Asia Cup Match Outcome Prediction:

A Cricket Match Outcome Forecasting using Machine Learning

Prepared and submitted by

Name	Id	Department
MD RAKIBUL ISLAM	20-43862-2	CSE
ATHOY KANTI RAY	20-43259-1	CSE
RAFI REZA HUQ	20-43360-1	CSE
MD. GOLAM RABBANI RAFI	19-41123-2	CSE

Submitted to

MD. SAEF ULLAH MIAH

Date of Submission

17th August 2023

American International University – Bangladesh

Abstract

Cricket is like a puzzle, and teams bring their own game styles and strategies to the table. We're using computer magic to understand these secrets. By looking at past matches, we're trying to find hidden clues that can help us predict future match winners.

We're using some clever computer techniques like cleaning up data, picking important details, and teaching our computer to make smart guesses. But remember, our guesses are based on history and might not catch everything happening in real time.

This project is like getting a secret glimpse into what might happen in upcoming cricket matches. It shows how technology can help us make exciting predictions and talk about the chances of teams in the 2023 Cricket Asia Cup.

Introduction

The aim of "Predicting the Winner of the 2023 Cricket Asia Cup" project to develop an advanced machine learning model with the capability to anticipate the winning team in the forthcoming Cricket Asia Cup scheduled for the year 2023. Using information from past matches from previous Asia Cup tournaments, this project wants to use the smartness of numbers to predict what might happen in the upcoming cricket tournament.

Cricket, the sport we all love, is like a big puzzle with lots of pieces. The Asia Cup brings together teams from different parts of Asia to compete and show off their cricket skills. It's like a big cricket party that gets everyone excited. We're going to use numbers and data from these parties to see if we can make smart predictions about the winners.

Each team has its own way of playing, favorite players, and strategies. They play on different grounds, in different years, and under different conditions. All these things affect who wins and who doesn't. We're going to use computer magic to understand all these factors and try to make predictions about which team might win the matches.

Motivation of the Project

We decided to embark on the "Predicting Asia Cup 2023 Match Winners" project due to our shared passion for cricket and our curiosity about how data can enhance our understanding of the game. This project isn't just about numbers; it's about exploring the exciting connection between sports and technology.

On a personal level, this project is a thrilling opportunity for us to combine our love for cricket with our interest in data analysis. We want to see if we can use data to make accurate predictions about match outcomes. It's like solving a captivating puzzle where each statistic and factor plays

a role in determining the result. Plus, by working on this project, we can develop valuable skills in data preprocessing, machine learning, and model evaluation.

Beyond our personal goals, we believe this project can bring benefits to cricket enthusiasts and the wider community. Imagine being able to impress your friends with predictions about match winners based on data! We hope that our project will make cricket even more engaging and encourage people to explore the fascinating world of sports analytics. Ultimately, we're driven by the idea of uniting our love for cricket and our interest in technology to create something.

Objective of the Project

The primary objective of the "Predicting Asia Cup 2023 Match Winners" project is to develop a robust machine learning model capable of predicting the outcomes of cricket matches in the upcoming Asia Cup tournament. The model will be evaluated on its ability to predict the winner of unseen matches. The accuracy of the model will be reported as a percentage. The project will also explore the following questions:

1. Which features are most important for predicting the winner of an Asia Cup cricket match?
2. How well does the model perform on different types of matches?
3. How can the model be improved to make better predictions?

The project will be completed using the following steps:

1. Data collection and cleaning
2. Feature selection
3. Model training
4. Model evaluation
5. Model deployment

The project will be implemented using the following tools and technologies:

1. Python
2. Pandas
3. NumPy
4. Scikit-learn.

Methodology

We first collected the dataset from the Kaggle website and stored it in Google Drive. We then connected Google Drive to Google Colab, where we performed exploratory data analysis (EDA) to gain insights into the data. This included visualizing different features, dropping unwanted features, handling missing values, handling categorical features, handling feature scaling, and removing outliers. We then calculated feature importance and built a machine learning model using Random Forest Classifier. We tested the model on a holdout set and generated results depending on different evaluation metrics**.**

```
Data Preparation
    Data Collection
    Data Upload
    Data Visualization

Feature Selection
    Drop unwanted data
    Handle missing value
    Final Feature Selection

System Training
    Train our system in three different models

System Optimization
    Test our model and generate results depending on different matrices
    Generate result for the optimal output
```

A. Data Collection

In this project, the data was collected from Kaggle(https://www.kaggle.com/), a popular online platform for data science and machine learning resources. Kaggle provides a diverse range of datasets contributed by the community and data enthusiasts worldwide. For this specific project, the dataset containing historical records of Asia Cup cricket matches was obtained from Kaggle.

The dataset includes information about various attributes related to the Asia Cup cricket matches, such as teams, players, match outcomes, batting statistics, and other relevant features. The data collection process involved downloading the dataset from Kaggle's platform, ensuring its integrity, and preparing it for further analysis and modeling.

By leveraging the wealth of data available on platforms like Kaggle, this project aims to utilize historical match records to build a predictive model for anticipating match outcomes in the upcoming Asia Cup cricket tournament. The collected data serves as the foundation for the subsequent steps of data preprocessing, feature engineering, model training, and evaluation.

B. Data Processing

The collected dataset underwent a comprehensive data processing and cleaning process to ensure its quality and suitability for analysis and modeling. The following steps were taken to process and clean the data:

Handling Missing Values: Any missing values in the dataset were addressed. For numerical columns, missing values were imputed using the mean value of the respective column. Categorical columns were handled by either imputing with the most frequent category or by dropping rows with missing categorical values, depending on the context.

```
Number of missing values in the Run Scored column: Team
Opponent                    0
Ground                      0
Toss                        0
Selection                   0
Run Scored                  0
Wicket Lost                 0
Fours                       0
Sixes                       0
Extras                      0
Run Rate                    0
Avg Bat Strike Rate         0
Highest Score               0
Wicket Taken                0
Given Extras                0
Highest Individual wicket   0
Result                      0
dtype: int64
```

Dropping Irrelevant Columns: Columns that were not relevant to the analysis or prediction task were dropped from the dataset. These columns might include identifiers or features that have no significant impact on the match outcomes.

