# Sport Result Prediction Using Classification Methods

Arash Mazidi[1*], Mehdi Golsorkhtabaramiri[2], Naznoosh Etminan[3]

**Abstract**–Traditional sport was based on the ability of the players and less science and knowledge was considered. However, sport has become a profession and an industry. Therefore, the use of technology and analysis on data in order to achieve goals is very important. Classification is one of technologies to classify new incoming samples. Furthermore, sports produce considerable information about each season, teams, matches and players. Classification on sport data helps managers and coaches in order to predict the match result, evaluate the player performance, predict the player injury, identify the sports talent and evaluate the match strategy. There are many algorithms to predict the basketball results, track the health of players and determine the strategy of the match against different opponents, which help coaches a lot. Further, preprocessing procedure makes better dataset. In this paper, we use classification methods on sport dataset using preprocessing procedure and without preprocessing. The results show an improvement was obtained results using preprocessing.

**Keywords:** Data Mining, Classification, Result Prediction, Basketball match

## 1. Introduction

Data mining is the process of discovering hidden, interesting, unexpected, and valuable patterns within a data. It is an activity related to the precise analysis of unstructured data that statistics is incapable of analyzing. Data mining is performed by special equipment, which performs the exploration based on repeated data analysis. In data mining methods, data collection, data integration, and clearing, all of which are preprocessing stages of data mining. Data mining tools are able to predict future, behaviors, and the potential for active business and knowledge-based decisions. Data mining techniques can be quickly implemented on existing software and hardware platforms to increase the value of available information resources[1].

Also, in the education industry, which is called educational data mining, it is related to evolving methods that discover information obtained from educational environments and from techniques such as decision trees, neural networks, nearest neighbor and so on. There are many data mining applications in Table 1.

Data mining has been used since ancient times in popular sports such as baseball, football and basketball. One of the researches is about using data mining for more suitable substitutions of football players. It has been shown that the most effective substitution to return the result of a football match is as follows: the first substitution earlier than the 58th minute, the second substitution earlier than the 73th minute and the third substitution earlier than the 79th minute. Therefore, it shows that coaches should make a substitution at early second half. There was a belief that knowledge existed only in minds of experts, educators, managers and analysts, while there are a lot of statistical information in sports for every player, team and match that are not used effectively. Sports organizations are turning to data mining techniques to gain more knowledge of the existing data. Sports data mining systems are created to assist sports coaches and managers in the result prediction of matches, player performance evaluation, player injuries prediction, talent identification and evaluating the strategy used in the matches. Today's, predicting the results of sport matches is one of the most interesting and hot topics among sport enthusiasts and has attracted many attentions. Therefore, according to the data collected in sports, data mining techniques have been used as methods to analyze the data[11].

One of the sports that has received many attentions is basketball, which a lot of data is collected due to mobility and a large number of events during a match. These events create large data sets that can be analyzed and predicted using data mining methods. In this article, we will use various data mining techniques in order to extract a model to predict the result of basketball match[12].

The rest of paper is organized as follows: Section 2 presents data mining in sport. A literature study of solutions provided for result prediction in sports is surveyed in section 3. The classification algorithms are provided in section 4. Section 5 discusses the experimental setup and results. Finally, section 6 concludes and provides the future directions of this study.

[1*] **Corresponding Author:** Department of Computer Engineering, Faculty of Engineering, Golestan University, Gorgan, Iran.
E-Mail: arash_mazidi_67@yahoo.com
[2] Department of Computer Engineering, Babol Branch, Islamic Azad University, Babol, Iran
[3] Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran

*Received:2020.08.10 ; Accepted: 2020.12.15*

**Table 1:** Applications of data mining

| *Application* | *Example* |
|---|---|
| Common business applications | Market analysis, oil price forecasting, target marketing, understanding customer behavior and risk analysis[2] |
| Fraud detection | Telephone fraud detection and car insurance fraud detection[3] |
| Text mining | Summarization and sentiment analysis[4] |
| Medicine | Discovering the relationship between symptoms and disease, DNA analysis[5] |
| Web mining | Related pages suggestion and improving search engines, web traffic analytics[4] |
| Network intrusion detection | Network traffic analysis[6] |
| Bioinformatics | Multiple sequence alignment[7] |
| Cloud | Resource allocation[8]–[10] |
| Sport | Recommendations for virtual training and result prediction |

## 2. Data Mining in Sport

There is no doubt that professional sports have become a business. Michael Jordan has played for the Chicago Bulls for twenty years and is one of the most famous and well-known athletes in the world. He earned 37 million dolor in a year, which is more than the domestic income of many countries. The market of sports in the world has become much wider than before due to the existence of large television programs and sponsors. Therefore, result prediction of matches has attracted many fans[12].

There are different data elements that will influence the result of a match. Some data elements that directly affect the result include player's goals, type of sport, environment, and so on. Some teams use human experts to analyze and predict the results of sports. Using automated methods and data mining algorithms to predict results can produce results that are more accurate. Data mining aims to help coaches and players to predict results as well as to assess player performance, predict player injuries, and evaluate game strategy. The knowledge discovered from sport data may be used to describe situations or predict situations[13].

