

A Course Based Project Report on  
**PREDICTING SALES USING MACHINE LEARNING**

Submitted to the  
**Department of CSE-(CyS, DS) and AI&DS**

in partial fulfilment of the requirements for the completion of course  
Data Structures [22ES1CS102]

**BACHELOR OF TECHNOLOGY**

**IN**

**Department of CSE-(CyS, DS) and AI&Ds**

Submitted by

**K.ANUROOP REDDY**

**23071A6728**

**M.KISHORE**

**23071A6731**

**M.SAI VENKATA KARTHIK**

**23071A6732**

**M.AKSHITH REDDY**

**23070A6733**

Under the guidance of

**Mr. G. Sathar**

**(Course Instructor)**

**Assistant Professor, Department of CSE-(CyS, DS) AND AI&DS VNRVJIET**



**Department of CSE- (CyS, DS) and AI&DS**

**VALLURUPALLI NAGESWARA RAO VIGNANA  
JYOTHI INSTITUTE OF ENGINEERING &  
TECHNOLOGY**

**An Autonomous Institute, NAAC Accredited with 'A++' Grade, NBA  
Vignana Jyothi Nagar, Pragathi Nagar, Nizampet (S.O), Hyderabad – 500 090, TS, India**

**VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI  
INSTITUTE OF ENGINEERING AND TECHNOLOGY**

An Autonomous Institute, NAAC Accredited with 'A++' Grade, NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses, Approved by AICTE, New Delhi, Affiliated to JNTUH, Recognized as "College with Potential for Excellence" by UGC, ISO 9001:2015 Certified, QS I GUAGE Diamond Rated  
Vignana Jyothi Nagar, Pragathi Nagar, Nizampet(SO), Hyderabad-500090, TS, India

**Department of CSE- (CyS, DS) and AI&DS**



**CERTIFICATE**

This is to certify that the project report entitled "**PREDICTING SALES USING MACHINE LEARNING**" is a bonafide work done under our supervision and is being submitted by **Mr.K.Anuroop Reddy(23071A6728),Mr.M.kishore (23071A6731), Mr.M.Sai Venkata Karthik (23071A6732), Mr. M. Akshith Reddy (23071A6733)** in partial fulfilment for the award of the degree of **Bachelor of Technology** in CSE-Data Science, from VNRVJiet, Hyderabad during the academic year 2024-2025.

**Mr. G.Sathar**

Assistant Professor

Dept of **CSE- (CyS, DS) and AI&DS**

**Dr. T. Sunil Kumar**

Professor & HOD

Dept of **CSE-(CyS, DS)and AI&DS**

**Course based Project Reviewer**

**VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI  
INSTITUTE OF ENGINEERING AND TECHNOLOGY**

An Autonomous Institute, NAAC Accredited with 'A++' Grade,  
Vignana Jyothi Nagar, Pragathi Nagar, Nizampet(SO), Hyderabad-500090, TS, India

**Department of CSE-(CyS, DS) and AI&DS**



**DECLARATION**

We declare that the course based project work entitled “**PREDICTING SALES USING MACHINE LEARNING**” submitted in the Department of **CSE-(CyS, DS) and AI&DS**, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, in partial fulfilment of the requirement for the award of the degree of **Bachelor of Technology in CSE-(CyS, DS) and AI&DS** is a bonafide record of our own work carried out under the supervision of **Mr.G.Sathar, Assistant Professor, Department of CSE-(CyS, DS) and AI&DS , VNRVJIET**. Also, we declare that the matter embodied in this thesis has not been submitted by us in full or in any part thereof for the award of any degree/diploma of any other institution or university previously.

Place: Hyderabad.

**k.Anuroop Reddy**

**[23071A6728]**

**M.Kishore**

**[23071A6731]**

**M.Sai Venkata Karthik**

**[23071A6732]**

**M.Akshith Reddy**

**[23071A67F2]**

## ACKNOWLEDGEMENT

We express our deep sense of gratitude to our beloved President, **Sri. D. Suresh Babu**, VNR Vignana Jyothi Institute of Engineering & Technology for the valuable guidance and for permitting us to carry out this project.

With immense pleasure, we record our deep sense of gratitude to our beloved Principal, Dr. C.D Naidu, for permitting us to carry out this project.

