Sports Data Predictive Analysis Using Machine learning

Dipti Patnayak[1], Samprathi B E[2], C Anil Kumar[3], Vamshi K V[4], Naveen Kumar M N[5]
[1,2,3,4,5]*Computer Science and Engineering, M.S. Engineering College, Bangalore, Karnataka*

***Abstract*— A prediction system is critical for reducing curiosity about anything. Many sports prediction techniques are popular, and data analysis plays an important role in forecasting. Previous attempts in sports data analysis have resulted in the prediction of sports such as football, tennis next shot location prediction, Olympic athlete performance, basketball slam dunk shot frequency, and many more. Cricket prediction is tough due to the numerous variables that might affect the result or outcome of a cricket match. Previously, simple cricket match prediction systems focused just on the venue, ignoring aspects such as weather, stadium size, captaincy, and so on. Factors such as the match's location, pitch, weather conditions, first-pitch batting, and fielding all play a role in forecasting the match's outcome. To predict, suitable models are required, and data mining allows the required information to be extracted from data sets. This project is a review of techniques used for predicting the winners of three different games. The statistical data of the game can be exploited using various machine learning techniques to predict various information related to a particular match namely the result of a particular game, injury of a player, performance of a player in a particular match, spotting new talents in the game, etc. The aim is to accurately predict the winner of a particular game.**

I. INTRODUCTION

Cricket is the most popular sport in the world after football. Cricket is a popular sport in India, Pakistan, Australia, England, South Africa, and several other nations, with billions of supporters worldwide. "Cricket is my religion," many enthusiasts declare, especially in India. The game is played between two teams of 11 players each (15 if extra players are included). The majority of the game takes place on a 22-yard-long pitch. One Day International (ODI), Twenty Overs International (T20), and Test cricket are the three forms in which it is played. A bowler can make six consecutive legal deliveries, which are referred to as an "over." Initially, one team will bat first and the other will bowl, depending on the toss. The first inning will be completed. The side that batted first will bowl in the second innings, while the other team will bat. Each innings lasts 50 overs in One Day cricket. The test is played over five days, with a minimum of 90 overs being bowled on the first four days and a minimum of 75 overs being bowled on the final day. Because each innings is only 20 overs long, T20 is known as a restricted overs game. The pitch will be 22 yards long, with a 30-yard inner circle. based on a set of specified rules He can bat till the batter is out, or until the innings is over. When each batsman is dismissed, the bowling team receives a wicket. When a team loses 10 wickets or reaches the specified number of overs, the innings is declared over. To win, the second team must chase down the runs scored by the first team. The team that bats first wins if they defend their total.

Furthermore, Football is a family of team sports that involve, to varying degrees, kicking a ball to score a goal. A football match is played by two teams, with each allowed no more than 11 players on the field at any one time, one of whom is a goalkeeper. A match is played in two 45-minute halves. The game begins with the toss of a coin, and the winning captain decides which goal to defend or to take the first kick-off. All players must use their feet head or chest to play the ball. Only the goalkeeper is allowed to use their hands, and only within their designated goal area. The aim of the game is to score a goal, which is achieved by kicking or heading the ball into the opposing team's goal. If the ball touches or crosses the sideline, it is thrown back in by the team that was not the last to touch the ball.

Prediction system works on the principles of machine learning. There are two types of machines learning namely supervised machine learning and unsupervised machine learning. In supervised machine learning, we must train the machine by providing huge data sets and outcomes. In this project, one such prediction method is introduced which is used to make predictions of the outcome of all the matches using a machine learning algorithm.

II. EXISTING SYSTEM

Presently prediction methods are introduced which are used to make predictions of the outcome of a cricket match using Google Prediction API. Google API is a black box prediction technique. It is a form of supervised learning and hence it is required to provide huge data and train the models. Google Prediction APIs make use of Regression Algorithms when numerical predictions must be made. And Classifiers when the target output can assume only a limited set of values, either numbers or strings, based on the application content. This API can only account for relatable data. If the attributes are not related to one other, then a correct probability curve will not be drawn. By providing a CSV file of previous cricket matches and using appropriate queries to extract the required data and train the model, predictions can be made. This is a brief overview of the data set used.

Disadvantages of Existing System

•The user cannot control the exact features, and false features are identified.

•It imposes a large burden on the computation.

•The algorithms are have been trained, experimented and forced on single sports data.

III.PROPOSED SYSTEM

The main procedure of the work is to first, collect data from researching different available sources. Then construct a labeled data set from the historical sports data to classify the winner and loser. Extract a set of features from sports data and apply it to machine algorithms. Random forest that operates by constructing a multitude of decision trees will be used to make the prediction. K-nearest neighbors' algorithm is a nonparametric method proposed by Thomas cover used for classification and regression. Support vector machine (SVM) is a supervised machine learning algorithm used in classification problems. Training and testing the model and then evaluating the model by different parameters, delivering a conclusion by the study.

