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): 368-377
DOI: 10.2478/bjlp-2022-004039

Higher Accuracy of Spam Mail Prediction using Random Forest Algorithm Comparing with Multinomial Naive Bayes 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 experiment will primarily collect samples from two groups. The Random Forest Algorithm belongs to Group-1, while the Multinomial Naive Bayes Algorithm belongs to Group-2. The sample sizes were all taken at the same time for both the Algorithms. The G-Power in the test set will be at 80%. **Result:** Data is processed in the given model so that Machine learning can function effectively. Emails are used as inputs for the Multinomial Naive Bayes algorithm, which generates a probabilistic index and determines whether the email is spam or not. The Random Forest Algorithm outperforms the Multinomial Naive Bayes Algorithm, and our hypothesis is significant with a significance value of 0.002 ($p<0.1$). **Conclusion:** These results were achieved through machine learning models such as Multinomial Naive Bayes, and Random Forest Algorithms. In this paper, here demonstrated that for the spam filtering method the most efficient algorithms are Random Forest Algorithm and MNB were given as they have the highest level of accuracy.

Keywords

Innovative Spam Prediction**,** Random Forest Algorithm, Classifier, Filtering, Machine learning, Multinomial Naive Bayes.

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, compared the performance of two machine

learning techniques for spam detection including the Random Forest Algorithm classifier Compared with the Multinomial Naive Bayes classifier. Multinomial Naive Bayes 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 Random Forest Algorithm comparing with Multinomial Naive Bayes Algorithm. In GoogleScholar this article is published 1310 times, and in IEEE Explore, this article is published 80 times in the past 5 years. In these 2 databases, the most cited articles and their findings are, Comparing different supervised machine learning algorithms for disease prediction (Uddin et al. 2019). That the preliminary discussion in the research background looks at how Automatic Spam Detection on Gulf Dialectical Arabic Tweets (Alorini and Rawat 2019). This suggested method identifies e-mail spam in both textual and speech-enabled e-mails. In terms of text extraction speed, performance, cost efficiency, and accuracy, the suggested GDP NLP technique gives a greater spam detection rate (Ismail et al. 2022). Here The technology recognizes the required features for categorizing spam emails automatically. The suggested system is based on the Genetic Algorithm and the Random Weight Network (Faris et al. 2019). From the above literature analysis and study, the paper (Kontsewaya, Antonov, and Artamonov 2021) is most relevant to our study and done most of the analysis.

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 (Gaurav et al. 2019).

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 Random Forest Algorithm and Group-2 Belongs to the Multinomial Naive Bayes Algorithm. The sample of both groups is 30% and 70% of the total samples. 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 (Rafat et al. 2022). 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 (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.

Random Forest Algorithm

Random Forest Algorithm is a probabilistic learning method and it is one of the most important algorithms in Supervised Machine learning. It can be used for both classification and regression purposes. The algorithms made with high dimensionality can be capable of handling large datasets. The Random forest algorithm has more no of trees which helps prevent overfitting the model. It can handle missing values easily. Random forests are very flexible and possess higher accuracy values. A Random forest is a predictive tool, not a descriptive tool. Normalization is not required as it uses a rule-based approach. The regression problems can be solved using Mean Square Error(MSE) (1) and classification problems can be solved using the Gini Index Equation (2) used to decide how many nodes are on the decision tree branch.

MSE Equation

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (fi - yi)^2 \text{----------(1)}$$

From the above Equation, The N is the number of data points. fi is the value returned by the model and Yi is the actual value of data point i.

Gini Index Equation

$$Gin = 1 - \sum_{i=c} (pi)^2 \text{----------(2)}$$

In the above Equation, Pi represents the Relative frequency of the class you are observing in the dataset and C represents the number of classes.

Pseudocode for Random Forest Algorithm

Input: Training dataset
Output: Classifier accuracy
A training set S:= (x1, y1), . . ,(xn, yn), features F, and a number of trees in forest B.
function RandomForest(S, F)
H ← ∅ 3
    for i ∈ 1, . . . , B do
        S (i) ← A bootstrap sample from S
        hi ← RandomizedTreeLearn(S (i) , F)
        H ← H ∪ {hi}
    end for
return H
end function
function RandomizedTreeLearn(S , F)
At each node:
    f ← very small subset of F
    Split on best feature in f
return The learned tree
end function

Multinomial Naive Bayes Algorithm

Multinomial Naive Bayes is a probabilistic learning method used in Natural Language Processing (NLP). Using the Bayes theorem, this approach guesses the tag of a

text, such as an email or a news item. It computes the likelihood of each tag for a given sample and returns the tag with the highest likelihood. The Naive Bayes classifier is a group of algorithms that all follow the same basic principle: each feature being classified is unconnected to any other feature. One character's existence or absence has no bearing on the presence or absence of another. The Equation for Naive Bayes efficiency and increase is used for text data analysis and multi-class scenarios (3). To understand how the Naive Bayes theorem works, you must first comprehend the Bayes theorem concept, as it is based on it. The Bayes theorem, developed by Thomas Bayes, states that previous knowledge of event-related circumstances does not affect the likelihood of an event occurring. It is calculated using the following equation:

$$P(A|B) = P(A) \star P(B|A)/P(B)\text{-------} (3)$$

The probability of class A when predictor B is already provided.
P(B) = prior probability of predictor B
P(A) = prior probability of class A
P(B|A) = occurrence of predictor B given class A probability
This Equation helps in calculating the probability of the tags in the text.

Pseudocode for Multinomial Naive Bayes

Input: Training dataset
Output: Classifier accuracy
The first step is Data collection.
Pre-processing and text cleaning of the train data.
Fit the Training Data Set to the Multinomial Naive Bayes.
Now Predict the Results for test split data.
Define class
Def MultinomialNB()
if(condition satisfies)
return accuracy
else
return previous step
End
Create the Confusion Matrix and find the Test Accuracy Results.
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 Random Forest Algorithm and Multinomial Naive Bayes 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 Random Forest Algorithm and Multinomial Naive Bayes. 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 compared. It gives the Mean value for the Group statistics. 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 Random Forest Algorithm Classifier and Multinomial Naive Bayes 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 % (Wood and Krasowski 2020). Finally sent all the Test sizes and also their Accuracy into the Spss tool and found the Average Accuracy values of the Random Forest Algorithm Classifier and Multinomial Naive Bayes Classifier.

RESULT

In the proposed model, data is trained so that Machine learning can work properly. After applying the Multinomial Naïve Bayes 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 Random Forest Algorithm gave us an accuracy of 91% and Multinomial Naive Bayes gave us an accuracy of 90% 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 Random Forest Algorithm are 90.299 and 0.47259. The Multinomial Naive Bayes Algorithm's mean accuracy is 87.997 and the standard deviation is 2.14172. The Random Forest Algorithm also obtained a standard error mean rate of 0.14945 whereas the Multinomial Naive Bayes Algorithm obtained an error mean rate of 0.67727 as given in Table 2.

Then an independent sample test of 10 samples was performed, Random Forest Algorithm obtained a mean difference of 2.302 and a standard error difference of 0.69356. When compared to other algorithm's performance, the Random Forest Algorithm performed better than the Multinomial Naive Bayes Algorithm and the significance value of 0.002 shows that our hypothesis is valid 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, Modeling Classifier, Implementation, and Predicted Accuracy.

The GGraph represents a bar chart of the simple bar mean accuracy, with the Random Forest Algorithm achieving an accuracy of approximately 91%, and the Multinomial Naive Bayes Algorithm achieving 90%. 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 0.002<0.05. When comparing the two algorithms the performance of the Random Forest Algorithm achieved a better performance than Multinomial Naive Bayes Algorithm as given in Fig. 1.

DISCUSSION

The Random Forest Algorithm has better accuracy than Multinomial Naive Bayes. 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 Random Forest Algorithm has an accuracy of approximately 91%, which is higher than that of the Multinomial Naive Bayes Algorithm which is 90%. Random Forest Algorithm has a better significance of 0.002 while using the independent samples T-test.

The mean accuracy and standard deviation for the Random Forest Algorithm are 90.299 and 0.47259 using a missing value imputation and a machine learning model to get an accuracy of 91%. The Multinomial Naive Bayes Algorithm's mean accuracy is 87.997 and the standard deviation is 2.14172. In the paper, (Kontsewaya, Antonov, and Artamonov 2021) the Random Forest Algorithm obtained an accuracy of 84%, and (Sharaff and Rao 2020) the Multinomial Naive Bayes Algorithm achieved an accuracy of 89% accuracy. Based on the literature survey, it is evident that the Random Forest Algorithm performs better than Multinomial Naive Bayes. By running independent sample tests in IBM's SPSS statistical program, it can be seen that the difference between the two algorithms is statistically significant at 0.002. Using IBM's SPSS statistical tool, independent sample analysis confirmed that the difference between the two methods is statistically significant at 0.002<0.05. The mean and standard deviation are determined using the SPSS statistical tool. Random Forest Algorithm outscored other algorithm classification accuracy by 91% 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 (Liu, Lu, and Nayak 2021). Attributes like this can result in improved accuracy and exact precision values. There exists a strong relationship between the content and the subject of the emails (Dada et al. 2019). 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. 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 Multinomial Naive Bayes, and Random Forest Algorithms. In this paper, demonstrated the spam filtering method the most efficient algorithms are the Random Forest Algorithm and MNB was 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, A. A. 2021. "Advances in Spam Detection for Email Spam, Web Spam, Social Network Spam, and Review Spam: ML-Based and Nature-Inspired-Based Techniques." https://doi.org/10.3233/JCS-210022.

Alorini, Dema, and D. Rawat. 2019. "Automatic Spam Detection on Gulf Dialectical Arabic Tweets." http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8685659.

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.

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.

Ismail, Safaa S. I., Romany F. Mansour, Rasha M. Abd El-Aziz, and Ahmed I. Taloba. 2022. "Efficient E-Mail Spam Detection Strategy Using Genetic Decision Tree Processing with NLP Features." *Computational Intelligence and Neuroscience* 2022 (March): 7710005.

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.
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.
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.
Sharaff, Aakanksha, and Ulligaddala Srinivasa Rao. 2020. "Towards Classification of Email through Selection of Informative Features." http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9071488.
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.
Uddin, Shahadat, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. 2019. "Comparing Different Supervised Machine Learning Algorithms for Disease Prediction." *BMC Medical Informatics and Decision Making* 19 (1): 1–16.
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 Algorithms. The Data Accuracy for the Random Forest Algorithm (Group-1) and Multinomial Naive Bayes (Group-2) with different Test sizes has been taken. In these different Test Sizes, the Accuracy value for Random Forest Algorithm is 91.16 and the Multinomial Naive Bayes is 90.35.

S. No.	Test Size	Group-1 Accuracy	Group-2 Accuracy
1	0.2	91.16	90.32
2	025	91.03	90.35
3	0.3	90.26	89.97
4	0.35	90.11	89.31
5	0.4	90.58	88.64

6	0.45	89.95	87.94
7	0.5	89.97	87.36
8	0.55	90.18	86.47
9	0.6	89.96	85.58
10	0.7	89.79	84.03

Table 2. Group Statistics the mean accuracy and standard deviation for Random Forest Algorithm are 90.2990 and 0.47259 and For Multinomial Naive Bayes(MNB) Algorithm is 87.9970 and 2.14172.

Group Statistics					
	RF, MNB	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	RF	10	90.2990	.47259	.14945
	MNB	10	87.9970	2.14172	.67727

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 0.002<0.05 and Std. Error Difference is noticed as .69356.

		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	13.295	.002	.69356	.84487	3.75913
	Equal variances not assumed			.69356	.75397	3.85003

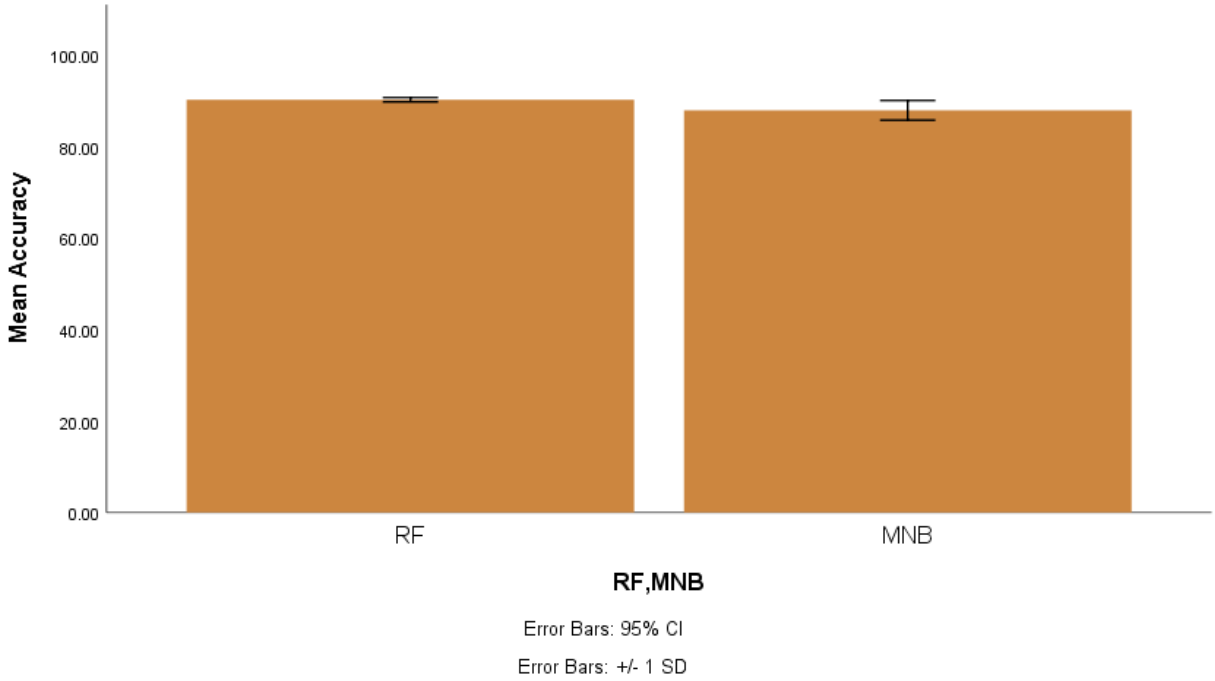

Fig. 1. Simple Bar Mean of Accuracy by Random Forest Algorithm and Multinomial Naive Bayes(MNB), the bar chart representing the comparison of mean accuracy of Random Forest Algorithm is 90.2990 and Multinomial Naive Bayes Algorithm is 87.9970. X-Axis: Random Forest Algorithm vs Multinomial Naive Bayes Algorithm. Y-Axis: Mean accuracy. The error bars are 95% for both algorithms. The Standard Deviation Error Bars are +/- 1 SD.