📊 ML Project Report: Profit Prediction Using 50 Startups Dataset

Project Title: *Profit Prediction from Startup Investment Data*
Developer: *Prashik Bhimte*
Role: Machine Learning Learner
Date: *April 2025*

🧠 Objective

To predict the profit of startups based on their R&D Spend, Administration Spend, and Marketing Spend using various regression models. The goal was to compare performance across different ML models and understand how feature relationships impact prediction accuracy.

🧾 Dataset Overview

Source: 50_Startups.csv
Features:

- `R&D Spend`
- `Administration`
- `Marketing Spend`
  Target Variable:
- `Profit`

🔍 Preprocessing Steps

- Split dataset into training and test sets (80:20 ratio).

🤖 Models Developed

Model	R^2 Score (Accuracy)
Multiple Linear Regression	0.9394
Polynomial Regression	0.9202
Decision Tree Regression	0.9764
Random Forest Regression	0.9603
Support Vector Regression	0.8246

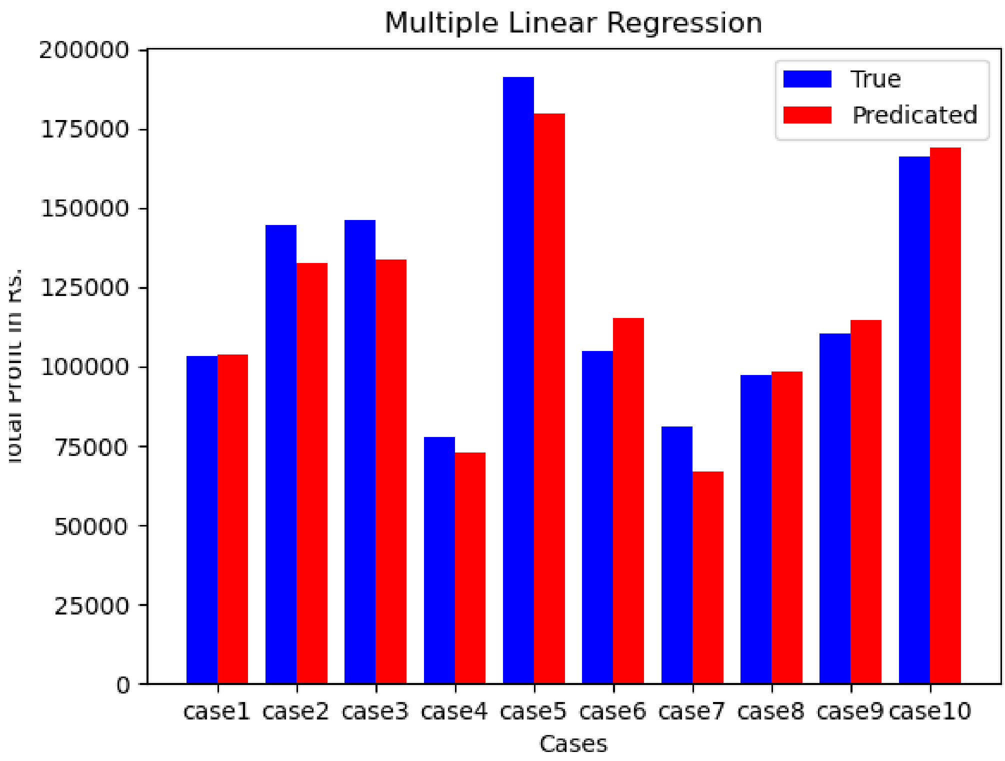
Accuracy evaluated using `sklearn.metrics.r2_score` on the test set.

📊 Model Visualizations

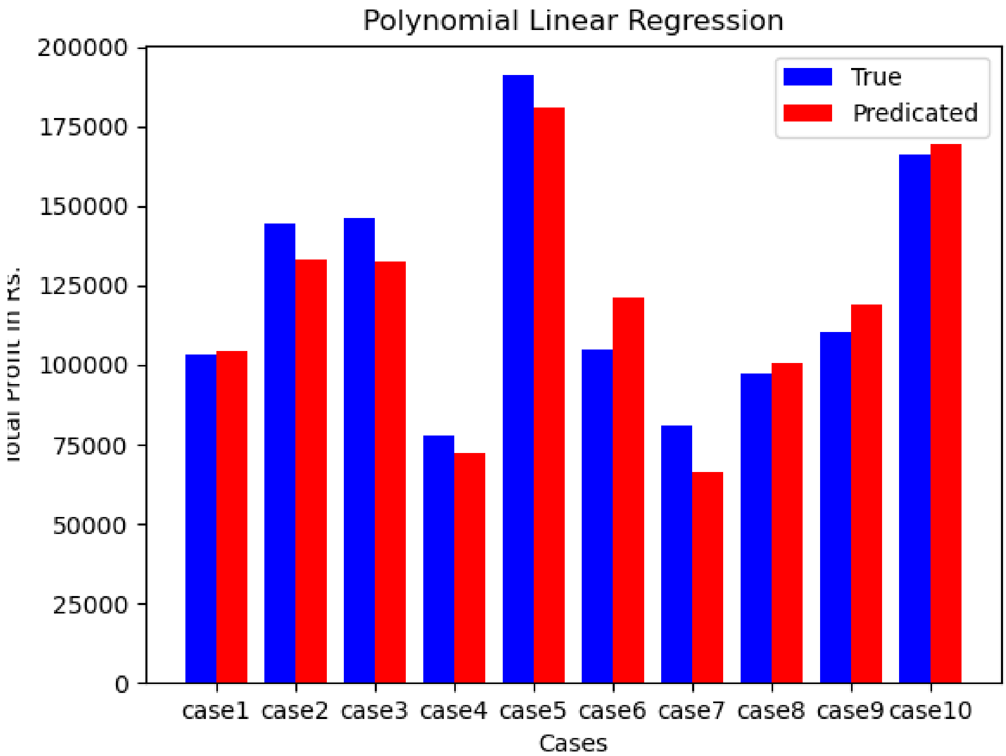
Each model prediction was visualized using a **bar chart** representing 10 random test cases:

- **X-axis**: Index of test cases (0 to 9)
- **Y-axis**: Profit values
- **Blue bars**: True Profit
- **Red bars**: Predicted Profit

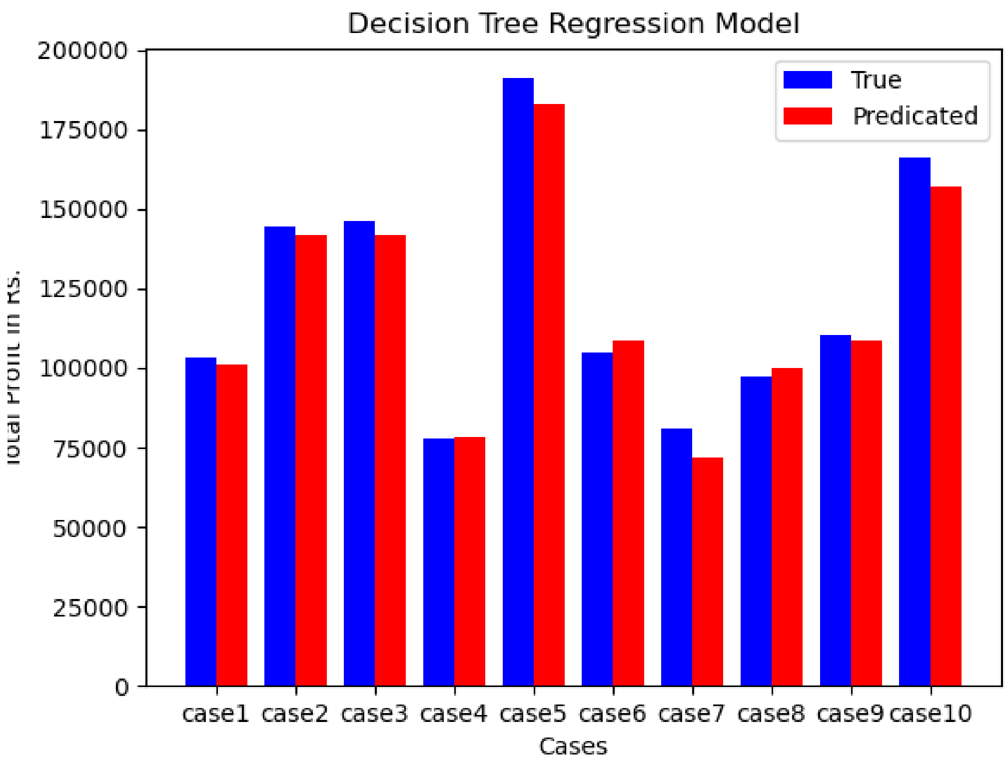
📈 Multiple Linear Regression



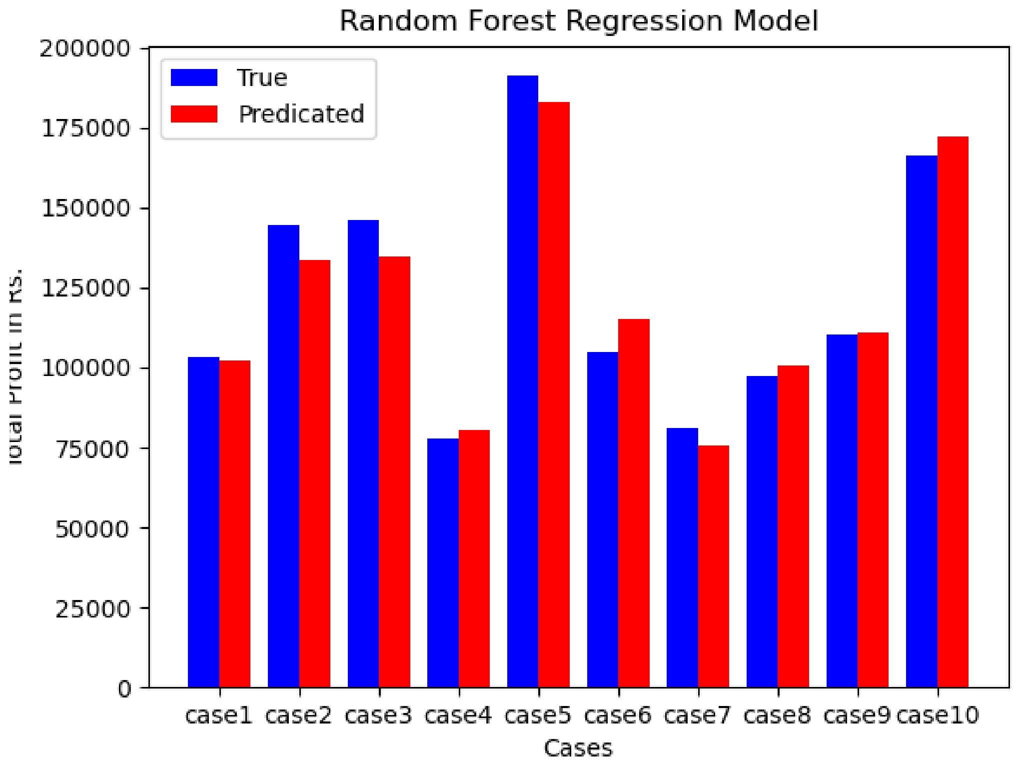
📈 Polynomial Regression



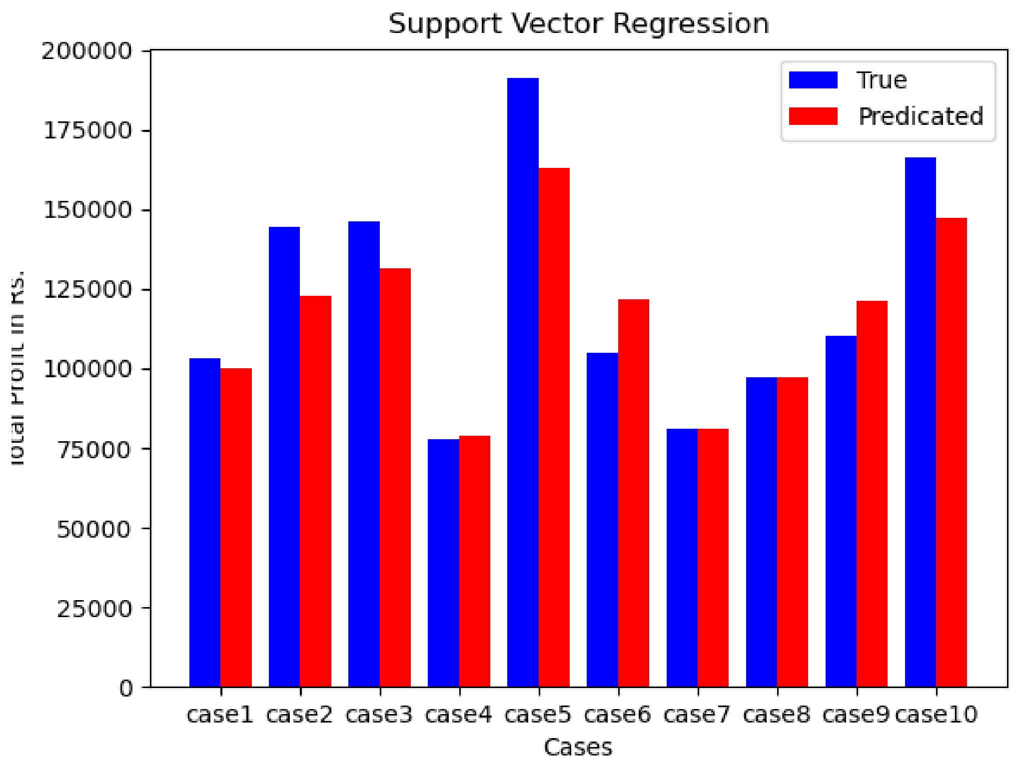
📈 Decision Tree Regression



📈 Random Forest Regression



📈 Support Vector Regression



🔍 Observations

- **Decision Tree** provided the **highest R^2 score (0.9764)**. It could perfectly fit the dataset due to overfitting, which worked well given the small sample size.
- **Random Forest** offered excellent performance while maintaining generalization.
- **SVM** underperformed, possibly due to the need for better feature scaling or parameter tuning.
- **Polynomial Regression** didn't outperform Linear Regression, hinting at a linear relationship among variables.

📚 Key Learnings

- Hands-on experience with regression techniques and model evaluation.
- Understood when and why certain models perform better.
- Importance of visualizing results to compare actual vs. predicted values.
- Gained insights into bias-variance trade-offs by comparing decision trees and random forests.
- Improved understanding of feature engineering and model tuning.