Dhirubhai Ambani
Institute of Information and Communication Technology

Documentation

for

CP-03

By Group -16

PROJECT PIPELINE

Understanding of Data (Problem Statement)

Data Preprocessing - checked null values, attribute types, duplicate values etc

Data Visualization -

1. *Distribution plots*
2. *Scatter Plots*
3. *Box Plots*
4. *Count Plots*
5. *Barplots*
6. *Heatmap*

Contributions :

Chinmaya (202218054)	TASK 2 and 3 : Predicting Finalist and Winner (Multiclass Classification)
Riya (202218049)	TASK 1 : Preprocessing, EDA and Predicting cricketers likely to achieve the highest runs, wickets, and catches along with their respective countries.
Swayista (202218035)	TASK 2 and 3 : Predicting Finalist and Winner (Regression)
Asish (202218022)	TASK 2 and 3 : Predicting Playing 11 for Finalist
Kashish (202001425)	TASK 2 and 3 : Scrapping, EDA, Documentation

Task 1 :

FILE : Data_Mining(Scrapped)_CP_03_Riya.ipynb

ML pipeline

- *Feature selection using correlation analysis*
- *Splitting of data*
- *Scaling of data*
- *Model Fitting*
  - → Linear Regression
  - → Decision tree
  - → Random Forest Regression
  - → ANN
- *Model Evaluation and Testing*

Problem Statement :

1. **Predicting the batsman who will score most runs in the tournament.**

- **The data has been scraped from the provided link -** *https://www.espncricinfo.com/records/tournament/batting-most-runs-career/icc-cricket-world-cup-2023-24-15338*

- **Dataset Description for most runs :**

  - **Player:** The name of the cricket player who took the wickets.
  - **Country:** The country for which the player represents or plays international cricket.
  - **Span:** The time period (range of years) during which the player's career spanned.
  - **Matches:** The total number of matches the player participated in.
  - **Innings:** The total number of bowling innings bowled by the player.
  - **Balls:** The total number of balls bowled by the player.
  - **Overs:** The total number of overs bowled by the player. An over consists of six legal deliveries.
  - **MDNS:** The total number of maidens bowled by the player. A maiden over is an over in which no runs are scored.
  - **Runs**: The total number of runs conceded by the player.
  - **Wickets:** The total number of wickets taken by the player.
  - **BBI (Best Bowling in an Inning)**: The player's best bowling performance in a single inning, represented as the number of wickets taken for the fewest runs.

- **AVE (Bowling Average):** The average number of runs conceded per wicket taken by the player.
- **Econ (Economy Rate):** The average number of runs conceded per over bowled by the player.
- **SR (Strike Rate):** The average number of balls bowled per wicket taken by the player.
- **Fours**: The total number of times the player's deliveries were hit for four runs by the batsmen.
- **Fives:** The total number of times the player took five wickets in a single inning.

	Task	Model Name	Mean Squared Error	Mean Absolute Error	R-squared	Batsman	Country
0	Batsman Making Most Runs	Linear Regression	178.072010	10.021132	0.991364	RG Sharma	IND
1	Batsman Making Most Runs	Random Forest	353.113566	11.401379	0.979047	R Ravindra	NZ
2	Batsman Making Most Runs	Decision Tree	573.448276	15.724138	0.966399	DA Warner	AUS
3	Batsman Making Most Runs	ANN (Batch Size: 64, Learning Rate: 0.1)	79.361768	5.989191	0.992180	SA Yadav	IND

2. Predicting the player who will have the most catches in the tournament.

- **The data has been scraped from the provided link -** *https://www.espncricinfo.com/records/tournament/fielding-most-catches-career/icc-cricket-world-cup-2023-24-15338*

- **Dataset Description for most catches :**
    - **Player:** The name of the cricket player who took catches.
    - **Country:** The country for which the player represents or plays international cricket.
    - **Span:** The time period (range of years) during which the player's career spanned.
    - **Matches:** The total number of matches the player participated in.
    - **Innings:** The total number of fielding innings the player was involved in.
    - **CT (Catches):** The total number of catches taken by the player.
    - **Max:** The maximum number of catches taken by the player in a single match.
    - **Ct/Inns (Catches per Inning):** The average number of catches taken by the player per fielding inning.

	Task	Model Name	Mean Squared Error	Mean Absolute Error	R-squared	Player	Country
0	Most Catches	Linear Regression	0.318335	0.449491	0.912725	C Green	AUS
1	Most Catches	Random Forest	0.602705	0.501500	0.834762	C Green	AUS
2	Most Catches	Decision Tree	2.600000	0.800000	0.287183	C Green	AUS
3	Most Catches	ANN (Batch Size: 32, Learning Rate: 0.01)	0.053335	0.108997	0.985378	C Green	AUS

3. **Predicting the bowler who will have the most wickets in the tournament.**

- **The data has been scraped from the provided link -** *https://www.espncricinfo.com/records/tournament/bowling-most-wickets-career/icc-cricket-world-cup-2023-24-15338*

- **Dataset Description for most wickets :**
    - **Player:** The name of the cricket player who took the wickets.
    - **Country:** The country for which the player represents or plays international cricket.
    - **Span:** The time period (range of years) during which the player's career spanned.
    - **Matches:** The total number of matches the player participated in.
    - **Innings:** The total number of bowling innings bowled by the player.
    - **Balls**: The total number of balls bowled by the player.
    - **Overs:** The total number of overs bowled by the player. An over consists of six legal deliveries.
    - **MDNS:** The total number of maidens bowled by the player. A maiden over is an over in which no runs are scored.
    - **Runs:** The total number of runs conceded by the player.

- **Wickets**: The total number of wickets taken by the player.
- **BBI (Best Bowling in an Inning):** The player's best bowling performance in a single inning, represented as the number of wickets taken for the fewest runs.
- **AVE (Bowling Average):** The average number of runs conceded per wicket taken by the player.
- **Econ (Economy Rate):** The average number of runs conceded per over bowled by the player.
- **SR (Strike Rate):** The average number of balls bowled per wicket taken by the player.
- **Fours:** The total number of times the player's deliveries were hit for four runs by the batsmen.
- **Fives:** The total number of times the player took five wickets in a single inning.

	Task	Model Name	Prediction Name	Prediction Country	Mean Squared Error	Mean Absolute Error	R-squared
0	Most Wickets	Linear Regression	GD Phillips	NZ	4.162301	1.562096	0.784194
1	Most Wickets	Random Forest	GD Phillips	NZ	0.997218	0.866471	0.948296
2	Most Wickets	Decision Tree	GD Phillips	NZ	2.117647	1.058824	0.890205
3	Most Wickets	ANN(Batch Size: 128, Learning Rate: 0.01)	TA Boult	NZ	1.959607	1.030218	0.947698

Task 2.1 and 3 : Predicting the Finalist Teams and Winner

- You are required to predict the two finalist teams in the ICC Cricket World Cup 2023

DATASET DESCRIPTION

1. ICC CWC 23 Points Table (Scrapped)
Dataset Link : Web_Scrapping_DM_Challenge.ipynb

Source: The dataset was collected through web scraping from Cricbuzz, a prominent sports platform. It contains cricket performance details for various countries, including the number of matches played, matches won and lost, points earned, and net run rate. The use of web scraping allows for a comprehensive and up-to-date overview of cricket statistics, providing valuable insights into the performances of different teams in specific matches.
Purpose: Comprehensive overview of cricket performances by different nations.

2. ODI Men's Cricket Match Data (2002-2023)
Dataset Link :
https://www.kaggle.com/datasets/utkarshtomar736/odi-mens-cricket-match-data-2002-2023

Description : This dataset provides comprehensive information about ODI cricket matches, making it suitable for analysis and research in the field of cricket statistics and performance evaluation.

ML pipeline

- *Extracting semifinalist teams from scrapped points table*
- *Feature selection using correlation analysis*
- *Splitting of data*
- *Scaling of data*
- *Model Training on historic data*

→ Logistic Regression
→ Decision tree Classifier
→ Random Forest Classifier
→ ANN

- *Model Evaluation and Testing*
- *Prediction for finalist*
- *Prediction For Winner*

We have used two approaches:

1. Considering the problem statement as Multiclass Classification task

File : Final_ICC_WC23_predictions.ipynb

Prediction :
Finalists :

```
logreg Predicted Finalists: ['India', 'India']
rf Predicted Finalists: ['Australia', 'South Africa']
dt Predicted Finalists: ['New Zealand', 'South Africa']
ANN Predicted Finalists: ['India', 'Australia']
```

Choosing ANN Model Predicted finalists
Based on the finalists
Winner :

```
ANN Predicted Finalists: ['Australia']
```

2. Considering the problem statement as Regression task

File: Final_Regression_TaskDM_3.ipynb

Prediction :

Task 2

Model Applied	Predicted_Winner_Encoded
ANN	[1,2]

TASK 3

Model Applied	Predicted_Winner_Encoded
Polynomial Linear Regression	2
Random Forest	2
XG Boost	2

Performing Regression Task for both Task 2 and Task 3

1. India is encoded as 1
2. Australia is encoded is 2
3. South Africa is encoded as 3
4. New Zealand is encoded as 4

Task 2.2: Predicting the finalist teams

File :

Batting Dataset Link: : https://drive.google.com/file/d/1v6YTvUfGlPnohr0DtIcBQ2vQmAAtt-jY/view?usp=sharing

Bowling Dataset Link: https://drive.google.com/file/d/1hs5pvLrj-vpCWEfRPSnFgK58ZOdiS1Nr/view?usp=sharing

Source:The dataset presented here was obtained through the process of web scraping from ESPNcricinfo, a prominent and reliable source for comprehensive cricket-related information. Web scraping is a technique that automates the extraction of data from web pages, allowing us to compile a detailed dataset on various aspects of cricket matches. ESPNcricinfo, being a renowned platform, provides up-to-date and in-depth statistics, including the number of matches played, matches won and lost, points earned, and net run rate for each country. Through this web scraping process, we aim to deliver accurate and timely insights into the performance of cricket-playing nations, enabling a thorough analysis of team dynamics and tournament standings.

Prediction:

Prediction of playing eleven for Team India:-

```
The predicted 11 for India(1st Finalist):
V Kohli
RG Sharma
SS Iyer
KL Rahul
Shubman Gill
RA Jadeja
SA Yadav
JJ Bumrah
Mohammed Shami
Kuldeep Yadav
Mohammed Siraj
```

Prediction of playing eleven for Team Australia:-

```
The predicted 11 for Australia(2nd Finalist):
DA Warner
MR Marsh
GJ Maxwell
M Labuschagne
SPD Smith
JP Inglis
TM Head
JR Hazlewood
A Zampa
MA Starc
PJ Cummins
```