DMPA PROJECT REPORT

Asia cup 2022 Win Prediction

Rupesh Koushik – 200953018
Josh Jaison – 200953033
Suhas Gowda-200953014

Introduction

A men's one-day international and twenty-twenty international cricket competition is called the Asian Cricket Council Asia Cup.
In order to forecast the winners of each game, we try to identify patterns from data.
In order to predict the tournament winners, we can also utilize sophisticated algorithms and methodologies to analyses the individual and team information.

We make an effort to create and test a data mining decision system that examines the dataset.
The dataset has been divided into three categories:

Historical (All Asia Cup matches between 1984 to 2022) (All Asia Cup -A men's one-day international and twenty-twenty international cricket competition is called the Asian Cricket Council Asia cup.)
In order to forecast the winners of each game, we try to identify patterns from data.
In order to predict the tournament winners, we can also utilise sophisticated algorithms and methodologies to analyse the individual and team information.

We make an effort to create and test a data mining decision system that examines the dataset.
The dataset has been divided into three categories:

Historical (All Asia Cup matches between 1984 and 2018) (All Asia Cup matches between 1984 and 2018)

Analytical (Players Data) (Players Data)

predictive (predicts the outcome of each game using historical data, player statistics, and the one-to-one win/loss ratio of the competing teams)

p matches between 1984 and 2018)

Analytical (Players Data) (Players Data)

predictive (predicts the outcome of each game using historical data, player statistics, and the one-to-one win/loss ratio of the competing teams).

Literature Review

1.

The goal of this study was to generate a model that could be used to forecast an ODI cricket match's result while it was still being played. The goal was to try to forecast the result of a cricket match in the Indian Premier League (IPL). The Prediction Modeling with Multiple Linear Regression algorithm is employed. In data mining, regression is a crucial statistical method that is frequently utilized. The logistic regression approach, which is frequently used for classification issues, was also applied. The model's effectiveness was measured by how well it predicted the toss, the score, and the run rate. It then plotted those predictions against the final score predictions and displayed the results on a graph. The eventual match winner and the scores for each innings might be predicted using multiple linear regression at regular intervals. This information might be useful in creating a more precise forecasting tool in the future.

2. The purpose of this work was to solve the issue of properly predicting match outcomes by modelling game evolution. By mining the existing game data, the authors created a model to learn the one-day format games. A one-day cricket dataset was used to train a number of algorithms for both supervised and unsupervised learning. Home-run and No-home run prediction model was the algorithm employed. According to this data, the error margin was less than or equal to 10 runs in both innings in more than 55% of the games. This study reported the effectiveness of the closest neighbor approach along with attribute bagging. Advantages include a 68–70% accuracy rate in predicting the outcome of both innings. Limitations of the paper was lack of prediction of fall of wickets along with the bowler’s and batsmen’s features.

3.

The concept was to develop a model had two methods, firstly to predict the score of first innings not only on the basis of current run rate but also considered the number of wickets fallen, venue of the match and batting team. The goal of this research was to use game evolution modelling to address the problem of precisely forecasting match outcomes. The authors developed a model to learn the one-day format games by mining the already-existing game data. Several supervised and unsupervised learning algorithms were trained on a one-day cricket dataset. The method used was a home run and no-home run prediction model. This information shows that in more than 55% of the games, the error margin was less than or equal to 10 runs in both innings. In addition to attribute bagging, this study reported on the effectiveness of the nearest neighbor strategy. Benefits include an accuracy percentage of 68–70% in forecasting the results of both innings.

4.

The purpose of the study was to make predictions on the outcome of a One Day International (ODI) cricket match following the conclusion of the first inning. The goal was to develop an initial model and then use feature selection methods such as univariate, recursive elimination, and principal component analysis to increase the accuracy of predicting the winner (PCA). This prediction was made using the data obtained using 15 features, which were made up of characteristics relating to batting, bowling, team makeup, and other factors. The technique utilized was the Naive Bayes (NB) approach. Furthermore, by dividing the inning into five over periods, NB and linear regression employ the same features to forecast the result of the game. Their research shows that NB correctly predicts the game's outcome 68% of the time using the first five overs, and 91% of the time by the conclusion of the 45th over. The benefit is that we discover that using the univariate feature selection approach with 85%:15% of training to testing sample sizes results in the greatest accuracy in predicting the winning team. The results confirm those found in the literature on the effectiveness of NB for relatively small sample sizes; the main negative was that they were unable to employ very large sample sizes.

5.

The numerous elements that determine the result of a cricket match were examined in this paper, and it was found that the home team, the opponent team, the venue, and the toss winner all affect a team's likelihood of winning. The suggested prediction model uses multivariate regression to determine the points earned by each player in the league and to determine each team's overall strength based on the players' prior performances. Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbor computations are used. Random Forest proved to be the most reliable classifier for both datasets, predicting runs scored by a batsman with an accuracy of 90.74% and wickets taken by a bowler with a 92.25%.

6.

