ISSN 2347 - 3983

Volume 8. No. 5, May 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter57852020.pdf

https://doi.org/10.30534/ijeter/2020/57852020

Predictive Analytics of Performance of India in the Olympics using Machine Learning Algorithms

Varagiri Shailaja[1], Rayala Lohitha[2], Sreethi Musunuru[3] ,K Deepthi Reddy[4] , J Padma Priya[5]

[1]Assistant Professor, GRIET, India, shailaja.v25@gmail.com

[2]Student, GRIET, India, lohitaarayala2000@gmail.com

[3]Student, GRIET, India, musunurusreethi@gmail.com

[4]Student, GRIET, India, deepthireddy2462000@gmail.com

[5]Student, GRIET, India, jankalapriya9@gmail.com

ABSTRACT

India is a country which all the time maintained an exciting pitch and an intense exhilaration for sports. Different reasons have been insinuated for India's lack of tendency to stand atop the podium in the Olympics. Natural endowment can take the athletes only to some extent, but support and encouragement in either financial, emotional or physical form are essential aspects that are necessary to an athlete and without them, the source can often be hopeless. The potential of education is substantial, and sports are considered as a source of relaxation and amusement for millions of youth who are aspiring to be at the top of their class in a country where there are millions of unemployed. The combination of this with terrible food habits, inefficient coaching, bad rehab facilities, increase in competition in schools, shortfall of exercise with physical education and long commutes from work results in many talents getting wasted. Despite all this, we see that India's cricket team is considered as one of the world's best cricket teams of all time. But at the same time we are not able to bring the same commitment and rectitude to the other sports, Olympics in particular. When all the nations are ranked in the order of the number of medals they have won over the years, India ranks at fifty fifth position. To understand the medal deficiency, various attributes are considered. This study, by using Data Science and Machine Learning algorithms, builds a model for predicting why India performs well or poorly in the Olympics by taking into account country-wise GDP, population, Gini index, health and education expenditures, literacy rate and prevalence of undernourishment as the features assuming that they contribute to the performance of the athletes and their effect is studied.

Key words: Data pre-processing, Data Science, Machine Learning, Olympics, Regression.

1. INTRODUCTION

The Olympics is considered as a prominent sporting event where thousands of athletes from around the world come together to take part in a variety of competitions. The nations from around the globe participate and the Olympic games are considered as the world's most popular sports competitions. Data Science and Machine Learning approaches are to be considered as a great help in the decision- making process of trainers, athletes and the government of these nations. Machine Learning is widely recognized as the methodology of choice in analyzing the Olympics Games data. In order to acquire a convincing result, data of the previous Olympics seasons are collected for model training and testing. This Project aims at getting good accuracy and specific insights using the concept of Correlation. This paper embarks on predicting the outcomes of a country's performance in sports using a supervised learning approach. The insights gathered can be used to reinforce the need to devise new policies to improve the quality of physical education in a country. The statistics show that a lot of attributes contribute to the performance of these countries in the games.

2. RELATED WORK

Proposed to predict a nation in view of medals owned by 2012. To develop a novel technique, they used a combination of Pearson correlation coefficient, Spearman correlation, coefficient and linear regression. They used the same dataset with features like GDP and number of medals won to compare the values obtained by using these methods to find the most suitable machine learning algorithms for the prediction [1]-[4].

Proposed to use data from the 1996 Summer Olympic Games and 2002 Olympic Winter Games to test the predictions of regional input-output models. The input-output have been used to understand the impact of short- duration sporting events. Public subsidies have been proposed to be improved using these models [5].

Proposed to use Artificial Neural Networks for sports result prediction. A novel sports prediction framework had been devised using Machine Learning. They have attempted to find the right ways of model evaluation and identifying the required data sources [6]-[8].

Has done performance analysis on Olympic games datasets using Python to evaluate the contributions of each country in the Olympics. To get a deeper insight on the performances of the countries in the Olympics over the years, exploratory data analysis has been performed by visualizing the dataset [9].

Used data envelopment analysis to understand how economic active population and corruption factors can work with the traditional system for medal prediction. The model structure proposed by decomposing the World Bank's income classification enables us to measure performances of the countries [10].

Proposed an approach that considers two inputs (GNP and population) and three outputs (number of golds, silver and bronze medals won). This was used to measure the performance of the nations in the Olympics [11].

3. PROPOSED MODEL

When the trend of the winning rate of a country is observed, we understand that factors like GDP, health index, literacy rate etc. affect the performance of a country in the Olympics. To prove the assumption, in this project, we make use of a reliable dataset and machine learning algorithms. In order to acquire a convincing result, data of the previous Olympics seasons is collected for model training and testing. The results of this project can be used to prepare a report to persuade a country to improve its positioning in prestigious sporting events. On a broader perspective, the observations achieved by correlating will help one understand how a country's performance in sports can affect other fields too. The insights gathered can be used to reinforce the need to devise new policies to improve the quality of physical education in the country. This project aims at getting good accuracy and specific insights using the concept of Correlation to understand if the features that are directly affecting the performance are the only ones to be taken into account or if there's more than what meets the eye of the analyst. To understand the methodology or implementation better, the architecture of proposed model is shown in figure 1,

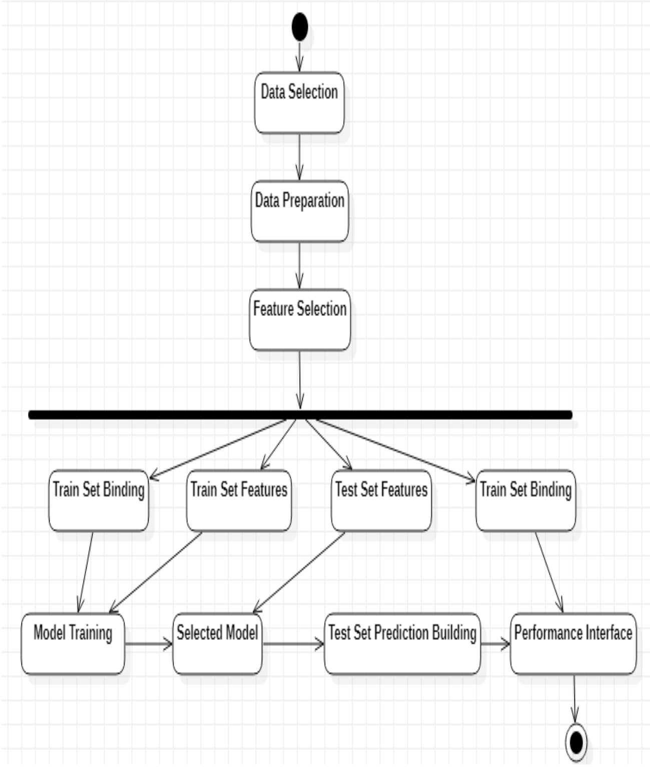

Figure 1: Architecture of Proposed Model

4. METHODOLOGY

4.1 Data Collection

The dataset for prediction and analysis of India's performance in the Olympics is collected from various sources on the web. The dataset contains 227 rows and 10 attributes. The attributes of the dataset include countries, population, GDP, education expenditure, literacy rate, health expenditure , the prevalence of undernourishment, Gini index, number of games participated and medals won. This research analyzes if the external attributes affect a country's performance in the Olympics.

4.2 Data Cleaning and Preprocessing

The dataset is built by collecting the data from the real world. Hence it contains missing values. These missing values may reduce the model accuracy and so they need to be handled. Also, the data is having values over a wide range and it needs to be normalized. There are many ways to fill the missing values like simply filling the values with zero or calculating the central tendencies. But since our dataset contains all continuous values over a wide range using the KNN algorithm is considered as an efficient solution.

4.2.1 Missing Values

Using KNN algorithm, a value can be approximated by the values of the points closest to it based on the other attributes. The number of neighbours to consider, the aggregation method to use and the distance function are the parameters to be focused. For this dataset, the k value is taken as 10, Euclidean for numeric distance and hamming for categorical distance are used, and the median is taken as the aggregation method.

4.2.2 Normalization

Normalization is an approach applied to the data as a part of data preparation to convert the values of numeric columns to a common range, not changing the differences in the range of values. Sigmoid Activation function is used to standardize the dataset which also helped in improving the accuracy during the training process.

4.3 Model Selection

Model Selection is an important step in the life cycle of any data science project since the algorithm we choose will influence the results. Data scientists use different Machine Learning algorithms to their datasets. We can divide those algorithms into supervised and unsupervised algorithms. Depending on the output label supervised is again classified into classification and regression.

The output label of the dataset is identified as continuous and hence this becomes a regression problem. The following regression algorithms are applied to the dataset:

- Decision tree regression: It builds a decision tree by breaking down the data into sets and subsets. The final output can be read from the decision nodes and leaf nodes. This has been considered as it considers all outputs of a decision and helps trace all paths to a conclusion.
- K nearest neighbours Algorithm: It stores all cases and classifies them based on distance function. This is considered as it can be used for pattern recognition and statistical estimation.
- Linear Regression: It is a method used to model scalar response and one or more explanatory variables. This has been considered as it models the independent variables and dependent variables.
- Random forest regression: A random forest can perform both regression and classification tasks on the data by building multiple decision trees. This has been considered as it offers efficient estimates of test error.
- Bayesian ridge regression: A probabilistic model of the regression problem can be estimated using Bayesian ridge regression.

The task of these algorithms is to predict the state of an outcome variable at a particular time point with the help of other correlated independent variables. There are various metrics used to evaluate the results of these predictions. The following metrics are considered for this study:

1. Mean Squared Error (MSE)
2. Root-Mean-Squared-Error (RMSE)
3. Mean-Absolute-Error (MAE)

5. RESULTS

The predicted values and the actual values of all the algorithms are plotted against each other using a scatter plot. Studying scatter plots: The points in neither of the graphs is either a straight line or a curve. Hence, we cannot directly estimate the correlation. We need to divide the graph into four quadrants and use a trend test table. The quadrants are generated by dividing the graph such that either horizontally or vertically there are equal numbers of points on both sides. Then we take the minimum of the sum of the upper left quartern and lower right quartern, and upper right quartern and lower left quartern. This is verified against the trend test table and the correlation is obtained.

Decision Tree

The scatter plot of the predicted values against actual values using Decision Tree algorithm is shown in figure 1

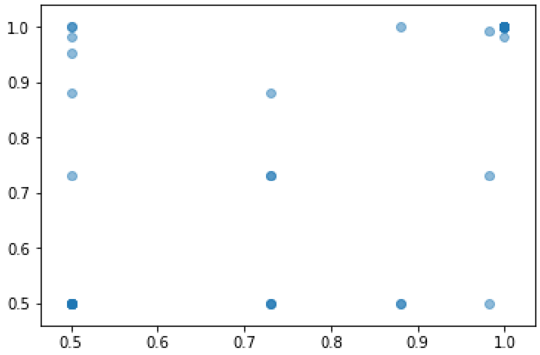

Figure 1: Scatter plot for Decision Tree Algorithm

KNN Algorithm

The scatter plot of the predicted values against actual values using KNN algorithm is shown in figure 2

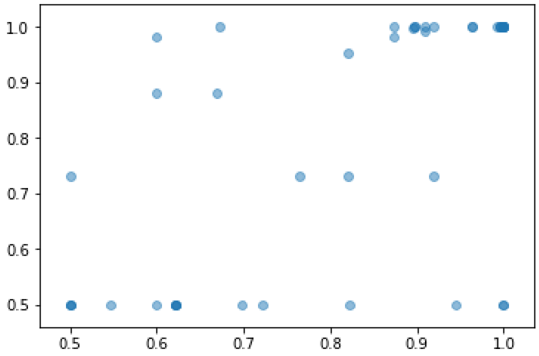

Figure 2: Scatter plot for KNN Algorithm

Linear Regression

The scatter plot of the predicted values against actual values using Linear Regression algorithm is shown in figure 3

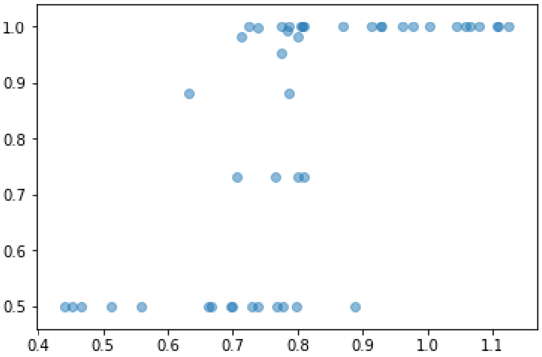

Figure 3: Scatter plot for Linear Regression

Random Forest Algorithm

The scatter plot of the predicted values against actual values using Random Forest algorithm is shown in figure 4

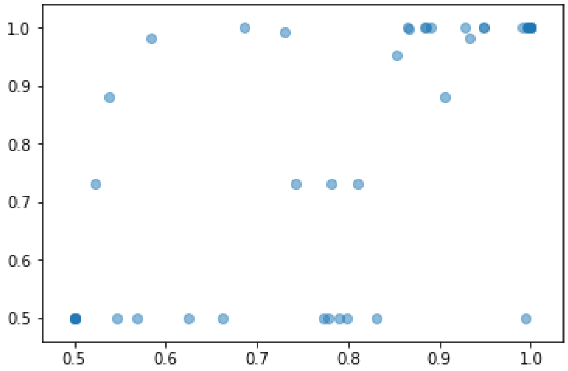

Figure 4: Scatter plot for Random Forest Algorithm

Bayesian Ridge Algorithm

The scatter plot of the predicted values against actual values using Bayesian ridge algorithm is shown in figure 5

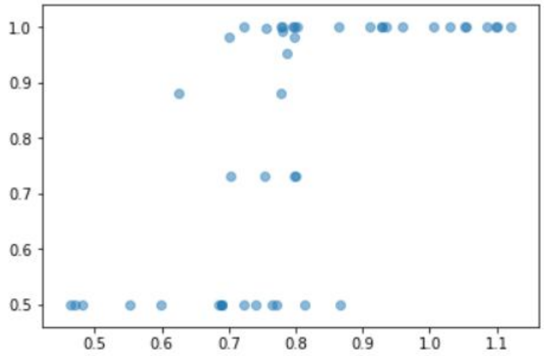

Figure 5: Scatter plot for Bayesian Ridge Algorithm

The mean absolute error (MAE), mean square error (MSE) and root square error (R2) calculated for the algorithms are given in table 1.

Table 1: Error Measures

ALGORITHM	MAE	MSE	R2
Decision Tree Regressor	0.109	0.042	0.182
KNN Regressor	0.122	0.033	0.344
Linear Regression	0.143	0.028	0.426
Random Forest Regressor	0.111	0.028	0.442
Bayesian Ridge	0.144	0.029	0.422

The error given by the decision tree regressor is low when compared to the other algorithms. But when we observe the scatter plots of every algorithm, we can derive the correlation between the predicted and actual values.

Thus, considering both the error values and the correlation values, KNN Regressor is the final selected algorithm.

Testing using a validation dataset: A new dataset is created for the purpose of testing the performance of the selected algorithm and the results are obtained.

Correlation between attributes: A pairwise scatter plot of each attribute with every other attribute is obtained and the graphs are studied to understand the correlation. Taking hue as countries the following observations are made: The number of medals won is strongly positively correlated with the population and literacy rate. Prevalence of undernourishment, GDP and Gini index is positively correlated with medals with some outliers. Education expenditure is in a weak positive correlation on the other hand health expenditure is not correlated. But health expenditure is in positive correlation with the prevalence of undernourishment. Thus, all the considered attributes affect the performance of a country in the Olympics either directly or indirectly.

6. CONCLUSION

The main objective behind the study is to understand the effect of factors that are external to sport on a country's performance in the Olympics. The observations drawn are: the number of medals won is strongly positively correlated with the population and literacy rate; Prevalence of undernourishment, GDP and Gini index is positively correlated with medals with some outliers; Education expenditure is in a weak positive correlation on the other hand health expenditure is not correlated; Health expenditure is in positive correlation with the prevalence of undernourishment. Thus, all the considered attributes affect the performance of a country in the Olympics either directly or indirectly.

When we compare India with China, which is among the top ten countries in the Olympics and whose population is nearly equal to India, we can understand how important the other factors are to upgrade a country's performance in the Olympics and how important it is to make sport mandatory rather than considering them as recreational activities. From these insights, we see that education is given higher importance in India and yet the literacy rate and economic status of its people are low. Hence sports should be given higher importance.

REFERENCES

1. Chandrasegar Thirumalai, S Monica, A Vijayalakshmi, "Heuristics Prediction of Olympic Medals Using Machine Learning" IEEE- International Conference on Electronics, Communication and Aerospace Technology ICECA 2017, At Coimbatore, India. https://doi.org/10.1109/ICECA.2017.8212734
2. Anto Germin Sweetha J, Dr B Ramakrishnan, "Novel Heterogenous Data Integration Using KNN Algorithm" in IJETER(Volume 7, Issue 4, April(2019))
3. Ajmol C, Dr. R. Kavitha Jaba Malar, "Biometric Fingerprint Spoof Identification Using Neural Networks" in IJETER(Volume 7, Issue 5, May(2019)).
4. Y. Jeevan Nagendra Kumar, B. Mani Sai, Varagiri Shailaja, Singanamalli Renuka, Bharathi Panduri, Python NLTK Sentiment Inspection using Naïve Bayes Classifier IJRTE ISSN: 2277-3878, Volume-8, Issue-2511, September 2019.
5. Philip K Porter, Deborah Fletcher, "The Economic Impact of the Olympic Games Ex Ante Predictions and Ex Poste Reality", Journal; of Sports Management, Volume 22, Issue 4. https://doi.org/10.1123/jsm.22.4.470
6. Rory P. Banker, Fadi Thabtah, "A machine learning framework for Sport result prediction", Applied Computing and Informatics(Volume 15, Issue 1, January 2019). https://doi.org/10.1016/j.aci.2017.09.005
7. A Joseph Selvanayagam, Dr S John Peter, "A Visually Impaired Double Watermarking using Discrete Wavelet Transform" in IJETER(Volume 7, Issue 5, May(2019)).
8. Prasanna Lakshmi, K Reddy, C.R.K, "A Survey on different trends in Data Streams (2010)", ICNIT 2010-2010 International Conference on Networking and information Technology, art.no.5508473, pp. 451-455. Cited 15 times. 2-s2.0-77955591448 Document Type: Conference Paper Publication Stage: Final Source: Scopus.
9. Yamunathangam D, Kirthicka G, Shahanas Parveen, "Performance Anlaysis in Olympic Games Using Exploratory Data Analysis Techniques", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878(Volume-7 Issue-4S, November 2018).
10. M. Flegl, L.A.Andrade(2018) Measuring Countries Performance at the Summer Olympic Games in Rio 2016, OPSEARCH, Springer, Operational Research Society of India, vol. 55(3), pages 823- 846, November. https://doi.org/10.1007/s12597-018-0347-8
11. Sebastian Lozano, Gabriel Villa, Fernando Guerrero, Pablo Cortes "Measuring the Performance of Nations at the Summer Olympics Using Data Envelopment", Journal of the Operational Research Society, 53(5): 501-511, May 2002. https://doi.org/10.1057/palgrave.jors.2601327