iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XI **Month of publication:** November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65138

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 12 Issue XI Nov 2024- Available at www.ijraset.com

Comparative Analysis of Machine Learning Algorithms: KNN, SVM, Decision Tree and Logistic Regression for Efficiency and Performance

Binita Brijal Acharya[1], Ghevariya Devam Shaileshbhai[2]
[1]*HOD (Information Technology), Tapi Diploma Engineering College, Surat, Gujarat, India*
[2]*Diploma Student, Information Technology, Tapi Diploma Engineering College, Surat, Gujarat, India*

Abstract: ***This paper presents a comparative study of machine learning models namely K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Logistic Regression on two datasets: 20Newsgroups and Wine datasets. Based on basic accuracy, and precision, and recall of the models, F1 measures were assessed. When compared with other classifiers such as KNN and Decisions Trees, SVM and Logistic regression gave better results especially in the case of the 20Newsgroups dataset dominated by textual high dimensional data. KNN had poor recall results and Decision Tree was moderate. In the Wine dataset, since the structure of data is comparatively less complex in our context then all the models yielded almost similar results with accuracy and precision factors very close to 1.0 which of course manifested that the choice of the model does not affect much on simple data. These results stress the importance of the choice of the appropriate model for tasks with a certain level of data complexity; detailed models show the greatest efficiency at the accomplishment of complicated tasks, though at the same time, they are not required for simple, structured data.***

Keywords: Machine Learning, Model Comparison, Text Classification, SVM, Logistic Regression, KNN, Decision Tree.

I. INTRODUCTION

In the ever-widening field of machine learning, the choice of the appropriate model on which to train is decisive when it comes to getting the best results. Such algorithms enumerated above make it necessary to evaluate the strengths and limitations of different models in view of various criteria including accuracy, computation time and generalization ability. From the four most commonly used algorithms, K-Nearest Neighbors is employed in numerous projects, Support Vector Machines has various uses, and Decision Trees and Logistic Regression depend on quite different methods. Today, artificial intelligence and more specifically machine learning has become a necessity on today's solving data problems among a variety of different disciplines ranging from healthcare, finance and marketing. Therefore, among the existing suspected algorithms, its selection when determining a specific model is critical to achieving the best balance between accuracy and speed (Gupta et al., 2022). In the given context, the measure of achievement is determined by the testing of accuracy, precision, recall or F1 score, while efficiency factors pertain to computation time, processor utilization, and capacity to scale up a model (Ul Hassan et al., 2018). K-Nearest Neighbors (KNN) is a technique that operates in a non-parametric instance base learning technique for classifying data based on the closeness of data points to some pre-specified neighbors (Altman, 1992). As is easy to understand KNN has its merits and demerits Indeed, KNN is highly sensitive to the value of K used and computational intensity can sometimes be a problem as data grows large. On the other hand, Support Vector Machine(SVM), a good supervised learning algorithm, is aimed at identifying a hyperplane that separates two classes with the largest margin between them (Cortes & Vapnik, 1995). Originally developed for two-class problems and showing specific strength in high dimensional spaces, SVM has already been used in textual categorization as well as image recognition. But it has to do well by the choice of the kernel and may not be efficient when used with large data sets (Bennett & Campbell, 2000). An exploration of practical details of machine learning algorithms of establishing the possibility of increasing the intelligence and functionality of applications have also been provided by Sarker (2021) explaining the viability of applying them across practical domains. Decision Trees are semi-graphical structures that consist of a tree, recursively separating data based on values of the features (Quinlan, 1986). Although the model is fairly easy to read and understandable, Decision Trees are only suitable for working with discrete data, and are inclined to overtraining, which also may cause the model to be replaced by a random guess when working with high cardinality features and imbalanced datasets.

As with most of the current statistical techniques in data mining, Logistic Regression does not assume the homogeneity of variance but supposes a linear relationship between the input attributes and the log-odds of the output class (Cox, 1958). On the advantage, despite of its simplicity in computation, Logistic Regression may not work so well in situations where there are non-linear relationships, especially where the data is more complicated. Therefore, the purpose of this work to compare all four of them and determine their effectiveness and performance on a wide variety of datasets. We will use a wide range of measurement metrics, including accuracy, precision, recall, F1 score, and AUC-ROC as our assessment criteria to evaluate the results (Powers, 2020). In the aspect of performance, we will compare and contrast their computational time, flexibility of scaling up, and space complexity respectively. By doing so, this study seeks to present useful information on the costs and benefits of these models to help practitioners choose the right machine learning algorithms for certain problems.

II. REVIEW OF LITEARTURE

Several of machine learning algorithms like KNN, SVM, DT and LR are implemented a lot in various fields like health, computer science, finance and more. This section summarizes the literature related to these algorithms paying attention to their comparisons with respect to their efficiency and performance in various domains. In computer field, many researches have been done to analyze KNN, SVM, DT and LR in different applications. These algorithms were compared for text classification according to Shah et al. (2020). According to their study, they found that logistic regression yielded better accuracy and efficient computation than others for text data. As well, Pathak and Pathak (2020) synchronized KNN with Decision Trees in IDS and exposed that KNN was more precise in the classification process compared with Decision Trees that was more efficient in computational time. In another analysis of Charbuty and Abdulazeez (2021), Decision Trees had higher versatility than KNN in using huge data to make intricate decisions in learning techniques even though KNN was quite effective for particular classification issues at that time. Moreover, Kiranashree et al. (2021) pointed out that there are possibilities for employing machine learning technologies, such as SVM and Decision Trees, in identifying stress in employees and the author found that decision trees have a shorter time of classification while having almost the same level of accuracy as SVM. In the health care industry, we have used KNN, SVM, Decision Tree for prediction, especially for the diagnosis of diseases and the prognosis of patients. Bansal et al. (2022) made a comparison between KNN, SVM, and Decision Trees yielded about diabetes prediction and the study showed that SVM and Decision Trees offered high accuracy than KNN and especially large data sets. This is in contrast with Maniruzzaman et al. (2020) study, which showed that SVM yielded highly accurate and exploitable diabetes classification outcomes even when dealing with high dimensions of the dataset. Some of the comparisons made have also been informative regarding the educational domain. Alghurair and Mezher (2020) proposed an investigation of KNN, SVM, and Decision Trees for predicting the performance of students. In conclusion SVM was most accurate in terms of prediction of academic performance KNN was best when small number of cases were to be dealt with which makes it efficient with less complicated model. Furthermore, Decision Trees were mentioned as performing the function of offering interpretability of the decision rules that can be employed by the education decision makers. The experiments comparing the general performance of KNN, SVM, and Decision Trees have included domains of machine learning other than text classification. Mujumdar and Vaidehi (2019) explained that KNN is more efficient for small sets of the problem with low dimensions; while SVM and Decision Trees work more effectively for large sets, they are suitable to face non-linearity and high dimensionality. In the converged world, data mining has emerged as one of the driving forces for decision-making; Sheth et al. (2022) conducted a current research analysis of the well-known data mining techniques, including clustering, classification, and association. Classification was underlined as essential for data identification, access, risk management and compliance, as well as and security purposes. The authors pointed out that classification enhances data organizations and recovery because they enhance categorization. They pointed out that the performance of a model is defined by data rather than a model in machine learning techniques. They contrasted the four classifiers; Decision Tree, SVM, NB, and k-NN over five heterogeneous data sets. According to their results, they showed that Naive Bayes algorithm provided higher accuracy compared to other algorithms that have been tested.

III. RESEARCH GAP

However, as has been discussed there are still more important research gaps with regard to KNN, SVM, Decision Trees, and Logistic Regression. Studies usually target distinctive industries, for instance, healthcare or finance, while few attempt to compare their outcome across several domains, for example, IoT and cybersecurity. Further, there is limited attention paid to computational complexity specifically, time and space complexity in large datasets. Moreover, their integration, and more generally, more complex models based on these algorithms are not yet developed.

Most of the studies also fail to take into account dynamic and current data and are mainly concerned with only database static data. Finally, literature on algorithm interpretability is scarce and this is especially significant when it comes to the various industries such as finance and health. This study aims at filling these gaps.

IV. RESEARCH OBJECTIVES

The primary objective of this study is to conduct a comprehensive cross-domain comparison of four widely used machine learning algorithms: K-Nearest Neighbors, Support Vector Machine, Decision Tree and Logistical Regression. In this case, the researchers have focus on discussing different algorithms' performances and effectiveness in different sectors, i.e., 20newsgroup dataset and Wine dataset. Based on the different data types and characteristics within each domain of study, this work intends to give an overview of how effective or ineffective these algorithms are in real life applications. To realize this a number of the following subsidiary aims have been envisaged. First, the study will aim at comparing the results that KNN, SVM, DT, and LR generate with regards to accuracy and performance with different datasets. This concern involves rate analysis for accuracy, precision and recall, F1 score to arbitrate under which condition which algorithm is best. In sum, these objectives are to offer comparisons of algorithms' performances with KNN, SVM, DT, and LR for various functions in fields to identify which algorithms are most suitable for certain application.