To put out a model in which, at the start of the second inning, the result of the game is predicted ball by ball. SVM and Random Forest Classifier perform exceptionally well when compared to classical machine learning techniques. The accuracy of the train is 76.47%, with a confidence interval of 3.77%, while the accuracy of the test is 72.74%. Both models have extremely high accuracy, 83.25% and 81.90%, respectively. The accuracy of the XGB Classifier is likewise quite excellent at 83.92%. One-to-one sequence accuracy for vanilla LSTM in deep learning models is 75.05%, with a confidence interval of 3.12%. The mean train and test scores are significantly smaller in Gaussian Nave Bayes and Bernoulli Nave Bayes models. This could be as a result of the presumption that qualities are continuous and independent.

7.

The main goal is to use five alternative ways to create a data mining system to forecast the winning cricket team given two teams for ODI matches. Five methods were used to conduct the research. They are determining how features affect how well a player performs individually, rating players based on how well they perform individually, determining how well a player or group of players work together, determining common player combinations, and forecasting the result of matches. The findings indicate that while Logistic Regression and MLP had the best test accuracy, Gaussian Process had the highest train accuracy.

8.

This research study will focus on predicting the result of an IPL match before it ever starts. To forecast the IPL winner, machine learning models are trained on the selected aspects. For this model-building purpose, many machine learning techniques, such as Random Forest, SVM, Naive Bayes, Logistic Regression, and Decision Tree, have been used to test and training datasets of diverse sizes. To perform the study and predict the winner of the IPL, several data science fields have come together, including feature selection, data pre-processing, data visualization, data preparation, and machine learning model implementation.

9.

To predict the second innings of ODIs matches, Singh et al, created a model to predict the cricket match's ultimate score and the first innings score. employed Linear Regression to forecast the first innings score using a 5 over interval technique and significantly outperformed the standard run rate metric in terms of accuracy. Jhawar et al. created a model to forecast cricket match results depending on team makeup. High levels of accuracy were attained by their model, with the average accuracy reaching 71%.

10.

It is possible to rank bowlers and batsmen according to how they performed in one-day international matches using a variety of methodologies. To overcome the difficulty encountered in learning the IPL match specifics and features, they have utilized a variety of features, including Number of Wickets Lost, Four Hitting Frequency, Six Hitting Frequency, Boundary Run Percentage, and Dot Ball Percentage, as well as playerperformance, toss decision, venue, and relative team strength.

Methodology

The dataset has been compiled to address following three aspects:

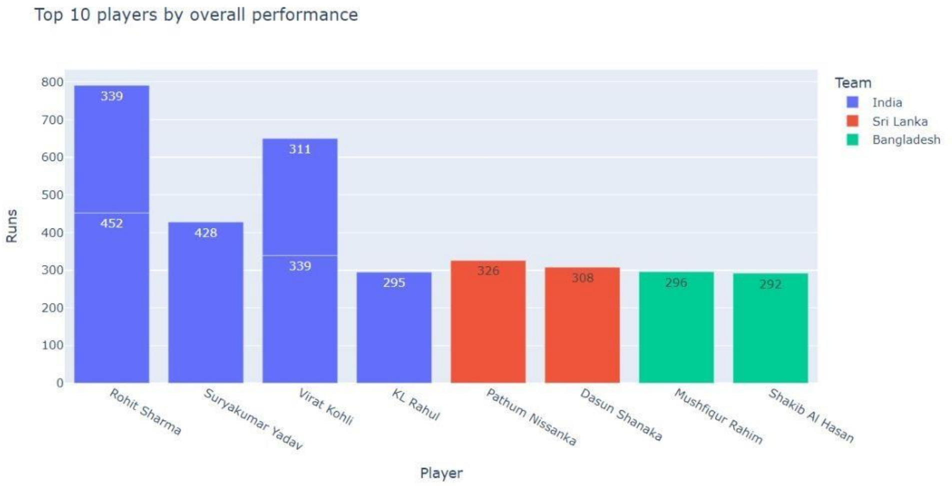
1. Historical (All Asia Cup matches between 1984 and 2022)
2. Analytical (Players Data)
3. Predictive (Utilize historical data, players data and one vs. one win/loss ratio of participating teams to predict winner of each match)

The definitions and workings of the different algorithms used to simulate the batters, bowlers, and teams are covered in length in this part as well as how we went about solving the issue. We use Python along with data manipulation frameworks like NumPy and Pandas. We also plan to use software's such as Rapid miner for data processing. The data will be procured from Kaggle.

Data processing: Generate a subset from raw data on which knowledge discovery will be performed. This also removes outliers and redundant data, and converts data to a form relevant to the next stage.

Data mining: Converts the processed data into decision trees, which contain useful patterns.

Evaluation: Evaluate the consistency of patterns via a testing set. This model can then possibly be used on a real-world situation.

