A Project/Dissertation Review Report

on

SPORTS PREDICTION

Submitted in partial fulfillment of the
requirement for the award ofthe degree of

B.TECH CSE AIML

GALGOTIAS
UNIVERSITY
(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
MR. SHAMSH TABAREJ
ASSISTANT PROFESSOR

Submitted By
AMAN KUMAR MISHRA
20SCSE1180125

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
December, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"sports prediction "** in partial fulfillment of the requirements for the award of the btech cse aiml-submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of month, Year to Month and Year, under the supervision of MR. SHAMSH TABAREJ assistant professor, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

-Aman kumar mishra,20SCSE1180125

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

MR. SHAMSH TABAREJ

Assistant professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Aman kumar mishra 20SCSE1180125 has been held on _________________ and his/her work is recommended for the award of b tech cse aiml.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December ,2021

Place: Greater Noida

Abstract

Sports prediction is usually treated as a classification problem, with one class (win, lose, or draw) to be predicted. In sports prediction, large numbers of factors including the historical performance of the teams, results of matches, and data on players, have to be accounted for to help different stakeholders understand the odds of winning or losing.

Machine learning (ML) is one of the intelligent methodologies that have shown promising results in the domains of classification and prediction. One of the expanding areas necessitating good predictive accuracy is sport prediction, due to the large monetary amounts involved in betting. In addition, club managers and owners are striving for classification models so that they can understand and formulate strategies needed to win matches.

These models are based on numerous factors involved in the games, such as the results of historical matches, player performance indicators, and opposition information. This paper provides a critical analysis of the literature in ML, focusing on the application of Artificial Neural Network (ANN) to sport results prediction. In doing so, we identify the learning methodologies utilized, data sources, appropriate means of model evaluation, and specific challenges of predicting sport results. This then leads us to propose a novel sport prediction framework through which ML can be used as a learning strategy.

Our research will hopefully be informative and of use to those performing future research in this application area. The focus of our project is basically based on the production .we use many platform to perform this project i.e. colab, Jupiter.

The result /outcome that come from our project is very simple i.e. if we give a particular value then our model show his outcome based on his training. Hence we calculate the score of model to know how efficient our model work

Sports prediction is usually treated as a classification problem, with one class (win, lose, or draw) to be predicted. In sports prediction, large numbers of factors including the historical performance of the teams, results of matches, and data on players, have to be accounted for to help different stakeholders understand the odds of winning or losing.

List of Tables

List of Figures

Data Flow Diagram:-

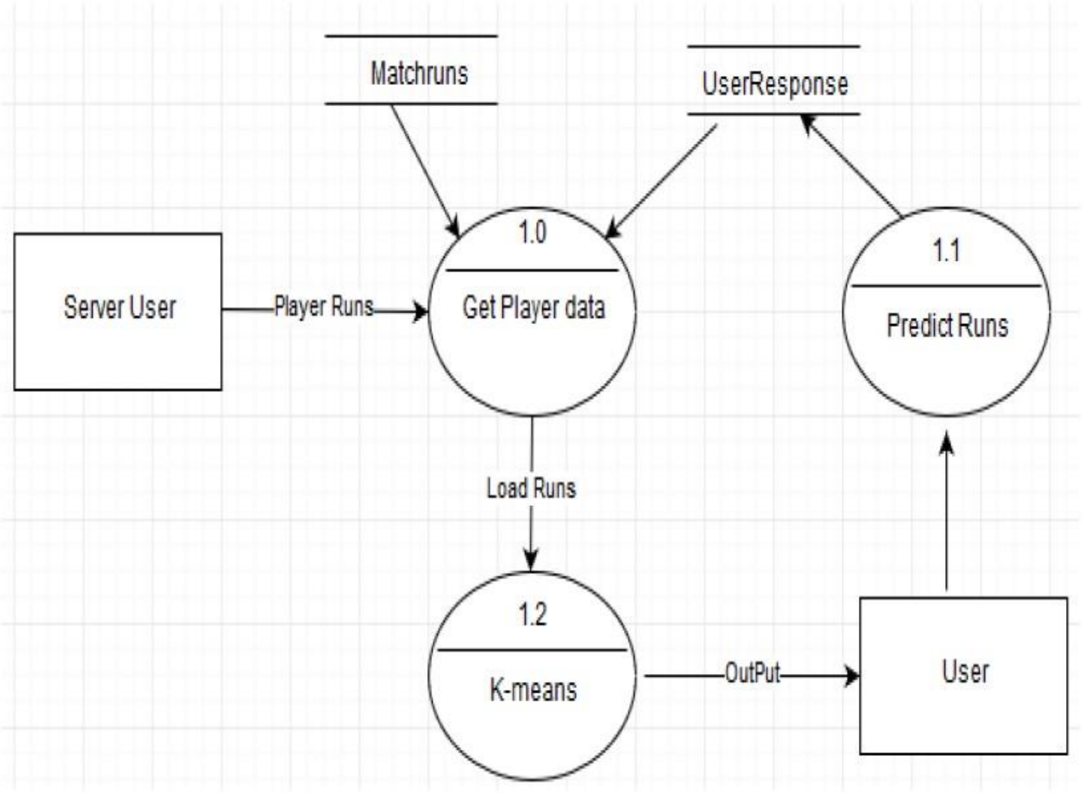


Table of Contents

CHAPTER 1
INTRODUCTION

Machine learning (ML) is one of the intelligent methodologies that have shown promising results in the domains of classification and prediction. One of the expanding areas necessitating good predictive accuracy is sport prediction, due to the large monetary amounts involved in betting.

In addition, club managers and owners are striving for classification models so that they can understand and formulate strategies needed to win matches.

There is huge commercial interest in player performance prediction. This has motivated many analysis of individual and team performance, as well as prediction of future games, across all formats of the game.

Currently, strategists rely on a combination of player experience, team constitution. The data set must be labelled to easily understand

One method of predicting results is the use of Deep Neural Networks (DNN) in conjunction with Artificial Neural Networks (ANN). These use multiple datasets which typically include training, validation, and testing.

Usage of this method proved particularly fruitful during the 2018 FIFA World Cup whereby over 63% of matches were proved accurate. Expectedly, this percentage would display greater accuracy through the use of increased datasets and team information.

Algorithm used:-

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

For example,

- To predict whether an email is spam (1) or (0)
- Whether the tumor is malignant (1) or not (0)

Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time.

From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

Simple Logistic Regression

Model

Output = 0 or 1

Hypothesis => $Z = WX + B$

$h\Theta(x) = \text{sigmoid}\ (Z)$

Sigmoid Function

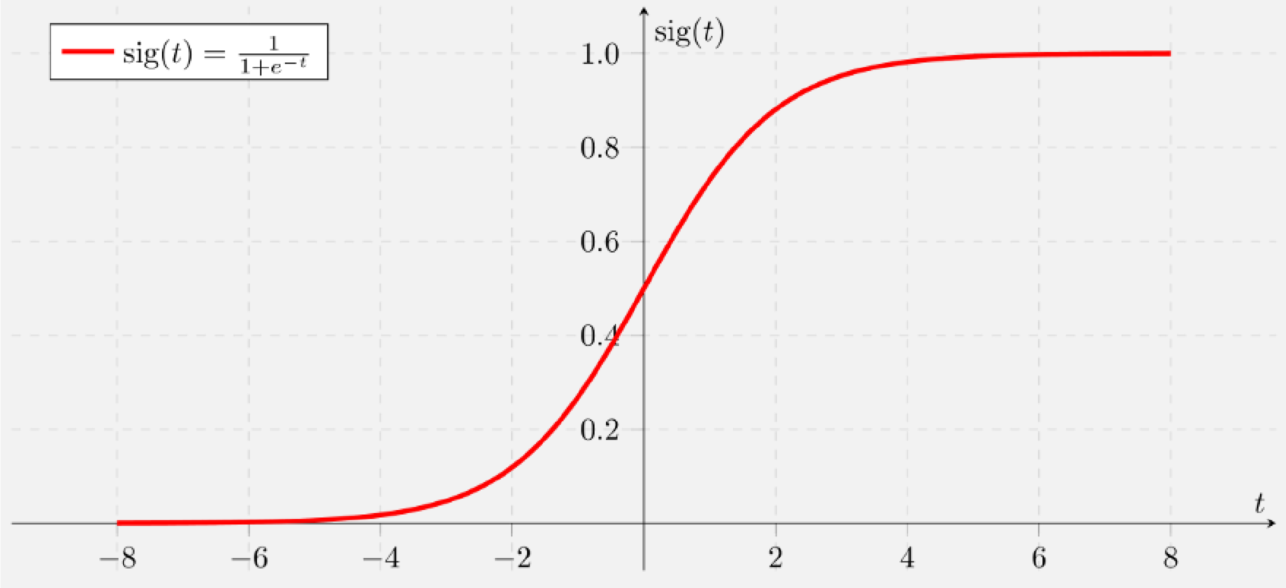


Figure 2: Sigmoid Activation Function

If ‘Z’ goes to infinity, Y(predicted) will become 1 and if ‘Z’ goes to negative infinity, Y(predicted) will become 0.

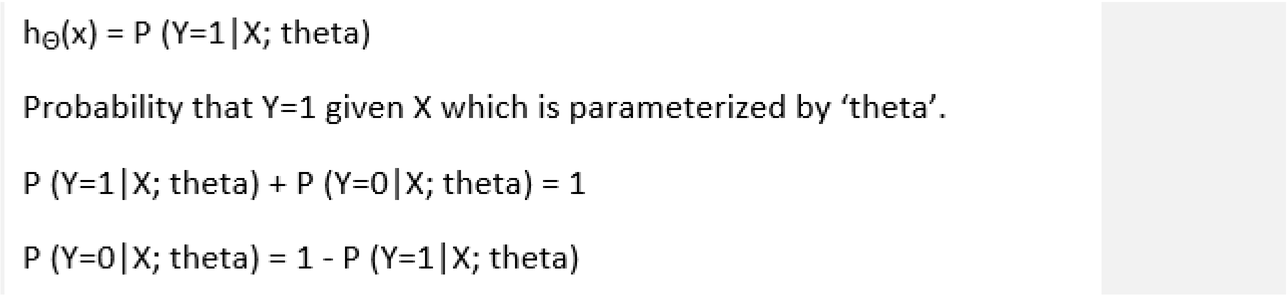
Analysis of the hypothesis

The output from the hypothesis is the estimated probability. This is used to infer how confident can predicted value be actual value when given an input X. Consider the below example,

X = [x0 x1] = [1 IP-Address]

Based on the x1 value, let’s say we obtained the estimated probability to be 0.8. This tells that there is 80% chance that an email will be spam.

Mathematically this can be written as,



$h_{\Theta}(x) = P\ (Y=1|X;\ theta)$

Probability that Y=1 given X which is parameterized by ‘theta’.

$P\ (Y=1|X;\ theta) + P\ (Y=0|X;\ theta) = 1$

$P\ (Y=0|X;\ theta) = 1 - P\ (Y=1|X;\ theta)$

Figure 3: Mathematical Representation

This justifies the name ‘logistic regression’. Data is fit into linear regression model, which then be acted upon by a logistic function predicting the target categorical dependent variable.

Types of Logistic Regression

1. Binary Logistic Regression

The categorical response has only two 2 possible outcomes. Example: Spam or Not

2. Multinomial Logistic Regression

Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)

3. Ordinal Logistic Regression

Three or more categories with ordering. Example: Movie rating from 1 to 5

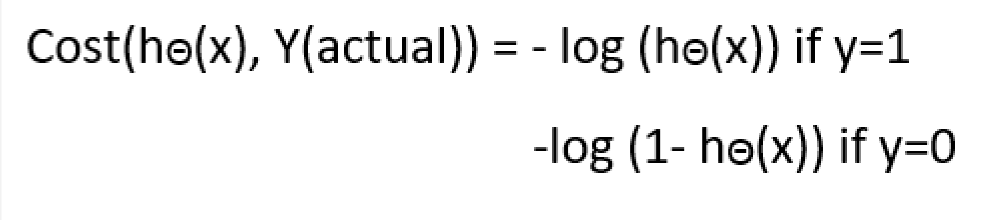
Decision Boundary

To predict which class a data belongs, a threshold can be set. Based upon this threshold, the obtained estimated probability is classified into classes.

Say, if predicted_value ≥ 0.5, then classify email as spam else as not spam.

Decision boundary can be linear or non-linear. Polynomial order can be increased to get complex decision boundary.

Cost Function



Cost(hΘ(x), Y(actual)) = - log (hΘ(x)) if y=1

-log (1- hΘ(x)) if y=0

Figure 4: Cost Function of Logistic Regression

Why cost function which has been used for linear can not be used for logistic?

Linear regression uses mean squared error as its cost function. If this is used for logistic regression, then it will be a non-convex function of parameters (theta). Gradient descent will converge into global minimum only if the function is convex.

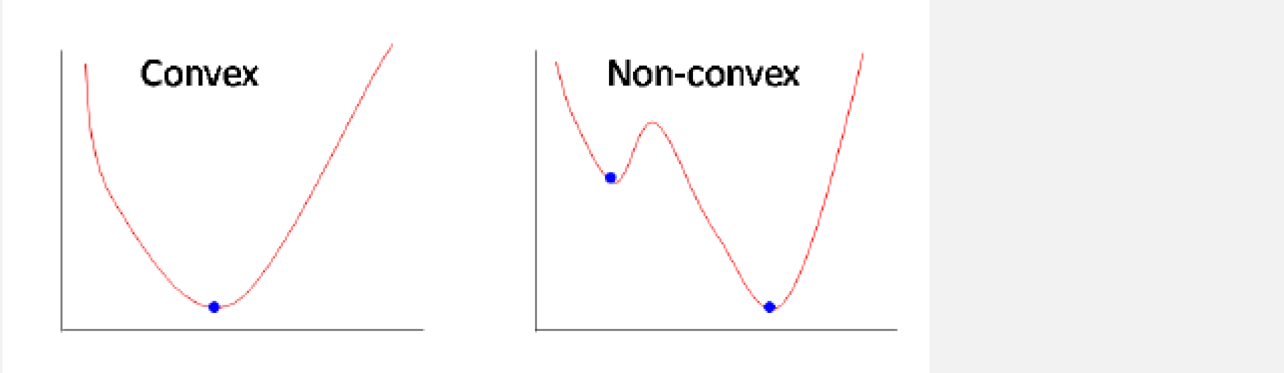
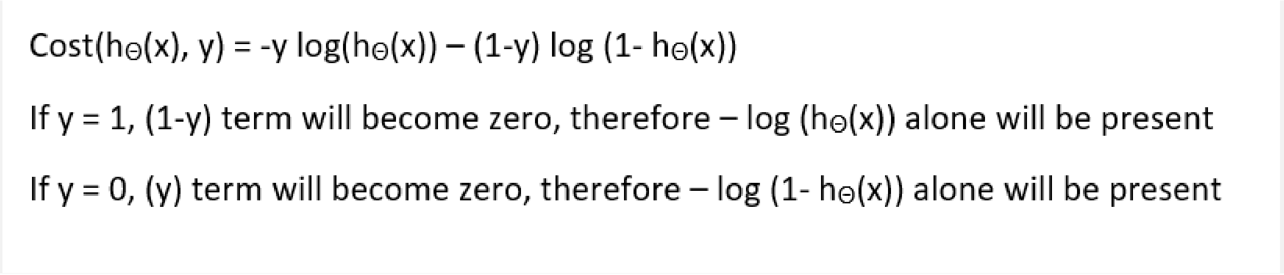


Figure 5: Convex and non-convex cost function

Cost function explanation

Simplified cost function



$\text{Cost}(h_\Theta(x), y) = -y \log(h_\Theta(x)) – (1-y) \log (1- h_\Theta(x))$

If $y = 1$, $(1-y)$ term will become zero, therefore $– \log (h_\Theta(x))$ alone will be present

If $y = 0$, (y) term will become zero, therefore $– \log (1- h_\Theta(x))$ alone will be present

Figure 8: Simplified Cost Function

Why this cost function?

This negative function is because when we train, we need to maximize the probability by minimizing loss function. Decreasing the cost will increase the maximum likelihood assuming that samples are drawn from an identically independent distribution.

Logistic regression, despite its name, is a classification model rather than regression model.

Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes.

It is an extensively employed algorithm for classification in industry.

The logistic regression model, like the Adaline and perceptron, is a statistical method for binary classification that can be generalized to multiclass classification.

Scikit-learn has a highly optimized version of logistic regression implementation, which supports multiclass classification task

Logistic regression is another powerful supervised ML algorithm used for binary classification problems (when target is categorical). The best way to think about logistic regression is that it is a linear regression but for classification problems.

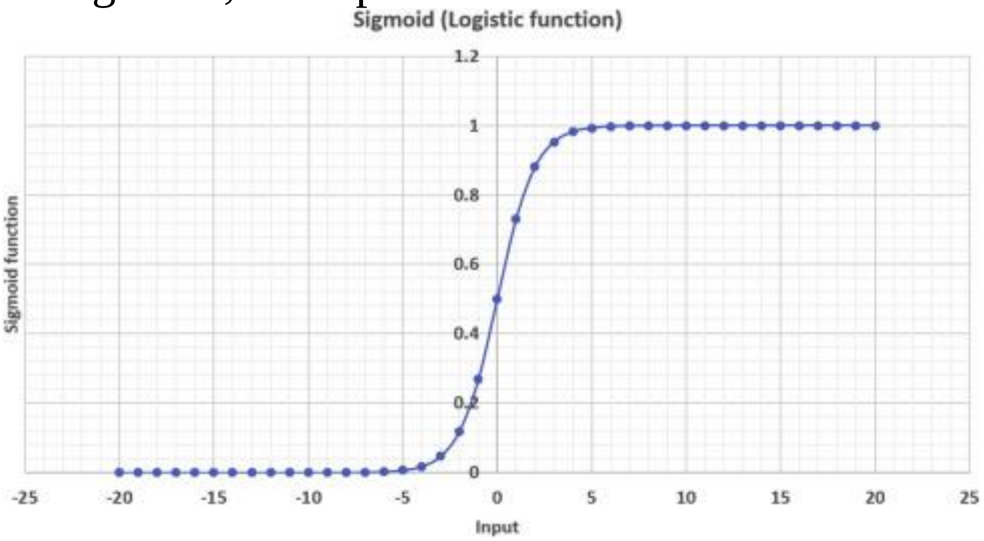
Logistic regression essentially uses a logistic function defined below to model a binary output variable (Tolles & Meurer, 2016). The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1.

In addition, as opposed to linear regression, logistic regression does not require a linear relationship between inputs and output variables.

This is due to applying a nonlinear log transformation to the odds ratio (will be defined shortly).

(5.6)Logisticfunction=11+e−x

In the logistic function equation, *x* is the input variable. Let's **feed in values** −20 to 20 into the logistic function. As illustrated in Fig. 5.17, the inputs have been transferred to between 0 and 1.



Logistic regression is another technique borrowed by machine learning from the field of statistics.

It is the go-to method for binary classification problems (problems with two class values). In this post you will discover the logistic regression algorithm for machine learning.

After reading this post you will know:

- The many names and terms used when describing logistic regression (like log odds and logit).
- The representation used for a logistic regression model.
- Techniques used to learn the coefficients of a logistic regression model from data.
- How to actually make predictions using a learned logistic regression model.

- Where to go for more information if you want to dig a little deeper.

This post was written for developers interested in applied machine learning, specifically predictive modeling. You do not need to have a background in linear algebra or statistics.

Logistic Regression for Machine Learning

Logistic regression is another technique borrowed by machine learning from the field of statistics.

It is the go-to method for binary classification problems (problems with two class values). In this post you will discover the logistic regression algorithm for machine learning.

After reading this post you will know:

- The many names and terms used when describing logistic regression (like log odds and logit).
- The representation used for a logistic regression model.
- Techniques used to learn the coefficients of a logistic regression model from data.
- How to actually make predictions using a learned logistic regression model.
- Where to go for more information if you want to dig a little deeper.

This post was written for developers interested in applied machine learning, specifically predictive modeling. You do not need to have a background in linear algebra or statistics.

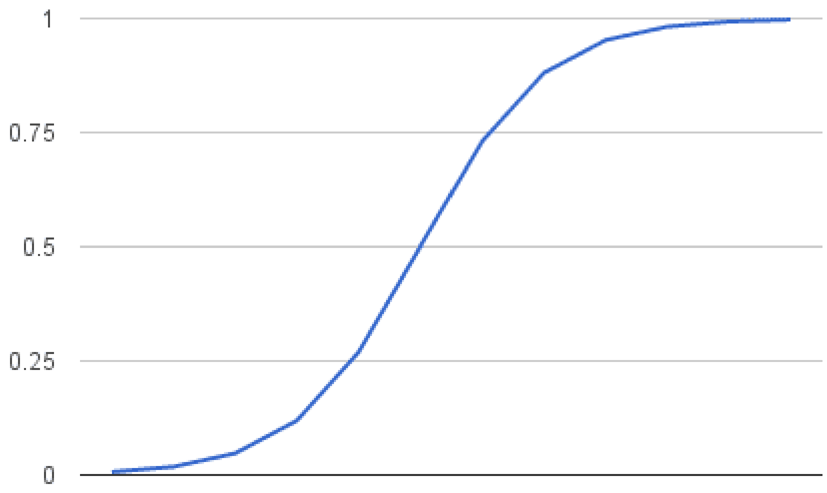
Let’s get started.

Logistic Function

Logistic regression is named for the function used at the core of the method, the logistic function.

The Logistic function , also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.
1 / (1 + e^-value)

Where e is the (Euler’s number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.



Logistic Function

Now that we know what the logistic function is, let’s see how it is used in logistic regression.

Representation Used for Logistic Regression

Logistic regression uses an equation as the representation, very much like linear regression.

Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

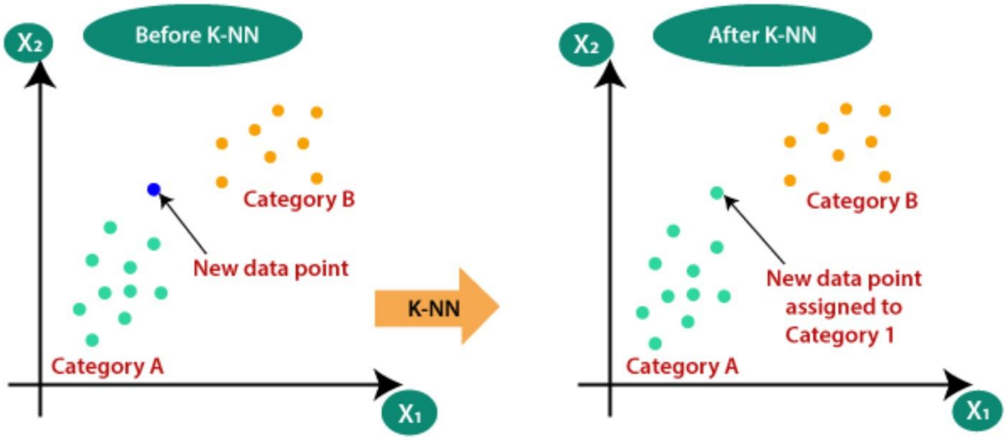
y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

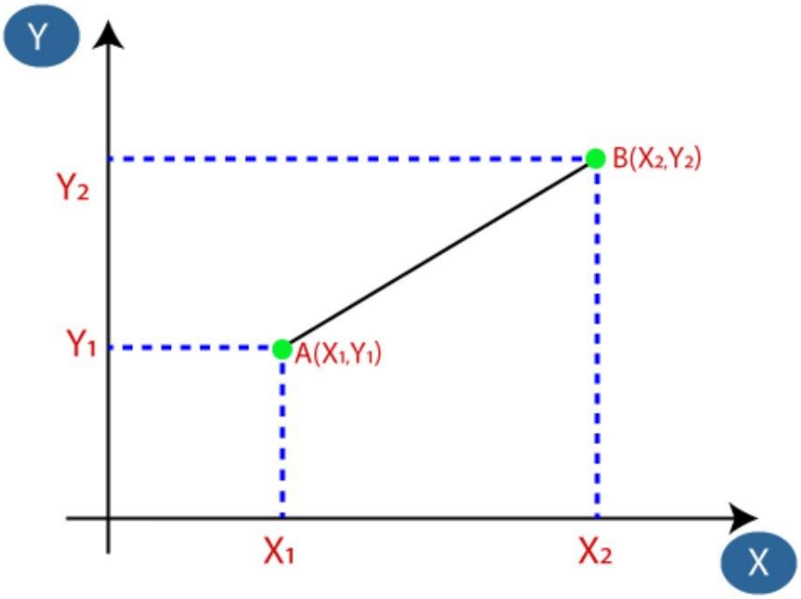
Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b’s).

K -nearest neighbour:-

1. K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
2. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
3. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
4. K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
5. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.





Euclidean Distance between A_1 and B_2 = $\sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

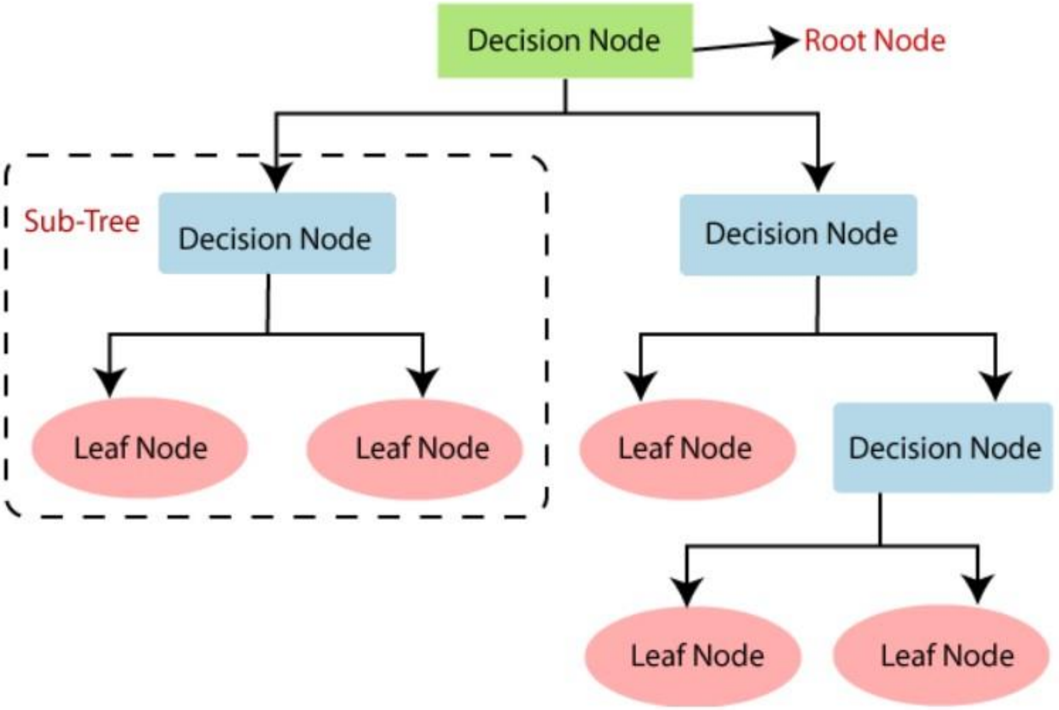
Decision Tree Classification Algorithm

- Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**

- In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

- The decisions or the test are performed on the basis of features of the given dataset.

- ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***


Decision Node
Root Node
Sub-Tree
Decision Node
Decision Node
Leaf Node
Leaf Node
Leaf Node
Decision Node
Leaf Node
Leaf Node

Chapter -2
Literature survey

This section comprises some of the literatures used for Sports Prediction using Supervised Learning. In this paper, the author has used Regression, which is a type of supervised learning method. The paper attempts to predict results for baseball, basketball, football, and hockey games.

Here, a straightforward machine learning model, logistically-weighted regularized linear method of least squares regression was used for predicting the sports outcomes. the information used here during this is thirty years previous taken from websites. The strategy used works best in basketball, possible as a result of it's a high scoring and it's long seasons.

The soccer predictions are sensible however can be made far better, additionally the hockey predictions can be made somewhat higher. The limitation is that it works higher in soccer and hockey than in baseball, however in baseball, the result predictions are nearer to a theoretical optimum. Predictor outcomes are very close however a lot of work is important for predicting baseball, football, and hockey a lot of optimal .

In this paper, the authors have used a network-based approach for predicting the sports outcomes, using a TeamRank method. A TeamRank method defines directed graphs of sports teams based on the observed outcomes of individual games, and use these networks to infer the importance of teams that determine their rankings.

The method also proposes some extensions to TeamRank that consider overall team wins records and shifts in momentum over time. It has provided 70% accuracy. The limitation is that it doesn’t provide useful rankings of teams that are not clustered by league divisions .

In this paper, the authors has used a probabilistic graphical model to learn the team Skills. The model decomposes the relative weights of luck and ability in every sporting event. The paper has introduced the skill coefficient, denoted by phi(ø), that measures the departure between the observed final season score distribution and what one could expect from a competition based on pure luck plus contextual circumstances. The larger the value of phi, the more distant from a no-luck competition. The paper also proposed a technique based on phi that identifies which teams are significantly more (and less) skilled than the others in the league. The limitation is that the Bigger teams are less cohesive but have more conflicts and are difficult to manage .

Functionality/Working of Project:-

The attempt to explore three simple to grab factors and uses doe design of experiment methodology to see the consequences on sports result prediction.

It has used Design of Experiment, Predict modeling and Data quality management as the performance measures and also randomization method is used, which falls under supervised learning.

These three factors consist of Controlled input, Uncontrolled input, and Output/Response. The analysis information modeling used here is predicated on the NBA (National Basketball Association) information. All the information is extracted from the official and public data supply of 2015 to 2016 regular season. The limitation is that solely 2 factors are enclosed during this model, they're factor-R and H, all the variance inflation factors (VIF) are smaller than ten and bigger than zero, proves that no collinearity existed. Also, no interaction candidate terms enclosed in this model .

In this paper, the attempt to present an advanced interface for sports tournament predictions that uses direct manipulation to allow users to make nonlinear predictions. The Direct manipulation interface is easy to understand for the sports enthusiasts. The paper presents an advanced interface for sports results that uses direct manipulation to give users to form nonlinear predictions.

The planned technique higher matches the manner individuals truly build predictions, like initial selecting the winner and so filling up the remainder of the bracket. The papers have used the subsequent performance measures:

1. outlined a listing of standard to grasp the present wants from football game tournament prediction input interfaces and also the state of the art in human-computer interaction (HCI), as well as direct manipulation.

2. designed and implemented an advanced interface which complies with the design to allow users drag and drop team icons (or badges) toward their final position in a single view.

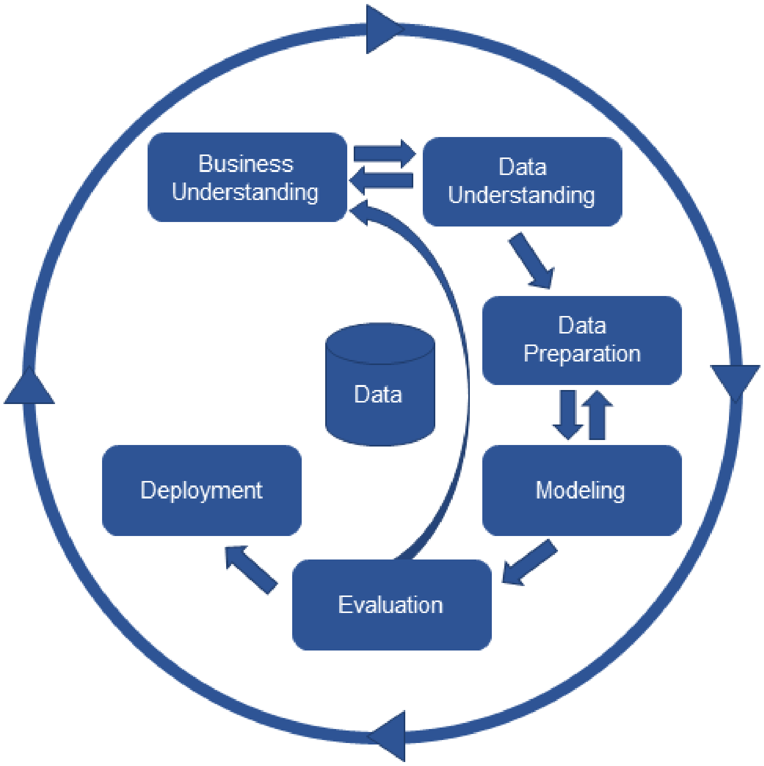
3.Also designed and implemented a set of novel visualization logs to make sense of the 504,307 recorded interaction logs from the 3,029 visitors in order to identify and discuss interesting behaviors. The limitation is that Only successful predictions are being analyzed here .

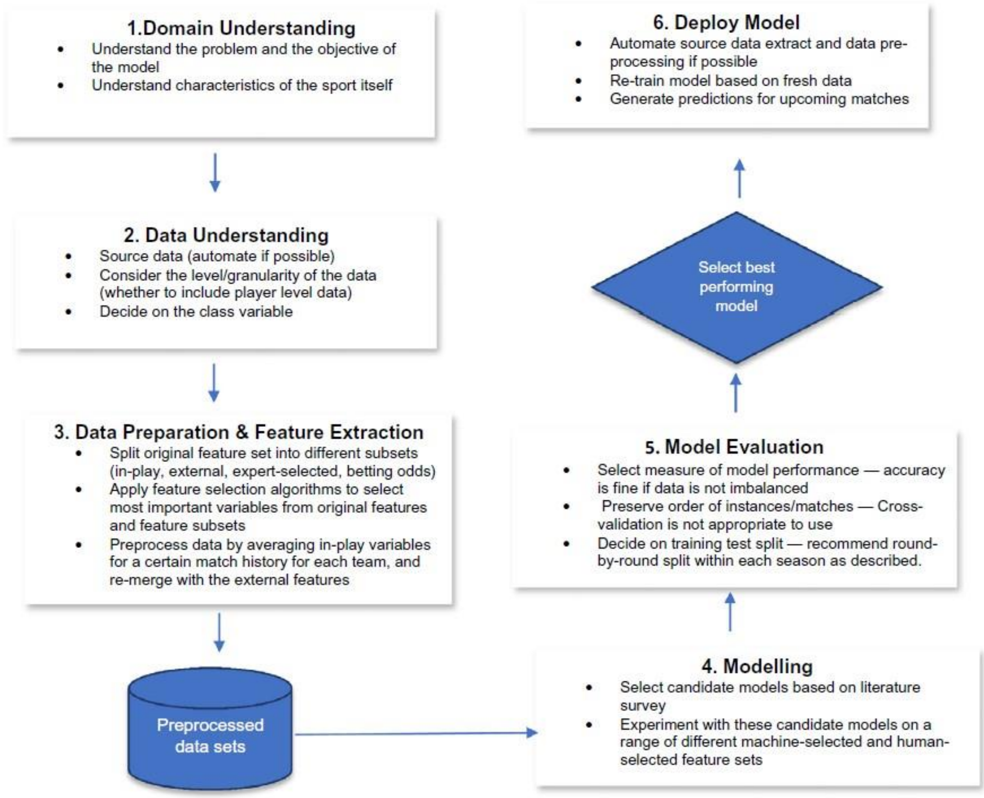
The paper attempts to establish a social-network sports lottery system to support Users in predicting and simulating sports lottery betting. The principal purpose of developing a social-network-based sports lottery community system was to exploit the main functions of social network to: obtain potential users and related information; use a self-service mechanism to enrich website content; promote frequent use to encourage users to continually use the proposed application; acquire numerous users to enhance the prediction accuracy; and design the interaction and prediction mechanisms to enhance the features of the developed sites

This is a type of classification problem which come under supervised learning in which labeled dataset i.e. the data set in which all attributes and tuples are sorted (in the proper order).this labeled data set goes under training using various methods like data mining, KNN, decision tree, ANN.Sports prediction is one of the popular topic in the recent time .by predicting it we generally predict the outcome in advance.
Artificial Neural Networks (ANNs) are perhaps the most

commonly applied approach among ML mechanisms to the sport result prediction problem. Thus, for this review, we focus on studies that have applied ANNs. An ANN usually contains interconnected components (neurons) that transform a set of inputs into a desired output. The power of ANN comes from the non-linearity of the hidden neurons in adjusting weights that contribute to the final decision.
ANN output often relies on input features and other components associated with the network, such as these weights. The ANN model is constructed after processing the training dataset that contains the features used to build the ANN classification model. In other words, weights associated with interconnected components are continuously changing to accomplish high levels of predictive accuracy. These changes are performed by the ANN algorithm to fulfill the desired model’s accuracy given earlier by the user. This may lead in some cases to the problem of over fitting, as well as wasting computing resources such as training time and memory

To deploy the model in model /steps to perform prediction using machine learning:-




1.Domain Understanding
Understand the problem and the objective of the model
Understand characteristics of the sport itself
2. Data Understanding
Source data (automate if possible)
Consider the level/granularity of the data (whether to include player level data)
Decide on the class variable
3. Data Preparation & Feature Extraction
Split original feature set into different subsets (in-play, external, expert-selected, betting odds)
Apply feature selection algorithms to select most important variables from original features and feature subsets
Preprocess data by averaging in-play variables for a certain match history for each team, and re-merge with the external features
Preprocessed data sets
4. Modelling
Select candidate models based on literature survey
Experiment with these candidate models on a range of different machine-selected and human-selected feature sets
5. Model Evaluation
Select measure of model performance — accuracy is fine if data is not imbalanced
Preserve order of instances/matches — Cross-validation is not appropriate to use
Decide on training test split — recommend round-by-round split within each season as described.
Select best performing model
6. Deploy Model
Automate source data extract and data pre-processing if possible
Re-train model based on fresh data
Generate predictions for upcoming matches

Over the past years, a variety of data-capturing technologies have become available in sport business . These technologies allow sport management businesses to capture and collect data on games, bidding, bookmaker odds, playing styles, scores, and many other sport attributes .

Such a repository of data allows firms to garner invaluable insights through the leveraging of data analytics. There have also been several discussions around this issue in literary and business circles .

The studies suggest that a data-driven approach to sport business and marketing is an interesting area to investigate. In this context, data analytics could be of immense value. Data analytics in sport has become an integral part of sport business .

Data mining-based models are also being explored in sport . Other forms of analytical techniques being explored include, page rank models , numerical algorithms and machine learning . Recently, the Fédération Internationale de Football Association approved the use of Electronic Performance and Tracking Systems (EPTS) during competition.

Now, the physical data collected using EPTS, from both training sessions and live matches, can be used to evaluate the extent to which match performance can be predicted [12]. Scholars have proposed a normalised root mean square error metric for analysing the results obtained from the application of machine-learning algorithms to the data collected for various physical variables. Specific physical variables can also act as representatives of several other variables, which are highly correlated, to further reduce the number of variables that must be periodically analysed by coaches .Moreover, continuously growing sports platforms related to games incentivize bookies and bettors to bet on match results as a game changes ball by ball. Hence, attempts have been made to predict match results based on historical match data . Further, the emotional expressions of sports team members, and their correlation with the team's performance, can be analysed to draw conclusions about the psychological mind-set of players . Design of experiments, is applied to frame experiments explain the variation under certain conditions to predict the outcome in certain variables . Finally, researchers have introduced variables such as possession (ball occupation) and territory (dominance of territory) and a novel visual analytics system to analyse tactical transitions in a continuous ball match

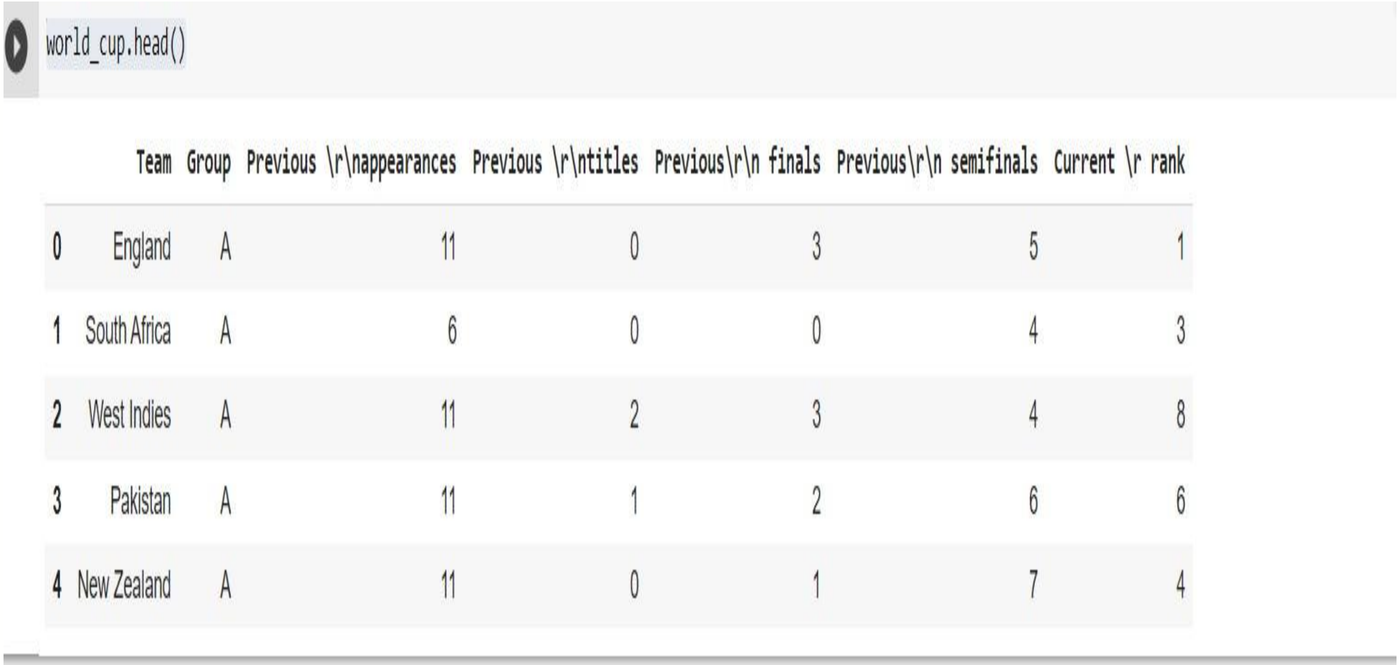
Sport analytics has been receiving significant research attention and studies have suggested that sport analytics could be used to a greater degree in businesses engaged in sport . A review in this field is of relevance as it can lead to understanding the state of research and classifications of study within this topic . There have been scoping reviews in sport management in the areas of sport governance and spectator sport . However, a Systematic Literature Review (SLR) is an appropriate approach as it enables the researcher to structure a research field and understand the state of research and emerging research themes . This, in turn, discovers the application of the area and provides guidance toward understanding or evaluating the area . SLR is also considered to be a valid approach as it is an integral part of research and a vital process in structuring a research field . It identifies the conceptual content of the field and is found to be an effective approach to understanding the intellectual foundations of a research field and to classify studies .

Chapter -3
Result and working project

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
```

data set used:-

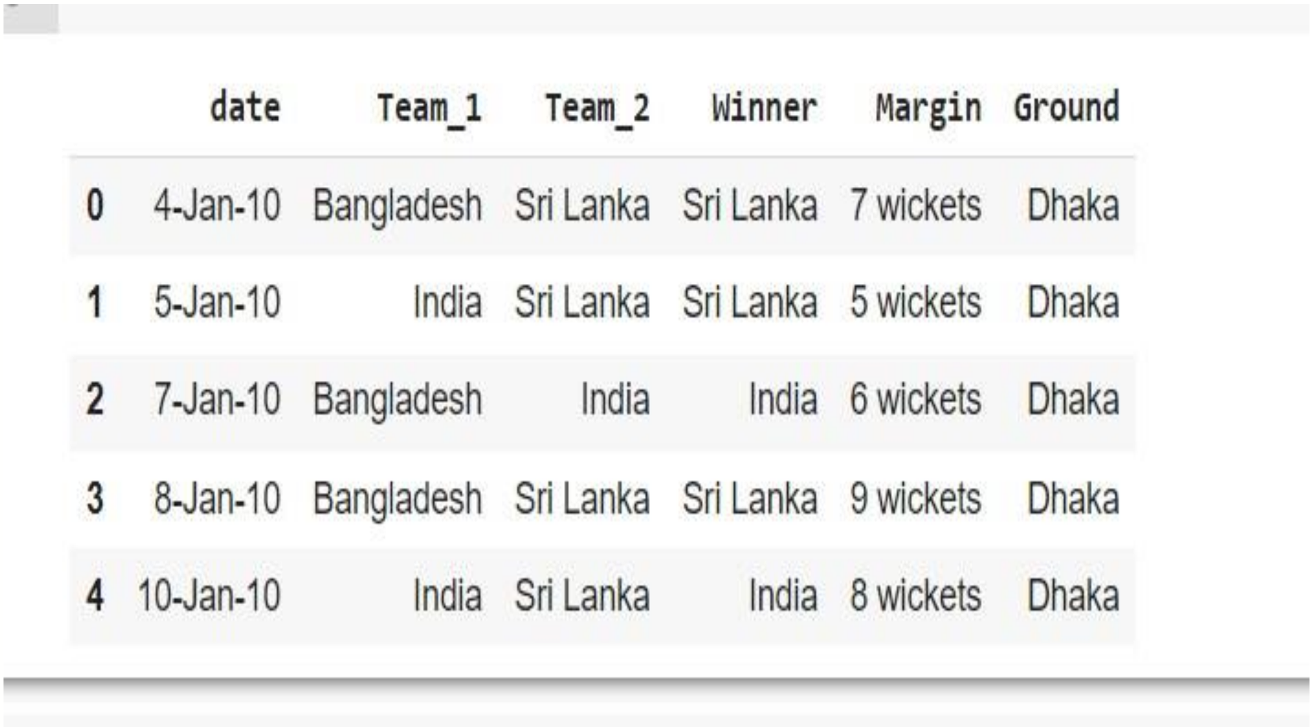
```
world_cup=pd.read_csv('World Cup 2019 Dataset.csv')
result=pd.read_csv('results.csv')
fixtures=pd.read_csv('fixtures.csv')
ranking=pd.read_csv('icc_rankings.csv')

world_cup.head()
```



```
world_cup.head()
```

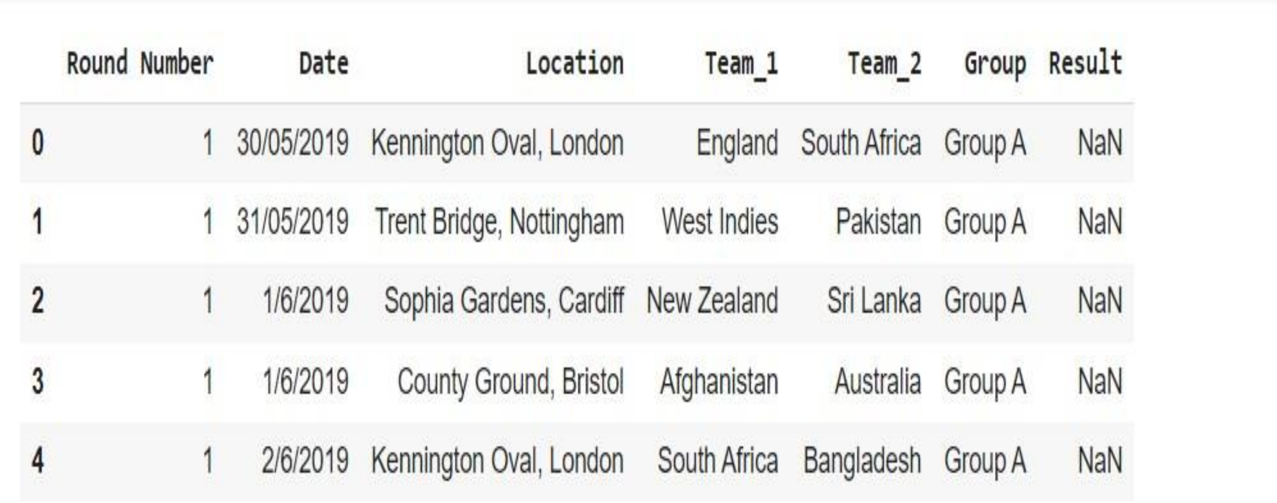
	Team	Group	Previous \r\nappearances	Previous \r\ntitles	Previous\r\n finals	Previous\r\n semifinals	Current \r rank
0	England	A	11	0	3	5	1
1	South Africa	A	6	0	0	4	3
2	West Indies	A	11	2	3	4	8
3	Pakistan	A	11	1	2	6	6
4	New Zealand	A	11	0	1	7	4

```
result.head()
```



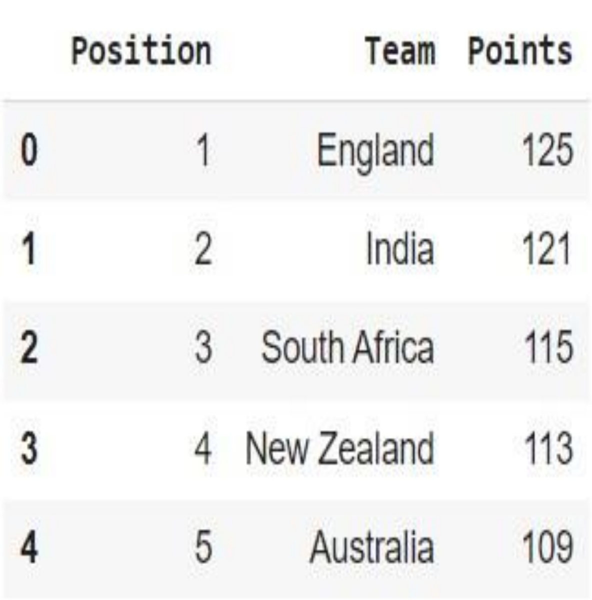
	date	Team_1	Team_2	Winner	Margin	Ground
0	4-Jan-10	Bangladesh	Sri Lanka	Sri Lanka	7 wickets	Dhaka
1	5-Jan-10	India	Sri Lanka	Sri Lanka	5 wickets	Dhaka
2	7-Jan-10	Bangladesh	India	India	6 wickets	Dhaka
3	8-Jan-10	Bangladesh	Sri Lanka	Sri Lanka	9 wickets	Dhaka
4	10-Jan-10	India	Sri Lanka	India	8 wickets	Dhaka

```
fixtures.head()
```



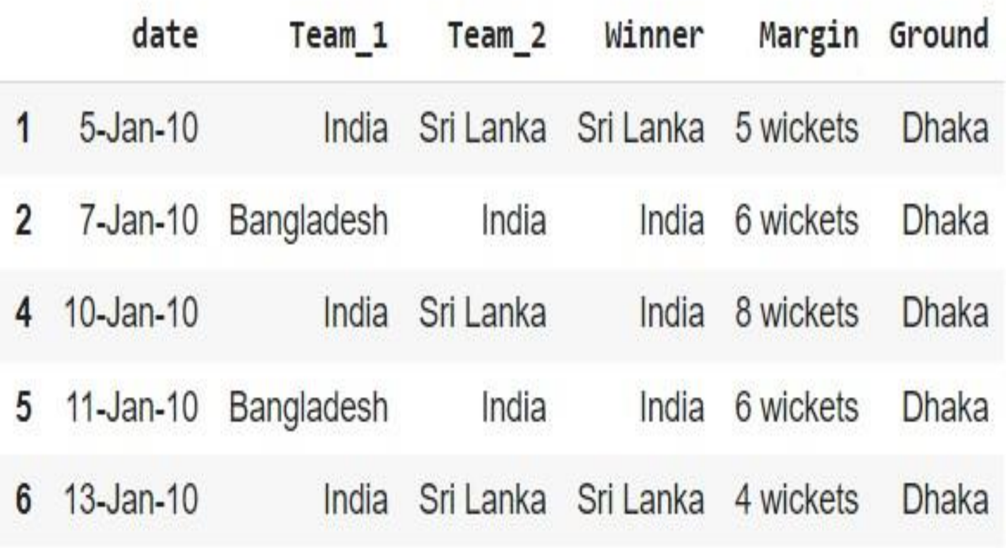
	Round Number	Date	Location	Team_1	Team_2	Group	Result
0	1	30/05/2019	Kennington Oval, London	England	South Africa	Group A	NaN
1	1	31/05/2019	Trent Bridge, Nottingham	West Indies	Pakistan	Group A	NaN
2	1	1/6/2019	Sophia Gardens, Cardiff	New Zealand	Sri Lanka	Group A	NaN
3	1	1/6/2019	County Ground, Bristol	Afghanistan	Australia	Group A	NaN
4	1	2/6/2019	Kennington Oval, London	South Africa	Bangladesh	Group A	NaN

```
ranking.head()
```



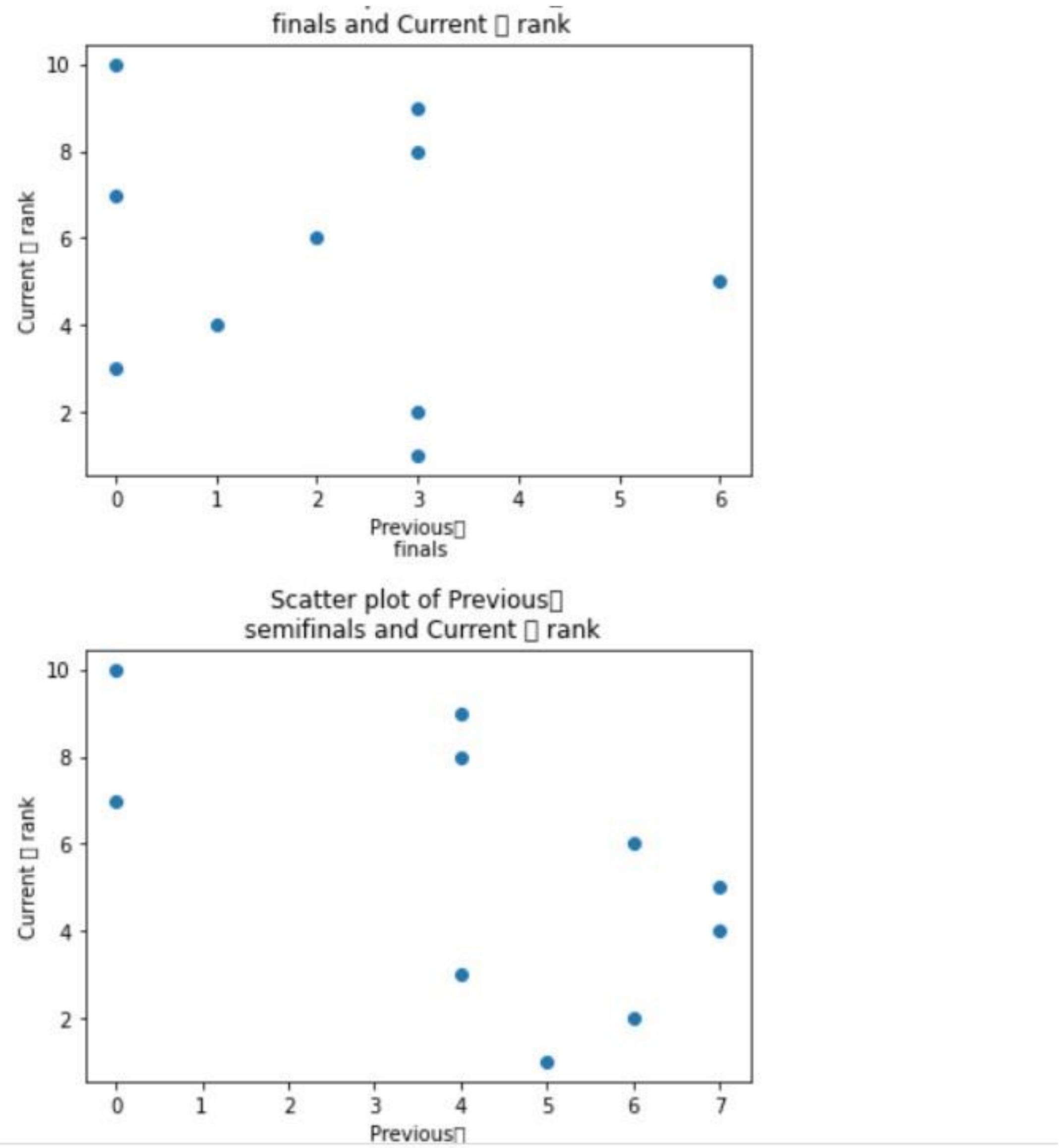
	Position	Team	Points
0	1	England	125
1	2	India	121
2	3	South Africa	115
3	4	New Zealand	113
4	5	Australia	109

```
india=result[(result['Team_1']=='India')|(result['Team_2'
]=='India')]
india.head()
```

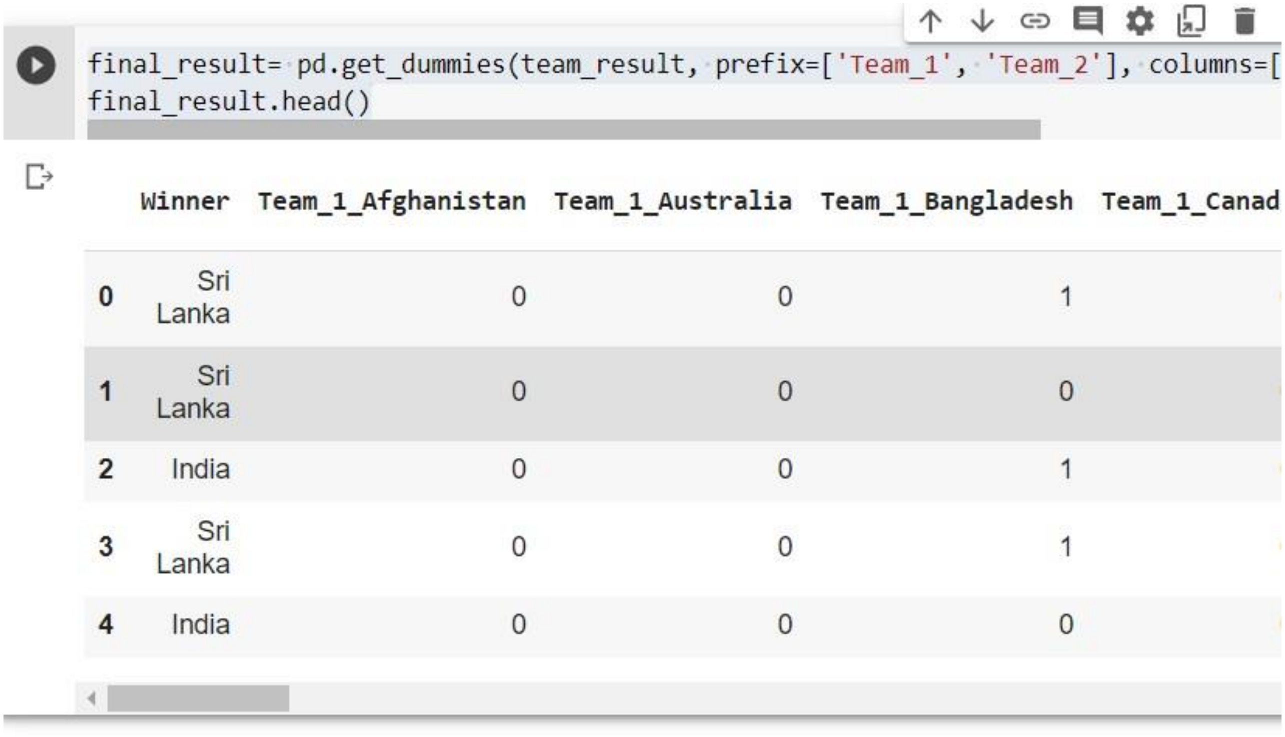


	date	Team_1	Team_2	Winner	Margin	Ground
1	5-Jan-10	India	Sri Lanka	Sri Lanka	5 wickets	Dhaka
2	7-Jan-10	Bangladesh	India	India	6 wickets	Dhaka
4	10-Jan-10	India	Sri Lanka	India	8 wickets	Dhaka
5	11-Jan-10	Bangladesh	India	India	6 wickets	Dhaka
6	13-Jan-10	India	Sri Lanka	Sri Lanka	4 wickets	Dhaka

```
World_cup_teams=['England', ' South Africa', 'West Indies
', 'Pakistan', 'New Zealand', 'Sri Lanka', 'Afghanistan',
 'Australia', 'Bangladesh', 'India']
team1=result[result['Team_1'].isin(World_cup_teams)]
team2=result[result['Team_2'].isin(World_cup_teams)]
teams=pd.concat((team1,team2))
teams=teams.drop_duplicates()


print(world_cup.columns[0:13])
x_names = world_cup.columns[0:13]          # ploting all
 x variables with y
y_name = world_cup.columns[-1]
def pllot(x,y):
  plt.scatter(world_cup[x],world_cup[y])
  plt.xlabel(x)
  plt.ylabel(y)
  plt.title("Scatter plot of "+x+" and "+y)
  plt.show()
for i in x_names:
  pllot(i, y_name)
```



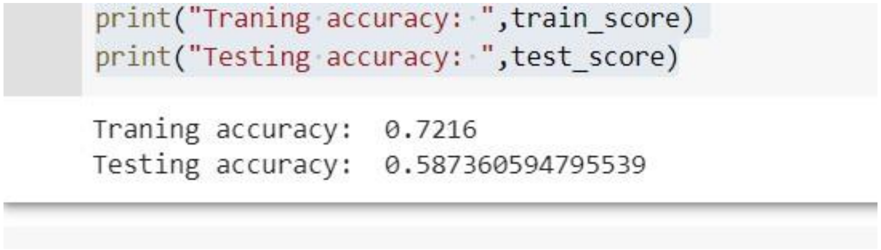
```
final_result= pd.get_dummies(team_result, prefix=['Team_1
', 'Team_2'], columns=['Team_1', 'Team_2'])
final_result.head()
```



```
final_result= pd.get_dummies(team_result, prefix=['Team_1', 'Team_2'], columns=[
final_result.head()
```

	Winner	Team_1_Afghanistan	Team_1_Australia	Team_1_Bangladesh	Team_1_Canad
0	Sri Lanka	0	0	1	
1	Sri Lanka	0	0	0	
2	India	0	0	1	
3	Sri Lanka	0	0	1	
4	India	0	0	0	

```
X=final_result.drop(['Winner'],axis=1)
y=final_result['Winner']
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.30,random_state=42)

model=LogisticRegression()
model.fit(X_train,y_train)
train_score=model.score(X_train,y_train)
test_score=model.score(X_test,y_test)
print("Traning accuracy: ",train_score)
print("Testing accuracy: ",test_score)
```

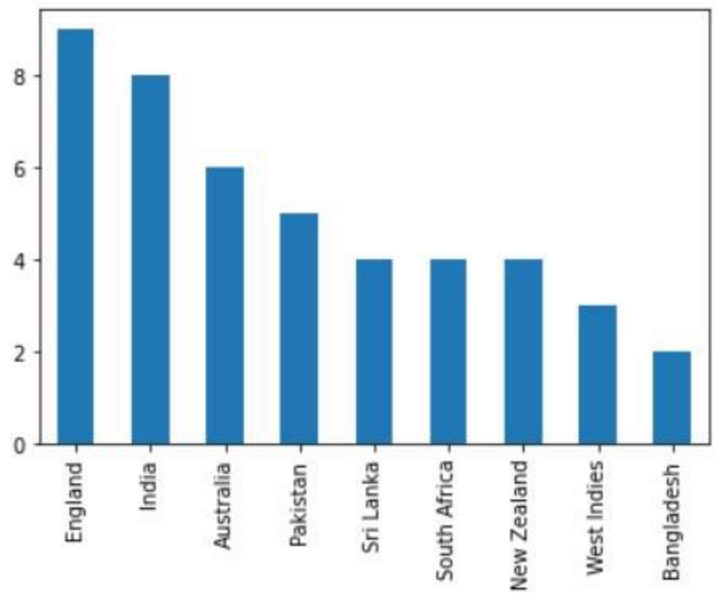


```
print("Traning accuracy: ",train_score)
print("Testing accuracy: ",test_score)
```

```
Traning accuracy:  0.7216
Testing accuracy:  0.587360594795539
```

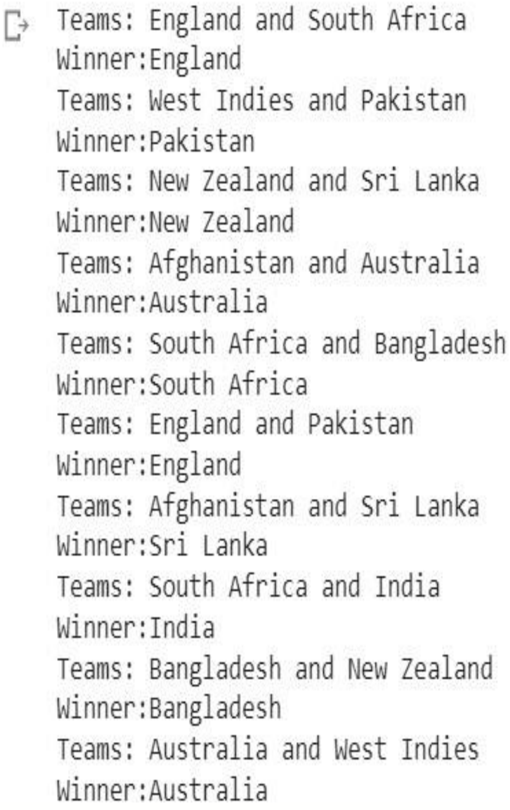
```
final_set=fixture[['Team_1','Team_2']]
final_set = pd.get_dummies(final_set, prefix=['Team_1', '
Team_2'], columns=['Team_1', 'Team_2'])
for col in (set(final_result.columns)-
set(final_set.columns)):
    final_set[col]=0
final_set=final_set.sort_index(axis=1)
final_set=final_set.drop(['Winner'],axis=1)
final_set.head()
```

```
fixture['Result'].value_counts().plot(kind='bar')
```



```
def predict_result(matches,final_result,ranking,model,mat
ch_type):
    #obtaining team position
    team_position=[]
    for match in matches:
        team_position.append([ranking.loc[ranking['Team']
 == match[0],'Position'].iloc[0],ranking.loc[ranking['Tea
m'] == match[1],'Position'].iloc[0]])

    #transforming data into useful information
    final=pd.DataFrame()
    final[['Team_1','Team_2']]=pd.DataFrame(matches)
    final_set=final
    final_set = pd.get_dummies(final_set, prefix=['Team_1', 'Team_2'], columns=['Team_1', 'Team_2'])

    for col in (set(final_result.columns)-set(final_set.columns)):
        final_set[col]=0
    final_set=final_set.sort_index(axis=1)
    final_set=final_set.drop(['Winner'],axis=1)


    #predict winner
    prediction=model.predict(final_set)


    #Results from League mathes
    if match_type == 'League':
        print("League Matches")

        final_fixture=fixtures[0:45]
        for index,tuples in final_fixture.iterrows():
            print("Teams: " + tuples['Team_1']+ " and " + tuples['Team_2'])
            print("Winner: "+ prediction[index])
            fixtures['Result'].iloc[index]=prediction[index]

        Semi_final_teams=[]
        for i in range(4):
            Semi_final_teams.append(fixture['Result'].value_counts().index[i])
        matches=[(Semi_final_teams[0],Semi_final_teams[3]),(Semi_final_teams[1],Semi_final_teams[2])]
        match_type="Semi-Final"
        predict_result(matches,final_result,ranking,model,match_type)

    #Result from semi-final
    elif match_type == 'Semi-Final':
        print("\nSemi-Final Matches")
        final_fixture=fixtures[45:47]
        for index,tuples in final_fixture.iterrows():
            fixtures['Team_1'].iloc[index]=final['Team_1'
].iloc[index-45]
            fixtures['Team_2'].iloc[index]=final['Team_2'
].iloc[index-45]
            fixtures['Team_1_position'].iloc[index]=team_
position[index-45][0]
            fixtures['Team_2_position'].iloc[index]=team_
position[index-45][1]
        final_fixture=fixtures[45:47]
        for index,tuples in final_fixture.iterrows():
            print("Teams: " + tuples['Team_1']+ " and " +
 tuples['Team_2'])
            print("Winner: "+ prediction[index-45])
            fixtures['Result'].iloc[index]=prediction[ind
ex-45]
        matches=[(prediction[0],prediction[1])]
        match_type="Final"
        predict_result(matches,final_result,ranking,model
,match_type)

    #Result of Final
    elif match_type == 'Final':
        print("\nFinal Match")
        final_fixture=fixtures[47:48]
        for index,tuples in final_fixture.iterrows():
            fixtures['Team_1'].iloc[index]=final['Team_1'
].iloc[index-47]
            fixtures['Team_2'].iloc[index]=final['Team_2'
].iloc[index-47]
            fixtures['Team_1_position'].iloc[index]=team_
position[index-47][0]
            fixtures['Team_2_position'].iloc[index]=team_
position[index-47][1]
        final_fixture=fixtures[47:48]
        for index,tuples in final_fixture.iterrows():
            print("Teams: " + tuples['Team_1']+ " and " +
 tuples['Team_2'])
            print("Winner: "+ prediction[0]+"\n")
            fixtures['Result'].iloc[index]=prediction[ind
ex-47]
        print("Winner Of the tournament is: " + fixtures[
'Result'].iloc[47])
```

```
def predict_result(matches,final_result,ranking,model,match_type):
    #obtaining team position
    team_position=[]
    for match in matches:
        team_position.append([ranking.loc[ranking['Team'] == match[0],'Position'].iloc[0],ranking.loc[ranking['Team'] == match[1],'Position'].iloc[0]])

    #transforming data into useful information
    final=pd.DataFrame()
    final[['Team_1','Team_2']]=pd.DataFrame(matches)
    final_set=final
    final_set = pd.get_dummies(final_set, prefix=['Team_1', 'Team_2'], columns=['Team_1', 'Team_2'])

    for col in (set(final_result.columns)-set(final_set.columns)):
        final_set[col]=0
    final_set=final_set.sort_index(axis=1)
    final_set=final_set.drop(['Winner'],axis=1)


    #predict winner
    prediction=model.predict(final_set)


    #Results from League mathes
    if match_type == 'League':
        print("League Matches")

        final_fixture=fixtures[0:45]
        for index,tuples in final_fixture.iterrows():
            print("Teams: " + tuples['Team_1']+ " and " + tuples['Team_2'])
            print("Winner: "+ prediction[index])
            fixtures['Result'].iloc[index]=prediction[index]

        Semi_final_teams=[]
        for i in range(4):
            Semi_final_teams.append(fixture['Result'].value_counts().index[i])
        matches=[(Semi_final_teams[0],Semi_final_teams[3]),(Semi_final_teams[1],Semi_final_teams[2])]
        match_type="Semi-Final"
        predict_result(matches,final_result,ranking,model,match_type)

    #Result from semi-final
    elif match_type == 'Semi-Final':
        print("\nSemi-Final Matches")
        final_fixture=fixtures[45:47]
        for index,tuples in final_fixture.iterrows():
            fixtures['Team_1'].iloc[index]=final['Team_1'].iloc[index-45]
            fixtures['Team_2'].iloc[index]=final['Team_2'].iloc[index-45]
            fixtures['Team_1_position'].iloc[index]=team_position[index-45][0]
            fixtures['Team_2_position'].iloc[index]=team_position[index-45][1]
        final_fixture=fixtures[45:47]
        for index,tuples in final_fixture.iterrows():
            print("Teams: " + tuples['Team_1']+ " and " + tuples['Team_2'])
            print("Winner: "+ prediction[index-45])
            fixtures['Result'].iloc[index]=prediction[index-45]
        matches=[(prediction[0],prediction[1])]
        match_type="Final"
        predict_result(matches,final_result,ranking,model,match_type)


for index,tuples in fixture.iterrows():
    print("Teams: " + tuples['Team_1']+ " and " + tuples['Team_2'])
    print("Winner:"+ prediction[index])
```



```
Teams: England and South Africa
Winner:England
Teams: West Indies and Pakistan
Winner:Pakistan
Teams: New Zealand and Sri Lanka
Winner:New Zealand
Teams: Afghanistan and Australia
Winner:Australia
Teams: South Africa and Bangladesh
Winner:South Africa
Teams: England and Pakistan
Winner:England
Teams: Afghanistan and Sri Lanka
Winner:Sri Lanka
Teams: South Africa and India
Winner:India
Teams: Bangladesh and New Zealand
Winner:Bangladesh
Teams: Australia and West Indies
Winner:Australia
```

CONCLUSION :-

We find that different kind of analytical methods and techniques have been applied in diverse context of sport. We also observe that frequency of research in sport analytics is consistently increasing as we could gather from the increasing frequency of such studies in the recent years. Several analytical methods and techniques, through which data analytics has evolved in the field of sport business management, have been studied and explored in the literature. The paper contributes to theory in two specific ways. This study has used systematic analysis of the literature in journals, conferences and other fora in the field of sport analytics. The analysis helps identify impactful studies in the field for scholarship and theoretical understanding. Second, the study classifies and conducts a taxonomy on variety of analytical methods used in different sport contexts. It helps in development of a structure to classify the studies and also understand the theoretical base. Certain practical implications are drawn from the review of studies. The review shows that several authors have demonstrated the application of data analytics in variety of practical contexts like fan-base marketing, consumer sentiments, player bidding, sport injury, player performance, promotions, bidding for games and others. It is also observed that variety of analytical methods have been used. These methods include multivariate regression, descriptive analysis, optimization and even machine-learning methods (logistic regression, support vector machines, random forest etc.). Increasingly, these methods are being adopted in practice with the application of open source technologies like Python, R, Gephi and others. There are certain limitations to the study which also point towards future research directions. The study uses a systematic review to determine the current position of studies published in conferences proceedings, journals, books and chapters in edited volumes. While a systematic literature review is useful to develop a scholarship base, it is still limited by the sample of literature as selected by the researcher. A systematic review cannot always be exhaustive as it is not possible to enumerate all the research historically. Further, the author does not attempt to perform a longitudinal analysis in this study as the author’s focus was on taxonomical review and uncover range of analytical methods and business contexts. As a future research, longitudinal analysis of review can be attempted to discover the trend in research. As future research direction, it would also be useful to use semantic analytics and bibliometrics to perform citation and co-citation analysis on sport analytics. Such an analysis, based on citations, co-citations and co-authorship, would uncover intellectual structure of this field while also discovering emerging trends, challenges and prospects.

One of the vital applications in sport that requires good predicting accuracy is match result prediction. Traditionally, the results of the matches are predicted using mathematical and statistical models that are often verified by a domain expert. Due to the specific nature of match-related features to different sports, results across different studies in this application can generally not be compared directly. Despite the increasing use of ML models for sport prediction, more accurate models are needed. This is due to the high volumes of betting on sport, and for sport managers seeking useful knowledge for modelling future matching strategies. Therefore, ML seems an appropriate methodology for sport prediction since it generates that can predict match results using predefined features in a historical dataset.

References :-

1. N. Abdelhamid, A. Ayesh, F. Thabtah, S. Ahmadi, W. Hadi

MAC: A multiclass associative classification algorithm
J. Info. Know. Mgmt. (JIKM), 11 (2) (2012), pp. 125001-1-1250011-10
WorldScinet

2. N. Abdelhamid, F. Thabtah

Associative classification approaches: review and comparison
J. Inform. Knowl. Manage. (JIKM), 13 (3) (2014)

3. A.C. Arabzad, M.E.T. Araghi, S.N. Soheil

Football match results prediction using artificial neural networks: the case of Iran pro league
Int. J. Appl. Res. Ind. Eng., 1 (3) (2014), pp. 159-179

4. D. Buursma, Predicting sports events from past results “Towards effective betting on football matches”, in: Conference Paper, presented at 14th Twente Student Conference on IT, Twente, Holland, 21 January 2011, 2001.

5. C. Cao, Sports data mining technology used in basketball outcome prediction. Master’s Thesis, Dublin Institute of Technology, Ireland, 2012.

6. E. Davoodi, A. Khanteymoori

Horse racing prediction using artificial neural networks
Recent Adv. Neural Networks, Fuzzy Syst. Evol. Comput., 2010 (2010), pp. 155-160

7. D. Delen, D. Cogdell, N. Kasap

A comparative analysis of data mining methods in predicting NCAA bowl outcomes
Int. J. Forecast., 28 (2) (2012), pp. 543-552

8. J. Edelmann-Nusser, A. Hohmann, B. Henneberg

Modeling and prediction of competitive performance in swimming upon neural networks

9. Eur. J. Sport Sci., 2 (2) (2002), pp. 1-10

10. M. Fernandez, B. Ulmer, Predicting Soccer Match Results in the English Premier League, 2014.

11. Y. Ishikawa, I. Fujishiro, Tidegrapher: visual analytics of tactical situations for rugby matches, Visual Inform. 2 (2018), 60–70.

12. A.H. Eagly, W. Wood, Using research syntheses to plan future research, In: H. Cooper, L.V. Hedges (Eds.), The Handbook of Research Synthesis, Russell Sage Foundation, New York, 1994, pp. 485–500.

13. M. Dowling, B. Leopkey, L. Smith, Governance in sport: a scoping review, J. Sport Manage. 32 (2018), 438–451.