V. RESEARCH METHODOLOGY

A. Data Collection

The datasets for this study have been collected from different domains to cover a wide range of domains and to compare KNN, SVM, Decision Trees and Logistic Regression Models in the best possible way. 20Newsgroup is another dataset containing newsgroup posts in 20 different topics which is used for text classification. While the Wine Quality dataset comprises information about red and white wines revealing their physicochemical characteristics and quality indicators, which are utilized for regression. They both present a set of unique and complex cases for testing the effectiveness of machine learning approaches, which allows the researchers to compare the outcomes in case of various methods and algorithms usage.

B. Data Preprocessing

After that, the datasets are elicited and they are ready for the training and assessment of the models after a series of preprocessing. The first and very important step in this case is dealing with the missing values either by imputation or by erasing the rows with more than one or more missing value. After missing value handling the process of feature scaling has been applied to normalize all numerical features, which is very important for distance-based algorithms such as KNN and SVM. Some of them have included Min-Max scaling or Standardization (Z-score normalization depending on the provided data) have been applied. Moreover, the categorical variables have undergone a process of one hot encoding or labelling in order to help feed models. Finally, each dataset has been split into training and testing sets, typically with an 80:20 split, which means that the performance of the models is calculated on the holdout samples to estimate generalization capability.

C. Model Training and Evaluation

For each of the selected ML models, namely KNN, SVM, Decision trees, and Logistic Regression, the findings of the training phase in the preprocessed training data have been tested on the testing data. In the training phase, the cross-validation technique was applied to optimize other hyperparameters of models as the number of neighbors in case of KNN, type and regularization of kernel in SVM and tree depth in Decision Trees. For Logistic Regression, researchers have used L1 and L2 to try and avoid overfitting of the model. After the models are trained, performance metrics as accuracy, precision, recall, F1-score for binary classification were used.

VI. RESULT ANALYSIS

The experiment compares the performance of four classifiers; K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Logistic Regression on various datasets thus shedding light on their appropriateness for text classification tasks. The results of performance between 20Newsgroups and Wine datasets shown the variation of model performance depending on the data and using task. Specifically, for the 20Newsgroups dataset which is all about text classification, the authors observed that SVM and Logistic Regression had the highest accuracy and highest score of precision, recall and F1. These models are particularly effective when dealing with the text data because of their ability to work with high dimensional sparse feature spaces.

The same results are derived in term of F1 score comparison in which the Logistic Regression and SVM have the dominant position and have higher values in each metric considered. Compared to SVM and Logistic Regression, Decision Tree has lower training accuracies and less effective for the text classification task Again, KNN gives the lowest F1 scores. However, it is observed that accuracy and recall are significantly low for KNN in case of sales data which is this type of text base data set.