Therefore, many sports coaches and managers use data mining to advance the sports. The results show that they have made significant progress in this area. A pilot program was conducted in 2002 by the Italian football club AC Milan by collecting data from training sessions to predict players' injuries over a period of time. This was done with the aim of predicting the players injuries as well as preventing injuries, which ultimately resulted in a lot of financial gain by preventing the players' injuries in favor of the club. The details of the test program were as follows: The head of the medical team of the AC Milan football team asked 18 players to connect sensors to predict the possibility of injury. These sensors were connected to a computer and stored and processed information about players. These data were related to the physiology data of the players and their nutrition. Data were collected over 18 months. 5,000 tests were performed on players. Further, the neural network was used to predict the occurrence of injuries. The results of data mining on player data were such that the neural network correctly predicted 84% of injuries. Mathematicians also speculated that this accuracy could be increased to 96%. Other variables also affect the probability of injury, including conditions, fatigue, weather and ground conditions[12].

Furthermore, at the American Basketball Association, Oliver Dean has been a consultant to the Seattle team for more than half a decade as one of the best basketball analysts. A book on the subject entitled "Rules and Tools for Performance Analysis" has been published, which has provided a unique perspective for organizations. Similarly, other NBA basketball teams began to employ analytical minds that could provide new insights into available data[14].

## 3. Related Work

One of the most important applications of data mining and machine learning is result prediction of sports matches. In many sports matches, fans are eager to predict the results of some matches. If one person has correctly predicted all the results of matches, he is winner. Today's, these results can be predicted automatically using machine learning algorithms and probabilistic rules. One of the sports prediction systems that use machine learning is designed at the University of Melbourne Australia, to predict the result of football matches. In this system, a rating is given to each prediction. This is a modified system of the Kullback-leibler rating system. The most important change made to this system is the addition of 1 to each formula. This change causes all prediction models that perform better than the normal predictor with a probability value of 0.5 to have a rating greater than zero, while models worse than normal will have a rating less than zero.

McCabe et al. [15] have used neural networks and artificial intelligence to predict the result of sports games. They have proposed a model with several functional layers, also one of the features of this method is the use of features that show the quality of each team. In the first phase, they have extracted the features, thus they eliminated the need for human judgment and examination of the team features. The extracted features are changing in each round of the tournament according to the match. They selected the best features for each team and then predicted the result of the matches using a neural network. In addition, they collected data from various leagues to evaluate their system.

Min et al. [16] have developed an algorithm that predicts football results using machine learning methods. Many fans of sports teams are very interested in predicting the results of matches. There are many factors that should be considered when predicting the result of a sport match, which complicates the prediction. Rules-based methods such as Bayesian rule are widely used to predict the future, including predicting the results of sports games as well as forecasting stock prices. The authors have proposed two methods in order to predict the result of football match. The proposed framework has two main components includes the rule-based analyst and Bayesian network-based component. The reason for combining these two different methods in one context is the fact that the results of sports games are completely random and probable, and at the same time the strategy and structure of the team can be presented using a series of rules and can affect the final result. Therefore, according to the rule-based analyst, the proposed method is able to make predictions with sufficient accuracy. Also, when we have little statistical data, it is able to predict the result. For example, when two teams or players do not have any previous encounters, other methods of machine learning are not able to predict the result, but the method suggested by authors in these cases has the ability to predict the results with sufficient accuracy. In addition, while most of previous methods considered just one factor, the proposed method considered some factors such as current score, skill level, mood and fatigue. Furthermore, unlike other previous methods, authors have predicted a knowledge-based method using a time series during the game. They have implemented the proposed method called FRES and have shown that the results of the proposed framework are reasonable and stable.

Trawinski et al. [17] have used fuzzy logic to predict the result of sports games. They collected data automatically by extracting information from the pages of sports websites that reflected the results of basketball games. Their data contained the results of the 2008-2009 ACB League season. They have selected ten fuzzy logic methods and algorithms and have examined the results of these methods in predicting the results of sports games. In addition, they compared the results with the regression method.

Haghighat et al. [18] have reviewed many methods in result predictions. They have included a lot of sports game statistics in their review research, including information about each player, team, their games and the season.

Miljkovic et al. [19] have used data mining methods to predict the results of NBA matches. They have formulated the problem of match result prediction as a classification. They also used a simple Bayesian algorithm to classify the results. In addition, authors have implemented the proposed system using data from a selection of 778 games in 2009-2010 NBA season. They represented each game with a record of 141 attributes. Two groups of features were considered for each team. The first group included the normal statistical dataset of a basketball game, including free throws, rebounds, blocking shots, and so on. The features of the second group describe the position of the team in the league. In this group, there are a set of attributes such as the number of wins and losses, hosts and guests and so on. In addition, they used SVM, K-NN and decision tree algorithms to classify the data with comparison Bayesian method. The results showed that the simple Bayesian classification method shows more favorable results than other methods.

