**Machine Learning in Sports**

Various machine learning algorithms have been applied and tested for their efficiency in solving the problems in sports. The relation between machine learning and games dates back to the initial days of artificial intelligence when Arthur Samuel, a pioneer in the field of gaming and artificial intelligence studied machine learning approaches using the game of checkers . A study was performed to predict the outcome (win, lose or draw) of football matches played by a professional English Premier League (EPL) team, Tottenham Hotspurs, based on matches that were played in the year between 1995 to 1997. It was observed that Bayesian networks relatively outperformed other machine learning algorithms which included MC4 - a decision tree learner, Naive Bayesian learner, Data-driven Bayesian, and K-nearest neighbor.

**The Proposed Work**

The literature survey concluded that there was a need for a machine learning model which could predict the outcome of an IPL match before the game begins. Among all formats of cricket, Twenty20 format sees a lot of turnarounds in the momentum of the game. An over can completely change a game. Hence, predicting an outcome for a Twenty20 game is quite a challenging task. Besides, developing a prediction model for a league which is wholly based on auction is another hurdle. IPL matches cannot be predicted simply by making use of statistics over historical data solely. Because of players going under auctions, the players are bound to change their teams; which is why the ongoing performance of every player must be taken into consideration while developing a prediction model. In sports, most of the prediction job is done using regression or classification tasks, both of which come under supervised learning.

In simple terms, y = f(x) is a prediction model which is learned by the learning algorithm from a set of dataset: D = ((X1, y1),(X2, y2),(X3, y3), ...(Xn, yn)). Based on the type of output (y) supervised learning is divided further into two categories, viz., regression, and classification. In Regression, the output is a continuous value; however, classification deals with discrete kind of output. For predicting continuous values, Linear Regression appeared to be quite effective, and for classification problems like predicting the outcome of matches or classifying players, learning algorithms like Naive Bayes, Logistic Regression, Neural Networks, Random Forests were found being used in most of the previous studies. In this work, the various factors that affect the outcome of a cricket match were analyzed, and it was observed that home team, away team, venue, toss winner, toss decision, home team weight, away team weight, influence the win probability of a team. The proposed prediction model makes use of multivariate.

**The Prediction Model Dataset**

The official website of Indian Premier League [35] was the primary source of data for this study. The data was scraped from the site and maintained in a Comma Separated Values (CSV) format. The initial dataset had many features including date, season, home team, away team, toss winner, man of the match, venue, umpires, referee, home team score, away team score, powerplay score, overs details when team reached milestone of multiple of 50 (i.e., 50 runs, 100 runs, 150runs), playing 11 players, winner and won by details. In a single season, a team has to play with other teams in two occasions, i.e., once as a home team and next time as an away team. For example, once KKR plays with CSK in its home stadium (Eden Gardens) next time they play against CSK in their home stadium (M Chinnaswamy Stadium). So, while making the dataset, the concept of home team and away team was considered to prevent the redundancy. Indian Premier League has just been 11 years old, which is why only 634 matches data were available after the pre-processing. This number is considerably less with comparison to the data available relating to the test or ODI formats. Due to certain difficulties with some ongoing team franchises, in some seasons the league has seen the participation of new teams, and some teams have discontinued. Presence of those inactive teams in the dataset was not really necessary, but if the matches data were omitted where the inactive teams appeared, the chances were that the valuable knowledge about the teams which were still active in the league would deteriorate. For better understanding and to make the dataset look somehow cluttered-free, acronyms were used for the teams. Table 1 lists the acronyms used in the dataset