

Machine Learning Approaches to Predict Basketball Outcomes

*Thesis submitted in partial fulfilment of the requirements for the
award of degree of*

Master of Engineering

In

Computer Science and Engineering

Submitted By

Harmandeep Kaur

(801532019)

Under the supervision of:

Dr. Sushma Jain

Assistant Professor



COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

JULY 2017

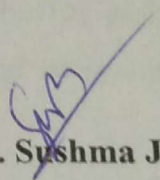
CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Machine Learning Approaches to Predict Basketball outcomes*", in partial fulfilment of the requirements for the award of the degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Sushma Jain* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for the award of any other degree of this or any other University.

Harmandeep Kaur
(Harmandeep Kaur)

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


(Dr. Sushma Jain)
Assistant Professor
Computer Science and Engineering
Department

Acknowledgement

The successful completion of any task would be incomplete without acknowledging the people who made it possible and whose constant guidance and encouragement secured the success.

First of all, I wish to acknowledge the benevolence of the omnipotent God who gave me strength and courage to overcome all obstacles and showed me the silver lining in the dark clouds. With the profound sense of gratitude and heartiest regard, I express my sincere feelings of indebtedness to my guide **Dr. Sushma Jain**, Assistant Professor, Computer Science and Engineering Department, Thapar University for her positive attitude, constant encouragement, keen interest, invaluable cooperation, generous attitude and above all her blessings. She has been a source of inspiration for me, I am grateful to **Dr. Maninder Singh**, Head of Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, Computer Science and Engineering Department, Thapar University for the motivation and inspiration for the completion of this thesis. I will be failing in my duty if I do not express my gratitude to **Dr. S. S. Bhatia**, Senior Professor and Dean of Academics Affairs in the University for making provisions of infrastructure such as library facilities, computer labs equipped with internet facility, immensely useful for the learners equip themselves with the latest in the field.

Later but not the least I would like to express my heartfelt thanks to my parents and my friends who with their thought-provoking views, veracity and whole-hearted co-operation helped me in doing this thesis.

Harmandeep Kaur
Harmandeep Kaur
(801532019)

Abstract

Sports prediction has always been a spellbinding research area for sports fans to know more about their favorite team and players, for teams and players to enhance their performance, for team managers and coaches to make strategies of the game and for a growing number of gamblers for making the predictions and betting on those predictions. Nowadays, companies are spending more effort in machine learning to predict the sports outcomes. The drastic increase in demand for sports advice, the presence of abundant data in sports and rapid growth of advanced technologies such as machine learning attracted a number of researchers for sports prediction. Support Vector Machines (SVMs) are powerful techniques that handle classification problems effectively and efficiently. However, SVM models lack in rule generation. So, this examination leads towards the development of Hybrid Fuzzy-SVM model (HFSVM) by integrating fuzzy approach and SVM technique for prediction of the basketball game outcomes that help the coaches, teams, and players to enhance their performance. The HFSVM model combines the advantage of both SVM technique and fuzzy approach, which is a unique strength of SVM and rule generation ability of fuzzy approach using fuzzy membership functions. In the proposed work the developed HFSVM model is applied to the data of 800 NBA games from 2015-2016 regular season to predict basketball game outcome. The basketball game is becoming more and more popular due to its high scoring and fast paced nature. Also, the HFSVM model is compared with SVM model and the empirical results showed that the HFSVM model not only provides better results than SVM model but also provides relatively satisfactory prediction accuracy. Therefore, promising results can be obtained using HFSVM model when analyzing the outcomes of basketball competitions.

Table of Content

| | |
|---|------------|
| CERTIFICATE..... | i |
| Acknowledgement..... | ii |
| Abstract..... | iii |
| Table of Content..... | iv |
| List of Figures..... | vi |
| List of Tables..... | vii |
| Chapter 1: Introduction..... | 1 |
| 1.1 National Basketball Association..... | 2 |
| 1.2 Basketball Game..... | 3 |
| 1.3 Game Outcome Prediction..... | 5 |
| Chapter 2: State of Art..... | 14 |
| 2.1 Statistical Approaches for Basketball Game Prediction..... | 19 |
| 2.2 Machine Learning Approaches for Basketball Game Prediction..... | 22 |
| 2.3 Fuzzy Approaches for Basketball Game Prediction..... | 27 |
| Chapter 3: Problem Statement..... | 31 |
| 3.1 Barriers in Previous Work..... | 31 |
| 3.2 Problem Statement..... | 32 |
| 3.3 Objectives of Proposed work..... | 32 |
| Chapter 4: Proposed Framework..... | 33 |
| 4.1 Data Pre-Processing..... | 35 |
| 4.2 Attribute Selection..... | 36 |
| 4.3 Classification..... | 36 |
| 4.3.1 Support Vector Machine (SVM)..... | 37 |

| | | |
|--|--------------------------------|-----------|
| 4.3.2 | Hybrid Fuzzy-SVM (HFSVM) | 42 |
| Chapter 5: Results and Discussions..... | | 50 |
| 5.1 | Data Used | 50 |
| 5.2 | Evaluation Parameters | 52 |
| 5.3 | Simulation and Results..... | 53 |
| Chapter 6: Conclusion and Future Scope..... | | 58 |
| 6.1 | Conclusion..... | 58 |
| 6.2 | Contribution Summary | 58 |
| 6.3 | Future Scope..... | 59 |
| References | | 60 |
| List of Publication | | 64 |
| Reflective Diary | | 65 |
| Video Presentation | | 68 |

List of Figures

| | |
|---|----|
| Figure 1.1: Types of Machine Learning Algorithm..... | 8 |
| Figure 1.2: Types of Machine Learning Problems | 10 |
| Figure 1.3: Applications of Machine Learning..... | 11 |
| Figure 4.1: Flowchart of Proposed Work | 34 |
| Figure 4.2: Fuzzy Logic System | 43 |
| Figure 4.3: Membership Degree of x Value | 44 |
| Figure 4.4: Union of Fuzzy Two Sets | 46 |
| Figure 4.5: Intersection of Fuzzy Two Sets | 46 |
| Figure 4.6: Compliment of Fuzzy Set..... | 46 |
| Figure 5.1: Attribute versus Corresponding Variable Importance..... | 53 |
| Figure 5.2: ROC and ROCH Plot (a) SVM and (b) HFSVM | 56 |
| Figure 5.3: The AUC w(c) Function, Corresponding to ROC Curve of (a) SVM and (b) HFSVM..... | 56 |
| Figure 5.4: The H measure w(c) Function versus Cost Curve of (a) SVM and (b) HFSVM .. | 57 |
| Figure 5.5: Smooth Score Distribution Curve of (a) SVM and (b) HFSVM..... | 57 |

List of Tables

| | |
|---|----|
| Table 4.1: Confusion Matrix | 48 |
| Table 5.1: Attributes of NBA..... | 50 |
| Table 5.2: Evaluation Parameters of SVM and HFSVM | 52 |
| Table 5.3: Attributes of NBA Selected from Boruta Algorithm | 53 |
| Table 5.4: Experimental Results of Testing Accuracy with SVM | 54 |
| Table 5.5: Experimental results of testing Accuracy with HFSVM | 55 |
| Table 5.6: Average Fivefold Cross-Validation Results of SVM and HFSVM | 55 |

Chapter 1

Introduction

Sports outcome prediction is a business that has popularly grown in recent years. Although, predicting sports game outcome is much difficult, but it has been considered for large and proximal attention for a sustainable time. The result of a sports game is not disclosed till the end of the match. Due to such unpredictability of the outcome of sports games, the excitement of sports competition increases.

Sports prediction is important for sports fans, coaches, media, sponsors and the growing number of gamblers. A huge amount of effort is yielded on predicting the playing event outcome. Due to increasing demand for professional advice regarding the sports event outcome, a variety of experts is involved in sports game prediction. The rapid advancement of high-performance computing devices and the presence of abundant data in sports attract more academic attention towards the quantitative study of professional sports. Moreover, the presence of abundant data regarding the sports event's outcome makes it possible to perform significant research about the sports prediction. Researchers predict the outcome of sports events through a variety of simulation models, mathematical formulas or quantitative analysis. Two important strands of sports prediction are: to obtain the factors that affect the game result and to learn how these factors can be changed so that profitable results can be obtained.

The basketball game is becoming more and more popular due to its high scoring and fast paced nature. Ranging from high school to the professional game level, the basketball game has attracted abundant of fans for live sports games along with television broadcast of its events like the National Basketball Association (NBA), National Collegiate the Athletic Association (NCAA) annual tournament and the Women's National Basketball Association (WNBA).

The NBA is the utmost level basketball league throughout the world. In addition to the popularity of NBA games, it is more professional, attended and marketed. The NBA has big following including experts anticipating results and abundant betting companies offering a large amount of money in gambling [1].

Although basketball game has gained large popularity, it has received less attention in prediction areas.

1.1 National Basketball Association

The National Basketball Association (NBA), the leading professional basketball league for men in the world established in 1946. The league was formed as the Basketball Association of America (BAA) in New York and it got its name after having merged with its rival National Basketball League (NBL). It is a USA Basketball's (USAB) active member, which is considered as the national governing body for basketball in the United States, by FIBA (also called International Basketball Federation). The NBA is among the four major professional sports leagues in the United States. The NBA has many followers around the world, with contenders predicting the result, in addition to numerous betting companies that are offering a large amount of money to the predictors. Professional basketball games are one of the most studied areas of many quantitative types of research. The NBA had 11 teams at its origin and having a series of team reductions, expansions and reallocations, it currently includes 30 teams. 29 teams are situated in the United States and one team is situated in Canada. On the basis of the population distribution of both United States and Canada, most of the teams are located in the eastern half of the country that is 13 teams are in the Eastern Time Zone, 9 teams are located in Central, 3 teams are in Mountain and 5 teams are in the Pacific.

The NBA played in Canada and USA is divided into two conferences that are a Western and Eastern conference. Every conference includes 3 divisions and the corresponding division includes 5 Teams each. The players having experience less than two years are assigned to the NBA development league. After training, pre-fall matches are held. The NBA regular season begins in October's last week. In NBA throughout the whole regular season, each team plays 82 games, 41 each home and away. There are 16 games that are scheduled within the division. A team plays with opponents of its own division four times a year. There are 36 games that scheduled within the conference, but out of division, that means each team plays with six teams from the other two divisions in its conference, four times in a regular season (24 games) and with the four other teams in the conference four times in a regular season (12 games). Finally, each team plays with all the teams of another conference, twice in a regular season that is there are 30 games that are scheduled against non-conference opponents. In this

way, the team plays with every other team in a regular season. In every season each team visits and hosts every other team at least once.

NBA playoffs start with eight teams from each conference in late April. The teams compete for a Championship. Three winners from each division and the team are selected for top four seeds on the basis of best record throughout the whole conference. Next four teams are given lower four seeds on the basis of next best record from the conference.

The tournament format is followed in the playoffs. Each team faces an opponent in a best-of-seven series. The team that is first to win four games is allowed to play in the next round and the other team is terminated from the playoffs. Next round is played with the successful team and another advancing team from the same conference that is the winning team from the series between the first and the eighth-seeded teams play with the winning team from the series between the fourth and the fifth-seeded teams and is the winning team from the series between the second and the seventh-seeded teams play with the winning team from the series between the third and the sixth-seeded teams. The best-of-seven series follows a 2 – 2 – 1 – 1 – 1 home court pattern in every round which means that one team plays on their home court in 1, 2, 3 and 7 games and the other team plays at home in 3, 4 and 6 games. The team with the best regular-season record in the league is assured with the advantage of home court in every series that it plays. The NBA Finals is the final playoff round with best-of-seven series played between the winners of both conferences that are held in June. The winner from NBA final wins the Larry O’Brein Championship Trophy. General Manager, coaches and each player of winning team receive a championship ring. Also, the player that is performing best is given the Bill Russell NBA Finals Most Valuable Player Award.

1.2 Basketball Game

Rules related with Basketball games are the ordinances and regulations that control the play, managing procedure of the basketball game. Although many of the fundamental rules are constant globally; however, variations always exist. In North America, many of the leagues or governing bodies such as the NBA and NCAA develop their own ordinances and regulations. In addition, the International federation’s technical commission decides the rules for international play. Most leagues outside North America follow the complete FIBA rule set.

The team's objective of the play is to make points greater than the differing team. Teams consist of fifteen players, with five players on the Basketball court at one time. They include two forwards, two guards and one center. The game begins with a tip off. The player after winning the possession of the ball has to shoot the ball towards the opposing basket within 24 seconds. The baskets are 10 feet above the ground. In the NBA the court is generally about 94 feet long by 50 feet wide and varies depending on where the player plays. The player can either pass the ball to a teammate or dribble the ball (player, bounce the ball up and down repeatedly when in motion) to move the ball up the court. The player must shoot the ball into the opponent's basket to score points. Players achieve two points for any shots scored within 3 point arc and 3 points from a shot outside this arc. The player scores 1 point for any free throws that are awarded to the team. The team blocks the shots, rebound a missed shot or steal the ball away from opposing team to get scores. In the NBA, the game is played in 4 quarters, each with 12 minutes. The international game is played in 4 quarters, each with 10 minutes. While in NCAA, the game is played in 2 quarters of 20 minutes each. At the end of time, the team with the highest score wins. If the scores are level at the end of game, overtime periods are put in action to decide the winner as there are no ties in basketball.

Violations generally occur when the player breaks any rule. The main violation includes short clock violation: a team player has 24 seconds to the ball on not doing so shot clock violation is called. The player is allowed to dribble the ball only once, but if the player dribbles the ball again, the violation occurs that is known as double dribble and the ball is passed to the opponent team. Travelling occurs when a player takes too many steps without dribbling the ball and another team is awarded the ball. Three key violations occur when the player stays in the key for more than 3 seconds. Charging violation occurs when an attacking player runs into a stationary defender. The defending team is given the possession of the ball.

Fouls come into scenario when a player does an illegitimate contact against another player. Whenever, a general foul is considered violent against another player, termed as Flagrant foul. Another team is awarded 2 free throws. Technical fouls can be awarded to the team for fighting, dishonest conduct, or misbehave from coaches and players against the referees. Two technical fouls mean ejection of the team automatically. A team with five or six individual fouls is fouled out and that team cannot participate in the rest of the game.

1.3 Game Outcome Prediction

The prediction of game outcome is one of the most studied areas due to the growing interest of fans, players, teams, coaches, and gamblers. Researchers use a variety of mathematical formulas, simulation models, and statistical approaches to predict the outcome of sports events.

Traditionally, researchers used different methods such as tipsters, betting odds and prediction markets. The betting odds and prediction markets yield comparable and better prediction accuracy than the tipsters that gives poor prediction accuracy. Interestingly, if the betting market charged a moderate fee then prediction markets could generate a profitable result from betting.

Machine learning provides better results in the form of accuracy. They overcome the disadvantages of statistical models by creating data-driven predictions or decisions using a model from sample input. They provide optimized results. In the proposed work, machine learning algorithms are used to predict the basketball game outcome. The data set is derived from the basketball game related sites. The data is pre-processed which means the missing values in the dataset are imputed and redundant data is removed. Important features are selected using machine learning algorithms. The models used for testing are built on training data set using machine learning techniques. Once the models are built they are used for predicting the basketball game outcome.

Machine learning is a form of data mining tool which generates specific algorithm in order to learn the problem and predict the results. Machine learning uses various techniques of data mining to solve the problem.

Data mining is a technique of discovering interesting patterns such as associations or patterns from the larger dataset and creating the relationships to solve the problem. Various Data Mining Techniques are listed below:

A. Data Pre-Processing

Data pre-processing is one of the data mining technique that transforms the raw data into a comprehensible format. Often, the raw data is not complete due to the presence of attribute missing values or lack of certain interesting attributes, noisy due to the presence of outliers and inconsistent due to certain discrepancies in the code. Data

pre-processing is used to resolving such issues. The various steps of data pre-processing are:

i. Data cleaning: Data is cleaned through the processes such as filling the missing values of the dataset, identifying outliers and smoothing the noisy data, correcting the inconsistencies in the data and removing the redundant data caused by data integration.

a) Filling missing values: Data is not always available because sometimes many tuples do not have any recorded value for certain attributes. Missing value can occur when there is malfunctioning in the equipment, inconsistent data are removed, data are not filled due to some misunderstanding, and data are not regarded as important at the entry time and when data schema is expanded.

Missing values can be handled by ignoring tuples, manually filling the missing values and filling attribute mean or a global constant automatically.

b) Removing noise data: Noise data can be due to the presence of random errors or some variance with respect to the measured value. Attribute values can be incorrect due to data collection by faulty instruments, problems at data entry or data transmission.