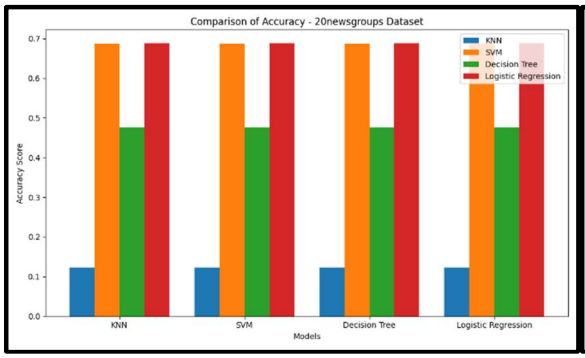


Figure 1. Accuracy for 20Newsgroup Dataset

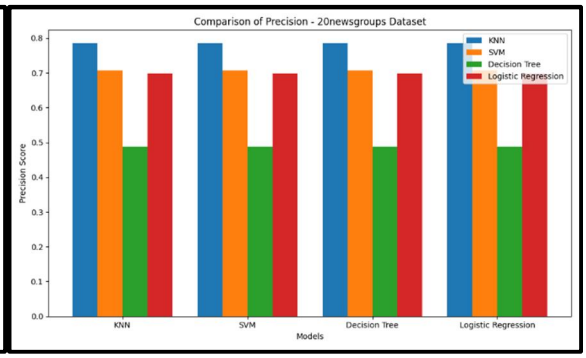


Figure 2. Precision Comparison for 20Newsgroup Dataset

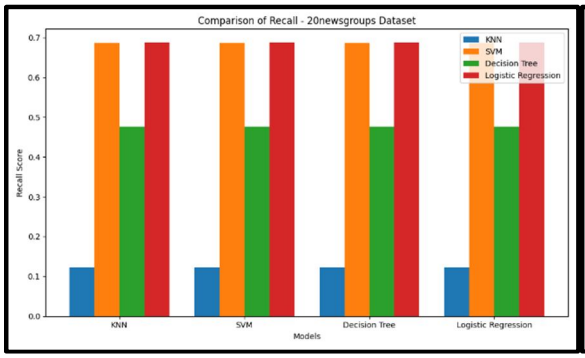


Figure 3. Recall Comparison for 20Newsgroup Dataset

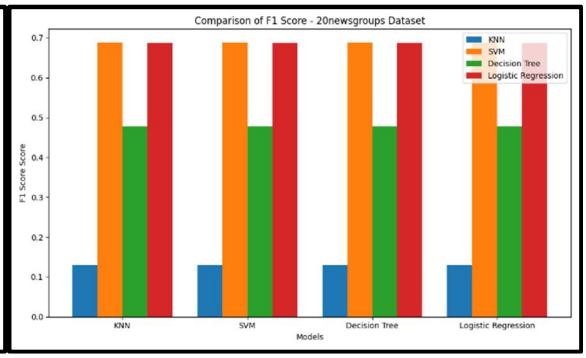


Figure 4. F1 Comparison for 20Newsgroup Dataset

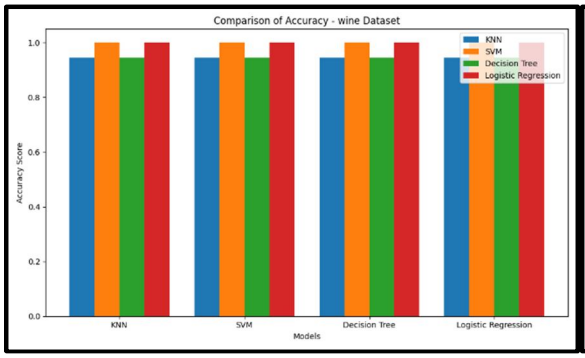


Figure 5. Accuracy comparison for Wine Dataset

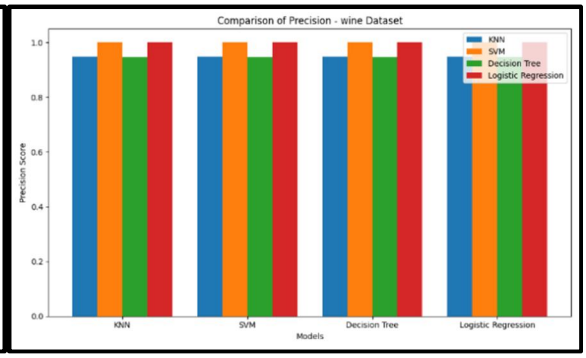


Figure 6. Precision Comparison for Wine Dataset

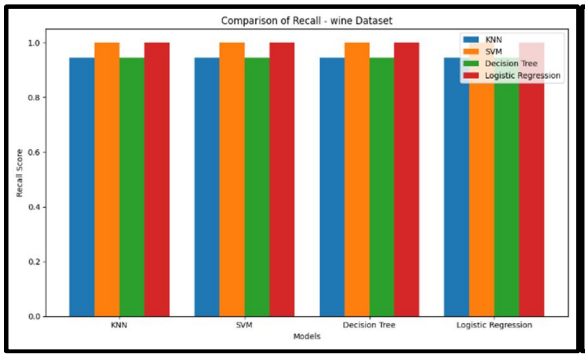


Figure 7. Recall comparison for Wine Dataset

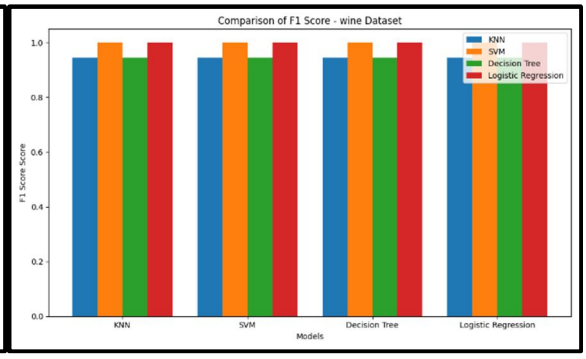


Figure 8. F1 Comparison for Wine Dataset

The Decision Tree model does reasonably well, yet, it cannot bring the desired level of accuracy and recall necessary for performing better at text classification tasks as in case with SVM and Log Reg where such difference is not characteristic. On the other hand, the Wine dataset, which is a classification problem, less complex compared to the previous dataset with no missing values and fairly normalized features, shows much lesser gaps in the accuracy, precision, recall and F-measure scores among the tested models. On a comparative analysis of accuracy, precision and recall, all the four models, KNN, SVM, Decision Tree and Logistic Regression are observed to be of fairly equal performance. This implies that for simple and clean data sets such as the Wine data, the differences in model choice are therefore minimal, since nearly all classification algorithms can achieve the best results. These near perfect scores represent results only possible when working with a simple dataset where the relations between features and classes are easily inferred. The systems formulated here show that SVM delivers demonstrably higher classification accuracy in the wine samples for the features provided as compared to KNN, Decision Tree and Logistic Regression. Surprisingly, KNN and Decision Tree are similar while Logistic Regression is slightly worse; this is elaborated by the thought that a linear decision boundary may not be the most optimal for this source separation task. SVM might have outperformed KNN because as seen earlier it can capture non-linear relationships between features. More detailed insights could be derived by analyzing precision and recall distinctively, or by making use of a confusion matrix.

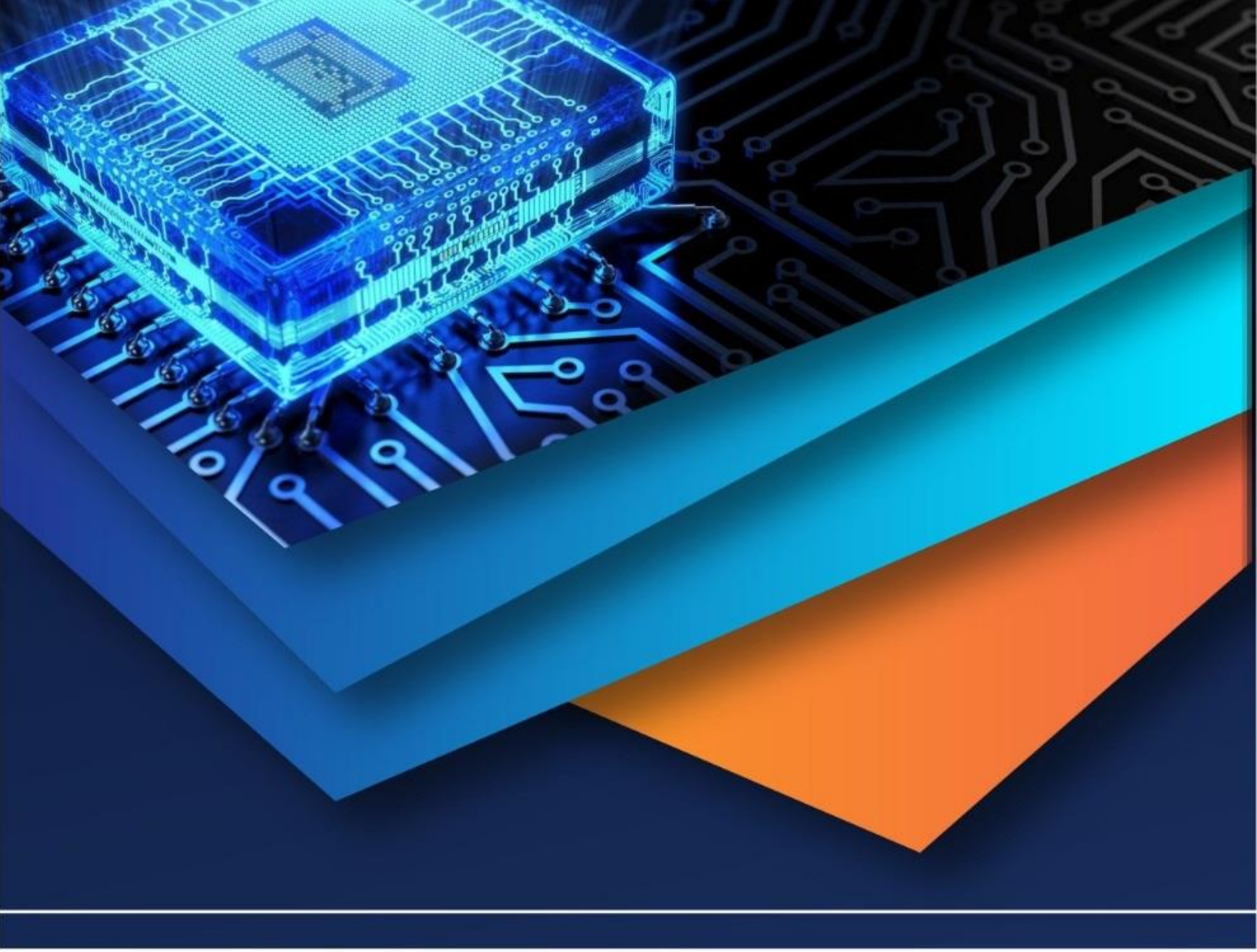
VII. CONCLUSION

In analyzing the results of the different models tested on the 20Newsgroups and Wine datasets, it is shown that the strategies employed do differ depending on the complexity and character of the two distinct datasets. In the case of the 20Newsgroups dataset, which is a more comprehensive text classification problem, the proposed models SVM and Logistic Regression achieve better accuracy, F1-score, precision, and recall among all the models in both selected metrics. It is also used in high dimensional and sparse text data which the basic models are established to handle. KNN appears to be the lowest scoring, and both recall and accuracy are low while the Decision Tree model though less impressive than the two most efficient models, but is still way ahead of the KNN model. On the other hand, Wine dataset, a much simpler classification problem with less pre-processing required for data, shows negligible variance among all models. All named algorithms, including KNN, SVM, Decision Tree, and Logistic Regression, yield near-perfect measurements of accuracy, precision, and recall, therefore, the easy classifier of the chosen dataset. From this analysis the argument can be made that for tasks such as text classification, more refined models like SVM, and Logistic Regression do best while for simpler and more clearly defined datasets all the models tested here are equally good. As a result, the choice of a model must be made depending on the nature of the data and the problem, which we're going to solve.

REFERENCES

[1] Alghurair, N. I., & Mezher, M. A. (2020). Generic Frameworks for SVM, ANN, LGBM and LR Algorithms. International Journal of Computer Science and Mobile Computing, 9(6), 132–140.
[2] Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician, 46(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879
[3] Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. Decision Analytics Journal, 3, 100071. https://doi.org/10.1016/j.dajour.2022.100071
[4] Bennett, K. P., & Campbell, C. (2000). Support vector machines: Hype or hallelujah? SIGKDD Explor. Newsl., 2(2), 1–13. https://doi.org/10.1145/380995.380999
[5] Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), Article 01. https://doi.org/10.38094/jastt20165
[6] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018
[7] Cox, D. R. (1958). The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x
[8] Gupta, S., Saluja, K., Goyal, A., Vajpayee, A., & Tiwari, V. (2022). Comparing the performance of machine learning algorithms using estimated accuracy. Measurement: Sensors, 24, 100432. https://doi.org/10.1016/j.measen.2022.100432
[9] Kiranashree, B. K., Ambika, V., & Radhika, A. D. (2021). Analysis on Machine Learning Techniques for Stress Detection among Employees. Asian Journal of Computer Science and Technology, 10(1), Article 1. https://doi.org/10.51983/ajcst-2021.10.1.2698
[10] Maniruzzaman, Md., Rahman, Md. J., Ahammed, B., & Abedin, Md. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8(1), 7. https://doi.org/10.1007/s13755-019-0095-z
[11] Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science, 165, 292–299. https://doi.org/10.1016/j.procs.2020.01.047
[12] Pathak, A., & Pathak, S. (2020). Study on Decision Tree and KNN Algorithm for Intrusion Detection System. International Journal of Engineering Research & Technology, 9(5). https://doi.org/10.17577/IJERTV9IS050303
[13] Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation (arXiv:2010.16061). arXiv. https://doi.org/10.48550/arXiv.2010.16061

[14] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106. https://doi.org/10.1007/BF00116251
[15] Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x
[16] Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. Augmented Human Research, 5(1), 12. https://doi.org/10.1007/s41133-020-00032-0
[17] Sheth, V., Tripathi, U., & Sharma, A. (2022). A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. Procedia Computer Science, 215, 422–431. https://doi.org/10.1016/j.procs.2022.12.044
[18] Ul Hassan, C. A., Khan, M. S., & Shah, M. A. (2018). Comparison of Machine Learning Algorithms in Data classification. 2018 24th International Conference on Automation and Computing (ICAC), 1–6. https://doi.org/10.23919/IConAC.2018.8748995



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY