression tasks that splits the data based on feature values to predict the target. It's interpretable and works well on both categorical and continuous data.
  - **Process**: The algorithm divides the dataset into subsets based on the best splits to predict a continuous target.
- **Random Forest Regressor**: An ensemble method that creates multiple decision trees and averages their predictions to improve performance and reduce overfitting.
  - **Process**: Random Forest builds multiple decision trees and outputs the average prediction from all trees. It's a more robust algorithm compared to a single decision tree.

Steps Involved:

1) **Data Preprocessing**:
   - Load the dataset and inspect it for missing values.
   - Handle missing data by filling numeric columns with mean or median values.
   - Drop columns that contain completely missing data or are not relevant (e.g., "Unnamed" columns).

2) **Feature Engineering**:
   - Apply one-hot encoding to categorical columns (such as Region, Country).

3) **Model Training**:
   - Split the dataset into training and testing sets (80%-20%).
   - Train Decision Tree and Random Forest models on the training data.

4) **Model Evaluation**:
   - Use metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) to evaluate both models.

Libraries/Tools Used:

- **Pandas**: Utilized for data manipulation and cleaning, including handling missing values, encoding categorical variables, and splitting datasets.

- **Scikit-Learn**: This library is used for implementing machine learning models such as Decision Tree and Random Forest, as well as evaluating their performance using regression metrics like MAE, MSE, and R^2.

- **NumPy**: Employed for numerical operations such as handling missing values with mean imputation and performing mathematical calculations during model evaluation.

***** The End *****