Noise data can be handled by a binning method in which the data is sorted and partitioned into equal sized bins where the data can be smoothed by bin mean, bin median or bin boundaries. Noise can be detected by clustering and then removed.

c) Removing outliers: Outliers are the data point inconsistent with the majority of the data. Outliers can be removed by clustering or curve-fitting.

ii. Data integration: Data integration is the process of combining the data from multiple sources and removing the conflicts between the data. It also involves schema integration in which metadata from different sources is integrated. Redundant and duplicate values are removed.

iii. Data transformation: Data is smoothed by removing the noise from the data. Data is normalized by scaling the data to fall within a small and specified range, aggregated and generalized. Data is smoothed by removing the noise

from the data. Data is normalized by scaling the data to fall within a small and specified range, aggregated and generalized.

iv. Data reduction: Data must be reduced when there are too many instances or features in the data. Data reduction is done to reduce the size of the dataset such that it produces the same analytical results. Data can be reduced by using clustering and aggregation for removing the redundant values, by using dimensionality reduction for removing unimportant attributes and by sampling.

v. Data discretization: It allows reducing the number of values for a continuous attribute by dividing the range of attributes into the interval. The actual data values can then be replaced by interval labels.

After pre-processing classification, clustering or prediction can be performed.

B. Classification

Classification is a data mining technique used to classify the data items into a predefined set of classes or groups. The mathematical techniques are used for classification such as neural network, linear programming or decision trees.

C. Clustering

Clustering is a data mining technique organizes the data items having similar characteristics into meaningful or useful clusters. The class labels are unknown and they are defined by clustering. The clustering technique puts the data objects into each class.

D. Prediction

It is a data mining technique that finds the relationship between the dependent and independent variables and the relationship between independent variables.

E. Sequential Pattern

It is a data mining technique that discovers a similar pattern, trends or regular events in transaction data over some business period.

Machine learning algorithms

Machine learning algorithms are of three types listed in Figure 1.1.

A. Supervised Algorithms

They are most commonly used machine learning algorithm. For any given input and output variables, the algorithm creates a model that predicts the output value for the corresponding input values. It is called “supervised” because the target variable is known. The model will learn again and again until it achieves an acceptable level. Supervised problems can be further categorized as:

- i. **Regression:** A supervised problem is said to be a regression problem if the output variable has a continuous value.
- ii. **Classification:** A supervised problem is said to be the classification problem when the output variable contains discrete (or category) value. Fraud detection is one of the examples of supervised classification.

Some of the supervised algorithms are decision trees, random forest, Support vector machine, neural networks.

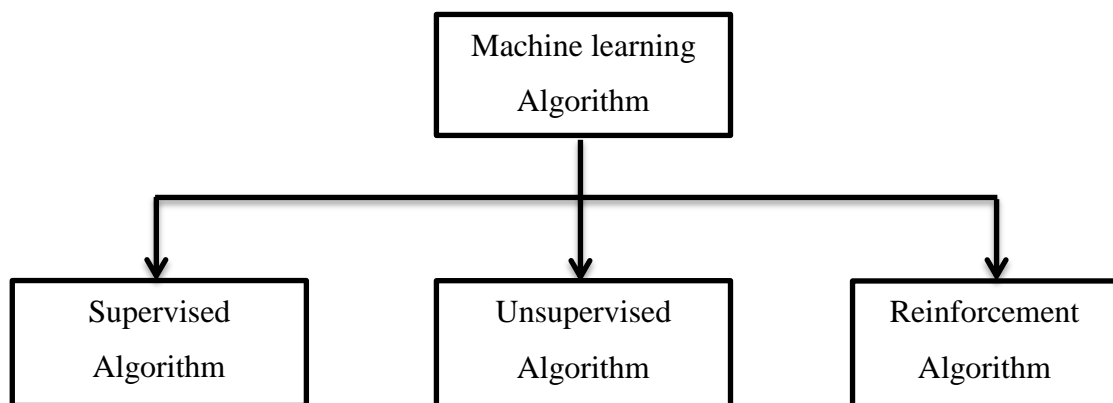


Figure 1.1: Types of Machine Learning Algorithm

B. Unsupervised Learning

In unsupervised machine learning algorithm, there is only input data available, but there is no corresponding output variable. The distribution in the dataset is modelled by the algorithm so as to learn about the data. It is known as “unsupervised” because the class labels are unknown and algorithms have to identify and present the structure

in the dataset. Unsupervised learning problems are divided into two groups, namely Cluster analysis, and Association.

- i. Cluster analysis:** A cluster analysis problem is an unsupervised problem where the built-in clusters or groupings in the data are discovered.
- ii. Association:** An association rule learning problem is an unsupervised problem where the presence of interesting relationships between variables in the dataset is discovered.

Some unsupervised clustering are Principal component analysis, K-means clustering and Apriori algorithm.

C. Reinforcement Learning

In reinforcement learning algorithm, an observation is given and the machine is trained to act or make specific decisions. It learns by interacting with an environment. The machine does not need to be explicitly taught, it learns from the repetitions of its actions. It is a trial-and-error learning in which, depending on the previous experiences and new choices, the machine chooses its actions. The machine learns from the actions and tries to get the best possible knowledge for accurate decision making. Markov Decision Process is a type of reinforcement learning algorithm.

Machine learning advantages

- 1) It is capable of handling multi-varied and multi-dimensional data in the uncertain or dynamic environment.
- 2) It allows efficient resources utilization and time cycle reduction.
- 3) Machine learning provides tools that continually lead to the quality improvement in complex and large process environments.
- 4) It easily consumes an unlimited amount of data along with timely analysis and assessment.
- 5) The machine learning system randomly initializes and trains the model on datasets that eventually learns good feature representation for the task.
- 6) The group of tuneable parameters can be visualized as a feature in parameter optimization, so the parameters can be learned same as feature learning.

Machine learning problems based on the number and types of classes

The range of the learning problems is very large and a growing number of templates to address the situations have been identified. These templates make the practical implementation of machine learning possible. The machine learning problems depend on the classes in target value. The list of templates is shown in Figure 1.2.

A. Binary Classification

Binary classification is the most commonly used classification in machine learning and it has led to the development of a large number of important algorithms and other theoretic developments. Given a pattern x from domain X , it is estimated that what value an associative binary random variable y will take. It can be +1 or -1. Different types of estimations can be made according to the protocol:

- i. Given a sequence of (x_i, y_i) pairs for which y_i is estimated through instantaneous online learning.
- ii. The value of X_0 can be known only at the time of learning called transduction.
- iii. For the purpose of model building, it is allowed to choose X and this is known as active learning.
- iv. The full information about X cannot be present that means some coordinates of x_i are not present, leading to the missing variable estimation problem.
- v. At the same time, there can be the observations stemming from two different problems having some related information, this is known as co-training.

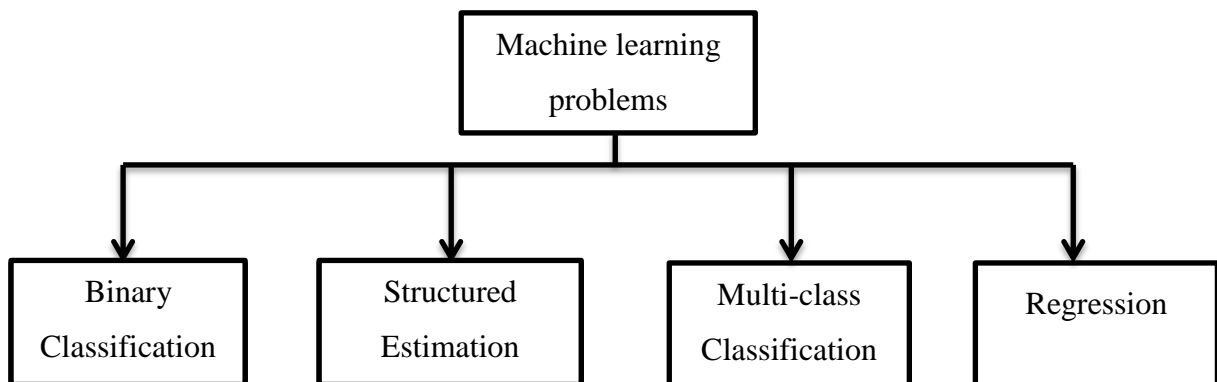


Figure 1.2: Types of Machine Learning Problems

B. Multiclass Classification

It is the logical extension of binary classification, where $y \in \{1, \dots, n\}$ assumes a range of different values. The main difference than binary classification is the type of error made decides the cost of error may. For example, in the problem of investing the risk of cancer, whether to miss-classify an early stage of cancer as being or an advanced stage of cancer makes a significant difference. The multi-class classification differs from the binary classification in the number of classes of target variables they contain.

C. Structured Estimation

In this, it is estimated that the labels y has some additional structure to be used in the estimation process. For example, y can be a path in ontology, on classifying webpages, y can be a permutation. Whereas, y can be an annotation of a text, on performing named entity recognition. In terms of the set of y , each of the problems has its own property.

D. Regression

Regression contains the real values for the variable y . Given, a number of instances, there is a need to obtain a function f to map the observations X to R so that $f(x)$ is near to the observed value.

Machine Learning Applications

The machine learning applications are listed in Figure 1.3.

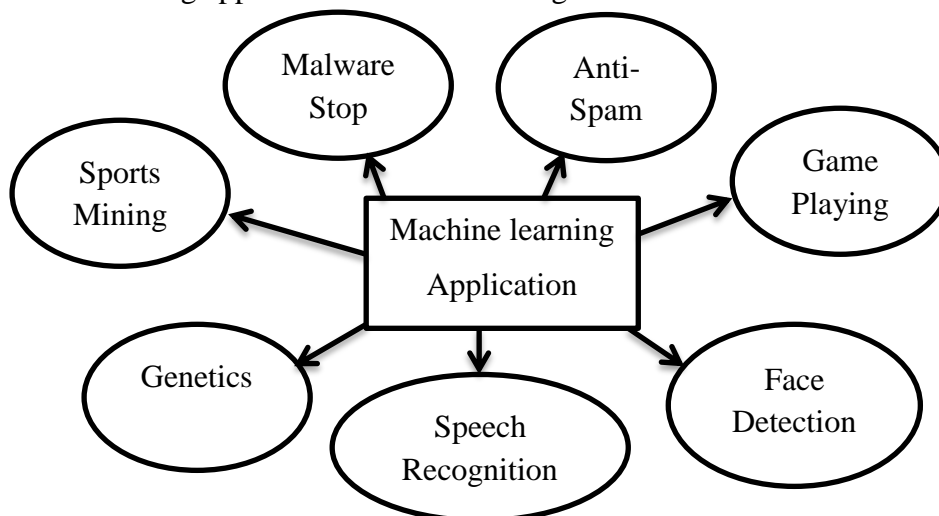


Figure 1.3: Applications of Machine Learning

A. Malware Stop/Anti-virus:

Due to the rapid increase in the number of malicious files, it is getting difficult for many security solutions and humans to keep up, and hence, machine learning is used for stopping malware or anti-virus. Machine learning helps in training anti-virus software to predict better.

B. Anti-spam:

Machine learning algorithms help the spam filtration algorithms to categorize spam emails from anti-spam mails. Machine learning uses the classification technique to categorize spam and anti-spam emails.

C. Game Playing:

There are two ways in which machine learning algorithms can be implemented in games, that is, during the design phase and during the runtime. In designing phase, the machine learning is applied before the game is rolled out. In runtime phase, machine learning is applied during runtime and it is fitted with a particular game session or a player. Forza Motorsports is a game where an artificial driver is trained depending on the player's own style.

D. Face Detection/Face Recognition:

Machine learning is used in mobile cameras or laptops for face detection or faces recognition. For example, cameras snap a picture automatically whenever a person smiles.

E. Speech Recognition:

Speech recognition systems are improving significantly due to of machine learning. For example, Google has developed speech recognition feature.

F. Genetics:

The genes associated with a particular disease can be found by using clustering algorithms of machine learning. For example, Medecision is a health management company that uses a machine learning platform to get a better understanding of patients who are at risk of diabetes.

G. Sports Mining:

Machine learning can be used in sports to generate the models that help in enhancing the performance of a player or team and predicting the game outcomes.

The remaining thesis is structured in the following way:

Chapter 2 discusses the state of the art which involves the research areas related to the basketball game prediction and discusses in detail the different techniques that are used to predict the basketball outcome. This chapter also discusses the important factors that affect the basketball game outcome. **Chapter 3** illustrates the problem statement and objectives of the proposed work. **Chapter 4** explains the framework of the proposed work and gives the insight knowledge of the work in detail. **Chapter 5** discusses the simulation parameters and compares the result obtained from both the models used in the proposed work and also compares the obtained results with existing approaches. **Chapter 6** discusses the conclusion, limitations of proposed work and its future scope.

Chapter 2

State of Art

From a long time, the aim of many researchers and gamblers is to precisely predict results of the sports in accordance with historical information. It has led to many sports-specific developments such as simulation models using statistical methods and machine learning models. Following are some researches related to sports analysis.

Schumaker et al. [2] concluded that Statistical simulations and machine learning techniques are essential for the development of modern sports. Statistical simulations include the prediction of new game outcome taking reference of historical data. The simulation model varies from the model that simulates an entire upcoming regular season data to predict the best chance of winning and the simulation models that find the weaknesses in motion and advises to correct them. Simulations can be analysed in the large domain of sports including basketball, football, baseball, and hockey. Besides simulations, machine learning approaches uncover hidden data trends. Although the algorithms used in each study differs but all of them have one similarity that they are able to beat the choices made by human experts and to create arbitrage opportunity by using the data.

Spann and Skiera [3] compared the prediction accuracy of different methods that are prediction markets, betting odds and tipsters as well as evaluate the potential of tipsters and production markets to systematically generate the profits in a betting market. In terms of prediction accuracy betting odds and prediction markets perform at an equal level. Both of them are better than tipsters. A weighting-based combination of the predictions of betting odds and production market gives slightly higher prediction accuracy, whereas rule-based combinations of predictions substantially improve prediction accuracy. On the betting market production market allows the punters to make more money if no fee is charged by betting company. *Forrest et al.* [4] employed a sample of football games compared the predictions based on published odds and predictions decided on statistical model as benchmark involving a large number of computable variables to compare the results. The experts' views represent publish odds that are increasingly effective over a period of five years. It was found that even in the case of financially pressured environment experts' subjective prediction is better than the prediction by statistical models. In contrast to previous research, *Song et al.* [5] had employed large samples of National Football League data. Also, the experts were either

professional football players or they consult to the coaches or players for the published information. Although the average success rates of system and experts are similar, however, the variation in success rates of experts is higher than among statistical systems. The worst records among experts are poorer than the worst records among statistical systems. It is an important factor to decide whether to use statistical systems or experts. It was suggested to use the combination of both of them. The results depict that both the models were outperformed by the final betting line.

Zhang et al. [6] presented a model which gains the interaction of Markov chains within a team. It is a two level influence model which learns the players' influence. The model is a Dynamic Bayesian Network (DBN) having a two-level structure: group-level and individual-level. Actions of each player are modelled by individual-level model while the actions of the team as a whole are modelled by group-level models. Experiments are performed on synthetically created multi-party meeting corpus and multi-player games, which gave effective results of the model.

Furnkranz [7] surveyed some of the previously published work related to machine learning in games. The survey shows the machine learning methods that are applied to different types of problems that arise in game playing. This approach helps the researcher to find the appropriate machine learning technique for solving their problems as well as to identify game playing domain areas that need to be further investigated. The paper involves techniques ranging from neural networks to decision trees in the games ranging from chess to poker. It can be concluded that research in game playing involves difficult and serious problems that need to be solved with machine learning techniques that are existing or yet to be developed.

Huilgol and Chhabra [8] described the popular data mining techniques and the way they are used for solving problems of sports domain. It discussed the application of decision trees, artificial neural network and fuzzy system. Sports require a higher degree of professionalism. Even a slight change in the variables that affect the match results can help the team or players to give a competitive edge to their opponents. The analysis provided by data mining techniques helps the players and teams to plan their strategies.

Soto Valero [9] compared the prediction accuracy of four different models applied for predicting outcomes (win or loss) for Major League Basketball (MLB) regular season. The research employed sabermetrics statistics to calculate team performance level and produced

30 datasets (one for each MLB game) using freely available data of ten years, in order to test the predictive model. Four models are applied on the reduced datasets and are evaluated by using 10-fold cross-validation. It was found that classification scheme is better than regression scheme and SVM resulted in best predictions with accuracy near to 60%. Feature selection methods used showed that the home team advantage is the most important predictor variable.

Delen et al. [10] compared three popular data mining techniques (namely decision trees, artificial neural networks and support vector machines) on eight years data of college football bowl games. To check the predictive ability of different procedures, both regression and classification models are developed. The results depict that the game's outcome is better predicted by classification models than regression models and out of the three different classification techniques, best results are produced by decision trees with accuracy better than 85% prediction accuracy on a 10-fold holdout sample, followed by prediction accuracy of support vector machines (79.51%) and neural networks (75%). A sensitivity analysis is conducted on trained prediction models to measure the comparative analysis of variables, through which inputting is done on the basis of the difference in modelling performance with or without including variables. The input variables on the top of variable importance list are: NCTW (non-conference team winning percentage), MARGOVIC (average margin of victory during the current season), HMWIN (Home Win Percentage), LAST7 (success in the last seven games of the season) and TOP25 (success against the top 25 teams during the season).

Demers [11] compared Relevance Vector Machine (RVM) approach with Support Vector Machine (SVM) algorithm on the dataset with 105 best-of-seven National Hockey League (NHL) playoff series (between 2008 and 2014 inclusively). Despite the RVM approach's potential, the SVM algorithm was proved to be superior. The probabilistic results of SVM are used to get playoff underachievers and overachievers. The traditional and advanced performance metrics are combined to form a new composite metrics in probabilistic SVM classifiers to provide additional predictive power. It was also found that the intangible or chance factors play important role in predicting NHL playoff outcomes.

Gu [12] described an expert system on a dataset of 1230 NHL regular season games in 2014-2015 to predict NHL game outcome (win or loss) of all the 89 post seasonal games before they actually happened. The expert system that is the Analytical Network Process (ANP) model includes data and judgments to predict the outcome. Important factors are filtered

through rank-sum tests. The significance of factors was confirmed by applying SVM to validate the regular season games whose outcomes are known. The tangible factors are incorporated with intangible factors (i.e., coach tactics, mental and physical states) into ANP model. The results showed 77.5% prediction accuracy with the combined factors.

Leung and Joseph [13] presented a sports data mining approach for discovering fascinating knowledge and predicting the winners as well as other results of college football bowl games. In place of applying the conventional approach that compared the two competing teams' statistics and projects the results, the approach used in the work predicts the results on the basis of the historical results of games. The approach makes the prediction depending on the amalgamation of four distinct measures applied on the historical results of the football game. The approach analyses historical game results and after analyzing statistics are kept in store in two data structures with a list contain every game that is played over a given time and a list containing all teams with their corresponding features from the season. The team lists are parsed and mapped with every point depicting a team such that distances between two points indicating their similarity. The competing teams are compared with other teams and the results of the similar teams are used for prediction. To achieve all the teams similar to a given team, the approach describes each team as a point in 4-dimensional space. Further, the four-dimensional space is represented as four different statistics that are: Pythagorean wins, turnover differential, RPI, and offensive strategy. The results show that the approach leads to relatively high accuracy in predicting the results.

Zeng and Li [14] concluded that the fuzzy logic and fuzzy set theory are suitable in developing knowledge-based systems for physical education tasks like an evaluation of different training approaches, the selection of athletes, the real-time monitoring of the game and the team ranking. For football team ranking the fuzzy set theory and fuzzy clustering analysis is applied. The ranking results are T7, T3, T1, T9, T10, T8, T11, T12, T2, T6, T5, and T4. The results show that the fuzzy logic method used for ranking is reliable even when there is a change in parameters in a certain range and the algorithm can be generalized for N number of teams, where N is an arbitrary positive integer. But when it becomes difficult to figure out that which teams have worse or better results then it will not rank the teams. The algorithm does not work on data having a large number of missing values.

Thakare et al. [15] explored the data mining techniques such as association rules and classification to investigate the players' performance for team selection. Data mining

techniques are divided into Predictive and Descriptive models. A descriptive model examines the existing properties of data and finds the relationships and pattern in data. The descriptive model involves clustering, summarization, sequence discovery and association rules. A Predictive model predicts the values of data using known outcomes that are found from different data. The predictive model involves regression, classification, Time Series Analysis, and Prediction. In this approach, association rules are used to get the performance of players and the classification algorithm is used to select best players for the final team. Quantitative association rules are used in the form of if/then statements. The rules contain numeric attributes with implicit ordering among values and equation for the rule has quantitative attributes on the left-hand side and one attribute that is categorical, on the right-hand side. Classification rules are the outcome of the classification process that predicts the selection of any player in the team.

Jian et al. [16] explored the principal of K-means clustering algorithm. Cluster analysis is one of the methods of data mining, which categorises the given data into distinct clusters on the basis of their similarities so that in the same type data have a high resemblance, different type data become high differential. As the algorithm is easy to converge into local minima and is sensitive to outliers so an isolated data point has the worst impact on the average value. Therefore, an improved clustering algorithm on the basis of ant colony optimization is presented.. It is inferred that the clustering outcomes achieved by the improved algorithm on the basis of ant colony optimization are more scientific, reasonable and fair.

Rajeswari [17] proposed a methodology where statistical analyses are utilized so as to analyze the performance of team and player effectively and provide the solution for player selection and recruitment in the football game based on the historical performance data. In order to explain each method properly, a number of scenarios are considered. A ranking method uses the data of performances of the players to rank the players of the given team on the basis of their previous performances. The statistical method is introduced to select the team on the basis of various tactical strategies depending on the preferences of the manager. The generation of Player Performance Score (PPS) allows the implementation of this concept for the task of player selection. Analyzing the historical data helps to compare the projections to the given target objective. This assists the managers to plan a strategy for the future games. The model identifies the weak areas of the player or team that would be required to improve. By using this information, the player that would be most suitable for filling the gap which is

observed by previous analysis, from the list of the players that is present for recruitment is identified. PPS and certain key factors are used by the ranking algorithm to get the rank for each player relative the players of the same group.

2.1 Statistical Approaches for Basketball Game Prediction

The statistical approach is a formalized way to estimate the reality and it can further optionally make predictions from the estimations. Statistical approaches can be used for estimation, testing or prediction.

Strumbelj and *Vracar* [18] applied possession-based Markov model to the predict progression of the NBA basketball game. The match was simulated using the model and the outcome prediction was produced. However, the model has some limitations related to what it can simulate and to deal with this issue non-homogenous model is required. The transition matrix of the model was calculated using NBA play-by-play data and summary statistics of the team. Both of these approaches as well as other forecasting approaches such as logistic regression model to infer home team wins probability from predictable variables, bookmaker odd and a latent strength rating method was evaluated. It was found that Markov model approach was better than other statistical approaches while providing more insight onto the basketball.

Vracar et al. [19] presented the methodology to three seasons of NBA game to generate simulation basketball match that is held between two different teams. Simulations involve a sequence of play-by-play in-game events at the team level. The simulations are created by using a disorganised sequence through a state space where in-game events represent a state. The state description is extended to seize the current context of the game's progression, which includes game time, the characteristics of opposing team and the point difference. Modeling the non-homogenous part improves the prediction results and generates the simulation results that better seize the dynamics of the basketball game's progression. The model is good at estimating the win probability of individual team. The model is better at predicting the prediction accuracy of the winner.

Harville [20] applied a modified least square system to the 1999-2000 basketball and 1999-2001 football seasons. The accuracy of prediction of outcomes for 93 postseason basketball games and 73 postseason football games was not diminished by the amendments and it was comparable to that with the betting line. The system with all the attributes except obtrusiveness can be made by applying statistical model with least squares. In each game, the

expected difference in score is generated as a distinction in team's consequences plus or minus a home field/ court advantage.

The researchers *Leake* [21], *Stern* [22] and *Stefani* [23] from different researches applied least squares to obtain the ratings for college basketball game and professional and college football game. It was applied on 2000, college basketball games, 3000 college football games and 1000 professional football games. The predictions were made from the ratings and results showed that the accuracy was 69% for college basketball games, 72% for college football games and 68% for professional football games. It was also found that the accurate prediction in college basketball, college football, and the professional basketball can be achieved by implementing least square on a digital computer.

Zak [24] estimated a production function for teams of professional basketball. The production function techniques are applied on the dataset of National Basketball Association season from 1976 to 1977 to verify the production frontier. The production frontier estimates provide information regarding the impact of various inputs applied production process and allows inter-team comparisons of average team efficiency and potential output. Richmond Techniques estimates each team's potential output. The inputs that give better result are field goal percentage, rebounding and free throw percentage. Other affecting inputs are turnovers and personal fouls. The same set of inputs contributes differently to different teams so the players can be evaluated by considering their contribution to output. So, by investigating team's potential and its efficiency performance can be evaluated and results of the teams can be ranked.

Chen and Fan [25] investigated the progress of score difference between home and away teams of professional basketball games by applying functional data analysis (FDA). Modelling score difference as latent intensity process using FDA has two main advantages that are among score change increments it allows for random dependent structure and statistical inferences involving FDA estimates do not suffer from inconsistency. It is found that the momentum in a basketball game is important to predict basketball game outcomes which are numerically characterized on the basis of FDA estimates. Home court advantage also plays important role in predicting basketball game outcomes.

Dezman et al. [26] empirically verified the expert system that is designated for effective organisation of basketball players to certain roles and/or position in the game. Data were

collected from ten basketball coaches that confirmed the overall enactment of 60 randomly selected basketball players (12 players per each position) from 12 Croatian 1st league teams of season 1998/99. Coaches estimated the performance of players on the basis of offense (12 variables) and defense (7 variables). Body's height measure was added to the defense group of variables. It was found that the players achieve their highest grades of overall performance at the primary positions in the game, they play. The great differences were determined between those playing at position 1 (point guards) and position 5 (centres). It was most difficult to evaluate the optimal position for the player at position 3 (small forwards), then for a player at position 2 (shooting guards) and last, for players at position 4 (power forwards). The players at these three positions are versatile and system's reliability is least when players are selected and oriented at these three positions. Convenient body height is the most important factor for the players to assume multiple roles and play multiple positions. The variable body height has the greatest influence on the orientation of players.

Cervone et al. [27] proposed a framework by using player-tracking data from the 2012-13 data to assign each moment of a possession a point value by computing the number of points the offense is expected to score up to the end of the possession which is termed as Expected Possession Value (EPV). The EPV derived metrics answers real basketball questions. EPV evaluates every decision generated during a basketball game whether it is to dribble, pass or shoot or it is allowing a multitude of new metrics and analyzing basketball to quantify value in the terms of points. EPV framework is congruent to the flowing nature of the sport. The player tracking data not only expands upon conventional approaches but when statistical modelling and computation are combined with them the new analysis are developed for answering the questions are generated since the beginning of basketball but yet not answered.

Ruiz and Cruz [28] modified the classical model used for soccer to predict the basketball outcomes. The classical model is classified to capture each National Collegiate Athletic Association (NCAA) conference's behavior and different approaches of teams and conferences. It is shown that the extension to soccer model provides better predictive performance from simulated bets on six different online betting houses. The model provides better results than the implicit probabilities of the betting houses and even provides higher benefits. Independent Poisson random variables are used to model outcomes at each game. The means of these variables rely upon the defense and attack coefficients of teams and conferences. The conference specific coefficients provide information of overall behavior of

each conference while team specific coefficients provide information about each team. However, different strategies of both conferences and teams are captured by vector valued coefficients. The model identifies weaker but undervalued teams.

Manners [29] re-evaluated the model of NBA basketball team strength on the data of eight seasons from 2006 to 2014. The model can be further extended for heteroscedasticity and for team related error variances and time variation in the strength of the team. With the introduction of dynamic state space model for the strength of the team, time variation can be introduced. Although the combination of regular seasons and paly-off games lead to slight improvement in the results, still it is typically tough to work efficiently than the betting market. Persistent-time variation is worst affected by injuries or trades. The offensive and defensive strengths in both static and dynamic setting did not lead to improvement. The error variation of the state-equation was considered as zero that implies a constant strength.

2.2 Machine Learning Approaches for Basketball Game Prediction

Statistical approaches are not an appropriate method for understanding issues in depth and for identifying the ways of solving the problem. They are also complex and time-consuming.

Therefore, to overcome this problem machine learning techniques can be used. Machine learning techniques are more powerful than statistical techniques and consume less time to solve the problem.

Cheng et al. [30] formalized NBA game outcome prediction problem as a classification problem and applied Maximum Entropy principle to build an NBA Maximum Entropy model (NBAME). This model fits discrete statistics for NBA game to predict NBA playoff outcomes using the model. The model was applied on the NBA playoffs from the 2007-08 seasons to the 2014-15 seasons. It was found that the difficulty in the prediction of the NBA playoff outcomes is due to many unforeseeable factors such as the presence of the injured player, the relative strengths of the team, players' attitude and operations that determine winner and loser by team's managers. The model use the each basic technical feature's mean, respectively, from the most recently, played six games for both sides from the starting of the game to predict the outcome of the upcoming game. The model predicts the winning team with accuracy 74.4% and outperforms other machine learning algorithms that predicted the winning team with a maximum prediction accuracy of 70.6%.

Dragan et al. [31] used data mining techniques to predict the basketball game outcomes in NBA (National Basketball Association) league. The problem for predicting the game outcomes is formalized as a classification problem by using Naive Bayes method. The system calculates the spread beside actual results by using multivariate linear regression. The system was evaluated on the NBA dataset involving 778 games from the 2009-10 season with the prediction accuracy of about 67%. It gives the predictions about match's winner that is home or visiting team and the spread value for the match.

Hermann and Ntoso [32] applied machine learning to sports so as to gain an edge over the average player. Basketball players' scores were predicted using naive Bayes with discretized state space as well as stochastic gradient descent and regression algorithm. By formulating the problem as a constraint satisfaction problem, a team of eight players was selected. Although regression optimizations were complex, they provided better prediction accuracy and gained the advantage of around 8% over DraftKings' regular users. The algorithm makes profits with a large enough volume of teams.

Markoski et al. [33] developed a solution known as BBFBR (Basketball Board for Basketball referees). The model used for this solution is a neural network which takes the movement of the ball on the court as an input vector and the output vector of the neural network involves the movement coordinates of the referees. The research describes the structure of these input and output vectors and the profits and flaws during neural network training. The solution allows calculation of most stable ways of the movement of the referee, but the movement of the ball in an action cannot consist of more than 15 key points. The solution can only be used for the educational purpose to train young basketball referees as well as enabling them to be aware of an action. The methods and software were also developed on the basis of the action and basketball referees' movement on the court so as to gain a better view of the action. BBFBR solution provided better results.

Wang and Zemel [34] focussed on the problem of classification of offensive strategy in the basketball game. Variants of the neural network are used to the dataset of the 2013-2014 season for testing the 2014-2015 season. Although the data used is limited, however, the methods still acquire the reasonable results and hold for representations learned from the previous year. The strategy with which the team plays and also understanding strategy of opposite team greatly affects the outcome of the game. The strategies include complex interactions among the players. The intrinsic factors that affect the strategy classification

problem are a diversity of the target classes and distinctiveness while the extrinsic factors are fouls, reactions to defense and consistency of executions. The variants of the neural network provided promising results. Simple gave a top-1 classification accuracy of more than 50% which indicates that decent model can also perform well on this problem, RNN model with more data representational understanding achieved 66% top-1 accuracy and on unseen examples, it achieved 80% top-3 accuracy. It can be concluded that the system can achieve good results by training on one season and testing on next season.

Ivankovic et al. [35] applied neural network on the data from 2005-06, 2006-07, 2007-08, 2008-09 and 2009-10 seasons of First B basketball league for men in Siberia. The total of 890 games was played during five seasons. Data were collected from each individual player in order to represent statistics for a whole team. Feed forward technique in neural networks was used to analyze these data, which is the popular technique to analyze nonlinear sports data. Due to dynamic nature of basketball game, a large number of events occur during a game and to provide analysis of them statisticians notes as many events as possible. So, the neural network is used to analyze sports events. It is found that defensive rebound and two-point shots under the hoop are important elements in basketball. In defense, after opponent's shot, it is important to catch a ball and preventing from next offense while in the offense, to be precise under the hoop is considered important. It can be concluded that the game under the hoop is essential for winning the game.

Ivankovic et al. [36] designed a framework for the automatic player position detection (APPD) in the basketball game. The videos broadcasted via television stations provide the images to detect the Court players. The view is from the only single camera at any point in time, which makes the process of detection much more difficult. The detection of player is done on the basis of non-oriented pictorial structures' mixture. The Support Vector Machine (SVM) algorithm is used in the detection of body parts. Both detection results are combined together with their location constraints. The latent form of SVM known as the latent SVM (LSVM) is used to train the whole model. Instead of detecting players the algorithm detects some other false positive objects which are mostly the referees or people from the audience. On the basis of histogram comparison and the playing court boundaries, these false results can be corrected. K-means clustering is used to separate the players in different teams. A spatial transformation is determined by the actual court measures and the detected play court boundaries. The court boundaries are detected using the canny edge detection algorithm. The

points that represent detected players' location in images collected through TV video are mapped to the players' actual location on the court. The approach generated the accuracy of about 82 % which is one of the best results in the framework of basketball player detection.

Jackie and Lu [37] applied a machine learning technique, Support Vector Machine (SVM) for the prediction of the National Basketball (NBA) playoff results. The attributes include the historical statistics of regular seasons. The samples that did not occur in the history are also included in the form of pseudo-labels. The predicting accuracy of one game can be improved, but playoff involves only fifteen game series with the single-elimination system. So, every game of playoff plays important role in prediction. It is found that even the champion of the year can be thumped home at the beginning of the game. By using 10-year records, the SVM classifier provides the prediction accuracy of 55%.

Markoski et al. [38] used data mining techniques to predict NBA game outcome. The collection and management of the data are fully automated. A cloud solution maintains the budget of the project. The basketball game is analyzed using AdaBoost algorithm which includes a linear combination of weak classifiers that is decision tree with only one level. This algorithm is commonly used for recognizing face and body parts. The AdaBoost algorithm is examined on the basis of its capability for a video footage from a single moving camera. The images of the entire body of a basketball player are used for first training while the images of a head and torso are used for second training. The algorithm applied to the set of images that involve head and torso provides an accuracy of 70.5%. Due to the presence of a large amount of background representing the noise, training process on the set of entire body image failed. Therefore, AdaBoost cannot be used in player detection, but when applied on simpler objects (like face recognition) it detects them successful.

Backler et al. [39] applied different machine learning methods for predicting game outcomes to guide for decision making in professional basketball, identifying outstanding players to support betting and sponsorship and choosing the optimal position of the player. The algorithms used for game outcome prediction are: i) Linear Regression, ii) Logistic Regression, iii) Support Vector Machines, and iv) Artificial Neural Networks. The best accuracy is achieved by using linear regression methods with 73% accuracy. It can be concluded that simple approaches perform well in the specific problem domain. For decision making the algorithms used are: i) Outlier Detection for identifying outstanding players and ii) K-means Clustering for predicting optimal player positions with 75% accuracy. The

outlier detection methods were able to find out the majority of all-star players in the history of NBA.

Shi et al. [40] have found that most of the NCAAB (National College Athletics Association Basketball) predictions are performed by using statistical techniques. But trusting on capabilities of machine learning techniques the importance of features is uncovered and their relationships are learned. It is found that using additional attributes can give worst results and attributes are more important models and for obtaining predictive quality there is an upper limit that should be considered. Although complex models tease out relationships that the simpler models would miss, it is shown that naive Bayes performed well.

Shah and Romijnders [41] applied a neural network to predict the success of three-pointer shot. The neural network model learns the trajectory of a ball without the knowledge of physics. The model is compared with a baseline static machine learning model which in addition to the positional data also involves a full set of features like velocity and angle. Using the dataset with more than 20,000 three pointers from NBA basketball game, the models that were applied on sequential positional data outperform the machine learning model that depends on feature rich statics. The RNN obtained an AUC of 0.843 predicting a make or miss of the shot for the ball that is 8 feet away from the basket which outperforms the traditional approaches having an AUC of 0.719 and 0.558 for a gradient boosted machines and a general linear model, respectively. It can be concluded that RNNs have a better understanding of sequential data. It is found that deep learning model improves the prediction accuracy than traditional machine learning models applied on feature-based data.

Perse [42] proposed a trajectory-based approach to determine complex multi-player behavior in the basketball game. The play-trajectory data is segmented by considering only the average position of all players in the team into game phases such as offense, defense and time out, by using probabilistic play model. Every phase is modelled by using Gaussian mixture model having two components. One component represents a transition to a particular phase and another component represents the behavior during that phase. A method to automatically recognize specific basketball activity is presented. In this method, the key elements of a basketball game are detected in order to analyze the activity of the team in more detail. According to basketball theory, the key elements such as starting formation, screen and move are considered as building blocks of basketball play. The temporal order of key elements produces a semantic description of the observed activity. The consistent behavior of this

approach is demonstrated on the dataset of 71 repetitions of three types of basketball offense. At last, the activity is identified after comparing semantic description with the manually defined template's description. The approach is not sufficient to study the activities of opposing team as the template covering all the interesting behavior is not available. The approach used on each trajectory segment provides us the clustering of the activities that represent the common team behavior.

Loeffelholz et al. [43] applied neural networks as a tool to predict basketball teams' success in National Basketball Association (NBA). A dataset with 620 games was gathered and used for training different neural networks such as radial basis, feed forward, generalized regression and radial basis neural network. Fusion of neural network is also evaluated using probabilistic neural network and Bayes belief networks fusion. It was also investigated that which are the better subset of features used for prediction. The obtained results were compared to predictions generated by experts in the field of basketball. The results show that the average prediction accuracy of the best networks is 74.33% and the prediction accuracy of the experts is 68.67%. It is found that the neural networks have the ability to use common box score statistics for accurately classifying the outcome of the game that is un-played. The models show that using feed forward neural network with only four variables, they can obtain average prediction accuracy up to a 5.66% better or for the validation set 1 they can obtain up to a 13.33% better accuracy.

2.3 Fuzzy Approaches for Basketball Game Prediction

The fuzzy approach is a powerful technique in the decision-making process. It can be embedded with any model due to its flexibility. Fuzzy approaches are better in rule generation. So, they can be used with machine learning models that lack in rule generation, to create better results than individual machine learning models.

Papic et al. [44] presented a fuzzy expert system for evaluating and scouting young sports talent. On the basis of knowledge of the human experts in the sports field several morphological characteristic measurements, functional tests and motoric skill tests are quantized with respect to their importance for a given set of sports. The stored knowledge in the knowledge base is derived by 97 experts. The expert system is fully web oriented which means it is developed by using ASP.NET applications. The system gives the prediction about the person's acceptability and the kind of sport that is best suitable for the person that is tested. The outcomes of the system were tested by 4 experts by using the real data collected.

The development of expert system leads to the gaining of top results in sports, improving the efficiency of finances related to sports and reducing the frustration during the poor performance. The sports like basketball and athletics need to be separated into new entities depending upon specialization in athletics and player's position in basketball.

Balli and Korukoglu [45] developed a decision support framework to select the candidates that are eligible to become basketball player by using Fuzzy Multi-Attribute Decision Making (MADM) algorithm. The attributes are selected by the experts. The model was applied to the seven junior players of basketball between age 7 and 14, from Youth and Sports Center of Mugla, Turkey. The model depends on the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) methods and Fuzzy Analytic Hierarchy Process (FAHP). The selection process often involves linguistic variables and imprecise data. FAHP, TOPSIS, and MADM approach overcome the problem. FAHP is applied to determine the weight criteria and the observational values of technical skills. The measured values of physical fitness are converted to fuzzy values with the help of fuzzy set approach. TOPSIS method ranks the candidate player. It is found that the selection process of the players based on scientific methods motivates the players and increases the efficiency of training provided by the coach.

Trawinski [46] presented a primary method for creating a fuzzy model to predict the basketball game outcomes. By using the KEEL (Knowledge Extraction based on Evolutionary Learning) system 10 fuzzy rule learning algorithms are organized, generated and analyzed with a standard linear regression model. Feature selection algorithms are executed and a majority voting method is used to get the relevant features. The experiment was performed with four fuzzy models that include ad-hoc data-driven approach and genetic fuzzy systems as presented in KEEL system.

Necati and Ermis [47] introduced a new MVP (Most Valuable Player) method to determine the most valuable player of the basketball team. MVP determinations method, when used for important leagues such as NBA or Euroleague, takes into account the statistics such as assists, points or rebounds of the players. The purpose of MVP determination method is to assess the player's contribution to the team in the best possible way. The method ignores the statistics such as rates or percentages of the team. It includes both individual percentages of the player as well as players' percentages within the team. Instead of using a stable method, a realistic calculation method is used that changes from game to game because the calculations that were made with category coefficients as well as the players' performance changes from

game to game. So, fuzzy clustering analyzing was involved for calculation and determination for this method. It is found that the fuzzy clustering approach can be applied as an alternative method for the MVP calculations.

Mourad and Zhang [48] developed an intelligent software agent that assists NBA scouting agent by using web sources. The software agent only retrieves the information that is relevant to NBA players from web sources for assisting scouting agents but also it extracts metadata from the information. The metadata consists of player static prediction and player performance evaluation. The software agent is improved by equipping it with some additional tool such as its capability of anticipating its client's request and suggesting players for a given team. With the generalization of the XML technologies like SOAP, WSDL and UDDI and the application servers' maturity, more resources are available to get the information they need. The Neuro-Fuzzy network is used for the prediction purpose. The genetic algorithm is used for optimization to predict more accurate prediction. The genetic algorithm is slower than the Neuro-Fuzzy Network.

Pai et al. [49] developed a hybrid model by combining the SVM technique and the decision tree approach (HSVMDT) for the prediction of basketball game outcome and to help the coaches planning the strategies and players to enhance their performances. Although Support Vector Machines (SVMs) are powerful technique when dealing with classification problem they lack in rule generation. So, this leads to the development of HSVMDT model. The model comprises of the unique strength of SVM and rule generation property of decision tree. The model allows both forward and backward reasoning algorithms. The forward reasoning is used to predict basketball game outcomes and the backward reasoning provides the suggestions to coaches to adjust the play strategy so as to win.

Wang [50] concluded that the essential factor to affect the game outcome was free throw percentage. It was also found that defense is more essential than offense.

From the literature survey, it can be concluded that Machine learning algorithms overcome the disadvantages of statistical models by creating data-driven predictions or decisions using a model from sample input. It has strong bonds to mathematical optimization. Plenty of companies are investing heavily in machine learning to predict the sports results. But this technique has not been extensively used in the basketball game analysis. It is also found that Support Vector Machines are one of the most powerful classification techniques of machine

learning [51], [52]. Although the SVM technique gives an effective result for classification problem, this approach has the main disadvantage that it is incapable of yielding rules for decision-making [53].

This disadvantage can be overcome by using Hybrid Fuzzy-SVM (HFSVM) model for rule generation. Nowadays, a large number of applications are using the SVM Technique. However, in many applications, some of the input points may not be exactly appointed to one of the given classes. So, they need to be assigned to one of the given class, for SVM to perform classification of these points more correctly. In this paper, a fuzzy membership function is applied to each of the input points of SVM. These input points make different contributions to the decision surface learning. Thus, it enhances the SVM to subtract the influence of noises and outliers in data inputs which directly reduces the net error effect.

Chapter 3

Problem Statement

3.1 Barriers in Previous Work

With the growing advancement of the internet, the quantifiable data is also growing. Due to the presence of abundant data, a number of people are devoted to exploring the meaningful information from the data.

Although the database can handle hundreds of millions of data, still there is a need for some technique to understand and analyze the data in order to get meaningful information from them. Traditionally, experts use statistical techniques to compare and filter the database knowledge in order to extract the rules for getting the meaningful information. But today's business needs are not satisfied by these statistical techniques. Transforming the historical data into useful information depends on the coaches and managers. But this could not lead to the complete development of information. Also, there is always a limitation of experts and leaders that lead to the problem of collection of information. A large amount of data cannot be handled by the statistical techniques. Most of the statistical techniques ignore the uncertainties in the game while game predictions work on these uncertainties to predict the outcomes. They focus on rules but predictions involve a lot of factors that influence the outcome. These problems can be overcome by using machine learning techniques. It can handle a large amount of data as well as extract the meaningful information. Now, Machine learning techniques are used in different areas.

Although Machine Learning is a powerful technique that handles the prediction problem efficiently still there are many machine learning algorithms that lack in generating rules for decision making. In these algorithms, some of the input points may not be exactly appointed to one of the given classes. Also, the presence of the outliers severely affects the outcome of the prediction. There is the need of the technique that can yield the rules required for decision making, to properly classify the input points and to reduce the effect of outliers on the prediction.

Also, the research area lacks in the study of machine learning in the basketball game for predicting the game outcome.

3.2 Problem Statement

Although basketball is one of the popular games due to its dynamic and high scoring nature, it lacks in the attention of researchers. Therefore, the proposed work focuses on the prediction of the basketball game outcome.

Nowadays, large numbers of machine learning techniques are applied to the prediction problem for decision making. One amongst them is Support Vector Machine (SVM), which is one of the powerful techniques used for classification. It lacks in the generation of the rules required for decision making. Therefore, in the proposed work Hybrid Fuzzy SVM (HFSVM) model is used for rule generation in which fuzzy approach is integrated with SVM. The model is successful in reducing the net effect of outliers as well as in properly classifying the given points.

Thus, using machine learning approaches and fuzzy approach as a tool for carrying out NBA game outcome prediction is the major research problem of the work.

3.3 Objectives of Proposed work

Keeping in mind the proposed research problem and various challenges that are posed in previous researches, the objectives of the proposed work can be outlined as follows:

- To gather dataset and pre-process the dataset for the experiment.
- To generate the rules using fuzzy approach for the machine learning models.
- To build the SVM and HFSVM model for classification and for making predictions.
- Analyze and compare the results of SVM and HFSVM model.

Chapter 4

Proposed Framework

The flow diagram of proposed framework is shown in Figure 4.1. The first step is the collection of raw data from NBA websites such as “NBA.com”, “Basketball-reference.com”. The data set contains some missing values. The second step is data pre-processing. In this step, data segregation is performed according to their data types and after that, the missing values for the numeric data are imputed by using caret algorithm. The segregated data are then combined together into a complete dataset. The third step is the feature selection process. Feature selection is a crucial step before the classification process so as the features that are unimportant are deleted from the original dataset. This reduces computational complexity and increases accuracy. In the proposed model, boruta algorithm is applied to select the important condition attributes. The attributes with their variable importance higher than shadow variable importance can be selected while the attributes with their variable importance lower than shadow variable importance can be rejected. The attributes that are tentative can be selected by adjusting the number of iterations. The data is normalized to avoid the variation in the range of all features. The third part of the work is to classify the data. Classification plays an important role in predicting the results of the basketball game. Here, classification is done by using two different ways that are using an SVM model and HFSVM model. In SVM, firstly the processed data is loaded into the model. The data is partitioned into a training dataset and testing dataset. The target variable and input variables are set. The model is built by using radial basis kernel function. The training data set was used to model SVM with essential attributes and classification performance of the trained SVM model is evaluated by employing testing dataset on a model that is trained. The model is evaluated on the basis of evaluation parameters: confusion matrix, accuracy, the time taken, sensitivity, specificity, precision and recall and results are represented in the form of plot. A fivefold cross-validation is carried out to obtain average accuracy. Finally, the results are saved. Thus, the well-trained SVM model can be used for forecasting the competition result that can be used by coaches or players to increase performance in the game.

In HFSVM model the input data is fuzzified into the linguistic variables and their corresponding membership function is calculated that gives the degree to which the input value belongs to the fuzzy set. The rules generated using fuzzy approaches are evaluated and

the aggregation approach is performed in which outputs of all the rules are unified. The rules are then defuzzified using centroid approach to obtain crisp data. This dataset is then used as the input to build the SVM model. This is how the fuzzy approach is integrated with SVM technique. The rest of the process to obtain the basketball game results is same as followed in the SVM model. At last, both models are compared through their prediction results. The methods used in steps of the flowchart given in Figure 4.1. are explained as:

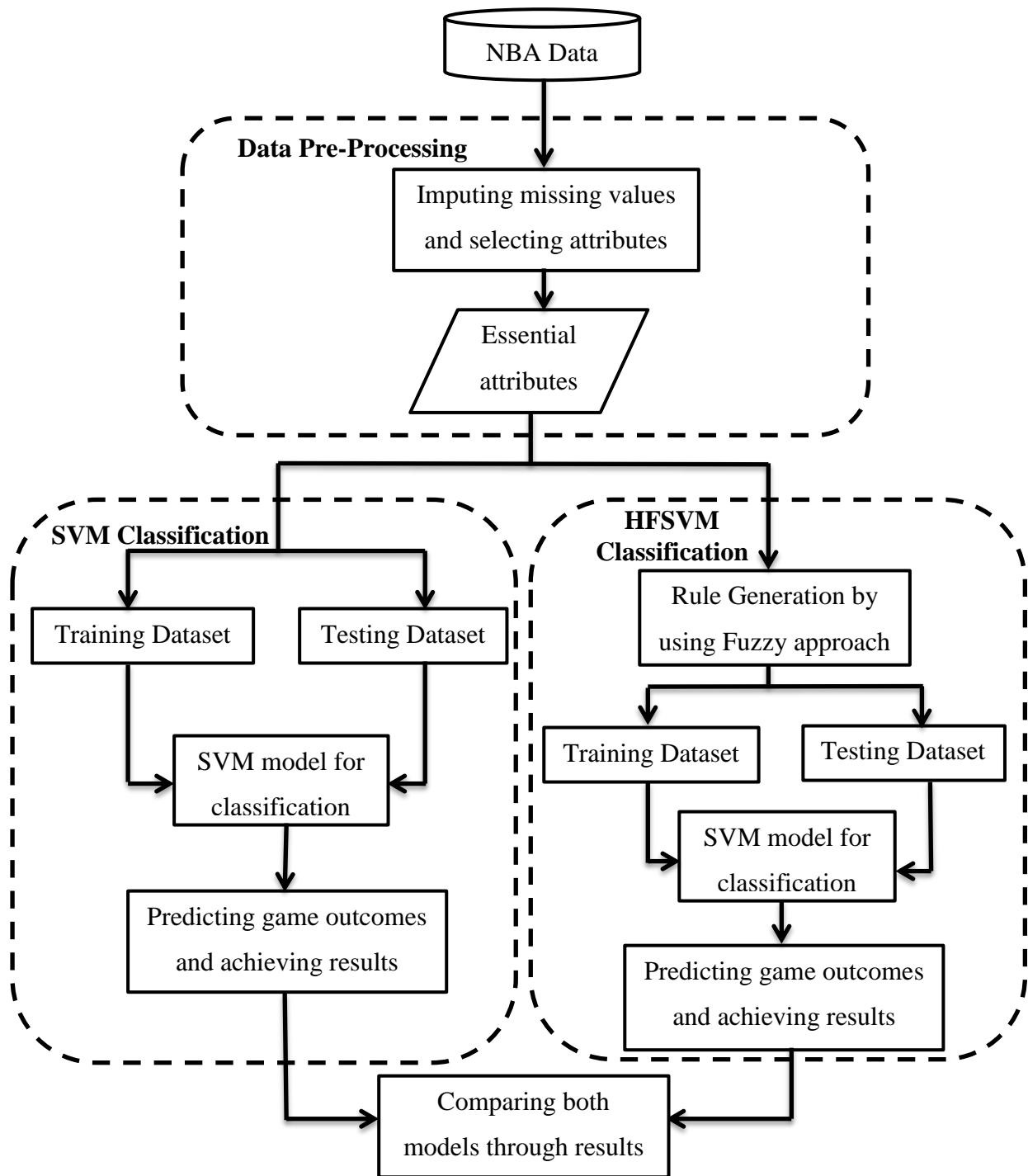


Figure 4.1: Flowchart of Proposed Work

4.1 Data Pre-Processing

The data pre-processing is a technique to remove the redundant values, impute missing values, removing correlated attributes and splitting as well as normalizing the data set.

This segment in the flowchart utilizes caret algorithm (short for classification and regression training) which includes a set of functions that to make the process of creating predictive models more efficient and effective. This algorithm contains tools for functionalities such as:

A. Data Splitting

Data splitting is done to select the random sample for the purpose of analysis. Data splitting can be performed in three different ways:

- i. Simple splitting based on the outcome:** It creates balanced splits of data based on the outcome. It uses `createDataPartition` function that creates balanced splits of the data. After splitting the data the function selects a random sample of data for analysis.
- ii. Splitting based on the predictors:** The function `maxDissim` is used to generate sub-samples by using a maximum dissimilar approach. For dataset X with n sample and a larger dataset Y with m samples, a sub-sample can be created from Y that must be diverse on comparing with X. For this, for each Y's sample, the function `maxDissim` calculates the n dissimilarities between each point in X. The most dissimilar point in Y is added to X and the process continuous.
- iii. Data splitting for time series:** The problem where training and testing data set moves with time `createTimeSlices` can be used to create the indices for such type of splitting.

B. Imputing Missing Values

It includes imputation, centering/scaling data, removing correlated predictors. Caret package segregates the attributes in the data set according to their data types with the help of `unique` function. Then the `preProcess` function is used to impute the missing value for the data set with numeric data type with the help of a `medianImpute` method. In this process, it estimates the parameters that are required for each operation and applies them to a specific dataset. It imputes the dataset based only on information about training dataset. The `preProcess` function scales the data set between zero and one. At last, the data set is combined using a function such as `cbind` and then it is

passed to the feature selection process. The Caret package allows removing of the correlated predictors by using a findCorrelation function that analyses a correlation matrix of data's attributes report that contains the attributes to be removed.

C. Variable Importance Estimation

In caret, the importance of features is estimated from the dataset by building a model. The model like decision trees has a built in mechanism that provides variable importance. Caret can also use some feature selection methods such as wrapper method and filter method.

4.2 Attribute Selection

Attribute selection methods are used to create models with different subsets of a dataset and determine the attributes that are necessary to build an accurate model.

In this segment, Boruta algorithm is used to select relevant features. It can work with any classification method that gives variable important measures as an output , but by default, boruta uses Random Forest. For obtaining the relevant features the method performs a top-down search. In each iteration, Boruta compares Z-Score of a condition attribute with Z-Score of shadow attribute that is created by rearranging original ones. Attributes having significant worst importance than shadow attributes are consecutively eliminated and attributes with significantly better importance are accepted. An algorithm running in default light mode drops unimportant attributes along with their random shadows. On the other hand, in the force mode, all shadow attributes are preserved until the termination of iterations.

The attributes that are important can be derived by passing boruta results as the argument to the getSelectedAttributes function. At last, a graphical representation of the results can be obtained by using plot function with boruta results as the argument.

The algorithm stops on two conditions:

- i. When only confirmed attributes are left or when the last iteration is reached.
- ii. The attributes may be left without a decision, called tentative attributes.

To avoid that, the number of iterations can be extended.

4.3 Classification

There are two different classification techniques that are used to predict the basketball game outcomes:

4.3.1 Support Vector Machine (SVM)

SVMs are a type of supervised learning methods that are used for regression, classification and outlier detection. Support vector machine is a technique that is used to classify both linear and nonlinear data. SVMs are capable of modeling complex nonlinear boundaries and are highly accurate. They can be used for classification as well as numeric prediction and avoid over-fitting. SVM can perform linearly inseparable classification with global optimization.

The support vector machines have many advantages:

1. It is efficient in high dimensional spaces.
2. It can be even effective with a greater number of dimensions and the comparatively smaller number of samples.
3. It is memory efficient as it uses a subset of training points in the dataset (called support vectors).
4. It is flexible as it has different kernel functions can be used for the decision function. Kernel functions can be customized according to requirement.
5. It reduces overfitting.

SVMs are used to get optimal separating hyperplane with the help of support vectors and margins. This hyperplane depicts the clear separation of different classes in the dataset. For nonlinear data, SVM uses nonlinear mapping to convert existing training dataset into a higher dimension. By using this new dimension, SVM searches for linearly optimal hyperplane separating the classes of the dataset.

Given the dataset D as:

$$\{x_i, y_i\}, \quad i = 1, 2, \dots, n, y_i \in \{+1, -1\} \quad (1)$$

Where, x_i is the set of training tuples corresponding to their class labels y_i .

SVM discovers the hyperplane with the largest margin called Maximum Marginal Hyperplane (MMH). This margin gives the maximum separation between the classes. A pair (w, b) exists such that w denotes weight vector and b refers scalar bias that satisfies the equation of hyperplane that can be written as:

$$w \cdot x_i + b = 0, \quad \text{for } i = 1, 2, \dots, n \quad (2)$$

Therefore, any point lying above the separating hyperplane satisfies the following equation:

$$w \cdot x_i + b < 0, \quad \text{for } i = 1, 2, \dots, n \quad (3)$$

Similarly, any point lying below the separating hyperplane satisfies the following equation:

$$w \cdot x_i + b > 0, \quad \text{for } i = 1, 2, \dots, n \quad (4)$$

The equations (3) and (4) can be written as the following equations respectively, that satisfies the condition [51], [52]:

$$H_1: \quad w \cdot x_i + b \geq 1 - \varepsilon_i, \quad \text{if } y_i = +1 \quad (5)$$

$$H_2: \quad w \cdot x_i + b \leq 1 - \varepsilon_i, \quad \text{if } y_i = -1, \quad \text{for } i = 1, 2, \dots, n \quad (6)$$

Where ε_i , is a slack variable that allows error in the training set and also makes a soft marginal hyperplane for classification. This slack variable is zero for two class separation. This indicates that tuple falling on or above hyperplane H_1 belongs to class +1 and tuple falling on or below H_2 belong to class -1. Combining both equations (5) and (6) we get:

$$y_i(w \cdot x_i + b) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0, \quad \text{for } i = 1, 2, \dots, n \quad (7)$$

Tuples falling on H_1 and H_2 hyperplane are called support vectors. These are tuples that are the most difficult to classify, but they provide more information regarding classification.

So, the size of the maximal margin can be obtained. The distance from hyperplane that separates two classes, to any point on H_1 or the distance from hyperplane that separates two classes, to any point on H_2 is:

$$d = \frac{1}{\|w\|} \quad (8)$$

Where $\|w\|$ is the Euclidean norm of w that is $\sqrt{(w \cdot w)}$

The distance describes above is equal to the distance from any point on H_1 or H_2 to the separating hyperplane. So, the size of maximal margin is:

$$D = 2 * d = \frac{2}{\|w\|} \quad (9)$$

The margin separating two classes can be maximized by solving the following quadratic problem.

$$\text{Minimize } \frac{1}{2} w \cdot w + C \sum_{i=1}^n \varepsilon_i \quad (10)$$

Subjected to the equation (7), where C is the constant used for controlling the scale of the margin and classification error. Equation (7) and (10) can be solved by using Lagrange multiplier α and then implementing Karush-Kuhn-Tucker (KKT) condition [54], [55] to the solution, the optimization problem can be written as:

$$\text{Maximize, } \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j k(x_i, x_j) \quad (11)$$

$$\text{Subjected to, } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^n \alpha_i y_i = 0, \quad \text{for } i = 1, 2, \dots, n \quad (12)$$

When searching a linear SVM in the new higher dimensional space, training tuples appear in the form of dot product of ϕ . The kernel function $k(x_i, x_j) = \phi(x_i) \times \phi(x_j)$ is used for converting non-linearly separable problem to the linearly separable problem by mapping non-linearly distinguishable data into higher dimensional feature space. After finding a solution, the distinguishing function can be given by:

$$d(x_i) = \text{sign} \left[\sum_{i=1}^n \alpha_i y_i k(x_i, x) + b \right], \quad i = 1, 2, \dots, n \quad (13)$$

The kernel function used in the model is a Gaussian function and can be written as the following equation (8):

$$k(x_i, x_j) = e^{\left[-\frac{(x_i - x_j)^2}{2\sigma^2} \right]} \quad (14)$$

The properties of SVM are:

A. Soft Margin Classifier

Practically the real data is sometimes complex as well as unorganized that cannot be separated completely with a hyperplane. To avoid this, the constraint used for maximizing the margin of the line which separates the classes must be relaxed.

Therefore, some set of coefficients is used that allows the margin to move side to side or up or down with small rapid movements. The margins are provided wiggle room in each dimension. These coefficients are known as slack variables. But due to this, the complexities of the model increases as to provide this complexity perfectly more parameters are used by the model to fit the data.

Tuning parameters are used to adjust the complexity allowed in the model. The C parameter determines the amount of wiggle that can be allowed across all dimensions as well as defines the margin's violation amount allowed. When C parameter is zero then no violation is allowed. With the increase in the value of C parameter the violations of the hyperplane increases.

During the process of generating the hyperplane from the data, the training instances that exists within the distance of the margin determines the placement of the hyperplane. These training instances are called support vectors. The C parameter determines the number of instances that within the margin. So, numbers of support vectors that are used by the model are influenced by the C parameter which means:

- i. The smaller is the value of C, the more sensitive is the algorithm to train the data which indicates higher variance and lower bias.
- ii. The larger is the value of C, the less sensitive is the algorithm to train the data which indicates lower variance and higher bias.

B. Support Vector Machines (Kernels)

Kernels are used to implements the SVM algorithms in practical. In linear SVM, the hyperplane is learned by transforming the problem with the help of linear algebra.

By using the inner product of any two given observations the linear SVM can be re-casted rather than the observations themselves.

The equation used for prediction of a new input is written as follows:

$$f(x) = b + \sum_{i=1}^n (w_i * K(x, x_i)) \quad (15)$$

The equation involves the dot product of a new input (x) and each support vector (x_i) in the data. For each input, the coefficient b and w_i are estimated from the training dataset by learning the algorithm. Three different types of kernel functions are:

- i. Linear kernel SVM:** The dot-product is known as the kernel. The kernel determines the similarity or a measure of distance between the support vectors and new data. Linear SVM or linear kernel use the dot product to measure the similarity. The dot product is given by the equation(16)

$$K(x, x_i) = \sum_{i=1}^n (x * x_i) \quad (16)$$

Other kernels such as Polynomial Kernel and Radial Kernel transform the input space into higher dimensional space. This is known as Kernel Trick. The more complex kernels are used to separate the classes that are complex or even more curved. This leads to the development of more accurate classifiers.

ii. Polynomial kernel SVM

Instead of using the dot product the polynomial kernel can be used. The polynomial kernel function can be represented as in equation (17):

$$K(x, x_i) = 1 + \sum_{i=1}^n (x * x_i)^d \quad (17)$$

Where d is the degree of the polynomial that is specified during the learning of the algorithm. When $d = 1$ the polynomial kernel function acts as the linear kernel function.

iii. Radial kernel SVM

The Radial Basis Function (RBF) is a type of radial kernel SVM. The RBF is most popular among the kernel types that are used in SVM. This is due to their finite and localized responses across an entire range of real x -axis. The radial kernel function can be classified as:

$$k(x_i, x_j) = e^{\left[\frac{(x_i - x_j)^2}{2\sigma^2} \right]} \quad (18)$$

Where, σ is a parameter that is passed to the learning algorithm.

C. Data Preparation for SVM

When learning the SVM model the training data can be prepared depending on:

- i. **Numeric inputs:** SVM accepts only numeric input value. If the input value is categorical then it is needed to be converted to binary dummy variables considering one variable for each category.
- ii. **Binary Classification:** Most of the SVMs are intended for binary classification problems although the extensions have been developed for multi-class classification and regression.

4.3.2 Hybrid Fuzzy-SVM (HFSVM)

Fuzzy logic is a multi-valued logic which includes any real number between 0 and 1 as the truth value of variables. In fuzzy logic, linguistic variables are used to facilitate the rules and facts expression. Fuzzification operation maps mathematical input values into membership functions. Membership functions allow quantifying linguistic variables with a degree of membership. Finally, fuzzy rules are generated as well as evaluated and defuzzification is used to map a fuzzy output into a crisp output. The output after defuzzification is used as the input to build SVM model. In this way, the fuzzy approach is integrated with SVM technique.

The fuzzy approach used in the proposed work is explained as follows:

A. Fuzzy Logic System

Fuzzy logic is applied to handle the partial truth concept which means true value lies between completely true value and completely false value. Whereas, in Boolean logic variable's truth value lies can be the integer value 0 or 1.

A fuzzy logic system (FLS) is described as the nonlinear mapping of an input dataset to the scalar output data. The FLS involves four major components: fuzzifier, fuzzy rules, fuzzy inference engine and defuzzifier. These components form the general architecture of an FLS as shown in figure 4.2:

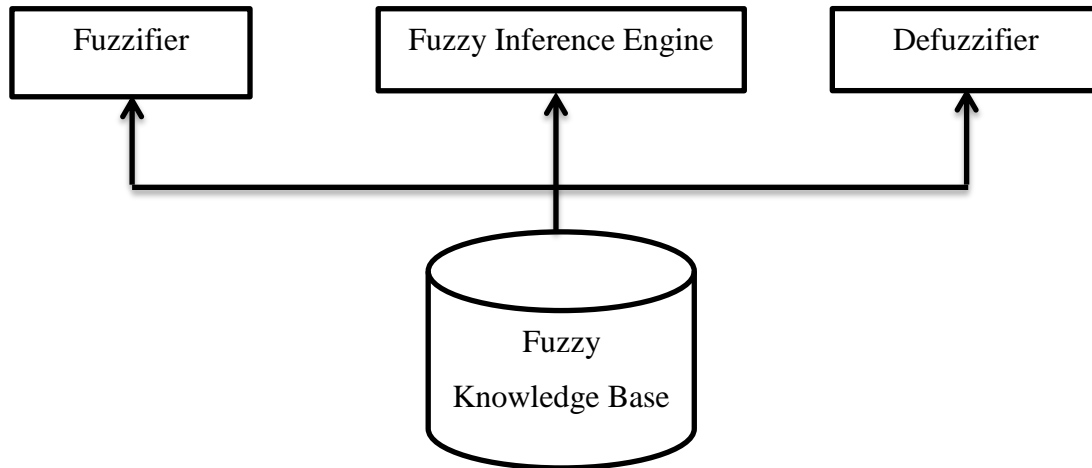


Figure 4.2: Fuzzy Logic System

Fuzzy logic process

Step 1: A crisp set of input is collected.

Step 2: It is converted to a fuzzy set of input with the help of fuzzy linguistic terms, fuzzy linguistic variables and the membership functions. This step is called fuzzification.

Step 3: All applicable rules existing in fuzzy rulebase are executed to compute the fuzzy output value.

Step 4: The outputs of all the rules are combined using aggregation process. The single fuzzy set is formed by combining the membership functions of all the rules.

Step 5: Defuzzification is performed by mapping fuzzy output values to obtain crisp data.

Fuzzy logic has many advantages:

1. Fuzzy logic concepts are easier to understand and use as it is based on natural language. The mathematics concepts used in fuzzy logic reasoning are simple.
2. Due to the flexibility of fuzzy logic, it can be used with any system and it is easy to add or manage more functions. Fuzzy logic is merged with conventional control techniques. The fuzzy system does not replace conventional control methods, but fuzzy systems can enhance them and make their implementation simple.

A. Fuzzification

It is the process of converting the crisp input values to the linguistic variables using membership functions that are stored in the fuzzy knowledge base.

- i. **Linguistic variables:** Linguistic variables are fuzzy system's input or output variables that contain the value in the form of word or sentence from a natural language, instead of having a numeric value.
- ii. **Membership functions:** A membership function is a measure of a linguistic term that gives the degree to which the linguistic term belongs to a fuzzy set. In fuzzification and defuzzification, the membership functions of fuzzy logic systems are used to map crisp input data to fuzzy linguistic terms and vice versa.

Given, the element x that belongs to set X in the fuzzy set A . Then the membership of x element can be described by membership function: $\mu_A(x)$, where μ is the membership degree. The degree of membership is the grade to which the element corresponds to the set. The value of the membership function lies between 0 and 1. This is shown in Figure 4.3.

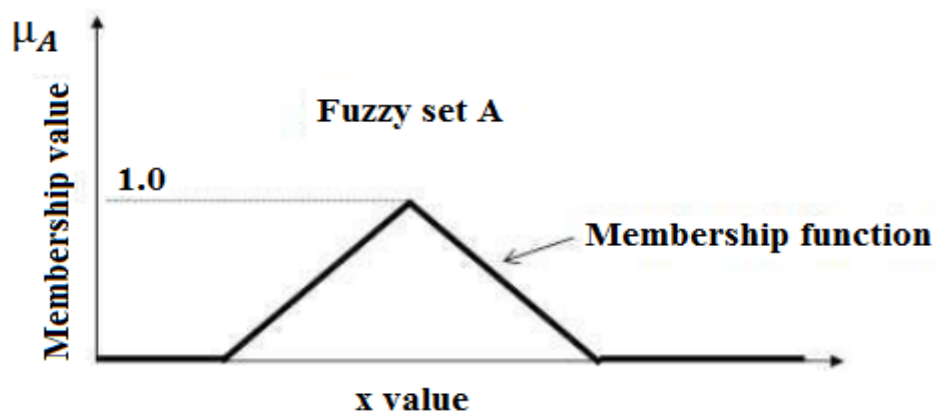


Figure 4.3: Membership Degree of x Value

If X is a collection of elements x , then a fuzzy set A can be described as a set of ordered pairs:

$$A = \{(x, \mu(x)) \mid x \in X\} \quad (18)$$

Where, $\mu(x)$ is called the membership function.

Different forms of membership functions are trapezoidal, triangular, Gaussian or singleton.

B. Fuzzy Rules

In a Fuzzy logic system, to control the output variable a rule base is constructed. A fuzzy rule is the form of IF-THEN rule having a condition and a conclusion. If x is A and y is B , a fuzzy rule can be implemented by a fuzzy relation, given by equation (19).

$$R: A \rightarrow B \quad (19)$$

The expression defines a relation between x and y where A and B are the linguistic values defined on universe discourse X and Y by fuzzy sets and “ x is A ” is known as the premise or antecedent as well as “ y is B ” is known as the conclusion or consequence. Therefore, the fuzzy rule can be described as a relation R which is binary, on the product space $X \times Y$. The relation can be treated as a fuzzy set having a two dimensional membership function which is expressed as in equation (20).

$$\mu_R(x, y) = f(\mu_A(x), \mu_B(y)) \quad (20)$$

Where, f is the function that is known as fuzzy implication function. It transforms the membership degrees of x in A and y in B into those of (x, y) in $A \times B$.

i. Fuzzy set operations: The Fuzzy set is a set without rigid boundaries. It has the flexibility to model linguistic expressions. It proposes the degree parameter with which a given element is linked to set. Fuzzy set operations perform fuzzy rule evaluations and the grouping of the outputs of the individual rules.

Given, μ_A as the membership function of fuzzy set A and μ_B as the membership function of fuzzy set B . The fuzzy set operations that can be performed on set A and B are:

a) OR fuzzy operation: OR fuzzy operation shown in Figure 4.4 is used to get the disjunction of the rule antecedents. For this, the fuzzy logic system uses the classical fuzzy operation that is union. OR fuzzy operation on set A and B can be expressed in equation (21).

$$\mu_{A \cup B}(x) = \max [\mu_A(x), \mu_B(x)] \quad (21)$$

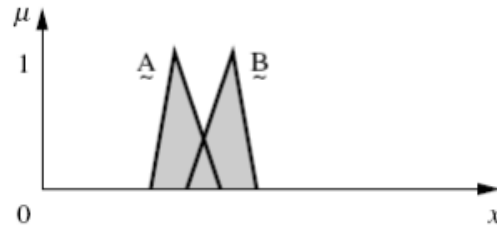


Figure 4.4: Union of Fuzzy Two Sets

- b) **AND fuzzy operation:** AND fuzzy operation shown in Figure 4.5 is used to get the conjunction of the rule precursors. For this, the fuzzy logic system uses the classical fuzzy operation that is an intersection. AND fuzzy operation on set A and B can be expressed in equation (22).

$$\mu_{A \cap B}(x) = \min [\mu_A(x), \mu_B(x)] \quad (22)$$

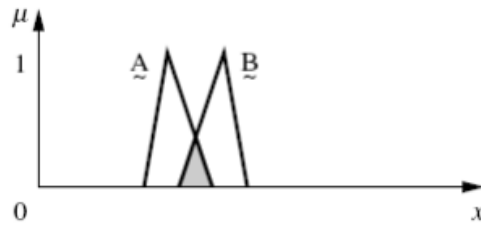


Figure 4.5: Intersection of Fuzzy Two Sets

- c) **NOT fuzzy operation:** NOT fuzzy operation shown in Figure 4.6 is used to get the complement of the rule antecedents. NOT fuzzy operation on set A and B can be expressed in equation (23).

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x) \quad (23)$$

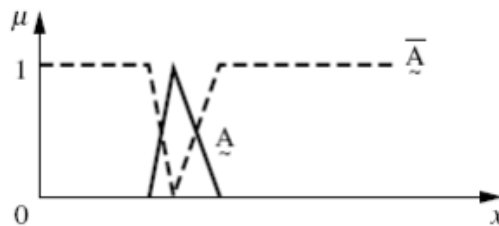


Figure 4.6: Compliment of Fuzzy Set

After evaluating of each rule's result, these results are combined to attain the final result. This process is known as inference. Different ways can be used to combine the results of individual rules with the help of fuzzy relations.

ii. Fuzzy relations: Binary fuzzy relations are the fuzzy sets A and B in X and Y , respectively. They map each element in $X \times Y$ to a membership degree that lies between 0 and 1. Fuzzy relations are used in decision making and fuzzy control. The most commonly used relation is max-min composition of relations.

The max-min composition of relations: Let X, Y and Z be the universal sets and R be a relation that relates elements from X to Y , in equation (24) and (25).

$$R = \{((x, y), \mu_R(x, y))\} \quad x \in X, y \in Y, R \subset X \times Y \quad (24)$$

$$S = \{((y, z), \mu_Q(y, z))\} \quad y \in Y, z \in Z, Q \subset Y \times Z \quad (25)$$

Then T is the relation that relates element in X that R contains to the element in Z that S contains, in equation (26).

$$T = R \bullet S \quad (26)$$

Where “ \bullet ” suggest the membership degree of R and S in max min sense.

C. Aggregation and Defuzzification

Aggregation is the process in which the outputs of all rules are unified. The membership functions of all rule consequents previously scaled or clipped are taken and combined into a single fuzzy set. It is done to obtain a final crisp output. Defuzzification depends on the output variable’s membership function. Different aggregation operations to perform defuzzification are:

i. Mean of maximum method (MOM): The MOM method creates a quantity that is the mean value of all outputs having membership functions that reach the maximum as in equation (27).

$$Z_0 = \sum_{j=1}^k \frac{z_j}{k} \quad (27)$$

z_j : The output with membership functions reaching the maximum.

k : Number of such outputs

ii. Centre of area method (COA): The COA method creates the center of gravity of all the possible distributions of a fuzzy set C that is defined on an output dimension z .

$$Z_0 = \left(\sum_{j=1}^n \mu_c(z_j) \cdot z_j \right) / \left(\sum_{j=1}^n \mu_c \cdot z_j \right) \quad (28)$$

n : Number of quantization levels of the output C

iii. Bisector of area (BOA): The BOA method creates the action (z_0) which partitions the original area into two equal regions as given in equation (29).

$$\int_{\alpha}^{z_0} \mu_c(z) dz = \int_{z_0}^{\beta} \mu_c(z) dz \quad (29)$$

Evaluation Parameters

The models used for predictions are evaluated through the confusion matrix and AUC. The confusion matrix consists of information about the actual and predicted results of classification performed by the model. It contains:

1. **True Negative (TN):** It depicts the number of events which are correctly predicted as ‘LOSS’.
2. **False Positive (FN):** It depicts the number of events which are predicted as ‘WIN’ but actually they are ‘LOSS’.
3. **False Negative (FN):** It depicts the number of events which are predicted as ‘LOSS’ but actually they are ‘WIN’.
4. **True Positive (TP):** It depicts the number of events which are correctly predicted as ‘LOSS’.

The Confusion Matrix is shown in Table 4.1.

Table 4.1: Confusion Matrix

| | | Predicted Condition | |
|----------------|--------------------|----------------------------------|--------------------------------------|
| | | Prediction Positive | Prediction Negative |
| True Condition | Condition Positive | True Positive (TP) | False Negative (FN) Type II Error |
| | Condition Negative | False Positive (FP) Type I Error | True Negative (TN) |

From Confusion Matrix, various evaluation parameters are derived to measure the performance of the models. The derived evaluation parameters are:

1. **Accuracy:** It is the ratio of events that are correctly predicted by the total number of events and is expressed in equation (30).

$$Accuracy = \frac{TP + TN}{TN + FP + FN + TP} \quad (30)$$

2. **True Negative Rate:** It is also termed as specificity. It is the ratio of events that are correctly predicted as 'LOSS' to the sum of the condition negative events as expressed in equation (31).

$$TNR = \frac{TN}{TN + FP} \quad (31)$$

3. **True Positive Rate:** It is also termed as sensitivity or recall. It is the ratio of events that are correctly predicted as 'WIN' to the sum of the condition positive events as expressed in equation (32).

$$TPR = \frac{TP}{TP + FN} \quad (32)$$

4. **False Positive Rate:** It is the complement of true negative rate as given in equation (33).

$$FPR = 1 - TNR \quad (33)$$

5. **False Negative Rate:** It is the complement of true positive rate as given in equation (34).

$$FNR = 1 - TPR \quad (34)$$

6. **Positive Predictive Value:** It is also known as precision. It is the ratio of relevant instances to the total number of retrieved instances as given in equation (35).

$$PPV = \frac{TP}{TP + FP} \quad (35)$$

5.1 Data Used

In proposed work, the dataset is collected from websites such as “NBA.com”, “basketball-reference.com” which provide very valuable and informative data. In this study, Data with total 800 games is collected from 2015-2016 regular season. This data contains one decision attribute and 33 are condition attributes.

The attributes employed in this study are depicted in Table 5.1 along with their corresponding description. First, nineteen attributes in this table are basketball game’s fundamental attributes and rests of the attributes except for a decision attribute, are advanced attributes of the game. The attributes from X1 to X33 are condition attributes and Y is a decision attribute.

Table 5.1: Attributes of NBA

| Attributes | Abbreviation | Description |
|-------------------|---------------------|-------------------------------|
| X1 | MP | Minutes Played |
| X2 | FG | Field Goal |
| X3 | FGA | Field Goal Attempts |
| X4 | FG% | Field Goal Percentage |
| X5 | 3P | 3-Point Field Goal |
| X6 | 3PA | 3-Point Field Goal Attempts |
| X7 | 3P% | 3-Point Field Goal Percentage |
| X8 | FT | Free Throw |
| X9 | FTA | Free Throw Attempts |
| X10 | FT% | Free Throw Percentage |
| X11 | ORB | Offensive Rebound |
| X12 | DRB | Defensive Rebound |
| X13 | TRB | Total Rebound |
| X14 | AST | Assists |
| X15 | STL | Steals |
| X16 | BLK | Blocks |
| X17 | TOV | Turnover |

| | | |
|-----|------|---------------------------|
| X18 | PF | Personal Fouls |
| X19 | PTS | Points |
| X20 | TS% | True Shooting Percentage |
| X21 | eFG% | Effective Field Goal %age |
| X22 | 3Par | 3-Point Attempt Rate |
| X23 | FTr | Free Throw Attempt Rate |
| X24 | ORB% | Offensive Rebound %age |
| X25 | DRB% | Defensive Rebound %age |
| X26 | TRB% | Total Rebound Percentage |
| X27 | AST% | Assist Percentage |
| X28 | STL% | Steal Percentage |
| X29 | BLK% | Block Percentage |
| X30 | TOV% | Turnover Percentage |
| X31 | USG% | Usage Percentage |
| X32 | ORtg | Offensive Rating |
| X33 | DRtg | Defensive Rating |
| Y | W/L | Win/Loss |

The fundamental parameters are: MP is the duration of the game, FG is the basket scored on any shot, worth two points, FGA is the total attempts for the field goal (including failed attempts), FG% is the ratio of field goals made to field goal attempted, 3P is the field goal in basketball that is made beyond the 3 point line, 3PA is the total three-point field goals attempted (including failed attempts), 3P% is the ratio of 3P and 3PA, FT is the free throw that is unopposed attempt to score points, FTA is the total free throw attempts including failed ones, FT% is the ratio of FT to FTA, ORB is the rebound secured by an offensive player on missing the shot from the teammate, DRB is the rebound secured by defending team after shot missed by offending team, AST is a pass that directly leads to the basket, STL is taking the ball away from the opponent, BLK is deflecting a field goal attempt from an offensive player, TOV is the losing possession of the ball before taking the shot in basket, PF is the personal foul (illegal contact with an opponent) and points is the scores made by team.

The advanced parameters are: TS% is true shooting percentage which is a measure of shooting efficiency, eFG% is effective field goal percentage that adjusts FG% so that 3P only counts for three points while FG only counts for two points, 3Par is three point rate means

percentage of FG attempts from 3P range, FTr is free throw attempt rate means number of FT attempts per FG attempts, ORB% is offensive rebound percentage that is an estimate of percentage of available offensive rebounds grabbed, DRB% is an estimate of percentage of available defensive rebounds grabbed, TRB% an estimate of percentage of total available rebounds, AST% is an estimated percentage of teammate FG's a player assisted, STL% is an estimated percentage of opponent possessions, BLK% is an estimated percentage of FGA's blocked, TOV% is an estimate of turnovers per 100 players, USG% is an estimated percentage of team plays used by a player, ORtg is an estimation of points scored per 100 possessions and DRtg is an estimation of points allowed per 100 possessions.

5.2 Evaluation Parameters

The evaluation parameters (used in the proposed work) and their values are given in Table 5.2.

Table 5.2: Evaluation Parameters of SVM and HFSVM

| | Abbreviation | SVM | HFSVM |
|------------|--------------------------|------------|--------------|
| TNR | True Negative Rate | 0.9 | 0.846 |
| TPR | True Positive Rate | 0.846 | 0.929 |
| FPR | False Positive Rate | 0.1 | 0.154 |
| FNR | False Negative Rate | 0.154 | 0.071 |
| PPV | Positive Predictive Rate | 0.917 | 0.839 |

The true negative rate (also known as specificity) of SVM is 0.9 and that of HFSVM is 0.846. The true positive rate (also known as sensitivity or recall) of SVM is 0.846 and that of HFSVM is 0.929. It depicts that the SVM predicts more accurately the true 'LOSS' events than HFSVM. On the other side, HFSVM predicts more accurately the true 'WIN' events than SVM. But the accuracy depends on the net effect of sensitivity and specificity. The net effect of sensitivity and specificity of HFSVM is more than SVM. Therefore, HFSVM is better in this case. The false positive rate of SVM is 0.1 and that of HFSVM is 0.154. HFSVM miss-classifies more 'WIN' events than SVM. The false negative rate of SVM is 0.154 and that of HFSVM is 0.071. On the other hand, SVM miss-classifies more 'LOSS' events than HFSVM. The positive predictive rate of SVM (0.917) is more than that of HFSVM (0.839).

5.3 Simulation and Results

The boruta algorithm was used for feature selection. It provides the result in the form of a plot with the importance of attributes against shadow attributes as shown in Figure 5.1. The attributes with high variable importance are shown in green color, the attributes in yellow color are tentative attributes and attributes in red color are rejected attributes and are of low variable importance. This means that the attributes beyond ORB% are selected attributes. BLK, FTA and STL% are the tentative attributes and remaining attributes are rejected.

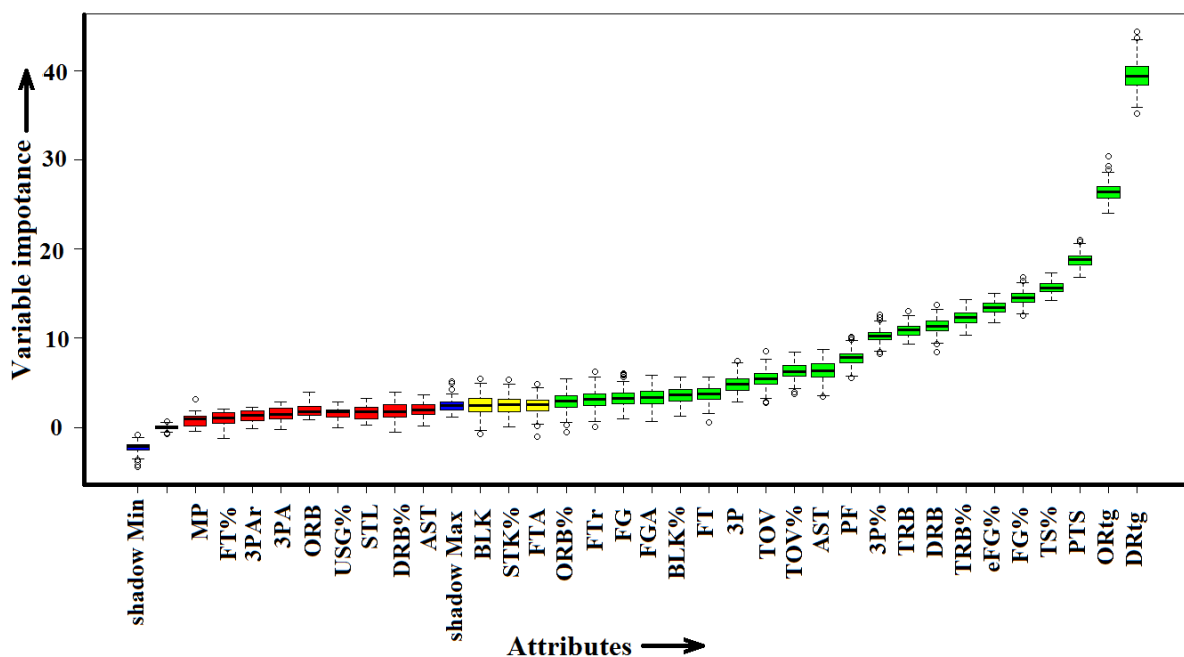


Figure 5.1: Attribute versus Corresponding Variable Importance

The Boruta algorithm selects 21 condition attributes that are shown in Table 5.3. After selection of attributes, the partitioning of data was done in two parts: the training dataset contains 640 games and testing dataset contains 160 games of the regular season.

Table 5.3: Attributes of NBA Selected from Boruta Algorithm

| Attributes | Abbreviation | Description |
|------------|--------------|-------------------------------|
| A1 | FG | Field Goal |
| A2 | FGA | Field Goal Attempts |
| A3 | FG% | Field Goal Percentage |
| A4 | 3P | 3-Point Field Goal |
| A5 | 3P% | 3-Point Field Goal Percentage |
| A6 | FT | Free Throw |

| | | |
|-----|------|---------------------------|
| A7 | DRB | Defensive Rebound |
| A8 | TRB | Total Rebound |
| A9 | TOV | Turnover |
| A10 | PF | Personal Fouls |
| A11 | PTS | Points |
| A12 | TS% | True Shooting Percentage |
| A13 | eFG% | Effective Field Goal %age |
| A14 | FTr | Free Throw Attempt Rate |
| A15 | ORB% | Offensive Rebound %age |
| A16 | TRB% | Total Rebound Percentage |
| A17 | BLK% | Block Percentage |
| A18 | TOV% | Turnover Percentage |
| A19 | AST | Assists |
| A20 | ORtg | Offensive Rating |
| A21 | DRtg | Defensive Rating |
| Y | W/L | Win/Loss |

Table 5.4 and 5.5 represent the result for each of the classifications that are SVM and HFSVM respectively. The measuring parameters are accuracy, computational time (in seconds) and a number of support vectors for both Table 5.4 and Table 5.5. The highest testing accuracy (87.82%) for the SVM model is given in Table 5.4 by Cross-Validation 2 (CV2). Similarly, the highest testing accuracy (89.26%) for the HFSVM model is given in Table 5.5 by Cross-Validation 3(CV3).

Table 5.6 outlines average testing accuracy, average computation time and an average number of support vectors of both SVM and HFSVM models. In the table, it is deduced that

Table 5.4: Experimental Results of Testing Accuracy with SVM

| | Accuracy | C/T(s) | Number of Support Vectors |
|------------|----------|--------|---------------------------|
| CV1 | 85.71 | 1.84 | 540 |
| CV2 | 87.82 | 1.79 | 545 |
| CV3 | 85.09 | 1.75 | 538 |
| CV4 | 85.71 | 1.74 | 551 |
| CV5 | 86.34 | 2.04 | 559 |

Table 5.5: Experimental results of testing Accuracy with HFSVM

| | Accuracy | C/T(s) | Number of Support Vectors |
|------------|-----------------|---------------|----------------------------------|
| CV1 | 87.60 | 0.93 | 390 |
| CV2 | 88.43 | 0.94 | 401 |
| CV3 | 89.26 | 1.09 | 381 |
| CV4 | 87.26 | 1.22 | 392 |
| CV5 | 88.43 | 102 | 393 |

HFSVM model can achieve higher average testing accuracy (88.26%) than can the SVM model (86.21%). The computation time of the HFSVM model is shorter than the SVM model.

Table 5.6: Average Fivefold Cross-Validation Results of SVM and HFSVM

| | Average Accuracy | Average C/T(s) | Average No. of Support Vectors |
|--------------|-------------------------|-----------------------|---------------------------------------|
| HFSVM | 88.26 | 1.04 | 391 |
| SVM | 86.21 | 1.83 | 547 |

Therefore, we can conclude that the net error effect is reduced when fuzzy membership is implemented to each input point of the dataset as these input points make different contributions to the decision surface learning. In the previous paper [49], the testing accuracy for predicting outcomes of basketball games was (85.25%). Thus, the testing accuracy attained by the HFSVM model is quite adequate. In addition to this, the average type 1 and type 2 prediction error rates of fivefold cross-validation with SVM are 6.21% and 7.56% correspondingly. SVM model's total average error rate is 13.77%. The average type 1 and type 2 of fivefold cross-validation with HFSVM are 6.44% and 5.28% correspondingly. HFSVM model's total average error rate is 11.72%, which is less than the total average error rate of SVM. Also, in HFSVM the type 2 error is smaller than type one error which is considered better in predicting the outcome. Type 1 error depicts the probability when the true outcome is "win" but the result by prediction model is "loss". Type 2 error depicts the probability when the true result is "loss", but the result but prediction model is "win".

The ROC (Receiver Operating Characteristic) plot is the plot between true positive rate and false positive rate. The ROC plot of SVM is shown in Figure 5.2 (a). The area under the

curve is 0.937 and area under convex hull is 0.946 which is indicated by dotted line. Similarly, the Roc plot of HFSVM is shown in Figure 5.2 (b). The area under the curve is 0.958 and the area under the convex hull is 0.97. Clearly, both Area under the curve and the area under the convex hull of HFSVM is more than SVM.

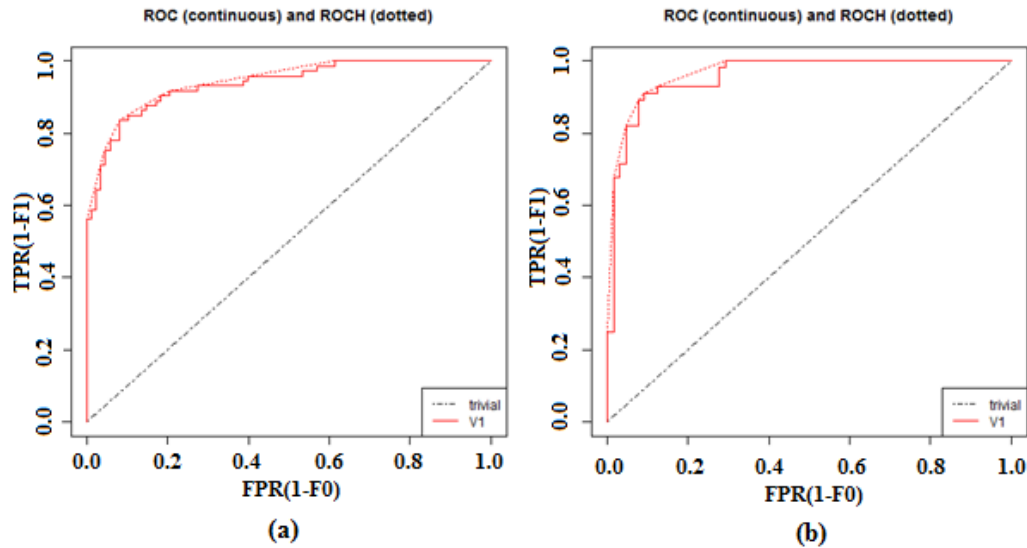


Figure 5.2: ROC and ROCH Plot (a) SVM and (b) HFSVM

The AUC plot is the plot of the cost function and its corresponding weight. From Figure 5.3 (a) SVM puts no weight on $C=0.5$ on the other hand in Figure 5.3 (b) HFSVM put a weight of approximately 0.05 on $C=0.5$. Thus, the probability of severity rate is equal to 1, is more in HFSVM than SVM.

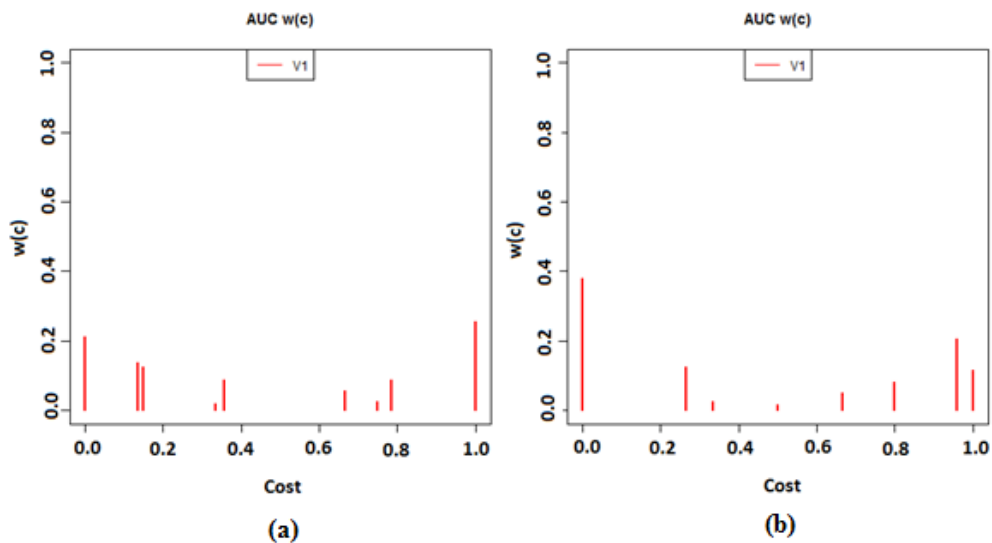


Figure 5.3: The AUC $w(c)$ Function, Corresponding to ROC Curve of (a) SVM and (b) HFSVM

As shown in Figure 5.4 (a) and (b), the H-measure curve of HFSVM is more symmetric than SVM and also severity rate of SVM is 0.83 and that of SVM is 0.862.

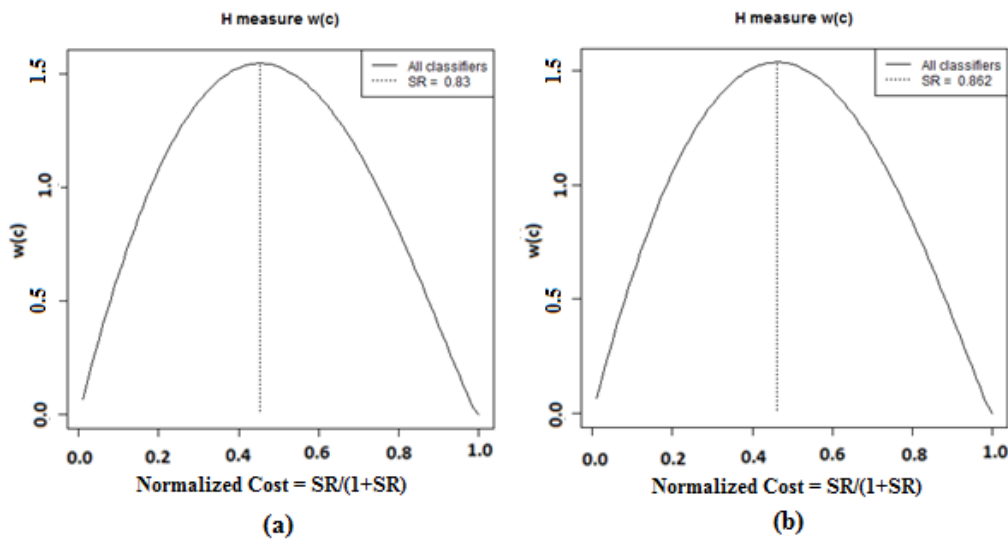


Figure 5.4: The H measure $w(c)$ Function versus Cost Curve of (a) SVM and (b) HFSVM

The smooth score distribution curve is the score versus density curve. The curve of HFSVM in Figure 5.5 (b) is more smooth and in bell shape than the curve of SVM in Figure 5.5 (a). The common area between the intersecting curve of HFSVM in Figure 5.5 (b) is less than the common area between the intersecting curve of SVM in Figure 5.5 (a). Thus, sensitivity and a specificity value of HFSVM is 0.821 and 0.723, respectively and as that of SVM is 0.753 and 0.602, respectively.

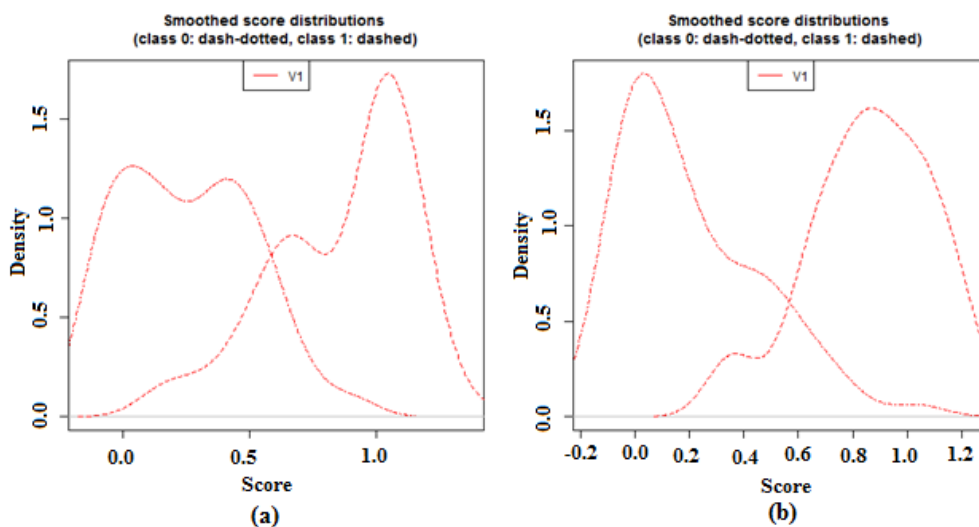


Figure 5.5: Smooth Score Distribution Curve of (a) SVM and (b) HFSVM

6.1 Conclusion

Analyzing the basketball game is an interesting research area for the researchers because of the arousing curiosity of fans and social media for the prediction of outcomes of the basketball competitions. This is also helpful for the teams and players to enhance their performance. But predicting the outcome of the basketball game is a challenging task. It is found that taking advanced attributes of NBA game results in increasing the accuracy of the model. So, this leads towards the development of the HFSVM model for predicting the outcome in the NBA. Comparing SVM model with HFSVM model, it is found that HFSVM provides better results. The attributes that plays very important role in predicting outcomes are defensive rating and offensive rating while the attributes minutes played and the field goal attempt rate has least effect on the results. HFSVM model can achieve higher average testing accuracy (88.26%) than can the SVM model (86.21%). The computation time of the HFSVM model is shorter than the SVM model. Also, on comparing the HFSVM model with previous studies in predicting the basketball games' outcomes, the accuracy achieved by HFSVM model is quite adequate. Therefore, HFSVM model can be used as a promising alternative for predicting the basketball game outcomes.

6.2 Contribution Summary

The contribution of the proposed work is that the Machine Learning models can be used to predict the outcome of the basketball game. The used models iteratively learn from the NBA dataset and automatically find the hidden pattern each time without being explicitly programmed. Machine learning algorithms used, create a simulation model that highly increased the classification accuracy for prediction. Also, the accuracies and computational time of the classification algorithms are compared by executing them on NBA data set used in the proposed work. The rules are generated using fuzzy approach for the machine learning algorithms that lack in rule generation for building the model to be used for classification problem.

6.3 Future Scope

The HFSVM model is only capable of predicting the win and loss outcomes of basketball games. Further extending HFSVM model the scores of win and loss can be investigated. This model can also be further applied to other sports such as soccer, baseball, and golf, in order to check the feasibility of HFSVM model. Further, machine learning algorithms that are used help to get the important factors that are required to predict the NBA game as well as provide the importance of these factors in a particular NBA game. The proposed work helps the coaches to manage their team and helps the players to increase their performance. It provides the suggestion that which factors can be changed by how much so that the game can be improved. It can be used by experts to analyze the sports and it can be used by many companies that spend a large amount of money in predicting the outcomes of the game for betting. It can also be used by fans that are interested in knowing more about the game, or their favorite teams or players.

References

- [1] D. Andrews. "The (Trans) National Basketball Association: American commodity-sign culture and global-local conjuncturalism." *Articulating the Global and the Local*. Boulder, CO: Westview, pp 72-101, (1997).
- [2] R.P. Schumaker, O.K. Solieman, and H. Chen. "Predictive modeling for sports and gaming." *Sports Data Mining*. Springer US, pp 55-63, (2010).
- [3] M. Spann and B. Skiera. "Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters." *Journal of Forecasting*, vol 28(1), pp 55-72, (2009).
- [4] D. Forrest, J. Goddard, and R. Simmons. "Odds-setters as forecasters: The case of English football." *International journal of forecasting*, vol 21(3), pp 551-564, (2005).
- [5] C. Song, B.L. Boulier, and H.O. Stekler. "The comparative accuracy of judgmental and model forecasts of American football games." *International Journal of Forecasting*, vol 23(3), pp 405-413, (2007).
- [6] D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy. "Learning influence among interacting Markov chains." *Advances in Neural Information Processing Systems*, pp 1577-1584, (2006).
- [7] J. Fürnkranz. "Machine learning in games: A survey." *Machines that learn to play games*, pp 11-59, (2001).
- [8] R.R. Huilgol and S.S. Chhabra. "A Review of Data Mining in Sports."
- [9] C. Soto Valero. "Predicting Win-Loss outcomes in MLB regular season games—A comparative study using data mining methods." *International Journal of Computer Science in Sport*, vol 15(2), pp 91-112, (2016).
- [10] D. Delen, D. Cogdell, and N. Kasap. "A comparative analysis of data mining methods in predicting NCAA bowl outcomes." *International Journal of Forecasting*, vol 28(2), pp 543-552, (2012).
- [11] S. Demers. "Riding a probabilistic support vector machine to the Stanley Cup." *Journal of Quantitative Analysis in Sports*, vol 11(4), pp 205-218,(2015).
- [12] W. Gu, T.L. Saaty, and R. Whitaker. "Expert System for Ice Hockey Game Prediction: Data Mining with Human Judgment." *International Journal of Information Technology & Decision Making*, vol 15(04), pp 763-789, (2016).
- [13] C.K. Leung and K.W. Joseph. "Sports data mining: predicting results for the college football games." *Procedia Computer Science*, vol 35, pp 710-719, (2014).
- [14] W. Zeng and J. Li. "Fuzzy logic and its application in football team ranking." *The Scientific World Journal*, vol 2014, pp 1-6, (2014).
- [15] I.S. Thakare, S.R. Suyal, and K.Y. Pandav. "Performance Evaluation for Sports Team Selection Using Data Mining Techniques." *AADYA-National Journal of Management and Technology (NJMT)*, vol 5, pp 102-108, (2015).

- [16] W. Jian, H. Zhi-Hua, and Z. Zhi-Yong. "Clustering Analysis of Sports Performance Based on Ant Colony Algorithm." *Intelligent Systems Design and Engineering Applications (ISDEA), 2014 Fifth International Conference on*. IEEE, pp. 288-291, (2014).
- [17] C. Rajeswari. "Performance analysis of sports person using data mining based ranking and classification methods." *IJPT*, vol 16(3), pp 18070-18097, (2016).
- [18] E. Štrumbelj and P. Vračar. "Simulating a basketball match with a homogeneous Markov model and forecasting the outcome." *International Journal of Forecasting*, vol 28(2), pp 532-542, (2012).
- [19] P. Vračar, E. Štrumbelj, and I. Kononenko. "Modeling basketball play-by-play data." *Expert Systems with Applications*, vol 44, pp 58-66, (2016).
- [20] D.A. Harville. "The selection or seeding of college basketball or football teams for postseason competition." *Journal of the American Statistical Association*, vol 98 (461), pp 17-27, (2003).
- [21] R.J. Leake. "A Method for Ranking Teams With an Application to College Football." in *Management Science in Sports*, eds. R. E. Machol, S.P. Ladany, and D.G. Morrison, Amsterdam: North-Holland, pp 27-46, (1976).
- [22] H.S. Stern. "Who's Number One? Rating Football Teams," in *Proceedings of the Section on Statistics in Sports*, American Statistical Association, pp 1-6, (1992).
- [23] R.T. Stefani. "Football and basketball predictions using least squares." *IEEE Transactions on systems, man, and cybernetics*, vol 7, pp 117-121, (1977).
- [24] T.A. Zak, C.J. Huang, and J.J. Siegfried. "Production efficiency: the case of professional basketball." *Journal of Business*, pp 379-392, (1979).
- [25] T. Chen and Q. Fan. "A functional data approach to model score difference process in professional basketball games." *Journal of Applied Statistics*, pp 1-16, (2016).
- [26] B. Dežman, S. Trinić, and D. Dizdar. "Expert model of decision-making system for efficient orientation of basketball players to positions and roles in the game—Empirical verification." *Collegium antropologicum*, vol 25(1), pp141-152, (2001).
- [27] D. Cervone, A. D'Amour, L. Bornn, and K. Goldberry "POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data." *8th Annual MIT Sloan Sports Analytics Conference, February*, vol 28, (2014).
- [28] F. JR Ruiz and F. Perez-Cruz. "A generative model for predicting outcomes in college basketball." *Journal of Quantitative Analysis in Sports*, vol 11(1), pp 39-52, (2015).
- [29] H. Manner. "Modeling and forecasting the outcomes of NBA basketball games." *Journal of Quantitative Analysis in Sports*, vol 12(1), pp 31-41, (2016).
- [30] G. Cheng, Z. Zhang, M.N. Kyebambe, and N. Kimbugwe. "Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle." *Entropy*, vol 18(12), p 450, (2016).
- [31] D. Miljković, L. Gajić, A. Kovačević, and Z. Konjović. "The use of data mining for basketball matches outcomes prediction." *Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on*. IEEE, pp 309-312, (2010).

- [32] E. Hermann and A. Ntoso. "Machine Learning Applications in Fantasy Basketball." pp 1-5, (2015).
- [33] B. Markoski, P. Pecev, L. Ratgeber, M. Ivković, and Z. Ivanković. "A new approach to decision making in basketball-BBFBR program." *Acta Polytechnica Hungarica*, vol 8(6), pp 111-130, (2011).
- [34] K.C. Wang and R. Zemel. "Classifying NBA offensive plays using neural networks." *Proc. MIT SLOAN Sports Analytics Conf*, pp 1-9, (2016).
- [35] Z. Ivanković, M. Racković, M. Markoski, D. Radosav, and M. Ivković. "Analysis of basketball games using neural networks." *Computational Intelligence and Informatics (CINTI), 2010 11th International Symposium on*. IEEE, pp 251-256, (2010).
- [36] Z. Ivankovic, M. Rackovic, and M. Ivkovic. "Automatic player position detection in basketball games." *Multimedia tools and applications*, vol 72(3), pp 2741-2767 (2014).
- [37] J.B. Yang and C.H. Lu. "Predicting NBA championship by learning from history data." *Proceedings of artificial intelligence and machine learning for engineering design*, pp 1-4, (2012).
- [38] B. Markoski, Z. Ivanković, L. Ratgeber, P., Pecev, and D. Glusac. "Application of AdaBoost algorithm in basketball player detection." *Acta Polytechnica Hungarica*, vol 12(1), pp 189-207, (2015).
- [39] M. Beckler, H. Wang, and M. Papamichael. "Nba oracle." *Zuletzt besucht am*, vol 17, pp 2008-2009, (2013).
- [40] Z. Shi, S. Moorthy, and A. Zimmermann. "Predicting NCAAB match outcomes using ML techniques—some results and lessons learned." *ECML/PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics*, (2013).
- [41] R. Shah and R. Romijnders. "Applying Deep Learning to Basketball Trajectories." *arXiv preprint arXiv 1608.03793* (2016).
- [42] M. Perše, M. Kristan, S. Kovačič, G. Vučković, and J. Perš. "A trajectory-based analysis of coordinated team activity in a basketball game." *Computer Vision and Image Understanding*, vol 113(5), pp 612-621, (2009).
- [43] B. Loeffelholz, E. Bednar, and K.W. Bauer. "Predicting NBA games using neural networks." *Journal of Quantitative Analysis in Sports*, vol 5(1), pp 1-15, (2009).
- [44] V. Papić, N. Rogulj, and V. Pleština. "Identification of sport talents using a web-oriented expert system with a fuzzy module." *Expert Systems with Applications*, vol 36(5), pp 8830-8838, (2009).
- [45] S. Balli and S. Korukoğlu. "Development of a fuzzy decision support framework for complex multi-attribute decision problems: A case study for the selection of skilful basketball players." *Expert Systems*, vol 31(1), pp 56-69, (2014).
- [46] K. Trawinski. "A fuzzy classification system for prediction of the results of the basketball games." *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*. IEEE, (2010).
- [47] N.A. Erilli and E. Ermis. "A New Measuring Efficiency for Basketball Using Fuzzy Clustering Analysis." *Kasmera*, vol 43(1), pp 161-175, (2015).

- [48] M. Atlas and Y.Q. Zhang. "Fuzzy Neural Agents for Online NBA Scouting." *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, vol 4, pp 58-67, (2004).
- [49] P.F. Pai, L.H. ChangLiao, and K.P. Lin. "Analyzing basketball games by a support vector machines with decision tree model." *Neural Computing and Applications*, pp 1-9 (2016).
- [50] J.N. Wang. "The fuzzy regression analysis application of offensive and defensive techniques in the basketball game." *Natl Sci Counc Repub China Part C Humanit Soc Sci*, vol 10, pp 287-298, (2000).
- [51] V. Vapnik and C. Corinna. "Support Vector Networks", *Machine Learning* vol 20, pp 273-297, (1995).
- [52] V. Vapnik. *The nature of statistical learning theory* Springer sciences and business media, (2013).
- [53] N. Barakat and A.P. Bradley. "Rule extraction from support vector machines: a review." *Neurocomputing*, vol 74(1), pp 178-190 (2010).
- [54] W. Karush. "Minima of functions of several variables with inequalities as side conditions." *Traces and Emergence of Nonlinear Programming*, pp 217-245, (2014).
- [55] H.W. Kuhn and A.W. Tucker. "Nonlinear programming." *Proceeding of 2nd Berkerey Symposium on Mathematical Statistics and Probabilities*. University of California Press, Berkeley, California, pp 481-492, (1951).

List of Publication

1. Harmandeep Kaur and Dr.Sushma Jain, “Machine Learning Approaches to predict Basketball Game outcome”, International Conference on Advances in Computing, Communication & Automation, IEEE Explore, 2017 [Accepted]
2. Harmandeep Kaur and Dr.Sushma Jain, “Basketball Game Outcome Prediction using Machine Learning”, International Conference on Advanced Computational and Communication Paradigms, Springer, 2017 [Communicated]

7.1 January 2015

- **Descriptive**

During this phase I initialized my work on my Master's Thesis. It was the time when the word "research" seems to be very new to me. I first time felt the difference between Bachelor's and Master's degree. It was the phase when my mentor cleared me the meaning of "Research".

- **The Reflection**

I initialized with this event as in the beginning of Master's, many persons didn't know much about research. My mentor guided me in the proper direction by making me aware that research is not just about writing or publishing thesis paper.

- **The Outcome**

My mentor asked me about the area in which I have interest and made me free so that I can choose any of the research topic related to my field (Machine Learning). I started searching on Google and downloaded some related survey papers from where I come to know about the possible directions in which I can work. Then, I met my mentor with my initial research findings.

7.2 February, 2015

- **Descriptive**

After completion of my initial research, I and my mentor started discussing it. We discussed one week on each and every topic that I had searched last month. This was the phase when I attended regular research session with my guide.

- **The Reflection**

I had selected this event as I want to analyze the importance of discussions in choosing the research topic. These sessions helped me to choose the best research topic related to my interest, under the guidance of my mentor. I come to know that I was going in right direction in choosing the research topic. We ended up these healthy discussions with a very demanding and significant research topic "Sports prediction". I began my work on this topic. I searched for good journal papers and shortlisted few

of them. These papers were only surveyed papers to get familiar with new terms related to my topic. After this when I met my guide she suggested me with some books to get into the work in more detail.

- **The Outcome**

Now, this was the time to convert these rational ideas into practical work. But there was a need to know about the way of implementing it. This was the phase when I was known with the topic thoroughly.

7.3 March, 2015

- **Descriptive**

I am discussing this event because at this point of time I was at a standstill and had no idea in which direction to move. At this time I took a step to meet my mentor and discuss the point where I felt the problem. I didn't know how to collect my dataset.

- **The Reflection**

This event was full of frustrations but still, I opt for a fresh start every time I got stuck in between. My mentor suggested me some sites to explore for the dataset. I was learning about the tool R and was facing problem in implementing my research work properly. I was not aware of the tool properly. I started reading some R documentaries and started exploring some websites such as Analytics Vidhya and R-blogger. My mentor boosted me up and I started implementing some of the machine learning approaches on the sample data.

- **The Outcome**

Working hard does not go in vain, it will pay off someday but one must not have to lose hope. One needs to discuss their problem with their mentor without any hesitation.

7.4 April, 2015

- **Descriptive**

It was the time when I completely knew about the direction which I followed. At this phase of the time, I had no doubts in my mind and all I wanted was to implement the best way I had searched after so many experiments.

- **The Reflection**

Before this, I was experimenting with my sample data. When I came to know the best possible way to get my results I started working on it. I found that by using large dataset I can improve my work, Using advanced features also lead to the increase in efficiency of my work. This was the point I was able to choose the best technique for my research,

- **The Outcome**

There is no shortcut to the success. Doing work systematically does not lead to wastage of time and efforts.

7.5 May, 2015

- **Descriptive**

This was the time to bind up my work with the implementation of techniques that were chosen on the dataset. After that, thesis writing process began that ended in the month of June.

- **The Reflection**

There was also the time when I knew about the research. I didn't know much about the tool R. During this complete thesis phase, when I lost patience my mentor boosted me up and ask me to give a try. This always encouraged me to work. I learned new techniques and finally felt relief when I got expected results for my proposed work.

- **The Outcome**

Machine Learning is a popular and rapidly growing area. I got the opportunity to learn this new technique. So, there is always a need to explore more and more and never to give up when doing a thesis work. This thesis had not only made me learn new techniques and tools but also some moral values that how to have patience in difficult situations and how to work systematically.

Video Presentation

<https://www.youtube.com/watch?v=XRT1W34Tyuk>