Optimal Predictive Model for FIFA World Cup

Durgansh Sharma, Vaibhava Sharma

Abstract*: In this research, the proposed deep learning network uses H2O framework using Multi-layer Feed Forward Network. Statistics of the 592 FIFA world cup Matches collected and used to train Naïve Bayes, k-NN and deep learning networks. As observed, the efficiency of deep learning based network is superior as compared with Naïve Bayes and k-NN for prediction of sports especially football/ soccer using dataset of Match result and status. The prediction observed with 97.68% accuracy while keeping the training-testing ratio as 20:80 using Deep Learning using "Max-out with Dropout" activation function. The model k-NN and Naïve Bayes were trained with 80:20 training-testing ratio.

***Index Terms*: Predictive Analytics, Deep Learning, Naïve Bayes, k-NN, FIFA.**

I. INTRODUCTION

An international football tournament was only a fantasy for much of the world population, since the First World War embarked in 1914. 16 years passed, and in 1930 the world witnessed its first ever Football World Cup held in Uruguay, as de-cided by FIFA - Fédération Internationale de Football Association – the governing body of world football. FIFA has been there since 1904, guiding all the confederations and country-based organizations, in promoting footballing culture all over the world. It has been famously organizing the world-popular international tournament - World Cup - both for Men, from 1930 and for Women from 1991. FIFA's headquarters are in Zurich, Switzerland and consists of 211 member association countries. Gianni Infantino is the current FIFA president. Under his reign, FIFA earned USD 734 million as revenue in year 2017. Many other tournaments organized by FIFA, like Club World Cup, U-20 & U-17 World Cups and Confederations Cup, but the World Cup is the most prestigious of them all. Recently, in June-July of 2018, World Cup planned in Russia; organized in every 4 years. The FIFA World Cup 2022 will held in Qatar and the 2026 version in United States, Canada & Mexico.

A. Predictive Sports Analytics

Talking about future tournaments, the term prediction comes to mind, at the very first thought. The big question, who will win the next World Cup? Arises. Most of the people just guess and move on, but not in our case. An efficient usage of predictive analytics would definitely help us to solve this question and probabilistically reach closest to the reality. Predictive Analytics is an uprising concept, which uses various techniques such as Data Mining, Machine Learning and Modelling, to train itself from the past records, and analyze them to predict the future results. Data Analytics consists of three types of models Descriptive, Predictive and Decision Models. Descriptive models are applicable to the problems of classifications, where certain data requires grouping into certain clusters. Decision Models makes relations between the decisional elements – Known Data, Decision and Forecast Results. Our case is an example of the Predictive Model, which relates one or more known attributes of the unit and the performance of a unit in a sample. Application of Predictive Analytics in real-life are vast including CRM (Customer Relationship Management), Clinical DSS (Decision Support System), Payment Collection Process, Project Risk Management, Sports analytics and many more. We will use this Predictive Model, in testing the modalities of foretelling the FIFA World Cup 2018 Winner with nearest predictive model, while using certain known attributes.

B. Deep Learning for Sports Prediction

Deep learning is a type of machine learning, which is usually implemented using neural network. Unlike the earlier neural networks, deep learning has better scalability and stability with sports related datasets. It is becoming popular due to highest predictive accuracy. It is based on Multi-layer feed forward networks initiated with input layers for matching the features of the dataset, followed by various layers of non-linearity and terminating with linear regression or layer of classification to map with outcome related space. The basic framework of multi-layer neural networks supports in accomplishing deep learning tasks. Accurate sports predictions, based on high level of non-linear outcomes, which require deep learning architecture based models comprised of hierarchical feature extraction. The ability of Deep Learning model is to acquire needful exemplifications of raw data and to exhibit high performance on complex datasets, in our case its FIFA world cup statistics.

C. k-NN and Naïve Bayes for Sports Predictions

Sports prediction based on various independent variables, incorporating their individual impact over the result of any Match. While searching for a classifier we narrowed down to k-NN and Naïve Bayes classifiers.

k-Nearest Neighbors (k-NN) is a non-parametric method through which any kind of classification and regression problem can be solved. This algorithm is the simplest of all kinds of Machine Learning techniques.

Revised Manuscript Received on December 22, 2018.

