International Journal of Information Technology and Applied Sciences
ISSN (2709-2208)
Int. J. Inf. Tec. App. Sci. 3, No.3 (July-2021)
https://doi.org/10.52502/ijitas.v3i3.114

Machine Learning for sport results prediction using algorithms

Said Lotfi[1], Mohamed Rebbouj[2]

[1] *Multidisciplinary Laboratory In Education Sciences and Training Engineering (LMSEIF). Sport Science Assessment and Physical Activity Didactic. Normal Superior School (ENS), Hassan II University of Casablanca, BP 50069, Ghandi, Morocco*
ORCID ID: https://orcid.org/0000-0002-0008-6145
[2] *Multidisciplinary Laboratory In Education Sciences and Training Engineering (LMSEIF). Sport Science Assessment and Physical Activity Didactic. Normal Superior School (ENS), Hassan II University of Casablanca, BP 50069, Ghandi, Morocco*

Abstract: This paper describes the use of machine learning in sports. Given the recent trend in Data science and sport analytics, the use of Machine Learning and Data Mining as techniques in sport reveals the essential contribution of technology in results and performance prediction. The purpose of this paper is to benchmark existing analysis methods used in literature, to understand the prediction processes used to model Data collection and its analysis; and determine the characteristics of the variables controlling the performance. Finally, this paper will suggest the reliable tool for Data mining analysis technique using Machine Learning.

Keywords: Sport results prediction, Machine learning, Data Mining.

1. INTRODUCTION

Computer science has evolved as never before to give a birth to the artificial intelligence, which is an emerging science integrating the theory, methods, techniques, models, systems, and applications of the human intelligence. This field of research includes many disciplines and fields such as Machine learning (ML). it has application in healthcare, education, agriculture, space, energy resources, industry and sport.

Sport, an activity that requires full set enumeration of parameters (data) in order to understand the game, its strategies and how to take decisions by minimizing unpredictability [1]. This type of data processing is called supervised learning since the data processing phase is guided toward the class variable while building the model [2].

Sports is a system in which multiple agents work together. Compared with a single agent, the learning space of multiple agents increases sharply as the number of agents increases, so the learning difficulty increases. Therefore, the use of machine Learning (ML) technology allow researchers to build models and simulate systems to predict results. Sport prediction is usually treated as a classification problem, with one class (win, lose, draw) to be predicted [3], so the researchers are looking to use many features including the historical performance of a team, historical matches results, and data collected on players and stuff. Therefore, the sport team managers work hardly with strategies and tactics to model the appropriate structure to gain the match; hence a correlation must be put in place to reconcile machine learning prediction outputs and sport team manager's strategy.

This paper aims to provide a critical description of the literature on Machine Learning for sport results prediction using in majority the deep learning algorithms such as the Artificial Neuronal network (ANN), this algorithm has proven to be effective in deriving highly accurate classification models in other domains [4].

This paper is organized as follows: in section 2, we describe the studies using different techniques for sport results prediction generally, then we specify about those using different algorithms in prediction with the specific data class variables. Section 3 is dedicated to a critic of conventional works on using algorithms for sport results prediction. Finally, Section 4 concludes the paper.

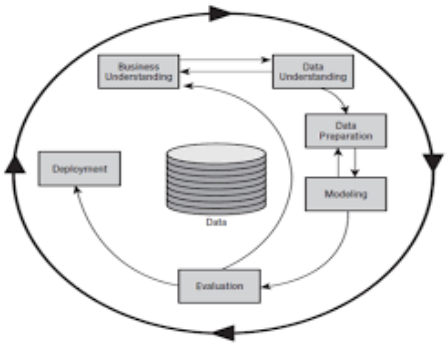
2. LITERATURE REVUE CRITICS

A. Big data

Big data is a larger Datasets that are growing in size and complexity and can require balance to avoid classification problems as a sample issue. So, the existence of imbalanced classification methods focuses on normal-sized datasets and not on the reality of the Big Data [5].

