Cricket Data scraping and analysis for robust data-driven decisions

LAIBA NADEEM

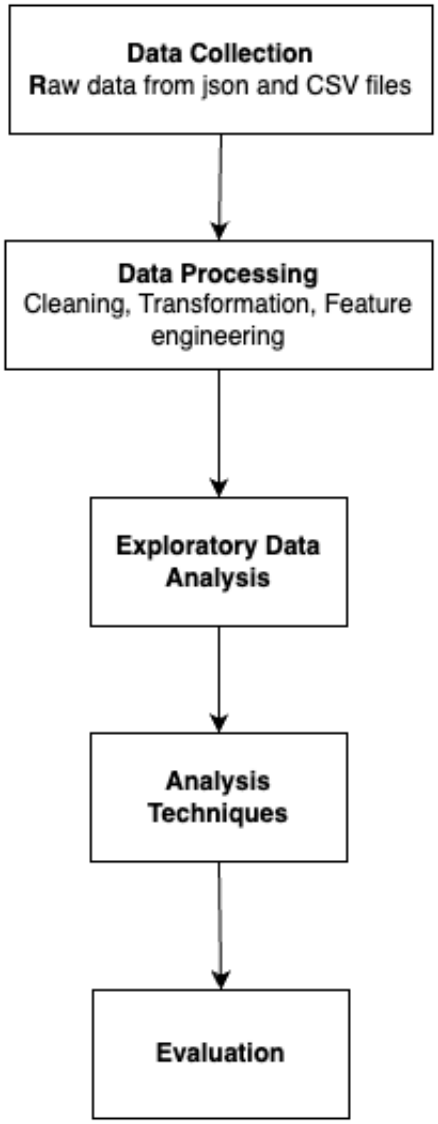
UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 3

3.1 **Introduction:**

The data analysis used in this research to study cricket data to provide solutions to some of the challenges including team selection, evaluation of players, and improving match strategies is described in this chapter. The methodology covers data collection of different types of sources, data cleaning to make them standardized for analysis, and usage of various analytical tools to draw useful inferences. This approach increases the validity and reliability of the study and follows the goals highlighted in the previous chapters of this dissertation.

3.2 **Research Framework**



The research was conducted in five key phases:

Data Collection: Collection of data at different levels: both in terms of structure and content.

Data Preprocessing and Preparation: Cleansing of data and making the collected data in a format that would be easy to use.

Exploratory Data Analysis: Exploratory data analysis to gain knowledge regarding the numerous distributions and relations as well as the first appearances of the data.

Analysis Techniques: Optimization of certain expanded machine learning as well as statistical modeling techniques.

Evaluation: Thus, we evaluate the quality and efficiency of the models up to their accuracy.

By following this phased approach, no aspect is left uncovered, and the loops that can be run in each phase guarantee progressive refinement.

3.3 **Data Collection**

The information for this study was obtained in scraped JSON files and CSV datasets containing information about matches, players, batting and bowling statistics, and results.

Sources:

JSON files: It comes usually in the form of web scraping which involves the use of Python libraries such as BeautifulSoup and Selenium.

CSV datasets: Cricket statistics existing databases before the analysis.

Match details: Participating teams, the result of a particular match (directing the winner, the number of Goals), the date and location of the match.

Player profiles: There are various items including names and surnames and batting and bowling positions and teams and roles.

Performance metrics: Total matches, number of innings, number of fours and sixes, total wickets, total runs conceded, bowling economy rate, and batting strike rate.

The datasets include historical matches as well as modern ones to get a better view of the matches before analyzing them.

3.4 **Data Preprocessing**

To ensure data quality and consistency, the following preprocessing steps were performed:

Cleaning:

Removing additional records, such as null values for players who are not participating allows the list to be concise to other managers. Subsequently, the author also filters duplicates from the results to eliminate distortion of results. Be it in more complicated cases where formats have to be standardized, for example, from JSON to CSV.

Transformation:

Categorical features transformation (for example, team names, and player positions) into formats understandable for machine learning algorithms. These include the scaling of numerical data since the activity requires consistent scaling across numerical sites.

Feature Engineering:

From them, we derive other features like player efficiency scores, match impact etc, and momentum. Developing overall means such as means of performance by the player or the team per season or tournament.

Data Integration:

Combining data sets from one or more sources based on keys that include match IDs and players' names where necessary.

3.5 **Exploratory Data Analysis**

- Before applying advanced analytics, exploratory data analysis (EDA) was conducted to:

- Highlight different trends and trends in batter and bowler rates.

- A major area is to analyze the contributions of the players based on the type of match and conditions it is played.

- Understand how various performance measures, [for example, strike rates compared with economy rates] are distributed.

- Identify trends that will warrant further research to check the validity of findings.

- To also better interpret the data histograms, scatter plots, and heat maps were employed.

3.6 **Analytical Techniques**

The study employed a combination of machine learning, statistical modeling, and data visualization techniques:

Team Selection:

- Assigning players to groups through the K-means clustering technique so that the characteristics of the players match the overall team and the players' prior games.

- Adaptive methods to delicately adjust the composition of a professional team to improve its efficiency ex. Genetic algorithms.

Performance Prediction:

- Random Forest, alongside Logistic Regression models, to forecast performance variation of individuals and teams across conditions.

- Regression analysis for evaluating the changes of a player throughout the tournaments.

Real-Time Insights:

- Internet of things- based analytics to evaluate the movements and fatigue rates of the players and motions during shots.

- Feeding of data collected by the sensors together with the existing models when the matches are ongoing.

3.7 Evaluation Metrics

To evaluate the performance of the models, the following metrics were used:

Accuracy: The proportion or percentage of correct forecasts or classification results by the specimen.

Precision and Recall: Measures for the assessment of prediction accuracy in general and cases with the existence of class imbalance.

F1-Score: The overall mean of precision and recall coefficients which could provide a fine tradeoff between true positive and false positive rates and true negative and false negative rates.

Cross-Validation: The main problem to be addressed was how to guarantee that a model works properly for different datasets.

Area Under the Curve (AUC): Applied in assessing the ability of classification models to correctly separate between classes.

3.8 Tools and Technologies

The research utilized the following tools and platforms:

Programming: Popular languages: Python (libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn).

Visualization: Power BI and other libraries of Python.

Data Management: Processing framework for the two structured data formats which are JSON and CSV.

Machine Learning: includes procedures for clustering, regression, and classification as well as for model evaluation.

Real-Time Processing: Challenges and opportunities for implementing frameworks for inserting IoT sensor data in the match analytics system.

3.9 Challenges and Mitigation

Challenge 1: Lack of data and data perturbing.
Solution: Along with missing values, special data imputation methods, and strict validation were applied.

Challenge 2: The integration of data from different formats.
Solution: Synchronized all data in a similar format (CSV) and employed good quality data merge approaches.

Challenge 3: Considering the computational cost for the suggested model.
Solution: Optimized algorithms and used structurally sound cloud resources for the evaluation.

Challenge 4: Handling imbalanced datasets.
Solution: Techniques that were used include SMOTE (Synthetic Minority Oversampling Technique) to balance this data.

3.10 Chapter Summary

This chapter focused on the description of the utilized methodological approach, employing data collection, preprocessing, exploratory analysis, advanced analysis, and assessment. The phased approach guarantees a structured analysis of the cricket data for insights about the subject of analysis; thus, filling the gaps that are stated while reviewing the literature. In this research, other sophisticated techniques and equipment have been incorporated to enhance the knowledge of data science in the game of cricket.