Leung et al. [20] have proposed a method based on data mining. The proposed method extracts the useful knowledge from any sport match, and can also predict the final result of the football match. They have also evaluated their method using the dataset of football matches. The dataset consists of the number of passes, the number of errors, the loss of the ball, the number of attacks and other statistical data from the teams. After collecting statistical data, they were stored in two data structures. A list contains all the matches played over time and a list contains information about all the teams and their statistical data in the season. They were able to accurately predict the results of football matches with accuracy 91.4%.

Bhandari et al. [21] have tried to collect data on NBA basketball matches and predict their results. First, they have pre-processed the data and then extracted the necessary knowledge from them using data mining methods. NBA basketball data is recorded by advanced recording systems. This data includes shots, type of shots, results and rebounds, and so on. Each action has a time code that specifies the time of its occurrence. After the match, this data is uploaded in electronic bulletins and each team can access the necessary data. They used this dataset to evaluate their method. One of the necessary tests for data preprocessing is to check their compatibility. The data is then prepared in a structured format. This structured format prepares data for electronic processing and also makes it easier for educators to extract the necessary knowledge from structured data. At the end of the preprocessing phase, they performed the data enrichment step by adding a number of composite information as fields to the existing data as needed. Then, using data mining methods, they extracted effective knowledge for the analysis of trainers.

Ivanković et al. [22] have used data mining to extract information and knowledge from data related to basketball matches. Data collected from the Serbian Basketball League from 2005 to 2010. By observing the 5 seasons of this competition, they have collected statistical data related to 890 matches. In the first phase, they analyzed the data collected for each player using preprocessors. Therefore, using these preprocessors, the data related to the players will be summarized and the data related to a team will be created. These data were analyzed using the feed-forward method. The feed-forward method is one of the non-linear methods for analyzing statistical data obtained from sports matches. Eventually they came to the conclusion that two-point shots and defensive rebounds are the most effective moves in a basketball game. Finally, they determine a number of parameters of each team and predict the results using existing neural networks. The designed neural network had 12 input nodes and one output node. Neural network is also a type of feed-forward networks. Each layer within the network is fully connected to all nodes of the previous layer as well as the next layer. The final results of the predicted results have an accuracy of 80.96%.

Kampakis et al. [23] have used a set of Twitter messages to predict the result of football matches. Twitter is a significant source for forecasting in various areas such as the stock market, diseases, as well as the results of sports matches. They have developed a set of predictive models for predicting English League results over a three-month period. They compared the results of the proposed model with existing models that use statistical data and football history. The results showed that the textual data available on Twitter will be useful for predicting the results of professional football. The combined model has performed better than the two models based on statistical data and based on Twitter data. Among the features that have been considered for statistical data are the average number of goals, the average number of corners, the average of targeted shots, and so on. Each sample of dataset records was divided into three groups of attributes of host team, guest team and related variables. The results showed that the use of combined features has improved the accuracy of prediction of match results. Prediction accuracy using textual data collected from Twitter was 25%. Also, the accuracy of predicting the results using statistical data was 23%, but the results of combined data were 28%.

Peace et al. [24] have developed a system for predicting the result of football matches relying on the importance of football in other societies and the interest of fans using neural networks and logistic regression. They used the RapidMiner tool to perform data mining tasks. The performance accuracy of the system for predicting results based on neural networks was 85% and the system based on logistic regression was 93%.

Davoodi et al. [25] have proposed a method to predict the results of horseback riding using artificial intelligence and data mining methods. They have used neural networks to predict the results of matches. The dataset used to evaluate the proposed method was collected from New York Horse Racing. They used statistical data related to matches in 2010. This dataset contains information about each horse participating in this match. A horse network has been used for each horse so that the output of the neural network is at the end of the race by that horse. The proposed method has a functional accuracy of 77% that was acceptable for a horse race.

Serveet al. [26] have used the previous results of football teams to predict the results of future matches. They used the Dutch football tournament set to test their proposed method. This collection contains statistical data related to the matches of the Dutch league 15 years ago. This dataset contains information such as the results of the first half, the final results of the match, the goals scored in the previous 7 matches and the number of wins of the team in the last 10 matches. They first preprocessed the data using Weka in such a way that they removed the useless attributes from the set and put together a reduced set of attributes. They have a set of attributes scored by the home team, goals scored by the away team, goals scored by the home team, goals scored by the home team, average points scored by the home team and average points scored by the away team in the last x matches to represent a team. The variable x is inserted with the appropriate value during the testing and evaluation process. Authors have used classification methods such as regression algorithm, logistic regression algorithm, decision tree, Boosting method and Bayesian network. They have shown that using the last 20 and 30 matches is better to determine the features. Also, the best algorithm to predict the results is the regression method, which has the highest accuracy in predicting the results and is able to predict with accuracy 55% of the results.

Ravichandran et al. [27] have proposed a graph-based model to predict the results of NBA basketball matches. They showed the data for each team using a graph. They used data from three consecutive NBA basketball seasons in 2007-2008, 2008-2009, and 2009-2010 to evaluate their method. In the first season 1183 matches, the second season 1176 and the third season 1215 matches have been played. They used attributes such as number of personal faults, throws fault, total number of rebounds, defensive rebounds, offensive rebounds, and so on. To compare the proposed method with previous methods, the basic method of linear regression, which has been one of the methods in previous researches has been examined. The results showed that the linear regression method for predicting the results of future matches had an accuracy of 41.5%, while their proposed method with a graph-based had a prediction accuracy of 65.5%.

## 4. Basketball Match Result Prediction

Many data in society do not convey useful information. In order to obtain useful information from the data, processing and analyzing the data is important. One of the methods of data analysis is the use of data mining methods. In sports, there will be a lot of information about the match and the players that in order to improve the quality of team performance, data needs to be processed and analyzed to make good predictions. Different data elements will influence the result of a sport match. Some data elements that directly affect the results are the goals of the players, the type of sport and environment. While some teams use human experts to analyze and predict the result, using automated methods to predict results can produce more accurate results. In the proposed method, we first collect the required data and then perform the data preprocessing operation, which is one of the most important parts of the work. Without quality data, no quality results can be extracted and the quality of decisions must be based on quality data. Further, duplicate or missing data may give inaccurate or even misleading statistics. Therefore, preprocessing in this paper includes four steps: data cleaning, data conversion and integration, data reduction and data discretization. Data cleaning involves filling in lost data and identifying and deleting noise data. Also, reducing data or straw data reduces the volume of data. Discretization is part of data reduction, especially for numerical data. It divides a wide range of continuous data into meaning intervals, reducing data volume and preparing for analysis.

After the preprocessing step, we perform the algorithms on the preprocessed data using Weka software. We selected 14 algorithms from the existing algorithms in weka. The main reason for the selection of these algorithms is successful result, which we will introduce one by one.

### 4.1.Bagging algorithm

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.

Each base classifier is trained in parallel with a training set which is generated by randomly drawing, with replacement, N examples (or data) from the original training dataset – *where N is the size of the original training set*. Training set for each of the base classifiers is independent of each other. Many of the original data may be repeated in the resulting training set while others may be left out. Bagging reduces overfitting (variance) by averaging or voting, however, this leads to an increase in bias, which is compensated by the reduction in variance though[28].

### 4.2. Multilayer perceptron (MLP) algorithm

This algorithm is an artificial neural network model that maps a set of input data to a suitable set of output data. This algorithm consists of several layers of nodes in a directional graph, each of which connects to all subsequent nodes. Except for the input node, all nodes are a nerve (processing element) with a nonlinear activation function. The MLP algorithm uses a supervised learning method called post-diffusion to train the network. This improved algorithm is the standard linear perceptron algorithm that is able to isolate data that cannot be separated linearly. The MLP algorithm is mostly used in computational neuroscience, parallel distributed processing, speech and image recognition. This algorithm is used to obtain appropriate approximate solutions in a short time in complex problems[29].

### 4.3. Voted Perceptron

The voted perceptron method is based on the perceptron algorithm of Rosenblatt and Frank. The algorithm takes advantage of data that are linearly separable with large margins. This method is simpler to implement, and much more efficient in terms of computation time as compared to SVM. The algorithm can also be used in very high dimensional spaces using kernel functions. In a general comparison of this algorithm with SVM classification algorithm in the field of natural language processing, it is shown that this algorithm has been much better than SVM algorithm in terms of accuracy, computational time, learning time and prediction speed[30].

### 4.4. LogitBoost algorithm

LogitBoost is a popular Boosting variant that can be applied to either binary or multi-class classification. From a statistical viewpoint LogitBoost can be seen as additive tree regression by minimizing the Logistic loss. Following this setting, it is still non-trivial to devise a sound multi-class LogitBoost compared with to devise its binary counterpart. The difficulties are due to two important factors arising in multiclass Logistic loss. The first is the invariant property implied by the Logistic loss, causing the optimal classifier output being not unique, i.e., adding a constant to each component of the output vector won’t change the loss value. The second is the density of the Hessian matrices that arise when computing tree node split gain and node value fittings. Oversimplification of this learning problem can lead to degraded performance. For example, the original LogitBoost algorithm is outperformed by ABC-LogitBoost thanks to the latter’s more careful treatment of the above two factors[31].

### 4.5. MultiBoostAB algorithm

The MultiBoostAB algorithm is a classification algorithm. The MultiBoosting algorithm is a successful extension of the AdaBoost algorithm for decision committee formation. This algorithm can be seen as a combination of AdaBoost and Wagging algorithms. This algorithm is able to control the high bias of AdaBoost and reduce the variance of the Wagging algorithm. This algorithm generates decision committees with less error than other AdaBoost-based algorithms[32].

### 4.6.Raced Incremental LogitBoost algorithm

This algorithm is a classification algorithm that competes with the LogitBoost algorithm committees and incrementally trains the entire larger dataset by processing smaller sets of dataset[33].

### 4.7. ConjunctiveRule algorithm

Conjunctive Rule algorithm implements a single conjunctive rule learner that can predict for numeric and nominal class values. Conjunctive rule uses the relation of logical AND to link stimulus attributes. The rule involves "AND"ing the antecedents together and the consequent (class value) for the classification. In this case, the consequent is the distribution of the available classes (or mean for a numeric value) in the dataset. If this rule does not enclose the test instance, then the default class distributions/value of data that is not enclosed by the rule in the training data is used to predict it. This learner selects an antecedent by calculating the Information Gain of each antecedent and the generated rule is pruned using Reduced Error Pruning (REP) or simple pre-pruning depending on the number of antecedents. The weighted mean of the entropies of both the data covered and not covered by the rule is the Information of one antecedent used for classification. Single conjunctive rule learner is one of the machine learning algorithms and is normally known as inductive Learning[34].

### 4.8. DecisionTable algorithm

Decision Table is an accurate method for numeric prediction from decision trees. It is an ordered set of If-Then rules that is more compact and understandable than the decision trees. Selection to explore decision tables because it is a simpler, less compute intensive algorithm than the decision-tree-based approach. The algorithm, decision table, is found in the Weka classifiers under Rules. The simplest way of representing the output from machine learning is to put it in the same form as the input. It summarizes the dataset with a "decision table" which contains the same number of attributes as the original dataset. The use of the classifier rules decision table is described as building and using a simple decision table majority classifier. The output will show a decision on a number of attributes for each instance. The number and specific types of attributes can vary to suit the needs of the task. Decision Table classifier algorithm is used to summarize the dataset by using a decision table containing the same number of attributes as that of the original dataset. A new data item is allocated a category by searching the line in the decision table that is equivalent to the values contained in the non-class of the data item[35].

### 4.9. DTNB algorithm

DTNM algorithm is combination of Decision Tree (DT) and Naïve Bayes (NB) algorithms. At each point in the search space, the algorithm evaluates the criterion for dividing attributes into two distinct subsets. One for the decision tree algorithm and the other for the NB algorithm. A forward selection is used for the search, in each step, the NB algorithm models the selected attributes and the rest are modeled by the decision tree algorithm, and finally the decision tree algorithm models all the attributes[36].

### 4.10. NNGE algorithm

NNGE is an algorithm that takes a broad view of exemplars devoid of nesting or overlapping. NNGE creates a generality every time a new example is inserted to the database, by merging it to its nearest neighbor of the same class. Hyper rectangles are not permitted to nest or overlap by NNGE. To achieve this, every potential recent generalization is tested. NNGE uses a heuristic that executes this post-processing in an even manner. Since NNGE averts extrapolation by remedying any overlay or nesting, there is no need for the "second chance" heuristic. As an alternative, it usually attempts to extrapolate new examples to their nearest neighbors of identical class. However, if this is not instantly possible owing to superseding negative examples, no generalization is executed. In situations where a generalization afterwards differs from a negative example, it is adjusted to preserve uniformity and reliability. NNGE is an incremental learner. It does this by initially classifying, and afterward generalizing every new example. It employs a modified Euclidean distance function that processes hyper rectangles, symbolic features, and exemplars and feature weights[37].

### 4.11. OneR algorithm

OneR is a simple algorithm that simply predicts the class of a sample by finding the most frequent class for the feature values. OneR is shorthand for One Rule, indicating we only use a single rule for this classification by choosing the feature with the best performance. While some of the later algorithms are significantly more complex, this simple algorithm has been shown to have good performance in some real-world datasets[38].

### 4.12. PART algorithm

One of the rule-based classification algorithms is PART algorithm that uses a set of If-Then rules for classification. PART is a partial decision tree algorithm, which is the developed version of C4.5 and RIPPER algorithms. The main specialty of the PART algorithm is that it does not need to perform global optimization like C4.5and RIPPER to produce the appropriate rules[39].

### 4.13. RIDOR algorithm

Ripple down-rule-learner (RIDOR) is a rule base classification which used for classification. The RIDOR classifier first generated the default-value for certain situation. Whenever the situation occur the default value generated with minimum error rate. The default value continuously iterated the value by using incremental rule which minimize the errors pruning and generated most accurate result. The RIDOR is also work on the concept of learning in the previous value and generated future value. The RIDOR mainly use the concept of if-else statement. RIDOR iterate value until its trues then generate output else generate default value as output. The RIDOR classifier is also useful for uncertainties because uncertainties values are set before iterating the training and testing data[40].

### 4.14. ZeroR algorithm

The ZeroR algorithm is the simplest category-based algorithm that eliminates all predictions. This algorithm simply predicts the majority class. Although there is no strong prediction in this algorithm, it is very useful and used to determine the baseline of performance or compare it with the performance of other algorithms[41].

## 5. Implementation and Evaluation

In this section, the implementation process of the proposed algorithm for classification is described. It needs some standard measures to compare algorithms. Most of the papers use standard measures like precision, recall and F-measure that are calculated in Equations 1, 2 and 3. In addition, introduced algorithms have been implemented with the WEKA. A confusion matrix (Table 2) is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known[42], [43].

**Table 2.** Standard measures in the outlier detection

| | | *Predicted Class* | |
|---|---|---|---|
| | | Class = Yes | Class = No |
| Actual Class | Class = Yes | **False Negative (FN)** | **True Positive (TP)** |
| | Class = No | **True Negative (TN)** | **False Positive (FP)** |

True Positives (TP): These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

True Negatives (TN): These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

False Positives (FP): When actual class is no and predicted class is yes.

False Negatives (FN): When actual class is yes but predicted class in no.

According to Table1, measures of precision, recall and F-measure can be calculated with equations 1, 2 and 3 respectively.

$$Precision = \frac{TP}{(TP+FP)} \tag{1}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{2}$$

$$F - measure = \frac{2\times(Precision \times Recall)}{(Precision+Recall)} \tag{3}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. Furthermore, Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. F-measure is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

The dataset used in this research is related to basketball matches. The data is collected from http://www.basketball-reference.com by crawler4j, which is an open-source application. This dataset contains information about the two host and guest teams. The dataset with 21 features was formed, which are the first ten features for the host team and the same features for the guest team. A final feature is provided to predict the result of the match between the guest and the host based on the last 10 matches. This dataset contains 1167 data samples, the features are listed in the Table 3.

**Table 3**: Description of data set features

| *2-point throws percentage* | *Average percentage of successful 2-points throws of the team* |
|---|---|
| Penalty Throw Percentage | Average percentage of successful penalty throws |
| Rebound Percentage | Average percentage of successful rebounds |
| 3-point throws percentage | Average percentage of successful 3-points throws of the team |
| Throw Percentage | Average percentage of successful throws |
| Assists | Total number of assists or goal passes that are passes to a teammate that results in a quick score |
| Steals | Total number of steals when dribbling, passing or moving the team |
| Blocks | Total number of changing the direction of the ball by the defender touching when the attacker throws the ball. |
| Offensive Rating | Average percentage of successful team offensive |
| Defensive Rating | Average percentage of successful team defenses |
| Result | Result prediction |

The results of classification algorithms on preprocessed data and without preprocessing are in Table 4. The results show that preprocessing is important and we will have better results.

## 6. Conclusions and future work

Nowadays, sport is out of the traditional mode and has become a kind of science that requires the use of modern technologies. Therefore, data mining and other technologies are used in sports science to achieve better results. There are many algorithms to predict the basketball results, track the health of players and determine the strategy of the match against different opponents, which help coaches a lot. In this paper, we have reviewed basketball match prediction algorithms. First, preprocessing procedures including data cleaning, data conversion, data reduction and data discretization are performed on the dataset and then we use classification algorithms to classify the data. In another scenario, in order to show the effects of preprocessing, we classify the dataset without preprocessing. After comparing the data without preprocessing with the preprocessed data, a relative improvement was obtained and it is suggested that the statistics of the whole season of a team be considered both in home matches and in away matches.

In future work, it is suggested that the proposed method be used on other sports as well as other classification methods and compared with the methods in this article.

**Table 4:** Classification results on preprocessed data and without preprocessing

| | *Before Using Preprocessing* | | |
|---|---|---|---|
| Algorithm | **Precision** | **Recall** | **F-Measure** |
| Bagging | 0.59 | 0.638 | 0.599 |
| MultiLayer Perceptron | 0.558 | 0.581 | 0.549 |
| Voted Perceptron | 0.507 | 0.646 | 0.417 |
| LogitBoost | 0.555 | 0.639 | 0.584 |
| MultiBoostAB | 0.509 | 0.647 | 0.772 |
| RacedIncrementalLogitBoost | 0.507 | 0.646 | 0.417 |
| ConjunctiveRule | 0.507 | 0.646 | 0.417 |
| DecisionTable | 0.507 | 0.646 | 0.417 |
| DTNB | 0.515 | 0.641 | 0.541 |
| NNge | 0.555 | 0.559 | 0.551 |
| OneR | 0.57 | 0.627 | 0.577 |
| PART | 0.577 | 0.63 | 0.584 |
| Ridor | 0.527 | 0.641 | 0.568 |
| ZeroR | 0.508 | 0.619 | 0.532 |
| | **After Using Preprocessing** | | |
| | **Precision** | **Recall** | **F-Measure** |
| Bagging | 0.574 | 0.619 | 0.574 |
| MultiLayer Perceptron | 0.564 | 0.579 | 0.57 |
| Voted Perceptron | 0.555 | 0.616 | 0.554 |
| LogitBoost | 0.573 | 0.637 | 0.543 |
| MultiBoostAB | 0.417 | 0.646 | 0.507 |
| RacedIncrementalLogitBoost | 0.417 | 0.646 | 0.507 |
| ConjunctiveRule | 0.417 | 0.646 | 0.507 |
| DecisionTable | 0.554 | 0.632 | 0.534 |
| DTNB | 0.543 | 0.599 | 0.55 |
| NNge | 0.547 | 0.545 | 0.546 |
| OneR | 0.417 | 0.646 | 0.507 |
| PART | 0.568 | 0.599 | 0.575 |
| Ridor | 0.548 | 0.64 | 0.519 |
| ZeroR | 0.413 | 0.616 | 0.501 |

## References

[1] D. J. Hand and N. M. Adams, “Data Mining,” in Wiley StatsRef: Statistics Reference Online, Chichester, UK: John Wiley & Sons, Ltd, 2015, pp. 1–7.

