Cricket Data Analysis Report

Introduction

Cricket is a sport that heavily relies on individual player performances, which significantly impact match outcomes. This report details the development of an unsupervised learning algorithm designed to predict the top performer in a cricket match based on publicly available data. The algorithm is capable of analyzing player statistics, match conditions, and performance metrics to identify the player likely to have the most significant impact on the match outcome.

Task Overview

A. Data Collection and Preprocessing

A.1. Data Collection

The data for this analysis was meticulously collected from publicly available sources, specifically from Cricinfo's searchable cricket statistics database. This database is a comprehensive resource for cricket records and statistics, covering a wide range of cricket formats, including Tests, One-Day Internationals, Twenty20 Internationals, and women's cricket.

We focused our data collection efforts on the Indian Men’s Cricket Team's One-Day International (ODI) matches played in the 21st century. This selective approach helps

maintain the relevance and specificity of the dataset, allowing us to draw meaningful insights from the analysis.

The data collection process involved extracting information related to individual player statistics (including batting, bowling, and fielding), playing styles (such as batting hand, bowling arm, and bowling style), match conditions (such as Toss Outcome, Toss Decision, Match Result, Day/Night Match, Home/Away/Neutral Match, Opposition, and Ground), and various performance metrics for each match (e.g., Runs Scored, Runs Conceded, Wickets Taken, and Start Date).

A.2. Data Preprocessing:

1. To ensure the accuracy and reliability of our analysis, the collected dataset underwent a comprehensive preprocessing phase. This crucial step involved addressing missing values, handling outliers, and performing necessary data transformations. For instance, we converted the "Start Date" column into the datetime data type, enhancing the dataset's temporal analysis capabilities.
2. To maintain data consistency and integrity, null values were filled with appropriate values, and numerical columns were standardized using z-score standardization. Categorical columns were encoded to make them suitable for machine learning analysis. These preprocessing efforts set the stage for robust and insightful data analysis.
3. By following a systematic data collection and preprocessing approach, we ensured the dataset's quality and suitability for subsequent analyses, model development, and the generation of valuable insights.

B. Feature Selection and Engineering

B.1. Feature Selection

In the process of building our cricket performance prediction model, we meticulously selected a wide array of variables encompassing player statistics, match conditions, and individual player performance metrics within a match. These variables were chosen based on their potential influence on a player's performance and their significance in cricket analytics.

B.1. Feature Engineering

Feature engineering played a pivotal role in augmenting our analysis and predictive capabilities. To empower our model, we introduced a set of carefully crafted derived features that enhance the depth and precision of our understanding of player statistics and their performance. The newly created features include a selection of vital metrics:

1) Player Statistics

- **Centuries To Fifties Ratio**:
    - This feature calculates the ratio of total centuries to total half centuries scored by a player. It serves as a valuable indicator of a player's ability to convert fifties (50 runs) into centuries (100 runs). A higher ratio suggests a more consistent performance in reaching the coveted milestone of 100 runs.
- **Runs Scored Per Boundary**:
    - This metric provides insight into a player's boundary-hitting prowess. It computes the average runs scored per boundary hit, which includes both fours (4s) and sixes (6s). This feature is instrumental in evaluating a player's ability to score efficiently by finding the ropes or clearing the boundary.
- **Runs Scored per Match**:
    - This feature calculates the average runs scored by a player per match. It allows us to gauge a player's typical contribution to their team's total score in each match. This insight is invaluable for assessing and comparing batting performances across various matches.
- **Wickets Taken per Match**:
    - In the context of player performance, this metric measures the average number of wickets taken per match by a player. It provides a comprehensive view of a player's bowling performance, indicating their ability to consistently take wickets in different matches.
- **Bowling Economy Rate per Wicket**:
    - This feature quantifies the economy rate per wicket taken by a bowler. It reveals how economically a bowler can secure wickets. A lower value is indicative of a more efficient bowler who can both contain runs and take wickets effectively.
- **Dismissals-to-Catches Ratio**:
    - This ratio is a crucial metric that calculates the proportion of total dismissals made by a player to the number of catches taken. It enables us to evaluate a player's fielding skills and their contribution to the team's overall fielding performance. A higher ratio suggests a more substantial involvement in dismissals beyond standard catches.