Advantages of Proposed System

•Considerably better accuracy compared to the previous models.

•Considerably less requirement of raw computation and Storage space requirements.

•Usage of machine learning model leads to less human intervention.

•Can process big data.

IV. OBJECTIVES

The main objectives of our project are:

• Classifying the winner of the match with great accuracy.

• Comparing different machine learning algorithms by their performance.

• Proposing an efficient way to classify the winner of the match.

V. LITERATURE SURVEY

1. Outcome Prediction of ODI Cricket Matches using MLP Networks Jalaz Kumar, Rajeev Kumar, Pushpender Kumar

Applications of machine learning supplemented with data mining techniques has become a hot topic for research worldwide, sports analytics is no exception though. Cricket is one of the most popular sports in Australia, Caribbean, UK and South Asian nations with a net fan base of around 2.5 billion. The game has tremendous spectator support in more than 100 nations and the masses show great interest in predicting the game outcomes. There are lots of pre-game and in-game attributes which decides the outcome of a cricket match. Pregame attributes like the venue, past track-records, innings(first/second), team strength etc. and the various in-game attributes like toss, run rate, wickets remaining, strike rate etc. influence the result of a match in a predominant manner. In this study, 2 different ML approaches namely Decision Trees and Multilayer Perceptron Network have been used to analyze the effect produced on the outcome of a cricket match due to these varied factors. Based on these results CricAI: Cricket Match Outcome Prediction System has been developed. The designed tool takes into consideration the pregame attributes like the ground, venue (home, away, neutral) and innings (first/second) for predicting the final result of given match.

2. The use of data mining for basketball matches outcomes prediction Dragan Miljković; Ljubiša Gajić; Aleksandar Kovačević; Zora Konjović

Sport result prediction is nowadays very popular among fans around the world, which is particularly contributed to the expansion of sports betting. This makes the problem of predicting the results of sporting events, a new and interesting challenge. Consequently, systems dealing with this problem are developed every day. This paper presents one such system, which uses data mining techniques in order to predict the outcomes of basketball games in NBA (National Basketball Association) league. The problem of predicting the game result is formalized as a classification problem, where the Naive Bayes method is used. Besides actual result, for each game system calculates the spread, by using multivariate linear regression. The MVC Model 2 pattern-based software system is implemented. The system was evaluated on the dataset comprising 778 games from the regular part of the 2009/2010 NBA season and it correctly predicted the winners of about 67% of matches.

3. Hybrid Basketball Game Outcome Prediction Model by Integrating Data Mining

Methods for the National Basketball Association Wei-Jen Chen, Mao-Jhen Jhou, Tian-Shyug Lee and Chi-Jie Lu The sports market has grown rapidly over the last several decades. Sports outcomes prediction is an attractive sports analytic challenge as it provides useful information for operations in the sports market. In this study, a hybrid basketball game outcomes prediction scheme is developed for predicting the final score of the National Basketball Association (NBA) games by integrating five data mining techniques, including extreme learning machine, multivariate adaptive regression splines, k-nearest neighbors, eXtreme gradient boosting (XGBoost), and stochastic gradient boosting. Designed features are generated by merging different game-lags information from fundamental basketball statistics and used in the proposed scheme.

4. Sports Data Mining Technology Used in Basketball Outcome Prediction Chenjie Cao Technological University Dublin Driven by the increasing comprehensive data in sports datasets and data mining technique successfully used in different area, sports data mining technique emerges and enables us to find hidden knowledge to impact the sport industry. In many instances, predicting the outcomes of sporting events has always been a challenging and attractive work and is therefore drawing a wide concern to conduct research in this field. This project focuses on using machine learning algorithms to build a model for predicting the NBA game outcomes and the algorithms involve Simple Logistics Classifier, Artificial Neural Networks, SVM and Naïve Bayes.

VI. SYSTEM ARCHITECTURE

System Architecture is an organized description that defines the structure, behavior, and the system views. The system architecture describes the major components, their relationships, structures and how they interact with each other. Software architecture and design includes several contributory factors such as Business strategy, quality attributes and many more. Software Architecture and Design can be classified into two phases - Software Architecture and Software Design. Software architecture refers to the fundamental structure of a software system where each structure comprises of elements of the software, the relationship between them and the properties of elements and the relations.

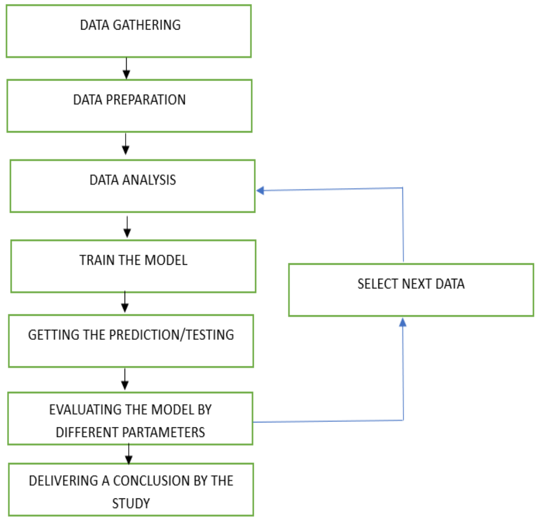

DATA GATHERING
DATA PREPARATION
DATA ANALYSIS
TRAIN THE MODEL
SELECT NEXT DATA
GETTING THE PREDICTION/TESTING
EVALUATING THE MODEL BY DIFFERENT PARTAMETERS
DELIVERING A CONCLUSION BY THE STUDY

Fig 6.1 System Architecture

VII. METHODOLOGY

• The main procedure of the work is to first, collect a data set from social networks. Then construct a labeled data set to predict the winner.
• Extract a set of important features from different data sets and apply it to machine algorithms.
• Random forests are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees.
• K-nearest neighbors' algorithm is a nonparametric method proposed by Thomas cover used for classification and regression.

- Support vector machine (SVM) is a supervised machine learning algorithm used in classification problems.
- Training and testing the model and then evaluating the model by different parameters, delivering a conclusion by the study.

VIII. IMPLEMENTATION

Implementation of system means the process of converting a new or revised system design into operational one.

Module Descriptions:

- Data gathering: Data gathering is defined as the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis on the basis of collected data. Extract a set of features from message content social behavior and apply it into machine algorithms.

- Data preparation: Data preparation is the process of collecting, cleaning, and consolidating data into one file or data table, primarily for use in analysis.

- Data analysis: Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains.

- Model training: Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range.

The main procedure of the work is to first, collect a data set from social networks. Then construct a labeled data set of users and manually classify into.

IX. DATA DIAGRAM

A Data flow diagram (DFD) is a way of representing a data flow of a system or a process graphically by representing the processes or functions which captures, manipulates, stores and distributes the data between a system and its environment or between the system components. DFD also provides information about the process and inputs and outputs of each entity. The visual representation of DFD makes it a good communication tool between the User & System designer. The DFD allows its structure to expand it to a hierarchy of detailed diagrams from a broad overview.

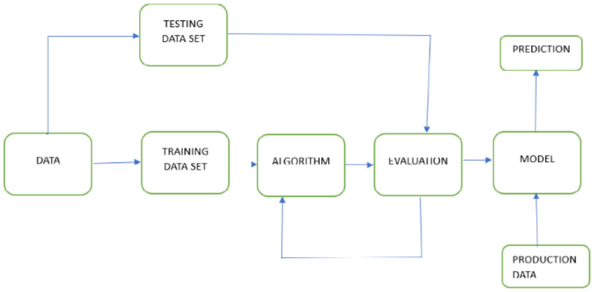


Fig.8.1 DFD

X. TESTING

Software testing is the process used to help identify the correctness, completeness, security and quality of developed computer software. With that in mind, testing can never completely establish the correctness of arbitrary computer software. In computability theory, a field of computer science, an elegant mathematical proof concludes that it is impossible to solve the halting problem, the question of whether an arbitrary program will enter an infinite loop, or halt and produce output. In other words, testing is criticism or comparison that is comparing the actual value with an expected one.

There are many approaches to software testing, but effective testing of complex products is essentially a process of investigation, not merely a matter of creating and following rote procedure. One definition of testing is “the process of questioning a product in order to evaluate it”, where the “questions” are things, the tester tries to do with the product, and the product answers with its behavior in reaction to the probing of the tester. Although most of the intellectual processes of testing are nearly identical to that of review or inspection, the word testing is connoted to mean the dynamic analysis of the product, putting the product through its paces. The quality of the application can, and normally does, vary widely from system to system but some of the common quality attributes includes reliability, stability, portability, maintainability and usability.

10.1 TESTING STRATEGIES

Designing effective test cases is important but so is the strategy to use them to execute them. If it is conducted in haphazard manner time is wasted and unnecessary effort is expended. Thus, it seems reasonable to establish a systematic strategy for testing software.

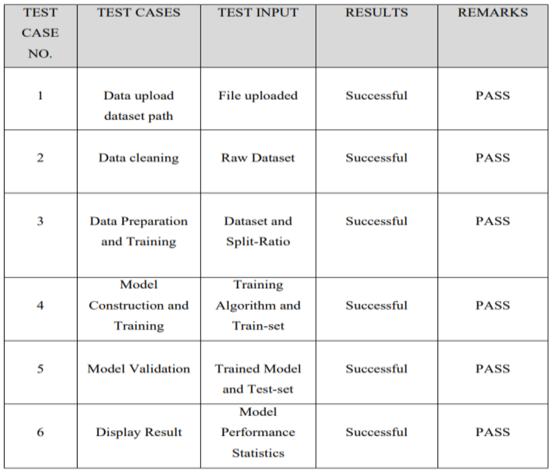
10.1.1 Unit Testing

Unit testing focuses verification effort on smallest unit of software designthe software component or module. The test that occurs as part of unit testing is given below: • The module interface is tested to ensure that the information flows into and out of the program and the test. • The local data structure is examined to ensure that data stored temporarily maintains its integrity during all steps in an algorithm's execution. • Boundary conditions are tested to ensure that the module operates properly at boundaries established to limit or restrict processing. • And finally, all error paths are tested. • Unit testing is normally considered an adjunct to coding step. After source level coding has been developed, reviewed and verified for correspondence to component level test case begins.

10.1.2 Functional Testing

At performed on hardware products to verify that the product functions exactly as designed. The general purpose of hardware functionality testing is to verify if the product performs as expected and documented, typically in technical or functional specifications. Developers creating a new product start from a functional specification, which describes the product's capabilities and limitations. This type of testing is beneficial to product developers who are creating a new product or an existing product which has undergone significant enhancements or changes in capabilities.

10.2 TEST CASE SPECIFICATION



TEST CASE NO.	TEST CASES	TEST INPUT	RESULTS	REMARKS
1	Data upload dataset path	File uploaded	Successful	PASS
2	Data cleaning	Raw Dataset	Successful	PASS
3	Data Preparation and Training	Dataset and Split-Ratio	Successful	PASS
4	Model Construction and Training	Training Algorithm and Train-set	Successful	PASS
5	Model Validation	Trained Model and Test-set	Successful	PASS
6	Display Result	Model Performance Statistics	Successful	PASS

Table 1 Test Case Specification

X. RESULT

This reviews three sports namely cricket, football, basketball using different machine learning algorithms. The comparing using algorithms have been analysed.

Cricket

Gaussian, Decision Tree, Support Vector Machine, Random Forest algorithms have been employed in the sport of cricket. Table 1: Illustrates the accuracy results of all 4 algorithms.

Algorithms	Accuracy
Gaussian	86.00
Decision Tree	91.00
SVM	56.99
Random Forest	87.00

Table 1 Accuracy Results

Football

Football related algorithms like logistic regression, Support Vector Machine, Random Forest have been implemented

Algorithms	Accuracy
Logistic Regression	65.0
SVM	54.0
Random Forest	64.0

Table 2 Accuracy Results

Basketball

Algorithms like Random Forest, KNN, Support Vector Classifier are executed. Table 3: Illustrates the accuracy results

Algorithms	Accuracy
Random Forest	49.00
KNN	54.05
SVM	64.86

Table 3 Accuracy Results

XI. CONCLUSION

Many works are being done in the field of prediction of sports matches. Analysis of sports data and foretelling the future is a hectic task. By the following project, we can derive an efficient way to predict the winner of the match. The study also provides a detailed review of the performance of algorithms on basis of several evaluation parameters. This study can be used for further studies. It is anticipated that the presented review will help researchers and the information on state-of-the-art match prediction techniques in a consolidated form. The future scope of this project will be to consider sentimental analysis to understand the mood of the players and combine sentimental analysis and statistical data to provide an even better prediction system.

REFERENCE

[1] Outcome Prediction of ODI Cricket Matches using MLP Networks Jalaz Kumar,
Rajeev Kumar, Pushpender Kumar.
[2] The use of data mining for basketball matches outcomes prediction Dragan
Miljković; Ljubiša Gajić; Aleksandar Kovačević; Zora Konjović.
[3] Hybrid Basketball Game Outcome Prediction Model by Integrating Data Mining
Methods for the National Basketball Association.
[4] Sports Data Mining Technology Used in Basketball Outcome Prediction.
[5] Using Machine Learning to Predict the Outcome of English County twenty over
Cricket Matches Stylianos Kampakis, University College London, William z
[6] Football Match Statistics Prediction using Artificial Neural Networks K. Sujatha, T. Godhavari and Nallamilli P G Bhavani