An Empirical Comparison of Some Multivariate Classifiers

Vivian U. Emeli[1], and Mbanefo S. Madukaife[2]

[1,2]Department of Statistics, University of Nigeria, Nsukka, Enugu State, Nigeria
Corresponding author's email: mbanefo.madukaife@unn.edu.ng

Abstract

This work evaluates and compares the performance of different parametric and non-parametric classifiers for the classification of datasets obtained from the multivariate normal, multivariate *t* and multivariate exponential distributions. The training datasets were generated using sample sizes of n = 10, 20, 50, 100 and 1000, with the number of variables set at p = 2, 3, and 4 for each sample size. Additionally, the Mahalanobis squared distance between mean vectors were set at 5 and 10, while maintaining equal covariance matrices across groups. In each case of these variations of datasets outlined above, a separate sample of 100 observations was generated for determining the classifiers' performances. The performance measure used throughout was the misclassification rate of classifiers and the classifiers considered were the linear discriminant function (LDF), logistic regression, naïve Bayes, support vector machines (SVM), *k*-nearest neighbour (*k*NN) and decision trees. The results obtained showed that the performance of the classifiers depends on the underlying distribution of the dataset as well as on the number of variables, p.

Keywords: Empirical comparison, Mahalanobis squared distance, Misclassification rate, Parametric and nonparametric classifier, Training dataset.

1. Introduction

As datasets grow in complexity and dimensionality, traditional statistical approaches may face limitations in capturing the intricate interplay among multiple variables as most traditional statistical methods often focus on analyzing data with a limited number of variables or dimensions, making them suitable for simpler datasets. However, many real-world problems involve complex systems with multiple interrelated variables, necessitating more advanced multivariate analysis techniques. They offer a comprehensive toolkit for exploring relationships and patterns within datasets characterized by multiple interdependent variables. They extend the capabilities of statistical analysis by simultaneously considering multiple variables within a dataset. Rather than examining each variable in isolation, multivariate analysis seeks to explore the interactions and dependencies among them. Rencher (2002) notes that the exclusive use of univariate procedures with multivariate data is no longer excusable, given the availability of multivariate techniques and inexpensive computing power to carry them out. According to Neomí et al. (2015), multivariate analysis, in a broad sense, is the set of statistical methods aimed to simultaneously analyze datasets. That is, for each individual or object being studied, several variables are analyzed. The essence of multivariate thinking is to expose the inherent structure and meaning revealed within these sets if variables through application and interpretation of various statistical methods.

One of the many multivariate statistical tools employed in model building problems is classification, which involves the practice of allocating an observed random object into one of the identified groups where it is expected to have come from. It is a supervised machine learning and statistical method where the model tries to predict the correct group of a given input data. According to Gareth et al. (2017), predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category or class, a process that is known as classification. It is a method that is used to group data based on predetermined characteristics. It is utilized to classify the item as indicated by the features for the predefined set of classes. The main significance of classification is to group data from large datasets to find patterns out of it (Nurshahirah et al., 2019). Specifically, the problem of classification arises when a researcher wants to classify an observation into one of several categories based on certain features. Unlike in regression where the response variable is quantitative, classification is used for prediction processes where the response variable is qualitative or categorical. In other words, predicting a qualitative or categorical response for an observation can be referred to as classification, which is, classifying or assigning the observation to a category.

As the complexity and scale of data continue to grow, the need for an efficient classification algorithm becomes increasingly vital and the selection of an appropriate classifier is crucial for the success of various applications and predictive modelling. In fact, there are a good number of classifiers in the literature, developed from both parametric and non-parametric standpoints. Yugal and Sahoo (2012) state that parametric classifiers are based on the statistical probability distribution of each class while non-parametric classifiers are used in case of unknown density function and used to estimate the probability density function. A good number of researchers have used these classifiers to solve real-life problems. They include Gambo and Yusuf (2010), Kim et al. (2011), Lakshmi (2013), Nurit and Avi (2015), Yakubu and Ibrahim (2016), Akinmoladun et al. (2017), Lilima et al. (2017), Abubakar (2020) and Adenaike et al. (2022), to mention but a few. This study revolves around the need to discern and compare the performance of multivariate classifiers in various real-world scenario as the optimal selection of ideal classifiers stands as a critical challenge particularly in the context of parametric and non-parametric multivariate classifiers. Through empirical analysis and consideration of misclassification rate, this study seeks to provide practitioners and researchers with actionable insights to guide the selection and implementation of multivariate classifiers in their respective fields.