2) Player Performance in a Match

- **Batting Performance**:
  - The Batting Performance metric reflects a batsman's overall performance in a specific cricket match. Its formula is as follows:
  - $$BattingPerformance = (RunsScoredinMatch + DidNotBat)/(NotOut + 1)$$

    This formula combines the total runs scored during the match with any runs scored when the batsman did not get an opportunity to bat. The division by the number of times the batsman remained not out (increased by 1 to prevent division by zero) provides a holistic assessment of a batsman's performance, considering their ability to score runs and their opportunities to bat. A higher Batting Performance value signifies a better performance, indicating the batsman's valuable contributions with the bat during the match.

- **Bowling Performance**:
  - Bowling Performance assesses a bowler's effectiveness in taking wickets while restricting the runs scored by the opposing team. Its formula is defined as:

    $$BowlingPerformance = WicketsTakeninMatch/(RunsConcededinMatch + 1)$$

    This metric reveals a bowler's ability to consistently take wickets and maintain control over the opposition's run rate. A higher Bowling Performance value highlights the bowler's proficiency in both taking wickets and preventing the opposition from scoring freely.

- **Fielding Performance**:
  - Fielding Performance is a composite metric that combines the counts of both catches taken and stumpings made by a player in a match. The formula simplifies to:

    $$FieldingPerformance = TotalCatchesTakeninMatch + TotalStumpingsMadeinMatch$$

This feature allows us to evaluate a player's contribution in the field, including their ability to create dismissal opportunities by taking catches. Additionally, for wicketkeepers, it considers their skill in executing successful stumpings. A higher Fielding Performance value indicates a player's proficiency in this vital aspect of the game, as it directly impacts the dismissal of opposing batsmen and the prevention of runs.

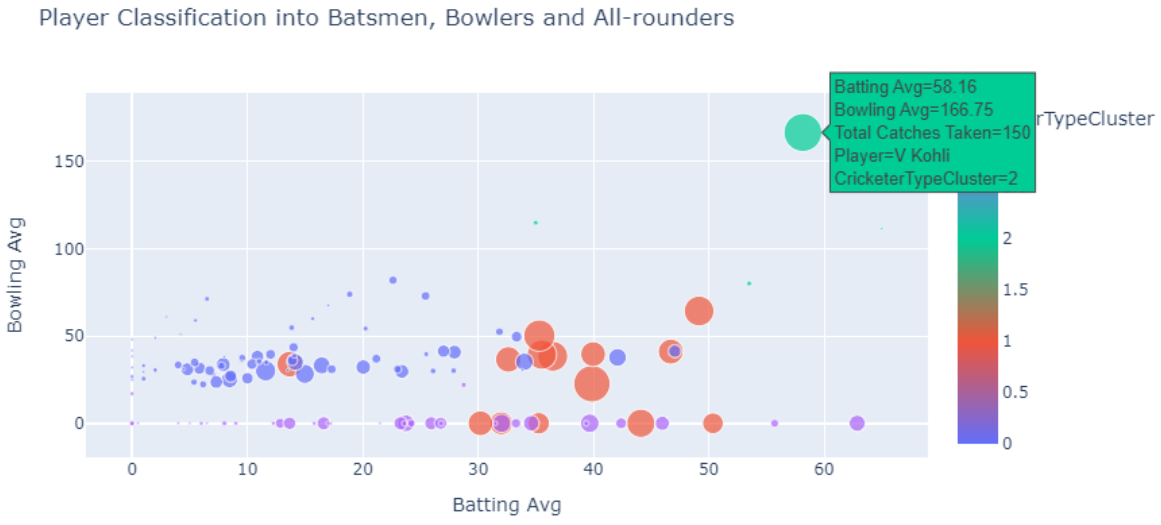
These meticulously engineered features, along with our carefully selected variables, collectively enrich our dataset and form the foundation for our predictive model, enhancing our understanding of player performance and its impact on cricket matches.