[2] I. Karakatsanis et al., “Data mining approach to monitoring the requirements of the job market: A case study,” Inf. Syst., vol. 65, pp. 1–6, Apr. 2017.

[3] A. Mazidi, F. Roshanfar, and V. Parvin Darabad, “A Review of Outliers: Towards a Novel Fuzzy Method for Outlier Detection ,” J. Appl. Dyn. Syst. Control, vol. 2, no. 1, pp. 7–17, Jun. 2019.

[4] S. Tofighy and S. M. Fakhrahmad, “A proposed scheme for sentiment analysis: Effective feature reduction based on statistical information of SentiWordNet,” Kybernetes, vol. 47, no. 5, pp. 957–984, 2018.

[5] A. Mazidi and F. Roshanfar, “PSPGA: A New Method for Protein Structure Prediction based on Genetic Algorithm,” J. Appl. Dyn. Syst. Control, vol. 3, no. 1, pp. 9–16, Jun. 2020.

[6] A. Mazidi, M. H. Saddredini, and H. Tahayori, “ProposingA NewAlgorithmtoDetectLocalOutliersinData Stream,” vol. 4, no. 4. JournalofSoft Computingand Information Technology (JSCIT), pp. 31–42, 01-Jan-2016.

[7] N. Etminan, E. Parvinnia, and A. Sharifi-Zarchi, “FAME: Fast And Memory Efficient multiple sequences alignment tool through compatible chain of roots,” Bioinformatics, 2020.

[8] A. Mazidi, M. Golsorkhtabaramiri, and M. Yadollahzadeh Tabari, “An autonomic risk- and penalty-aware resource allocation with probabilistic resource scaling mechanism for multilayer cloud resource provisioning,” Int. J. Commun. Syst., p. e4334, Feb. 2020.

[9] A. Mazidi, M. Golsorkhtabaramiri, and M. Y. Tabari, “Autonomic resource provisioning for multilayer cloud applications with K□nearest neighbor resource scaling and prioritybased resource allocation,” Softw. Pract. Exp., vol. 50, no. 8, pp. 1600–1625, Aug. 2020.

[10] A. Mazidi, E. Damghanijazi, and S. Tofighy, “An Energy-efficient Virtual Machine Placement Algorithm based Service Level Agreement in Cloud Computing Environments,” Circ. Comput. Sci., vol. 2, no. 6, pp. 1–6, 2017.

[11] H. Yan, N. Yang, Y. Peng, and Y. Ren, “Data mining in the construction industry: Present status, opportunities, and future trends,” Automation in Construction, vol. 119. Elsevier B.V., p. 103331, 01-Nov-2020.

[12] R. P. Bonidia, J. D. Brancher, and R. M. Busto, “Data Mining in Sports: A Systematic Review,” IEEE Latin America Transactions, vol. 16, no. 1. IEEE Computer Society, pp. 232–239, 01-Jan-2018.

[13] D. Rojas-Valverde, C. D. Gómez-Carmona, R. Gutiérrez-Vargas, and J. Pino-Ortega, “From big data mining to technical sport reports: The case of inertial measurement units,” BMJ Open Sport and Exercise Medicine, vol. 5, no. 1. BMJ Publishing Group, p. e000565, 01-Oct-2019.

[14] F. Thabtah, L. Zhang, and N. Abdelhamid, “NBA Game Result Prediction Using Feature Analysis and Machine Learning,” Ann. Data Sci., vol. 6, no. 1, pp. 103–116, Mar. 2019.

[15] A. McCabe and J. Trevathan, “Artificial intelligence in sports prediction,” in Proceedings -

International Conference on Information Technology: New Generations, ITNG 2008, 2008, pp. 1194–1197.

[16] B. Min, J. Kim, C. Choe, H. Eom, and R. Ian, "A Compound Framework for Sports Prediction: The Case Study of Football," undefined, 2007.

[17] K. Trawiński, "A fuzzy classification system for prediction of the results of the basketball games," in 2010 IEEE World Congress on Computational Intelligence, WCCI 2010, 2010.

[18] M. Haghighat, H. Rastegari, and N. Nourafza, "A Review of Data Mining Techniques for Result Prediction in Sports," Adv. Comput. Sci. an Int. J., vol. 2, no. 5, pp. 7–12, Nov. 2013.

[19] D. Miljković, L. Gajić, A. Kovačević, and Z. Konjović, "The use of data mining for basketball matches outcomes prediction," in SIISY 2010 - 8th IEEE International Symposium on Intelligent Systems and Informatics, 2010, pp. 309–312.

[20] C. K. Leung and K. W. Joseph, "Sports data mining: Predicting results for the college football games," in Procedia Computer Science, 2014, vol. 35, no. C, pp. 710–719.

[21] I. Bhandari, E. Colet, J. Parker, Z. Pines, R. Pratap, and K. K. Ramanujam, "Advanced scout: Data mining and knowledge discovery in NBA data," Data Min. Knowl. Discov., vol. 1, no. 1, pp. 121–125, 1997.

[22] Z. Ivanković, M. Racković, B. Markoski, D. Radosav, and M. Ivković, "Analysis of basketball games using neural networks," in 11th IEEE International Symposium on Computational Intelligence and Informatics, CINTI 2010 - Proceedings, 2010, pp. 251–256.

[23] S. Kampakis and A. Adamides, "Using Twitter to predict football outcomes," Nov. 2014.

[24] C. Peace and E. Okechukwu, "An Improved Prediction System for Football a Match Result," 2014.

[25] A. Khanteymoori, E. Davoodi, and A. R. Khanteymoori, Horse racing prediction using artificial neural networks Detection of minimum number of driver genes in gene regulatory networks for applying control signals to control the network View project Knowledge-based and Parallel Gene Network Reconstruction View project Horse Racing Prediction Using Artificial Neural Networks. 2010.

[26] S. Serwe and C. Frings, "Who will win Wimbledon? The recognition heuristic in predicting sports events," J. Behav. Decis. Mak., vol. 19, no. 4, pp. 321–332, Oct. 2006.

[27] K. Ravichandran, L. Gattani, A. Nair, and B. Das, "A Novel Graph Based Approach to Predict Man of the Match for Cricket," in Communications in Computer and Information Science, 2020, vol. 1241 CCIS, pp. 600–611.

[28] N. C. Oza, "Online bagging and boosting," in Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, 2005, vol. 3, pp. 2340–2345.

[29] J. Tang, C. Deng, and G. Bin Huang, "Extreme Learning Machine for Multilayer Perceptron," IEEE Trans. Neural Networks Learn. Syst., vol. 27, no. 4, pp. 809–821, Apr. 2016.

[30] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," Mach. Learn., vol. 37, no. 3, pp. 277–296, Dec. 1999.

[31] J. Otero and L. Sánchez, "Induction of descriptive fuzzy classifiers with the Logitboost algorithm," Soft Comput., vol. 10, no. 9, pp. 825–835, Jul. 2006.

[32] A. Nurzahputra, M. A. Muslim, and B. Prasetiyo, "Optimization of C4.5 algorithm using meta learning in diagnosing of chronic kidney diseases," in Journal of Physics: Conference Series, 2019, vol. 1321, no. 3, p. 32022.

[33] F. Hutter, K. Leyton-Brown, C. Thornton, and H. H. Hoos, "Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms Computer Go View project Sparkle: A PbO-based Multi-agent Problem-solving Platform View project Auto-WEKA: Automated Selection and Hyper-Parameter Optimization of Classification Algorithms," 2012.

[34] A. Afshar, A. Zahraei, and M. A. Mariño, "Large-Scale Nonlinear Conjunctive Use Optimization Problem: Decomposition Algorithm," J. Water Resour. Plan. Manag., vol. 136, no. 1, pp. 59–71, Jan. 2010.

[35] C. Kingsford and S. L. Salzberg, "What are decision trees?," Nature Biotechnology, vol. 26, no. 9. Nature Publishing Group, pp. 1011–1012, Sep-2008.

[36] C. Chen, G. Zhang, J. Yang, J. C. Milton, and A. D. Alcántara, "An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier," Accid. Anal. Prev., vol. 90, pp. 95–107, May 2016.

[37] E. J. Alqahtani, F. H. Alshamrani, H. F. Syed, and S. O. Olatunji, "Classification of Parkinson's Disease Using NNge Classification Algorithm.," in 21st Saudi Computer Society National Computer Conference, NCC 2018, 2018.

[38] Z. Muda, W. Yassin, M. N. Sulaiman, and N. I. Udzir, "Intrusion detection based on K-means clustering and OneR classification," in Proceedings of the 2011 7th International Conference on Information Assurance and Security, IAS 2011, 2011, pp. 192–197.

[39] Y. Cao and J. Wu, "Dynamics of projective adaptive resonance theory model: The foundation of PART algorithm," IEEE Trans. Neural Networks, vol. 15, no. 2,

pp. 245–260, Mar. 2004.

[40] D. H. Toneva, S. Y. Nikolova, G. P. Agre, D. K. Zlatareva, V. G. Hadjidekov, and N. E. Lazarov, “Data mining for sex estimation based on cranial measurements,” Forensic Sci. Int., vol. 315, p. 110441, Oct. 2020.

[41] “Predictive Analysis in Agriculture to Improve the Crop Productivity using ZeroR algorithm | SCIA.” [Online]. Available:http://www.hindex.org/2016/article.php?page=172. [Accessed: 09-Jan-2021].

[42] A. Mazidi, M. Fakhrahmad, and M. Sadreddini, “A meta-heuristic approach to CVRP problem : local search optimization based on GA and ant colony,” J. Adv. Comput. Res., vol. 7, no. December, pp. 1–22, 2016.

[43] A. Mazidi, M. Mahdavi, and F.Roshanfar, “An autonomic decision tree□based and deadline□constraint resource provisioning in cloud applications” Concurrency and Computation: Practice and Experience., 2021, e6196.