SIDDHARTHA DEV SHRESTHA

Advertising Sales Prediction — Detailed Report

- **1. Objective**

The aim of this project is to **predict sales revenue** based on advertising spending across three different channels:

- **TV**
- **Radio**
- **Newspaper**

This is a **regression problem** because the target variable (sales) is continuous. We'll use **Linear Regression** as our main model.

- **2. Dataset**

We use the classic **Advertising Dataset** (commonly available from the book *"An Introduction to Statistical Learning"*).

Columns:

Feature	Description
TV	Advertising spend on TV (in thousands of dollars)
Radio	Advertising spend on Radio
Newspaper	Advertising spend on Newspaper
Sales	Product sales (in thousands of units)

Target Variable → Sales

- **3. Step-by-Step Process**

- **a) Importing Libraries**

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression

from sklearn.metrics import mean_squared_error, r2_score
```

- **b) Loading the Dataset**

```
df = pd.read_csv("advertising.csv")
```

- Loads the CSV file into a Pandas DataFrame.
- **df.head()** → preview first 5 rows.

- **c) Exploratory Data Analysis (EDA)**

Check dataset shape and info:

```
print(df.shape)

print(df.info())

print(df.describe())
```

Insights:

- No missing values.
- Sales seem correlated with TV and Radio spending.

- Newspaper may have weaker correlation.

- **d) Data Visualization**

sns.pairplot(df, x_vars=["TV", "Radio", "Newspaper"], y_vars="Sales", height=5, aspect=0.8)

plt.show()

- Shows scatter plots for each feature vs Sales.
- TV and Radio have a stronger linear relationship.

Correlation heatmap:

sns.heatmap(df.corr(), annot=True, cmap="coolwarm")

plt.show()

- **e) Feature Selection**

Separate features (X) and target (y):

X = df[["TV", "Radio", "Newspaper"]]

y = df["Sales"]

- **f) Splitting the Dataset**

```
X_train, X_val, y_train, y_val = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

- **g) Model Training**

```
model = LinearRegression()

model.fit(X_train, y_train)
```

- **h) Making Predictions**

```
y_pred = model.predict(X_val)
```

- **i) Model Evaluation**

```
mse = mean_squared_error(y_val, y_pred)

rmse = np.sqrt(mse)

r2 = r2_score(y_val, y_pred)


print("Mean Squared Error:", mse)

print("Root Mean Squared Error:", rmse)

print("R-squared:", r2)
```

- **4. Example Output**

Mean Squared Error: 3.174

Root Mean Squared Error: 1.781

R-squared: 0.897

Interpretation:

- **R-squared ~ 0.897** means ~89.7% of the variance in sales is explained by TV, Radio, and Newspaper spending.
- Lower RMSE means predictions are close to actual sales.

- **5. Model Coefficients**

```
print("Intercept:", model.intercept_)

print("Coefficients:", model.coef_)
```

Example output:

Intercept: 2.938889

Coefficients: [0.045765, 0.188530, -0.001037]

Interpretation:

- For each extra $1,000 spent on **TV**, sales increase by ~0.045 units (keeping others constant).
- Radio also increases sales, Newspaper has almost no effect.

- **6. Conclusion**
- TV and Radio ads significantly impact sales.
- Newspaper ads have little to no effect.
- Linear Regression is effective here with high R^2 value.
- Future improvements:
    - Remove less significant features (feature selection).
    - Try **Ridge** or **Lasso Regression** for better regularization.