Durgansh Sharma, Associate Professor, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India

Vaibhava Sharma, Accounts Executive, Finance Team, Next Level Business Services Inc., Noida, India.

Naïve Bayes classifier is a probabilistic classifier while applying Bayes theorem, simply along with the strong independence assumptions; its advantage is need of small amount of training data for the estimation of means and variances of defined variables, which are necessary for classification. As independent variables are assumed and the variances of the variables for each label is required to be calculated and not the entire covariance matrix**.**

II. LITERATURE REVIEW

Huang, K.-Y., & Chang, W.L [1] have used Neural Networks to predict the 2006 FIFA World Cup Match results. The model used was MLP – Multilayer Perceptron with a back propagation-learning rule implied upon it. Based on this model, the accuracy of prediction was 76.9% excluding the matches drawn. There were eight different features used to predict the outcome of every single match. Certain experimental results showed the architecture for MLP as 8-11-1, i.e. 8 Inputs, 11 Hidden Notes and 1 Output. There were 17 statistical terms, though they used eight important terms. Relative Ratio was an input feature. It normalized Training and Prediction samples. Training samples included only those teams, which won or lost all three matches in the group stage. It triggered our thought process to use ANN for our study.

Boulier, B. L., & Stekler, H. O. [2] used a different kind of technical strategy to get the predictive outcomes of the Matches from 1994-2000 of the NFL – National Football League. They stated that there are some kind of Power Scores in NFL, based on which the teams are given rankings. They compared the predictions generated from Probit Regressions using Power Rankings published in The New York Times, with the forecasts from Naïve Model, the Betting Market and the verdict of the Sports Editor of The New York Times. The Power Scores published by The New York Times generated using Win-Loss Record of the team, Home or Away Match-es, Winning Margin and the Opponent Quality. Recent Matches got more weight-age. Since Boulier, B. L., & Stekler, H. O. did not have access to these minute de-tails, at that time; they could not calculate these predictive Power Scores systematically, on their own end. Therefore, using the available resources they concluded that the accuracy of the Power Scores were inferior to the Naïve Model and Bet-ting Market but, slightly better than that of the Sports Expert. Naïve Model does not require a maximum of sports knowledge and predictive expertise, thus it just outclassed the Sport Expert predictions by little margin, when whole of the data was tested.

While performing their research in the real-time visual analytics of soccer data, Janetzko, H., Sacha, D., Stein, M., Schreck, T., Keim, D. A., & Deussen, O. [4] represents the usage of KNIME as the data-mining framework and the application of classifiers such as Neural Networks, Decision Trees and Support Vector Machines. They used N-Cross-fold Validation to evaluate the classifiers. This presents a flexible and elastic layer-based system, which allowed in-depth analysis. Evaluation of this approach was through comparison with real-world soccer matches. They used a 33% data sample as the training set, remaining being the testing data. The usage of defined ratio for the data sample has provided a clue towards our study.

Jiang, Wenhua, and Cun-Hui Zhang [5] measured the performance of some Empirical Bayes methods for predicting the batting averages by using the 2005 Major League Baseball dataset. They considered utilizing both Homoscedastic and Heteroscedastic Partial Linear Models. As mentioned by them they used Heteroscedastic Model because the explanation of the unknown matter is partial by a linear effect. The errors are Normal with zero mean variances, measured distinctively. In applying this model to Baseball, the linear component includes Pitcher-Non Pitcher Group and effect of number of bats. The main reason of using Empirical Bayes was the possibility of a greater reduction of the compound risk in a group of statistical decisional problems, with the observations combined from all the problems. Since, they used a model from the Bayesian Family; we got an idea to use one of the models from the Bayesian Family itself.

Loeffelholz, Bernard, Earl Bednar, and Kenneth W. Bauer [6] applied Neural Networks for creating a predictive modelling structure for foretelling the results of NBA matches. A pool of 650 2007-08 Season NBA matches used as the dataset for the model. They used Feed-Forward, Radial Basis, Generalized and Probabilistic Neural Networks Fusion. For selecting the inputs for the model, they opted Signal-to-Noise Ratios and Experts’ Feedback. This model showed 74.33% accuracy, whereas the experts were correct for 68.67% of the time. They took the 620 NBA Matches as the training dataset and the remaining 30 being the Testing dataset. The inputs included Offensive and Defensive Rebounds, Field Goal%, Steals, Blocks, Points and some more important statistical points, which affects the out-come of the match. There were two techniques employed, to predict the future Matches. First one being the effect of Home and Away Fixtures. Second on being the results of the past five matches played, this would also cover the teams on Hot/Winning or Cold/Losing Streak. Finally the data taken from the experts’ opinions. The four types of Neural Networks used, were finally combined and put into a Bayesian Model, which predicted the results. Since, they also put the Bayesian Model into use along with the Neural Networks, thus we got inspired and applied both the kind of techniques in our model.

McCabe, A., & Trevathan, J. [7] also applied artificial intelligence to predict soccer match results. They used a model, which is a form of Multi-Layer Perceptron and captured the quality of sporting teams through various features and attributes. According to them, predicting the match result for any sport is a less traditional application of Neural Networks; however, application of some concepts of machine learning made. They used datasets from four variant of sporting tournaments, namely, Australian Football League, Super Rugby consisting of Super 12 as well as Super 14, National Rugby League and English Premier League, from 2002. They also mentioned that it is a necessity to provide a Training and Testing dataset for the model and Learning Algorithms such as Back Propagation for the model to work efficiently.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Miljkovic, D., Gajic, L., Kovacevic, A., & Konjovic, Z. [8] used Naïve-Bayes Method to predict NBA Matches. For each Match, the system calculates the spread using Multivariate Linear Regression. Formalization of this prediction scenario is set as a classification problem. An evaluation dataset of 778 Matches from NBA Season 2009/2010 used and the prediction was accurate for 67% of times. It also paves our way to use this technique as well to cross check its feasibility in our study for prediction of FIFA latest world cup result.

MINA, B. [9] also referred to the Bayesian Approach, when predicting the Match Results in any sport. They also mentioned rule-based reasoning, as one of the techniques, which seems eligible for predictive modelling. They used Bayesian Inference and Rule-based inference along with a unique Time-Series Approach for creating a predictive model. The idea of combining two different techniques for creating this model was encouraged due to the nature sport results, which are highly stochastic. Due to the rule-based approach, prediction of scarred data done without any obstacles. For example, if a match result between two teams who have played one or two, or not even a single match, will still have a prediction through this technique. They also mentioned that their framework included many factors such as Team Morale, Fatigue, Skills, etc. According to their Time-Series Approach, they divided a match into ten periods. In every single frame, they tried to narrow down the reasoning of the Team Manager, according to their performance, and attempt to refine a real-time strategy for the team. Due to the use of Bayesian Inference in such an efficient way by them, encouraged us to use the same technique for our own model.

Rotshtein, Alexander P., Morton Posner, and A. B. Rakityanskaya [10] compiled previous results of the teams to predict the result of the upcoming matches of such football teams. They used Fuzzy Knowledge Bases to identify Non-Linear Dependencies. They found simulation results through some alteration of the Tournament Data. Alteration included the selection of parameters based on which the model would give results. The parameters were of the functions of Fuzzy-term membership functions and the mixture of some genetic and neural techniques for optimization. They claimed that the other predictive methods like Regressive Analysis, Bayesian Approach and Monte Carlo Method were complex and required large testing samples to provide results. The results were also hard to understand and interpret. They mentioned the recent use of Neural Networks for such kind of a Match Prediction Model, and encouraged its use, but did not use them-selves, since it required extensive statistical data, which might not be available at that time. Because of such a statement, we got encouraged in applying one of the techniques of the Neural Networks in our own approach.

Stefani, Raymond T. [11] implied the Least Squares to obtain ratings for both the games of football and basketball for College and Professional Teams, then used to predict the outcome of the Matches. Accuracy of the model was moderate, 72% in the 3000 college football Matches, 68% in the 1000 Professional Football Matches and 69% in the 2000 college Basketball Matches. Factors affecting the prediction generated by trial and error method through hundreds of matches played in the past. The margin of winning calculated as half the difference in team ratings. A home team advantage of 2-4 points given, points being greater for the better team. Adjustment of ratings done by comparison between predicted point spread and the actual point spread, along with the application of a threshold and Factor of Multiplication. As a matter of fact, least squares was a good technique to create predictive models, but in the present time, they have be-come obsolete due to the uprising of Machine Learning and Artificial Intelligence, therefore we did not opt for this technique.

Andreas Groll et.al [12] have compared three modeling approaches namely Poisson regression models, random forest and emphasized over the usage of ranking methods approach to establish a relationship over various parameters towards team ability and their performance in the match.

Marcio Alves Diniz et.al [13] have compared different probabilistic predictive models, whereas emphasized over Bayesian multinomial – Dirichlet models for the prediction of final football matches. The comparison is done between three sophisticated models, first, The Benchmark Model, which contains a Bivariate Poisson Distribution, second, Bradley-Terry Models, uses MLE (Maximum Likelihood Estimation) and thirdly, Multinomial-Dirichlet Model.

David Dormagen [14] while designing of a simulator for FIFA world cup 2014 used different methods such as Elo rating method, FIFA rating, Soccer power index, Market Value of players participating and the Home Advantage for the Teams playing. He has created a simulator, which incorporates all of the above-stated Rating methods together and provides the positional results of a team in percentage. It also takes in the Average goals scored by a team in the matches played.

M. J. Dixon and S. G. Coles [15] also put the Poisson regression model into use, for predicting games of English League. They reviewed Bookmaker's Odds and claimed that, if used, their model could provide positive results in betting.

III. DATASET & METHODOLOGY

Dataset comprises of various attributes, we have reduced them as mentioned in Table 1 below, for the predictive analysis of FIFA world cup Matches since 1982 [3]. The reduction of the dataset processed through a method of Data Cleaning termed as Mean of Class. It removes the missing values from the dataset; hence, the data cleaned for the purpose of analysis. Statistics of the 592 FIFA world cup Matches collected and used to train Naïve Bayes, k-NN and deep learning networks.

Feature Selection – This is an important step in the process of data analysis. Also known as Variable Selection or Subset Selection, it helped us in finding out the required features out of a no. of features available in the whole dataset. It enables the model to get itself trained effectively and efficiently. We used the Best First Feature Selection type for our dataset.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

It is a searching algorithm, which expands the most important node of a graph, selected through a specified rule.

Rapidminer Studio – We have used rapidminer studio for calculating predictive results for our three models. The framework provided by this software is user-friendly and easy-to-use, thus making it easier for us to have tedious calculations done in lesser than a minute.

Table 1: Dataset attributes

S.No.	Dataset - Attributes	Short Form
1	Date of Match	Date
2	Home Team	Home
3	Away Team	Away
4	Home Team Score	HTS
5	Away Team Score	ATS
6	Tournament Played	Tournament
7	City in which match played	City
8	Country in which match played	Country
9	Neutral Venue (True/False)	Neutral
Result for the match played		
1	Result – Home Team Win	H
2	Result – Away Team Win	A
3	Result – Match Drawn	D

A. k-NN:-

k-NN is a sort of instance-based learning, which approximates the function locally. The purpose of k-NN being a supervised learning algorithm is to use the dataset where the data points divided into various classes to predict the classification of a new sample point. K-NN being non-parametric, it does not make any assumptions on the given dataset distribution. Football being an instance-based sport, k-NN is one of the available techniques, whose implication for prediction of a Match is comparatively easy.

At k=4 and keeping the Measure type as Mixed Measures with Mixed Euclidean Distance, along with the training data and testing data in the ratio 0.8:0.2.

We kept the value of "k" equal to four, along with using the Measure type as Mixed Measures with Mixed Euclidean Distance. The ratio of training and testing data is 0.8:0.2. This was the most successful ratio, in our case, based on the accuracy we got after applying the model along with this ratio. The model shown in Figure 1 is the process used for prediction using k-NN.