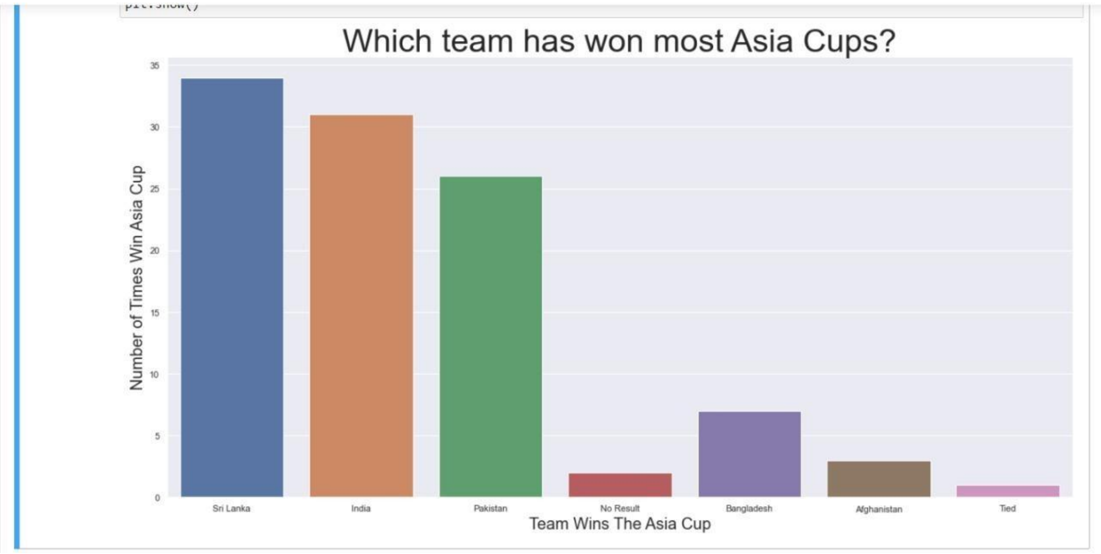


Modeling Batsmen:

The outcome of a game is significantly influenced by a player's hitting skills. A squad normally comprises of a group of 6-7 specialized hitters out of the 11 players. In order to analyses a player's traits and predict a batsman's ability, we may utilize two distinct sorts of data. To determine whether he has the potential to be a competitor, we first look at his past results.

Next, we assess his present form by looking at the outcomes of his most recent games. A batsman's confidence is reflected by how well he has performed recently for the team, which is determined by his form.

```
Algorithm 1 Modeling Batsmen

Input: Players p ∈{P (A, m) UP (B, m)}, Career Statistics of player p: ϕ(p) Output:
Batsmen Score of

For all players p∈{P(A, m)UP(B, m)}do ϕ ←ϕ(p)

u ← q ϕ Bat Innings ϕ Matches Played

v←20∗ϕNum Centuries +5∗ϕNum Fifties w←0.3∗v+0.7∗ϕBat Avg
ϕ Career Score ← u ∗ w

M ← Last 4 matches played by p ϕ Recent Score ← mean (Mp )

end for

for all players p∈{P(A ,m) UP(B, m)} do

ϕ Career Score ← ϕ Recent Score ←

ϕ Career Score max(ϕ Career Score)

ϕ Recent Score max(ϕ Recent Score)

Runs

Φ Batsmen Score = 0.35 ∗ ϕ Career Score + 0.65 ∗ ϕ Recent Score end for
```



```
plt.show()
```

Modelling teams: The core components of a team are the bowlers and the batsmen.

We therefore plan to establish a team's overall score in relation to the opposing team using the modelled batsmen and bowlers.

The total batting average of a team is determined by adding the batting averages of each of its members.

Similar to this, a team's bowling score is determined by adding the individual bowling scores of each member. Since the measured contributions of individual players to the team score are already handled by the variable u in Algorithms 1 and 2, we immediately used the player scores in the team score. Evaluation of the relative strength between two teams, A and B, playing the match is illustrated in Algorithm 3. Initially, the score of the bowler and batsman need to be normalized in the range [0,1] since their values are varied.

The potency of team A in comparison to team B is depicted by the variable S(A/B). The algorithm follows the game's core precept that one team's batsmen should face off against another team's bowlers, and vice versa.

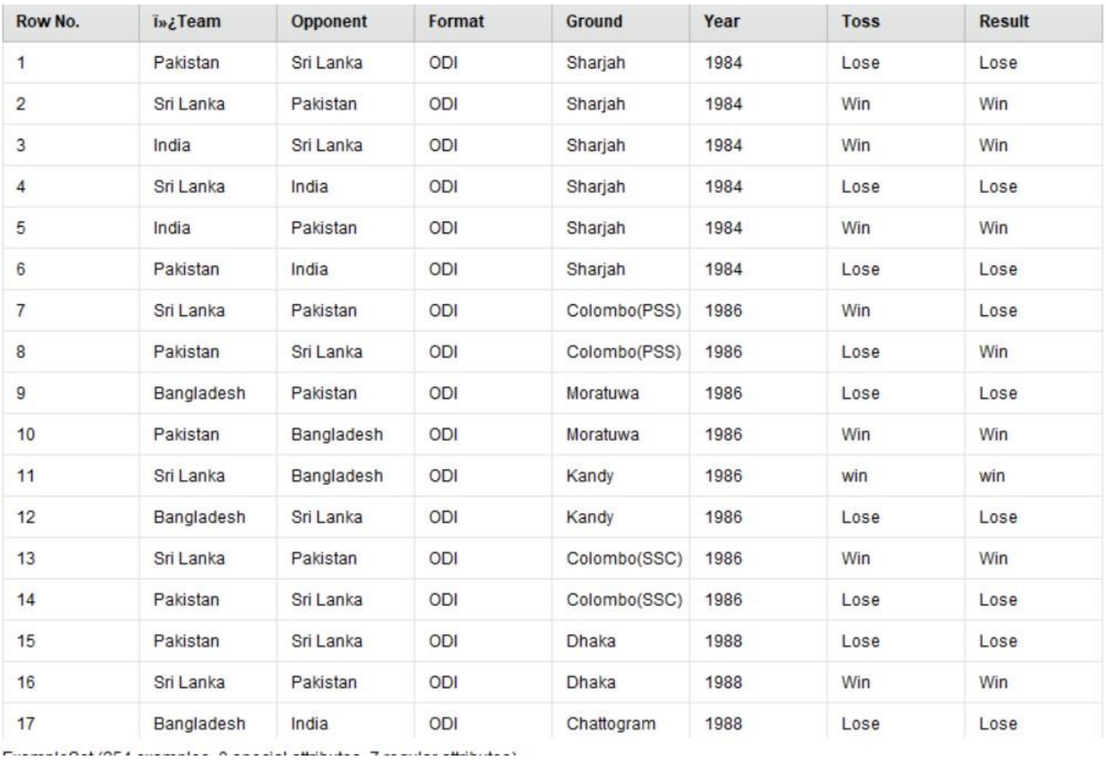
Algorithms

Decision tree:

The decision tree algorithm is the most efficient and well-liked

classification and prediction approach. Each internal node in a decision tree represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label. Decision trees are a sort of tree structure that mimics flowcharts. Our accuracy rate using the decision tree method was 61.33%.

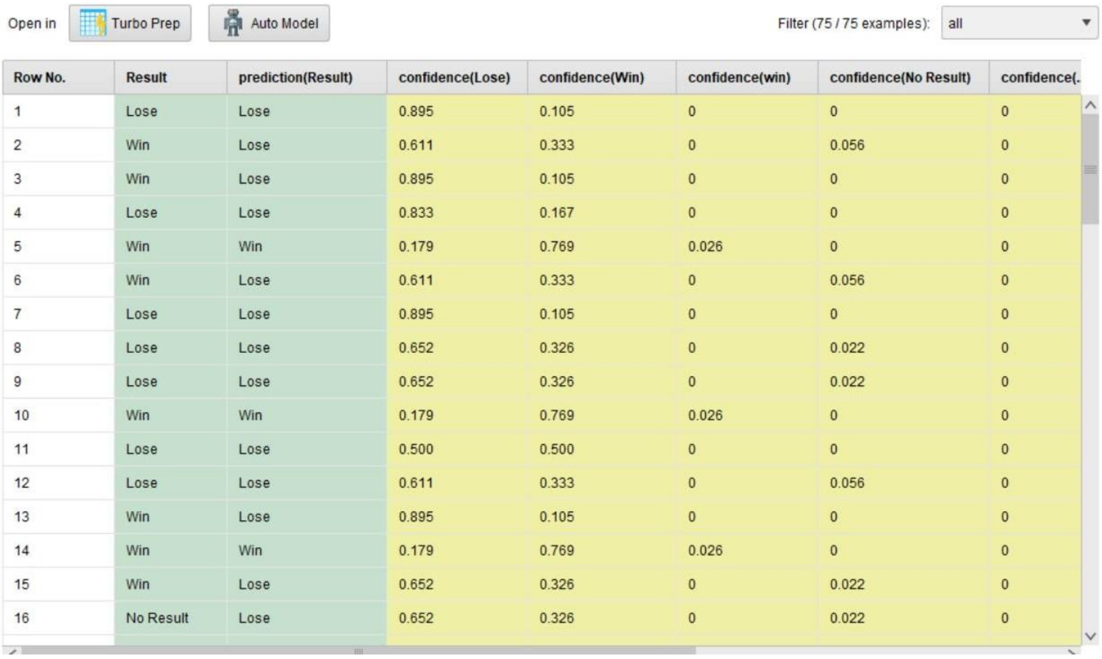
Asia cup – all matches dataset



Row No.	ï»¿Team	Opponent	Format	Ground	Year	Toss	Result
1	Pakistan	Sri Lanka	ODI	Sharjah	1984	Lose	Lose
2	Sri Lanka	Pakistan	ODI	Sharjah	1984	Win	Win
3	India	Sri Lanka	ODI	Sharjah	1984	Win	Win
4	Sri Lanka	India	ODI	Sharjah	1984	Lose	Lose
5	India	Pakistan	ODI	Sharjah	1984	Win	Win
6	Pakistan	India	ODI	Sharjah	1984	Lose	Lose
7	Sri Lanka	Pakistan	ODI	Colombo(PSS)	1986	Win	Lose
8	Pakistan	Sri Lanka	ODI	Colombo(PSS)	1986	Lose	Win
9	Bangladesh	Pakistan	ODI	Moratuwa	1986	Lose	Lose
10	Pakistan	Bangladesh	ODI	Moratuwa	1986	Win	Win
11	Sri Lanka	Bangladesh	ODI	Kandy	1986	win	win
12	Bangladesh	Sri Lanka	ODI	Kandy	1986	Lose	Lose
13	Sri Lanka	Pakistan	ODI	Colombo(SSC)	1986	Win	Win
14	Pakistan	Sri Lanka	ODI	Colombo(SSC)	1986	Lose	Lose
15	Pakistan	Sri Lanka	ODI	Dhaka	1988	Lose	Lose
16	Sri Lanka	Pakistan	ODI	Dhaka	1988	Win	Win
17	Bangladesh	India	ODI	Chattogram	1988	Lose	Lose

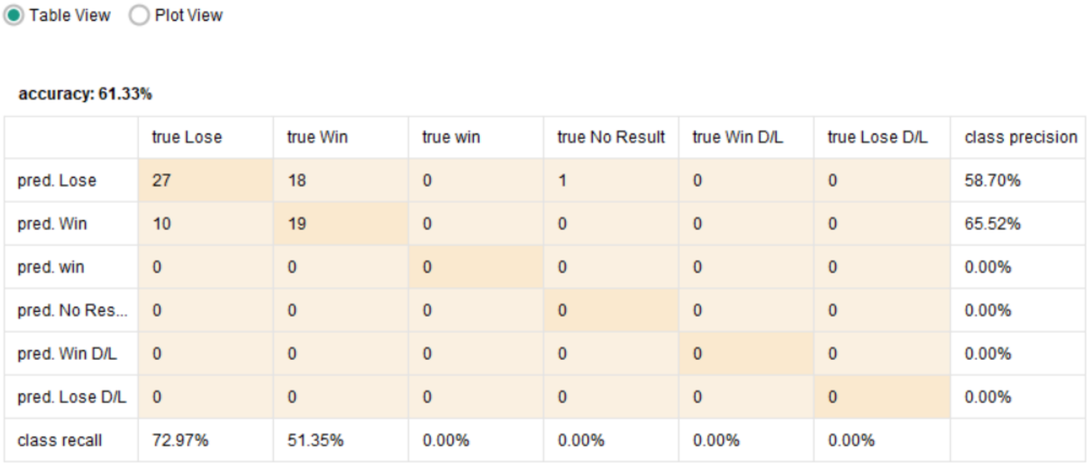
ExampleSet (254 examples, 0 special attributes, 7 regular attributes)

predicted result:



Row No.	Result	prediction(Result)	confidence(Lose)	confidence(Win)	confidence(win)	confidence(No Result)	confidence(.
1	Lose	Lose	0.895	0.105	0	0	0
2	Win	Lose	0.611	0.333	0	0.056	0
3	Win	Lose	0.895	0.105	0	0	0
4	Lose	Lose	0.833	0.167	0	0	0
5	Win	Win	0.179	0.769	0.026	0	0
6	Win	Lose	0.611	0.333	0	0.056	0
7	Lose	Lose	0.895	0.105	0	0	0
8	Lose	Lose	0.652	0.326	0	0.022	0
9	Lose	Lose	0.652	0.326	0	0.022	0
10	Win	Win	0.179	0.769	0.026	0	0
11	Lose	Lose	0.500	0.500	0	0	0
12	Lose	Lose	0.611	0.333	0	0.056	0
13	Win	Lose	0.895	0.105	0	0	0
14	Win	Win	0.179	0.769	0.026	0	0
15	Win	Lose	0.652	0.326	0	0.022	0
16	No Result	Lose	0.652	0.326	0	0.022	0

Performance of decision tree algorithm:



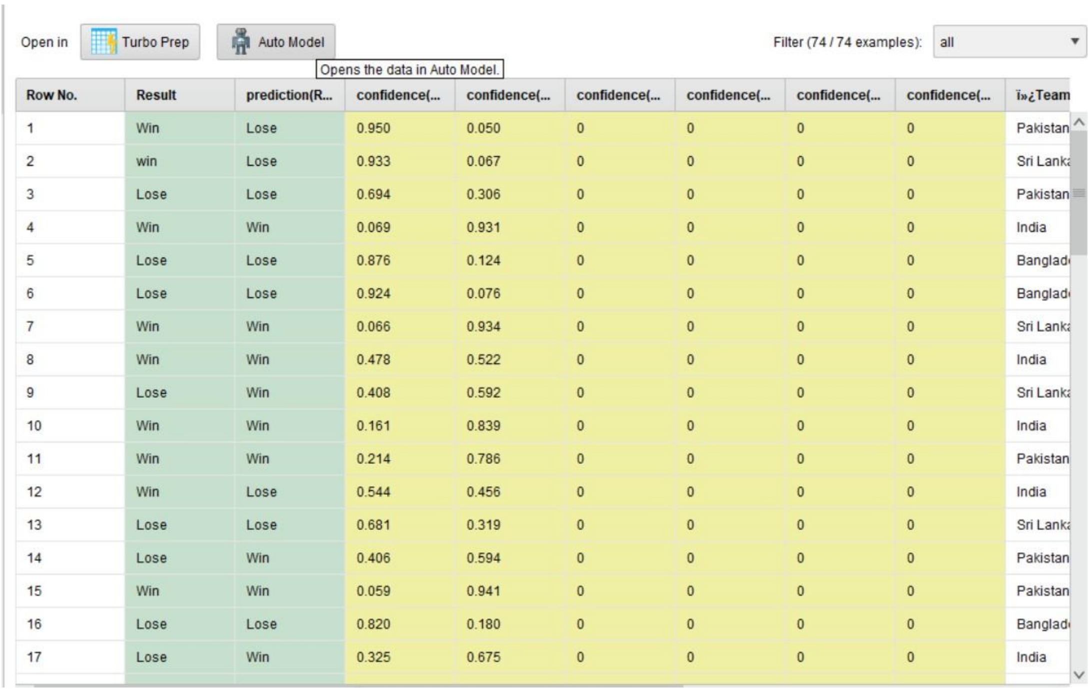
accuracy: 61.33%

	true Lose	true Win	true win	true No Result	true Win D/L	true Lose D/L	class precision
pred. Lose	27	18	0	1	0	0	58.70%
pred. Win	10	19	0	0	0	0	65.52%
pred. win	0	0	0	0	0	0	0.00%
pred. No Res...	0	0	0	0	0	0	0.00%
pred. Win D/L	0	0	0	0	0	0	0.00%
pred. Lose D/L	0	0	0	0	0	0	0.00%
class recall	72.97%	51.35%	0.00%	0.00%	0.00%	0.00%	

Naïve bayes algorithm:

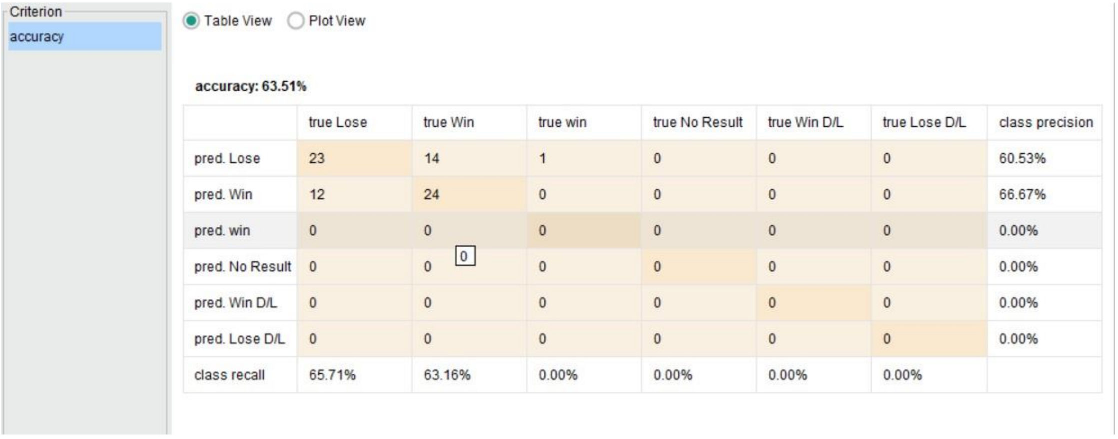
On the basis of Bayes theorem, the Naive Bayes algorithm is a supervised learning technique for classification problems. It primarily uses a huge training set for text classification. One of the most simple and effective classification algorithms now in use is the Naive Bayes Classifier. It facilitates the creation of efficient machine learning models that are capable of generating precise predictions. It provides predictions based on the likelihood that an object will occur since it is a probabilistic classifier. When the Naive Bayes technique is used on the dataset that is currently available, accuracy is 63.51%.

Predicted result:



Row No.	Result	prediction(R...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	ï»¿Team
1	Win	Lose	0.950	0.050	0	0	0	0	Pakistan
2	win	Lose	0.933	0.067	0	0	0	0	Sri Lank
3	Lose	Lose	0.694	0.306	0	0	0	0	Pakistan
4	Win	Win	0.069	0.931	0	0	0	0	India
5	Lose	Lose	0.876	0.124	0	0	0	0	Banglad
6	Lose	Lose	0.924	0.076	0	0	0	0	Banglad
7	Win	Win	0.066	0.934	0	0	0	0	Sri Lank
8	Win	Win	0.478	0.522	0	0	0	0	India
9	Lose	Win	0.408	0.592	0	0	0	0	Sri Lank
10	Win	Win	0.161	0.839	0	0	0	0	India
11	Win	Win	0.214	0.786	0	0	0	0	Pakistan
12	Win	Lose	0.544	0.456	0	0	0	0	India
13	Lose	Lose	0.681	0.319	0	0	0	0	Sri Lank
14	Lose	Win	0.406	0.594	0	0	0	0	Pakistan
15	Win	Win	0.059	0.941	0	0	0	0	Pakistan
16	Lose	Lose	0.820	0.180	0	0	0	0	Banglad
17	Lose	Win	0.325	0.675	0	0	0	0	India

performance of naïve bayes algorithm:



accuracy: 63.51%

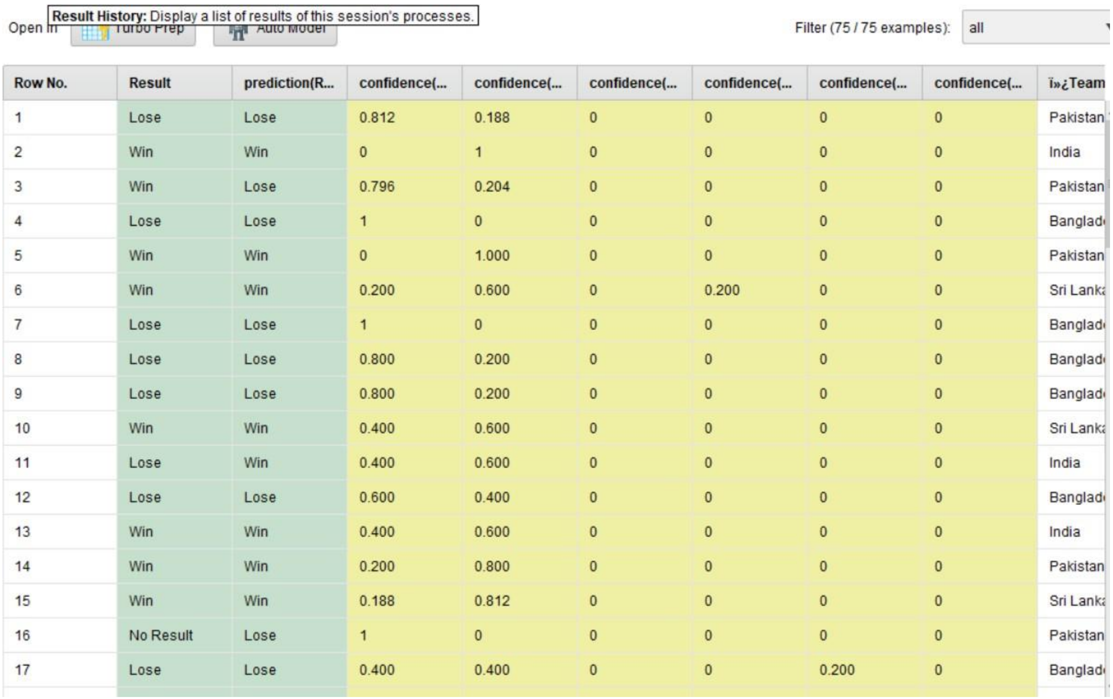
	true Lose	true Win	true win	true No Result	true Win D/L	true Lose D/L	class precision
pred. Lose	23	14	1	0	0	0	60.53%
pred. Win	12	24	0	0	0	0	66.67%
pred. win	0	0	0	0	0	0	0.00%
pred. No Result	0	0	0	0	0	0	0.00%
pred. Win D/L	0	0	0	0	0	0	0.00%
pred. Lose D/L	0	0	0	0	0	0	0.00%
class recall	65.71%	63.16%	0.00%	0.00%	0.00%	0.00%	

KNN algorithm:

K-Nearest Neighbor is one of the most basic supervised learning-based machine learning algorithms. The K-NN method places the new instance in the category that resembles the current categories the most, presuming that the new case and the previous instances are comparable. After storing all the previous data, a new data point is categorized using the K-NN algorithm based on similarity. This indicates that new data may be reliably and rapidly categorized using the K-NN approach. The K-NN technique may be used for regression even though classification issues are where it is most typically applied.

K-NN makes no assumptions about the underlying data because it is a non-parametric approach. As a result of saving the training dataset rather than instantly learning from it, the method is sometimes referred to as a lazy learner. Instead, it performs an action while categorizing data by using the dataset. The KNN approach simply stores the data during the training phase and categorizes fresh data into a category that is very similar to the training data. We were able to get the highest accuracy of 74.67% using the KNN approach.

