# MOVIE RECOMMENDATION SYSTEMS USING RANDOM FOREST AND COMPARING PREDICTION ACCURACY WITH NAÏVE BAYES BASED COLLABORATIVE FILTERING

**Chamala Karthik Reddy[1], Terrance Frederick Fernandez[2*]**

**Article History:** **Received:** 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

**Abstract**

**Aim:** The aim of the research article is to improve the accuracy of movie recommendation systems using a novel Random forest (RF) algorithm in comparison with a Naive Bayes (NB) algorithm. Materials and methods: The dataset used in this paper was collected from the Movie lens database. The sample size for the movie recommendation system was sample 20 (Group 1 = 10 and Group 2 = 10) and the calculation was performed utilizing G-power 0.8 with alpha and beta qualities of 0.05 and 0.2 with a confidence interval of 95%. The movie recommendation system is performed by the Random forest (RF) classifier with a number of samples (N=10) and Naive Bayes (NB) model with a number of samples (N=10).
**Results:** The Random forest (RF) classifier has a 94.30 percent higher accuracy rate when compared to the accuracy rate of the Naive Bayes (NB) model, which is 82.56 percent. The study has a significance value of p=0.037.
**Conclusion:** The Random forest (RF) classifier provides better outcomes in terms of accuracy rate when compared to the Naive Bayes (NB) model for movie recommendation systems.

**Keywords:** Innovative Recommendation System, Movie Recommendation, Novel Random forest, Naive Bayes, Machine Learning.

[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 Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode:602105.

## 1. Introduction

Many E-commerce websites and applications, such as YouTube and Netflix, now include a recommendation system (Jannach and Zanker 2022). The main goal of the recommendation engine is to estimate which products the user would be concerned about. The majority of recommendation systems are being researched using content-based and collaborative methodologies. However, scalability, data sparsity, overspecialization, and cold-start issues plague these systems, resulting in low-quality recommendations and coverage (Khan, Chan, and Chua 2019). In this study, a novel Random forest (RF) approach for a movie recommendation system is developed, and the results are compared to the Naive Bayes (NB) algorithm (H.-R. Zhang, Min, and He 2014).Recommendation systems are extensively employed in e-commerce, social media sites, Tourism, Digital advertising, and a variety of other industries, and they have higher impacts and prospects(Alhamid et al. 2016; Klašnja-Milićević et al. 2018; Ayata, Yaslan, and Kamasak 2018).

Over the last five years, the researchers have created several alternative methodologies and strategies for movie production recommendation systems. There are 175 research publications on IEEE Xplore, and 196 articles on Google Scholar. The major discussion on Novel Bayes and Hidden Bayes took place in another paper(Adomavicius and Kwon 2007). How they can give you a leg up on what you're doing now. In another paper(Jiang, Zhang, and Cai 2009), a method is proposed that shows how the Naive Bayes approach for text categorization can be improved.Another paper(W. Zhang and Gao 2011) discusses how they generally fail to match information content and proposes a model that learns word representations containing semantic conditions data as well as rich emotion content using a combination of unsupervised and supervised techniques. Matrix decomposition was introduced into the tree structure in(Karimi et al. 2013), and it was used to speed up tree creation and forecast tree node ratings. The authors of(H.-R. Zhang and Min 2016) present a paradigm for developing recommender systems that blends three-way decision with Random Forests. To plot user recommendations for items, a three-way decision was introduced: "recommend," "not recommend," or "actively consult the user" for his or her preference.The authors of(Theocharous, Thomas, and Ghavamzadeh 2015) offer a framework for Personalized Ad Recommendation (PAR) systems that uses reinforcement learning to develop optimal policies. To learn a PAR policy quickly, Random Forest regression is employed. Ajesh et al.(Ajesh, Nair, and Jijin 2016) suggested a system that employs clustering and random forest as multilevel techniques to forecast suggestions based on(Shah, Parikh, and Deshpande 2016)user ratings while focusing on the users' mindset and current trends. M Shah et al.(Shah, Parikh, and Deshpande 2016) propose a movie recommendation system based on latent graph properties in highly randomized trees Our team has extensive knowledge and research experience that has translated into high quality publications(K. Mohan et al. 2022; Vivek et al. 2022; Sathish et al. 2022; Kotteeswaran et al. 2022; Yaashikaa, Keerthana Devi, and Senthil Kumar 2022; Yaashikaa, Senthil Kumar, and Karishma 2022; Saravanan et al. 2022; Jayabal et al. 2022; Krishnan et al. 2022; Jayakodi et al. 2022; H. Mohan et al. 2022).

The major drawback with the existing Naïve Bayes (NB) method is that it has cold start and data sparsity problems. Recommended systems might not be able to provide accurate predictions when there is not a significant amount of data to work with. This research proposes a novel movie recommendation system based on the Random Forest (RF) algorithm with collaborative filtering to solve this drawback. In comparison to the NB algorithm, the experimental findings reveal that the suggested RF algorithm provides a dependable model that is accurate and generates more personalized movie recommendations.

## 2. Materials and Methods

The study was carried out at Saveetha School of Engineering. The segmentation dataset was collected from the Movie Lens repository. This research uses two different methods: the new Random Forest and the Naive Bayes. It entails two sample sets of ten samples each and a total of twenty samples, with a pretest power of 0.8. The test size for training the RF is around 20% of the whole dataset, with the remaining 80% utilized for the training dataset. For the movie recommendation system, Python software is used to obtain the results. Using earlier findings from (Khanvilkar and Vora 2018) at clinicalc.com, the sample size was calculated by setting the threshold at 0.05, G power at 80%, and confidence interval at 95%.

**Naive Bayes**

The Bayes theorem is used to create a probabilistic classifier with features that are independent of each other. Each attribute is thought to influence the likelihood that a given test instance belongs to a specific class. We can use Bayes theorem to calculate the conditional distribution of correctly predicting a class provided a feature.

$$P\ C|A\ = P(A|C) * \frac{P(C)}{P(A)} \qquad (1)$$

Where P(C/A) – Posterior Probability, P(A/C) – Likelihood, P(A) - Prior Probability.

The higher probability is chosen as the label of the review based on the four labels. Four labels are Positive, Negative, Over Positive, Over Negative. All the reviews are classified and the count for each label is determined for building a recommendation system.

**Pseudocode**

Step 1: Movie recommendation system_Input Features
Step2: Classification output
Step 3: Function: Naïve Bayes (Input features Var = 1…n)
Step 4: Training dataset for movie recommendation system
Step 5: for each (condition) do
Step 6: Read the dataset of the efficient predicting of movie recommendation system
Step 7: Compute the Mean and Standard deviation in each class
Step 8: Compute the probability of Input features Var1… Var n utilizing gauss density equation
Step 9: Compute the likelihood for each class
Step 10: Acquire the maximum probability
Step 11: End
Step 12: Return Classification outcomes for efficient predicting of movie recommendation system

**Random forest**

Random Forests(Persson 2015) is a popular ensemble learning technique for classification problems. It identifies based on the results of the countless decision trees it produces during training, where the forest output is the mode of the desired outcomes from each decision tree. Random Forests attempt to average out the number of decision trees because trees are known to overfit data due to their low bias and high variation. Random Forests create decision trees based on random tests of training data, reducing deviation in the estimated model, improving performance, and preventing overfitting.

**Pseudocode**

Input: size of forest (N), condition attributes (C), decision attribute (d)
Output: Classification on motion picture recommendation system
Method: build RandomForest(Input features F = 1…m)
Step 1: Create an aggregated training ($St$) and testing ($Se$) set
Step 2: $i = 0$
Step 3: while ($i < N$) do
Step 4: shuffle C
Step 5: $RTi = buildRandom(St, C, d)$
Step 6: $i = i + 1$
Step 7: end while
Step 8: $RF = \{RT0, RT1, \ldots, RTN{-}1\}$
Step 9: End while
Step 10: Return Classification outcomes for movie recommendation system

**Statistical Analysis**

Python software is used to generate the results(Sanner 1999). A monitor with a resolution of 1024x768 pixels was required to train these datasets (10th gen, i5, 12GB RAM, 500 GB HDD). NB and RF algorithms are statistically analyzed using SPSS software(Hilbe 2004). The mean, standard deviation, and standard error mean statistical significance between the groups were determined using the independent sample t test, followed by a comparison of the two groups using SPSS software. Accuracy is a dependent variable, while RF and NB are independent variables.

## 3. Results

The accuracy rate of the RF classifier is compared to that of the NB classifier in Figure 1. The RF classifier has a higher accuracy rate of 94.30 when compared to the NB classifier, which has 82.56. The RF classifier is significantly different from the NB classifier ($p<0.05$ independent sample test). On the X-axis, RF and NB accuracy rates are plotted. Y-axis: Mean accuracy rate for keyword identification, ± 1 SD with 95 percent confidence interval. Table 1 presents the evaluation metrics of the comparison of the RF classifier with the NB classifier. The RF classifier has a 94.30 accuracy rate, whereas the NB classifier has 82.56 accuracy rate. In all parameters, the RF classifier outperforms the NB in the recommendation of movies, with a higher accuracy rate.

Table 2 displays the statistical computations for the RF and NB classifier, such as mean, standard deviation, and standard error mean. In the t-test, the accuracy rate parameter is used. The RF classifier has a mean accuracy rate of 94.30, while the NB classifier has 82.56, respectively. The standard deviation of RF is 0.69113 and the NB algorithm is 1.67839. The standard error mean of RF is 0.32048 and the NB algorithm is 1.78293. Table 3 shows the statistical computations for independent samples of RF compared to the NB classifier. The

significance value for the accuracy rate is 0.037. An independent sample T-test is applied for comparison of RF and NB algorithms with a confidence interval as 95% and level of significance as 0.77838. This independent sample test consists of significance as 0.001, significance (2-tailed), mean difference, standard error difference, and lower and upper interval difference.

## 4. Discussion

A comparative study has been presented between the Random Forest (RF) and Naïve Bayes (NB) models (Joshi et al. 2019). An accuracy analysis has been performed to investigate the importance of each of the input parameters. RF provides better accuracy in output when compared to the NB algorithm. The accuracy of the results produced by RF is better than the NB method. RF can significantly improve classification accuracy and time efficiency. This shows that the maximum accuracy is obtained quickly in the RF algorithm. The RF has an accuracy of 94.30% whereas the NB has an accuracy of 82.56%. This study's resultant accuracy was compared with the different existing models like pre-computed clustering for movie recommendation systems in real-time and the model has shown accuracy is 84.7%(Li, Liao, and Qin 2014). In(Forsati, Meybodi, and Neiat 2009), the authors gave the review of the recommender systems using collaborative filtering. Burke's study(Burke 2002) was one of the first exploratory methods of hybrid RSs. This paper examines the benefits and drawbacks of several recommendation methods and proposes a taxonomy for categorizing the manner in which they interact to produce hybrid RSs. The authors of AA Kothari's(Kothari and Patel 2015) study the use of unexpected and coverage as RS features and quality indicators in 2015. They claim that unpredictability and coverage are better at accounting for the value and utility of recommendations than accuracy. The movie recommendation system employs a Random forest algorithm, which has been demonstrated to be 83% accurate(Thomas and Vaidhehi 2018).Wei1 et al.(Wei et al. 2017) suggested a Random forest algorithm-based technique and obtained an accuracy of 88.5%. This study, however, has significant limitations as well as potential research areas. The following are the restrictions and future research issues. First, the studies were conducted with a tiny dataset, resulting in insufficient learning. During data gathering for this suggested study, not much information was obtained. As a result, this is an interesting future research topic in terms of continuous data collecting to demonstrate the relationship between information and efficiency change. Second, because image processing takes a lengthy time, the training time was considerable. Enhancing training time efficiency in the future by optimizing the learning structure is also a viable research area.

## 5. Conclusion

The proposed model exhibits the Random forest and Naïve Bayes algorithms, in which the RF algorithm has the highest values. The accuracy rate of the RF algorithm is 94.30% higher compared with the NB algorithm,which has an accuracy rate of 82.56%. The accuracy rate of the RF algorithm is efficient when compared with the NB algorithm,which has lower values in the movie recommendation system.

**Declarations**

**Conflict of Interests**

No conflict of interest in this manuscript.

**Authors Contributions**

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

**Acknowledgment**

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. Sankar Industries
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

## 6. References

Adomavicius, Gediminas, and Youngok Kwon. 2007. "New Recommendation Techniques for Multicriteria Rating Systems." IEEE Intelligent Systems 22 (3): 48–55.

Ajesh, A., Jayashree Nair, and P. S. Jijin. 2016. "A Random Forest Approach for Rating-Based Recommender System." In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 1293–97. ieeexplore.ieee.org.

Alhamid, Mohammed F., Majdi Rawashdeh, Haiwei Dong, M. Anwar Hossain, and Abdulmotaleb El Saddik. 2016. "Exploring Latent Preferences for Context-Aware Personalized Recommendation Systems." IEEE Transactions on Human-Machine Systems 46 (4): 615–23.

Ayata, D., Y. Yaslan, and M. E. Kamasak. 2018. "Emotion Based Music Recommendation System Using Wearable Physiological Sensors." IEEE Transactions on. https://ieeexplore.ieee.org/abstract/document/8374807/.

Burke, Robin. 2002. User Modeling and User-Adapted Interaction 12 (4): 331–70.

Forsati, R., M. R. Meybodi, and A. Ghari Neiat. 2009. "Web Page Personalization Based on Weighted Association Rules." In 2009 International Conference on Electronic Computer Technology, 130–35.

Hilbe, Joseph M. 2004. "A Review of SPSS 12.01, Part 2." The American Statistician 58 (2): 168–71.

Jayabal, Ravikumar, Sekar Subramani, Damodharan Dillikannan, Yuvarajan Devarajan, Lakshmanan Thangavelu, Mukilarasan Nedunchezhiyan, Gopal Kaliyaperumal, and Melvin Victor De Poures. 2022. "Multi-Objective Optimization of Performance and Emission Characteristics of a CRDI Diesel Engine Fueled with Sapota Methyl Ester/diesel Blends." Energy. https://doi.org/10.1016/j.energy.2022.123709.

Jayakodi, Santhoshkumar, Rajeshkumar Shanmugam, Bader O. Almutairi, Mikhlid H. Almutairi, Shahid Mahboob, M. R. Kavipriya, Ramesh Gandusekar, Marcello Nicoletti, and Marimuthu Govindarajan. 2022. "Azadirachta Indica-Wrapped Copper Oxide Nanoparticles as a Novel Functional Material in Cardiomyocyte Cells: An Ecotoxicity Assessment on the Embryonic Development of Danio Rerio." Environmental Research 212 (Pt A): 113153.

Jiang, Liangxiao, Harry Zhang, and Zhihua Cai. 2009. "A Novel Bayes Model: Hidden Naive Bayes." IEEE Transactions on Knowledge and Data Engineering 21 (10): 1361–71.

Karimi, Rasoul, Martin Wistuba, Alexandros Nanopoulos, and Lars Schmidt-Thieme. 2013. "Factorized Decision Trees for Active Learning in Recommender Systems." In 2013 IEEE 25th International Conference on Tools with Artificial Intelligence, 404–11.

Klašnja-Milićević, A., M. Ivanović, B. Vesin, and Z. Budimac. 2018. "Enhancing E-Learning Systems with Personalized Recommendation Based on Collaborative Tagging Techniques." Applied Intelligence. https://link.springer.com/article/10.1007/s10489-017-1051-8.

Kothari, Aansi A., and Warish D. Patel. 2015. "A Novel Approach Towards Context Based Recommendations Using Support Vector Machine Methodology." Procedia Computer Science 57 (January): 1171–78.

Kotteeswaran, C., Indrajit Patra, Regonda Nagaraju, D. Sungeetha, Bapayya Naidu Kommula, Yousef Methkal Abd Algani, S. Murugavalli, and B. Kiran Bala. 2022. "Autonomous Detection of Malevolent Nodes Using Secure Heterogeneous Cluster Protocol." Computers and Electrical Engineering. https://doi.org/10.1016/j.compeleceng.2022.107902.

Krishnan, Anbarasu, Duraisami Dhamodharan, Thanigaivel Sundaram, Vickram Sundaram, and Hun-Soo Byun. 2022. "Computational Discovery of Novel Human LMTK3 Inhibitors by High Throughput Virtual Screening Using NCI Database." Korean Journal of Chemical Engineering. https://doi.org/10.1007/s11814-022-1120-5.

Li, Bo, Yibin Liao, and Zheng Qin. 2014. "Precomputed Clustering for Movie Recommendation System in Real Time." Journal of Applied Mathematics. https://doi.org/10.1155/2014/742341.

Mohan, Harshavardhan, Sethumathavan Vadivel, Se-Won Lee, Jeong-Muk Lim, Nanh Lovanh, Yool-Jin Park, Taeho Shin, Kamala-Kannan Seralathan, and Byung-Taek Oh. 2022. "Improved Visible-Light-Driven Photocatalytic Removal of Bisphenol A Using V2O5/WO3 Decorated over Zeolite: Degradation Mechanism and Toxicity." Environmental Research. https://doi.org/10.1016/j.envres.2022.113136.

Mohan, Kannan, Abirami Ramu Ganesan, P. N. Ezhilarasi, Kiran Kumar Kondamareddy, Durairaj Karthick Rajan, Palanivel Sathishkumar, Jayakumar Rajarajeswaran, and Lorenza Conterno. 2022. "Green and Eco-Friendly Approaches for the Extraction of Chitin and Chitosan: A Review." Carbohydrate Polymers 287 (July): 119349.

Persson, Karl. 2015. "Predicting Movie Ratings : A Comparative Study on Random Forests and Support Vector Machines." diva-portal.org. https://www.diva-portal.org/smash/record.jsf?pid=diva2:821533.

Sanner, M. F. 1999. "Python: A Programming Language for Software Integration and Development." Journal of Molecular Graphics & Modelling 17 (1): 57–61.

Saravanan, A., P. Senthil Kumar, B. Ramesh, and S. Srinivasan. 2022. "Removal of Toxic Heavy Metals Using Genetically Engineered Microbes: Molecular Tools, Risk Assessment and Management Strategies." Chemosphere 298 (July): 134341.

Sathish, T., R. Saravanan, V. Vijayan, and S. Dinesh Kumar. 2022. "Investigations on Influences of MWCNT Composite Membranes in Oil Refineries Waste Water Treatment with Taguchi Route." Chemosphere 298 (July): 134265.

Shah, Mit, Dhruvesh Parikh, and Bharat Deshpande. 2016. "Movie Recommendation System Employing Latent Graph Features in Extremely Randomized Trees." In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, 1–4. ICTCS '16 42. New York, NY, USA: Association for Computing Machinery.

Theocharous, Georgios, Philip S. Thomas, and Mohammad Ghavamzadeh. 2015. "Personalized Ad Recommendation Systems for Life-Time Value Optimization with Guarantees." In Twenty-Fourth International Joint Conference on Artificial Intelligence. https://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/viewPaper/11480.

Thomas, Leyo Babu, and V. Vaidhehi. 2018. "The Design of Web Based Car Recommendation System Using Hybrid Recommender Algorithm." International Journal of Engineering & Technology. https://doi.org/10.14419/ijet.v7i3.4.16772.

Vivek, J., T. Maridurai, K. Anton Savio Lewise, R. Pandiyarajan, and K. Chandrasekaran. 2022. "Recast Layer Thickness and Residual Stress Analysis for EDD AA8011/h-BN/B4C Composites Using Cryogenically Treated SiC and CFRP Powder-Added Kerosene." Arabian Journal for Science and Engineering. https://doi.org/10.1007/s13369-022-06636-5.

Wei, Jian, Jianhua He, Kai Chen, Yi Zhou, and Zuoyin Tang. 2017. "Collaborative Filtering and Deep Learning Based Recommendation System for Cold Start Items." Expert Systems with Applications 69 (March): 29–39.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Algal Biofuels: Technological Perspective on Cultivation, Fuel Extraction and Engineering Genetic Pathway for Enhancing Productivity." Fuel. https://doi.org/10.1016/j.fuel.2022.123814.

Yaashikaa, P. R., P. Senthil Kumar, and S. Karishma. 2022. "Review on Biopolymers and Composites – Evolving Material as Adsorbents in Removal of Environmental Pollutants." Environmental Research. https://doi.org/10.1016/j.envres.2022.113114.

Zhang, Heng-Ru, and Fan Min. 2016. "Three-Way Recommender Systems Based on Random Forests." Knowledge-Based Systems 91 (January): 275–86.

Zhang, Wei, and Feng Gao. 2011. "An Improvement to Naive Bayes for Text Classification." Procedia Engineering 15 (January): 2160–64.

**Tables and Figures**

Table 1: The evaluation metrics of the RF classifier with the NB classifier has been calculated. The RF classifier has a 94.30 accuracy rate, whereas the NB classifier has 82.56, respectively. In all parameters, the RF classifier outperforms the NB in the classification of movie recommendation, with a higher accuracy rate.

| SI.No. | Test Size | ACCURACY RATE | |
|---|---|---|---|
| | | RF | NB |
| 1 | Test1 | 90.23 | 80.10 |
| 2 | Test2 | 90.54 | 80.23 |

| | | | |
|---|---|---|---|
| 3 | Test3 | 91.36 | 80.19 |
| 4 | Test4 | 92.34 | 80.92 |
| 5 | Test5 | 92.12 | 81.92 |
| 6 | Test6 | 93.56 | 81.01 |
| 7 | Test7 | 94.30 | 81.85 |
| 8 | Test8 | 94.36 | 82.28 |
| 9 | Test9 | 94.45 | 82.58 |
| 10 | Test10 | 94.54 | 82.34 |
| Average Test Results | | 94.30 | 82.56 |

Table 2: The statistical calculation such as mean, standard deviation and standard error mean for RF and NB algorithm. Accuracy rate parameter used in the t-test. The mean accuracy rate of RF is 94.30% and the NB algorithm is 86.956%. The Standard Deviation of RF is 0.69113 and the NB algorithm is 1.67839. The Standard Error mean of RF is 0.32048 and NB algorithm is 1.78293.

| **Group** | | **N** | **Mean** | **Standard Deviation** | **Standard Error Mean** |
|---|---|---|---|---|---|
| **Accuracy Rate** | NB | 10 | 82.56 | 1.67839 | 1.78293 |

| | | | | | |
|---|---|---|---|---|---|
| | RF | 10 | 94.30 | 0.69113 | 0.32048 |

Table 3: The statistical calculations for independent samples test between RF and NB algorithm. The significance value for accuracy rate is 0.037. Independent samples T-test is applied for comparison of RF and NB algorithm with the confidence interval as 95% and level of significance as 0.38812. This independent sample test consists of significance as 0.001, significance (2-tailed), mean difference, standard error difference, and lower and upper interval difference.

| **Group** | | **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 (Lower)** | **95% Confidence Interval (Upper)** |
| **Accuracy Rate** | **Equal variances assumed** | 7.784 | 0.037 | 13.723 | 13 | .001 | 13.07012 | 0.38812 | 13.02342 | 13.19117 |
| | **Equal variances not assumed** | | | 12.467 | 12.391 | .001 | 12.21133 | 0.39180 | 12.31911 | 12.13022 |

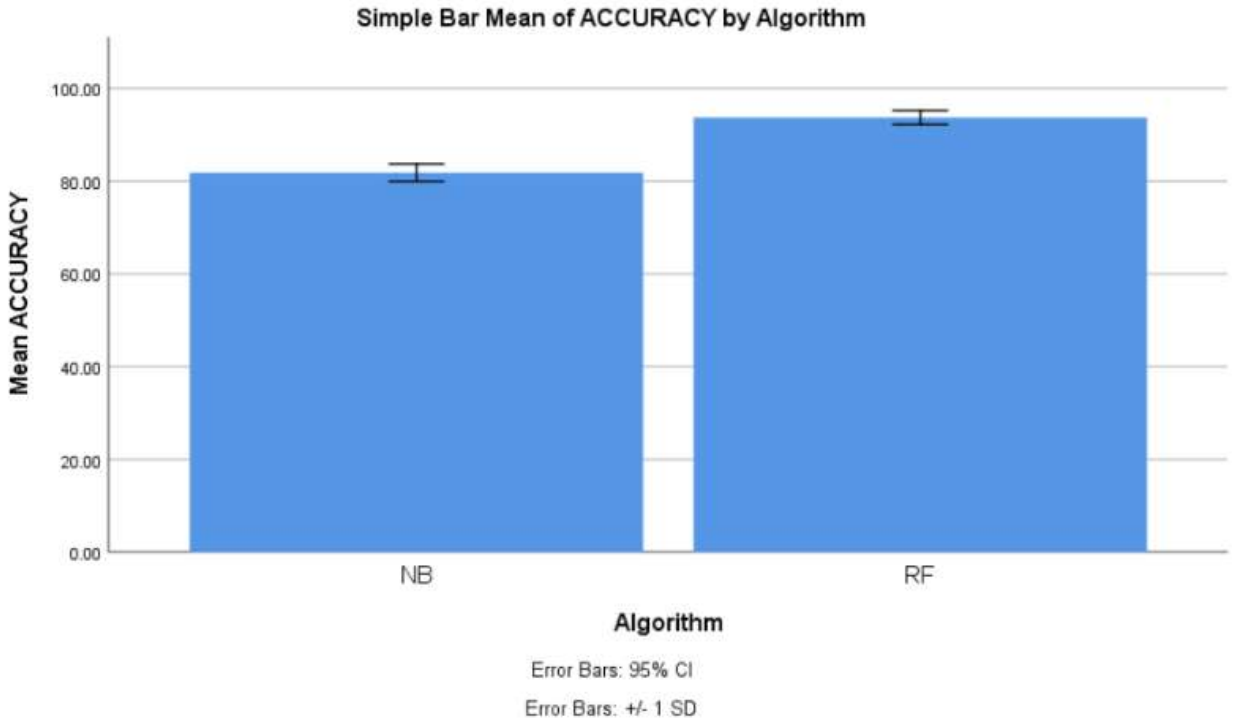


Fig 1: Simple Bar graph for RF classifier accuracy rate is compared with NB model. The RF classifier is higher in terms of accuracy rate 94.30 when compared with NB model 82.56. There is a significant difference between RF classifier and NB model ($p<0.05$ Independent sample test). X-axis: NB model accuracy rate vs RF classifier Y-axis: Mean of accuracy rate, for identification of keywords ± 1 SD with 95 % CI.