Industry and manufacturing field possess continuous and changing data related to each quarter and linked to business climate. From finance, operations, marketing, to customers and human resource management; data flows and changes continually affecting business performance. It

requires permanent follow up and deep analysis to avoid multiple risks. Big Data analysis is strongly connected with classical data analysis and mining approaches. From different sources of data, it needs to undergo sampling and querying procedures [6]. This amount of data, if well



modeled, assigns the possibility to discover the hidden patterns and give the insight to make decisions on the correct business [7] based on Key performance indicators (KPIs). This figure explains the processes used for a data mining project using the (CRoss-Industry Standard Process for Data Mining) also known as CRISP-DM.

Fig.1: tasks and outputs of the CRISP-DM Reference model

B. Machine Learning

Machine Learning (ML) is one of the intelligent methods shown promising results in the classification and prediction of structural domains [8]. ML is the subject of research which enables computers to learn without explicit programming, it gives an ability to think with high performance [9]. Moreover, ML is considered as the prominent area in identification of hidden patterns from the datasets, it focuses on the training of intelligent models so that the prediction can be accurate and fast [10].

With the use of social media, through applications on the internet, a source of rapid data generation is available for analysis. In this case, the use of an efficient algorithms is necessary to analyses this massive data [11]. In this area, machine Learning is used for teaching and/or setting up programs for Software or hardware focused on fundamental rules models of mining, forecasting and discovery. In order to obtain precision and performance from raw data sets, data researchers and studies take multiple tracked or unattended approaches [12]. As for example, and during this pandemic situation, machine learning has been applied to evaluate the mortality of Covid-19 by using many series of ML models as such as blood samples dataset to obtain what the machine learning models can provide. The important variables in mortality were considered as robust predictors [13].

C. Results prediction

Machine learning algorithms are used to establish a unified capacity prediction method. Through deep learning, the extension of ML technique, which is considered the effective prediction process; and its error rates are low when compared with other ML models [14]. Moreover, it allows concerned people to detect risk before happening, as used in the healthcare area, it shows an undertaking results to consider when dealing with patients [15] and in its application field in earth science shows prominent results to predict the accuracy of a variable [16]. However, existing prediction models do not sufficiently take user behavior into account [17].

The accuracy is related to the ML algorithm used. That is why a need for the deep learning to get effective prediction process compared to other ML models. Generally, the ANN are considered the most accurate one to predict results, can be employed wherever a relationship between explanatory variables (inputs) and explained variables (outputs) exists. It contains neurons as interconnected components that can convert a set of inputs into a preferred output [18]. The figure below shows an ANN structure.

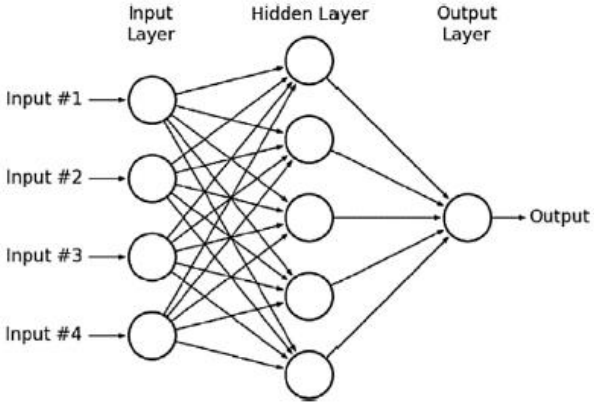


Fig.2: ANN structure with 4 input nodes in the input layer, 5 hidden nodes in the hidden layer and one output node in the output layer

The ANN model is constructed after processing the training dataset that contains the features used to build the ANN classification model. By way of illustration, weights associated with interconnected components are continuously changing to accomplish high levels of predictive accuracy. These changes are performed by the ANN algorithm to fulfill the desired model's accuracy given by the user [19].

3. SPORT RESULTS PREDICTION METHODS

Sport results prediction is related to the use of the appropriate structured experimental approach in order to obtain the best possible results based on a specific data set.

To predict in sport, we need to understand the pattern recognition, predictive systems, inference and data analytics, data collection… to prepare it, to model it, to

evaluate it and finally to deploy it; therefore, we can obtain an appropriate process model for data mining project [20].

In sport area of research, many factors can impact the matches results as such as previous goals for, goals against, home game or visitor, ranking score…. [21]. In the past 20 years, people settle for the results of the use of mathematical methods to analyze performance, recognize trends and patterns and predict results [22]. There are many indicators can be used as the judgment. Among them, accuracy is one of the most widely used indicators and considered as a key metric, it measures the ratio of correct predictions to the total number of cases evaluated. This prediction metric is based on many existing models as the Bayesian Logic (BLOG) and Markov Logic Networks (MLNs) that has shown 63% and 64% accuracy respectively for the 2006-2007 NBA season [23].

Many researchers used different algorithms to obtain the highest accuracy prediction results; For athletes' individual performances or training models to adopt, team results or even the fan's emotions towards a team or a player [24] , [25]. In this table, we present the most common algorithms used for classification and regression:

Regression algorithms	Classification algorithms
Simple linear regression	Logistic regression
Multiple linear regression	K nearest Neighbors
Polynomial linear regression	Support vector machine
support vector regression	Naïve bayes
Decision tree regression	Decision tree classification
Random forest regression	Random forest classification
Neural network	Neural network

Tab.1: commonly used algorithms for classification and regression.

We have to notice that in the regression studies, the outputs are generally numbers while in the classification problems, the outputs are class label (see fig.2). In regression the machine learning model comes up with a generalized function that approximately learns the trend of data. However, in classification, the machine learning model comes up with a generalized function that approximately divides the data into different classes that exists. This is known as decision boundary.

We can find more algorithms used to predict results or patterns to follow in training for example, such as:

A. C4.5 and C5.0 Algorithms

C4.5 and updated C5.0 version build decision trees from a set of training data using the concept of information entropy. They are based on ID3 (Iterative Dichotomiser 3). Both algorithms were proposed by Quinlan [26].

B. KStar

KStar is an instance-based classifier using an Entropic Distance Measure. It provides a consistent approach to handling of symbolic attributes, real valued attributes, and missing values [27].

C. Meta END

The main idea of meta-classification is to represent the judgment of each classifier (SVM based) for each class as a feature vector, and then to re-classify again in the new feature space. The final decision is made by the meta-classifiers instead of just linearly combining each classifier judgment [28].

D. HyperPipe

HyperPipes is a very simple algorithm that constructs a hyperpipe for every class in the data set; each hyperpipe contains each attribute-value found in the examples from the class it was built to cover. An example is classified by finding which hyperpipe covers it the best.

E. JRip

JRip is the WEKA version of RIPPER. RIPPER is a rule-based learner that builds a set of rules that identify the classes while minimizing the amount of error. The error is defined by the number of training examples misclassified by the rules.

The following table (Tab.2) explain the measurement instruments used in ML to analyze data of a study sample. The process is based on defining the variables used to reach the target variables with or without a software, and then discuss the results accuracy for each algorithm used.

Tab.2: algorithm measurement methods, variables features and results according to sport studies.

Field	Author, Date	Measurement				Variable features		Software used	Results
		Measurement instrument	Learning	% prediction	Sample feautures	Target	Used		
Sport	Adam Maszczyk, Artur Gołaś, Przemysław Pietraszewski, Robert Roczniok, Adam Zając, Arkadiusz Stanula, 2014	1/non-linear regression models. 2/Multilayer Perceptron neural models. 3/Levenberg-Marquardt algorithm	Supervised	$p \leq 0.05$. Absolute network error 16.77 m Absolute regression error 29.45 m	116 javelin throwers 18±0.5	the distance of the javelin throw from a full run-up.	CSATS expressed in seconds. SPAT expressed in meters. SPAM GP: Grip power.	STATISTICA 9.1; STATISTICA Neural Networks module (Release 4.0E); Excel 2010 from Microsoft Office 2010.	the created neural models offer much higher quality of prediction than the nonlinear regression model (absolute network error 16.77 m versus absolute regression error 29.45 m).
	Li Yongjun, Wang Lizheng, Li Feng, 2021	1/Data envelopment analysis (DEA) model with data-driven technique. 2/Logistic regression analysis and DEA-based player portfolio analysis.	Supervised	82.15%. The McFadden R^2 0.3791, and LR statistic 150.6680	2015–16 regular season based on a four-season dataset from the 2011–12 season to the 2014–15 season	"Win or Lose " results.	Home or visitor; Ranking score; Team in a conference (stronger/weak).	(DMUs) a mathematical programming method originally pro- posed to evaluate the relative performance of peer decision making units	DEA-based data-driven approach can predict the sports team performance very well and can also provide interesting insights into the performance prediction problem.
	Gabriel Fialho, Aline Manhães, João Paulo Teixeira, 2019	Litterature review: 1/ Bayesian and Logistic Regression; 2/ Artificial Neural Networks (ANN); 3/ Support Vector Machine (SVM); 4/ Fuzzy Logic and Fuzzy Systems;	Supervised	1/ Ranked Probability Score (RPS) = 0.208256 99.06% relative performance. 2/ 85% accuracy using Logistic Regression. 3/ 61.4% accuracy. 4/ 91% accuracy predicting.	1/ 216.743 match. 2/ 10 matches played in the 2014 and 2015 English Premier League season. 3/ Data of 10 Years Ukranian FB Cham. 4/ Learning sample: 1056 matches. Testing samples : 350 matches.	Predict the results of soccer matches	Season, country and league, date on which the game was played, name of the home and away team, number of goals scored by home and away team, goal difference and outcome of the game in terms home win (W), draw (D) and away win (L).		Based on the analysis of the related works, it was proposed a model and feature selection for predicting soccer outcomes. Furthermore, these systems can also be useful to make profit in betting industries, using science on our side.

Field	Author, Date	Measurement				Variable features		Software used	Results
		Measurement instrument	**Learning**	**% prediction**	**Sample feautures**	**Target**	**Used**		
Sport	Fangyao Liu, Yong Shi, Lotfollah Najjar, 2017	1/ Design of experiments DOE or DOX (systematic method-DOE).	Supervised	Low to predict the winning score differences, it just used to predict which team will win.	Data of 30 NBA teams of the 2015-2016 season	Win / lose	Home or visitor; Ranking score; Team in a conference (stronger /weak).	Minitab 17 edition DOE software	only two factors (Home and Ranking) have a significant impact on the basketball results. the more ranking difference, the bigger effect the team will have. ranking score difference is greater than four, then this factor has a higher impact than the second factor (home game or not) on team's win results.
	Edward Mezyk, Olgierd Unold, 2011	1/ Fuzzy logic; 2/ C4.5 and C5.0 Algorithms; 3/ SVM; 4/ Naïve Bayes; 5/ KStar; 6/ Meta END; 7/ HyperPipe; 8/ JRip;	Supervised	68.66% accuracy	Class II level swimmers with 7 years of training.	Trainee's feelings control in water.	Sexe, age, Hour of beginning of sleep, Amount of the sleep, pulse Before- Pulse After, Pulse vertical- pulse After, Stimulus of the day, Time of the stimulus, Water training, Land Trainning, Km during the swim, feelings	C# language under .NET 2.0.	The accuracy of the result of compared methods is significantly lower than the accuracy of fuzzy rules obtained by a method presented in this study (paired t-test, $P < 0.05$).
	Purucker M.C., 1996	ANN with backward-propagation (BP)	Unsupervised (clustring)	61% accuracy.	8 rounds of National Football league (NFL).		Yards gained, rushing yards gained, turnover margin, time of possession, beeting line odds.		The Backward-Propagation algorithm was found to be the most effective approach to predict results even if the study is limited in the use of a small number of features.

