A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS

PREETI SINGH
Registration No: 29821015
Research Scholar
Shri JJT University
Rajasthan.

DR. SHAILESH KUMAR
Professor
Shri JJT University
Rajasthan.

DR. MULUGANDLA SRIDEVI
Associate Professor
CVR College of Engineering
Hyderabad.

ABSTRACT

The main objective of human evolution has always been to look for ways to mold the nature to satisfy our needs. A key milestone in this regard is the invention of a machine – called the computer that can complete a task given to it in fraction of time taken by an average human. While that sounds great, the only drawback is that the decision must still be taken by a man who is bound by limitations of the human body. The run to reap the complete benefits has given rise to what is called the Artificial Intelligence. Machine learning is a part of AI, which deals with imparting knowledge to the computer through various related examples. Throughout the years, various machine learning algorithms have been developed each with their own merits and demerits. This paper is a consolidated effort to bring together different ML algorithms like linear regression, KNN (k- nearest neighbours) etc. This research paper discusses the most recent developments in these areas of study and tries to define the best applications for each of those based on previous researches.

Keywords: *Machine Learning Algorithms, KNN, Linear Regression, Deep Learning, SVM, RF, Activation functions*

INTRODUCTION

Machine learning is one of the fastest-growing areas of computer science with long-range applications, which refers to the automatic detection of significant patterns in data with machine learning tools, which give programs the ability to learn and adapt.

Machine learning has become one of the pillars of information technology and, with that, a reasonably central, though generally hidden, part of our life. With the increasing amount of data available, there is a good reason to believe that intelligent data analysis will be even more widespread as a necessary ingredient for technological progress.

Data mining and machine learning go hand in hand with which several ideas can be derived through appropriate learning algorithms. There has been significant progress in data mining and machine learning as a result of the evolution of nanotechnology, which generated curiosity to find hidden patterns in the data to obtain results. The fusion of math and statistics, machine learning and artificial intelligence, information theory and big data, and hight processing computation, has created a reliable science, with a firm mathematical base and compelling tools.

Since their evolution, humans have been using many types of tools to accomplish various tasks in a simpler way. The creativity of the human brain led to the invention of different machines. These machines made the human life easy by enabling people to meet various life needs, including travelling, industries, and computing. And Machine learning is the one among them. According to Arthur Samuel Machine learning is defined as the field of study that gives computers the ability to learn without being explicitly programmed. Arthur Samuel was famous for his checkers playing program. Machine learning (ML) is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the extract information from the data. In that case, we apply machine learning. With the abundance of datasets available, the demand for machine learning is in rise. Many industries apply machine learning to extract relevant data. The purpose of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves without being explicitly programmed. Many mathematicians and programmers apply several approaches to find the solution of this problem which are having huge data sets. Machine Learning relies on different algorithms to solve data problems. Data scientists like to point out that there‟s no single one-size-fits-all type of algorithm that is best to solve a problem. The kind of algorithm employed depends on the kind of problem you wish to solve, the number of variables, the kind of model that would suit it best and so on. Here‟s a quick look at some of the commonly used algorithms in machine learning (ML)

LITERATURE REVIEW

R Prakash[2022] As Machine Learning algorithms are becoming popular to solve everyday challenges, it is being used to make predictions on data across different industries. To obtain an advantage over the competitive pressures, a service or goods supplying firm needs a strong marketing strategy to understand their clients and grasp how people think, reason, and choose among the various items and services supplied in the market. By effectively understanding customer's purchasing behavior, a firm can predict the future purchases of the customer enhance user experiences, design marketing strategies to create new consuming markets, and boost sales revenue. The purpose of this research is to investigate the purchasing behavior of e-commerce retailer customers to enable to provide better services and products. The data analysis conducted on customer data enables understanding of the type of products the customers are likely to purchase based on the description of their previous transactions.

Rohan Kumar C L[2022] Fraudulent transactions have a huge impact on the economy and trust of a block chain network. Consensus algorithms like proof of work or proof of stake can verify the validity of the transaction but not the nature of the users involved in the transactions or those who verify the transactions. This makes a block chain network still vulnerable to fraudulent activities. One of the ways to eliminate fraud is by using machine learning techniques. Machine learning can be of supervised or unsupervised nature. In this paper, we use various supervised machine learning techniques to check for fraudulent and legitimate transactions. We also provide an extensive comparative study of various supervised machine learning techniques.

Kavita Bisht[2020] In this paper, we will explore and try to solve the bird species classification problem. We have use the publicly available bird datasets, which provides us with the bird images to classify the species. The difficulty of this task arises from the inter-class similarities of species of birds that often fool even the expert bird watchers. This dataset has been widely used to develop approaches towards finegrained classification, leaving us with the task of understanding how geometry in the datasets relates to such classification algorithms. Fine-grained classification often describes an end-to-end pipeline, from image to class prediction. The dataset, however, provide labels for image's classes of various breeds of birds. In our paper, we have use the supervised learning algorithm of Machine learning wherein we have make use of the labelled dataset to decide the class or breed of the bird.

Bharathi Ramesh[2019] In the present digital era massive amount of data is being continuously generated at exceptional and increasing scales. This data has become an important and indispensable part of every economy, industry, organization, business and individual. Further handling of these large datasets due to the heterogeneity in their formats is one of the major challenge. There is a need for efficient data processing techniques to handle the heterogeneous data and also to meet the computational requirements to process this huge volume of data. The objective of this paper is to review, describe and reflect on heterogeneous data with its complexity in processing, and also the use of machine learning algorithms which plays a major role in data analytics

Machine Learning

Machine learning, then, is about making computers modify or adapt their actions (whether these actions are making predictions, or controlling a robot) so that these actions get more accurate, where accuracy is measured by how well the chosen actions reflect the correct ones. Imagine that you are playing Scrabble (or some other game) against a computer. You might beat it every time in the beginning, but after lots of games it starts beating you, until finally you never win. Either you are getting worse, or the computer is learning how to win at Scrabble. Having learnt to beat you, it can go on and use the same strategies against other players, so that it doesn't start from scratch with each new player; this is a form of generalisation.

Linear Regression Algorithm

Regression is an approach of supervised learning. It can be used to model continuous variables and do the predictions. Examples of application of linear regression algorithm are the following : prediction of price of real-estate, forecasting of sales, prediction of students' exam scores, forecasting of movements in the price of stock in stock exchange. In Regression we have the labeled datasets and the output variable value is determined by input variable values - so it is the supervised learning approach. The most simple form of regression is linear regression where the attempt is made to fit a straight line (straight hyperplane) to the dataset and it is possible when the relationship between the variables of dataset is linear.

Multinomial Naïve Bayes Algorithm

Naive Bayes is a family of algorithms based on applying Bayes theorem with a strong (naive) assumption, that every feature is independent of the others, in order to predict the category of a given sample. They are probabilistic classifiers, calculate the probability of each category using Bayes theorem, and the category with the highest probability is output. MNB is a probabilistic classifier, meaning that for a document d, out of all classes ck ∈C the classifier returns the class ck which has the maximum posterior probability. MNB is always a preferred method for any sort of text classification as taking the frequency of the word into consideration, and get back better accuracy than just checking for word occurrence.

K Nearest Neighbour Algorithm

KNN is a non-parametric, lazy learning algorithm. To identify the sentiment of new test document KNN classifier computes the similarity between a new test document and every training document. Then KNN classifier sort the training documents in descending order of their similarity to the test document in order to pick the top K most similar training documents with a test document. Finally the KNN classifier assigns this new test document to a sentiment category that has the highest score of similarity.

Support Vector Machines Algorithm

A Support Vector Machine (SVM) performs classification by finding the hyper plane (classifier) that maximizes the margin between the two classes subject to the constraint that all the training tuples should be correctly classified. Hyper plane is defined by using the data points that are closest to the boundary. These points are called support vectors and the decision boundary itself is called support vector machine. The main advantage of SVM classifier is that it minimizes the training set error and the test set error. To obtain a SVM classifier with the best generalization performance, appropriate training is required. The most commonly used and popular algorithm for training SVM is the SMO algorithm. The main advantage of SMO algorithm is that it works analytically on a fixed size working set by decomposing the large training data set. So, that it can works fine even for large data sets and it also gives superb performances in almost all kinds of training data sets

METHODOLOGY

In research work mentioned in datasets are collected, on which classifier models are applied and results are evaluated. Whereas, in certain feature selection techniques are applied before the data classification along with fine tuning is done. Based on these methodologies the proposed methodology is constructed and described as follows: Step number two is conversion of values. Input parameters are demographic and academic data. Output is class consisting respective grade categories. The dataset is imported and converted to numeric format in order to apply Pearson correlation between all input attributes to target attribute. The attributes having correlation less than 0.05 are eliminated. Attributes with high correlation are used for further processing. Step number 4 is application of classification technique. The J48 and C5.0 are applied in selected attributes and predictions are performed. Various evaluation measures such as precision, recall, FPR (False Positive Rate), TPR (True Positive Rate) are used for analysis.

RESULTS

Figure 1 represents the outcome of linear regression where orange line indicates the actual value and the Green line indicates the predicted value and Figure 2 represents KNN outcome where Blue line indicates the actual value and the Orange line indicates the predicted value. Figure 3 represents the result of Fb Propnet where Orange line indicates the actual value and the Green line indicates the predicted value and Figure 4 represents LSTM outcome where Red line indicates the actual value and the Green line indicates the predicted value

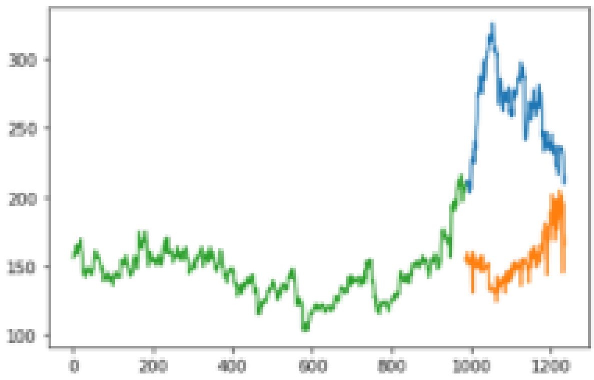

Figure 1: Linear regression outcome

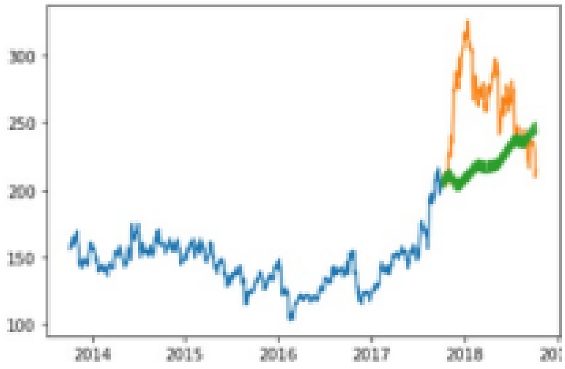

Figure 2: K nearest neighbor outcome

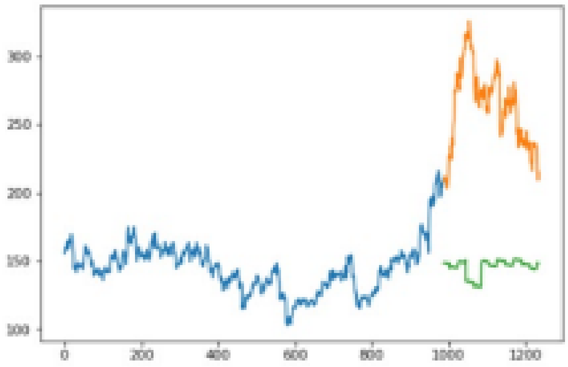

Figure 3: Fb prophet outcome

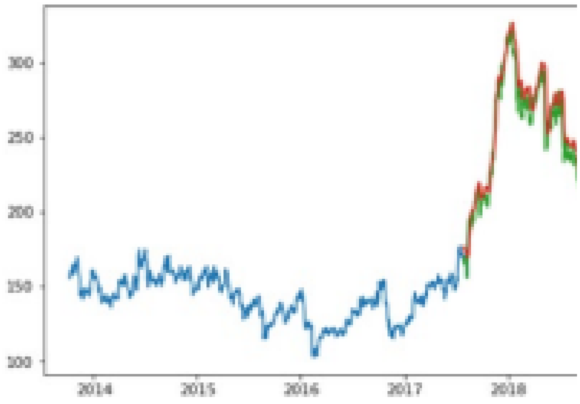

Figure 4: Long short-term memory outcome

RMSE scores

The RMSE is the square root of the difference of the residuals. It indicates the outright fit of the design to the data-- just how close the observed data factors are to the design's anticipated values. RMSE can be taken the standard deviation of the unusual difference, as well as has the beneficial property of remaining in the exact same units as the reaction variable. Reduced worth's of RMSE indicate much better fit. RMSE is an excellent measure of how properly the design forecasts the reaction, and also it is the most crucial requirement for fit if the primary function of the design is prediction.

CONCLUSION

This section focuses on various seasons of rainfall database from 1901 to 2020 by using Machine Learning (ML) classification Algorithm and studying each of them. Three classification methods are based on accuracy and kappa statistics and they are visualized with different levels of rainfall data collected from the India Meteorological Department. The purpose of this research is to determine which classifier is the most effective. The accuracy of various classifiers for the southern states of India is compared and the sensitivity, specificity, accuracy, true positive rate, and false positive rate of each classifier for all states are calculated. Furthermore, a comparison of kappa statistics is conducted by utilizing a confusion matrix. To analyze the performance of the most popular classification techniques, the training dataset is used to train the classifier using classification and regression. The accuracy of the Naïve Bayes approach, K Nearest

Neighbor algorithm and Support Vector Machine (SVM) are tested on the test dataset, and the results show that the SVM model has the best performance. The Naïve Bayes Classifier has also performed well, but the KNN algorithm did not.

Reference

1. *R Prakash[2022], "Comparative Study Of Machine Learning Algorithms For Product Recommendation Based On User Experience", ECS Transactions, ISSN 1938-6737 , Volume 107, Number 1*
2. *Rohan Kumar C L[2022], "Comparative Study of Machine Learning Algorithms for Fraud Detection in Blockchain", International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), ISSN 2581-9429,Volume 2, Issue 2,*
3. *Asif Raza[2021], "COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR FAKE REVIEW DETECTION", International Journal of Computational Intelligence in Control, ISSN: 0974-8571, Vol.13, No. 1*
4. *Muhammad Bello Aliyu[2020], "Comparative Study of Some Supervised Machine Learning Algorithms for Information Retrieval ", Saudi Journal of Engineering and Technology, ISSN 2415-6264, 5,(3), 106-113*
5. *Kavita Bisht[2020], "Comparative Study of Classification Performance Accuracies of Machine Learning Algorithms", International Research Journal of Engineering and Technology (IRJET) ISSN: 2395-0056,Volume: 07, Issue: 05*
6. *Nawaraj Paudel[2020], "Comparative Analysis of Machine Learning Based Classification Algorithms for Sentiment Analysis", International Journal of Innovative Science, Engineering & Technology, ISSN 2348 – 7968, Vol. 7, Issue 6,*
7. *G.ROBINSON PAUL[2020], "MACHINE LEARNING ALGORITHMS USING STOCK MARKET DATASET–A COMPARATIVE STUDY", JOURNAL OF CRITICAL REVIEWS ISSN- 2394-5125, VOL 7, ISSUE 15*
8. *Xin Fang[2019], "A Comparative Study of Machine Learning Algorithms in Predicting Severe Complications after Bariatric Surgery", J Clin Med, ISSN 2077-0383, 8,(5), doi: 10.3390/jcm8050668*
9. *Bharathi Ramesh[2019], "MACHINE LEARNING ALGORITHMS FOR HETEROGENEOUS DATA: A COMPARATIVE STUDY", International Journal of Computer Engineering & Technology (IJCET), ISSN: 0976–6375, Volume 10, Issue 3, pp. 9-19,*
10. *Chung-Chu[2017], "A Comparative Study on Machine Learning Algorithms for Network Defense", Virginia Journal of Science, ISSN: 0042-658X, Vol. 68, No. 3,*