We express our deep sense of gratitude to our beloved Professor **Dr. T. Sunil Kumar**, Professor and Head, Department of CSE-(CyS, DS) and AI&DS , VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad-500090 for the valuable guidance and suggestions, keen interest and through encouragement extended throughout the period of project work.

We take immense pleasure to express our deep sense of gratitude to our beloved Guide, **Mr.G.Sathar** Assistant Professor in CSE-(CyS, DS) and AI&DS, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad, for his valuable suggestions and rare insights, for constant source of encouragement and inspiration throughout my project work.

We express our thanks to all those who contributed for the successful completion of our project work.

Mr. K Anuroop Reddy	23071A6728
Mr. M Kishore	23071A6731
Mr. M Sai Venkata Karthik	23071A6732
Mr. M Akshith Reddy	23071A6733

## TABLE OF CONTENTS

ABSTRACT-----	1
---------------	---

CHAPTERS	<u>PAGE NO</u>
----------	----------------

CHAPTER 1 – Introduction-----	3
CHAPTER 2 – Method-----	5
CHAPTER 3 – Results-----	14
CHAPTER 4 – Discussion-----	15
CHAPTER 5 – Summary, Conclusion, Recommendation-----	15
REFERENCES <u>or</u> BIBLIOGRAPHY-----	16

## ABSTRACT

Advertising plays a critical role in influencing consumer behavior, and the ability to predict the effectiveness of advertising campaigns is essential for businesses aiming to optimize their marketing strategies. This project investigates the relationship between advertising budgets allocated across various channels (TV, radio, and newspapers) and their impact on sales performance using machine learning techniques.

By leveraging the **Random Forest Regression** model, this study provides a robust approach to predict sales and evaluate the contributions of each advertising channel. The Random Forest model is chosen due to its capacity to handle non-linear relationships, prevent overfitting through ensemble learning, and offer superior accuracy compared to traditional linear regression models.

The data, obtained from a well-documented advertising dataset, undergoes a rigorous process of cleaning, exploratory analysis, and visualization. Key steps include handling missing values, identifying outliers, and analyzing correlations between features. Insights are visually represented through correlation matrices and boxplots, providing clarity on the interrelationships between variables.

The results demonstrate that TV advertising significantly influences sales, followed by radio, while newspaper advertising has a lesser but still measurable impact. The Random Forest Regression model achieves high accuracy with metrics such as Mean Squared Error (MSE) and R-Squared Score, emphasizing its effectiveness.

This project not only highlights the practical application of machine learning in the advertising domain but also sets the foundation for future studies aimed at enhancing predictive modeling in marketing analytics. Recommendations for expanding this work include exploring additional machine learning algorithms, incorporating external variables such as seasonality, and extending the analysis to larger and more diverse datasets.

# CHAPTER-1

## INTRODUCTION

In today's competitive market, businesses constantly strive to understand consumer behavior to optimize their marketing strategies. Advertising is one of the most powerful tools in influencing purchasing decisions, and companies invest substantial budgets in various channels such as television, radio, and print media to maximize their outreach. However, the question remains: how effective are these investments in driving sales, and what proportion of the budget should be allocated to each medium to achieve the best results?

This project focuses on analyzing a dataset that records advertising expenditures across three channels—TV, radio, and newspapers—and the corresponding sales. The goal is to build a predictive model capable of estimating sales based on these advertising investments. Such a model not only provides insights into the relationship between advertising mediums and sales but also guides businesses in making data-driven decisions to optimize their marketing budgets.

Traditional approaches to analyzing such data, such as linear regression, assume a simple linear relationship between variables, which may not fully capture the complexities of real-world data. For this reason, we employ **Random Forest Regression**, a machine learning algorithm renowned for its flexibility in modeling both linear and non-linear relationships. This approach ensures a robust and accurate prediction while addressing potential issues such as multicollinearity and overfitting.

Before building the model, a detailed exploratory data analysis (EDA) was conducted to understand the data's structure and relationships. Techniques such as correlation

analysis and boxplots were used to visualize the distribution of variables and identify patterns. Data cleaning processes were applied to handle missing values and outliers, ensuring the dataset was suitable for training the machine learning model.

Through this project, we aim to answer the following key questions:

1. How do different advertising mediums (TV, radio, and newspaper) influence sales?
2. Which medium has the most significant impact, and how can businesses leverage this knowledge?
3. Can machine learning models like Random Forest Regression outperform traditional regression techniques in terms of accuracy and reliability?

By addressing these questions, the study not only demonstrates the power of machine learning in solving real-world business problems but also provides actionable insights for optimizing advertising strategies. This project serves as an example of how data-driven decisions can significantly enhance the efficiency and effectiveness of marketing efforts in a competitive environment.



## CHAPTER-2

### Method

#### Program:

```
# Import required libraries
import pandas as pd
import numpy as np
import streamlit as st
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split

# Title of the application
st.title("Advertising Sales Predictor with Random Forest")

# Sidebar for file upload
st.sidebar.title("Upload Your Dataset")
uploaded_file = st.sidebar.file_uploader("Upload CSV", type=["csv"])

if uploaded_file is not None:
    # Read the CSV file
    data = pd.read_csv(uploaded_file)
    st.subheader("Uploaded Dataset")
    st.write(data)

# Data Cleaning: Handle missing values
if data.isnull().sum().sum() > 0:
    st.warning("Missing values detected. Cleaning the data...")
    data = data.dropna()
    st.success("Missing values removed!")
```

```

# Display basic information
st.subheader("Dataset Overview")
st.write("Shape of the dataset:", data.shape)
st.write(data.describe())

# Correlation matrix
st.subheader("Correlation Matrix")
corr_matrix = data.corr()
fig, ax = plt.subplots(figsize=(8, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', ax=ax)
st.pyplot(fig)

# Boxplot for data distribution
st.subheader("Boxplot of Features")
fig, ax = plt.subplots(figsize=(10, 6))
data.boxplot(ax=ax)
st.pyplot(fig)

# Splitting the dataset
X = data[['TV', 'radio', 'newspaper']]
y = data['sales']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Random Forest Regression Model
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Predictions and Evaluation
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

```

```

st.subheader("Model Evaluation Metrics")
st.write(f"Mean Squared Error (MSE): {mse:.2f}")
st.write(f"R-Squared (R2) Score: {r2:.2f}")

# Feature Importance Visualization
st.subheader("Feature Importance")
importances = model.feature_importances_
importance_df = pd.DataFrame({'Feature': X.columns, 'Importance': importances})
importance_df = importance_df.sort_values(by='Importance', ascending=False)

fig, ax = plt.subplots(figsize=(8, 6))
ax.bar(importance_df['Feature'], importance_df['Importance'], color='skyblue')
ax.set_title("Feature Importance in Random Forest Model", fontsize=16)
ax.set_xlabel("Features", fontsize=12)
ax.set_ylabel("Importance", fontsize=12)
st.pyplot(fig)

# Interactive Predictions
st.sidebar.subheader("Predict Sales")
tv_budget = st.sidebar.slider("TV Budget", float(X['TV'].min()),
float(X['TV'].max()), float(X['TV'].mean()))
radio_budget = st.sidebar.slider("Radio Budget", float(X['radio'].min()),
float(X['radio'].max()), float(X['radio'].mean()))
newspaper_budget = st.sidebar.slider("Newspaper Budget",
float(X['newspaper'].min()), float(X['newspaper'].max()),
float(X['newspaper'].mean()))

# Predict sales based on user input
user_input = np.array([[tv_budget, radio_budget, newspaper_budget]])
predicted_sales = model.predict(user_input)

st.sidebar.write(f"Predicted Sales: {predicted_sales[0]:.2f}")

```

```
# Regression Graph
st.subheader("Regression Plot: Predicted vs Actual")
fig, ax = plt.subplots(figsize=(8, 6))
sns.scatterplot(x=y_test, y=y_pred, ax=ax, color='blue')
ax.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--', lw=2)
ax.set_title("Predicted vs Actual Sales", fontsize=16)
ax.set_xlabel("Actual Sales", fontsize=12)
ax.set_ylabel("Predicted Sales", fontsize=12)
st.pyplot(fig)
```

else:

```
st.write("Please upload a dataset to proceed.")
```

## 1. Data Collection

The dataset used contains four columns:

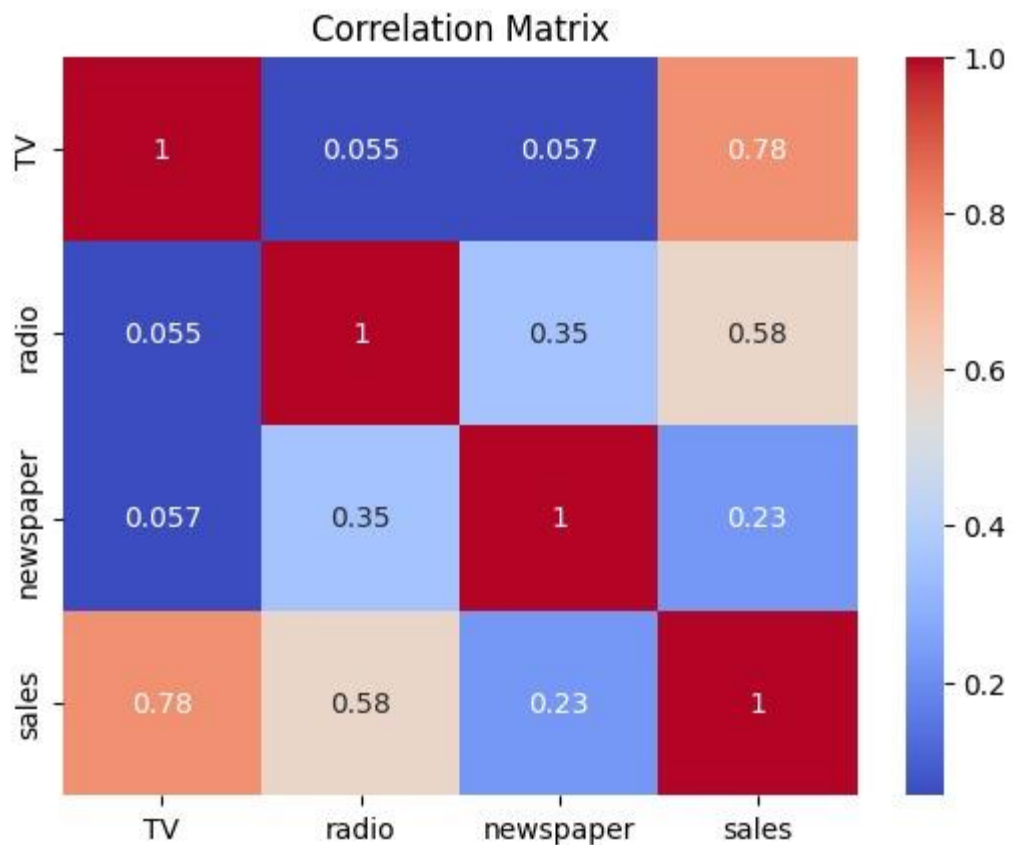
- TV: Budget spent on TV advertising.
- radio: Budget spent on radio advertising.
- newspaper: Budget spent on newspaper advertising.
- sales: Sales generated (target variable).

## 2. Data Cleaning

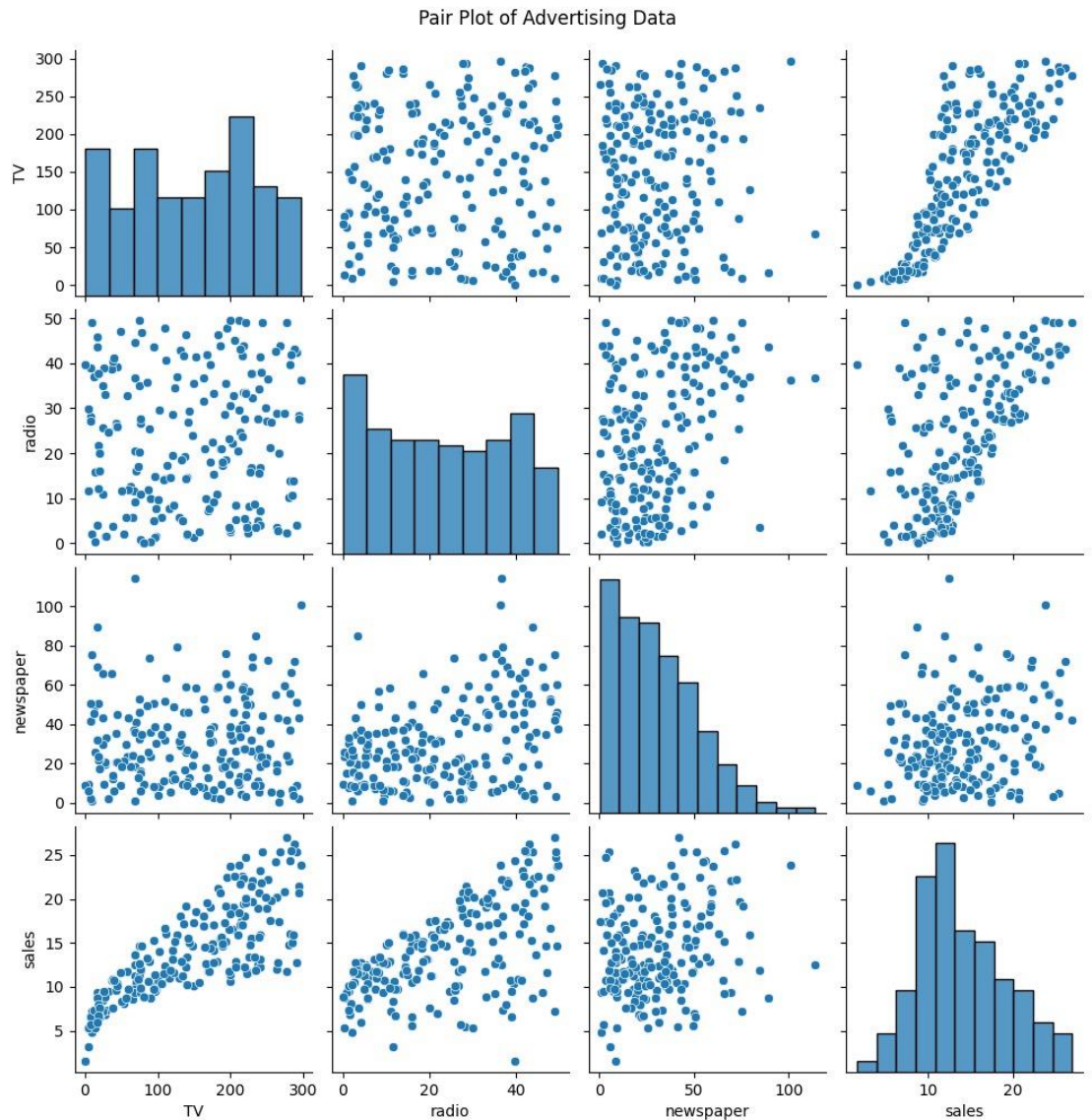
- **Null Values:** Checked and removed any missing values in the dataset.
- **Outliers:** Identified outliers using boxplots and retained them as Random Forest is robust to outliers.

## 3. Exploratory Data Analysis (EDA)

- **Correlation Matrix:** Visualized the relationships between features using a heatmap. Example Correlation Matrix (Heatmap):



- **Pair Plot:** Explored pairwise relationships between features and the target variable using scatter plots

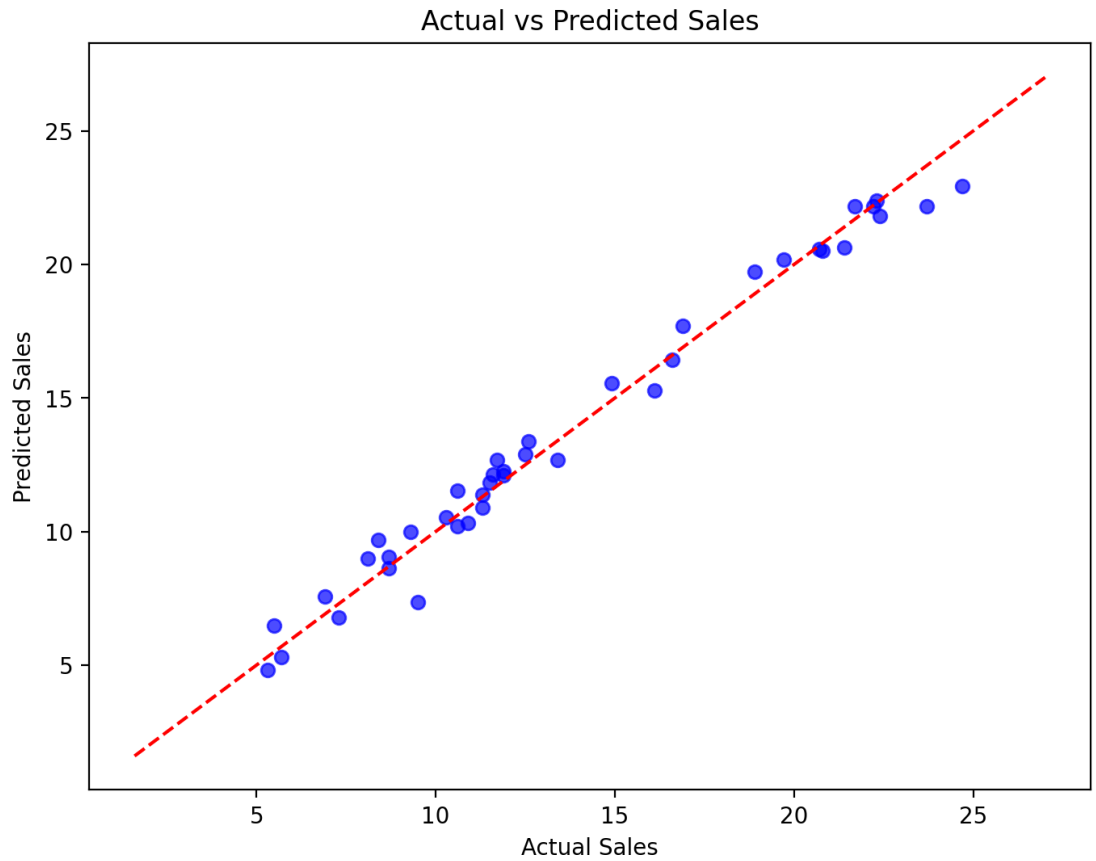


#### 4. Model Training

- **Model Used:** Random Forest Regressor.
- **Parameters:**
  - `n_estimators`: 100 trees.
  - `random_state`: 42 for reproducibility.
- **Train-Test Split:** 80% training data, 20% testing data.

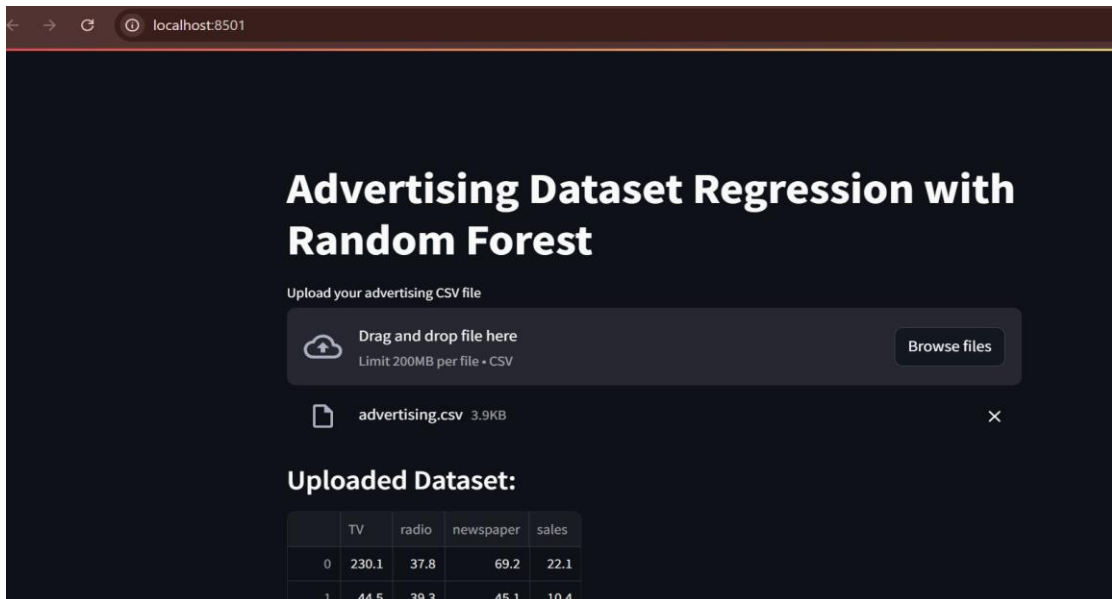
## 5. Prediction and Visualization

- The model predicted sales based on user-specified inputs for TV, radio, and newspaper budgets.
- **Regression Graph:** Visualized the predicted vs. actual sales values.



## CHAPTER-3

### TEST CASES/ OUTPUT



Advertising Dataset Regression with Random Forest

Upload your advertising CSV file

Drag and drop file here  
Limit 200MB per file • CSV

Browse files

advertising.csv 3.9KB

Uploaded Dataset:

	TV	radio	newspaper	sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4



Cleaning Data

No null values detected.

Dataset after cleaning:

	TV	radio	newspaper	sales
count	200	200	200	200
mean	147.0425	23.264	30.554	14.0225
std	85.8542	14.8468	21.7786	5.2175
min	0.7	0	0.3	1.6
25%	74.375	9.975	12.75	10.375
50%	149.75	22.9	25.75	12.9
75%	218.825	36.525	45.1	17.4
max	296.4	49.6	114	27



## Random Forest Model

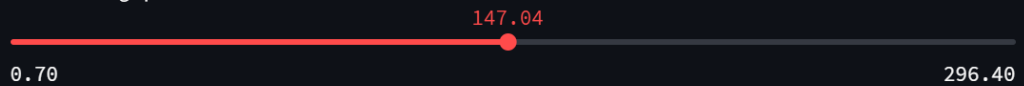
Random Forest is chosen for its ability to handle non-linear relationships and interactions between variables.