Field	Author, Date	Measurement				Variable features		Software used	Results
		Measurement instrument	Learning	% prediction	Sample feautures	Target	Used		
Sport	Kahn J. 2003	ANN with backward-propagation (BP) multi-layer perceptron.	Supervised	75% accuracy.	208 matches in the NFL 2003 season	Win (+1) or lose (-1).	Total yardage differential, rushing yardage differential, turnover differential, away team indicator, home team indicator.	The script: processStatistics.pl	Improved prediction data would lead to more accurate predictions. Still, some fallibility exists in the network, as the classification rate is not 100%.
	McCabe A. & Trevathan J., 2008	Multi-Layer perceptron algorithm; Backward-Propagation algorithm; Conjugative-Gardient algorithm.	Supervised	67.5%	Data from 2002 NFL, (AFL) Australian Rules Football, Super Rugby and English premier League Foorball (EPL).	Predict the winner	Points for, points against, overall performance, home performance and away performance, performance in previous game, performance in previous n games, team ranking, points for in previous n games, points against in previous n games, location, player availability		The multi-layer perceptrons used were able to adapt very quickly and perform well despite the limited information and the outside influences not included in the feature set.
	Davoodi E., Khanteymoori A., 2010	Gardient-descent BP; Gardient-descent with a momentum parameter (BPM); Levenberg-Marquadt (LM); Conjugate gardient-descent (CGD)	Supervised	77% with BP	Data of 100 races of 2010	The horse finishing time	Horse weight, type of race, horse trainer, horse jockey, number of horses in race, race distance,track condition and weather.		With 400 epochs, the BPM (with momentum parameter of 0.7) and the BP algorithm were most effective at predicting the winner of the race.
	Tax, N. & Joustra, Y., 2015	Naive Bayes LogitBoost (with Decision Stump) Neural Network (Multilayer Perceptron) Random Forest CHIRP FURIA DTNB Decision tree (J48) Hyper Pipes	Supervised	54.7% accuracy for naïve Bayes (used with a 3-component PCA), and the ANN (used with a 3 or 7-component PCA). 55.3% with the FURIA classifier. 56.1% for LogitBoost with ReliefF attribute selection.	Dutch football competition data from 13 years	Match outcome	Previous performance in current season, Performance in earlier encounters, Streaks, Managerial change, Home advantage, Matches with special importance, Fatigue, National team players, Promotion to higher league, Expert predictions, Football skills, Strategy, Travel Distance, Betting odds, Club budgets, Availability of key players,	WEKA Machine Learning Toolkit	Betting odds alone can be a reasonable predictor of match outcome.

4. CONCLUSION

This article critically analyses some recent research on sport prediction that have used several algorithms. Due to the nature of the field of study, the table 2 has shown that we can provides many alternatives to reach the best accuracy; on which relays the viability of any ML prediction model.

Traditionally, mathematical, and statistical models were used to predict the matches results. However, predicting the score is still difficult compared to Win or Lose prediction. That's why this study has compared many related works concerning results predictions across different sports. But the need for more accurate models last due to the high volumes of data in sport, which can be global for a team historic or specific taking into consideration each element (Player, coach strategy, training, weather condition…) as a key indicator. Therefore, ML seems an appropriate methodology for sport prediction since it generates predictive models that can predict match results using predefined features in a historical dataset.

REFERENCES

[1] «Sports analytics Evaluation of basketball players and team performance,» vol. 93, n° %1101562, 2020.

[2] F. Thabtah, S. Hammoud et H. Abdeljaber, «Parallel associative classification data mining frameworks based mapreduce,» *Parallel Processing Letters*, vol. 25, n° %102, 2005.

[3] D. Prasetio et D. Harlili, «Predicting football match results with logistic regression,» *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, 16-19 Aug. 2016.

[4] R. M. Mohammad, F. Thabtah et L. McCluskey, «An Improved Self-Structuring Neural Network,» chez *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Auckland, 2016.

[5] M. Juez-Gil, Á. Arnaiz-González, J. J.Rodríguez et C. García-Osorio, «Experimental evaluation of ensemble classifiers for imbalance in Big Data,» *Applied soft computing*, vol. 108, n° %1107447, 2021.

[6] M. Chen, S. Mao et Y. Liu, « Big Data: A Survey,» *Mobile Networks and applications* , vol. 19, pp. 171-209, 2014.

[7] X. Yang, «Business big data analysis based on microprocessor system and mathematical modeling,» *Microprocessors and Microsystems* , vol. 82, n° %1103846, 2021.

[8] D. Yang, «Online sports tourism platform based on PFGA and machine learning,» *Microprocessors and Microsystems*, vol. 80, n° %1103584, 2021.

[9] M. K. Vanam, B. A. Jiwani, A. Swathi et V. Madhavi, «High performance machine learning and data science based implementation using Weka,» *Materials today: Proceedings*, 2021.

[10] G. Arunakranthi, B. Rajkumar, V. C. S. Rao et A. Harshavardhan, «Advanced patterns of predictions and cavernous data analytics using quantum machine learning,» *Materialstoday: Proceedings*, 2021.

[11] B. T.K., C. S. R. Annavarapu et A. Bablani, «Machine learning algorithms for social media analysis: A survey,» *Computer science review*, vol. 40, n° %1100395, 2021.

[12] G. Ramkrishna Reddy, C. Vani et E. Hari Krishna, «Implementation of MLDB as high performance machine learning database for high performance applications,» *materials today: proceedings*, 2021.

[13] M. Smith et F. Alvarez, «Identifying mortality factors from Machine Learning using Shapley values – a case of COVID19,» *Expert Systems with Applications*, vol. 176, n° %1114832, 2021.

[14] A. A. Rasheed, «Improving prediction efficiency by revolutionary machine learning models,» *Materials today: proceedings*, 2021.

[15] S. L. Kausch, J. R. Moorman, D. E. Lake et J. Keim-Malpass, «Physiological machine learning models for prediction of sepsis in hospitalized adults: An integrative review,» *Intensive and critical care nursing* , n° %1103035, 2021.

[16] S. Chanda, M. Raghucharan, K. Reddy, V. Chaudhari et S. N. Somala, «Duration prediction of Chilean strong motion data using machine learning,» *Journal of south american earth science*, vol. 109, n° %1103253, 2021.

[17] K. Amasyali et N. El-Gohary, «Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings,» *Renewable and Sustainable Energy Reviews*, vol. 142, n° %1110714, 2021.

[18] I. H. Witten et F. Eibe, «Data mining: practical machine learning tools and techniques with Java implementations,» *ACM SIGMOD Record*, vol. 31, n° %11, 2002.

[19] R. P. Bunker et F. Thabtah, «A machine learning framework for sport result prediction,» *Applied Computing and Informatics*, n° %115, pp. 27-33, 2019.

[20] C. Shearer, «The CRISP-DM Model: The New Blueprint for Data Mining,» *Journal of data warehousing*, vol. 5, n° %14, 2000.

[21] C. T. A. S. Ley, «Analytic Methods in Sports – Using Mathematics and Statistics to Understand Data from Baseball, Football, Basketball, and Other Sports,» *Stat Papers* , n° %162, pp. 1091-1092, 2020.

[22] C. T. A. S. Ley, «Analytic Methods in Sports – Using Mathematics and Statistics to Understand Data from Baseball, Football, Basketball, and Other Sports.,» *stat papers*, n° %162, pp. 1091-1092, 2020.

[23] D. Orendorff et T. Johnson, «First-Order Probalistic Models For Predecting The Winners Of Professional Basketball Games,» *ResearchGate*, 2008.

[24] X. Gong et Y. Wang, «Exploring dynamics of sports fan behavior using social media big data - A case study of the 2019 National Basketball Association Finals,» *Applied Geography* , vol. 129, n° %1102438, 2021.

[25] Y. Yu et X. Wang, «World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets,» *Computers in human behavior* , vol. 48, pp. 392-400, 2015.

[26] J. R. Quinlan, C4.5: Programs for Machine Learning, San Mateo California: Morgan Kaufmann Publishers, 1993.

[27] J. Cleary et L. Trigg, «K*: An instancebased learner using an entropic distance measure.,» chez *Proceedings of the 12th International Conference on Machine Learning*, Shenzhen China, 1995.

[28] W.-H. Lin, R. Jin et A. G. Hauptmann, «Meta-Classification of Multimedia Classifiers.,» chez *International Workshop on Knowledge Discovery in Multimedia and Complex Data (KDMCD 2002), in conjunction with the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-02), Taipei, Taiwan, May 6-8*, Taipei, Taiwan, 2002.