TAHREEM MALIK

DATA SCIENCE PROJECT: SCORE PREDICTOR OF CRICKET MATCH

Introduction

This project is a **Data Science regression-based project** focused on analyzing a cricket dataset and building machine learning models to predict a numerical target variable (such as player performance or match-related outcomes) based on multiple input features. The project demonstrates the complete data science workflow starting from data loading and exploration to model building, evaluation, and comparison.

Aim and Objective

Aim

The primary aim of this project is to **apply data science and machine learning techniques to a real-world cricket dataset** in order to: - Understand patterns in cricket data - Train predictive models - Compare model performance using standard evaluation metrics

Objectives

- Load and explore the cricket dataset
- Perform basic data preprocessing
- Visualize relationships between features
- Build regression models
- Evaluate and compare model accuracy

Tools and Technologies Used

Category	Tools / Libraries
Programming Language	Python
Data Handling	Pandas, NumPy
Visualization	Matplotlib, Seaborn
Machine Learning	Scikit-learn
Development Environment	Jupyter Notebook / VS Code

Dataset Description

The dataset used in this project is a **cricket-related CSV dataset** containing numerical features related to player or match statistics. These features are used as input variables for predicting a target variable.

```
First 5 rows of dataset:
   PP_Runs  PP_Wkts  Venue_Avg  Final_Score
0       68        2        150          180
1       58        2        156          164
2       44        1        177          168
3       37        1        163          149
4       50        3        144          148

Dataset info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 120 entries, 0 to 119
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PP_Runs      120 non-null    int64
 1   PP_Wkts      120 non-null    int64
 2   Venue_Avg    120 non-null    int64
 3   Final_Score  120 non-null    int64
dtypes: int64(4)
memory usage: 3.9 KB
None
```

Key Characteristics:

- Structured tabular data
- Contains independent (input) features and one dependent (output) feature
- Suitable for regression analysis

Data Preprocessing

Data preprocessing is a crucial step to ensure model accuracy and reliability.

Steps Performed:

- Loaded dataset using **Pandas**
- Checked first few rows to understand structure
- Verified data types and missing values
- Selected relevant numerical features

- Split dataset into **training and testing sets** using `train_test_split`

Exploratory Data Analysis (EDA)

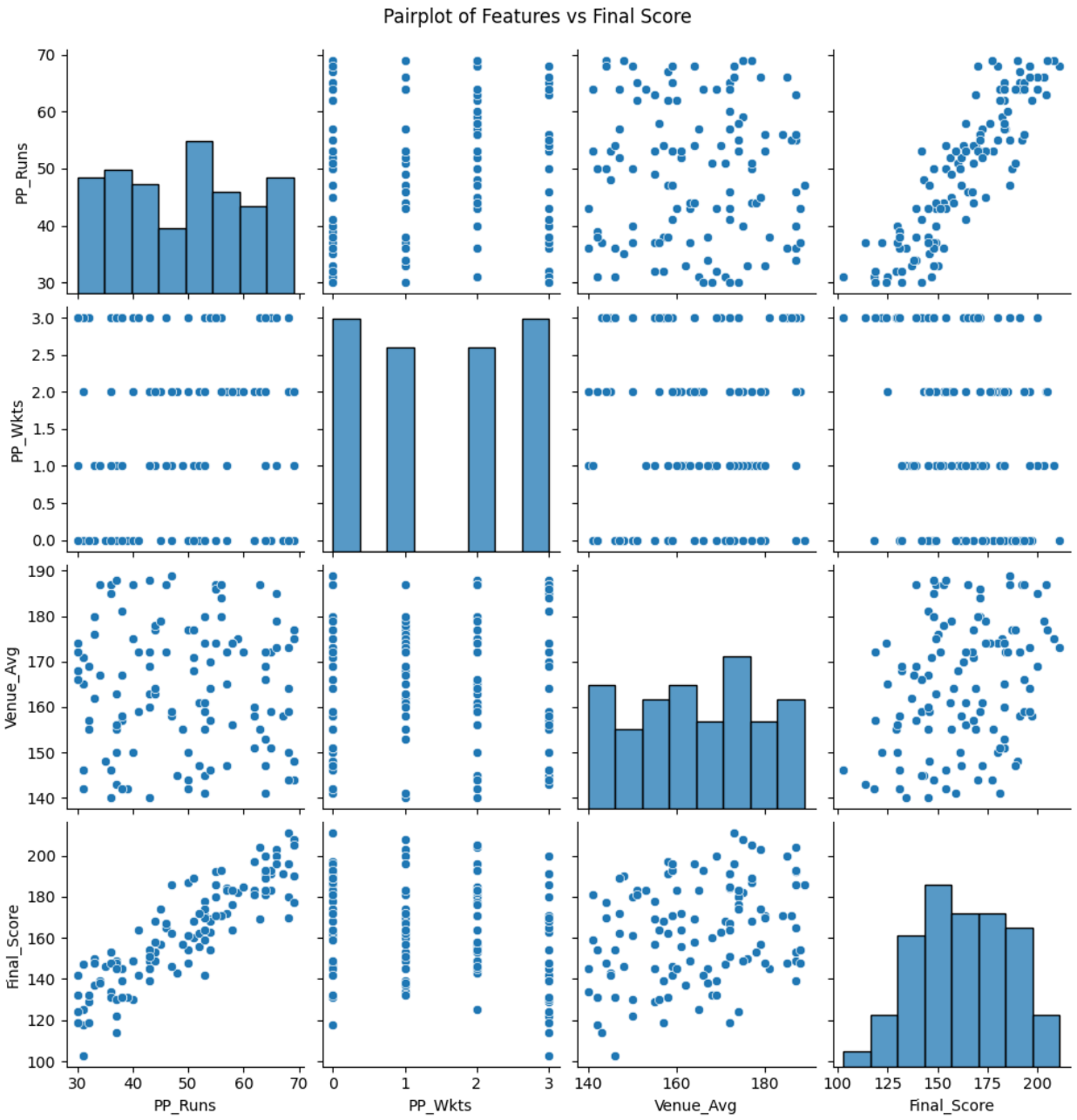
EDA helps in understanding data distribution and relationships.

Techniques Used:

- Descriptive statistics using `.describe()`
- Visualization using:
    - Histograms
    - Scatter plots
    - Correlation heatmaps (Seaborn)

Purpose of EDA:

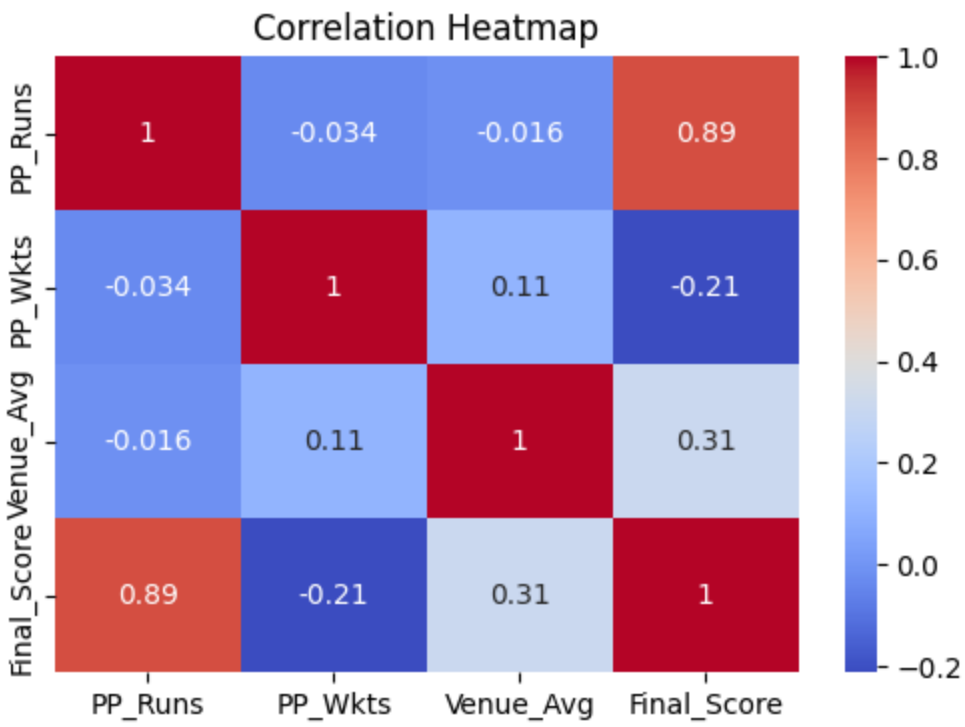
- Identify trends and patterns
- Detect outliers
- Understand feature relationships



Feature Selection

Features (independent variables) were selected based on: - Relevance to cricket performance - Numerical nature - Correlation with target variable

The target variable (dependent variable) represents the value to be predicted (e.g., runs, score, or performance metric).



Machine Learning Models Used

Linear Regression Model

Linear Regression is a basic and interpretable regression model that assumes a linear relationship between input features and the target variable.

Why Linear Regression?

- Simple and easy to interpret
- Acts as a baseline model

Implementation:

- Model: `LinearRegression()`
- Trained on training data
- Predictions made on test data

Evaluation Metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R^2 Score

Random Forest Regressor

Random Forest Regressor is an ensemble learning technique that builds multiple decision trees and averages their predictions.

Why Random Forest?

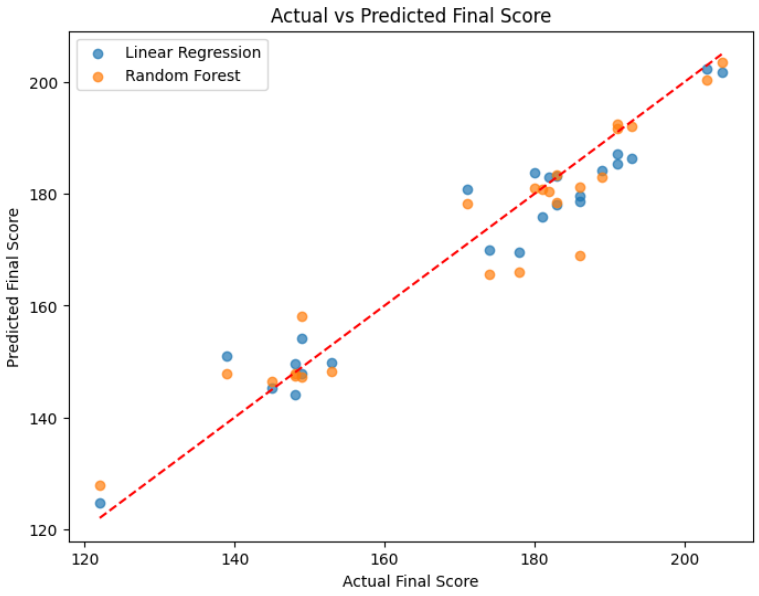
- Handles non-linear relationships
- Reduces overfitting
- Higher accuracy compared to simple models

Implementation:

- Model: `RandomForestRegressor(n_estimators=100)`
- Trained using training data
- Predictions made on test data

Evaluation Metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R^2 Score



Model Evaluation and Testing

Evaluation Metrics Explained:

MAE (Mean Absolute Error): Measures average absolute difference between predicted and actual values

RMSE (Root Mean Squared Error): Penalizes larger errors more heavily

R^2 Score: Indicates how well the model explains variance in data

Model Comparison:

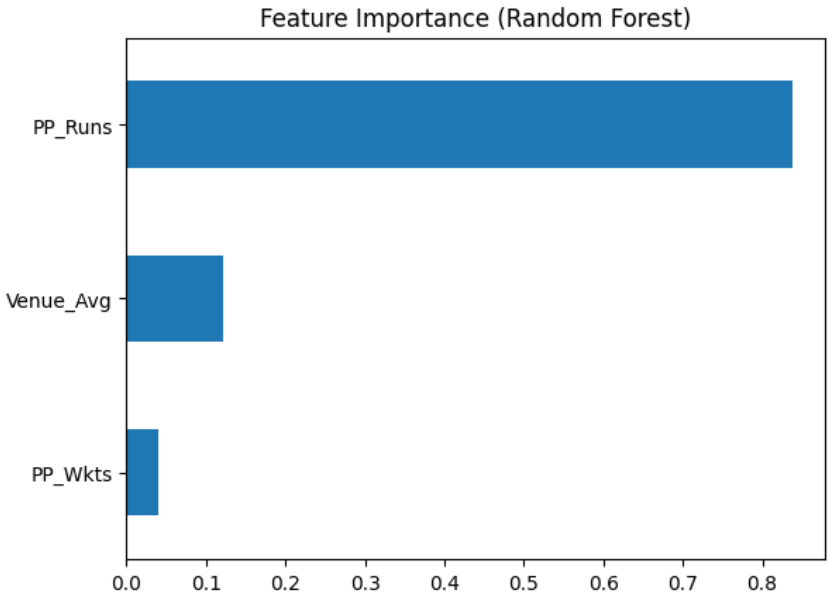
Model	MAE	RMSE	R^2 Score
Linear Regression	Lower baseline	Moderate	Moderate
Random Forest	Lower	Better	Higher

```
Training samples: 96, Test samples: 24

Linear Regression Evaluation:
MAE: 4.403423299957585
RMSE: 5.295950173485514
R2 Score: 0.9404886749713082

Random Forest Evaluation:
MAE: 4.290159722222221
RMSE: 6.05644133254697
R2 Score: 0.922170029027897
```

```
...
    Prediction for new match (PP_Runs=52, PP_Wkts=1, Venue_Avg=165):
    Linear Regression Prediction: 169
    Random Forest Prediction: 168
```



Conclusion

This project successfully demonstrates the application of data science techniques to a cricket dataset. By using both Linear Regression and Random Forest models, we were able to: - Analyze cricket-related data - Build predictive models - Compare model performance

The **Random Forest Regressor** outperformed Linear Regression, making it a better choice for this dataset.