COGNITIVE ANALYTICS (CSBA3009)

PROJECT REPORT

End to End Cricket Data Analytics using Web Scraping and Python

School of Computer Science

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

Name	SAP ID	Specialization
Abhishek Joshi	500090966	AI&ML (H) – B1

Submitted To:
Dr Sugandha Sharma

Title: Comprehensive Analysis of T20 World Cup Cricket Data: An End-to-End Sports Data Analytics Project

Abstract: This extensive project delves into the realm of sports data analytics, focusing on the T20 World Cup cricket data from 2022. It encompasses various stages, starting with web scraping using Bright Data's enhanced interface, followed by data collection, cleaning, transformation using Python's Pandas and NumPy libraries, and insightful visualization using Matplotlib. The goal is to derive meaningful insights and select the best 11 players team capable of challenging even extraterrestrial opponents.

1. Introduction: Cricket is a widely followed sport globally, generating vast amounts of data during tournaments like the T20 World Cup. Analyzing this data offers valuable insights into player performance and strategic decision-making for team composition. This project aims to showcase the power of data analytics in the context of sports, particularly cricket.

2. Problem Statement: The challenge involves analyzing the T20 World Cup cricket data retrieved from the ESPN Cricinfo website, with a focus on creating visualizations using Matplotlib and selecting the best 11 players team based on performance metrics.

3. Requirement Scoping: The project scope includes web scraping using Bright Data's interface, data cleaning, transformation using Python Pandas and NumPy, and visualization using Matplotlib to derive insights and aid in player selection.

4. Data Collection: Web scraping was employed to gather comprehensive data from the ESPN Cricinfo website, including match results, batting and bowling summaries, and player information. The collected data was stored in JSON and CSV formats.

5. Data Cleaning and Transformation: Utilizing Python Pandas and NumPy, the collected data underwent rigorous cleaning processes, including handling missing values, standardizing data formats, and performing statistical operations for data normalization and scaling.

6. Data Modeling and Parameters Building: While Power BI is a powerful tool for data visualization, this project focused on utilizing Matplotlib for creating visualizations such as bar charts, scatter plots, and histograms. Parameters were built using NumPy for dynamic calculations and filtering.

7. Visualizations Using Matplotlib: The project showcases a wide range of visualizations using Matplotlib, including:

- Bar charts depicting top batsmen by runs and top bowlers by wickets.
- Scatter plots illustrating the relationship between runs scored and strike rates for batsmen.
- Histograms displaying the distribution of player performances in batting and bowling.

8. Insights and Player Selection: The insights derived from the visualizations aided in selecting the best 11 players team, considering factors such as batting and bowling performance, player roles, and overall team balance.

9. Conclusion: This project exemplifies the application of data analytics in sports, specifically cricket, highlighting the importance of data-driven decision-making in team selection and strategy formulation. The visualizations created using Matplotlib add depth and clarity to the analysis, making it accessible and insightful for cricket enthusiasts, coaches, and analysts.

10. Future Scope: Future enhancements could include advanced statistical analysis using machine learning algorithms, real-time data integration for live match analytics, and interactive visualizations for deeper exploration of player performance trends.

This end-to-end sports data analytics project showcases technical proficiency in web scraping, data cleaning, transformation, and visualization using Python libraries such as Pandas, NumPy, and Matplotlib, making it a valuable asset for any data science portfolio.