2022 57th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST) | 978-1-6654-8500-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICEST55168.2022.9828583

Performance Comparison of Machine Learning Algorithms in Movie Recommender Systems

Nora PireciSejdiu[1], Blagoj Ristevski[2] and Ilija Jolevski[3]

Abstract - **We are all aware that the use of technology in every domain of life produces an enormous amount of information by overloading the amount of data on the Internet. To make data access easier, recommendation systems have been shown to be more efficient, especially performance enhancement has been significantly increased with the integration and use of machine learning algorithms. This paper compares the performance of three machine learning algorithms: Naïve Bayes, neural networks and logistic regression when applied on a movie recommender system. The movie recommender system is implemented in Python programming language using the MovieLens dataset.**

Keywords – **Recommender systems, Machine learning techniques, Naïve Bayes, Neural networks, Logistic regression.**

I. INTRODUCTION

The use of information and communication technology (ICT) in every domain of life produces a huge amount of data and this often becomes challenging for users to access the right information. Recommendation systems (RS) which by filtering information, especially in the most overloaded systems, based on the preferences or requests of users often based on their behavior, have shown efficiency and facilitated access to data. Of particular importance in recommendation systems are the use of accurate and efficient techniques in order to provide the most useful recommendation for users of that system [1].

Data mining is used by many researchers and organizations to extract the necessary data related to their requirements as data mining involves many techniques such as naïve Bayes, neural networks, logistic regression, k-nearest neighbors (kNN), decision trees, etc. [2]. To train an algorithm, machine learning uses the data as a training set. They improve the quality of recommendations [3].

This paper compares the confusion matrix, precision and accuracy of machine learning techniques in movie recommendation systems such as Naïve Bayes, neural network and logistic regression. Confusion matrix is a table with the combination of actual and predicted values. It is a performance measurement for the problem of machine learning classification. The precision is the proportion of relevant predictions among the retrieved predictions, while recall is the proportion of relevant predictions that were retrieved. Precision can be seen as a measure of quality, and recall as a measure of quantity; therefore, recall in this paper is not taken into account. Accuracy is the proportion of the total number of correct predictions and the total number of predictions [4].

The rest of the paper is structured as follows. Naïve Bayes classifier, neural network and logistic regression algorithms are described in Section III. The subsequent section depicts the performance analysis and comparison of the obtained results from naïve Bayes, neural network and logistic regression algorithms. Concluding remarks are highlighted in the last section.

II. RELATED WORKS

Due to the increase in Internet speed and the rapid development of ICT, almost any device can now be connected to the Internet. Moreover people use social networks for marketing, e-commerce, business meetings and even online conferences are held, and hence a huge volume of heterogeneous structured, semi-structured and non-structured data are generated.

Recommendation systems have changed and improved the communication way between users and the web pages. Recommendation systems classify large amounts of data and make the information search easier. The most popular areas where recommender systems are applied are books, news, articles, music, videos, movies, etc. [5].

In paper [6] the authors have used a Bayesian methodology that uses all available information including user ratings and features of articles and users in a unified framework. The authors in the paper [7] have used a hybrid approach by combining Bayes classification with collaborative filtering which has shown better performance in terms of accuracy and coverage. Paper [8] compares naïve Bayes, random forest, decision tree, support vector machines, and logistic regression by evaluating the accuracy implemented in Apache Spark. Authors in paper [9] have used kernel logistic regression, radial basis function classifier, multinomial naïve Bayes and logistic model tree to select sensitivity mapping to floods. Paper [10] compares the performance of logistic regression, naïve Bayes and kNN algorithms measuring accuracy, sensitivity, specificity, precision, F-measure and area under the curve (AUC value). In the paper [11] authors compared functional trees (FT), multilayer perceptron neural networks (MLP neural

[1]Nora Pireci Sejdiu is with University of St. Kliment Ohridski, Faculty of Information and Communication Technologies, 1 Maj bb., 7000 Bitola, Republic of Macedonia, E-mail: pireci.nora@uklo.edu.mk

[2]Blagoj Ristevski is with University of St. Kliment Ohridski, Faculty of Information and Communication Technologies, 1 Maj bb., 7000 Bitola, Republic of Macedonia, E-mail: blagoj.ristevski@uklo.edu.mk

[3]Ilija Jolevski is with University of St. Kliment Ohridski, Faculty of Information and Communication Technologies, 1 Maj bb., 7000 Bitola, Republic of Macedonia, E-mail: ilija.jolevski@uklo.edu.mk

978-1-6654-8500-5/22/$31.00 ©2022 IEEE

nets), and naïve Bayes (NB) for landslide susceptibility assessment at the Uttara hand area.

III. Machine Learning Techniques

Machine learning is considered a branch of artificial intelligence as it aims that systems learn from data and make their own decisions without human intervention or minimal human intervention. [12].

In machine learning, there are many types of classifiers, but the most popular techniques are naïve Bayes, decision tree classifier, neural network, k-nearest neighbor, logistic regression, support vector machines, etc.

Naïve Bayes classification is the method of supervised learning based on the Bayes theorem that has a principle where each pair to be classified are independent of each other. This assumption is called the conditional independence of the class which is made in order to simplify the calculations and therefore is called "naive". This algorithm has excellent generalization capabilities, it is simple and has a linear execution time, therefore it is very well known in pattern recognition and text categorization [13] [14].

A neural network is a special method in artificial intelligence inspired by the human brain that functions through interconnected neurons in layered structures consisting of an input layer, a hidden layer and an output layer, and each connection has a certain weight associated with it [12]. Artificial neural networks (ANNs) have the ability to work with multidimensional data therefore have applicability in many areas of face recognition, documents summarization, in traffic forecasting, etc. [15].

Logistic regression is a supervised learning classification algorithm similar to linear regression which predicts a binary result. The fact that more than two explanatory variables can be used simultaneously, is easy to implement and very efficient to train, make this algorithm to have a great advantage. The main disadvantage of this algorithm is the selection of which variables to include. The best practice is to use as many variables as possible and place them all in the model [16].

A. Performance analysis of naïve Bayes, neural network and logistic regression

To accomplish this paper, the code is implemented in Python using the MovieLens dataset [14]. The splitting criterion corresponds to 80% of the training set and 20% of the testing set. The performance evaluation of these three algorithms is done by evaluating the accuracy, confusion matrix and precision.

A confusion matrix, in this case, is 6x6 matrix used for evaluating the performance of naïve Bayes, neural network and logistic regression models, where each of the target class is represented by the 0 to 5 star rating, respectively. The confusion matrix compares the actual target values with those predicted by the algorithms. Given the fact that not all users have rated all the movies in the dataset, as a result, we have a lot of empty values, conditionally NaN values have been algorithms, however, the performance was evaluated with the results obtained from algorithms

Fig. 1 shows the results of the confusion matrix for the neural network, whereas Figs. 2 and 3 show the confusion matrix for logistic regression and logistic regression with cross-validation, respectively. As can be noticed from Figs. 1 and 2, the results of the confusion matrix are not satisfactory as the diagonals are not dominated by higher values except in the 4-star rated class prediction, where the error rate is also high. These results are the same for both neural network and logistic regression so we tried the logistic regression algorithm with cross-validation. As shown in Fig. 3, this algorithm has given us different results by slightly improving the performance of the diagonal where the values appear even in the prediction of the 3-star rating class with the maximum value on the diagonal but also the error rate is high. As seen in Table 1, the accuracy and precision for these three algorithms is 0.35% compared to the logistic regression CV which has the precision of 0.35% also for the predicted of the 3-star rating class.

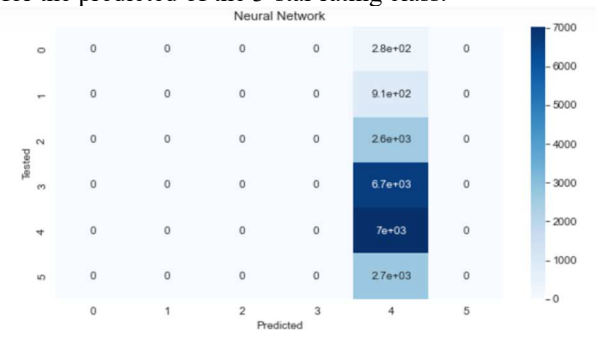

Fig. 1. Confusion matrix results for Neural Network technique

Fig. 2. Confusion matrix results for Logistic regression technique

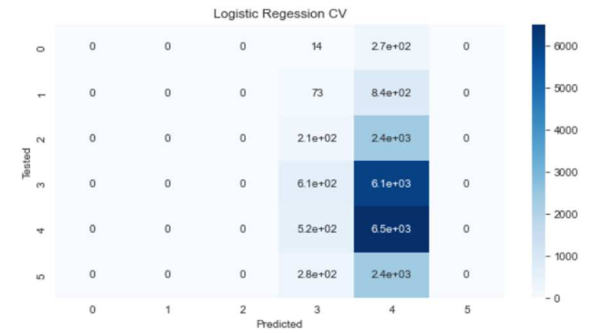

Fig. 3. Confusion matrix results for Logistic Regression CV

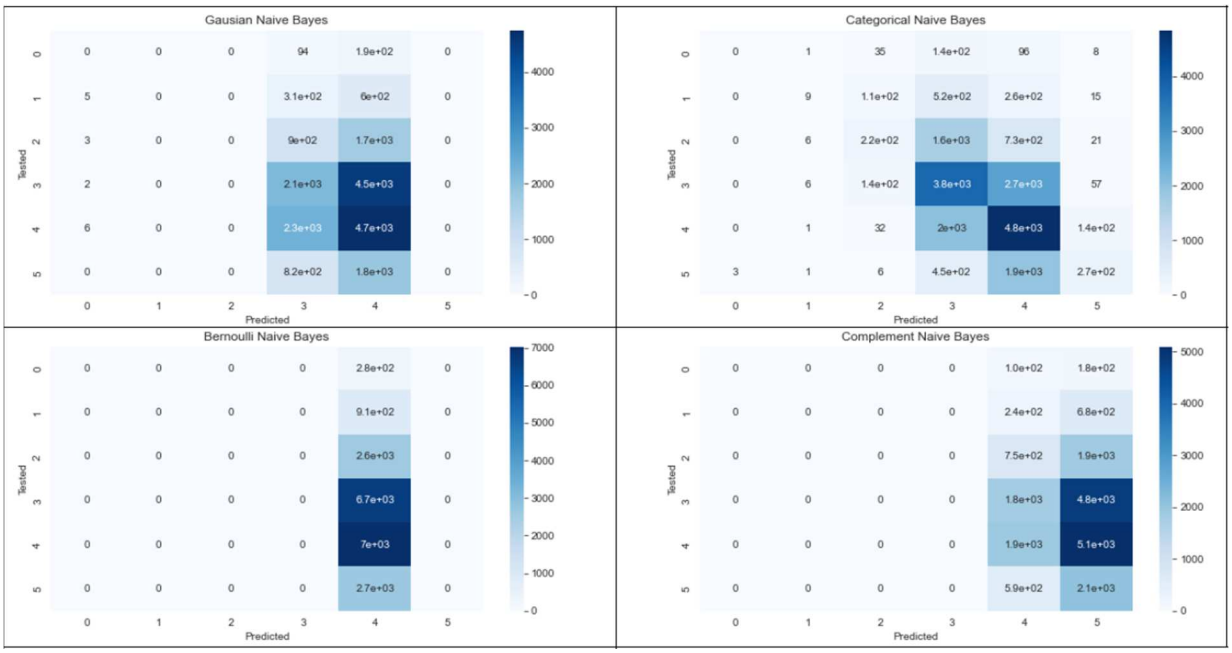

Fig. 4. Confusion matrix results for Gaussian naïve Bayes, categorical Naïve Bayes, Bernoulli naïve Bayes and complement naïve Bayes

One of the algorithms we have compared in this paper is the Naive Bayes algorithm which has several implementations depending on the kind of dataset. We compared Gaussian naïve Bayes, categorical naïve Bayes, Bernoulli naïve Bayes and complement naïve Bayes and as seen in Fig. 4 for the Movielens dataset, the best performance of the confusion matrix showed the categorical naïve Bayes classifier which also shows values for 1-star, 2-star, 3-star, 4-star and 5-star rating classes where in the diagonals the maximum values prevail followed by Gaussian naïve Bayes that shows values

TABLE I
ACCURACY AND PRECISION RESULTS FOR NEURAL NETWORK, LOGISTIC REGRESSION AND LOGISTIC REGRESSION CV

Classifiers	Accuracy	Ratings	Precision
Neural Network	0.35	0	0.00
		1	0.00
		2	0.00
		3	0.00
		4	0.35
		5	0.00
Logistic regression	0.35	0	0.00
		1	0.00
		2	0.00
		3	0.00
		4	0.35
		5	0.00
Logistic regression CV	0.35	0	0.00
		1	0.00
		2	0.00
		3	0.00
		4	0.35
		5	0.35

TABLE II
ACCURACY AND PRECISION RESULTS FOR NAÏVE BAYES CLASSIFIERS

Naïve Bayes Classifier	Accuracy	Ratings	Precision
Gaussian	0.34	0	0.00
		1	0.00
		2	0.00
		3	0.33
		4	0.34
		5	0.00
Categorical	0.45	0	0.00
		1	0.38
		2	0.41
		3	0.44
		4	0.46
		5	0.53
Bernoulli	0.35	0	0.00
		1	0.00
		2	0.00
		3	0.00
		4	0.35
		5	0.00
Complement	0.20	0	0.00
		1	0.00
		2	0.00
		3	0.00
		4	0.36
		5	0.14

only for the 4-star and 5-star rating classes, after this comes Bernoulli naïve Bayes that shows values only for the 4-star rating class dominating maximum values on the diagonal, while complement naïve Bayes has shown the poorest performance where the highest values have dominated outside the diagonal.

When we compare accuracy and precision for these algorithms as shown in Table 2, complement Naive Bayes has an accuracy of 0.45% and the precision value for 0-star rating class is 0, for 1-star rating class is 0.38%, for 2-star rating class is 0.41%, for 3-star rating class is 0.44 %, for 4-star rating class is 0.46% and for 5-star rating class is 0.53%. At the same time, it is the only algorithm that has the value of precision in almost all predicted classes except 0-star rating class. If we make a comparison between the three main algorithms elaborated in this paper, it is clear that complement naïve Bayes has the highest accuracy value compared to neural network and logistic regression CV. Also, the confusion matrix of this algorithm has a distribution of maximum values in almost the entire diagonal except 0-star rating class, which also results in precision which has higher values compared to precision in neural network and logistic regression CV. Besides having a higher value of precision, it is worth noting that in complement naïve Bayes almost all prediction classes have values of precision above 0 compared to neural network where only the 4-star rating class has a precision of 0.35% and in logistic regression CV the 3-star and the 4-star rating classes have 0.35% precision.

IV. Conclusion

Based on the research that has been done so far, it is well known that no algorithm has yet been found that satisfies all the requirements and needs of users. Therefore, in this paper, we have compared some of the machine learning algorithms in the context of movie recommendations. Based on performance evaluation of the selected algorithms for this research (naïve Bayes, neural network and logistic regression), the results have shown that neural network and logistic regression algorithms have almost the same results for the confusion matrix, precision and accuracy, while complement naïve Bayes has shown a better accuracy. The precision comparing to neural network and logistic regression CV is higher and all prediction classes have values above 0 of precision compared to neural network and in logistic regression CV.

It can be concluded that: for this kind of dataset in the context of movie recommendation, complement naïve Bayes has shown better performance of accuracy, precision and confusion matrix compared to neural network and logistic regression with cross-validation. The focus in the future work will be the evaluation of the most popular techniques, such as decision tree and kNN.

Acknowledgement

This work has been partially supported by the COST Action CA19122 – European Network for Gender Balance in Informatics (EUGAIN).

References

[1] F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh, “Recommendation systems: Principles, methods and evaluation”, Egyptian Informatics Journal, Volume 16, Pages 261-273, 2015.

[2] M. Mohanapriya, J. Lekha, “Comparative study between decision tree and knn of data mining classification technique”, Second National Conference on Computational Intelligence, 2018.

[3] J. Leskovec, A. Rajaraman, J. D. Ullman, “Mining of Massive Datasets”, Chapter 9, 2010.

[4] M. Buckland, F. Gey,” The Relationship between Recall and Precision”, Journal of the american society for information science, 45(1):12-19, 1994.

[5] P. Nagarnaik, A.Thomas, “Survey on Recommendation System Methods”, IEEE sponsored 2nd International Conference on Eletronics and Communication System, 2015.

[6] M. Condliff, D. Lewis, D. Madigan and C. Posse, "Bayesian Mixed Effects Models for Recommender Systems", Proc. ACM SIGIR '99 Workshop Recommender Systems: Algorithms and Evaluation, 1999.

[7] M. Ghazanfar, A. Prugel-Bennett “An Improved Switching Hybrid Recommender System Using Naïve Bayes Classifier and Collaborative Filtering”, Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS, Vol. 1, 2010.

[8] T. Pranckevicius, V. Marcinkevicius, “Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification”, Baltic J. Modern Computing, Vol. 5, 2017.

[9] Pham, Binh T., Tran V. Phong, Huu D. Nguyen, Chongchong Qi, "A Comparative Study of Kernel Logistic Regression, Radial Basis Function Classifier, Multinomial Naïve Bayes, and Logistic Model Tree for Flash Flood Susceptibility Mapping", Water 12, no. 1: 239. https://doi.org/10.3390/w12010239, 2020.

[10] F. Itoo, Meenakshi, S. Singh, “Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection”, International Journal of Information Technology volume 13, pages1503–1511, 2021.

[11] B. Thai Pham, D. Tien Bui, H. Reza Pourghasemi, P. Indra, M. B. Dholakia, “Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods”, Theoretical and Applied Climatology volume 128, pages255–273, 2017.

[12] R. Ahuja, A. Solanki, A. Nayyar, “Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor”, 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019.

[13] Charles Elkan, “Boosting and naïve Bayesian learning”. Technical Report No. CS97-557, 1997.

[14] A. Ashari, I. Paryudi, A. Tjoa, “Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool”, International Journal of Advanced Computer Science and Applications, Vol. 4, No. 11, 2013.

[15] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, Y. Wang, “Learning Traffic as Images: A deep convolutional neural network for large-scale transportation network speed prediction”, Special Issue Sensors for Transportation, 2017.

[16] Sperandei, Sandro, “Understanding logistic regression analysis”, Biochemia medica vol. 24,1 12-8. 15, 2014.