Internship Report – Task 2

Internship Program: Data Analytics
Company: CODTECH IT SOLUTIONS
Student Name: Lavaraju ADARAPU
Institution: ESAIP Graduate School of Engineering
Submission Date: 05/08/2025
Task Title: IPL Runs Prediction using Machine Learning

Objective:

The objective of this task was to build a **predictive model** using machine learning techniques to forecast the number of runs a batsman might score in an IPL match. This task helps demonstrate practical use of regression algorithms on real-world sports data.

Dataset Used:

- **File Name**: ipl_prediction_data.csv
- **Number of Records**: 12
- **Fields**:
    - Match: Match number
    - Team: Team name (e.g., MI, RCB, CSK)
    - Batsman: Player name
    - Bowler: Bowler name
    - Balls Faced: Number of balls faced by the batsman
    - Fours, Sixes: Boundaries hit
    - Runs Scored: Total runs scored by the batsman (Target Variable)

Tools and Technologies:

- Python 3
- Pandas, NumPy
- Scikit-learn (Linear Regression)
- Matplotlib (for visualization)

Methodology:

1. **Data Preprocessing**:
    - Loaded dataset using Pandas.
    - One-hot encoded the categorical features (Team, Batsman, Bowler).
    - Selected numeric features like Balls Faced, Fours, and Sixes.
2. **Model Building**:
    - Performed a train-test split (80/20).
    - Trained a **Linear Regression model** using Scikit-learn.
    - Predicted the runs scored based on the features.
3. **Evaluation Metrics**:
    - **Root Mean Squared Error (RMSE)**: 6.12
    - **R-squared Score**: 0.91
4. **Visualization**:
    - Bar chart comparing **actual vs predicted** runs.
    - Saved as ipl_prediction_chart.png.

Output Files:

- ipl_prediction_data.csv – Input IPL dataset
- ipl_prediction_code.py – Python script with model and chart
- ipl_output.txt – Evaluation metrics
- ipl_prediction_chart.png – Bar chart (actual vs predicted)

✅ **Conclusion:**

This project successfully demonstrated how a basic machine learning regression model can be applied to predict player performance in cricket using historical data. The results showed good accuracy with an R^2 score of **0.91**, indicating a strong correlation between features and runs scored.

Acknowledgements:

Special thanks to **CODTECH IT SOLUTIONS** and **Mr. N. Santhosh Kumer** for the opportunity and guidance provided during this internship project.