By systematically analyzing the performance of different classifiers, this study aims to provide valuable insights that can inform decision-making processes, enhance predictive accuracy, and facilitate algorithm selection in the quest for more effective classification techniques. The findings will equip researchers with the knowledge of the most effective classifiers for their specific needs, promoting efficiency and accuracy in various applications and fostering a better understanding of multivariate classifiers.

2. Methods

Empirical comparison of the six different classifiers carried out in this work was based on Monte Carlo simulation. The datasets were simulated to ensure they satisfy all the conditions and assumptions of the multivariate normal, multivariate *t* and multivariate exponential distributions as these conditions are essential for an accurate classification. The training datasets were generated using different sample sizes of *n* = 10, 20, 50, 100, 1000. Throughout the study, the datasets consist of observations from two groups (coded 0 and 1) and the number of variables were *p* = 2, 3, 4. Also, the distance between the mean vectors were set at 5 and 10 for all the sample sizes and number of variables considered. Each dataset generated at a sample size, number of variables and distance considered was then used to build six different classifiers; four parametric and two non-parametric (linear discriminant function (LDF), logistic regression, naïve Bayes, support vector machines (SVM) for the parametric and *k*-Nearest Neighbours (*k*NN) and decision trees for the non-parametric) using R statistical software. After building the classifiers to reflect the *n*, *p* and distance conditions as described above, a new dataset of size *n* = 100 (test data) was generated from the distributions under consideration, to determine the accuracy of the classifiers built. The six different classifiers were then used to classify the test observations into one of two classes and the misclassification rates (in probability) of all the six classifiers were computed.

3. Results

The computed misclassification rates are presented in Tables 1 to 9 according to the underlying distributions and the number of variables, *p*.

Table 1: Misclassification rates of the test dataset at different sample sizes under bivariate normal distribution, *p* =2.

		PARAMETRIC				NON-PARAMETRIC	
Distance	Sample Size	LDF	Logistic	Naïve Bayes	SVM	*k*NN	Decision Trees
5	10	0	0.3	0.16	0.02	0.08	0.5
	20	0	0	0.24	0	0.07	0.29
	50	0.01	0.01	0.07	0.01	0.06	0.11
	100	0	0	0.08	0.01	0	0.11
	1000	0.01	0.01	0.1	0.01	0.01	0.06
10	10	0	0	0.13	0	0	0.5
	20	0	0	0.06	0	0	0.16
	50	0	0	0	0	0	0.16
	100	0	0	0.01	0	0	0.03
	1000	0	0	0	0	0	0

Table 2: Misclassification rates of the test dataset at different sample sizes under trivariate normal distribution, p =3.

		PARAMETRIC			NON-PARAMETRIC		
Distance	Sample Size	LDF	Logistic	Naïve Bayes	SVM	*k*NN	Decision Trees
5	10	0.13	0.19	0.36	0.23	0.19	0.5
	20	0	0	0.19	0	0.12	0.29
	50	0	0	0.11	0	0.03	0.15
	100	0	0	0.17	0	0.03	0.15
	1000	0	0.01	0.05	0	0	0.02
10	10	0.01	0.01	0.1	0	0	0.5
	20	0	0	0.08	0	0	0.2
	50	0	0	0	0	0	0.16
	100	0	0	0.01	0	0	0.09
	1000	0	0	0.01	0	0	0.02

Table 3: Misclassification rates of the test dataset at different sample sizes under tetravariate normal distribution, p =4.

		PARAMETRIC			NON-PARAMETRIC		
Distance	Sample Size	LDF	Logistic	Naïve Bayes	SVM	kNN	Decision Trees
5	10	0.01	0.04	0.24	0.03	0.06	0.5
	20	0.01	0.05	0.15	0.01	0.03	0.35
	50	0.01	0.01	0.01	0.01	0.01	0.17
	100	0	0.02	0.01	0	0.01	0.13
	1000	0.01	0.01	0.01	0.01	0.01	0.13
10	10	0	0	0.06	0	0	0.5
	20	0	0	0.03	0	0	0.03
	50	0	0	0	0	0	0.03
	100	0	0	0	0	0	0
	1000	0	0	0	0	0	0.01

Table 4: Misclassification rates of the test dataset at different sample sizes under bivariate *t*- distribution, p =2.

		PARAMETRIC			NON-PARAMETRIC		
Distance	Sample Size	LDF	Logistic	Naïve Bayes	SVM	*k*NN	Decision Trees
5	10	0.21	0.19	0.46	0.2	0.37	0.5
	20	0.15	0.15	0.44	0.48	0.21	0.47
	50	0.11	0.11	0.46	0.27	0.15	0.32
	100	0.13	0.13	0.49	0.5	0.14	0.26
	1000	0.12	0.08	0.49	0.46	0.09	0.15
10	10	0.14	0.11	0.18	0.11	0.19	0.5
	20	0.07	0.06	0.41	0.35	0.08	0.22
	50	0.06	0.08	0.45	0.07	0.06	0.21
	100	0.07	0.05	0.49	0.05	0.08	0.1
	1000	0.05	0.03	0.49	0.03	0.05	0.11

Table 5: Misclassification rates of the test dataset at different sample sizes under trivariate *t*- distribution, p =3.

		PARAMETRIC			NON-PARAMETRIC		
Distance	Sample Size	LDF	Logistic	Naïve Bayes	SVM	*k*NN	Decision Trees
5	10	0.07	0.15	0.42	0.44	0.25	0.5
	20	0.12	0.13	0.49	0.48	0.16	0.44
	50	0.3	0.29	0.55	0.19	0.19	0.29
	100	0.15	0.16	0.5	0.16	0.14	0.13
	1000	0.33	0.27	0.5	0.51	0.17	0.16
10	10	0.09	0.08	0.42	0.06	0.1	0.5
	20	0.11	0.09	0.48	0.13	0.1	0.3
	50	0.19	0.19	0.4	0.12	0.08	0.16
	100	0.06	0.08	0.5	0.07	0.1	0.18
	1000	0.17	0.15	0.5	0.49	0.12	0.09

Table 6: Misclassification rates of the test dataset at different sample sizes under tetravariate *t*- distribution, *p* =4.

			PARAMETRIC			NON-PARAMETRIC	
Distance	Sample Size	LDF	Logistic	Naïve Bayes	SVM	*k*NN	Decision Trees
5	10	0.1	0.17	0.52	0.2	0.37	0.5
	20	0.48	0.52	0.49	0.46	0.21	0.35
	50	0.2	0.23	0.48	0.51	0.09	0.28
	100	0.21	0.24	0.54	0.14	0.11	0.23
	1000	0.13	0.12	0.48	0.12	0.1	0.09
10	10	0.1	0.09	0.2	0.08	0.13	0.5
	20	0.17	0.02	0.48	0.04	0.08	0.19
	50	0.16	0.06	0.41	0.11	0.04	0.14
	100	0.08	0.05	0.53	0.06	0.07	0.09
	1000	0.11	0.27	0.48	0.1	0.07	0.12

Table 7: Misclassification rates of the test dataset at different sample sizes under bivariate exponential distribution, *p* =2.