C. Clustering Analysis

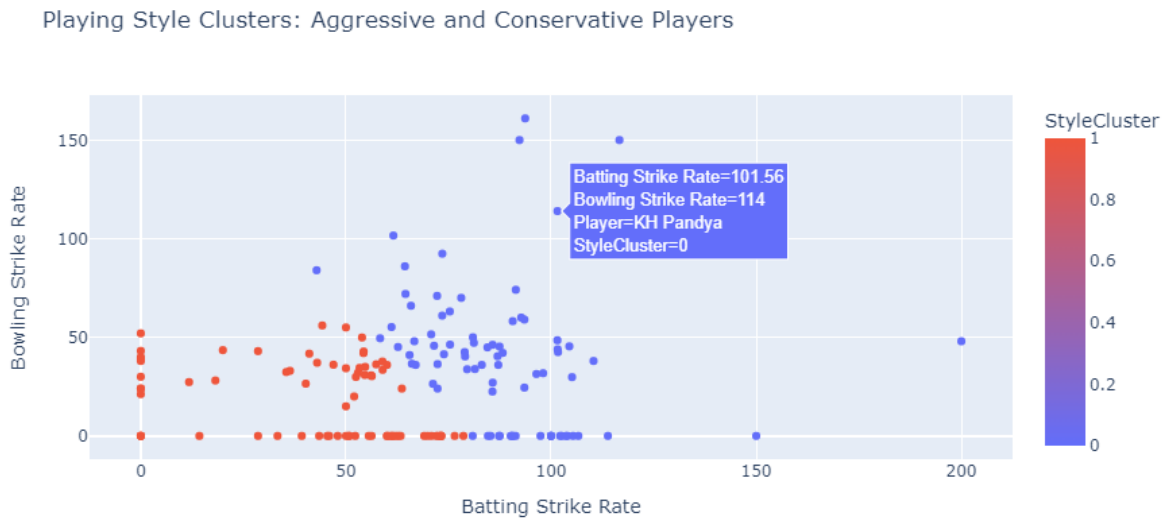
C.1. Player Clustering

The clustering analysis brought insights into grouping players according to their unique playing styles, strengths, and weaknesses. The k-means algorithm was our tool of choice to categorize players into distinct clusters, each characterized by its own set of attributes:

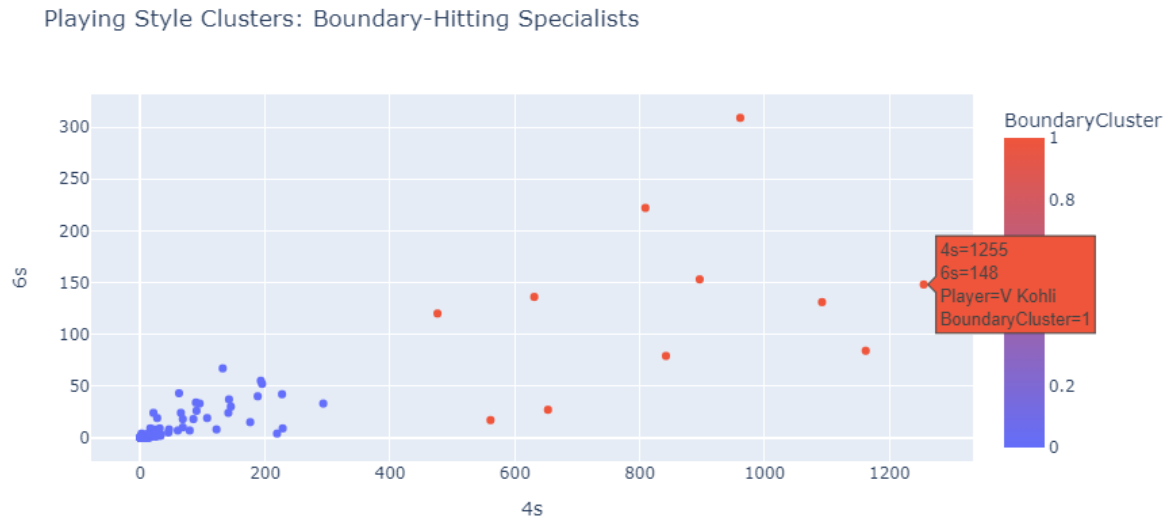
1. **Clustering 1: Player Classification**: In this cluster, we categorized players into Batsmen, Bowlers, and All-rounders based on their performance in key metrics like Batting Avg, Bowling Avg, and Total Catches Taken. This classification allowed us to identify the specialized roles each player embodied on the field.



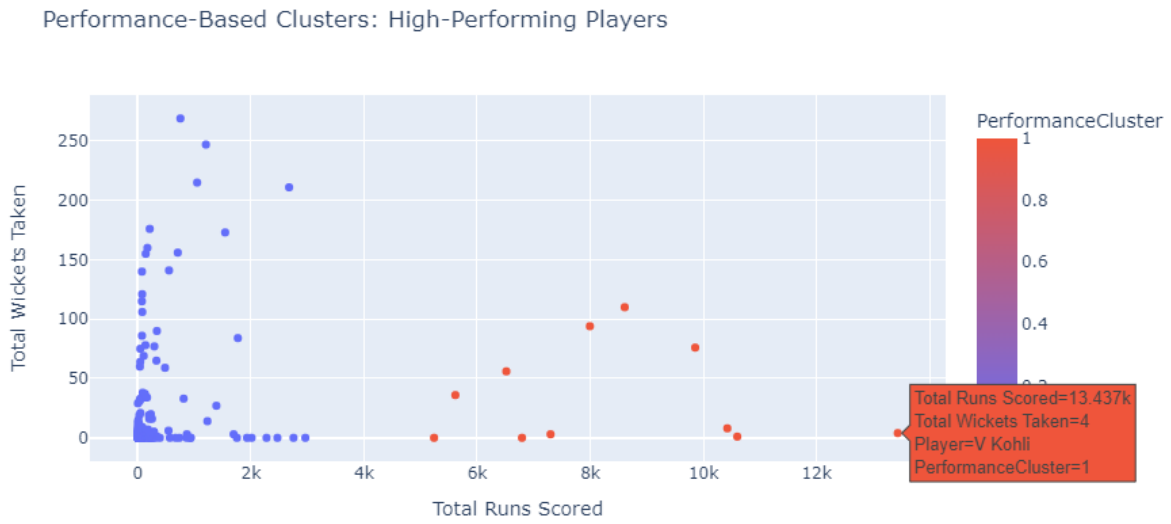
2. **Clustering 2: Playing Style Clusters (Aggressive and Conservative)**: By examining Batting Strike Rate and Bowling Strike Rate, we delineated players into clusters that revealed their playing styles. We identified aggressive and conservative players, each bringing their unique flair to the game.



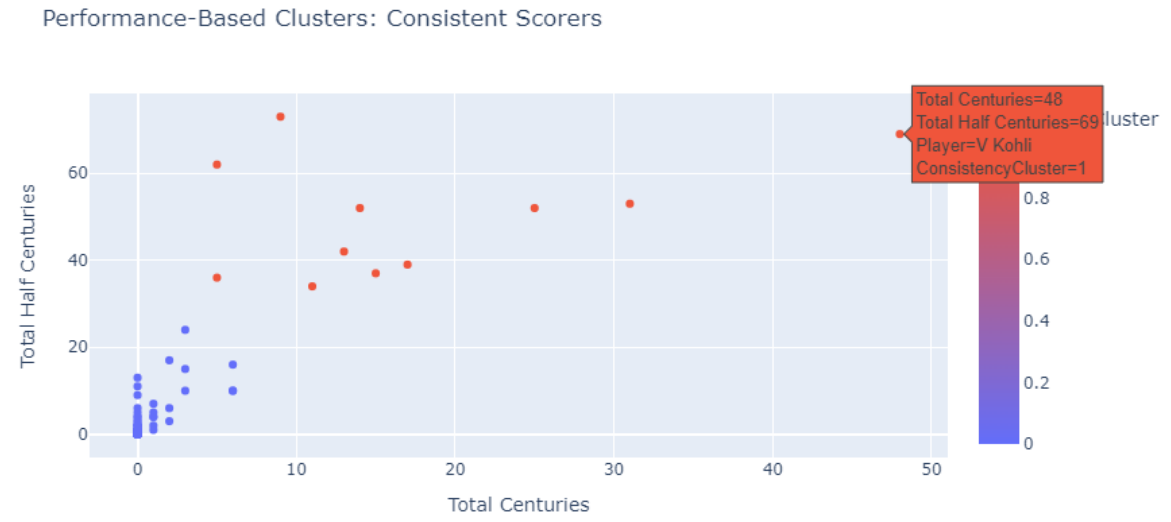
3. **Clustering 3: Playing Style Clusters (Boundary-Hitting Specialists and Others)**: We delved into players' ability to hit boundaries by analyzing Total Sixes and Fours scored by each player. This led to clusters distinguishing boundary-hitting specialists from others.



4. **Clustering 4: Performance-Based Clusters (High-Performing Players and Others)**: Players were grouped into clusters based on their Total Runs Scored and Total Wickets Taken, highlighting high-performing players who stood out from the rest.



5. **Clustering 5: Performance-Based Clusters (Consistent Scorers)**: To uncover players with a penchant for consistency, we scrutinized metrics such as Total Centuries and Half Centuries scored by the players. This clustering strategy helped us identify players with a knack for steady performance throughout their careers.



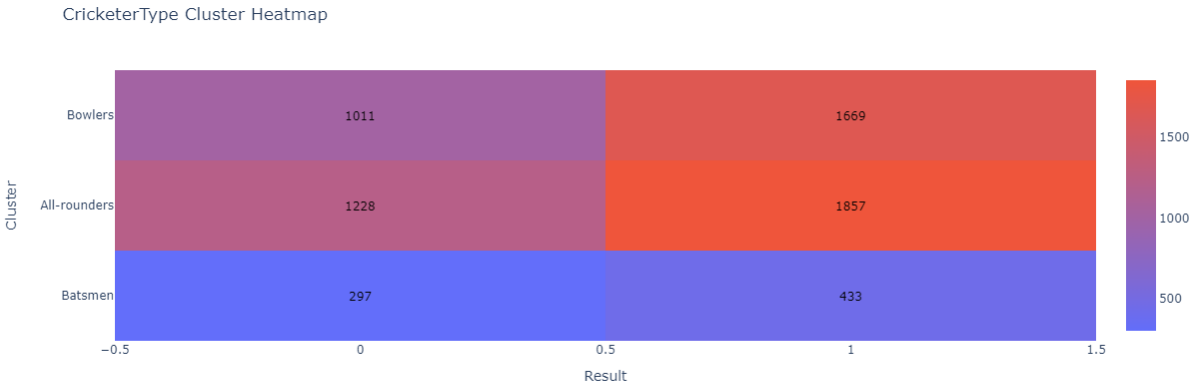
The various clustering analyses enabled us to understand and appreciate the diverse roles that players assume on the cricket field and provided valuable insights into their strengths and areas for improvement.

C.2. Analysis of Clusters

We analyzed how different player clusters correlated with match outcomes and individual player performances. Understanding the impact of player clusters on match results provided

valuable insights.

For e.g. Key Insights Unveiled for ***Cricketer Type Cluster Vs. Match Result****:*



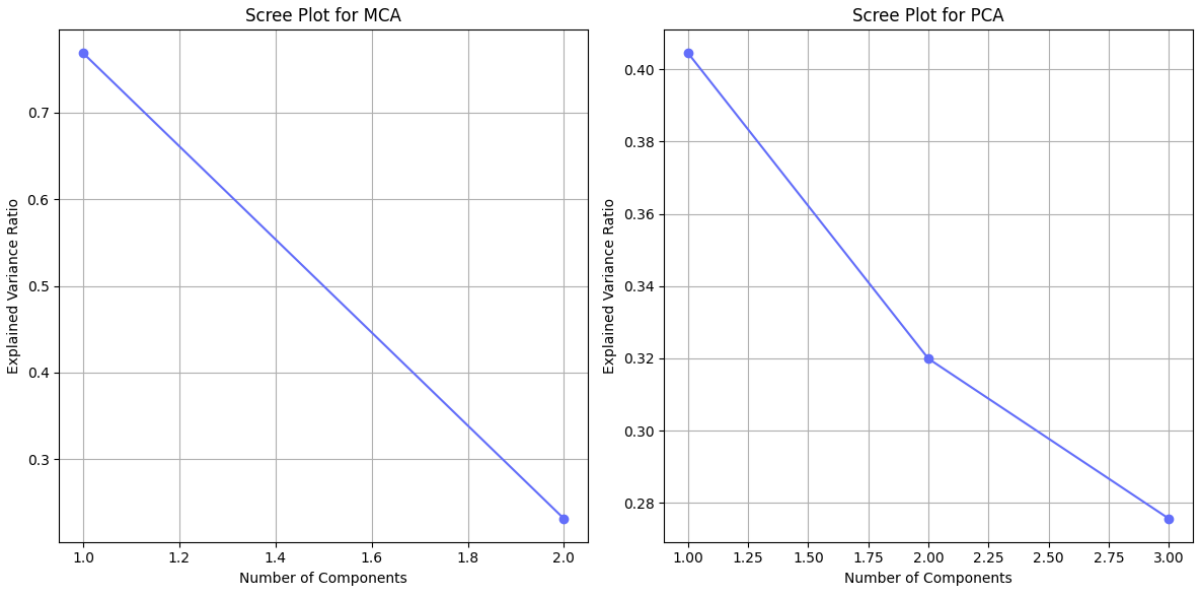
The cross-tab heatmap visualization, as exemplified above, showcases how various clusters of cricketers influence match results. For instance, most of the matches requires all-rounders cricketers to win and good bowlers to take out max wickets.

D. Dimensionality Reduction

Dimensionality Reduction Techniques: This phase of our analysis was dedicated to applying dimensionality reduction techniques to simplify the complexity of our dataset while retaining critical information for our analytical endeavors. Specifically, we leveraged two distinct techniques: Principal Component Analysis (PCA) for numerical columns and Multiple Correspondence Analysis (MCA) for categorical columns.

Optimal Number of Components: Our initial and crucial step was to ascertain the number of components required to capture 90% of the variance within the dataset. This threshold was chosen to ensure that our dimensionality reduction preserved a substantial portion of the original data's variability. It's important to maintain a balance between data reduction and information retention.

Scree Plots for Visualization: To facilitate a clear understanding of the contributions of different components to the total explained variance, we generated scree plots for both MCA and PCA. These scree plots depict the relationship between the number of components and the cumulative percentage of variance they explain. They serve as visual aids for assessing the trade-off between the number of components and the retained variance.



Scree Plot for MCA

From the scree plot for MCA, we can draw the following conclusions:

1. The 1st Component explains approximately 76% of the variance.
2. The 2nd Component accounts for approximately 23% of the variance.

Scree Plot for PCA

From the scree plot for PCA, we can make the following observations:

1. The 1st Principal Component explains about 42% of the variance.
2. The 2nd Principal Component contributes to approximately 32% of the variance.
3. The 3rd Principal Component encapsulates roughly 21% of the variance.

These scree plots provide valuable insights into the distribution of explained variance across components. They are instrumental in determining the optimal number of components needed for our dimensionality reduction, striking a balance between data simplification and information retention. This process ensures that we maintain essential data characteristics while reducing the dataset's complexity, ultimately enhancing the efficiency and effectiveness of our analysis and modeling.

E. Model Building and Evaluation

Model Development: With dimensionality reduction completed, we proceeded to model development. The initial step involved applying a K-Means clustering algorithm to the combined data to label players as top performers or otherwise, classifying them as per the

insights drawn from our analysis. These labels served as a valuable foundation for our subsequent supervised learning model. For our supervised learning algorithm, we selected the Random Forest Classifier, known for its robustness, versatility, and effectiveness, especially in handling class imbalances.

Handling Class Imbalances: Recognizing class imbalances within our dataset, we opted for a more robust scoring strategy, F1 scores, to measure our model's performance. This strategy ensures that the model's evaluation considers the precision and recall rates, providing a comprehensive assessment of its predictive capabilities.

Cross-Validation Score: To train the Random Forest Classifier, we divided the data into an 80:20 split between the training and testing sets. We applied hyperparameter tuning to optimize model performance, utilizing Grid Search Cross-Validation. The resultant best-fit Random Forest Classifier, with its optimized parameters, was then trained on the training set.

Model Evaluation: Our model underwent rigorous evaluation to validate its effectiveness in predicting match outcomes. The following classification report reveals key model evaluation metrics, offering insights into its precision, recall, F1 score, and support.

```
Fitting 5 folds for each of 108 candidates, totalling 540 fits
Best Accuracy: 0.99
Best Hyperparameters: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       935
           1       0.99      0.97      0.98       364

    accuracy                           0.99      1299
   macro avg       0.99      0.98      0.99      1299
weighted avg       0.99      0.99      0.99      1299

Confusion Matrix:
[[931   4]
 [ 10 354]]
```

Model Utility: After successful training and evaluation, our best-performing model, along with key utilities like StandardScaler, PCA, and MCA, was preserved. The model's predictions, match details, and player identifications were systematically stored in a CSV file, ensuring that the insights derived from our analysis could be accessed and leveraged for future decision-making and strategic purposes.

	Match ID	Result	Top Performer	Opposition	Ground	Start Date	Runs Scored in Match	Wickets Taken in Match	Catches Taken in Match	Stumpings Made in Match
57	59	won	Shubman Gill	Bangladesh	Colombo (RPS)	2023-09-15	121	1.0	0.0	0.0
56	58	NaN	Shubman Gill	Sri Lanka	Colombo (RPS)	2023-09-17	27	3.0	0.0	0.0
55	57	won	Shubman Gill	Australia	Mohali	2023-09-22	74	1.0	2.0	0.0
54	56	won	SS Iyer	Australia	Indore	2023-09-24	105	3.0	0.0	0.0
53	55	won	RG Sharma	Australia	Rajkot	2023-09-27	81	0.0	0.0	0.0
52	54	lost	V Kohli	Australia	Chennai	2023-10-08	85	1.0	2.0	0.0
51	53	won	RG Sharma	Afghanistan	Delhi	2023-10-11	131	4.0	0.0	0.0
50	52	won	RG Sharma	Pakistan	Ahmedabad	2023-10-14	86	2.0	0.0	0.0
49	51	won	Shubman Gill	Bangladesh	Pune	2023-10-19	53	0.0	2.0	0.0
48	50	lost	V Kohli	New Zealand	Dharamsala	2023-10-22	95	0.0	2.0	0.0

Here are some of the predicted labels for last 10 matches. The Top Performer shows the predicted top performer for the match played with the opposition team on the given start date. This comprehensive approach to model building and evaluation strengthens the foundation for data-driven decision-making in the world of cricket, providing valuable tools and insights for teams, analysts, and enthusiasts.

F. Interpretation and Recommendation

F.1. Interpretation of Results

Our report delves deep into the findings, decoding the intricate web of cricketing data to unveil valuable insights into the factors that catalyze a player's top performance in cricket matches. These insights transcend mere numbers and statistics, offering a profound understanding of the ever-evolving cricketing landscape.

- *Impact of Player Clusters*: Our analysis illuminates how different player clusters, each with its unique playing style, strengths, and weaknesses, influence match outcomes. The results underscore the significance of assembling a well-balanced team with a strategic mix of batsmen, bowlers, and all-rounders. This impact is not confined to match outcomes alone; it resonates with the individual performances of players within these clusters.
- *Match Conditions Matter*: Match conditions, especially the toss outcome and its ripple effect on the match result, are pivotal. We've unveiled the strategic importance of winning the toss, providing teams with insights into the decision to bat or field first. Match conditions, including the time of day (day or night) and the nature of the pitch (home, away, or neutral), also cast their shadows on player performance.

F.2. Recommendations

The report's significance extends beyond data analysis; it offers actionable recommendations to empower cricket teams and analysts in optimizing their strategies and elevating their game.

- *Strategic Team Composition*: The insights gained from player clusters shed light on the composition of a winning team. By strategically combining batsmen, bowlers, and all-rounders, teams can enhance their performance and chances of victory. The art of selecting the right mix becomes paramount.
- *Toss Decision Strategies*: Winning the toss is not just a stroke of luck; it's a strategic advantage. Teams are recommended to weigh their options carefully, taking into account the pitch conditions, playing style, and the opponent's strengths and weaknesses.
- *Performance-Driven Training*: With the nuances of player performance unveiled, teams can focus their training regimens to amplify their strengths and work on areas that need improvement. This data-driven approach can help players reach their full potential.
- *Pre-Match Tactical Adaptations*: Analysts can provide real-time insights and recommendations to teams during matches, helping them make dynamic decisions in response to evolving game situations. This proactive approach can be a game-changer.
- *Scouting and Recruitment*: Cricket teams can use data-driven insights to scout and recruit players who align with their strategic objectives. Identifying the right talent is a critical step in building a successful team.

The recommendations aren't just theoretical; they are grounded in the wealth of data and analysis encapsulated in this report. They empower cricketing stakeholders to embrace data as a strategic asset and a compass that can guide them towards triumph in the unpredictable world of cricket.

Conclusion

In the world of cricket, where every ball, every run, and every wicket can sway the course of a match, data analysis plays a pivotal role. This report demonstrates how cricket data analysis, driven by unsupervised learning algorithms, can unravel the intricacies of the game, shedding light on factors that influence player performance and match outcomes.

The ability to predict top performers based on player clusters and match conditions is a significant leap forward in the realm of cricket analytics. It's not just about runs scored or wickets taken; it's about strategy, decision-making, and game-changing moments.

This endeavor has not only deepened our understanding of the sport but has also brought forth actionable recommendations for cricket teams and analysts. It offers a roadmap to strategic success, encouraging teams to make data-backed decisions in team composition, toss strategies, and player recruitment.

As the digital age transforms the cricketing landscape, the fusion of data and cricket is more powerful than ever. It's not just about numbers; it's about insights. It's not just about past performance; it's about predicting the future. Cricket data analysis is more than a tool; it's a game-changer.

The journey of cricket data analysis is a dynamic and ever-evolving one, mirroring the sport itself. As the game continues to thrill and inspire fans worldwide, the promise of data-driven cricket remains an exciting horizon. With each match and each player, new data emerges, offering fresh perspectives and opportunities.

This report is not a conclusion but a beginning—a beginning of a data-driven era in cricket. It underscores the power of data analytics in enhancing the sport and shaping its future.

In the world of cricket, every delivery brings an opportunity. With data analytics, every data point brings insight. It's time for cricket to embrace the digital age, for a new era is dawning, and the pitch is set for data-driven triumphs.

The future of cricket belongs to those who understand it, those who analyze it, and those who harness its hidden potential.

Future Improvements

As with any data analysis project, there are several avenues for further improvement and exploration:

1. **Analysis of Opponent Teams**: To gain deeper insights, we can extend our analysis to consider the factors that lead to victories or losses against specific opponent teams. This would allow cricket teams to tailor their strategies to counter the strengths and exploit the weaknesses of specific rivals.
2. **T20 Match Analysis**: T20 cricket is a dynamic and high-paced format that demands unique strategies and skills. Analyzing T20 matches separately can provide valuable insights into how the Indian cricket team can further enhance their performance in this format. The analysis could focus on specialized skills, team compositions, and strategies that are particularly effective in T20 matches.
3. **Real-Time Data Integration**: For real-time decision-making during matches, integrating live data feeds can be immensely valuable. Real-time data can help teams make immediate tactical adjustments based on the current conditions, player form, and evolving match situations.
4. **Advanced Machine Learning Models**: Exploring more advanced machine learning models, such as deep learning and ensemble methods, can potentially improve prediction

accuracy. These models can capture complex relationships within the data and provide more accurate insights.

References

1. Cricinfo’s Searchable Cricket Data Base: https://stats.espncricinfo.com/ci/engine/stats/index.html?class=2;filter=advanced
2. Kaggle Notebooks:

https://www.kaggle.com/code/sonalgan/indian-cricket-team-analysis/notebook