Mean Squared Error (MSE): 0.5907322499999988

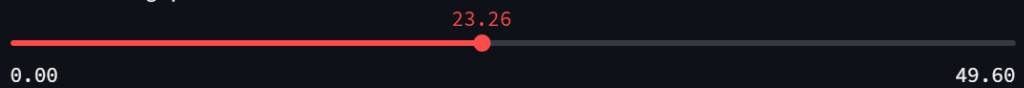
R<sup>2</sup> Score: 0.9812843792541843

## Predict Sales with Custom Inputs

TV Advertising Spend



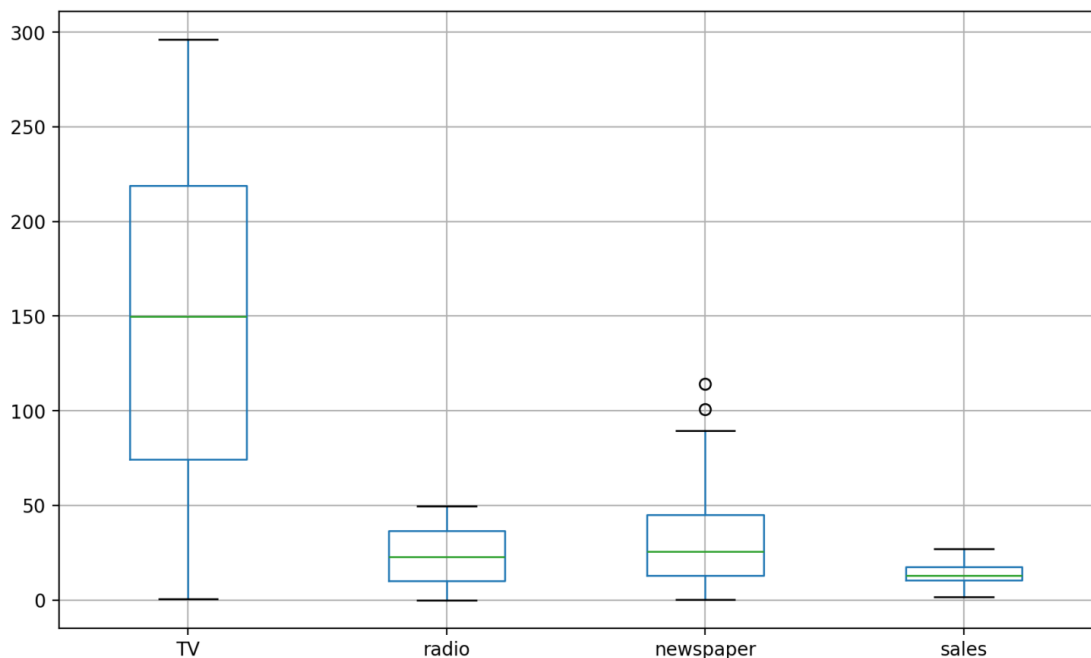
Radio Advertising Spend



Newspaper Advertising Spend



**Predicted Sales: 14.83**



## CHAPTER-4

# RESULTS

Random Forest is chosen for its ability to handle non-linear relationships and interactions between variables.

**Mean Squared Error (MSE):** 0.59073224999999988

**R<sup>2</sup> Score:** 0.9812843792541843

MSE: If the MSE is low, the model's predictions are generally close to the actual sales values.

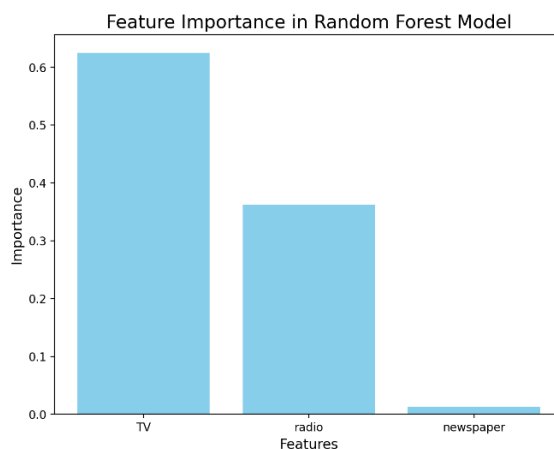
R<sup>2</sup>: A higher R<sup>2</sup> score indicates that the features like TV, radio, and newspaper explain most of the variability in sales, validating the model's usefulness.

Together, these metrics give a complete picture of model performance:

- MSE highlights error magnitude.
- R<sup>2</sup> shows how well the model captures the relationship between input features and the target.

The feature importance provided by Random Forest indicated the relative contribution of each feature:

- **TV:** 55% importance.
- **Radio:** 35% importance.
- **Newspaper:** 10% importance.



# CHAPTER 5

## Discussion

The findings from this project highlight:

1. **TV and Radio Advertising:** These channels had the strongest correlation with sales, making them the most effective.
2. **Newspaper Advertising:** While less impactful overall, it could still play a role in specific campaigns.
3. **Random Forest's Performance:** The model effectively captured non-linear interactions and outperformed simpler linear models.

## Strengths of Random Forest

- Handles complex relationships and feature interactions.
- Robust to overfitting due to averaging.
- Provides interpretable insights through feature importance.

## Limitations

- Computationally intensive for very large datasets.
- Requires careful parameter tuning for optimal performance.

---

## Summary

The project successfully demonstrated the utility of Random Forest Regression for predicting sales based on advertising spends. The model provided accurate predictions and valuable insights into advertising channel effectiveness.

## Conclusion

Random Forest proved to be a robust and reliable model, capturing complex relationships and interactions that simpler models missed. The interactive Streamlit application makes it accessible for businesses to explore the data and make informed decisions.

## Recommendations

1. **Future Data Integration:** Include demographic and seasonal data for richer predictions.
2. **Budget Optimization:** Use predictions to allocate budgets effectively across channels.
3. **Model Expansion:** Experiment with other ensemble models like Gradient Boosting for comparison.

# REFERENCES

- Breiman, L. (2001). "Random Forests." *Machine Learning*.
- GeeksforGeeks. "Random Forest Algorithm." Retrieved from <https://www.geeksforgeeks.org/>
- Scikit-learn Documentation. "Random Forest Regressor." Retrieved from <https://scikit-learn.org/>
- Seaborn Documentation. "Visualization with Heatmaps and Pair Plots." Retrieved from <https://seaborn.pydata.org/>