IJIEMR Transactions, online available on 26th Nov 2021. Link :http://www.ijiemr.org/downloads.php?vol=Volume-10&issue=Issue 11

**10.48047/IJIEMR/V10/ISSUE 11/77**

Title *A CRITICAL STUDY ON BIG DATA ANALYTICS USING MACHINE LEARNING FOR DATA PREDICTION*

Volume 10, ISSUE 11, Pages: 479-483

Paper Authors **P. RAJASHEKER REDDY, Dr. Mukesh Kumar**

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

# A CRITICAL STUDY ON BIG DATA ANALYTICS USING MACHINE LEARNING FOR DATA PREDICTION

**P. RAJASHEKER REDDY**
Research Scholar Monda University, Delhi Hapur Road Village & Post Kastla, Kasmabad, Pilkhuwa, Uttar Pradesh
**Dr. Mukesh Kumar**
Research Supervisor Monda University, Delhi Hapur Road Village & Post Kastla, Kasmabad, Pilkhuwa, Uttar Pradesh

## ABSTRACT

Big Data Analytics has emerged as a pivotal field in extracting valuable insights and predictions from vast and complex datasets. This paper explores the integration of Machine Learning techniques into the realm of Big Data Analytics for the purpose of data prediction. As the volume, variety, and velocity of data continue to increase exponentially, traditional methods of data analysis fall short in processing and deriving meaningful information from such large datasets. Machine Learning algorithms, with their capacity to learn patterns and relationships within data, offer a promising approach to tackle these challenges.

**Keywords: -** Data, Machine Learning, Big data, Volume, Framework.

## I. INTRODUCTION

Data analysis is the way toward dissecting the data to remove data. The removed data is useful for additional analytics. In light of the fast increment of data across different spaces handling those using manual framework is a difficult assignment. To overcome this issue robotized frameworks are familiar with measure on going data. This is accomplished via doing Machine Learning techniques. In this part the highlights of mechanization are empowered by using Spark an open source instrument that assists with resembling disseminated environment. The Machine learning algorithms are used on Spark to generate figure results. The exemplified resource that includes the exercises of proposed model is depicted in the going with portions.

## II. ENVIRONMENTAL SETUP

A natural game plan is set up across cloud stages by taking the occurrences from Amazon Web benefits on EC2. The virtual machine is outfitted with the establishment of Spark and Anaconda in which the Jupyter diary are gotten to by executing the Machine Learning algorithms. Orange is an open source device that unravels the delayed consequences of Machine Learning algorithms through cross approval. Table 1. Portrays the devices required for the execution.

**Table 1. Tools Applicable for Implementation**

| Work Flow | Tools | Version |
|---|---|---|
| Job Distribution | Spark | Spark-2.1.0-bin-hadoop2.7 |
| Prediction | Linear Regression | Anaconda3-4.3.1 |
| | Random Forest | |

| | | |
|---|---|---|
| | Decision Tree | |
| | Gradient Boosting Tree | |
| Canvas | Orange | Orange 3.2 |

## III. TOOLS

The Machine Learning frameworks analyze datasets and automatically catch the highlights while preparing the test data. The preparation set comprises of data, includes and needed yield. The yield of the capacity is a constant value or expectation from a class name.

Scikit Learn is a Python Machine Learning library used for executing the proposed work. It maintains both controlled and independent learning algorithms. Jupyter scratch pad is another instrument that runs in Apache Spark structure. This Spark system scatters occupations on virtual machines that run Jupyter scratch pad completing Machine Learning techniques.

Orange is open source programming that helps with stacking and change data. It is useful in applications like data mining, Machine Learning, farsighted analytics and Statistical Demonstrating. This product assesses the relative results of the algorithms drawn from the proposed model.

## IV. PHASES OF PROPOSED MODEL

The proposed model comprises of three significant stages. Starting stage concentrates on data assortment which incorporates of data and brief depiction of dataset. The calculation portrayal followed by the work process and execution are depicted in stage II the stage III displays the outcome and error portrayal between the learning algorithms with an excellent spotlight on the similar outcome investigation.

**Stage I:** Data Collection

**Stage II:** Proposed Model and Algorithm

**Stage III:** Comparative Result examination

### Phase I: Data Collection

The dataset comprises of Temperature data 'All India occasional Annual Temperature' series extricated from https://data.gov.in/lists/ministry_department/india-meteorological-division imd. It comprises of annual temperature perusing over 100 years which helps in expecting the temperature for the looming years. The removed dataset contains both annual occasional data with least and most extreme temperature recorded over years. It contains occasional perusing, Month-wise perusing and annual perusing which benefits the proposed work for handling the data to expect the future temperature. Table 3.2 addresses the prepared data dependent on the marked highlights.

### Phase II: Proposed Model and Algorithm

The proposed model illustrates errand of the AI in anticipating the value of the capacity for a substantial info object having countless preparing data.

## V. COMPARISON OF LEARNING ALGORITHMS

The noticed values of learning algorithms are thought about across different traits utilizing Orange device. Linear Regression, Random Forest and Decision tree algorithms are thought about utilizing two significant techniques (i) cross validation and (ii) random sampling.

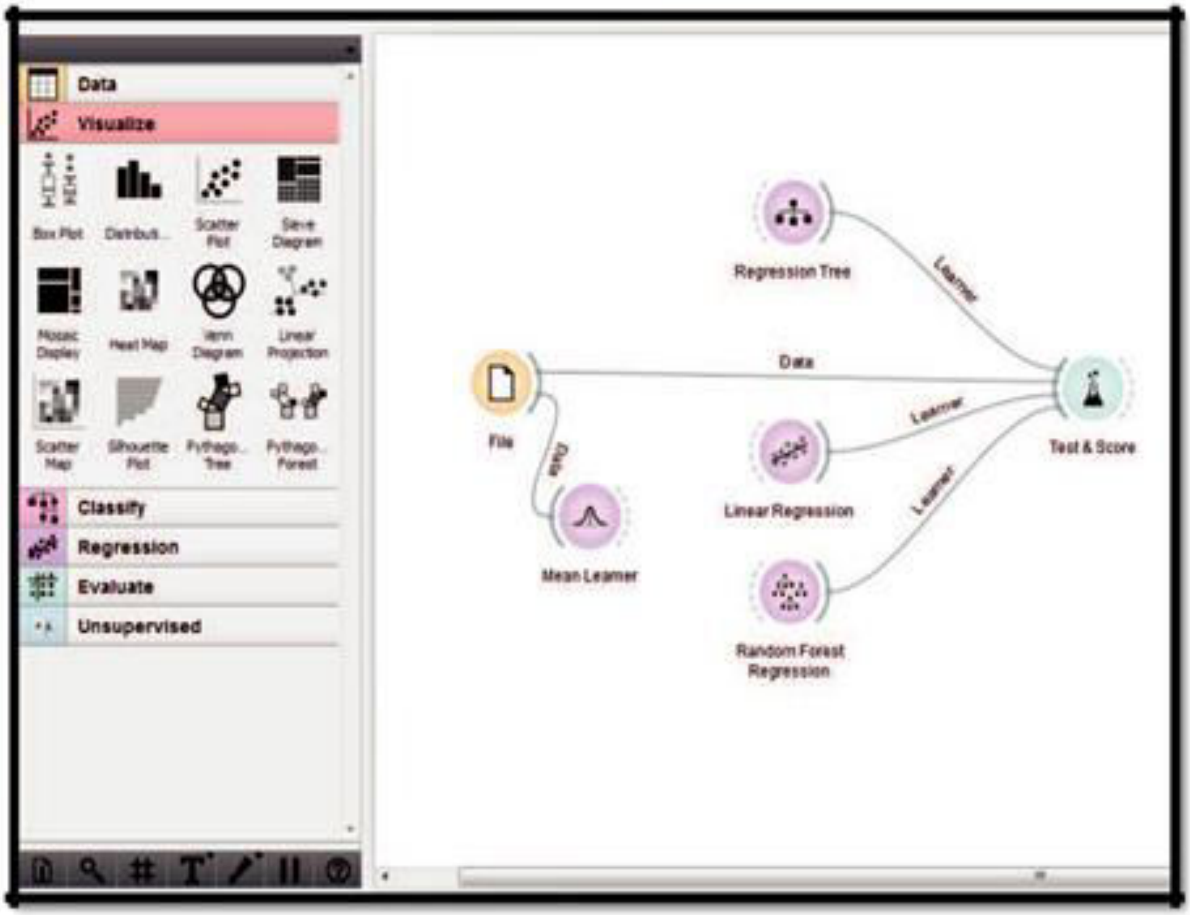

**Figure – 1: Comparison of Algorithms using Orange Tool**

Figure 1. Illustrates the material of orange apparatus that considers the test data for looking at the algorithms. A grade processed allots the quantity of folds as 10. Cross approval is assessed on Linear Regression, Random Forest Regression and Regression Tree (Decision Tree) for a delineated examining technique. The calculations are done for MSE, RMSE, MAE and R2.

Linear Regression bring about ideal positive relationship between the annual and forecast highlights from the model by processing R2=1.00. Random Woodland Regression shows no relationship between the highlights and ultimately regression tree shows negative connection. Mean Square error is more for regression tree when contrasted and Random boondocks.

## VI. CONCLUSION

In conclusion, the convergence of Big Data Analytics and Machine Learning has ushered in a new era of data prediction that holds immense promise for various industries and domains. The amalgamation of vast and complex datasets with the prowess of Machine Learning algorithms has demonstrated its potential to revolutionize decision-making, innovation, and efficiency. This symbiotic relationship addresses the limitations of traditional data analysis methods, enabling organizations to extract valuable insights, make informed predictions, and ultimately stay competitive in an increasingly data-driven landscape.

The journey of employing Machine Learning for data prediction within the realm of Big Data Analytics has been one of continuous exploration, adaptation, and refinement. Throughout this process, key considerations such as data quality, feature selection, algorithm suitability, and scalability have come to the forefront. The selection of appropriate algorithms has been a pivotal point, offering a spectrum of choices that cater to diverse data characteristics and prediction objectives. From the flexibility of decision trees to the complexity of neural networks, the toolbox of Machine Learning empowers analysts to tailor their approaches for optimal results.

Moreover, the advent of cloud computing and distributed processing frameworks has paved the way for handling the sheer volume and complexity of Big Data. The ability to process, analyze, and predict patterns within such vast datasets has unleashed a new level of potential, accelerating the pace of discoveries and insights. However, this progress has not been without challenges. Ethical concerns surrounding data privacy, security, and bias must be meticulously addressed to ensure that the benefits of predictive analytics are harnessed responsibly and equitably.

As we look to the future, the partnership between Big Data Analytics and Machine Learning is poised to continue shaping industries, transforming operations, and enhancing our understanding of complex systems. The predictive capabilities afforded by this integration enable businesses to anticipate trends, optimize processes, and make strategic decisions with confidence. Moreover, the potential extends beyond corporate domains, influencing fields such as healthcare, environmental science, social policy, and beyond.

In essence, the synergy between Big Data Analytics and Machine Learning for data prediction is an ongoing narrative of innovation, advancement, and realization. As technologies evolve, and new horizons emerge, the collaboration between data-driven insights and intelligent algorithms will undoubtedly catalyze breakthroughs that redefine what is achievable. The journey towards predictive excellence is one that demands vigilance, creativity, and ethical stewardship. By navigating these waters judiciously, we can unlock the true potential of data as a transformative force for a more informed and empowered world.

## REFERENCES

1. S. Landset, T.M. Khoshgoftaar, A.N. Richter, T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem." Journal of Big Data 2.1 (2015): 24
2. Ren, Y., Han, Z., Chen, K.C. and Hanzo, L., 2017. Machine learning paradigms for next-generation wireless networks.. IEEE Wireless Communications journal, Volume 24, pp. 98-105

3. Ren et al., “Ren, Y., Han, Z., Chen, K.C. and Hanzo, L., 2017. Machine learning paradigms for next-generation wireless networks.. IEEE Wireless Communications journal, Volume 24, pp. 98-105.
4. “Brocke, J.V. and Debortoli, S., 2016. Utilizing big data analytics for information systems research: challenges, promises and guidelines. European Journal of Information Systems, pp. 289-302
5. AMIS (2019, April 9). What is Apache Drill, and how to set up our Proof-of-Concept? AMIS, Data-Driven Blog -Oracle& Microsoft Azure.