Predicted result:

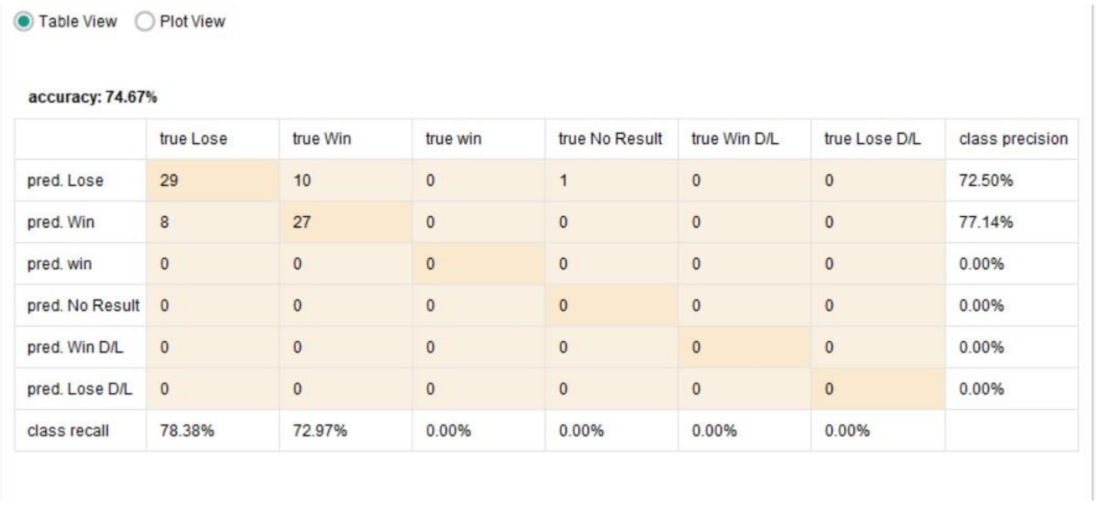


Result History: Display a list of results of this session's processes.

Filter (75 / 75 examples): all

Row No.	Result	prediction(R...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	ï»¿Team
1	Lose	Lose	0.812	0.188	0	0	0	0	Pakistan
2	Win	Win	0	1	0	0	0	0	India
3	Win	Lose	0.796	0.204	0	0	0	0	Pakistan
4	Lose	Lose	1	0	0	0	0	0	Banglad
5	Win	Win	0	1.000	0	0	0	0	Pakistan
6	Win	Win	0.200	0.600	0	0.200	0	0	Sri Lank
7	Lose	Lose	1	0	0	0	0	0	Banglad
8	Lose	Lose	0.800	0.200	0	0	0	0	Banglad
9	Lose	Lose	0.800	0.200	0	0	0	0	Banglad
10	Win	Win	0.400	0.600	0	0	0	0	Sri Lank
11	Lose	Win	0.400	0.600	0	0	0	0	India
12	Lose	Lose	0.600	0.400	0	0	0	0	Banglad
13	Win	Win	0.400	0.600	0	0	0	0	India
14	Win	Win	0.200	0.800	0	0	0	0	Pakistan
15	Win	Win	0.188	0.812	0	0	0	0	Sri Lank
16	No Result	Lose	1	0	0	0	0	0	Pakistan
17	Lose	Lose	0.400	0.400	0	0	0.200	0	Banglad

Performance of the algorithm:



● Table View ○ Plot View

accuracy: 74.67%

	true Lose	true Win	true win	true No Result	true Win D/L	true Lose D/L	class precision
pred. Lose	29	10	0	1	0	0	72.50%
pred. Win	8	27	0	0	0	0	77.14%
pred. win	0	0	0	0	0	0	0.00%
pred. No Result	0	0	0	0	0	0	0.00%
pred. Win D/L	0	0	0	0	0	0	0.00%
pred. Lose D/L	0	0	0	0	0	0	0.00%
class recall	78.38%	72.97%	0.00%	0.00%	0.00%	0.00%	

Results and Discussion

All Asia cup matches dataset contains the historical data from the year 1984 to 2022. 25% of Asia cup matches are played between 1997, 50% Asia cup matches are played between 2004 and 75% Asia cup matches are played between 2012.

The system predicts the success of a team by mining past team success data through a prediction methodology and data mining algorithms. we have taken the past data of ODI/T20 matches happened. We have analyzed the data based on the data points Teams, Toss, Ground, Year, Format. Using algorithms, we have predicted the winner of the match.

The full dataset has been scraped from the cric-info and Kaggle website in order to obtain all the necessary statistics.

The dataset contains the cricket events played between 2010 and 2022.This primarily includes match information related to the two combatant teams, result of coin toss, the venue and the final winner.Along with this, the outcomes of each game and the participation players' careerstatistics are also included.

Conclusion

Dataset: The Kaggle and cric info websites allowed users to download the whole dataset.

Every game played between 1984 and 2022 is represented in the dataset.

The dataset includes the fundamental match information for each game, inclusive of thetwo competing teams, the result of the coin toss, the game's time and location, and the game's winner.

The results of every game as well as the participating players' career statistics are also included.