# Task 05 – Sales Prediction using Machine Learning

## Objective

The goal of this project is to predict product sales based on advertisement spending across three different media channels — TV, Radio, and Newspaper. The project applies machine learning algorithms such as Linear Regression and Random Forest to analyze relationships and make accurate sales predictions.

## Dataset Information

The dataset used for this analysis is 'Advertising.csv', which includes the following columns:

- TV — advertising budget spent on TV (in thousands of dollars)
- Radio — advertising budget spent on Radio (in thousands of dollars)
- Newspaper — advertising budget spent on Newspaper (in thousands of dollars)
- Sales — units sold (in thousands)

## Data Preparation

1. Loaded and inspected the dataset for data types, missing values, and duplicates.
2. Checked dataset statistics using describe() and info().
3. Verified there were no missing or duplicate values.
4. Defined features (TV, Radio, Newspaper) as predictors and Sales as the target variable.
5. Split the dataset into training and testing sets using an 80/20 ratio.

## Model Training and Evaluation

Two different regression models were trained and compared:

1. Linear Regression — a simple, interpretable model that establishes a linear relationship between the features and the target.
2. Random Forest Regressor — an ensemble model that combines multiple decision trees to improve accuracy.

The performance of both models was evaluated using the following metrics:

- $R^2$ Score — goodness of fit
- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)

## Model Comparison Results

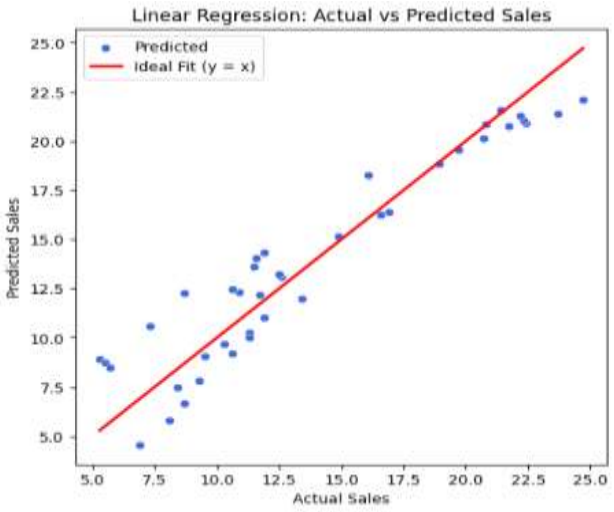
After training and evaluation, the results showed that the Random Forest model performed better in terms of prediction accuracy, achieving a higher $R^2$ score and lower errors

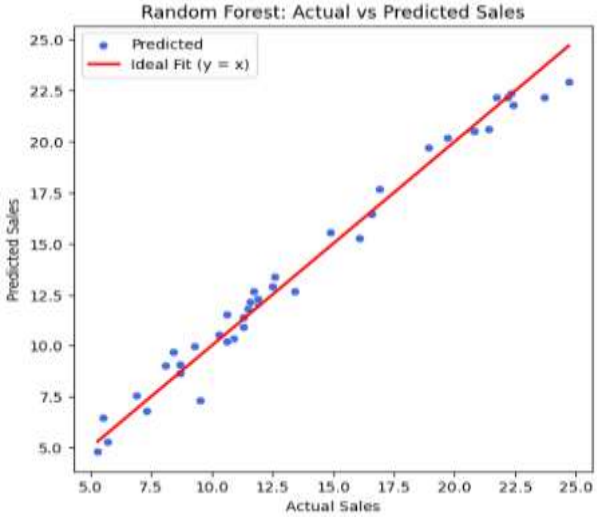
compared to the Linear Regression model. This indicates that Random Forest captured complex relationships between features and sales more effectively.

## Visualization and Analysis

Several visualizations were created to compare actual versus predicted sales and to assess model performance:

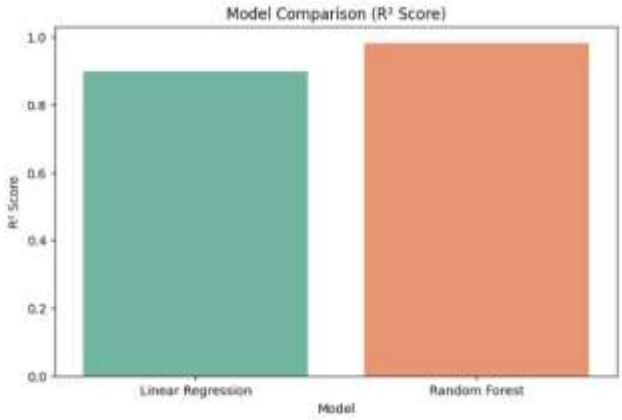
• Scatter plots of Actual vs Predicted Sales for both models.




• A reference line (y = x) was added to visualize prediction accuracy — the closer the points are to this line, the better the model's performance.

• A bar chart comparing $R^2$ scores for both models.




## Insights and Conclusion

• Advertisement spending directly influences product sales, especially TV and Radio budgets.

• Random Forest outperformed Linear Regression, showing its ability to handle nonlinear patterns.

• The analysis demonstrates how businesses can use data-driven approaches to optimize

advertising budgets.
• Predictive models like these can help allocate marketing resources efficiently for better ROI.

## Tools and Technologies Used

• Python (Pandas, NumPy, Matplotlib, Seaborn, scikit-learn)
• Jupyter Notebook / Google Colab for running and visualizing results
• Machine Learning Models — Linear Regression, Random Forest Regressor

## Final Remarks

This project showcases the use of machine learning for sales forecasting using advertising data. It highlights the process of data cleaning, model training, evaluation, and visualization. The insights gained can help organizations understand the most effective advertising channels and plan future marketing strategies accordingly.