Original Article

Naive Bayes Classifier Algorithm for Spam Detection of Email to Improve Accuracy and in Comparison with Decision Tree Algorithm

K.Varun Kumar[1], M. Ramamoorthy[2*]

[1]Research Scholar,Department of Computer Science and Engineering,Saveetha School of Engineering,Saveetha Institute of Medical And Technical Sciences,Saveetha University, Chennai, TamilNadu, India. Pincode: 602105

[2]Project Guide, Department of Artificial and Machine Learning,,Saveetha School of Engineering,Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, TamilNadu, India. Pincode: 602105

Abstract

Aim:The aim of the research is to detect spam in email using the Novel Naive Bayes Classifier (NB) and the Decision Tree algorithm (DT).

Material and Methods: We'll need two groups of 40 samples each to classify spam. The Decision Tree technique (DT) includes a sample size of 20, whereas the Novel Naive Bayes Classifier (NB) includes a sample size of 20 and G-power (value = 0.8).

Results: The accuracy of the Novel Naive Bayes Classifier is 98.05 %, which is higher than the Decision Tree algorithm with 91.80 %. All of us identified that the 2-tailed significant value of accuracy is 0.022 ($p<0.05$) in the Independent Sample T-Test analysis.

Conclusion: The Novel Naive Bayes Classifier has higher accuracy than the Decision Tree algorithm.

Keywords: Machine Learning, Supervised Learning, Spam Detection, Spam Filtering, Novel Naive Bayes Classifier, Decision Tree Algorithm.

DOI: 10.47750/pnr.2022.13.S04.006

INTRODUCTION

Email helps us communicate with other users and enterprises more quickly, and vice versa. Most job-specific details were only shared via email in multinational corporations. So the spam or worthless emails divert people's focus away from their conversation. Spam or worthless emails may include commercial ads or uninvited content, as well as potentially harmful programs or software that allow fraudsters or hackers to gain and utilize your personal information (Trivedi 2016). According to the current poll, an individual receives 20-40 emails each day on average, with 60-70 % of emails are being waste mails. Spam in the email is always increasing, causing users to become distracted, utilizing more storage space, burning time and energy, and producing more network traffic (Agarwal and Kumar 2018). There are multiple techniques and methods for analyzing spam and filtering it depending on ham (which comprises legitimate words) and spam (which comprises worthless words) (Gibson et al. 2020). The applications of spam detection and spam filtering are Yahoo Mails, Microsoft Outlook, Google Mails, and Internet Service Providers (ISPs) and are also utilized to protect employees and networks in Small-Medium Business organizations (SMBs)(Cichosz 2015).

Google Scholar has published 18,400 publications, IEEE Xplore holds 27 journal papers, and ScienceDirect holds 1,773 articles on spam detection and spam filtering using machine learning in the last five years. Researchers utilize a number of machine learning technologies to prevent spam. Logistic Regression, a supervised learning methodology, is one of the methodologies utilized. It predicts discrete values like True or False, 0 or 1, etc. Logistic Regression is used to predict the probability of an incident by compressing it into a logistic function. It will view '0', which represents ham, and '1', which represents spam (Dedeturk and Akay 2020; Gupta et al. 2018). The Random Forest algorithm, which is a supervised learning process, is another method utilized. It's

made up of multiple separate decision trees, each of which is composed of votes based on the overall classification of the data set. The tree with the most votes is preferred using this algorithm. A subset of the Random Forest approach is called a Decision Tree (Devi 2018; Wang et al. 2015). Another option is KNN, which is a supervised learning algorithm. The K-Nearest Neighbor algorithm is a technique for forming clusters based on shared characteristics. The model divides the clusters into classes and calculates whether each class is yielded or not. All of us must calculate the score from the K classes of documents. The scores that pass the threshold value after the statistical value of the class are taken into account (Ren and Shi 2016; Laksono, Basuki, and Bachtiar 2020). The Decision Tree algorithm, which is a supervised learning methodology, is another method utilized. It behaves well with both categorical and continuous variables. The root node is utilized as an input variable. The subset's values were repeatedly segmented until they fulfilled the target variables (Chakraborty and Mondal 2012).Our team has extensive knowledge and research experience that has translate into high quality publications(Bhansali et al. 2021; Jayanth et al. 2021; Sudhakar, Ravel, and Perumal 2021; Sathiyamoorthi et al. 2021; Deepanraj et al. 2021; Raju et al. 2021; Arun Prakash et al. 2020; Kamath et al. 2020; Shanmugam et al. 2021; Rajasekaran et al. 2020; Adhinarayanan et al. 2020; Rajesh et al. 2020; Aurtherson et al. 2021)

The existing system has a drawback in terms of accuracy. Due to inaccuracies in the existing system, various spam or trash emails turn up in the inbox of an email account, which aggravates the user. So that it requires to boost the proposed system's accuracy using machine learning. The aim of this research is to create the Novel Naïve Bayes Classifier algorithm for spam detection or spam filtering of email to improve accuracy and in comparison with the Decision Tree algorithm. The goal of this research is to identify spam more accurately.

MATERIALS AND METHODS

The research was conducted in the Image Processing Lab of the Saveetha School of Engineering at the Saveetha Institute of Medical and Technical Sciences, which has a huge infrastructure system for collecting experimental data. For this research, two groups are recruited, each of which needs 20 samples (Wei 2018). The Decision Tree technique is in group 2 while the Novel Naive Bayes Classifier is in group 1. The technique employs G-power 80 % (Agarwal and Kumar 2018) with an alpha value of 0.05 and a beta value of 0.95 with a 95 % confidence interval.

The spam dataset can be obtained from the Kaggle website. A 491-KB file is provided as a CSV file containing the spam dataset, which comprises ham and spam elements. The phrases ham (legitimate mail) and spam (worthless mail) assist us to decide whether an email is a ham (legitimate mail) or spam (worthless mail).

Google Colab was used to work on this project. Always Google Colab was utilized to deploy my script. All python scripts are run with full GPU access and no configuration using Google Colab. It's broken up into cells and is used to put together all of the Python procedures that can be invoked. The hardware of the system used in this project was an 8 GB RAM, 1.8 TB ROM Windows 10 64-bit operating system with an Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.70 GHz processor.

Novel Naive Bayes Classifier

The Novel Naive Bayes Classifier is among the supervised learning algorithms. The Bayes theorem is used to calculate an event's probability. High independence, the ability to handle big datasets, and reliance on probability distributions are among its properties (Cichosz 2015; Pooja and Bhatia 2018)**.** The Bayes theorem is used to determine the probability distribution based on the frequency of the dataset. The Novel Naive Bayes classifier picks the class with the maximum posterior probability from the probability distribution. The equation for the posterior probability is shown in equation (1) (Agarwal and Kumar 2018).

$$\square(\square|\square) = \frac{\square(\square|\square)\ \square(\square)}{\square(\square)} \quad (1)$$

Where

$X=(c_1,c_2,c_3,.....)$, $Y=(d_1,d_2,d_3,.....)$

$\square(\square)$is evidence probability,

$\square(\square)$is prior probability,

$\square(\square|\square)$ is a conditional probability,

$\square(\square|\square)$ is posterior probability.

Pseudocode

Input: Spam dataset

Output: Accuracy of spam detection

Step-1: To begin, download, and install all of the required programs and libraries.
Step-2: You should import all of the programs you've downloaded.
Step-3: The ham and spam keywords must be extracted from the dataset.
Step-4: Among other things, it needs to clean the dataset by removing single-letter words, shrinking all-white spaces, tokenizing all types of communications, deleting all punctuation, and turning all characters to lowercase.
Step-5: The data should be separated into two sections test and training datasets.
Train_variables: C_Train, D_Train.
Test_variables: C_Test, D_Test
Step-6: When they come all around the words spam and ham, they should instruct the machine which one is spam and which is ham.
Step-7: Build the Novel Naive Bayes classifier and use the training dataset to train it.
NB=NaiveBayesClassifier()
NB.fit (C_Train, D_Train)
Step-8: Using the Bayes theorem, predict the probability distribution P(D|C) for each class.
Step-9: Measure the accuracy by developing a confusion matrix.
accuracy = sum (C_Test.Label == D_ Test.predicted) /len (C_Test)

Decision Tree algorithm

The Decision Tree algorithm, which is a supervised learning methodology, is another method utilized. It behaves well with both categorical and continuous variables. The root node is utilized as an input variable. The subset's values were repeatedly segmented until they fulfilled the target variables (Chakraborty and Mondal 2012). By learning decision rules from past data, the purpose of employing a Decision Tree is to develop a training model that can be used to predict the class or value of the target variable (training data) (Saini 2021).

Pseudocode

Input: Spam dataset
Output: Accuracy of spam detection

Step-1: To begin, download and install all of the required programs and libraries.
Step-2: You should import all of the programs you've downloaded.
Step-3: The ham and spam elements must be selected from the dataset.
Step-4: To clean up the dataset, all types of communications were tokenized, single-letter words were deleted, all punctuation was erased, all characters were converted to lowercase, all-white spaces were reduced, and so on.
Step-5: The data should be separated into two sections test and training datasets.
Train: E_Train, F_Train.
Test: E_Test, F_Test.
Step-6: The Decision Tree model is installed, and the model is trained using the training dataset.
DT= DecisionTree algorithm()
DT.fit(E_Train.iloc[:,6:], F_Train)
Step-7: Estimate the accuracy of the test dataset using the confusion matrix.
confusion__matrix(D_Test, DT.predict(C_Test.iloc[:,6:])
DT.score(C_Test.iloc[:,6:], D_Test)

Statistical Analysis

The statistical variables mean, standard deviation, standard error mean, mean difference, sig, and F value are determined using IBM SPSS Version 28. In this paper, the Independent Sample T-Test analysis is done. The spam dataset, which contains 4862 ham words and 728 spam terms, is used to detect spam. Spam and ham are the dependent variables. Two independent variables are the number of words counted and the accuracy (Udge et al. 2019).

RESULTS

When comparing the Novel Naive Bayes Classifier to the Decision Tree technique, accuracy is utilized as a criterion. The spam dataset is crucial for comparing and analyzing algorithms. The results indicate that the Novel

Naive Bayes Classifier surpasses the Decision Tree algorithm model n terms of accuracy. The Novel Naive Bayes Classifier has a mean accuracy of 98.05 %, whereas the Decision Tree technique has a mean accuracy of 91.80 %.
Table 1shows the statistical measurements for both Novel Naive Bayes Classifier and Decision Tree algorithm and the measurements consist of Mean, Standard Deviation, and Standard Error Mean of accuracy. The average accuracy of the Decision Tree algorithm is 91.80 %, which is smaller than the average accuracy of the Novel Naive Bayes Classifier is 98.05 %. The Standard Deviation of the Decision Tree algorithm is 0.83351, which is larger than the Standard Deviation of the Novel Naive Bayes Classifier, which is 0.759.
Table 2 shows the Independent Sample T-Test, which calculates the Mean Difference and Standard Error Difference with a 95 % confidence interval for both the Novel Naive Bayes Classifier and the Decision Tree method. The accuracy of both models is 0.022 ($p<0.05$), which is a 2-tailed significant value.
Figure 1 shows the Bar graph with Confidence intervals of 95%, error bars, and SE for comparing the mean accuracy of both the Novel Naive Bayes Classifier and the Decision Tree algorithm. The mean accuracy is on the Y-axis, while the groups (NB, DT) are on the X-axis. The mean accuracy of the Decision Tree algorithm is 91.80%, which falls shorter than the mean accuracy of the Novel Naive Bayes Classifier, which is 98.05 %.

DISCUSSION

The data is subjected to an Independent Sample T-Test using IBM SPSS version 28 software. The results show that the Novel Naive Bayes Classifier is much more accurate than the Decision Tree algorithm. The Novel Naive Bayes Classifier has a mean accuracy of 98.02 %, which is relatively higher than the Decision Tree algorithm's mean accuracy of 91.80 %.

Similar findings related to the Novel Naive Bayes Classifier are (Agarwal and Kumar 2018; Pooja and Bhatia 2018). The spam dataset is used in those papers. The data is filtered and categorized into two groups: the train and test datasets. A training dataset is used to train the model after it has been loaded. The model is then run, and a test dataset is used to produce the confusion matrix. The Naive Bayes Classifier has a number of benefits, including the ability to predict class data quickly and easily, the ability to deal with independent assumptions, and the ability to work with categorical rather than numerical input elements (Cichosz 2015). The following are the drawbacks: The 'zero-frequency problem' occurs when the Naive Bayes Classifier, which assumes all predictors are independent, allots likelihood '0' to categorical variables that are existent in the test dataset and not in the training dataset (Trivedi 2016). Opposite findings related to spam detection and spam filtering are (Singh, Pamula, and Shekhar 2018). The spam dataset is used in those articles. The information was cleaned and separated into two groups: training and test datasets. They have used the Support Vector Machine approach, which is a type of supervised learning technology, in this example. It displays the data from the input as a point in n-dimensional space. Each feature is a value associated with a specific coordinate. The categorization is finished by drawing a hyperplane with vectors that clearly distinguish the two classes. The train and test datasets are used to load, train, and test this model. The accuracy computations and the confusion matrix have been completed.

The limitations in the detection of spam are avoiding the selection of ham mail to the dump rather than spam, detailed link evaluation, getting sufficient keywords, making a better text categorization, preserving data security for all users. In the future, We'll work to reduce the error rate to less than 1% and increment spam detection or spam filtering to over 99 %. We'll also seek to use Deep Learning techniques to generate spam detection or spam filtering.

CONCLUSION

According to this research project, the accuracy of the Novel Naive Bayes Classifier is bigger than the accuracy of the Decision Tree method in classifying spam. The Novel Naive Bayes Classifier has a 98.05 %, which is quite higher than the 91.80 % of the Decision Tree technique.

DECLARATION

Conflict of Interest

No conflict of interest in this manuscript.

Authors Contribution

Author KVK was involved in data collection, data analysis, and writing the manuscript. Author MRwas involved in the conceptualization, data validation, and critical review of the manuscript.

Acknowledgments

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

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

1. Cyclotron Technologies, Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. Adhinarayanan, Rajesh, AravindhRamakrishnan, Gopal Kaliyaperumal, Melvinvíctor De Poures, Rajesh Kumar Babu, and DamodharanDillikannan. 2020. "Comparative Analysis on the Effect of 1-Decanol and Di-N-Butyl Ether as Additive with diesel/LDPE Blends in Compression Ignition Engine." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, June, 1–18.
2. Agarwal, Kriti, and Tarun Kumar. 2018. "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization." *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. https://doi.org/10.1109/iccons.2018.8662957.
3. Arun Prakash, V. R., J. Francis Xavier, G. Ramesh, T. Maridurai, K. Siva Kumar, and R. Blessing Sam Raj. 2020. "Mechanical, Thermal and Fatigue Behaviour of Surface-Treated Novel Caryota Urens Fibre–reinforced Epoxy Composite." *Biomass Conversion and Biorefinery*, August. https://doi.org/10.1007/s13399-020-00938-0.
4. Aurtherson, P. Babu, Bhanu Teja Nalla, Karthikeyan Srinivasan, Kulmani Mehar, and Yuvarajan Devarajan. 2021. "Biofuel Production from Novel Prunus Domestica Kernel Oil: Process Optimization Technique." *Biomass Conversion and Biorefinery*, May. https://doi.org/10.1007/s13399-021-01551-5.
5. Bhansali, Karan J., Kamlesh R. Balinge, Subodh U. Raut, Shubham A. Deshmukh, M. Senthil Kumar, C. Ramesh Kumar, and Pundlik R. Bhagat. 2021. "Visible Light Assisted Sulfonic Acid-Functionalized Porphyrin Comprising Benzimidazolium Moiety for Photocatalytic Transesterification of Castor Oil." *Fuel* 304 (November): 121490.
6. Chakraborty, Sarit, and BikromadittyaMondal. 2012. "Spam Mail Filtering Technique Using Different Decision Tree Classifiers through Data Mining Approach - A Comparative Performance Analysis." *International Journal of Computer Applications in Technology* 47 (16): 26–31.
7. Cichosz, Paweł. 2015. "Naïve Bayes Classifier." *Data Mining Algorithms: Explained Using R*, January, 118–33.
8. Dedeturk, Bilge Kagan, and Bahriye Akay. 2020. "Spam Filtering Using a Logistic Regression Model Trained by an Artificial Bee Colony Algorithm." *Applied Soft Computing*. https://doi.org/10.1016/j.asoc.2020.106229.
9. Deepanraj, B., N. Senthilkumar, D. Mala, and A. Sathiamourthy. 2021. "Cashew Nut Shell Liquid as Alternate Fuel for CI Engine—optimization Approach for Performance Improvement." *Biomass Conversion and Biorefinery*, February. https://doi.org/10.1007/s13399-021-01312-4.
10. Devi, Khongbantabam Susila. 2018. "Random Forests Spam Email Classification System." *Journal of Computer Engineering & Information Technology*. https://doi.org/10.4172/2324-9307.1000190.
11. Gibson, Simran, Biju Issac, Li Zhang, and Seibu Mary Jacob. 2020. "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms." *IEEE Access*. https://doi.org/10.1109/access.2020.3030751.
12. Gupta, Mehul, Aditya Bakliwal, Shubhangi Agarwal, and Pulkit Mehndiratta. 2018. "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers." *2018 Eleventh International Conference on Contemporary Computing (IC3)*. https://doi.org/10.1109/ic3.2018.8530469.
13. Jayanth, BellappuVenkat, Melvin Victor Depoures, Gopal Kaliyaperumal, DamodharanDillikannan, DilipsinghJawahar, KumaranPalani, and Ganesha Prasad MeravanigeeShivappa. 2021. "A Comprehensive Study on the Effects of Multiple Injection Strategies and Exhaust Gas Recirculation on Diesel Engine Characteristics That Utilize Waste High Density Polyethylene Oil." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, June, 1–18.
14. Kamath, Manjunath, Subha Krishna Rao, Jaison, Sridhar, Kasthuri, Gopinath, Sivaperumal, and Shantanu Patil. 2020. "Melatonin Delivery from PCL Scaffold Enhances Glycosaminoglycans Deposition in Human Chondrocytes – Bioactive Scaffold Model for Cartilage Regeneration." *Process Biochemistry* 99 (December): 36–47.
15. Laksono, Eko, Achmad Basuki, and Fitra Bachtiar. 2020. "Optimization of K Value in KNN Algorithm for Spam and Ham Email Classification." *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*. https://doi.org/10.29207/resti.v4i2.1845.
16. Pooja, and Komal Kumar Bhatia. 2018. "Spam Detection Using Naive Bayes Classifier." *International Journal of Computer Sciences and Engineering*. https://doi.org/10.26438/ijcse/v6i7.712716.
17. Rajasekaran, S., D. Damodharan, K. Gopal, B. Rajesh Kumar, and Melvin Victor De Poures. 2020. "Collective Influence of 1-Decanol Addition, Injection Pressure and EGR on Diesel Engine Characteristics Fueled with diesel/LDPE Oil Blends." *Fuel* 277 (October): 118166.
18. Rajesh, A., K. Gopal, De Poures Melvin Victor, B. Rajesh Kumar, A. P. Sathiyagnanam, and D. Damodharan. 2020. "Effect of Anisole Addition to Waste Cooking Oil Methyl Ester on Combustion, Emission and Performance Characteristics of a DI Diesel Engine without Any Modifications." *Fuel* 278 (October): 118315.
19. Raju, P., K. Raja, K. Lingadurai, T. Maridurai, and S. C. Prasanna. 2021. "Glass/Caryota Urens Hybridized Fibre-Reinforced nanoclay/SiC Toughened Epoxy Hybrid Composite: Mechanical, Drop Load Impact, Hydrophobicity and Fatigue Behaviour." *Biomass Conversion and Biorefinery*, March. https://doi.org/10.1007/s13399-021-01427-8.
20. Ren, Biyi, and Yuliang Shi. 2016. "Research On Spam Filter Based On Improved Naive Bayes and KNN Algorithm." *Proceedings of the 2016 4th International Conference on Machinery, Materials and Computing Technology*. https://doi.org/10.2991/icmmct-16.2016.220.
21. Saini, Anshul. 2021. "Decision Tree Algorithm - A Complete Guide - Analytics Vidhya." August 29, 2021. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/.
22. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021.

"Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." *Environmental Progress & Sustainable Energy* 40 (6). https://doi.org/10.1002/ep.13696.
23. Shanmugam, Rajasekaran, DamodharanDillikannan, Gopal Kaliyaperumal, Melvin Victor De Poures, and Rajesh Kumar Babu. 2021. "A Comprehensive Study on the Effects of 1-Decanol, Compression Ratio and Exhaust Gas Recirculation on Diesel Engine Characteristics Powered with Low Density Polyethylene Oil." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects* 43 (23): 3064–81.
24. Singh, Manmohan, Rajendra Pamula, and Shudhanshu Kumar Shekhar. 2018. "Email Spam Classification by Support Vector Machine." *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*. https://doi.org/10.1109/gucon.2018.8674973.
25. Sudhakar, M. P., Merlyn Ravel, and K. Perumal. 2021. "Pretreatment and Process Optimization of Bioethanol Production from Spent Biomass of GanodermaLucidum Using Saccharomyces Cerevisiae." *Fuel* 306 (December): 121680.
26. Trivedi, Shrawan Kumar. 2016. "A Study of Machine Learning Classifiers for Spam Detection." *2016 4th International Symposium on Computational and Business Intelligence (ISCBI)*. https://doi.org/10.1109/iscbi.2016.7743279.
27. Udge, Ganesh, Mahesh Mohite, ShubhankarBendre, YogeshwarBirnagal, and DishaWankhede Mrs. 2019. "Statistical Analysis for Twitter Spam Detection." *International Journal of Scientific Research in Science, Engineering and Technology*, May, 624–29.
28. Wang, W. B., F. Yin, H. Sun, and P. Li. 2015. "Random Forest Algorithm for Spam Filtering Based on Machine Learning." In *Electronic Engineering and Information Science*, 225–28. CRC Press.
29. Wei, Qijia. 2018. "Understanding of the Naive Bayes Classifier in Spam Filtering." https://doi.org/10.1063/1.5038979.

TABLES AND FIGURES

Table 1. The statistical evaluations of the Novel Naive Bayes Classifier and the Decision Tree algorithm are computed. The Novel Naive Bayes classifier has a mean accuracy of 98.05 %, while the Decision Tree method has a mean accuracy of 91.80 %. Both the Novel Naive Bayes Classifier and the Decision Tree algorithm have standard deviations of 0.759 and 0.833, respectively.

	Groups	N	MEAN	Std. Deviation	Std. Error Mean
Accuracy	NB	20	98.05	0.75915	0.16975
	RF	20	91.80	0.83351	0.18638

Table 2. A Statistical Independent T-Test with a 95 % confidence interval was performed between the Novel Naive Bayes Classifier and the Decision Tree technique. The accuracy has a 2-tailed significant value which is 0.022($p<0.05$).

		Levene's Test for Equality of Variances		T-Test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95 % Confidence Interval of the Difference	
									Lower	Upper
Accuracy	Equal variances assumed	1.17	.285	24.792	38	.022	6.25	.2521	5.7395	6.7604

	Equal variances not assumed			24.792	37.67	.022	6.25	.2521	5.7395	6.7604

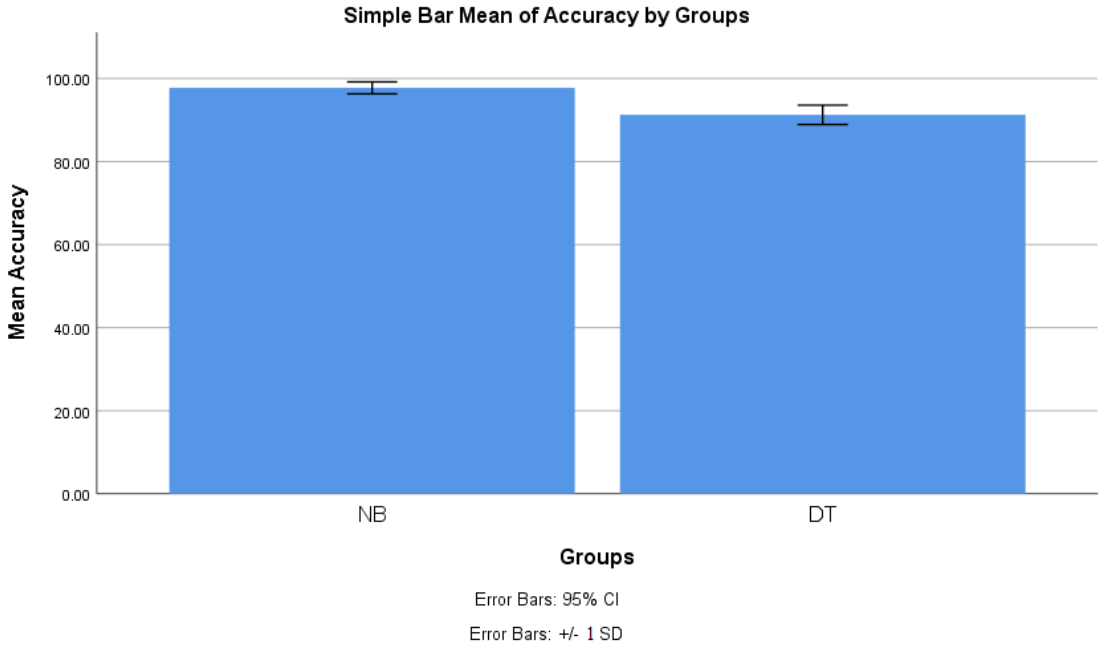

Fig. 1. Bar Graph for comparing the Novel Naive Bayes Classifier and Decision Tree algorithm with a confidence interval of 95% and with +/- 1 SD. The Novel Naive Bayes Classifier is slightly stronger than the Decision Tree algorithm. The X-axis shows the groups (NB, DT), whereas the Y-axis shows the mean accuracy.