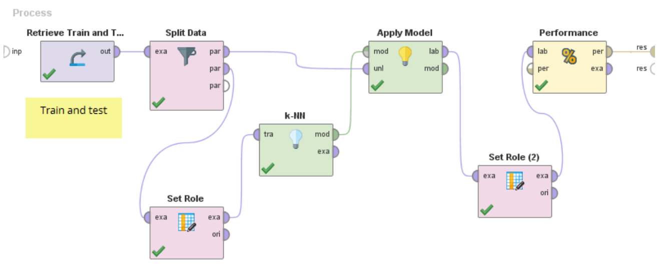

Figure 1: k-NN Model

B. Naïve-Bayes:-

Briefly, Naïve-Bayes is a conditional probability based model. Bayesian Classifiers assume that the value of any feature is independent of any other value in the feature subset of the dataset. Despite being a model consisting of oversimplified assumptions, it has worked well in many complex problems. Scalability of the Naïve Bayes classifiers are higher, therefore requiring a number of linear parameters in the variables of a learning problem. Thus, it is suitable for Sports Prediction, as every sport includes many independent variables.

Similar to k-NN, we used the same, 0.8:0.2, training and testing ratio while applying this model. The model shown in Figure 2 is the process used for predicting the results. The only weakness we encountered in this model was that if the training data includes a given attribute value, which never occurs in the whole dataset, then it sets the conditional probability to zero automatically. When this value multiplied with other probabilities it also results in zero, hence the model returns with biased values. Therefore, application of a Laplace Correction is important for a stable result. By implying it, one is added to each data entry, making every null value, if any, equal to one. Adding one to those entries, which are already having some value more than zero, faces a negligible effect. Thus, it was found that applying Laplace Correction along with the Naïve-Bayes model is important in most cases, like ours.

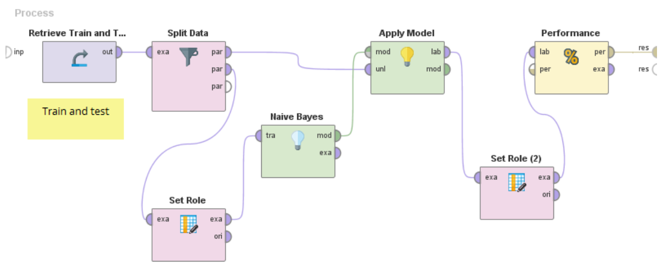

Figure 2: Naive-Bayes Model

C. Deep Learning:-

We applied Deep Learning a subset of Machine Learning, using an open source platform with H2O framework, having various inherent features of being in-memory, distributed, fast, and scalable machine learning used for predictive analytics. The activation function is the function used by the neurons in the hidden layer, in this case we have applied Maxout Activation Function along with the Range of Maxout with Dropout. The Maxout Model is a feed-forward architecture that implies a newer type of activation function: Maxout Unit. Whenever, an input like x ϵ R^d(x is a hidden layer's state) delivers the maxout hidden layer results mentioned in the following function:-

$$h_i(x) = \max Z_{ij}\ ;\ j\epsilon[1,k] \qquad (1)$$

Where,

$Z_{ij} = x^T W_{...ij} + b_{ij}$ and $W\epsilon\ \mathbb{R}^{d\times m\times k}$ and $b\epsilon\ \mathbb{R}^{m\times k}$ are learned parameters.

When we train the maxout activation function with dropout range, an elementwise multiplication performed with the dropout mask, prior to the multiplication by the weights in all cases–we do not drop inputs to the max operator. Maxout networks also learn about the activation function of each hidden unit of each hidden layer.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

It is an update to the famous Rectifier Function. The no. of layers used are four, namely, Input Layer, 2 Hidden Layers with neuron size of 50 per layer and lastly the Output Layer. Since, our data is of small-scale, we used reproducibility of data, which uses only one thread for whole process. Local Random Seed used for randomizing the data so that we could get better results. The auto-tuning mode used for training samples per iteration. Using this, each row is immediately used to update the model with Stochastic Gradient Descent. The dropout ratio is 0.5-0.5. The epochs is 100, i.e. the dataset iterated for 100 times. The model shown in Figure 3 is the process used for prediction using the Deep Learning Model. Whereas, this model was trained with 20% of dataset and tested on the other 80%.

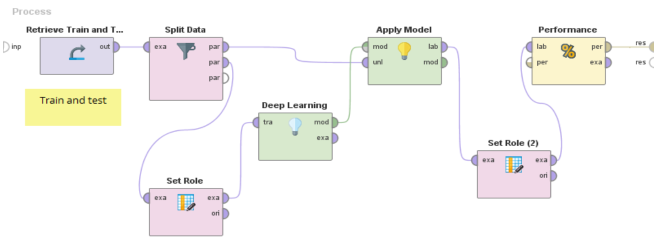

Figure 3: Deep Learning Model

IV. RESULTS

A. *k-NN*

Table 2: Confusion Matrix for k-NN Model

Predicted Result	True Result		
	A	H	D
A	**19**	12	13
H	17	**30**	16
D	2	5	**4**

Accuracy: 44.92%
Kappa: 0.141

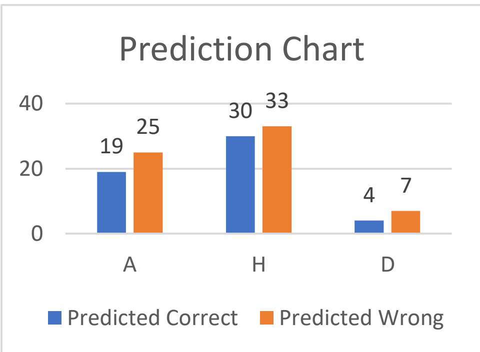

Figure 4: Prediction Chart for k-NN model

B. **Naïve-Bayes**

Table 3: Confusion Matrix for Naive-Bayes Model

Predicted Result	True Result		
	A	H	D
A	**25**	2	10
H	3	**33**	11
D	10	12	**12**

Accuracy: 59.32% , Kappa: 0.383

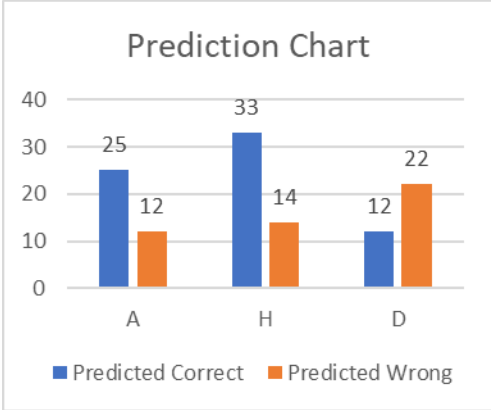

Figure 5: Prediction chart for Naive Bayes model

C. **Deep Learning**

Table 4: Confusion Matrix for Deep Learning Model

Predicted Result	True Result		
	A	H	D
A	**139**	0	4
H	0	**216**	4
D	0	3	**108**

Accuracy: 97.68%
Kappa: 0.964

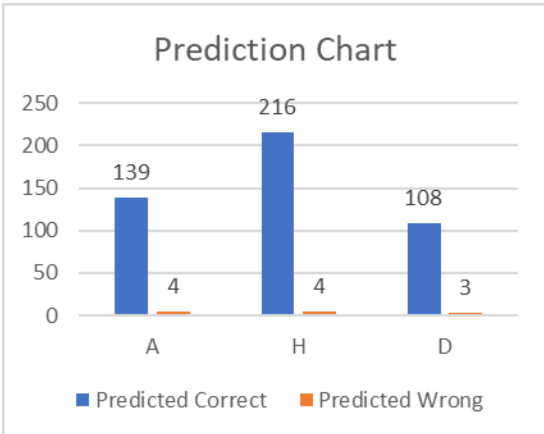

Figure 6: Prediction chart for Deep Learning model

V. CONCLUSION

According to the above stated results, it can be concluded that Deep Learning with H2O framework is far more superior to Naïve-Bayes and k-NN Models, in terms of accuracy. Referring to Table 4 and Figure 6, Deep Learning is 97.68% accurate in its predictions. Second to it is the Naïve-Bayes Model, with Laplace correction, having 59.32% accuracy, as seen in Table 3 and Figure 5. The least accurate model in our case is the k-NN Model, with "k" equals to four, barely reaching the halfway mark, with 44.92% accuracy, as observed from Table 2 and Figure 4.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

A graphical representation of such a comparison is shown in Figure 7. Another reason for such a powerful performance of the Deep Learning model is that it is still getting regular updates with firsthand machine learning tools and is a relatively newer version of predictive models, as compared to Naïve-Bayes and k-NN Models.

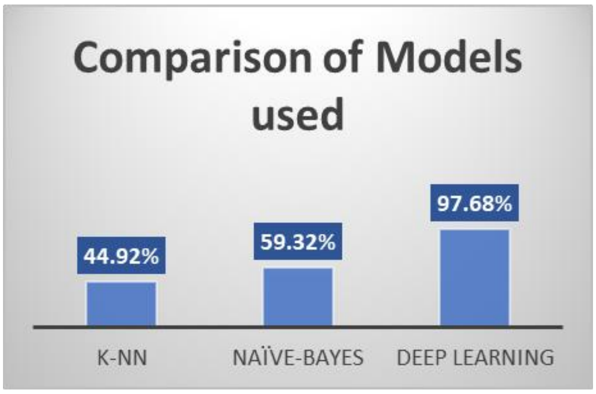

Figure 7: Comparison of Prediction Percentage of the three models

REFERENCES

1. K. Y. Huang and W. L. Chang, "A neural network method for prediction of 2006," The 2010 International Joint Conference on Neural Networks, 2010.
2. B. L. Boulier and H. O. Stekler, "Predicting the outcomes of National Football League Matches," International Journal of Forecasting, pp. 257-270, 2003.
3. Agostontorok, "Predicting the winner of the 2018 FIFA World Cup," Kaggle, 2018. [Online]. Available: https:// www.kaggle.com/agostontorok/soccer-world-cup-2018-winner/data. [Accessed 2018].
4. H. Janetzko, D. Sacha, M. Stein, T. Schreck and D. A. Keim, "Feature-driven visual analytics of soccer data," IEEE Conference on Visual Analytics Science and Technology (VAST), 2014.
5. Jiang, Wenhua and C.-H. Zhang, "Empirical Bayes in-season prediction of baseball batting averages," Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown, Institute of Mathematical Statistics, pp. 263-273, 2010.
6. B. Loeffelholz, B. Earl and B. W. Kenneth, "Predicting NBA Matches using neural networks," Journal of Quantitative Analysis in Sports, 2009.
7. A. McCabe and J. Trevanthan, "Artificial Intelligence in Sports Prediction," Fifth International Conference on Information Technology: New Generations, 2008.
8. D. Milijkovic, L. Gajic, A. Kovacevic and Z. Konjovic, "The use of data mining for basketball matches outcomes prediction," IEEE 8th International Symposium on Intelligent Systems and Informatics, 2010.
9. B. Mina, "Compound Framework for Sports Prediction: The Case Study of Football," Knowledge-Based Systems, pp. 551-562, 2008.
10. A. P. Rotshtein, P. Morton and A. B. Rakityanskaya, "Football predictions based on a fuzzy model with genetic and neural tuning," Cybernetics and Systems Analysis, pp. 619-630, 2005.
11. R. T. Stefani, "Football and basketball predictions using least squares," IEEE Transactions on systems, man, and cybernetics, pp. 117-121, 1977.
12. A. Groll, C. Ley, G. Schauberger and H. Van Eetvelde, "Prediction of the FIFA World Cup 2018 – A random forest approach with an emphasis on estimated team ability parameters," 13 June 2018. [Online]. Available: https://arxiv.org/abs/1806.03208.
13. M. A. Diniz, R. Izbicki, D. Lopes and L. E. Salasar, "Comparing probabilistic predictive models applied to football," 25 April 2018. [Online]. Available: https://doi.org/10.1080/01605682.2018.1457485.
14. D. Dormagen, "Development of a Simulator for the FIFA World Cup 2014," Bachelorarbeit FU Berlin, 2014.
15. M. J. Dixon and S. G. Coles, "Modelling Association Football Scores and Inefficiencies in the Football Betting Market," Journal of the Royal Statistical Society. Series C (Applied Statistics), pp. 265-280, 1997.

ABBREVIATIONS AND ACRONYMS

- H2O – It is an open source software used for Big Data Analytics.
- FIFA - Fédération Internationale de Football Association
- k-NN – k-Nearest Neighbors
- CRM - Customer Relationship Management
- DSS - Decision Support System
- NFL – National Football League
- MLP – Multi Layer Perceptron
- NBA – National Basketball Association

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication