iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.43366

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Performance Analysis of Machine Learning Algorithms using Fake News Detection

Dr. Senthil Kumar M[1], Mr. Sakthivel S[2], Mr. Sasitharan M[3], Mr. Shakir Ahamed M[4]
[1]*Associate Professor, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Chennai, Tamil Nadu, India*
[2, 3, 4]*Department of Computer Science and Engineering, SRM Valliammai Engineering College, Chennai, Tamil Nadu, India*

Abstract: ***Machine learning has vast algorithms in which each and everything is specialized to predict and compute certain functionalities and tasks. A neural network is a collection of neurons in which every neuron holds a numerical value as the output of other neurons. Furthermore, neural networks are classified more as regular neural networks, convolutional neural networks, feed-forward neural networks, and long-term memory networks. Each is specialized in unique scenarios, such as regular neural networks work better in position-based image detection and convolutional neural networks work better in edge-based detection. Therefore, this paper bases the comparison of different algorithms on a base prediction problem, which is fake news detection. This project displays various performance metrics such as accuracy score and various visualization plots like scatter, pie, and bar. so that users will know the different algorithms and their scalability on non-relatable text patterns.***
Keywords: Fake news detection, Machine learning, Naïve bayes Gaussian NB, Classification, Regression, Support Vector Classifier, Decision Tree Classifier.

I. INTRODUCTION

The fields of machine learning and deep learning basically read and understand the patterns and flaws in the data we feed into them and Predictions are based on the input we provide, and machines can be programmed to make their own decisions without any human interference. They learn from their errors and rectify them. there are algorithms that are specifically good at certain fields of prerequisite. For example, let us consider a dataset that is related to the chance of heart disease and has an attribute of whether a person is prone to heart attack or not. Considering the required feature sets, regular neural networks (RNN) can produce good accuracy with low loss, and on the other hand, ML algorithms such as logistic regression can project good metrics. These two above-mentioned algorithms perform at their best as long as the data is linear (numerical or categorical). What if the data we have is an image, video, or audio? In this case, videos are nothing but a continuous frame of images, so let us consider that images and videos are of the same format. Image processing totally depends on the position or the edge-based classification. Regular neural networks can deal with position-based prediction, but this may not help in every case, so here come convolutional neural networks. They classify things based on edges. To be more specific, depending on the dataset and the problem statements, different algorithms can perform differently.

II. RELATED WORKS

There are various machine learning algorithms used to solve the fake news detection problem. Xavier Jose et al. discusses the many features and forms of false news as well as a practical approach for detecting fake news on OSM networks. The two key components of the approach are the stance detection model and the created content classifier. The posture detection model achieved an accuracy of 90.37 percent using Logistic Regression, and the manufactured content classifier achieved an accuracy of 93.46 percent using Bi-directional LSTM [10]. Zhang Bao-wen et al. states that regression algorithms can be used in many classification issues. The purpose of this article is to look at how temperature affects the selling of iced items. To begin, the author gathered data on last year's anticipated temperature and iced product sales, followed by data compilation and purification. Using data mining theory, the author created a mathematical regression analysis model based on the cleaned data [5]. Sarker. I.H explains the concepts of numerous machine learning techniques and their usefulness in a variety of real-world application areas, including cybersecurity, smart cities, healthcare, e-commerce, agriculture, and many more. Based on his findings, he also emphasizes the obstacles and future research avenues [8]. Saima Akhter et al. proposes to computerize the fake news detection in Twitter datasets, this research provides a strategy for spotting fabricated news messages from Twitter tweets by figuring out how to anticipate accuracy evaluations. After that, the author compared five well-known Machine Learning techniques, including Support Vector Machine,

Nave Bayes Method, Logistic Regression, and Recurrent Neural Network models, to show how efficient the classification performance on the dataset was.

The SVM and Nave Bayes classifiers outperformed the other methods in the experiments [2]. R. Saravanan et al. examines supervised learning methods that are commonly employed in data classification. The strategies are evaluated in terms of their goals, methodologies, benefits, and drawbacks. Finally, readers will get an understanding of supervised machine learning approaches to data classification [6]. Hadeer Ahmed et al. proposes a false-news detection model based on n-gram analysis and machine learning approaches is presented in this study. Two different feature extraction strategies and six different machine classification algorithms are investigated and compared. The best results were obtained with Term Frequency-Inverted Document Frequency (TF-IDF) as a feature extraction technique and Linear Support Vector Machine (LSVM) as a classifier, with an accuracy of 92 percent [3].

III.MATERIALS AND METHODS

A. Various library packages are available to process text, to split the data into training and testing, and to feed the training data to the algorithms. Finally, some visualisation plots and performance metrics such as accuracy and losses are presented.

1) *Regular Expression:* A character sequence that generates a search pattern. The purpose is to check whether a string contains a specified search pattern.
2) *Natural Language toolkit:* It is a text processing package in Python where it plays only with categorical data. It performs various functions such as tokenization, classification, stemming, parsing, semantic reasoning, and tagging. NLTK provides access to many algorithms to get something done, whereas spacy provides a vision for the best way to do it.
3) *CountVectorizer:* The text data that users provide can't be taken as it is by the machine. To understand the text messages, the first and foremost thing is that they must be converted into machine-understandable language, and then it must perform the other task as per the code instruction. To achieve this, the CountVectorizer method is called. It converts text data into numbers. To make this point clear, let us consider a string str="How are you doing?" and it is transformed into a sparse matrix.
4) *Train_Test_Split:* One of the most important parts of machine learning programmes is data. How we compute and evaluate it matters a lot. Sci-kit Learn has a method where it splits the data into training and testing data. Training data is something we fed into the proposed model will learn a variety of parameters, including patterns, outliers, and insights. Testing data is provided by the users to check the model performance, like accuracy and loss. It totally depends on the performance of the model.
5) *accuracy_score:* The most significant factor in the model is accuracy. To achieve good accuracy, the model must have been properly manipulated. Datasets, model architecture, and last but not least, the methods and modules that are used in it. The more accurate the model provides, the more it is considered top-notch performance.

B. Algorithms

1) *Gaussian NB:* Gaussian NB is a special type of Naive Bayesian algorithm, which is specially developed to perform when the features have a continuous value. This is mostly used in text classification and problems with multiple classes. It is applied based on Bayes theorem and with the naive assumption of conditional independence between every pair of features.
2) *Logistic Regression*: Logistic regression is an algorithm for classification problems. It is similar to linear regression. At first, it draws a straight line in the graph using computed slope, and then comes a sigmoid function where an S-shaped curve is plotted. The point where the straight line and sigmoid curve intersect is called the pivot point.
3) *Decision Tree Classifier:* An algorithm that can be used for both classification and regression problems comprise nodes linked together and subdivided further downwards into sibling nodes or child nodes, and the terminating node is called the leaf node. Each node has a test on attributes, and each branch represents the possible outcomes. Finally, the leaf node represents the labels to which an attribute belongs.
4) *Support Vector Classifier:* A vector starts from the origin and connects to the data labels that are being plotted in the graph. Secondly, draw a vector from the farthest point in class A to the closest point in class B by drawing a unit perpendicular line that touches both the points. Now we have a line that separates labels. A hyperplane is created first, and then every other mathematical computation happens. Support vector machines are classified into 2 types: SVC (Support vector classifier) and SVR (Support vector regression).
5) *K- Neighbors Classifier:* The given data labels are plotted in the graph, and the points which are adjacent or neighboring to each of them are considered as one class, and so on, compared with every point being considered as the same class. If a user is entering new data, it is compared with the existing data for the shortest distance between them and considers that point to belong to that class.

IV. IMPLEMENTATION

The Jupyter Notebook is the platform used to implement the code. There is 50+ kernels running in the notebook, each containing visualization, performance metrics checking, and comparison. The final outcome of the project is a confusion matrix, which includes the accuracy and loss of each algorithm and plots those on a linear graph and scatters (Accuracy vs Loss). As the dataset used in the model is not balanced, let us consider an example. Suppose a person has a shoe factory and has 100,000 shoes, of which 99,000 are Adidas and 1,000 are Nike. Suppose a client needs a project where the model must be able to classify the shoe as two labels, whether it is Adidas or Nike, but the model always predicts Adidas no matter what brand and has an accuracy of 99%. It makes no sense in this case to provide high accuracy, but the model fails to predict correctly. So here the confusion matrix is projected. It has 2 columns and 2 rows, namely, true positive, true negative, and predicted positive and predicted negative. It shows a two-dimensional array of true and false values.

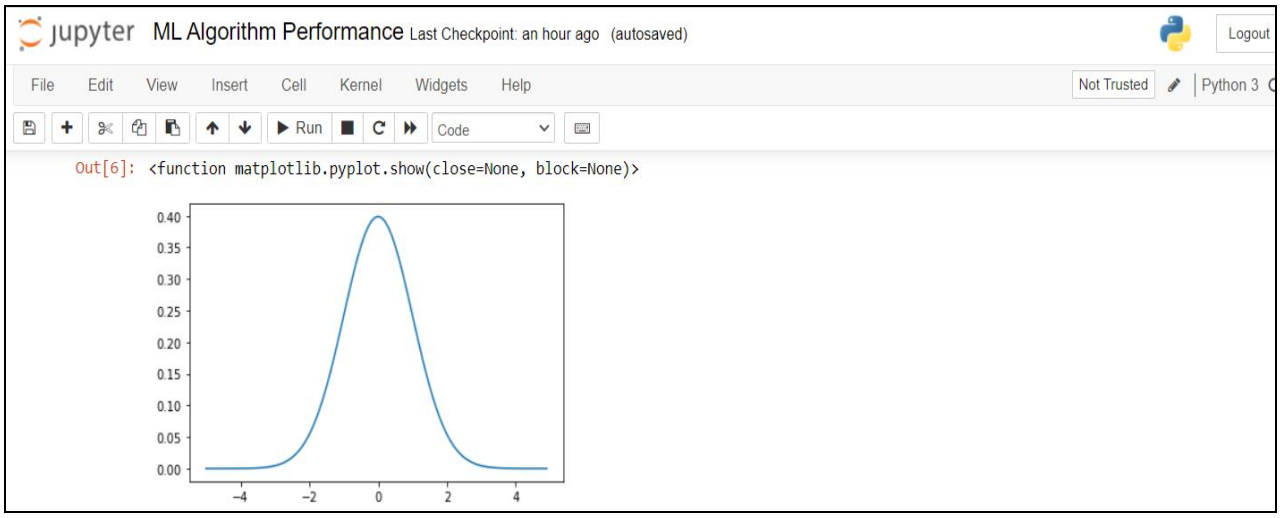


Fig. 4. 1

Fig.4.1 describes the bell-shaped graph which represents the Naive Bayes Graph with a model accuracy of 0.883, and an error of 0.117.

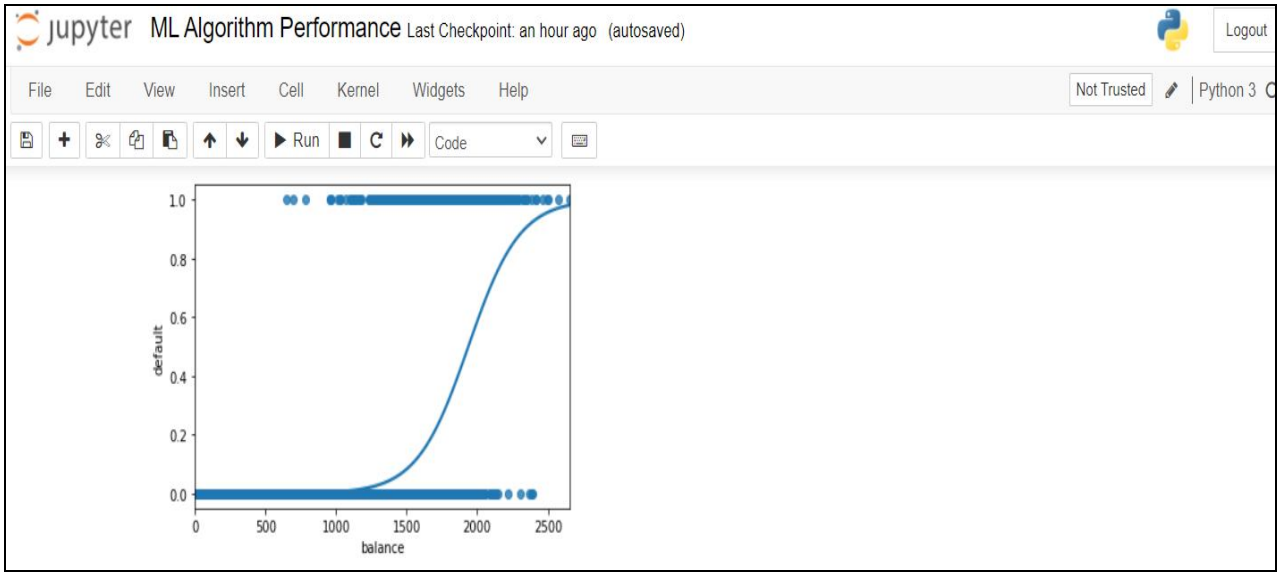


Fig. 4. 2

Fig.4.2 represents logistic regression graph, it first plots a linear graph with the computed slope and plots a sigmoid curve. the point at which the straight line and sigmoid curve intercepts are called a pivot point, Accuracy of 0.94, and loss of 0.05.

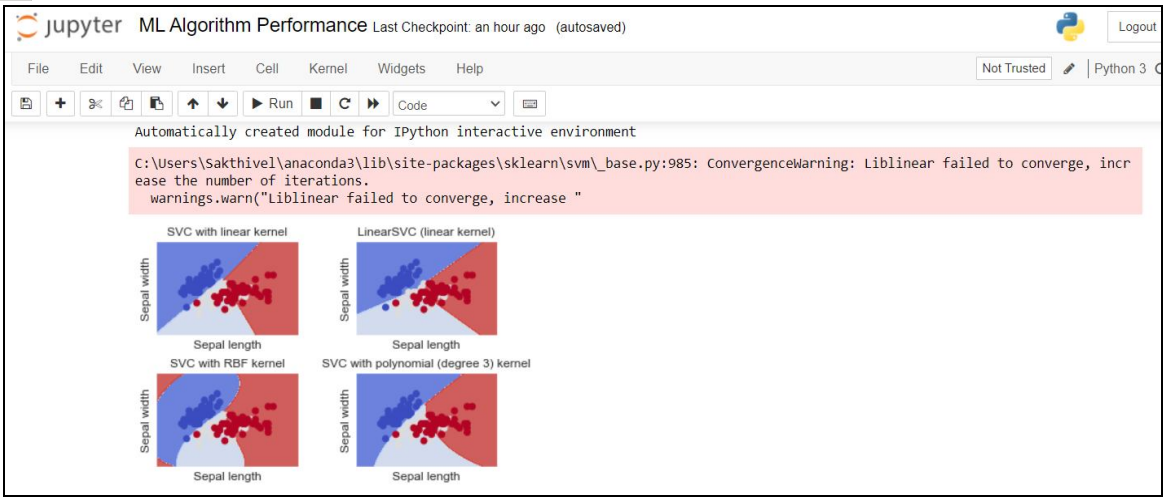


Fig. 4. 3

Fig 4.3 represents various subplots of the support vector machine; a hyperplane is used to separate the target labels differently.

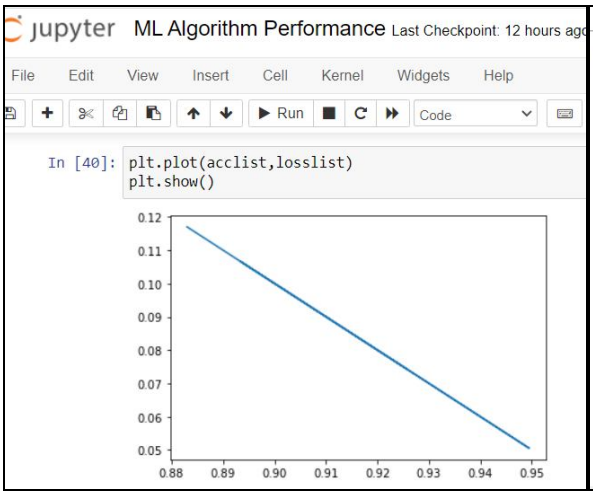


Fig 4.4

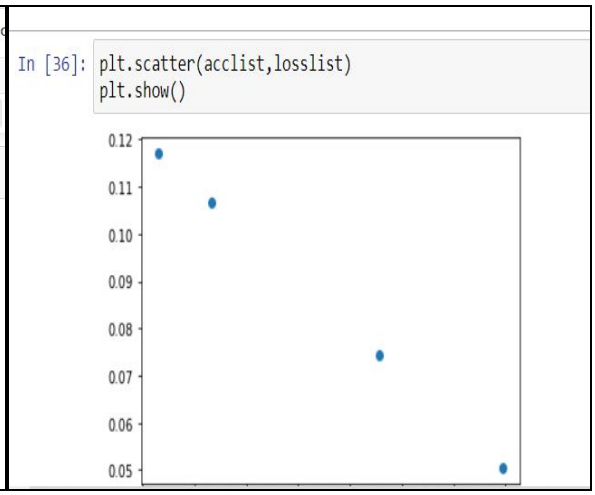


Fig 4.5

Fig. 4.4 is the performance graph (Linear Graph) in which it is plotted as accuracy on the x-axis and loss on the y-axis, Fig. 4.5 represents the scatter of same accuracy and loss represents similar linear points as like the Linear Graph.

V. DISCUSSION

Starting from the environment being coded to the final output system projects, everything is done in Python. The Jupyter notebook is used as an integrated development environment as it facilitates kernels that will be useful for machine learning and deep learning. And coming to the part of coding, there is the usage of various library packages, which include re (Regular Expression), nltk (Natural Language Processing Toolkit), Matplotlib, Pandas, and Scikit Learn. As this project deals with text processing, nltk plays a vital role. It cleans the text and forms corpus, such as removing symbols and stopwords by using the Porter-Stremer method. The program reads the data in the form of numbers, although the data provided by the user is in text, so no matter which format the data is in, it must be converted to the numerical format. CountVectorizer is a method in the sklearn.feature_extraction set that performs the required conversion of text to numbers. Furthermore, the data is split into training and testing derived from a method model_selection with a testing size of 20% of the data we feed, and after that, various algorithms are applied to check their performance by visualizing it, comparing its accuracy, and finally a confusion matrix. Performing with a low accuracy of 0.883 (Gaussian Naive Bayesian) and a high accuracy of 0.9495 (Logistic Regression). Logistic regression works based on linear regression.

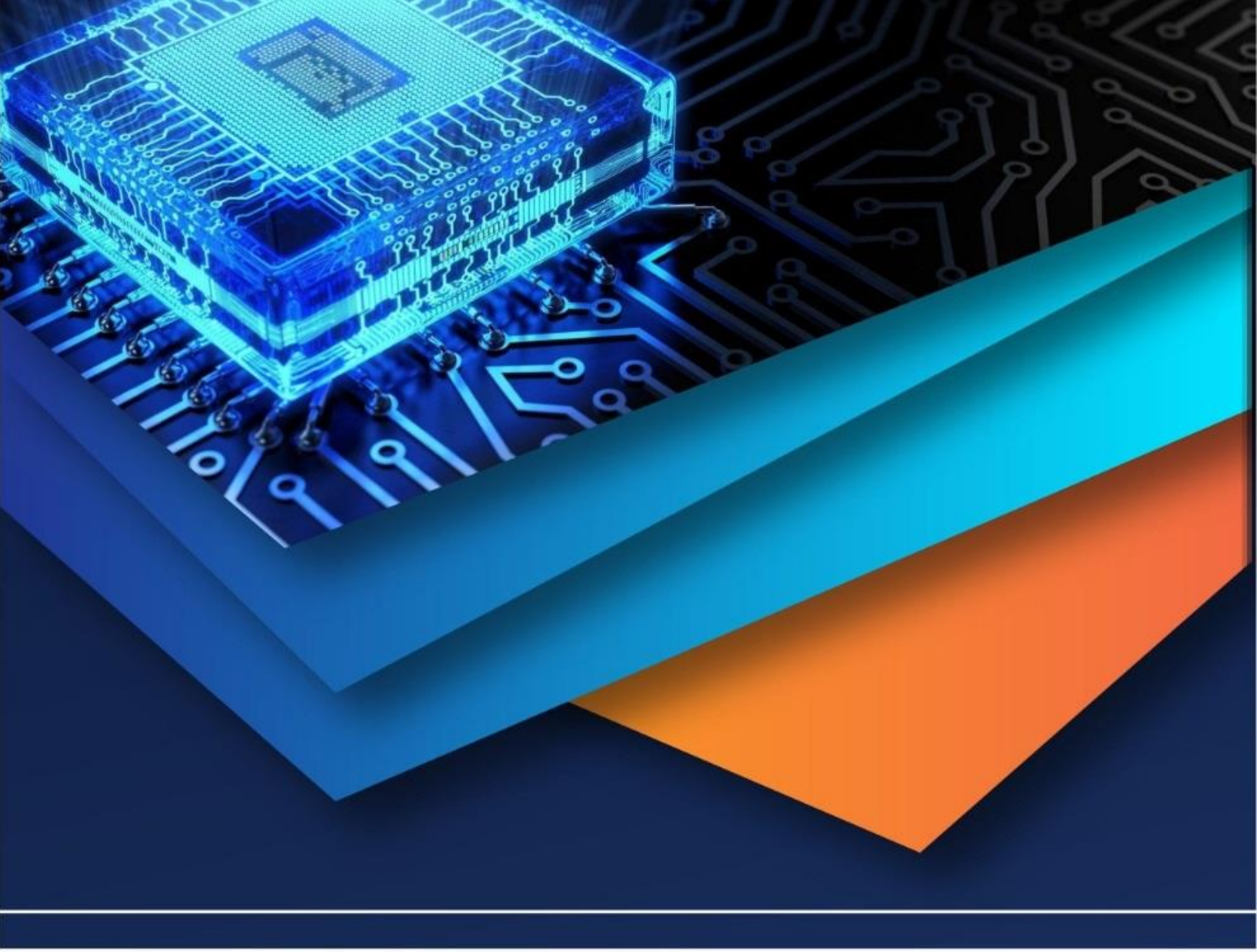
The only difference is that logistic regression has a sigmoid graph and a pivot point. On the other hand, Gaussian NB is based on conditional probability and the Bayes theorem. We can observe that logistic regression works for binary class labels (male and female, black and white, fake and legit). Gaussian NB performs worst when it comes to binary label predictions as they are performed on a given condition that has already occurred.

VI. CONCLUSION

Machine learning is a vast domain with numerous algorithms to work with, each specialising in a specific area problem statement. As a result, developers have a wide range of options for selecting an algorithm and incorporating it into their projects. our proposed model applies 6 different algorithms to the model, which are fake news detection and computing its performance. parameters including accuracy, loss, and the confusion matrix. Let's say convolutional neural networks (CNN) are good at image processing and NLTK is good at text processing. For image processing, it is good to use conventional neural networks whereas NLTK is good at processing text. In addition to this, the model deals with CountVectorized at first followed by six different algorithms to process data and accuracy in order to find out which algorithms are best performing and which are worst-performing. With the values of accuracy versus loss, a graph is plotted to find out the relationship between the accuracy of six different algorithms and to suggest which is best and worst-performing in text processing.

REFERENCES

[1] Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq et al., Detecting fake news using machine and deep learning algorithms, Intl Conf. on Smart Computing & Communications, 2019, Pp.43-75.
[2] Hadeer Ahmed, Issa Traore and Sherif Saad, Detection of online fake news using n-gram analysis and machine learning techniques, Intl conf. on intelligent secure and dependable system, 2017, Pp. 145-167.
[3] Xichen Zhang and Ali A.Ghorbani, An overview of online fake news Characterization detection, Information Processing and Management, 2020, Pp.45-56.
[4] Zhang Bao-wen, Shen Rong, The research of regression model in machine learning field, MATEC Web of Conferences, 2018, Pp-95-110.
[5] R. Saravanan, Pothula Sujatha, A Perspective of Supervised Learning Approaches in Data Classification, Second Intl Conf. on Intelligent Computing and Control Systems, 2019, Pp-78-102.
[6] Sarker.I.H, Machine Learning: Real-World Applications and Research Directions, 2021, Pp-76-89.
[7] Harsh Patel, urvi Prajapati, Study and Analysis of Decision Tree Based Classification Algorithms, 2018, Intl Jrnl of computer sciences and engineering, Pp-65-80.
[8] Xavier Jose, S. D Madhu Kumar, Priya Chandran, Characterization, Classification and Detection of Fake News in Online Social Media Networks, Mysore Sub Section Intl Conf. ,2021, Pp-30-43.
[9] Alessandro Bondielli And Francesco Marcelloni, A Survey On Fake News Detection Techniques, 2019, Information Science, Pp. 78-85.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY