International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XVIII): January - March 2026
ISSN 2394 - 7780

EXPLAINABLE MACHINE LEARNING: ENHANCING TRANSPARENCY AND RELIABILITY IN INTELLIGENT SYSTEMS

Mr. Dhanesh Kumar P. Singh
Assistant Teacher, Chandrabhan Sharma Junior College of Commerce & Science, Powai, Mumbai-400076

ABSTRACT
Artificial intelligence (AI) systems are computational systems with varying levels of autonomy that, given a set of human-defined objectives, can generate predictions, provide recommendations, or make decisions by leveraging large and diverse datasets, commonly referred to as "big data" (OECD, 2019). These datasets feed machine learning (ML) models, enabling them to identify patterns, learn from experience, and improve performance over time without explicit programming by humans.

In recent years, AI and ML have emerged as transformative technologies in the field of finance, offering new approaches for risk assessment, investment strategies, fraud detection, and decision support. This study examines the thematic structure of AI and ML research within finance, employing co-occurrence and confluence analyses to identify key trends, interconnections, and emerging research directions. The findings provide a comprehensive assessment of current research efforts, highlighting the evolution, applications, and potential of AI and ML in financial contexts, as well as guiding future investigations in this rapidly evolving area.

Keywords: *Artificial intelligence, Bibliometric analysis, Finance, Machine learning, Big Data*

INTRODUCTION
This study focuses on applying supervised learning techniques to real-time datasets using the MapReduce programming model. Machine learning methods will be optimized to handle large-scale data, enabling applications such as prediction, pattern recognition, deep learning, recommendation systems, and more.

The MapReduce framework processes data using parallel "map" and "reduce" operations and iterative algorithms. While it can handle large datasets, traditional MapReduce models often face challenges such as high disk usage and low throughput. Deploying advanced analytics within this framework is increasingly difficult, as organizations must extract meaningful insights from existing data for further analysis.

To develop efficient predictive intelligence systems, existing models must be tuned using machine learning techniques. These methods leverage existing data features to generate accurate predictions. Since data is distributed across cluster machines, optimizing analytics performance in Hadoop systems is a significant challenge. The proposed model aims to minimize overall execution time while delivering the most accurate predictions possible, demonstrating the effectiveness of supervised learning techniques in handling real-time data within constrained time limits.

OBJECTIVES OF STUDY
The primary objective of this research is to tune Big Data methodologies using supervised approach. This research is a fair attempt to study the complex learning techniques. Prediction outcome is extracted from the existing data sets by utilizing Machine Learning methods. The captivating objectives are discussed below:

1. To develop a predictive model that forecast from the test data
The supervised learning methods are used as automated tool for prediction on data set. It incorporates methods to transform, evaluate and predict from the data by computing RMSE.

2. To identify the framework for distributed platform for analytical modeling
Data is parallely distributed on cluster environment, which appropriately adapts the dataset for analytical processing.

3. To compare Machine Learning Algorithms for accurate prediction
RMSE is calculated on each learning method and collective evaluation is examined by considering time and space complexities.

4. To evaluate new model with MapReduce and Machine Learning techniques
This model is evaluated on Apache Spark framework by replacing the existing MapReduce for prediction through supervised algorithms.

5. To predict the data from existing feature selection

ISSN 2394 - 7780

Apart from the considered dataset there is an elite chance of possibility to append further features to increase the accuracy of prediction.

Study Background (Literature Review):
Data Analytics is the scientific and statistical tool for analyzing raw data to renovate information for acquiring knowledge. Data analytics collaborates with data to formulate complex decisions from diverse perspectives for facing the real world challenges. The role of analytics is to assemble, store, process and analyze data to address empirical methods in real world for decision making. It is broadly classified into descriptive, inferential, predictive and prescriptive analytics **[1].**

Data analytics extends as a process of analyzing massive real time streaming data, which varies in data structure called as Big Data Analytics. Big data acts as a frontier for innovation, competition, productivity and business forecasting since the data is exponentially growing**[2].**

The analytics on such huge data reveals hidden patterns, unfound correlations, market trends, consumer requirements and future recommendations, which assist in critical decision-making process **[3].**

The analytical tools applicable on Big Data facilitate both technical companies and academic users for processing an expected span of time and for acquiring knowledge from the data **[4].**

Significance of Study:
This research portrays the usage of Machine Learning techniques for Big Data analytics. The proposed model incorporates the parallel processing on distributed environment like Spark to predict various features from datasets. The central core of the thesis is the comparative study on various Machine Learning algorithms on a proposed model for predictive analytics.

The following are the unique features of research study:
- Proposing a novel model to obtain best prediction from various Machine Learning algorithms.
- Time and space are the prime parameters for assessing performance metrics of this model.
- MapReduce and Spark models are interpreted for computational operations on clusters after job allocation.
- The data drawn from past experience are trained as test data using supervised learning.
- At times applying certain irrelevant and redundant features on the model may affect the performance negatively due to outliers. For instance, considering the temperature dataset, the redundant and irrelevant features like amount of rainfall, salinity and salt concentration in the ocean water may impact negatively.

Research Methodology:

The various Phases of Proposed Model will be as follows:

Phase I: Data Collection

Phase II: Proposed Model and Algorithm

- Algorithm for Machine Learning using PySpark Framework
- Flowchart for the Proposed System
- Process of Transformation
- Machine Learning

Phase III: Comparative Result Analysis Comparison of Learning Algorithms An elaborate narration on various active phases that support the functionality of the model will be discussed in detail. A comparative analysis will be conducted and evaluated across various learning algorithms. The annual features and the learning algorithms will be examined for accurate prediction values. The Spark and MapReduce framework will be examined across typical job distribution for time and space complexities.

ISSN 2394 - 7780

Year	ANNUAL
1904	19.51
1908	19.44
1912	19.25
1916	19.22
1920	19.03
1924	19.51

1960	19.64
1964	19.34
1968	19.02
1972	19.02
1976	19.37
1980	19.07
1984	19.54

DATASET ANALYSIS

The dataset used for this research work contains temperature related information. It comprises of annual temperature computed from the month of January to December. It also includes seasonal variations occurred in various months of a year. These are categorized as January to February, March to May, June to September and October to December. The dataset measured for analysis has hundred and fifteen years of temperature data from the year 1904 to 2016. This information is applied for computing and predicting the future years. Table 5.1 consists of temperature dataset that is used as training data set. Machine Learning algorithms are trained on this data set to predict the values for later years. Figure graphically represents the annual temperature data for various years

MACHINE LEARNING ALGORITHMS FOR PREDICTION ON DATASETS

The model proposed in this research is implemented on various Machine Learning algorithms. The predictions obtained from the learning algorithms are illustrated through graphical representations. The various Machine Learning algorithms used for prediction are:

- Linear Regression
- Decision Tree
- Random Forest
- Gradient Boosting Tree

LINEAR REGRESSION ALGORITHM

The model is implemented using the linear regression algorithm. The graphical representation of data by comparing the annual temperature with prediction is shown in the figure.

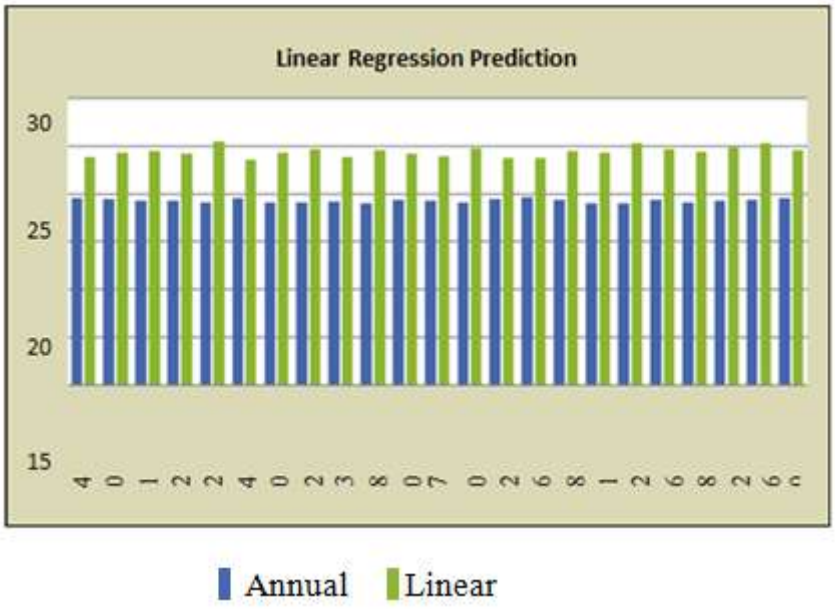


Figure –: Annual Temperature and Prediction

Observation: Annual Temperature is compared against the prediction using linear regression with an accuracy of 72.5%. The computations drawn from Linear Regression algorithms are predictions and RMSE. The obtained prediction measures are less accurate and the graphical mapping shows wide deviations, which results in higher error rate with minimal utilization of time and space.

ISSN 2394 - 7780

Figure illustrates the annual temperature and prediction for various years. It represents the annual temperature taken from the dataset where the prediction values are computed from the linear regression algorithm. The prediction values deviates showing large variance which results in higher error rate.

DECISION TREE

The tree structure is obtained through the prediction outcomes using Decision Tree algorithm. The graphical representation as in figure illustrates the variations

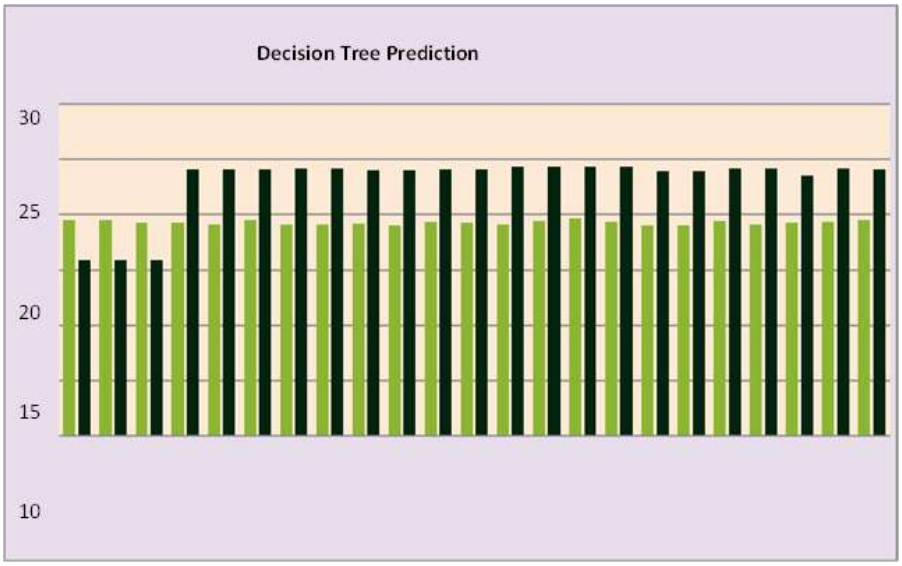
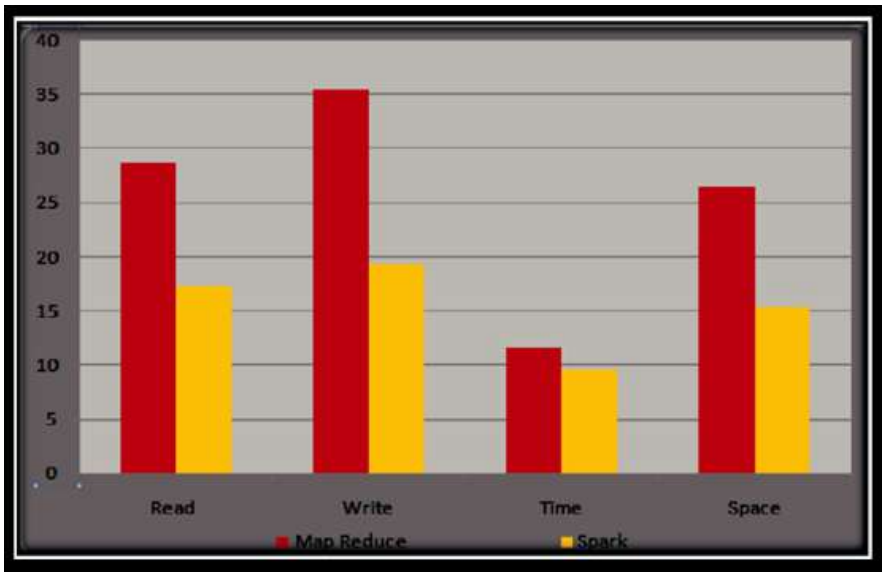


Figure –: Read, Write, Time and Space Complexity Comparison between Spark and MapReduce

The MapReduce model exhibits high disk rates for read operations with a result of 28.65 MB/Sec and for write operations with 35.43 MB/Sec. It shows reduced performance on a processor. While executing the jobs on Spark platform, the results are 17.23 MB/Sec on read and 19.41 MB/Sec for write operations. Table 5.7 and Figure 5.13 shows the combined measures of learning techniques considering the time, space, read and write operations. Hence Spark shows 65.21% of improvement on MapReduce framework.



Limitations of the Study:

The following could be the various limitations and delimitations of this research:

1. **Time and Space as Complexity Features:** The data collected for analysis is massive and multifaceted. Time and Space are considered for evaluating the complexities of read and write operations while implementing on Spark project.
2. **Data Size:** Enormous data is collected as a training set to learn Machine Learning methods. A collection of hundred and fifteen years of temperature data, which consist of both annual and seasonal, is used for predictive analysis.

ISSN 2394 - 7780

3. **Predictive Modeling:** Different models of analytics can be applied on a dataset. But this research work focuses on predictive modeling. According to this model, data is partitioned to train test sets for predictive analytics.

CONCLUSION

This study provides a clear understanding of model-based performance analysis for weather prediction. The proposed model is simple yet effective and identifies areas where performance can be improved. Incorporating additional features such as humidity, moisture, fog, and pollution can help achieve more accurate temperature predictions in the future. These models benefit from the rapid growth in computing power and can provide better prediction results. Applying this approach can make predictive models more accessible and useful for future forecasting. Accurate weather predictions are important for the economy, helping farmers plan and alerting emergency services to take action during extreme conditions. Future research can extend this work to hourly or daily predictions and include specific geographical regions to provide more localized forecasts

BIBLIOGRAPHY

[1] Burhan U.I.Khan., Rashidah F., and Hunain A. et.al, (2014) 'Critical Insight for MapReduce Optimization in Hadoop' in International J of Computer Science and Control Engineering, Vol.2, Issue-1, pp: 1-7

[2] Erik R., Christos F. and David N. (2001) 'Active disk for large scale data processing', IEEE. J., pp 68 - 74 June.

[3] Douglas T., Todd T., and Miron L. (2004) 'Distributed computing in practice: The condor experience' in Concurrency and Computation Conference.

[4] Dean J. and Sanjay G. (2004) 'MapReduce: Simplified data processing on large clusters' in ACM Conference OSDI.

[5] Milos Hauskrecht (2017),'Ensemble Learning' in Book Chapter 12 Data Mining pp: 479-501.

[6] John B., Douglas T. and Miron L. (2004) 'Explicit control in a batch aware distributed file system' in USENIX symposium on Networked system design and implementation, Conference March.

[7] Huston L., Sukthankar R. and Ailamaki A. (2004) 'Diamond: A storage architecture for early discard in interactive search' in USENIX file and storage technology Conference April.

[8] Alekh J. and Dittrich J et al. (2011) 'Trojan Data Layouts: Right Shoes for a Running Elephant' in SOCC ACM Conference.