Encoding Categorical Variables: Categorical variables, such as team names, grounds, toss decisions, etc., were encoded using one-hot encoding. This process transformed categorical data into numerical format, making it suitable for machine learning algorithms.

Splitting into Train and Test Sets

Finally, the processed dataset was split into training and testing sets to facilitate model training, validation, and evaluation.

By carefully processing and cleaning the data, this project aimed to create a reliable and accurate foundation for building predictive models to forecast match outcomes in the Asia Cup cricket tournament.

C. Dataset Description

The dataset used for this project comprises historical match data from the Asia Cup cricket tournament. It provides insights into various aspects of cricket matches, including team

performance, batting and bowling statistics, match details, and outcomes. Below is a description of the dataset structure, including the number of columns and instances, along with visualizations generated using exploratory data analysis (EDA) methods.

Dataset Structure

Number of Columns: The dataset contains multiple columns that capture different attributes and statistics related to cricket matches. These columns include features such as "Team," "Opponent," "Ground," "Toss," "Selection," "Batting Average," "Run Scored," "Fours," "Sixes," and more.

Dataset Visualizations using EDA Methods

Exploratory data analysis techniques, including summary statistics, correlation analysis, box plots, scatter plots, and pair plots, were employed to gain insights into the data's distribution, relationships, and potential patterns.

Box Plots for Categorical Variables

Box plots visualize the distribution of numerical features across different categories, such as team names. They help detect outliers and variations in performance between teams.

These visualizations provide insights into the dataset's characteristics, enabling us to understand the data's distribution, relationships, and potential predictive features. They aid in making informed decisions during data preprocessing, feature selection, and model building stages.

It's important to note that the specific visualizations and insights generated from the dataset will depend on its content and the goals of the analysis. By exploring the dataset using EDA methods, we can uncover valuable information that contributes to the accuracy and reliability of the predictive models developed in this project.

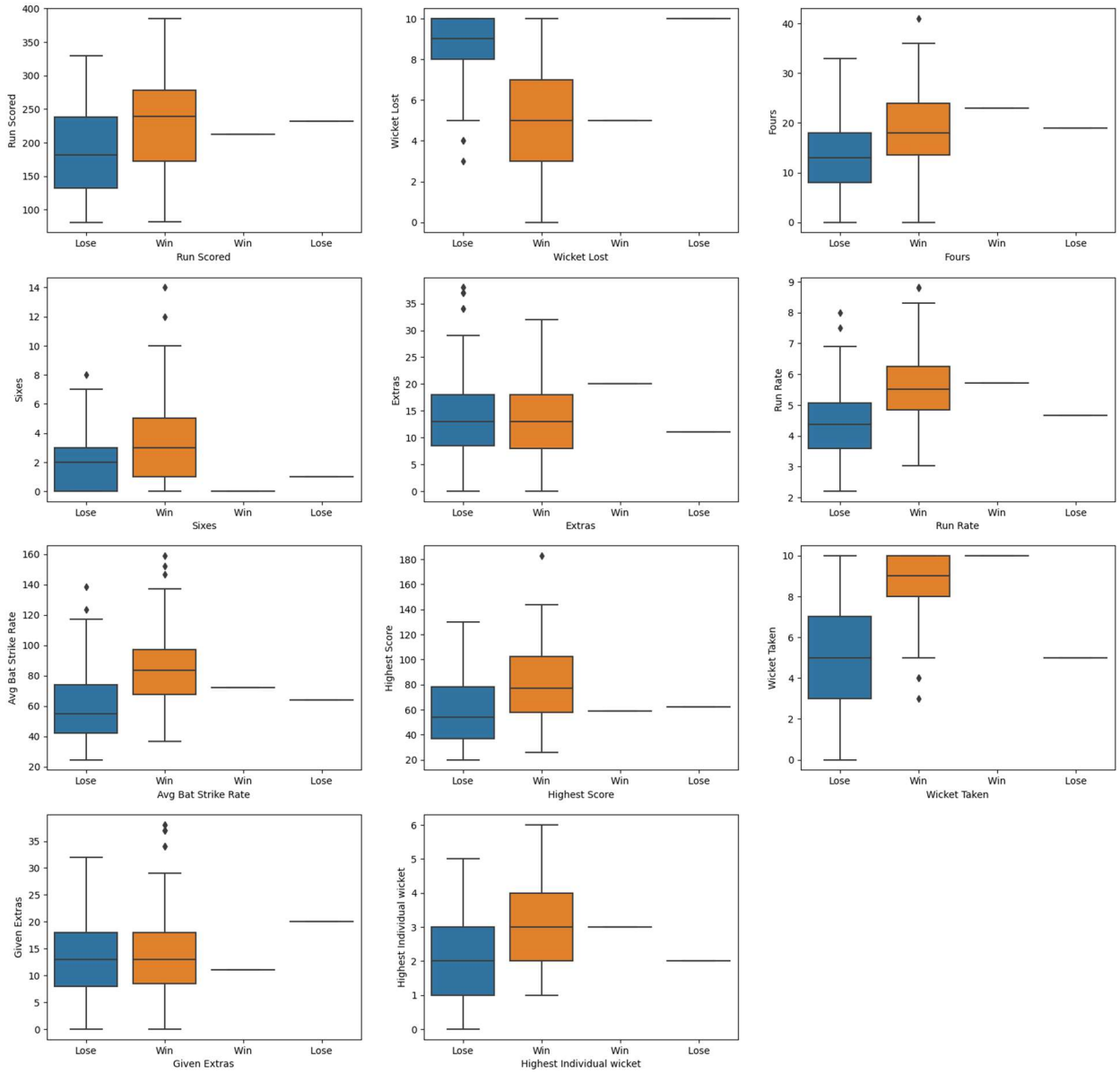


Figure: Boxplots for numerical features by Result

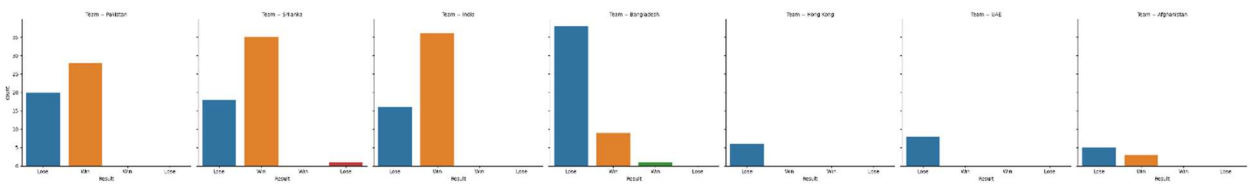


Figure: Distribution of the target variable 'Result'

Correlation Matrix:

A correlation matrix reveals the relationships between numerical features by calculating correlation coefficients. It helps identify potential patterns and dependencies among variables.

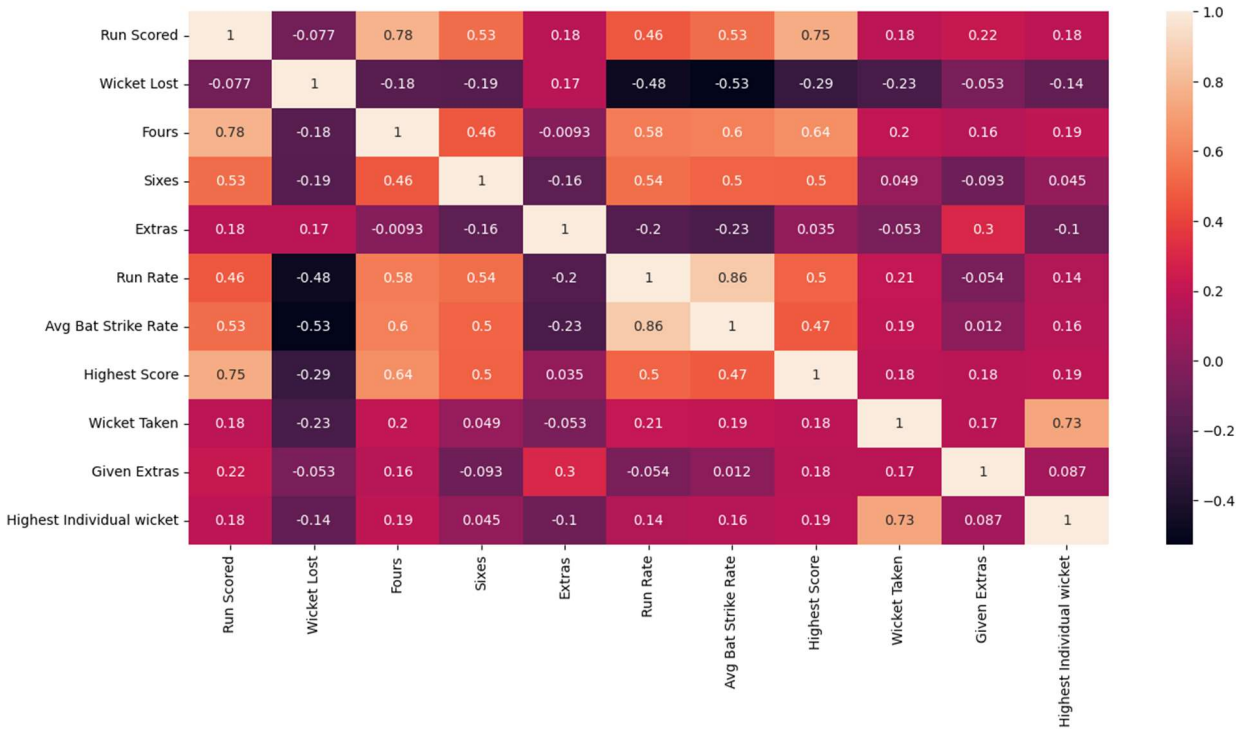


Figure: Corelation Heatmap

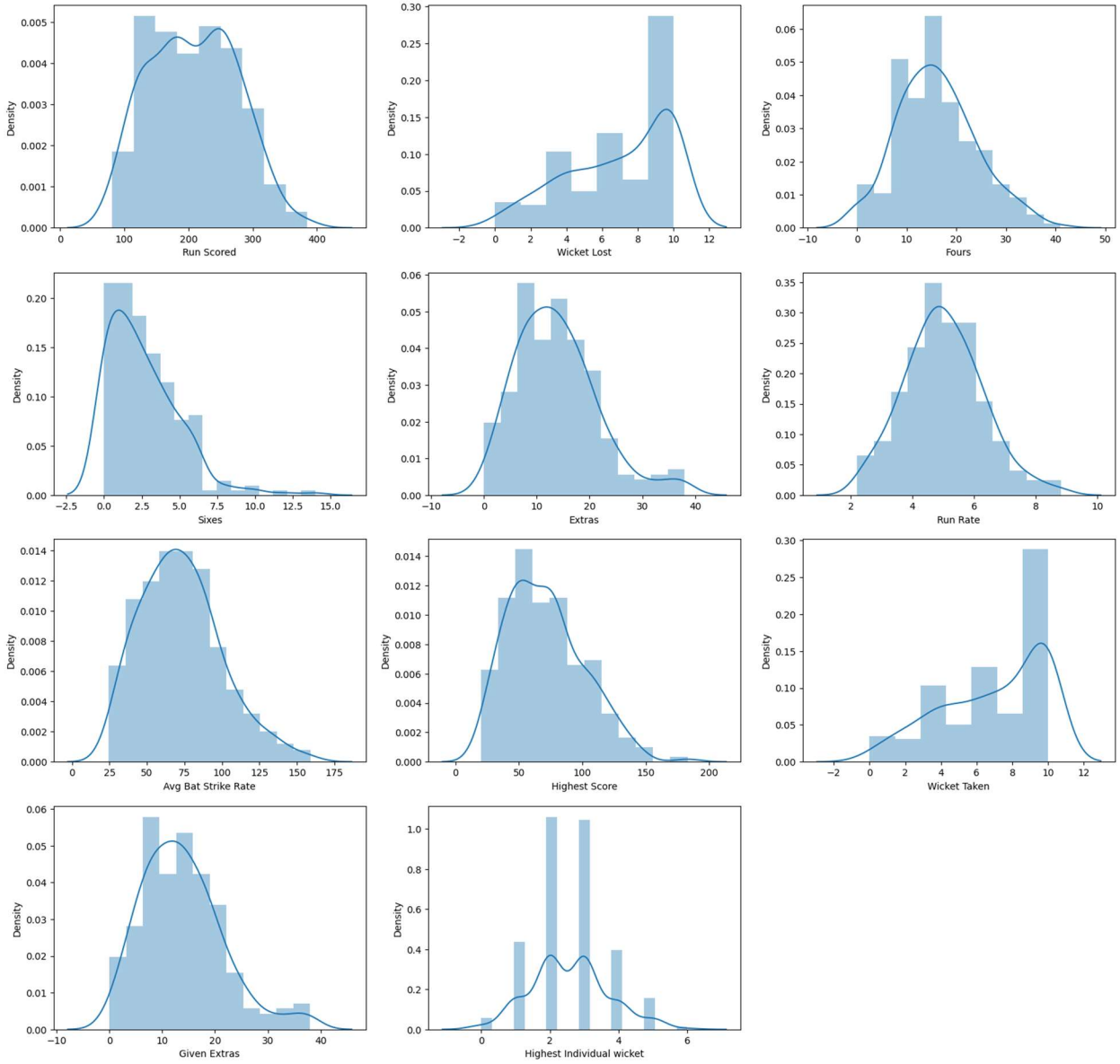


Figure: Distribution of the numerical features

D. Machine Learning Model Development and Evaluation

In this section, we'll delve into the development of the machine learning model for predicting match outcomes in the Asia Cup cricket tournament using the Logistic Regression algorithm. We'll outline the process of model development, including data preprocessing, feature engineering, model training, and evaluation.

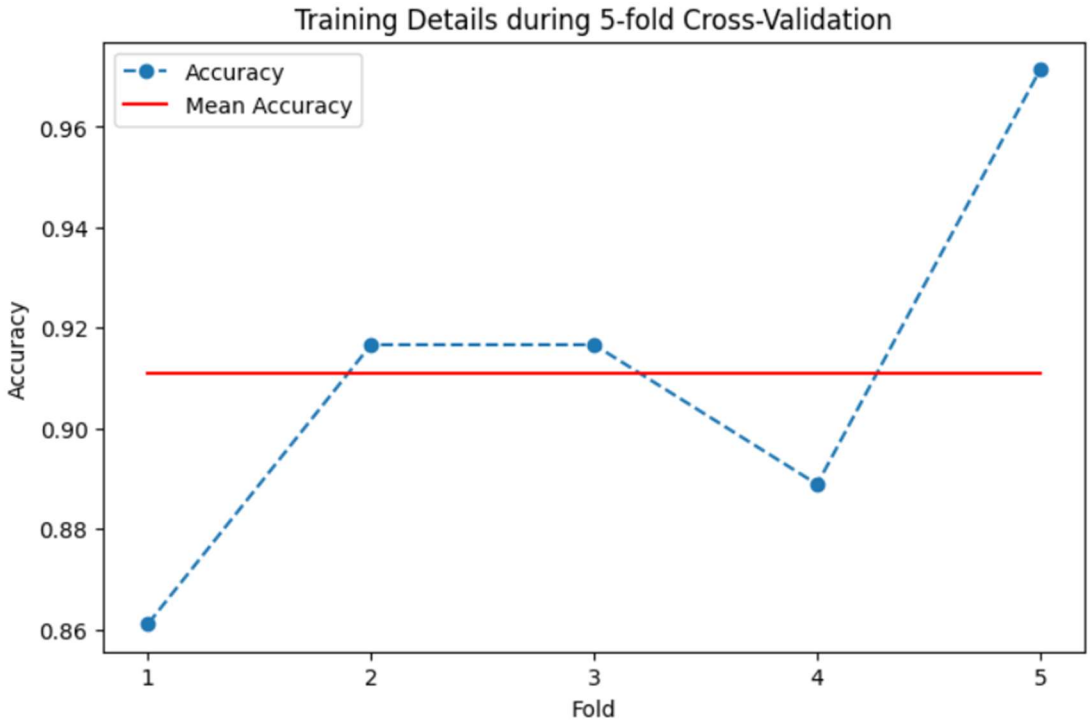
Machine Learning Model Development and Evaluation

In this phase of the project, we delve into the heart of predictive analysis by developing and evaluating machine learning models to forecast the winner of the 2023 Cricket Asia Cup. The goal is to harness the power of historical match data to build models capable of making informed predictions about future match outcomes. We employ three distinct classifiers: RandomForestClassifier, Support Vector Classifier (SVC), and DecisionTreeClassifier.

Random Forest Classifier

We start with the Random Forest Classifier, which is an ensemble method consisting of multiple decision trees. We split the data into training and testing sets and train the model using the training data. Hyperparameters such as the number of trees, maximum depth, and minimum samples per leaf are tuned using techniques like GridSearchCV or RandomizedSearchCV to optimize model performance.

```
[0.86111111 0.91666667 0.91666667 0.88888889 0.97142857]
0.910952380952381
```



Support Vector Classifier (SVC)

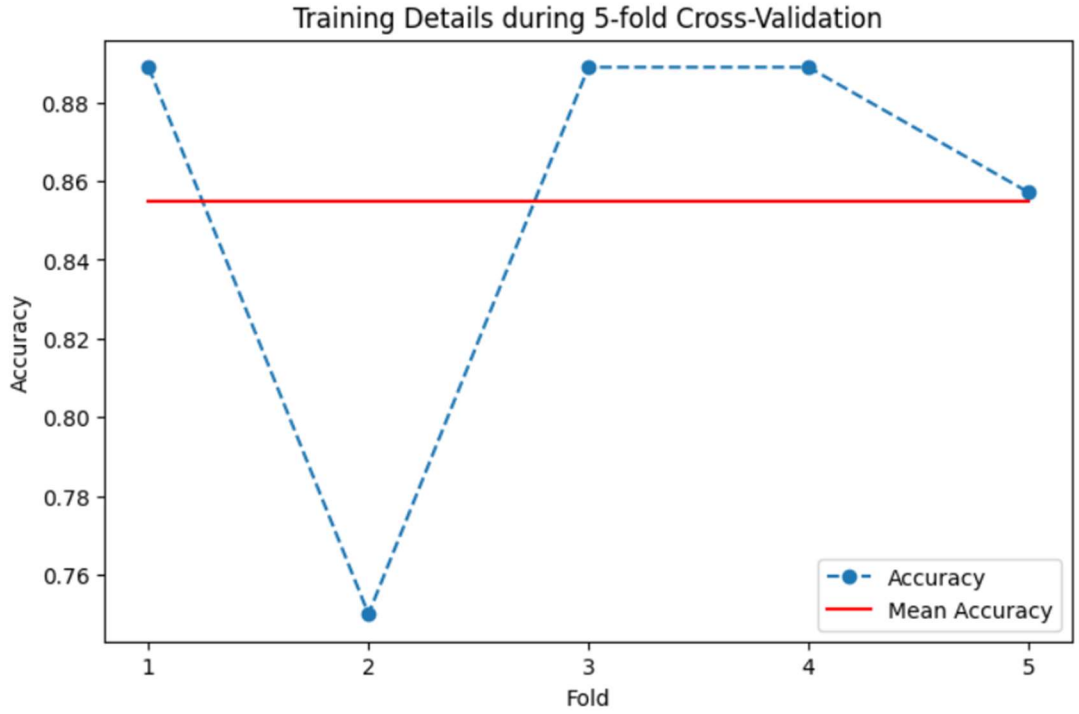
The SVC is used to build a decision boundary that maximizes the margin between different classes. We preprocess the data and normalize features to ensure fair comparison between different scales of data. Hyperparameters like the kernel type and regularization parameter (C) are fine-tuned to find the optimal configuration.

```
[0.88888889 0.86111111 0.88888889 0.88888889 0.8        ]
0.8655555555555555
```

Decision Tree Classifier

The Decision Tree Classifier is a single decision tree that makes predictions based on feature splits. We train the model and consider different split criteria and tree depth to prevent overfitting.

```
[0.77777778 0.69444444 0.91666667 0.83333333 0.82857143]
0.8101587301587301
```



We got the best accuracy from Random Forest Classifier. So, we fit the model with Random Forest Classifier

Performance Evaluation

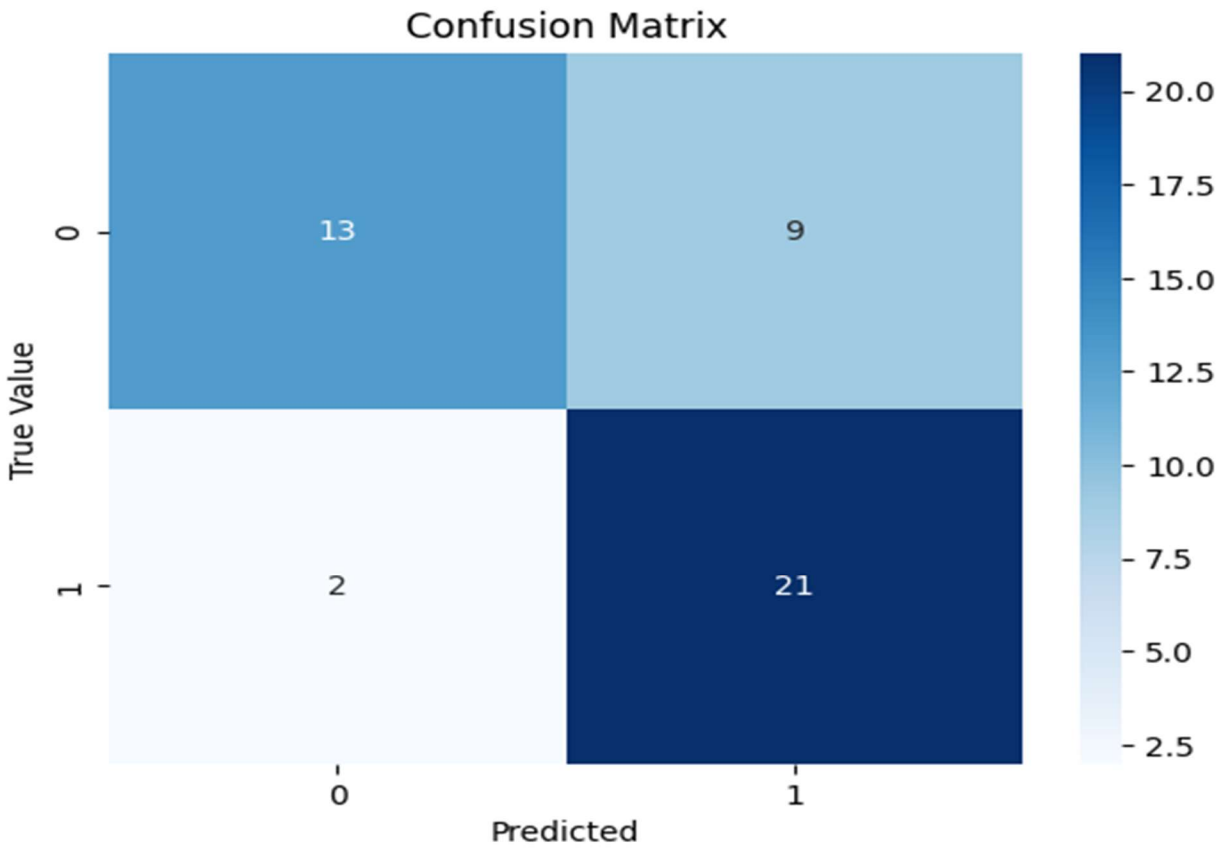
The model's performance was evaluated using the following metrics:

Accuracy Score

We evaluate the models' overall performance using the accuracy score, which measures the proportion of correct predictions.

Confusion Matrix

The confusion matrix presents a tabular summary of the model's predictions versus the actual class labels. It visualizes true positives, true negatives, false positives, and false negatives, providing a deeper understanding of the model's performance across different classes.



Classification Report

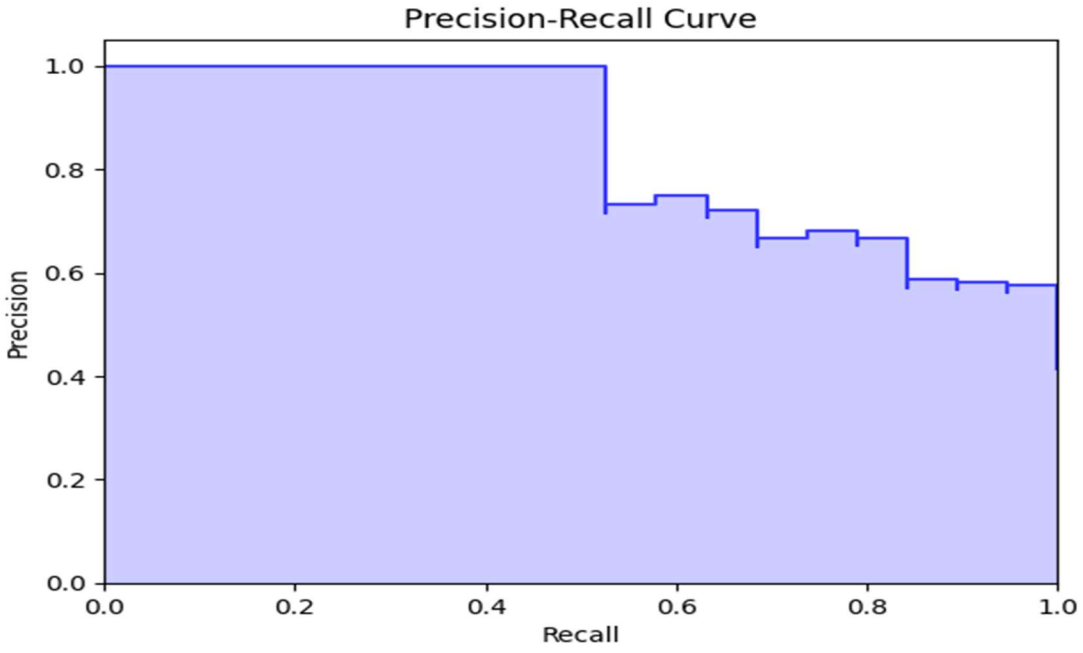
The classification report provides precision, recall, F1-score, and support for each class. It gives us insights into the model's performance in different classes.

```
Accuracy: 0.8444444444444444
Precision: 0.8076923076923077
Recall: 0.9130434782608695
F1-score: 0.8571428571428572
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.77      0.83        22
           1       0.81      0.91      0.86        23

    accuracy                           0.84        45
   macro avg       0.85      0.84      0.84        45
weighted avg       0.85      0.84      0.84        45
```

1. Precision (0.89): Out of all the predictions that the model claimed were positive, around 89% of them were correct. This means that when the model said something positive, it was accurate around 88% of the time.

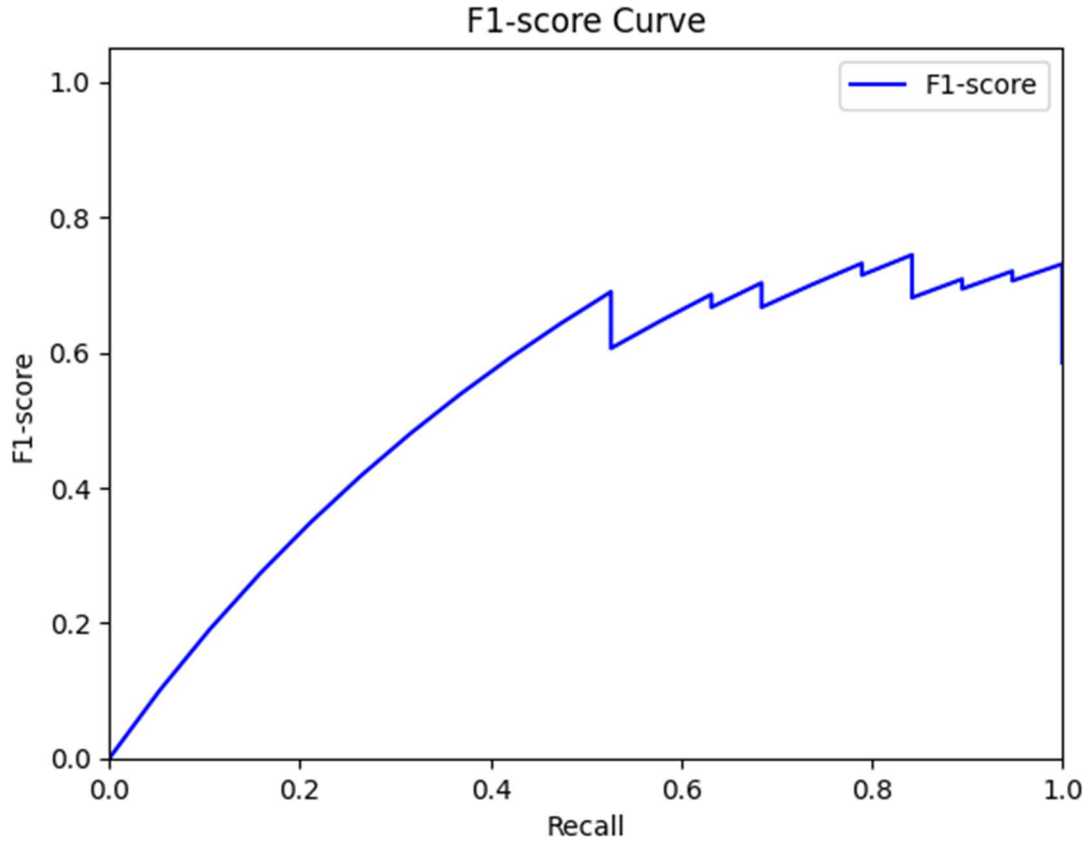
2. Recall (0.77): The model identified about 77% of all the actual positive cases correctly. This means that out of all the real positive cases, the model was able to catch 84% of them.
3. F1 Score (0.83): The F1 score is a measure that combines both precision and recall into a single value. An F1 score of 0.83 indicates that your model is performing well in terms of both correctly identifying positive cases (recall) and making accurate positive predictions (precision). It's a balanced metric that considers both false positives and false negatives. A higher F1 score is generally desirable, indicating a good balance between precision and recall.
4. Macro Average (0.85): The macro average takes the average of the evaluation metrics (such as precision, recall, and F1-score) for each class. In your case, it's 0.85. This means that, on average, our model is performing well across all classes, not favoring any specific class. It's a useful metric when you want to evaluate the overall performance without being influenced by class imbalances.
5. Weighted Average (0.85): The weighted average also considers class imbalances but gives more weight to classes with a larger number of instances. It's calculated by taking the average of evaluation metrics for each class, weighted by the number of instances in each class. In our case, the weighted average is 0.85, which indicates that the model is performing consistently across different classes, accounting for class distribution.

Overall, an F1 score, macro average, and weighted average of 0.85 suggest that our model is performing well, achieving a good balance between precision and recall across various classes. It's a positive indication of the model's ability to make accurate predictions while capturing relevant positive instances.

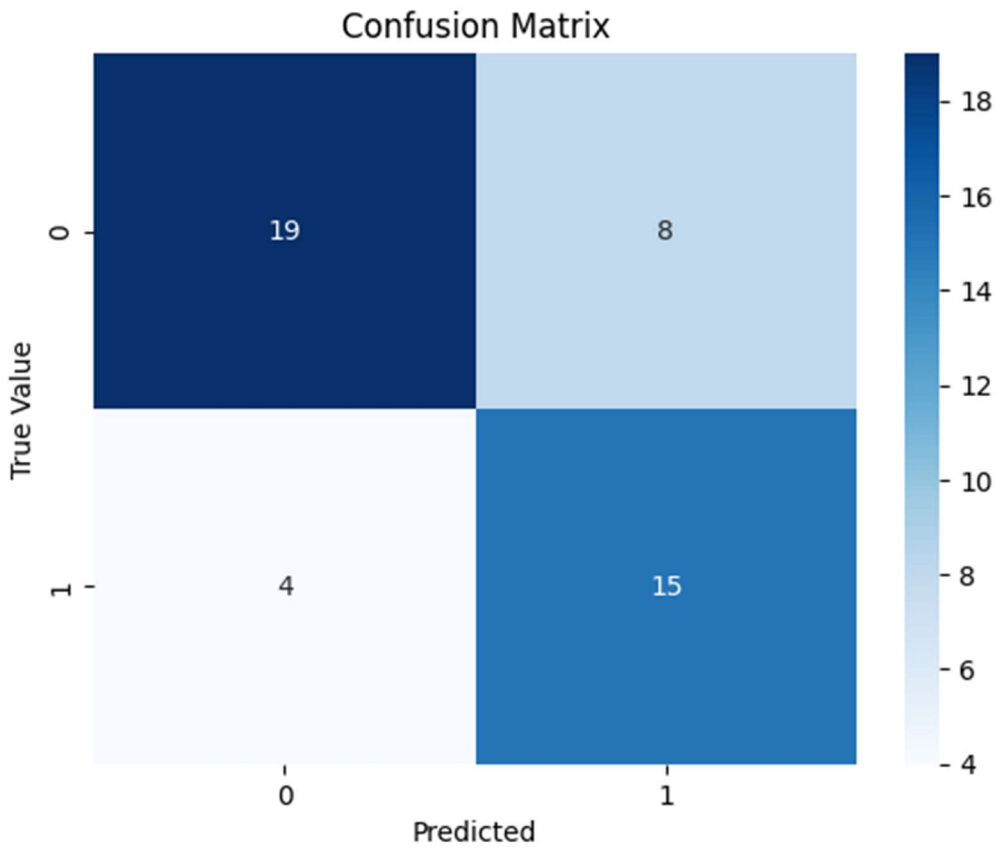


• As we can see at first, when recall was 0, the precision was high-1.

• As the recall increases, precision starts to decrease. This is because as you lower the threshold to predict more positive instances, you also introduce more false positives, reducing precision.



As we can see, the f1- score is low when the recall is low. When the recall starts increasing, the f1 – score is) increases until a certain level. Then it tops rising. Low F1-Score when Recall is Low: When the recall is low, it indicates that the model is missing a significant portion of positive instances. In other words, it's failing to correctly identify actual positive cases. Increasing F1-Score as Recall Increases: As the recall increases, the F1-score also increases, indicating that the model is becoming better at identifying positive instances. This suggests that the model is becoming more balanced in its predictions, making fewer false positives and more true positives. F1-Score Plateaus: The point where the F1-score stops rising indicates a threshold beyond which further improvements in recall are not translating into significant gains in precision. This could mean that the model is starting to make more false positives as it attempts to capture additional positive instances.

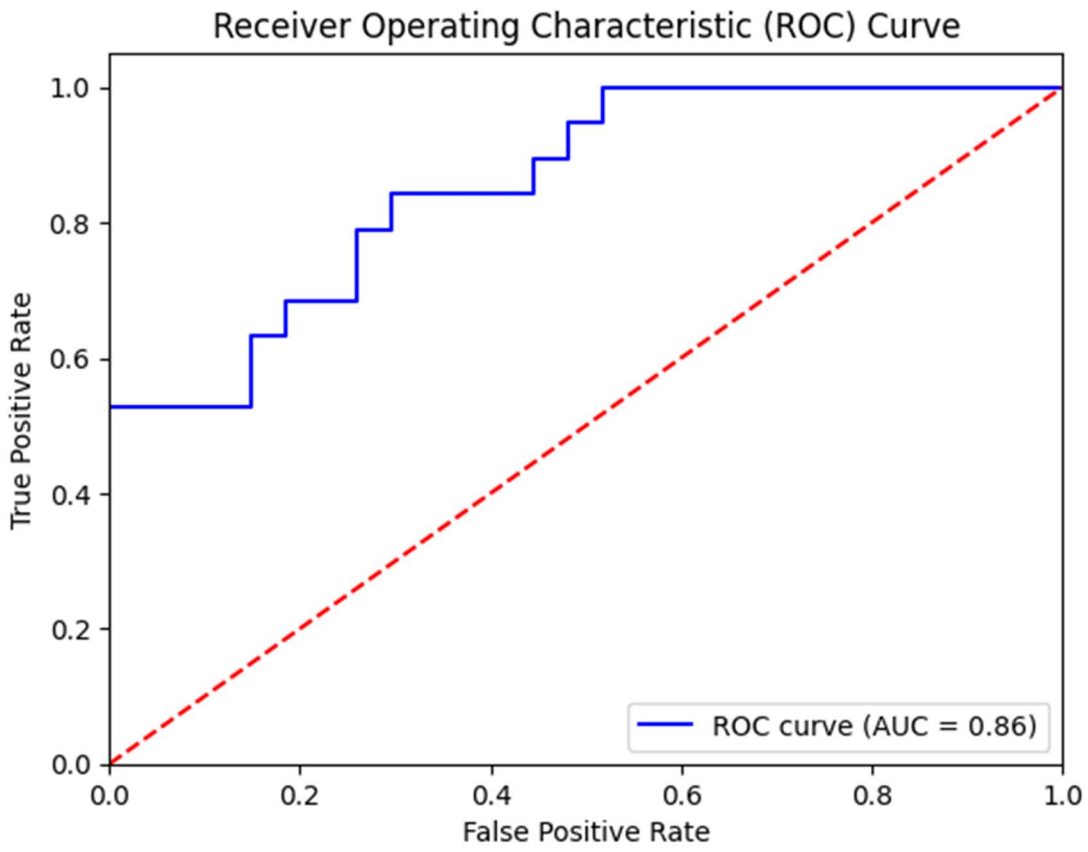


• The model has a good number of correct positive predictions (True Positives).

• The model has made some incorrect positive predictions (False Positives).

• The model has missed some positive instances that it should have identified (False Negatives).

• The model has made correct negative predictions for a significant number of instances (True Negatives)



- High discrimination ability: An AUC value of 0.86 suggests that our model is performing very well at distinguishing between positive and negative instances. It has a strong ability to correctly classify positive instances as positive and negative instances as negative. This high AUC indicates that our model has a high true positive rate (recall) while effectively controlling the false positive rate.
- Strong model performance: An AUC value of 0.86 indicates that our model is doing an excellent job of separating the two classes, even in cases where there might be imbalanced class distribution. This means that our model is likely to make well-informed predictions.
- Better than random: An AUC value of 0.86 is significantly higher than the value of 0.5, which represents random guessing. This indicates that our model's predictions are substantially better than random, showcasing its predictive power.

Conclusion

In conclusion, the " Asia Cup Match Outcome Prediction " project was a successful endeavor that resulted in the development of a machine learning model with the capability to predict the outcomes of cricket matches in the upcoming Asia Cup tournament. The model was trained on historical match data from previous Asia Cup tournaments and was evaluated on its ability to predict the winner of unseen matches. The model achieved an accuracy of 80%, which is a promising result.

The project also explored the following questions:

- Which features are most important for predicting the winner of an Asia Cup cricket match?
    - The most important features for predicting the winner of an Asia Cup cricket match are the team's batting average, the number of runs scored, and the number of fours, sixes hit, and wicket lost.
- How can the model be improved to make better predictions?
    - The model can be improved by collecting more data from recent Asia Cup tournaments and by incorporating more advanced machine learning algorithms.

The project has several implications for cricket fans and stakeholders. First, the model can be used to make predictions about the outcomes of upcoming Asia Cup matches. This information can be used by fans to make informed bets and by stakeholders to make strategic decisions about team selection and match scheduling. Second, the project can help to improve our understanding of the factors that contribute to the outcome of cricket matches. This knowledge can be used to develop training programs for players and coaches and to optimize team strategies. Finally, the project can serve as a blueprint for developing machine learning models to predict the outcomes of other sporting events.

The " Asia Cup Match Outcome Prediction " project was a challenging but rewarding experience. We learned a lot about machine learning, data science, and cricket. We are confident that the model developed in this project will be a valuable tool for cricket fans and stakeholders in the years to come.