BALTIC JOURNAL OF LAW & POLITICS
A Journal of Vytautas Magnus University
VOLUME 15, NUMBER 4 (2022)
ISSN 2029-0454

Cite: *Baltic Journal of Law & Politics* 15:4 (2022): 378-387
DOI: 10.2478/bjlp-2022-004040

Higher Accuracy of Spam Mail Prediction using Decision Tree Algorithm Comparing with K-Nearest Neighbor Algorithm

Putta Charan
Research scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, Pincode: 602105.

P.Sriramya
Project Guide, Corresponding Author, Department of Data Science, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, Pincode: 602105.

Received: August 8, 2022; reviews: 2; accepted: November 29, 2022.

Abstract

Aim: To make an Innovative Spam Prediction of spam emails using Machine learning modeling techniques and to evaluate their performance. **Materials and Methods:** The endeavor will primarily collect samples from two groups. The Decision Tree Algorithm belongs to Group-1, while the K-Nearest Neighbor Algorithm belongs to Group-2. For both Algorithms, the same sample sizes were used. The G-Power in the test set will be at 80%. **Result:** Data is trained in the given model so that Machine learning can function effectively. Emails are used as inputs after the K-Nearest Neighbor Algorithm is applied, which gives us a probabilistic index and determines whether the email is spam or not. The Decision Tree Algorithm outperforms the K-Nearest Neighbor Algorithm, and our hypothesis is insignificant with a significance value of 0.268. **Conclusion:** These results were achieved through machine learning models such as K-Nearest Neighbor Algorithm, and Decision Tree Algorithms. In this paper, we have demonstrated that for the spam filtering method the most efficient algorithms are the Decision Tree Algorithm and K-Nearest Neighbor Algorithm given as they have the highest level of accuracy.

Keywords: Classifier, Filtering, Innovative Spam Prediction**,** Decision Tree Algorithm, Machine learning, K-Nearest Neighbor Algorithm.

INTRODUCTION

The purpose of Spam Email classification is to automatically classify new emails as spam or ham based on their contents. There has been a significant growth in the number of emails received, necessitating effective approaches such as Text Mining and Natural Language Processing to automatically categorize emails as spam or ham nearly 4.1 billion Email accounts are created throughout the world and More than 196 billion Emails will be sent day by day. Spam-Emails are one of the main threats to Email Users (Kontsewaya, Antonov, and Artamonov 2021). In this paper, we compare the performance of two machine learning techniques for spam detection including the Decision Tree Algorithm classifier Compared with the K-Nearest Neighbor Algorithm classifier. K-Nearest Neighbor Algorithm Classifier takes more time during the training period but its classification speed

is better than other classifiers. An unwanted Email sent in bulk to an unknown recipient is referred to as a spam Email (Akinyelu 2021). It refers to the use of an email system to send unsolicited emails, particularly marketing emails to a large number of people. These accounts perform all email traffic worldwide. Unsolicited emails indicate that the receiver has not been permitted to receive them. Spam emails have grown in popularity over the last decade and are a problem that most email users confront for filtering methods. The applications of the research are Users and emails (Hossain, Uddin, and Halder 2021), (Kumar, Sonowal, and Nishant 2020). Botnets or networks of infected computers may send massive amounts of spam emails.

Innovative Spam Prediction using Decision Tree Algorithm comparing with K-Nearest Neighbor Algorithm. In GoogleScholar this article is published 4460 times, and in ScienceDirect, this article is published 111 times in the past 5 years. In these 2 databases, the most cited articles and their findings are, that suggested Machine intelligence-based algorithms A Survey of Existing E-mail Spam Filtering Methods Considering Machine Learning Techniques(Bhuiyan et al. 2018). That the preliminary discussion in the research background looks at how machine learning methods are used in the email spam filtering processes of the top internet service providers (Dada et al. 2019). A Quick Review of Machine Learning Algorithms is predominantly an area of Artificial Intelligence that has been a key component of digitalization solutions that has caught major attention in the digital arena (Ray 2019)and (Baaqeel and Zagrouba 2020)The suggested system is based on the Genetic Algorithm and the Random Weight Network (Faris et al. 2019). In my studies, this is the best (Kontsewaya, Antonov, and Artamonov 2021).

Previously our team has a rich experience in working on various research projects across multiple disciplines (Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020). On Daily basis, Spam Email is continuously increasing day by day. The rapidly increasing Spam Emails are responsible for over 77% of the whole global email traffic, these motivated me to do the Research on Spam mail Prediction. The team in the department has much experience in research on Machine learning models, so it's helpful to come up with innovative ideas in machine learning approaches for developing efficient algorithms with higher accuracy in the spam email prediction and this shows experience in our lab for research of spam email prediction. The aim is to increase the accuracy value of the email spam prediction using Machine Learning techniques and predict if the email is spam or not and make an Innovative Spam Prediction of spam emails using Machine learning modeling techniques and evaluate their performance (Rafat et al. 2022).

MATERIALS AND METHODS

This Research paper for Spam Email Prediction research is done in the Software Engineering Lab, Saveetha School of Engineering, SIMATS. The Dataset has been taken from Kaggle and this has an open-source license to download and use the data for the research. In this project, there will be mainly two groups of samples taken in the project. That Group-1 belongs to the Decision Tree Algorithm and Group-2 Belongs to the K-Nearest Neighbor Algorithm. The sample sizes of both groups are 30% and 70% total sample sizes taken are 100% of the data. The Same set of Sample sizes will have for both algorithms. Iteration-1 for the Train set and Iteration-2 for the Test set will have 80% of the G-Power (Wood and Krasowski 2020). This helps to create a more Accurate Prediction for the Spam Mail using Machine Learning models.

Data Collection

The Data Set for this Research is collected from Kaggle which is an Open source Platform for getting Machine Learning Datasets. The Url for the datasets is mentioned below as (ishansoni 2018). I got 10743 rows and 2 columns By combining the two datasets used in

the Algorithms. In the Datasets, different dependent and independent Variables are Considered to Perform Machine Learning Techniques.

Decision Tree Algorithm

Decision Tree Algorithm is one of the most widely used supervised learning algorithms. It can be used for both classification and regression problems but the decision tree algorithm is usually used to solve the classification difficulties. The test results of the decision tree are performed based on the features found in the given data set. The decision tree always starts with the root node and ends with the decisions made by leaves. The output of the decision tree always executes either yes or no. The decision tree algorithm always consists of Root Nodes, Decision Nodes, and Terminal Nodes. Decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules, and the fitter the model. It is known as a choice tree in light of the fact that, like a tree, it begins with the root hub, which develops further branches and builds a tree-like design.

Decision Tree Algorithm Equation

Entropy

$$E(S) = \sum_{i=1}^{c} \quad - P_i log_2 p_i \text{—----------(1)}$$

The impurity or randomness of a dataset is measured using entropy. The entropy value is always between 0 and 1. Its value is better when it equals 0, and it is worse when it equals 0, i.e. the closer it is to 0, the better. From the above Equation (1) we say that

S is the current state

Pi is the probability of an event I of state S or Percentage of class I in a node of State S.

$$Gain(S, A) = \sum \quad v \in V(A) \frac{|S_v|}{|S|} Entropy(S_v) \text{—----------(2)}$$

From equation (2) that Where the range of attribute A is (A), and Sv is a subset of set S equal to the attribute value of the attribute.

Pseudocode for Decision Tree Algorithm

Input: Training dataset
Output: Classifier predicted Accuracy
An attribute-valued dataset DT
Tree1={}
If DT is "pure" OR other stopping criteria had met then
    terminate
endif
for all attributes, b DT do
    Compute information-theoretic criteria if we split on b
end for
bbest = Best attribute according to the above-computed criteria
Tree = Create a decision node that tests bbest the root.
DTv= Induced sub-datasets from DT based on bbest
for all Dv do
    Treev= C4.5(DTv)
    Attach Treev to the corresponding branch of Tree
end for
return Tree

K-Nearest Neighbor Algorithm

The K-Nearest Neighbor method is one of the most fundamental Machine Learning algorithms and is based on the Supervised Learning methodology. The K-Nearest Neighbor approach assumes that new and existing data are comparable, and it assigns the new example to the category that is most similar to the existing categories. The K-Nearest Neighbor approach preserves all previously saved data and categorizes new data points

based on their similarity. This means that when new data is available for filtering, the K-Nearest Neighbor approach can swiftly categorize it into an appropriate category. The K-Nearest Neighbor approach may be used for both regression and classification, while classification is more typically utilized. It is straightforward to build because all that is required is a probability calculation. We will forecast the Accuracy value of the K-Nearest Neighbor using the data. This is an iterative procedure that must be done for each data point in the dataset. Assume we already have a cleaned dataset that has been separated into training and testing data sets. Use the distance metrics of (3) Euclidean distance to calculate the distance between test data and each row of training data.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots\ldots + (p_i - q_i)^2 + \ldots (p_n - q_n)^2} \quad \text{-------- (3)}$$

Pseudocode for K-Nearest Neighbor

Input: Training dataset
Output: Classifier accuracy
Read dataset as input
select 'k' samples from the total samples
classifier. fit(x_train,y_train)
Among 'k' tokens calculate the node 'd' using the best-split
Repeat 1 to 3 steps until many samples are reached
Build the K-classifier
Predict value using predict feature
prediction=model.predict(parameters, "")
Calculate vote for each predicted value
Get predicted accuracy
Get Test Results.

The platform used to evaluate the Machine learning Algorithm was Anaconda/Jupyter. The hardware used to perform the work is Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz with a RAM size of 8 GB. The system type used was 64 bit, Windows OS, X64-based processor with an SDD of 256 GB. The Operating System used was Windows 10, and the tool used was JupyterLabs with the Python programming language. The testing procedure was to split the data into train and test data and then implement the Machine learning classifier to build and train a model on our data. After training, the predictions are made and the performance of the model is evaluated using the available metrics.

The dataset for Innovative spam prediction is collected from Kaggle. Data preprocessing was performed to gain some context about the data using Statistical Analysis techniques. Data cleaning methods such as removing unnecessary attributes, and contents and filling null values are done. The comparison of the Decision Tree Algorithm and K-Nearest Neighbor Algorithm with data exploration gives us some context and valuable insight into the dataset. The Spam Email Prediction with two widely spread classification algorithms in machine learning was selected Decision Tree Algorithm and K-Nearest Neighbor Algorithm. The algorithms will be trained with some data when the test data is given then it will predict the output whether the given email is spam or not. The testing data is used to give the predicted output and analyzes the data according to that.

Statistical Analysis

The IBM Spss is the Statistical Software Tool that is used for Spam Email data analysis. The IBM Statistical Tool can analyze the data and helps to create Graphs and Charts to display it quite easily. Before sending results into the Spss tool the Data sets are standardized and then the data is converted into arrays. The IBM tool can easily handle large data because it consists of a wide array of characteristics. The number of clusters required is pictured and analyzed and therefore the existing algorithms are obtained. It gives the Mean value for the Group statistics (Gaurav et al. 2019). The Group-1 and Group-

2 Accuracy as shown in Table 1 the different Test sizes and their average accuracy values that are acquired after being tested with the Decision Tree Algorithm Classifier and K-Nearest Neighbor Algorithm Classifier with 10 Sample test sizes. The Data Sets for the Spam Email Prediction are taken from the kaggle which consists of Both Dependent Variables and In-Dependent Variables in Table-2 and Table-3. The Statistical Comparison of The Spam Email Prediction using two Sample groups was done with the SPSS Version 25. The Analysis was done using the Mean, Median, Independent T-Test, and Deviation. For each sample size of data, the Accuracy is deviating between 3% to 5 %. So that we finally sent all the Test sizes and also their Accuracy into the Spss tool and found the Average Accuracy values of the Decision Tree Algorithm Classifier and the K-Nearest Neighbor Algorithm Classifier.

RESULT

In the proposed model, data is trained so that Machine learning can work properly. After applying the K-Nearest Neighbor Algorithm, emails are taken as inputs which will give us the probabilistic index of that and will identify whether the Email is spam or not. This necessitates the development of a sensible method for detecting or identifying such spam emails, therefore saving a significant amount of time and memory space for the system. Spammers may easily create a false profile and email account by pretending to be a legitimate person in their spam emails. This paper will discuss machine learning algorithms and apply all of these algorithms to our data sets, and the best algorithm is selected for email spam detection with the highest precision and accuracy.

The Innovative Spam Prediction using the Decision Tree Algorithm gave us an accuracy of 94.73% and the K-Nearest Neighbor Algorithm gave us an accuracy of 91.03% compared with their accuracy rate. Each algorithm was repeated 10 times, for each algorithm and the accuracy varies for different test sizes in decimals. The accuracy varies due to random changes in the test sizes of the algorithm as given in Table 1.

The observed values for the metrics of Group Statistics, the mean accuracy, and the standard deviation for the Decision Tree Algorithm are 93.31000 and 1.01763. The K-Nearest Neighbor Algorithm's mean accuracy is 90.1870 and the standard deviation is 0.79363. The Decision Tree Algorithm also obtained a standard error mean rate of 0.32180 whereas the K-Nearest Neighbor Algorithm obtained an error mean rate of 0.25097 as shown in table 2.

Then an independent sample test of 10 samples was performed, Decision Tree Algorithm obtained a mean difference of 3.123 and a standard error difference of 0.40810. When compared to other algorithm performance, the Decision Tree Algorithm performs better than the K-Nearest Neighbor Algorithm and the significance value of 0.268 shows that our hypothesis is insignificance as given in Table 3.

It is called the Innovative Spam Prediction architecture. The architecture defines the steps which are performed to develop a spam email prediction. It consists of the steps as Data Pre-processing, Database, Data Extraction, Modelling Classifier, Implementation, and Predicted Accuracy.

The GGraph represents a bar chart of the simple bar mean accuracy, with the Decision Tree Algorithm achieving an accuracy of approximately 95%, and the K-Nearest Neighbor Algorithm achieving 91%. The 95% error bars represent the variation in the corresponding coordinates of the point. Independent t-tests were performed to compare the accuracy of the two algorithms and a statistically significant difference was noticed between the two algorithms 0.268. When comparing the two algorithms the performance of the Decision Tree Algorithm achieved a better performance than K-Nearest Neighbor Algorithm as given in Fig. 1.

DISCUSSION

The Decision Tree Algorithm has better accuracy than the K-Nearest Neighbor Algorithm. The results are collected by performing multiple times for identifying different scales of accuracy rates. Independent samples t-tests are performed on the dataset. In this study of spam email prediction, the Decision Tree Algorithm has an accuracy of approximately 95%, which is higher than that of the K-Nearest Neighbor Algorithm which is 91%. The Decision Tree Algorithm has insignificance of 0.268 while using the independent samples T-test. The mean accuracy and standard deviation for the Decision Tree Algorithm are 93.3100 and 1.01763 using a missing value imputation and a machine learning model to get an accuracy of 94%. The K-Nearest Neighbor Algorithm's mean accuracy is 90.2990 and the standard deviation is 0.47529. In the paper, (Kontsewaya, Antonov, and Artamonov 2021) the Decision Tree Algorithm obtained an accuracy of 94%, and the K-Nearest Neighbor Algorithm achieved an accuracy of 90% accuracy. Based on the literature survey, it is evident that the Decision Tree Algorithm performs better than K-Nearest Neighbor Algorithm. By running independent sample tests in IBM's SPSS statistical program, it can be seen that the difference between the two algorithms is statistically insignificant at 0.268. The SPSS statistical program is also used to compute the mean and standard deviation.

Using IBM's SPSS statistical tool, independent sample analysis confirmed that the difference between the two methods is statistically insignificant at 0.268>0.05. The mean and standard deviation are determined using the SPSS statistical tool. The decision Tree Algorithm outscored other algorithm classification accuracy by 95% percentage in the paper (Kontsewaya, Antonov, and Artamonov 2021).

The main limitation is that the attributes in the dataset contain fewer data to predict accuracy (%) for spam email classification. The more the independent and dependent variables the more accuracy will be improved. For future work, the dataset contains many attributes the classifier can work efficiently and can improve the prediction accuracy. Attributes like this can result in improved accuracy and exact precision values (Liu, Lu, and Nayak 2021). There exists a strong relationship between the content and the subject of the emails. With the help of this relationship, one can easily classify the documents. Positive value tells us how strongly that word belongs to the subject and negative tells how much it differs from a subject(Maguluri et al. 2019). With the help of a negative score also the accuracy of the classifier has been improved and this paper is to improve the relationship between the subject and content of the email by identifying the most relevant words using evolutionary computation of Email.

CONCLUSION

These results were achieved through machine learning models such as K-Nearest Neighbor Algorithm, and Decision Tree Algorithms. In this paper, we have demonstrated that for the spam filtering method the most efficient algorithms are the Decision Tree Algorithm and the K-Nearest Neighbor Algorithm, given as they have the highest level of accuracy. These spammers target those who are unaware of these scams and have filtering issues. So, it is necessary to identify those spam emails that are fraudulent, this project will identify those spam by using machine learning techniques. The results can be used to create a more intelligent spam detection classifier by combining algorithms of filtering methods.

DECLARATIONS

Conflict of Interests

No conflict of interest in this manuscript.

Author Contribution
Author PCR was involved in data collection, data analysis, and manuscript writing. Author PSR was involved in conceptualization, data validation, and critical review of the manuscript.

Acknowledgement
The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding
We thank the following organizations for providing financial support that enabled us to complete the study.

1. Vee Eee Technologies Solutions Pvt. Ltd.
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES

Akinyelu, Andronicus A. 2021. "Advances in Spam Detection for Email Spam, Web Spam, Social Network Spam, and Review Spam: ML-Based and Nature-Inspired-Based Techniques." *International Journal of Information and Computer Security* 29 (5): 473–529.

Baaqeel, Hind, and Rachid Zagrouba. 2020. "Hybrid SMS Spam Filtering System Using Machine Learning Techniques." http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9300071.

Benin, S. R., S. Kannan, Renjin J. Bright, and A. Jacob Moses. 2020. "A Review on Mechanical Characterization of Polymer Matrix Composites & Its Effects Reinforced with Various Natural Fibres." *Materials Today: Proceedings* 33 (January): 798–805.

Bhuiyan, Hanif, Akm Ashiquzzaman, Tamanna Islam Juthi, Suzit Biswas, and Jinat Ara. 2018. "A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques." *Global Journal of Computer Science and Technology*, May. https://computerresearch.org/index.php/computer/article/view/1705.

Dada, E., J. Bassi, H. Chiroma, S. Abdulhamid, A. O. Adetunmbi, and Opeyemi Emmanuel Ajibuwa. 2019. "Machine Learning for Email Spam Filtering: Review, Approaches and Open Research Problems." https://www.ncbi.nlm.nih.gov/pubmed/31211254.

Faris, Hossam, Ala' M. Al-Zoubi, Ali Asghar Heidari, Ibrahim Aljarah, Majdi M. Mafarja, Mohammad A. Hassonah, and H. Fujita. 2019. "An Intelligent System for Spam Detection and Identification of the Most Relevant Features Based on Evolutionary Random Weight Networks." https://doi.org/10.1016/J.INFFUS.2018.08.002.

Gaurav, Devottam, Sanju Mishra Tiwari, Ayush Goyal, Niketa Gandhi, and Ajith Abraham. 2019. "Machine Intelligence-Based Algorithms for Spam Filtering on Document Labeling." *Soft Computing* 24 (13): 9625–38.

Gudipaneni, Ravi Kumar, Mohammad Khursheed Alam, Santosh R. Patil, and Mohmed Isaqali Karobari. 2020. "Measurement of the Maximum Occlusal Bite Force and Its Relation to the Caries Spectrum of First Permanent Molars in Early Permanent Dentition." *The Journal of Clinical Pediatric Dentistry* 44 (6): 423–28.

Hossain, Fahima, M. N. Uddin, and Rajib Kumar Halder. 2021. "Analysis of Optimized Machine Learning and Deep Learning Techniques for Spam Detection." http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9422508.

ishansoni. 2018. "SMS Spam Collection Dataset." Kaggle. October 6, 2018. https://kaggle.com/ishansoni/sms-spam-collection-dataset.

Kontsewaya, Yuliya, Evgeniy Antonov, and Alexey Artamonov. 2021. "Evaluating the Effectiveness of Machine Learning Methods for Spam Detection." 6074. EasyChair. https://easychair.org/publications/preprint_open/Th28.

Kumar, N., Sanket Sonowal, and Nishant. 2020. "Email Spam Detection Using Machine Learning Algorithms." http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9183098.

Liu, Xiaoxu, Haoye Lu, and Amiya Nayak. 2021. "A Spam Transformer Model for SMS Spam Detection." *IEEE Access*. https://doi.org/10.1109/access.2021.3081479.

Maguluri, Lakshmana Phaneendra, R. Ragupathy, Sita Rama Krishna Buddi, Vamshi Ponugoti, and Tharun Sai Kalimil. 2019. "Adaptive Prediction of Spam Emails : Using Bayesian Inference." *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. https://doi.org/10.1109/iccmc.2019.8819744.

Nalini, Devarajan, Jayaraman Selvaraj, and Ganesan Senthil Kumar. 2020. "Herbal Nutraceuticals: Safe and Potent Therapeutics to Battle Tumor Hypoxia." *Journal of Cancer Research and Clinical Oncology* 146 (1): 1–18.

Rafat, Khan Farhan, Qin Xin, Abdul Rehman Javed, Zunera Jalil, and Rana Zeeshan Ahmad. 2022. "Evading Obscure Communication from Spam Emails." *Mathematical Biosciences and Engineering: MBE* 19 (2): 1926–43.

Ray, Susmita. 2019. "A Quick Review of Machine Learning Algorithms." http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8862451.

Reddy, Poornima, Jogikalmat Krithikadatta, Valarmathi Srinivasan, Sandhya Raghu, and Natanasabapathy Velumurugan. 2020. "Dental Caries Profile and Associated Risk Factors Among Adolescent School Children in an Urban South-Indian City." *Oral Health & Preventive Dentistry* 18 (1): 379–86.

Sathish, T., and S. Karthick. 2020. "Gravity Die Casting Based Analysis of Aluminum Alloy with AC4B Nano-Composite." *Materials Today: Proceedings* 33 (January): 2555–58.

Sathish, T., D. Bala Subramanian, R. Saravanan, and V. Dhinakaran. 2020. "Experimental Investigation of Temperature Variation on Flat Plate Collector by Using Silicon Carbide as a Nanofluid." In *PROCEEDINGS OF INTERNATIONAL CONFERENCE ON RECENT TRENDS IN MECHANICAL AND MATERIALS ENGINEERING: ICRTMME 2019*. AIP Publishing. https://doi.org/10.1063/5.0024965.

Sivasamy, Ramesh, Potu Venugopal, and Rodrigo Espinoza-González. 2020. "Structure, Electronic Structure, Optical and Magnetic Studies of Double Perovskite Gd2MnFeO6 Nanoparticles: First Principle and Experimental Studies." *Materials Today Communications* 25 (December): 101603.

Venu, Harish, and Prabhu Appavu. 2021. "Experimental Studies on the Influence of Zirconium Nanoparticle on Biodiesel–diesel Fuel Blend in CI Engine." *International Journal of Ambient Energy* 42 (14): 1588–94.

Wood, Kelly E., and Matthew D. Krasowski. 2020. "Academic E-Mail Overload and the Burden of 'Academic Spam.'" *Academic Pathology* 7 (January): 2374289519898858.

TABLES AND FIGURES

Table 1. Accuracy Values for the Algorithm. The Data Accuracy for the Decision Tree Algorithm (Group-1) and K-Nearest Neighbor Algorithm(Group-2) with different Test sizes have been taken. In these different Test Sizes, the best Accuracy value for the Decision Tree Algorithm is 94.73 and the K-Nearest Neighbor Algorithm is 91.03.

S No	Test Size	Group-1 Accuracy	Group-2 Accuracy
1	0.2	94.28	90.55
2	025	94.3	90.72
3	0.3	94.73	90.53
4	0.35	93.94	89.84
5	0.4	93.62	90.85

6	0.45	93.24	91.03
7	0.5	92.59	90.65
8	0.55	92.06	89.6
9	0.6	92.49	89.65
10	0.7	91.85	88.45

Table 2. Group Statistics the mean accuracy and standard deviation for the Decision Tree Algorithm are 93.310 and 1.01763 and For K-Nearest Neighbor Algorithm is 90.1870 and .79363.

Group Statistics					
	DT, KNN	N	Mean	Std.Deviation	Std.Error Mean
Accuracy	DT	10	93.3100	1.01763	.32180
	KNN	10	90.1870	.79363	.25097

Table 3. Independent Samples Test. Independent t-tests were performed to compare the accuracy of the two algorithms and a statistically significant difference was noticed between the two algorithms 0.268>0.05 and Std. Error Difference is noticed as .40810.

		Levene's Test for Equality of Variances		**t-test for Equality of Means**		
					95% Confidence Interval of the Difference	
		F	**Sig.**	**Std. Error Difference**	**Lower**	**Upper**
Accuracy	Equal variances assumed	1.306	.268	.40810	2.26562	3.98038
	Equal variances not assumed			.40810	2.26196	3.98404

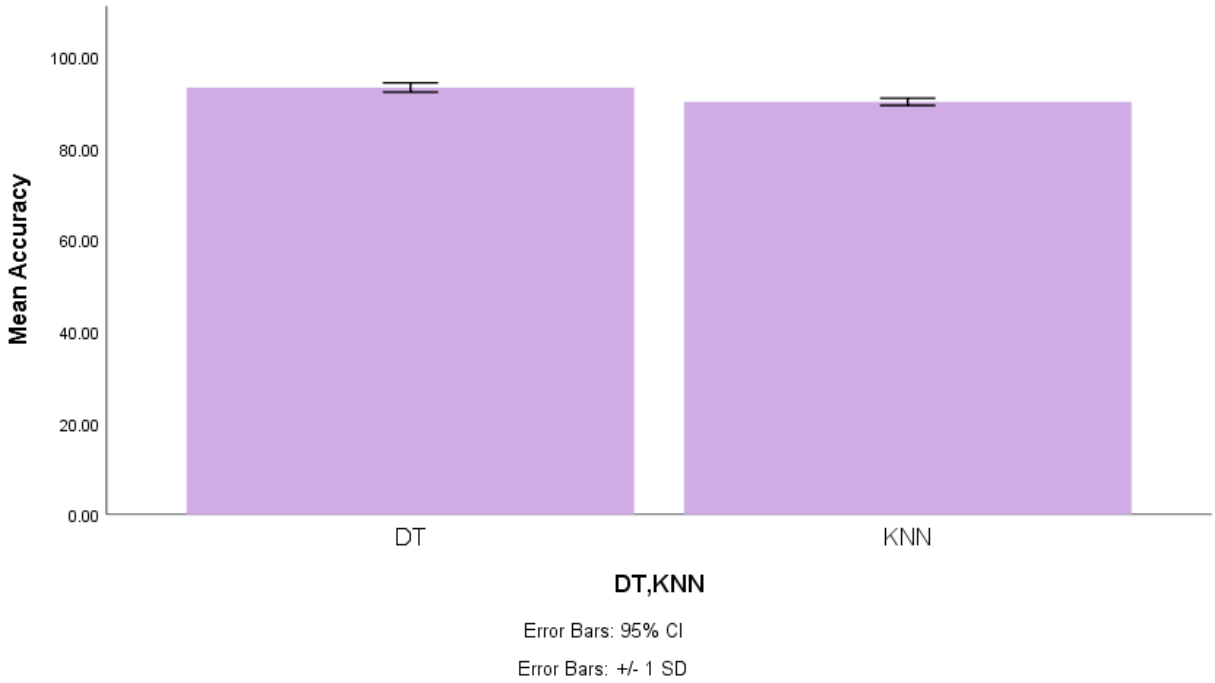

Fig. 1. Simple Bar Mean of Accuracy by Decision Tree Algorithm and K-Nearest Neighbor Algorithm(MNB), the bar chart representing the comparison of mean accuracy of Decision Tree Algorithm is 93.3100 and K-Nearest Neighbor Algorithm is 90.1870. X-Axis: Decision Tree Algorithm vs K-Nearest Neighbor Algorithm. Y-Axis: Mean accuracy. The error bars are 95% for both algorithms. The Standard Deviation Error Bars are +/- 1 SD.