			PARAMETRIC			NON-PARAMETRIC	
Distance	Sample Size	LDF	Logistic	Naïve Bayes	SVM	*k*NN	Decision Trees
5	10	0.23	0.15	0.17	0.3	0.18	0.5
	20	0.21	0.19	0.2	0.2	0.21	0.38
	50	0.2	0.21	0.18	0.21	0.22	0.26
	100	0.24	0.17	0.21	0.19	0.16	0.26
	1000	0.21	0.18	0.22	0.21	0.18	0.19
10	10	0.2	0.15	0.13	0.19	0.23	0.5
	20	0.15	0.16	0.15	0.13	0.18	0.36
	50	0.16	0.17	0.16	0.17	0.18	0.25
	100	0.22	0.12	0.14	0.17	0.12	0.17
	1000	0.17	0.17	0.18	0.17	0.17	0.2

Table 8: Misclassification rates of the test dataset at different sample sizes under trivariate exponential distribution, *p* =3.

			PARAMETRIC			NON-PARAMETRIC	
Distance	Sample Size	LDF	Logistic	Naïve Bayes	SVM	*k*NN	Decision Trees
5	10	0.18	0.15	0.18	0.14	0.13	0.5
	20	0.29	0.31	0.09	0.23	0.19	0.13
	50	0.13	0.12	0.11	0.12	0.1	0.12
	100	0.24	0.2	0.16	0.17	0.2	0.21
	1000	0.2	0.22	0.2	0.18	0.19	0.22
10	10	0.15	0.12	0.11	0.09	0.14	0.5
	20	0.31	0.31	0.08	0.15	0.17	0.15
	50	0.07	0.08	0.08	0.04	0.05	0.07
	100	0.11	0.07	0.09	0.07	0.1	0.1
	1000	0.17	0.08	0.08	0.07	0.08	0.1

Table 9: Misclassification rates of the test dataset at different sample sizes under tetravariate exponential distribution, *p* =4.

			PARAMETRIC			NON-PARAMETRIC	
Distance	Sample Size	LDF	Logistic	Naïve Bayes	SVM	*k*NN	Decision Trees
5	10	0.19	0.28	0.15	0.15	0.18	0.5
	20	0.15	0.14	0.14	0.13	0.1	0.14
	50	0.15	0.12	0.12	0.12	0.17	0.14
	100	0.14	0.15	0.14	0.13	0.16	0.17
	1000	0.14	0.13	0.14	0.14	0.13	0.15
10	10	0.36	0.4	0.11	0.03	0.07	0.5
	20	0.04	0.05	0.07	0.05	0.11	0.2
	50	0.04	0.08	0.01	0.04	0.05	0.12
	100	0.08	0.06	0.04	0.02	0.02	0.19
	1000	0.04	0.01	0.05	0.01	0.02	0.02

4. Discussion

From Table 1, at distance = 5, LDF was a better parametric classifier for all the samples sizes considered, having comparatively smallest misclassification rates. Logistic regression and SVM also showed very low misclassification rates. Naïve Bayes showed high but reducing misclassification rates as sample size increased. *k*NN was a better non-parametric classifier and both *k*NN and decision trees had a reducing rate as sample size increased. In other words, for a normally distributed multivariate observation to be classified, with groups not well separated, LDF is an ideal classifier. When distance increased to 10, LDF, logistic regression and SVM classified all observations to their right category, naïve Bayes also had reducing misclassification rates as sample size increased for parametric classifiers. *k*NN was a better non-parametric classifier and classified all observations to the right category. Decision trees also had a reducing classification rate as sample size increased. At sample size of 1000, all classifiers grouped all the points to their correct group. That means for a normally distributed data whose classes are well separated, LDF, logistic regression and SVM are ideal parametric classifier and *k*NN is an ideal non-parametric classifier to be used.

From Table 2, at distance = 5, LDF was the best parametric classifier having comparatively smallest misclassification rate. SVM and logistic regression also showed very low misclassification rates, while naïve Bayes showed the highest misclassification rate for parametric classifiers but had reducing misclassification rate as sample size increased. *k*NN was again a better classifier among the non-parametric classifiers and both *k*NN and decision trees had reducing rates as sample sizes increased. That means for a normally distributed data with not so well separated classes, LDF is an ideal classifier. When distance increased to 10, SVM classified all observation to their right category. LDF and Logistic regression had the same misclassification rates at all sample sizes and classified all observations correctly as the sample size increased. Naïve Bayes also showed low misclassification rates but performed poorly compared to the other parametric classifiers. For non-parametric classifiers, *k*NN classified all observations correctly and decision trees had reducing classification rates as sample size increased. That means for a normally distributed data whose classes are well separated, SVM, LDF and logistic regression are ideal parametric classifiers and *k*NN would also suffix if a non-parametric classifier is to be used.

From Table 3, at distance = 5, LDF was the best parametric classifier having comparatively smallest misclassification rate. SVM and logistic regression also showed very low misclassification rates. The naïve Bayes showed the highest misclassification rate but had reducing misclassification rate as sample size increased. For non-parametric classifiers, *k*NN was again a better classifier, both *k*NN and decision trees had reducing rates as sample size increased. That means for a normally distributed data with not so well separated classes, LDA is an ideal classifier. When the distance was increased to 10, LDF, logistic regression and SVM classified all observations to their right category for parametric classifiers, naïve Bayes also had reducing misclassification rates and classified all observations to the right group as sample size increased. *k*NN was a better non-parametric classifier and classified all observations to the right category. Decision trees also had a reducing classification rate as sample size increased. That means for a normally distributed data whose classes are well separated, LDF, logistic regression and SVM are ideal parametric classifiers and *k*NN would also suffix if a non-parametric classifier is to be used.

From Table 4, at distance = 5, all the four parametric classifiers had a fairly high misclassification rate, but logistic regression was the best classifier having comparatively smallest misclassification rate. Naïve Bayes had very high misclassification rates. For the non-parametric classifiers, misclassification rate was fairly high too and both *k*NN and decision trees had reducing classification rates as sample size increased. Also, *k*NN was a better non-parametric classifier. That means given a *t*-distributed multivariate dataset with not so well separated classes to be classified, logistic regression is an ideal classifier. When distance increased to 10, logistic regression was still the best classifier having comparatively smallest misclassification rate. LDF and SVM had close misclassification rates but slightly higher than the logistic regression. All the classifiers had decreasing misclassification rates as sample size increased but naïve Bayes had the highest values at all sample sizes considered. For non-parametric classifiers, both of them performed poorly but had decreasing rates as sample size increased with *k*NN being a better classifier at large sample size. That means for a *t*-distributed multivariate dataset whose classes are well separated to be classified, logistic regression is an ideal classifier.

From Table 5, all the parametric classifiers considered had a fairly high misclassification rates, but LDF and logistic regression showed the least misclassification rates. All the classifiers had reducing misclassification rates as sample size increased but naïve Bayes had very high misclassification rates. For non-parametric classifiers, misclassification rates were fairly high too and both *k*NN and decision trees had reducing classification rates as sample size increased. *k*NN was a better classifier. That means for a *t*-distributed multivariate dataset, with not so well separated classes, but a higher number of variables to be classified, logistic regression or LDF is an ideal classifier. When distance increased to 10, LDF, logistic regression and SVM showed low misclassification rates, with the logistic regression showing lowest rates, especially as n increased to 1000. Also, SVM showed a very good misclassification rates. LDF, logistic regression, naïve Bayes and SVM had decreasing misclassification rates as sample size increased but naïve Bayes had the highest value at all sample sizes considered. For non-parametric classifiers, both of them performed poorly but had decreasing rates as sample size increased with *k*NN being a better classifier. That means for a *t*-distributed multivariate dataset, whose classes are well separated to be classified, logistic regression is an ideal classifier irrespective of sample size considered.

From Table 6, at distance = 5, all parametric classifiers had a fairly high misclassification rate, but LDF and logistic regression showed the least misclassification rates. Naïve Bayes showed the highest value at all sample sizes. For non-parametric classifiers, both classifiers performed poorly but had decreasing rates as sample size increased with *k*NN being a better classifier. That means for a *t*-distributed multivariate dataset, with not so well separated classes and larger number of variables to be classified, LDF and logistic regression are ideal classifiers. When distance increased to 10, logistic regression was a better parametric classifier having comparatively smallest misclassification rate, LDF, logistic, naïve Bayes and SVM all had reducing classifiers as sample size increased, naïve Bayes had very high misclassification rates. For non-parametric classifiers, misclassification rate was fairly high too and both *k*NN and decision trees had reducing classification rates as sample sizes increased. However, *k*NN was a better classifier. That means given a *t*-distributed multivariate dataset whose classes are well separated, logistic regression is a good classifier.

From Table 7, all the classifiers performed relatively poorly under the multivariate exponential distribution and had very high misclassification rates. At a distance of 5, the logistic regression and naïve Bayes showed very close misclassification rates but logistic regression showed the least misclassification rate for parametric classifiers. For non-parametric classifiers, the *k*NN showed lower misclassification rates and the values were close to that of logistic regression. That means for a multivariate exponentially distributed dataset, with p = 2 and not well separated mean vectors, logistic regression is an ideal classifier. When distance was set at 10, naïve Bayes and logistic regression showed the least misclassification rates for parametric classifiers. For non-parametric classifiers, *k*NN was a better classifier and had close values with naïve Bayes. That means for a multivariate exponentially distributed dataset, with p = 2 and well separated mean vectors, the naïve Bayes is an ideal classifier.

From Table 8, all the classifiers performed relatively poorly with very high misclassification rates. At a distance of 5, naïve Bayes showed the least misclassification rate for parametric classifiers. For the non-parametric classifiers, *k*NN showed lower misclassification rates and the values were close to that of naïve Bayes. That means for a multivariate exponentially distributed dataset, with p = 3 and not well separated mean vectors, naïve Bayes is an ideal classifier. When distance was set at 10, SVM showed comparatively smallest misclassification rates for parametric classifiers. For non-parametric classifiers, *k*NN was a better classifier and had close values with SVM. That means that the SVM is an ideal classifier for a 3-dimensional exponentially distributed dataset with well separated mean vectors.

Lastly from Table 9, all the classifiers performed poorly under the multivariate exponential distribution with very high misclassification rates. At a distance of 5, SVM and naïve Bayes showed very close misclassification rates but SVM showed the least misclassification rate for parametric classifiers. For non-parametric classifier, *k*NN showed lower misclassification rates and the values were close to that of logistic and naïve Bayes. When distance was set at 10, all the classifiers performed better. SVM showed the least misclassification rates for parametric classifiers. For non-parametric classifiers, *k*NN was still a better classifier and had close values with SVM. As a result, for a 4-dimensional exponentially distributed dataset with not well separated mean vectors, SVM and naïve Bayes are ideal classifiers. Also, when the mean vectors are well separated, only the SVM is an ideal classifier.

In general, the parametric classifiers considered in this work performed better than their non-parametric counterparts under all the conditions considered. Also, all the classifiers performed better with increasing sample sizes.

5. Conclusion

From the results obtained in this empirical study, LDF was best when the observations were from a multivariate normal distribution irrespective of the sample size, distance between the mean vectors or number of variables. However, it performed poorer with datasets that were not p-dimensional normally distributed. Logistic regression acted like LDF but was more flexible than LDF and performed better in datasets that were not normally distributed like the multivariate *t*- and the multivariate exponential distributions. Naïve Bayes performed poorly overall. It was not an ideal classifier for normally and *t* distributed datasets but was better for exponential distribution, especially as the number of variables and distance between the mean vectors increased. SVM was a very good choice for classification when the classes were very well separated for normally and exponentially distributed datasets but performed poorly for *t*-distributed datasets irrespective of distance and number of variables. It also performed better as sample sizes increased but when sample size was too large, SVM was not ideal.

Generally, parametric classifiers performed better than non-parametric classifiers in all the distributions considered. However, if a non-parametric classifier was to be used, *k*NN was a more flexible classifier and performed better than decision trees for all the distributions considered. It was also very ideal when the groups were well separated. On the other hand, decision trees performed better as sample size increased irrespective of the distance between the mean vectors, number of variables or distribution of dataset. Finally, all the classifiers performed better when larger sample sizes were used for training data sets and with large distance between their mean vectors.

References

Abubakar, A. (2020). Comparative analysis of classification algorithms using CNN transferable features: A case study using burn datasets from black Africans. *Applied System Innovation, 3*(4), 43, doi.org/10.3390/asi3040043.

Adenaike, A. S., Oloye, O. S., Emmanuel, H. O., Bello, K. O., & Ikeobi, C. N. O. (2022). Comparison of linear discriminant analysis, support vector machine and artificial neural network in classifying Nigerian local turkeys based on plumage colours using biometric traits. *Journal of Agriculture and Rural Development in the Tropics and Subtropics, 123*(2), 197–204, doi.org/10.17170/kobra-202210116964.

Akinmoladun, O. M., Kareem, R., & Ekum, M. I. (2017). Criteria for selecting prospective students into higher institutions using discriminant and artificial neural network analysis. *IOSR Journal of Mathematics, 13*(6), 77-90.

Gambo, A. I., & Yusuf, M. W. (2010). An application of multivariate analysis in modeling students placement in Nigerian higher institutions. *Journal of Mathematics and Statistics, 6*(3), 350-356.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2017). *An introduction to statistical learning with applications in R* (8th ed.). New York: Springer.

Kim, J., Le, D. X., & Thoma, G. R. (2011). Naïve Bayes and SVM classifiers for classifying databank accession number sentences from online biomedical articles. *Proc. SPIE, 7534*(XVII), doi.org/10.1117/12.838961.

Lakshmi, D. C. (2013). Classification of multivariate datasets without missing values using memory based classifiers – An effectiveness evaluation. *International Journal of Artificial Intelligence & Applications, 4*(1), 129-142, doi.org/10.5121/ijaia.2013.4110.

Lilima, P., Taneja, N. A., Dixit, C., & Suhag, M. (2017). Comparison of text classifiers on news article. *International Research Journal of Engineering and Technology, 4*(3), 2513-2517.

Neomi, M., Pedro, J. M., Rafael, G., & Diego, G. (2015). Multivariate analysis in thoracic research. *Journal of Thoracic Disease, 7*(3), E2-E6, doi.org/10.3978%2Fj.issn.2072-1439.2015.01.43.

Nurit, O., & Avi, O. (2015). Comparison of two multivariate classification models for contamination event detection in water quality time series. *Journal of Water Supply: Research and Technology-Aqua, 64*(5), 558-566, doi.org/10.2166/aqua.2014.033.

Nurshahirah, M. A., Othman, M. S., & Yusuf, L. M. (2019). Comparative analysis of classifiers for education case study. *International Journal of Software Engineering and Computer Systems, 5*(1), 67-76, doi.org/10.15282/ijsecs.5.1.2019.5.0055.

Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed.). New York: John Wiley & Sons, Inc.

Yakubu, A., & Ibrahim, I. A. (2011). Multivariate analysis of morphostructural characteristics in Nigerian indigenous sheep. *Italian Journal of Animal Science, 10*(2), 83-86, doi.org/10.4081/ijas.2011.e17.

Yugal, K., & Sahoo, G. (2012). Analysis of parametric & non parametric classifiers for classification technique using WEKA. *International Journal of Information Technology and Computer Science, 7*, 43-49, doi.org/10.5815/ijitcs.2012.07.06.