Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

An empirical Performance evaluation and comparison of different classifiers on standard datasets

Kamlesh Kumar Raghuvanshi, Subodh Kumar and Arun Agarwal*
Department of Computer Science, Ramanujan College, University of Delhi
kamlesh@ramanujan.du.ac.in, subodhkumar588@gmail.com, arun.agarwal@ramanujan.du.ac.in

** Corresponding Author*

Abstract— Supervised Machine Learning (SML) refers to the mapping of the input variable to output variable using an algorithm. The correctness of learning is the number of correct predictions/classifications after training a model. This study shows the behavior of different supervised classifiers namely distance-based non-parametric algorithm KNN, statistical-based Naïve Bayes classifier, parametric method SVM, and Neural Network on linear and non-linear data. The performance of the algorithms is evaluated for accuracy score parameter on some authentic dataset repository of UCI Machine Learning Repository and compared to find out which algorithm is best suitable for which type of data sets. One of the important steps is also included before analyzing the accuracy score. This is calculated using Standard scalar libraries present within Anaconda software.

Keywords— Classification, K-Nearest Neighbor, Naïve Bayes, parametric, prediction, supervised, SVM

I. INTRODUCTION

Machine learning is an interesting topic that has applicability in almost every area of science, medical, business (making decisions) that help solve social, biological, and industrial problems [1,2]. The two most important parts of machine learning are data and mathematics. It is a critical task of selecting a particular algorithm for a dataset. This study gives a way to show the behavior of different classification algorithms on different data sets [3]. The preprocessing of data is always needed to make the decision more confident as machine learning is Data-Driven AI. Data has everything that can lead us to a new perception of things. In machine learning main importance is given to data, the results can be as good depending on the data. Data can be modeled and stored in a database. Relevant data and patterns can be extracted from operational data stores according to the analyzing purposes. The insights need to be visualized and communicated so that they can help to make decisions more confidently. The applicability of machine learning in almost every area of science and medicine has given a faith to start this study. People often make errors while analyzing or, possibly, when obtaining to establish relations among multiple features [4].

Classification is an important concept in machine learning. It is the method by which a similar group of objects can be combined based on certain criteria (referred to as traits, variables, characters, etc.). Every classification algorithm calculation has its intrinsic biases, and no unique classification model appreciates predominance on the off chance that we don't make any assumption about the given approach. It is accordingly crucial for contrast calculations with train and selects the best performing model. However, before we can analyze various models, we initially need to settle on a measurement to perform execution for the best-performing model. One ordinarily utilized measurement is precision, which is characterized as the extent of accurately classified cases.

$$Accuracy = \frac{TP + TN}{TN + FP + TP + FN} \qquad \text{.............}\textbf{Eq1}$$

The issue of classification can be analyzed as: a dataset X= {*, o} producing a classifier C: X->Y, where each X variable maps to its respective classification label Y. The logic for this classification deals on the nearest label using KNN classifier. The important aspect is the number of voters used to decide the classification group for a particular object, these voters are known as K in this algorithm. If the value for variable K is very small, so the output results may be obtained as sensitive to variations whereas for a greater value of K, the neighborhood might have a variety of points from other classes. The option of the distance measure is also important to consider for results [5]. The distance between two points' p and q can be calculated using the following formula: [6]

$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

…………..**Eq2**

Due to complex calculation with storage requirements for the calculation of this distance for each point this algorithm is said to be non-parametric. The issues with this method are resolved using Naïve Bayes classification that is fully dependent on conditional probability and it contains three algorithms namely Gaussian, Multinomial, and Bernoulli. In these classifiers, the conditional probability for each possible value of all attributes is to be calculated to resolve this issue we have to adopt a discretization technique to discretize numerical attributes into several classes. The algorithm KNN is a statistical classification approach, and is based on pattern recognition. KNN is widely used in machine learning as well as in many other domains sharing accurate and positive results alone as well as in comparison with other algorithms also.

Data correlation and visualization plays a very important role in choosing machine learning algorithm. In the linear data correlation, neural network and linear regression both can fit the data properly but linear regression is less costly and trains faster as compared to the neural network.

Table 1: Correlation matrix feature dataset

Dataset	**battery power**	**blue**	**Clock_ speed**	**Dual_ sim**	**fc**
battery power	1.0000	-0.0466	-0.0390	-0.0611	-0.0078
blue	-0.0466	1.0000	0.0347	-0.0111	-0.0560
Clock_ speed	-0.0390	0.0347	1.0000	-0.0124	0.0101
Dual_ sim	-0.0611	-0.0111	-0.0124	1.0000	0.5760
fc	-0.0078	-0.0560	0.0101	0.5760	1.0000

The study is divided into four sections introduction, dataset description, results, and conclusion. Sometimes when the results are not acceptable, the problem is not with the model, the data is who the culprit is. So the identification of data correlation plays an important role to make a correct choice of the machine learning algorithm. In this paper, the focus is that the data drives the choice of the machine learning algorithm. This shares an important procedure of obtaining the relation among data. The technique linear regression is useless for training any model on dataset having non-linear correlation. Some of the renowned programming software like pandas, NumPy, matplotlib, sklearn plays an important role for data analysis and is easy to use with any model for getting better results.

In general, assumption of a model with machine learning features having function f: X->Y, mapping from x € X → y € Y, where variable X defines the input space while Y defines the output space of the given function. The situation where f (x->y) is present, the machine is said to be as trained sharing an estimated output value for a specific pattern having variable x. To analyze the positivity of the learning mechanism of the trained model, there is a need to maintain some of the possible parameters namely errors, accuracy score, and losses. In supervised learning, the often-used criteria are the minimization criteria which involve potential loss into a classification decision made. The classification based on machine learning requires accuracy of the parameters as well as the size of the variables for the residing dataset. For any accurate algorithm to work the model developed should be precise and correct with classification procedure. [7].

A. Naïve Bayes classifier

Naïve Bayes classifiers are from the family of “probabilistic classifiers” with independent assumptions between the features. The features are not related to each other and also this is a text classification featured algorithm. Naïve-Bayes algorithm consists of simplicity but can outperform more sophisticated classification algorithms [8]. These classifiers perform well only with categorical attributes. Continuous attributes should be binned (input is divided into a specified range of equally-sized intervals) and transformed into categorical variables before the naïve Bayes classifiers are applied to process them [9]. Three types of naïve Bayes classifiers are used for the classification of the dataset namely multinomial, Gaussian and Bernoulli. The applicability of one over the other on a dataset is tested.it is a non-parametric method that uses Bayes theorem as the model and estimates the priors P (A) and likelihoods P (X|A) for an unseen sample X from the dataset [10].

B. Support Vector Machine

SVM is a directed learning technique utilized for grouping, regression, and exception detection. It is a flexible procedure with various methods working with different kernels to obtain decision capability and outcomes [11]. The essential hypothesis of a SVM is to plan the information onto a higher layered element space nonlinearly connected with the information space and decide an isolating hyperplane with the most extreme edge between the two classes in the component space [12]. SVM is one of the compelling calculations that perform characterization by building a N-layered hyperplane that ideally isolates the information into two classes [13]. SVM can be considered as the binary classifiers [14] although applicable for multi-class classification also. SVM outperforms many datasets but naïve Bayes gave competition to it. In SK-learn package has SVC(C-Support Vector Classification). Svc shares the “one-against-one” method for multi-class classification. If the variable n defines total classes, then n*(n-1)/2 classifiers are obtained and every data is being trained form classes present within a model. To provide an interface with other classifiers, the classifiers decide to map (n sample, n classes) [15]. The kernel decision is also important, three kernels are there linear, sigmoidal, RBF.

II. METHODOLOGY

The proposed methodology is simple yet useful, the datasets taken are of different types. Before training the model. The dataset is divided into the train, split and test datasets (x_train, y_train, x_test, y_test). The percentage of data taken usually for testing is 30%. Then the model is trained on 70% of the data. The accuracy score is calculated. The datasets from the UCI repository [16] taken contains continuous, categorical, and discrete values. The simplest relationship is linear [17].

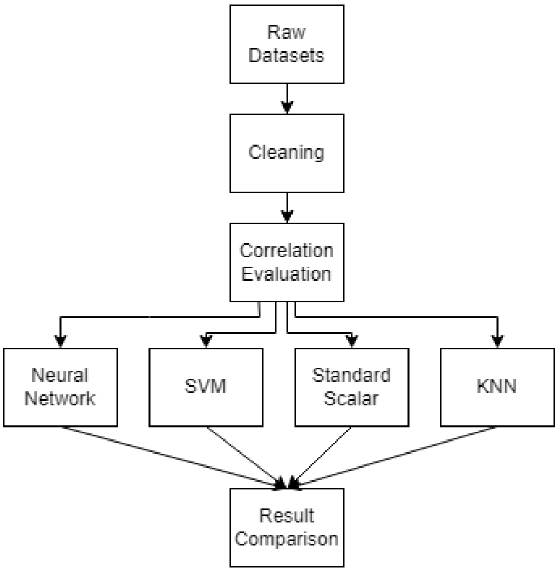


Fig 1: Methodology of the proposed work

In this study, we have taken an iris data set from a machine learning repository that is in categorical form and applied a machine-learning algorithm to it. After applying KNN on the IRIS dataset the accuracy score achieved is 93.33 percent but when the same dataset is applied on the neural network it gives an accuracy score for this classification dataset as (13.33 to 93.33]. But due to the simplicity and less number of features in data, KNN can be used as compared to Neural Network (multilayer perceptron classifier) [18, 19]. The accuracy score of 93.33 is achieved with 250 iterations that are not recommendable. When Naïve Bayes (multinomial model) is applied to this dataset the accuracy score achieved is 97.77%. The situation where properties are discrete and also more than one classes reside in the dataset. Multinomial gives best results and Bernoulli worst. In the iris dataset, three categories are there in which features are to be classified.

On a dataset like mobile price classification, the number of samples is nine times the number of samples in the iris dataset. KNN performance achieved is 23.3 to 39.9 after preprocessing is performed using standard scalar. The memory requirement is more because of the non-parametric algorithm so KNN is less preferred on the dataset with a large number of samples.

Another dataset named body mass index in Table 2 contains data of males and females with height and weights having 5 classes (weak, obesity, extreme obesity, and overweight, normal) [20, 21]. Since this is a multi-class classification data is continuous and Gaussian outperforms other Naïve Bayes classifiers (Bernoulli with 42.3% and multinomial with 76.9%). consequently when the properties are in a flow continuously the most preferred model is Gaussian for features and the least favorable model is KNN whereas SVM seems to have more positive outcomes in many complex situations. [22, 23].

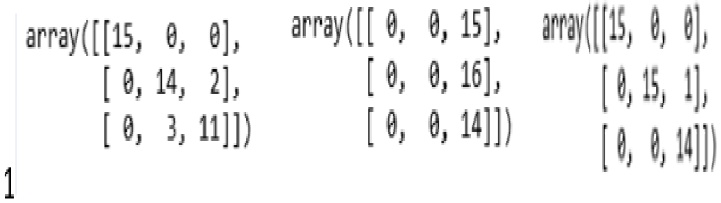
In wine data classification, based on certain parameters the quality of wine is to be decided, before applying the Naïve Bayes classifier the scaling of data using standard scalar is performed. Then the performance by KNN reaches 92.33% whereas neural network showed 35.66% only.

Table 2: Datasets used in the study

Dataset	No. of Classes	No. of Attributes/ sample	Data Relationship	Missing Value
Iris Dataset	3	4 attribute 1 class 150 sample	Non Linear Categorical	No
Petrol Consumption	3	4	Non Linear Numerical	No
Mobil Price Classification	2	20	Linear Categorical Binary	No
Wine data Classification	3	13 177 samples	Non Linear Categorical	No
Fruit Classification	2	2	Linear	No
Pollution	3	4 729 samples	Non Linear	Yes
Body Mass Index	5	3 500 samples	Non Linear Continuous	No

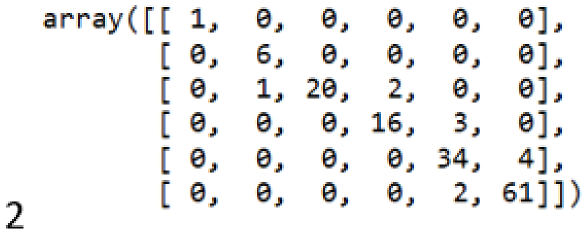
III. CONFUSION MATRIX ON DIFFERENT DATASETS

Confusion matrix on iris dataset after applying (i) Gaussian (ii) Bernoulli (iii) Multinomial shown by index 1 (figure 2) depicts that there is only one misclassified data and hence shows the best performance. The index given in figure 3 shows on the BMI dataset. The variation of correlation on the Neural Network model and KNN model respectively on the dataset is given in figure 4. Figure 5 shows the index of the confusion matrix on dataset pollution. Confusion matrix on dataset mobile classification on multinomial KNN and Bernoulli model respectively (best and worst) performance indicated in figure 6 as shown in the following study.



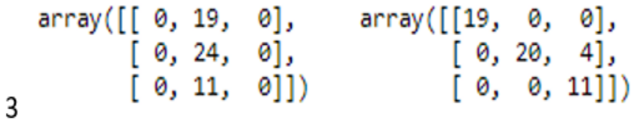
```
array([[15,  0,  0],      array([[ 0,  0, 15],    array([[15,  0,  0],
       [ 0, 14,  2],             [ 0,  0, 16],           [ 0, 15,  1],
       [ 0,  3, 11]])            [ 0,  0, 14]])          [ 0,  0, 14]])
1
```

Fig 2: Confusion Matrix: IRIS Dataset



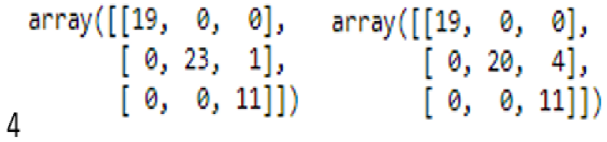
```
array([[ 1,  0,  0,  0,  0,  0],
       [ 0,  6,  0,  0,  0,  0],
       [ 0,  1, 20,  2,  0,  0],
       [ 0,  0,  0, 16,  3,  0],
       [ 0,  0,  0,  0, 34,  4],
2      [ 0,  0,  0,  0,  2, 61]])
```

Fig 3: Confusion Matrix: BMI Dataset



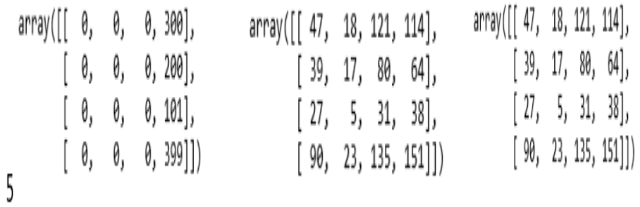
```
array([[ 0, 19,  0],      array([[19,  0,  0],
       [ 0, 24,  0],             [ 0, 20,  4],
3      [ 0, 11,  0]])            [ 0,  0, 11]])
```

Fig 4: Confusion Matrix: KNN Model Dataset



```
array([[19,  0,  0],      array([[19,  0,  0],
       [ 0, 23,  1],             [ 0, 20,  4],
4      [ 0,  0, 11]])            [ 0,  0, 11]])
```

Fig 5: Confusion Matrix: Pollution Dataset



```
array([[  0,   0,   0, 300],    array([[ 47,  18, 121, 114],    array([[ 47,  18, 121, 114],
       [  0,   0,   0, 200],           [ 39,  17,  80,  64],           [ 39,  17,  80,  64],
       [  0,   0,   0, 101],           [ 27,   5,  31,  38],           [ 27,   5,  31,  38],
       [  0,   0,   0, 399]])          [ 90,  23, 135, 151]])          [ 90,  23, 135, 151]])
5
```

Fig 6: Confusion Matrix: Mobile Classification Dataset

IV. PERFORMANCE MEASURE

When the data is in numeric and target is also in numeric (dataset in Table no.2) form then the task associated with it is prediction not classification. Classification technique is applied when targets are in categorical form.

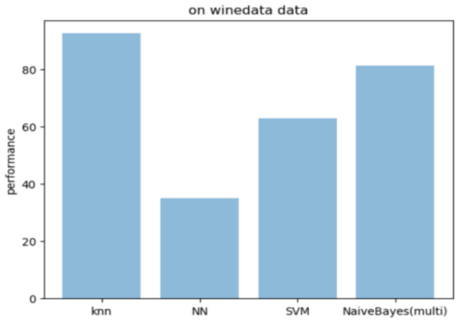


Fig 7: Performance Results: Wine Data

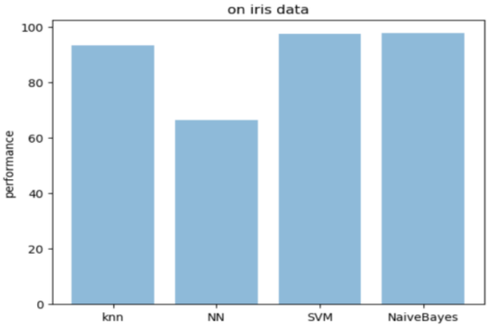


Fig 8: Performance Results: Iris Data

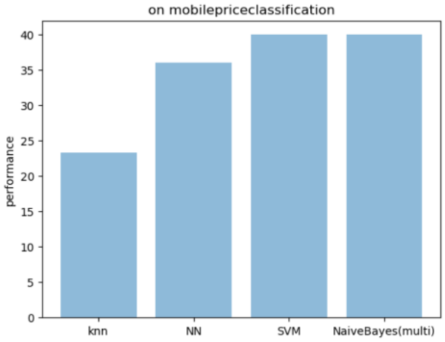


Fig 9: Performance Analysis: Mobile Price Data

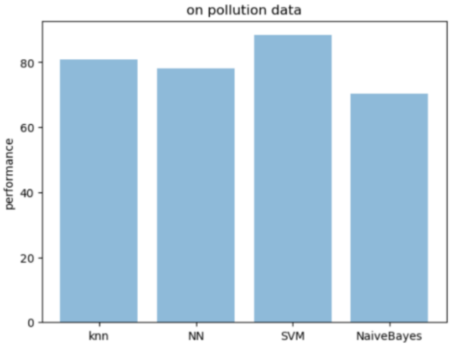


Fig 10: Performance Analysis: Pollution Data

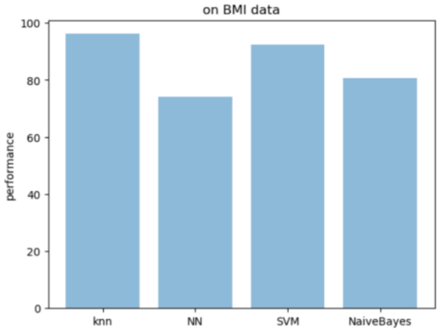


Fig 11: Performance Results: BMI Data

The performance parameter in this case used is root mean square (RMS) and for the performance evaluation of the classification task by a model accuracy score. Task association relation is an important aspect to be identified before applying the algorithm.

V. CONCLUSION

The above study represents the results after applying popular supervised machine learning algorithms on datasets. The applicability of one in a situation is also shown. Naive Bayes outperforms in most of the cases. When the features are continuous Gaussian distribution is to be applied to the datasets for better classification, when the features are discrete and are to be classified in more than two classes, a Multinomial distribution algorithm is used whereas when the features are discrete and have exactly two classes (0/1) Bernoulli gives best results. Though KNN gives good results due to non-parametric algorithm and more calculation on large datasets causes classification to be slow, not appreciated.

Other classification algorithms like Random Tree, hyper pipe, J48, K stark means can be applied and checked on these datasets to find out which is better than these algorithms shown above in the study. Other matrices for deciding the better model after training can also be applied.

REFERENCES

1. Etaiwi, W., Biltawi, M., & Naymat, G. (2017). Evaluation of classification algorithms for banking customer’s behavior under Apache Spark Data Processing System. *Procedia computer science*, *113*, 559-564.
2. Wang, C. C., & Chang, C. D. (2010, July). SVD and SVM based approach for congestive heart failure detection from ECG signal. In *The 40th International Conference on Computers & Indutrial Engineering* (pp. 1-5). IEEE.
3. Chen, S. L., Tuan, M. C., Chi, T. K., & Lin, T. L. (2015). VLSI architecture of lossless ECG compression design based on fuzzy decision and optimisation method for wearable devices. *Electronics Letters*, *51*(18), 1409-1411.
4. Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.
5. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda and Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.
6. Gopal, M. (2019). *Applied machine learning*. McGraw-Hill Education.
7. Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138.
8. Tjoa, A. M., Paryudi, I., & Ashari, A. (2013). Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. *Journal of IJACSA, IJACSA (International Journal of Advanced Computer Science and Applications)*, *4*(11).
9. Al-Qahtani, M. F., Almansour, R., Alharbi, A., Aljasser, M., & Alsunaid, H. (2013). Employer perceptions of workforce preparation of the graduates of the health information management and technology program. *Journal of American Science*, *9*(12).
10. Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.
11. https://scikit-learn.org/stable/modules/svm.html
12. Ferdousy, E. Z., Islam, M. M., & Matin, M. A. (2013). Combination of naive bayes classifier and K-Nearest Neighbor (cNK) in the classification based predictive models. *Computer and information science*, *6*(3), 48.
13. Suykens, J. A., & Vandewalle, J. (1999, July). Multiclass least squares support vector machines. In *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)* (Vol. 2, pp. 900-903). IEEE.

14. Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145-1159.
15. Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), 1-27.
16. Blake, C. L., & Merz, C. J. (1998). UCI repository of machine learning databases, 1998.
17. Huang, J., Lu, J., & Ling, C. X. (2003, November). Comparing naive Bayes, decision trees, and SVM with AUC and accuracy. In *Third IEEE International Conference on Data Mining* (pp. 553-556). IEEE.
18. Fayyad, U., & Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning.
19. Ashari, A., Paryudi, I., & Tjoa, A. M. (2013). 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 (IJACSA)*, *4*(11).
20. Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., ... & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, *97*(1), 262-267.
21. Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining, addison wesley publishers.
22. Nematzadeh, B. Z. (2012). Comparison of Decision Tree and Naive Bayes Methods in Classification of Researcher's Cognitive Styles In Academic Environment.
23. Entezari-Maleki, R., Rezaei, A., & Minaei-Bidgoli, B. (2009). Comparison of classification methods based on the type of attributes and sample size. *J. Convergence Inf. Technol.*, *4*(3), 94-102.
24. Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137-142). Springer, Berlin, Heidelberg.
25. Gandhi, V. C., & Prajapati, J. A. (2012). Review on comparison between text classification algorithms. *Int. J. Emerg. Trends Technol. Comput. Sci*, *1*(3), 1-4.
26. Langley, P., Iba, W., & Thompson, K. (1992, July). An analysis of Bayesian classifiers. In *Aaai* (Vol. 90, pp. 223-228).
27. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.
28. Ding, C., & He, X. (2004, July). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning* (p. 29).