DraftCheck - Web Application For Article Analysis

**Rayala Lohitaa, Sreethi Musunuru, K Deepthi Reddy, J Padma Priya**

***Abstract: Digitisation is the process of converting information from physical formats into a digital one so that we can increase data quality and apply advanced techniques like summarising, extracting keywords etc. Text or article summarization refers to the technique of lessening long pieces of text. The intention is to create a well organized and fluent summary having only main points outlined in the document. Similarly, the keyword extraction process not only separates the articles but also helps in saving time when we are searching for a document. Also when there is a limit on the number of words that can be used in a document, we need the help of a word counter. Spell Checking is an important task as we might accidentally spell the words incorrectly and this makes the article less accurate. DraftCheck provides a single platform for all the analysis tasks of articles. This project focuses on using flask to provide a framework for the web application, natural language processing to perform tasks like summarising, keyword extraction etc, web designing and services to maintain the application and other techniques to achieve the article analysis. In order to provide a better user experience, a software intermediary or an api is used so that a smooth communication is achieved between various parts of the application. The main aim of this project is to increase the accuracy of the existing systems and provide a one stop solution to the user requirements. Adding other functionalities and customising the application to make it more user friendly can be considered as the future work for this project.***

***Keywords: Flask, Digitisation, Web application, Summarisation, Keyword extraction, NLP.***

**I. INTRODUCTION**

DraftCheck is a web application and hence it needs a software intermediary. We need to understand that a software intermediary needs to exist for two applications to communicate with each other. Imagine you are checking weather or using an App for instant messaging, so what you are doing is using an API (Application Programming Interface).If you send a request to the provider, the API delivers it to the provider and returns with the response from the provider.

APIs also provide a layer of security as they are standardized, governed and monitored for performance.APIs make complex processes reusable with just a little bit of code. This makes the process of app development faster.

**How to send and retrieve data using an API?**

The architecture of an API consists of three parts-

User, who makes a request, Client, the computer which sends the request to the server and server, which responds

to the request made by the client.

In order for the server to send a response and hold data it has to be built by someone first. Once the server starts running, the programmer can publish its documentation and endpoints so that other developers can understand the structure of data on the server.

Then the developer can send a query to search the data in the database and transform the data into a different desired format.

**II. LITERATURE SURVEY**

Chandrasegar Thirumalai, S Monica, A Vijayalakshmi (2017): To predict a nation in view of medals owned by 2012, they used a combination of Pearson correlation coefficient, Spearman correlation coefficient and linear regression.

Philip K Porter, Deborah Fletcher (2008): They used data from the 1996 Summer Olympic Games and 2002 Olympic Winter Games to test the predictions of regional input-output models.

Rory P. Bunker, Fadi Thabtah (2017): They used Artificial Neural Networks for sports result prediction. A novel sports prediction framework had been devised using Machine Learning.

Yamunathangam D. (2018): Performance Analysis was done on Olympic games datasets using Python to evaluate the contributions of each country in the Olympics.

M. Flegl, L. A. Andrade (2018): They used data envelopment analysis to understand how economic active population and corruption factors can work with the traditional system for medal prediction.

Sebastian Lozano, Gabriel Villa, Fernando Guerrero, Pablo Cortes (2015): The proposed approach considers two inputs (GNP and population) and three outputs (number of gold, silver and bronze medals won).

**III. PROPOSED SYSTEM**

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.

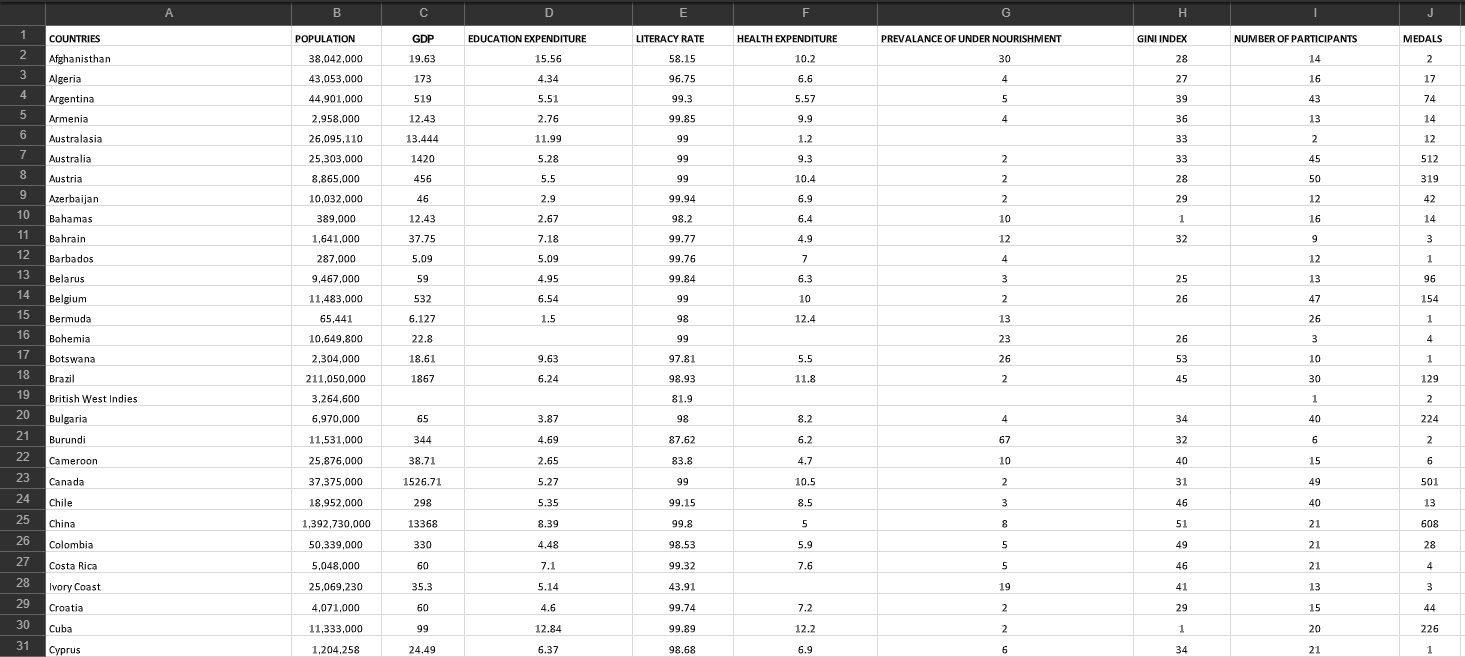
In order to acquire a convincing result, data of the previous Olympics seasons will be collected for model training and testing. The results of this project can be used to prepare a report to persuade a country to improve its positioning in prestigious sporting events.

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.

**IV. Methodology**

*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 consists of 227 rows and 10 attributes. The attributes of the dataset include countries, population, GDP, expenditure on education, literacy rate, expenditure on health, the prevalence of undernourishment, Gini index, number of games participated and medals won. This research would analyze if the external attributes affect a country’s performance in the Olympics.

** **Fig 1: Dataset snapshot**

*Data Cleaning and Pre Processing:* 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.

*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.

*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 scale, without 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.

Model Selection is an important phase in Applying Machine Learning to any Dataset as the results depend on how effectively the algorithms selected will fit the data. 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:

1. *Decision tree regressor:*

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.

1. *K nearest neighbours regressor:*

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.

1. *Linear Regression:*

It is a method used to model scalar response and one or more explanatory variables.

The linear regression line has the form of Y= a+bX.

This has been considered as it models the independent variables and dependent variables.

1. *Random forest regressor:*

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.

1. *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 project :

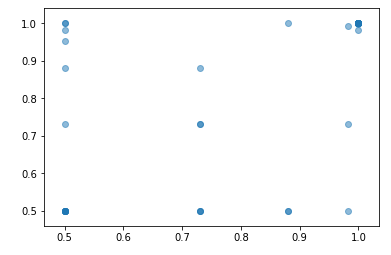
1. Mean Squared Error(MSE)
2. Root-Mean-Squared-Error(RMSE).
3. Mean-Absolute-Error(MAE).

**V. RESULTS**

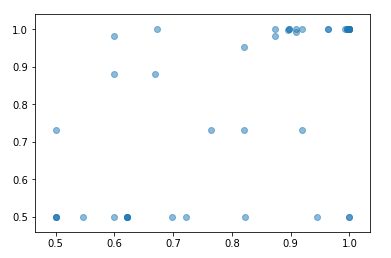
The predicted values and the actual values of all the algorithms are plotted against each other using a scatter plot.

*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 quadrant and lower right quadrant, and upper right quadrant and lower left quadrant. This is verified against the trend test table and the correlation is obtained.

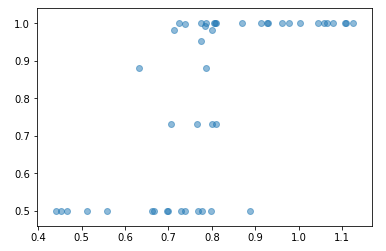
The scatter plot for Decision Tree regressor:

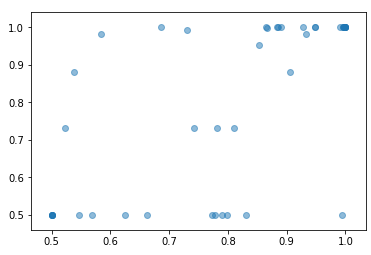


The scatter plot for KNN regressor:

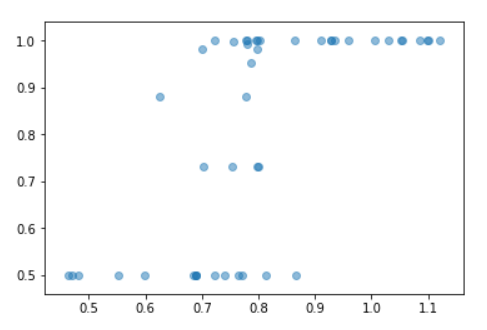


The scatter plot for Linear regression:



The scatter plot for Random forest regressor:

The scatter plot for Bayesian ridge regression:



The mean absolute error(MAE), mean square error(MSE) and root square error(R2) calculated for the algorithms are:

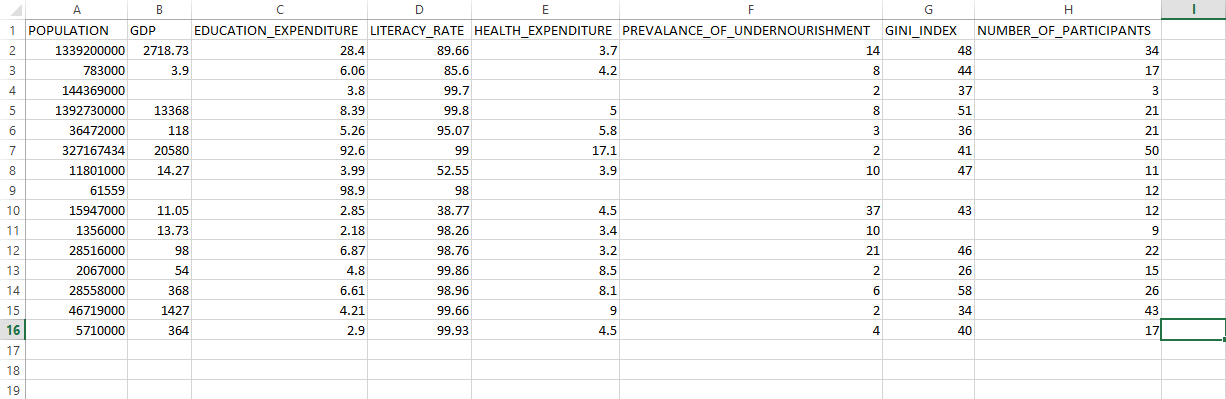
| **Algorithm** | **MAE** | **MSE** | **R2** |
| --- | --- | --- | --- |
| Decision tree regressor | 0.109 | 0.042 | 0.182 |
| KNN regressor | 0.122 | 0.033 | 0.344 |
| Linear regression | 0.143 | 0.028 | 0.426 |
| Random forest  regressor | 0.109 | 0.029 | 0.438 |
| Bayesian ridge | 0.109 | 0.029 | 0.438 |

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 algorithm.



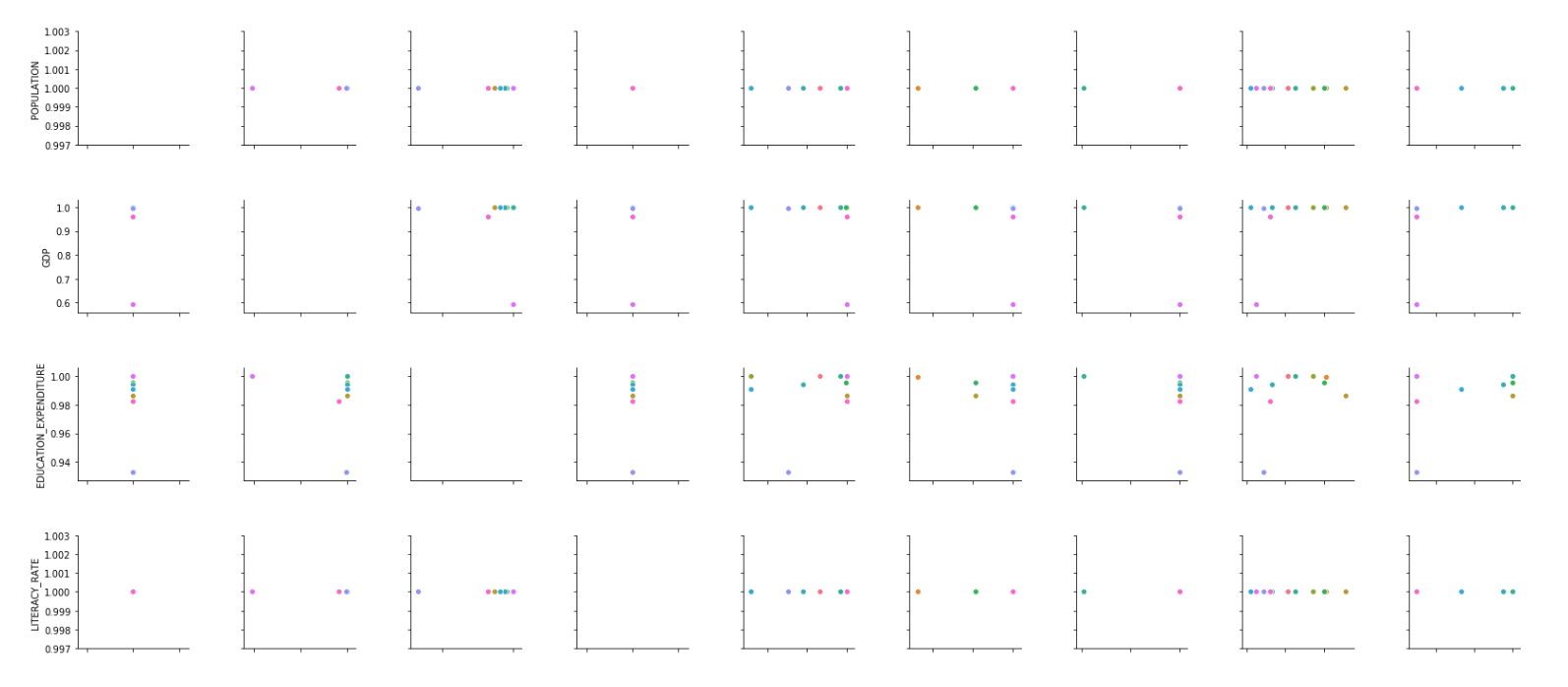


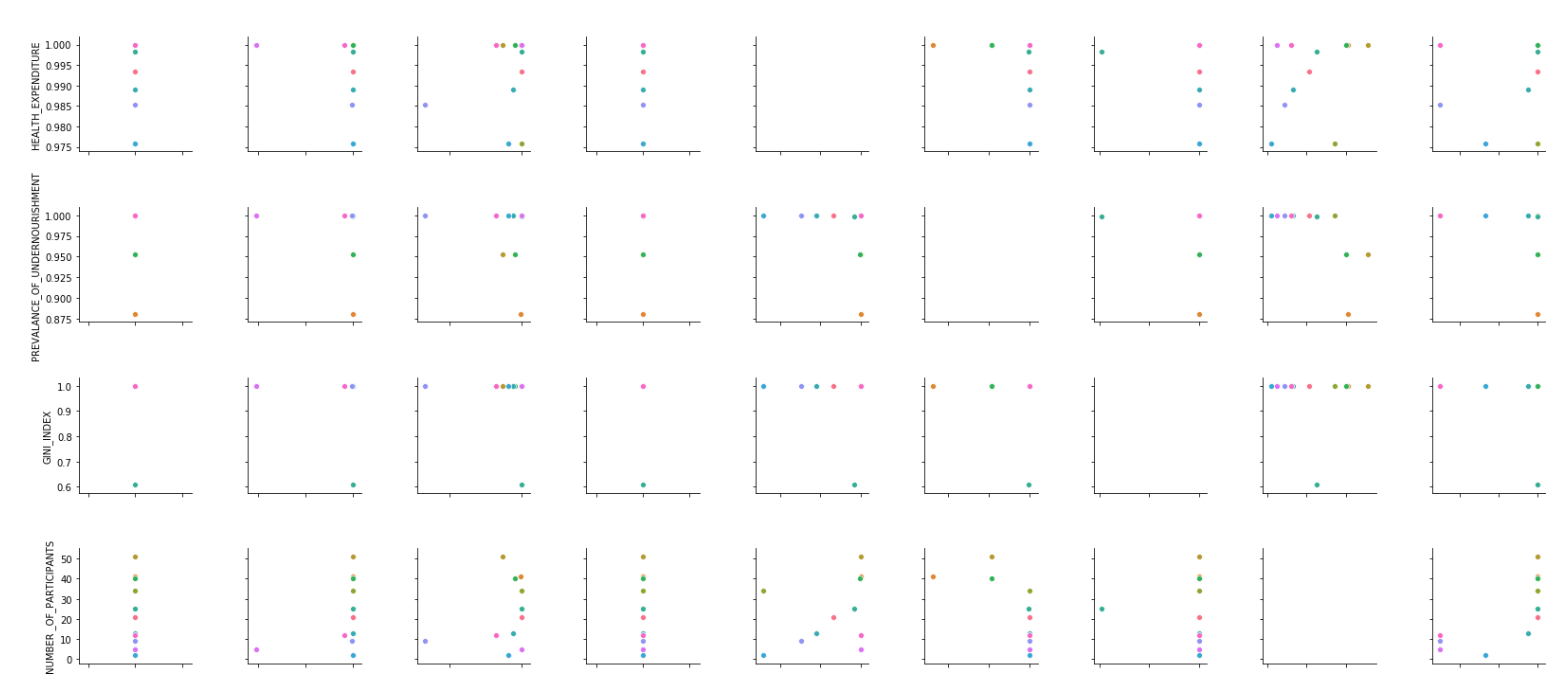
*Correlation between attributes:* A pairwise scatter plot of each attribute with every other attribute is obtained and the graphs are studied to understand the correlation. Taking *hue* as the countries the following visualization is obtained. The observations made are:

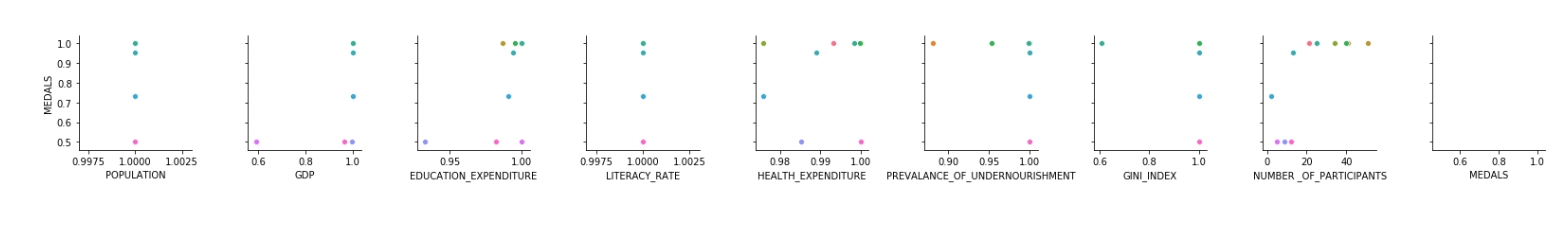
The number of medals won is strongly positively correlated with the population and literacy rate.

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. 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.









**VI. CONCLUSION**

The main objective behind the study is to understand the effect of external factors 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 improve 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.

**VII. REFERENCES**

1. Chandrasegar Thirumalai, S Monica, A Vijayalakshmi, Heuristics Prediction of Olympic Medals using Machine Learning(2017) IEEE - International Conference on Electronics, Communication and Aerospace Technology ICECA 2017, At Coimbatore, India.
2. Philip K Porter, Deborah Fletcher, The Economic Impact of the Olympic Games: Ex Ante Predictions and Ex Poste Reality(2008), Journal of sports management
3. Rory P. Bunker, Fadi Thabtah, A machine learning framework for sport result prediction (2017)
4. Yamunathangam.D, Kirthicka.G, Shahanas Parveen, Performance Analysis 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
5. Padmavathi, K., Krishna, K.S.R. Myocardial infarction detection using magnitude squared coherence and Support Vector Machine (2014) 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems, MedCom 2014, art. no. 7006037, pp. 382-385. Cited 9 times. 2-s2.0-84988299735 Document Type: Conference Paper Publication Stage: Final Source: Scopus
6. 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
7. Padmavathi, K., Sri Ramakrishna, K. Classification of ECG signal during Atrial Fibrillation using Autoregressive modeling (2015) Procedia Computer Science, 46, pp. 53-59. Cited 29 times. 2-s2.0-84931436900 Document Type: Conference Paper Publication Stage: Final Source: Scopus
8. Dhanalaxmi, B., Apparao Naidu, G., Anuradha, K. Adaptive PSO based association rule mining technique for software defect classification using ANN (2015) Procedia Computer Science, 46, pp. 432-442. Cited 7 times. 2-s2.0-84931310946 Document Type: Conference Paper Publication Stage: Final Source: Scopus
9. Y.Jeevan Nagendra Kumar, B. Mani Sai, Varagiri Shailaja, Singanamalli Renuka, Bharathi Panduri, Python NLTK Sentiment Inspection using NaïveBayes Classifier IJRTE ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019.
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.
11. Sebastian Lozano, Gabriel Villa, Fernando Guerrero, Pablo Cortes Measuring the Performance of Nations at the Summer Olympics Using Data Envelopment
12. Dragan Miljković, Ljubisa Gajic, Aleksandar Kovacevic, Zora Konjovic, The use of data mining for basketball matches outcomes prediction IEEE 8th International Symposium on Intelligent Systems and Informatics 2010
13. Josip Hucaljuk, Alen Jakupovic, Predicting football scores using machine learning techniques, 2011 Proceedings of the 34th International Convention MIPRO 2011
14. Darwin Prasetyo, Dra. Harlili Predicting football match results with logistic regression, 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA) 2016
15. Alan McCabe, Jarrod Trevathan, Artificial Intelligence in Sports Prediction Fifth International Conference on Information Technology: New Generations (itng 2008) 2008