Comparative analysis of Malignancy prediction of Breast Cancer cells using Logistic Regression& K Means Algorithm

ABSTRACT

Cancer is treated as one of the major health concerns of the country and it is caused when some abnormality happens to a cell. Among cancer, one of the most common cancer that affects the female gender is breast cancer and there are many new cancer cases as well as death reported globally, and it is one of the major concern of public health. So, an early diagnosis of cancer can provide clinical suggestions to recover or improve the survival rate significantly. This paper suggests a comparative analysis of the detection of breast cancer using one of the machine learning algorithms like logistic regression and K Means Algorithm. An open dataset provided by the University of Wisconsin Hospital at Madison, Wisconsin, USA will be used to implement the algorithm. There are 569 instances of data available in the open dataset with its classification as malignant and benign. The datasets undergo preprocessing followed by the implementation of algorithms from which results are visualized using the orange tool. Models are trained and tested from which results are displayed in different forms. The classification is depicted using a confusion matrix as well as through a ROC curve while MDS and Silhouette plot is used for K-means clustering. Comparison between logistic regression and k means clustering is observed along with a comparison of different literature reviews with research in hand. It was concluded that logistic regression is a better predictive model based on the given requirements and usage of different machine learning tools significantly affects the results and accuracy of the model.

Introduction

Artificial intelligence is a field of computer science that describes machines that work like human beings in the sense of operations and functionalities. This field of study is mainly divided into three parts which are artificial narrow intelligence, artificial general intelligence, and artificial super intelligence [11]. Based on levels, narrow intelligence is the least with implementations search engines, social media, etc. while superintelligence is the highest that has yet to exist. General intelligence implementations include voice recognition in smart devices and virtual assistants. Artificial intelligence is widely used in various fields of study and sectors. AI application in health care is mainly used for diagnosis of diseases based on medical tests and results. AI application in the medical industry is used to discover new medicines and support the interaction of different drugs [3]. Machine learning is a subset of Artificial Intelligence that uses previously recorded data to train a model that will be used to predict results based on new data entry. Machine learning is mainly divided into four categories which are supervised learning, semi-supervised learning, unsupervised learning, and reinforced learning [1]. Supervised learning focuses on classification and regression methods. On the other hand, unsupervised learning involves clustering techniques while semi-supervised deals with both labeled and unlabeled data. Lastly, a method that focuses on a goal throughout the process of learning is called reinforcement learning. Furthermore, machine learning practices have been applied in various sectors which include health care, web, business, and security. Some of the most common applications of machine learning are the diagnosis of diseases, profiling customers, marketing, and fraud detection. Modern applications of machine learning have been observed that include biological sequence learning, text learning, and web learning [2]. Text learning application includes predicting words in search engines while web learning involves the study of user's behavior in the web. The human genome project involves biological sequence learning. Breast cancer is a very common cancer in women around the world. The number of patients with breast cancer keeps on increasing throughout the years. Most of these patients are women of urban areas. Among different types of cancers in Oman, breast cancer is mostly observed in Omani women with around 16 percent of female cancer patients

diagnosed with breast cancer [4]. Most of these cases are diagnosed at a later stage which decreases the possibility of treatment. To ensure early diagnosis different machine learning algorithms are used to predict breast cancer. The algorithms include Naïve Bayes, Artificial Neural Network, Random Forest, Support vector machine, logistic regression, K means clustering, etc. These algorithms use breast cancer results to predict if the tumor is Malignant or Benign.

Literature Review

This section will review different resources of research which will be divided into three parts. The parts are machine learning review, logistic regression review and k means clustering review. According to Ayon Day machine learning consists of four main categories which are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. The following categories will be discussed one by one accordingly. A research paper elaborates that this learning method uses labeled data to train the models. The models include regression and classification-based algorithms. It states that supervised learning uses labeled data which is fit for this type of learning [9]. A highlight on the most used supervised algorithms are Naïve Bayes, Support vector machine, and Decision tree. Naïve Bayes classifies the data and depends on conditions from which different trees are formed called Bayesian network. Support vector machine draws margins between classes to the maximum which minimizes the errors in the classification of data. Lastly, a decision tree algorithm consists of many trees which contain nodes and branches that classify data based on values. The nodes and branches are linked to each other and a big tree diagram is formed [7]. This learning technique is mostly used for new and unlabeled data [8]. Prediction is made based on patterns found on the new data which is determined by clustering methods such a k means clustering and dimension reduction methods such as principal component analysis. The grouping of data in an automated form along with separating the data into clusters with a k distance between them is called the k means algorithm. PCA reduces the dimension and uses two axes that plot the data accordingly. The reduction of dimension makes it easier and faster to process the data. In this learning method, a combination of techniques from supervised and unsupervised learning are used. It also deals with less labeled and more unlabeled data. A few semi-supervised learning features are generative models, transductive SVM, and self-training [7]. All these features deal with a mixture of unlabeled and labeled data that mainly focuses on unsorted data and few labeled data. Here the data is processed after the learner has decided the type of learning method (supervised or unsupervised) to be used [9]. The learner picks the method that will bring better results based on the given input. This method solely depends on two main areas delayed outcome and trial and error search Ayon.

A model that uses the probability of 1s and 0s to classify the outcome based on data input. This model can be further used to classify a variety of groups using extended features. It is mostly used when data is categorical, the data is used to train and test the model for prediction purposes [10]. A research was conducted from which a dataset was created at the University of Wisconsin by a physician named Dr. William Wolberg who collected fluid samples which were then computerized and 10 features were deduced. The data was preprocessed by eliminating the missing values and empty spaces. The tool used to conduct this implementation of research is the Jupiter tool. The supervised models like Decision tree, K nearest neighbors, Support vector machine, Random forest, Naïve Bayes along with logistic regression were used to determine the patterns, and results were found that the Random forest method had the highest accuracy of 96.5% after the testing phase and logistic regression had an accuracy of 94.4%. Random forest was seen to be the best method to determine malignancy in breast cancer patients [11]. Another research was conducted on the same dataset using Logistic regression, KNN, K-star, MLP, Random forest and Decision tree, Decision table, PART, and REP tree. These machine learning algorithms were used to test the accuracy, sensitivity, specificity, RMSE, Kappa statistics, F-measure, and ROC curve area of each model. The preprocessing technique used was the removal of duplicate and empty data. The tool used to conduct this research was the WEKA explorer tool. The results showed that logistic regression had the highest accuracy of 97% and

lowest RMSE value of 0.14. hence concluded that the logistic regression was the best technique to predict malignancy on breast cancer inputs [13]. Third research that used the same dataset from Wisconsin University showed an implementation of logistic regression by manipulating sigmoid function and weighting factors to observe the changes between a classical method and the newly proposed method. The preprocessing of data involves image preprocessing, image segmentation, and feature extraction and selection. The tool used in this research was MATLAB. The results proved that the proposed method had shown better results than the normal method. The accuracy obtained from the classical logistic regression method was 95.7% [12]. If we observe the implementation phase of each research, we come to know that tools used to implement the same dataset were different. While the data preprocessing methods used in each research were also different. These factors can be seen to have affected the results of logistic regression while using the same dataset.

k-means clustering is simply the grouping of data based on the features given as input. The data will be classified into clusters and the distance between clusters is given as 'k' while the center of the cluster value is the mean value. Research that shows the prediction of malignancy and benign based on FNA biopsies from the University of Wisconsin uses k- means clustering as a predictive model [15]. The dataset is based on 699 breast cancer cases with 16 cases containing missing values from which preprocessing is done by eliminating the missing values. MATLAB tool is used to train and test the data using the k-means model. Accuracy of 96.4% is obtained which is concluded to be a high accuracy showing that the model can be used to predict breast cancer. Another research uses a similar dataset from the University of Wisconsin but with 30 attributes defining texture, smoothness, compactivity, area, etc. No missing values were found the dataset was observed to be complete hence further steps were taken. Accuracy of 85.7% was observed compared to the SOM algorithm which had a higher accuracy of 88.8%. A hybrid algorithm of k-means and SOM was seen to have the highest accuracy of 92.1%. Hence concluded that the hybrid was the best method to predict breast cancer [14]. The third research was done on the Breast cancer Wisconsin diagnostic dataset which had 11 attributes from which different factors like cell size, cell shape, nuclei mitoses, etc. The research was based done to observe the changes of accuracy by manipulating different computerized measures of k-means clustering [16]. The tool used to experiment is NetBeans IDE. The results concluded that the highest variance with the same centroid gave the highest accuracy of about 92%. Three different datasets acquired from three different medical methods were used from which different tools were applied but with the same cause which is predicting breast cancer. Here we can see that the choice of a dataset is important, and all three choices determine different accuracy results.

Methods and Implementation

The Breast Cancer Wisconsin Diagnostic Data Set is used containing 569 instances and 32 attributes along with 0.1% missing values. The data is categorized into three components which are ID, Group, and features. The ID simply resembles a unique number for each test result while the group classifies the record as either M (Malignant) or B (Benign). The features are obtained from the cell nucleus which are computerized. The features are:

- region (average of distances from middle to spots on the boundary)
- surface (standard deviation of gray-scale values)
- boundary
- region
- softness (local variation in radius lengths)
- density (boundary^2 / region - 1.0)
- incurvation (severity of concave portions of the contour)
- incurvation points (count of concave portions of the contour)
- symmetry
- fractal portion ("coastline estimation" - 1)

ID	Group	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9
842302	M	17.990	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.1471	0.2419
842517	M	20.570	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812
84300903	M	19.690	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.1279	0.2069
84348301	M	11.420	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.1052	0.2597
84358402	M	20.290	14.34	135.10	1297.0	0.10030	0.13280	0.198	0.1043	0.1809
843786	M	12.450	15.70	82.57	477.1	0.12780	0.17000	0.1578	0.08089	0.2087
844359	M	18.250	19.98	119.60	1040.0	0.09463	0.10900	0.1127	0.074	0.1794
84458202	M	13.710	20.83	90.20	577.9	0.11890	0.16450	0.09366	0.05985	0.2196
844981	M	13.000	21.82	87.50	519.8	0.12730	0.19320	0.1859	0.09353	0.2350
84501001	M	12.460	24.04	83.97	475.9	0.11860	0.23960	0.2273	0.08543	0.2030

Figure 1. Breast Cancer Wisconsin Diagnostic Data Set in a table format [6].

Data preprocessing – Missing values and Outliers

The dataset consists of 569 instances with 0.1% missing values. The missing values were eliminated from the data table. Eliminating the values would also eliminate the corresponding values related to that case. This technique was chosen based on the research in hand, compared to other imputation methods removing the missing values showed to have maximum accuracy over the other methods. hence, the missing values were eliminated from the data table for future processes.

Table 1. Comparison of different imputation methods using accuracy percentage.

Imputation method	Accuracy percentage
Remove missing values	95.6%
Average	95.4%
Random values	95.4%
Simple tree values	95.5%
Without imputation	95.4%

On the other hand, we conducted a test to check for outliers from which two methods were chosen. The local outlier factor and isolation forest. It was found that the usage of both outlier methods at default showed isolation forest to have greater accuracy of 95.3% than the local outlier factor. But with the usage of different distance measurement techniques, we found that the cosine distance technique showed higher accuracy than that of isolation forest. We concluded that usage of local outlier factor with 20 neighbors and cosine distance technique portrayed the highest accuracy.

Table 2. Comparison of different LOF distance techniques using accuracy percentage.

Local outlier factor distance metric	Accuracy percentage
Euclidean	95.1%
Manhattan	95.2%
Cosine	97.0%
Jaccard	94.6%
Hamming	95.6%
Minkowski	95.1%

Integration of data

The tool used for implementation, evaluation, and prediction is the Orange tool. Many programming languages like C++, C, Python, etc. were used to develop such tool [17]. It is a user-friendly tool and contains many visualization and algorithmic methods. The tool works simply by linking various widgets based on user and algorithmic requirements. Many formats can be read by the tool which includes CSV, basket, .tab, and much more. Here, the tool was used to implement logistic and k-means clustering accordingly.

Implementation of logistic regression

In logistic regression, the data was retrieved and read by the orange tool from which missing values were identified along with numeric and classified data. Data was then divided into target variables which were the group (M & B) and features which were the 10 features discussed above ignoring the ID since it did not have any importance to the process. The process was followed by the imputation of data from which the missing values were eliminated. After preprocessing, the logistic regression model was trained and tested. The test score results were obtained along with the confusion matrix, prediction, and ROC analysis.

Implementation of k-means clustering

K means clustering has similar initial steps as logistic regression in which data is retrieved displaying the missing values, classified and numeric data. The data is then imputed to remove the missing values followed by the k means clustering model which divides the data into different two different clusters (C1 and C2). The data is then compared from the actual values with the predicted values. The results are then visualized using MDS distribution and silhouette chart.

Results and Discussion

After the implementation of logistic regression and k-means algorithm, results were obtained. For logistic regression, the confusion matrix was first obtained which showed an equal number of Benign values of 688 for predicted and actual data. A slight difference in the values of Malignant which showed 14 for predicted and 17 for actual. Based on the result, we can deduce that the algorithm could identify the benign values at 100% accuracy with some errors on the malignant prediction.

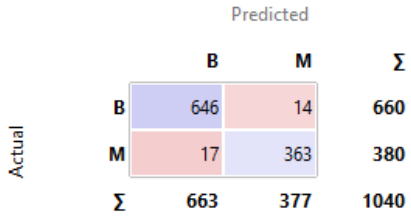

Figure 4. Confusion matrix of logistic regression.

Logistic regression results are followed by the ROC curve. The area under the curve determines the ability of the model to predict values on a binary form. The higher the area the better the model, for the AUC of this curve is 0.991 which is 99.1% area this shows that the algorithm has a great potential to predict results.

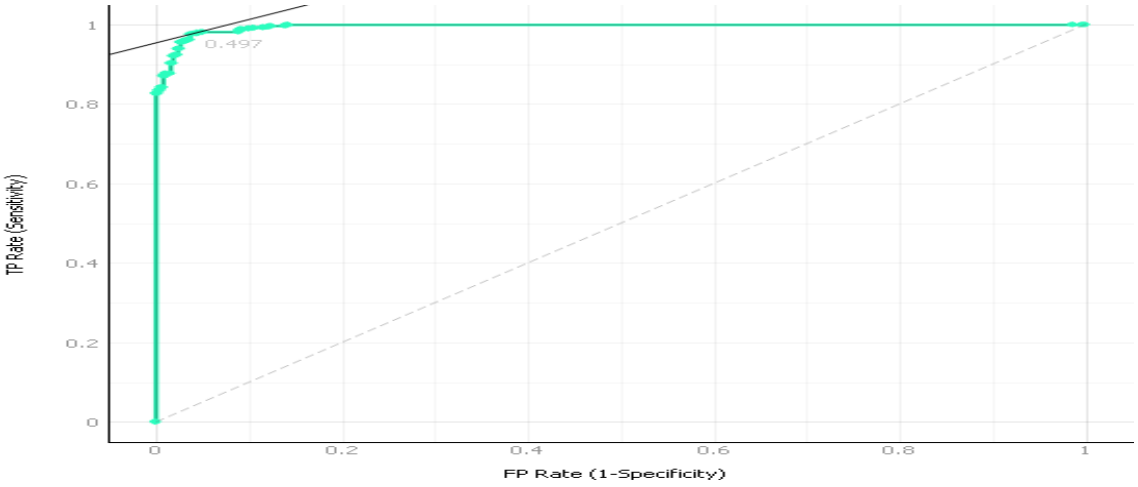

Figure 5. ROC curve of logistic regression.

Based on the scores we can see that the area under graph (AUC) is at 99.5%, classification accuracy (CA) at 97.0%, harmonic mean (F1) at 97.0%, Precision at 97.0% and Recall at 97.0%. We can deduce that from the values that a higher number of all these factors means that the model has a high potential of predicting breast cancer.

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.995	0.970	0.970	0.970	0.970

Figure 6. Test and score of logistic regression.

K-means clustering result

On the other side, for k-means, a silhouette chart was used to showcase the results. It was observed that blue represented C1 which was Benign and the color red represented C2 which was Malignant. The values above zero for clusters one and two were successfully interpreted while the ones below zero were misinterpreted. C1 showed good results with few data misinterpreted while C2 portrayed more misinterpreted values. This shows that the model was not able to classify the data with high precision.

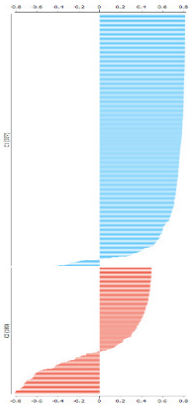

Figure 7. Silhouette chart of K-means clustering.

Another visualization method used was the MDS diagram. The diagram showed that the points of C1 points to be closer to each other which represents a good interpretation of benign data. While spread points of C2 portray bad data interpretation. We can deduce that the model was not able to classify C2 values properly.

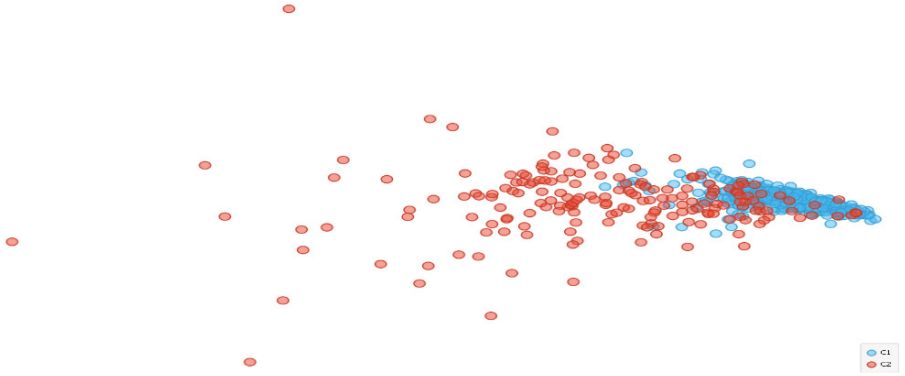

Figure 8. MDS plot of K-means clustering.

Comparative analysis

Given that the dataset is the same for both algorithms with which the data preprocessing is the same for both methods. We observe that linear regression is a better choice to predict breast cancer as it shows higher precision levels and other test score values while k means shows poor clustering of data hence regression is a better choice. The table below shows a general comparison between two algorithms.

Table 2. General comparison of logistic regression and k means.

Logistic regression	K-means clustering
"Group" is taken as a target variable and the 10 features are taken as feature variables	"Group" is taken as a target variable and the 10 features are taken as feature variables
Processes continuous and categorized data	Processes continuous and categorized data
Predictive algorithm	Predictive algorithm
Binary classifier system	Clustering system
S-shaped curve	K distance value
Binary output	Clustered output

We deduced that logistic regression was a better method to predict breast cancer compared to k-means clustering. But here we compared different accuracies of the literature review along with the accuracy of this paper and the results were given as shown. We observe that the same dataset was used for all research along with similar preprocessing methods. The difference observed was the usage of different tools. Hence the difference in usage of tools would differentiate the results as well.

Table 3. Comparison of literature review accuracy.

S. No	Paper title	Author	Dataset used	Tool used	Result (Accuracy)
1	Breast Cancer Detection and Prediction using Machine Learning	A. Chowdhury	Breast Cancer Wisconsin Diagnostic Data Set	Jupiter tool	94.4%
2	Classification of malignant and benign tissue with logistic regression	L. Khairunnahar et al.	Breast Cancer Wisconsin Diagnostic Data Set	MATLAB simulation tool	95.7%
3	Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers	J. Sultana and A. Jilani	Breast Cancer Wisconsin Diagnostic Data Set	WEKA explorer tool	97.2%
4	Comparative analysis of Malignancy prediction of Breast Cancer cells using Logistic Regression& K Means Algorithm	Shariff A. and Dr.C.Jayakumari	Breast Cancer Wisconsin Diagnostic Data Set	Orange tool	97.0%

Conclusion

Breast cancer has been a big challenge for women nowadays. This cancer is mostly diagnosed at the late stage where no treatment is found. This mostly leads to removal of the organ or worse no solution at all. An early diagnosis is required to prevent such events from happening. Here machine learning comes to picture, with such advanced methods an early diagnosis can occur saving many lives. These machine learning techniques proved to function efficiently compared to human diagnosis and analysis. In this research, we implemented supervised and unsupervised machine learning methods. We concluded that the supervised method proved better prediction than that of unsupervised. We also concluded that the usage of different tools can significantly affect the accuracy of the model. Overall, a breast cancer diagnosis dataset was used to compare to machine learning algorithms and different literature reviews were compared with this research in terms of accuracy and usage of various tools.

Limitations

The limited exploration of different factors of logistic regression and k-means clustering. This limited the ability of models to portray various results depending on factors like normalization, error analysis, imbalance, etc. for logistic regression. For k-means factors like initialization, normalization, several iterations, etc. On other hand, exploring different unsupervised and supervised learning methods at a deeper level to realize better methods for maximum

accuracy is another limitation. Lastly, the implementation of various data preprocessing methods could have improved the results including accuracy.

Acknowledgment

I would like to thank the UCI machine learning repository for providing the dataset based on real cases. I would also like to thank all authors referred to this paper for providing knowledge and resources which were used to support this research. Lastly, I would like to thank Middle East College along with my co-author for the support to complete this paper.

References

[1]Ö. Çelik, "A Research on Machine Learning Methods and Its Applications", *Journal of Educational Technology and Online Learning*, 2018. Available: 10.31681/jetol.457046 [Accessed 27 April 2021].

[2]G. Tzanis, I. Katakis, I. Partalas, and I. Vlahavas, “Modern Applications of Machine Learning,” , 2021

[3]S. Mannam, “Journal of Young Investigators Science News,.” [Online]. Available:

[4]M. Al-Azri et al., "Psychosocial Impact of Breast Cancer Diagnosis Among Omani Women", *Oman Medical Journal*, vol. 29, no. 6, pp. 437-444, 2014. Available: 10.5001/omj.2014.115 [Accessed 27 April 2021].

[5]N. Rane, J. Sunny, R. Kanade, and P. Devi, "Breast Cancer Classification and Prediction using Machine Learning", *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 02, 2021. [Accessed 27 April 2021].

[6]“UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set,” *Uci.edu*, 2021.

[7]A. Dey, “Machine Learning Algorithms: A Review II. TYPES OF LEARNING,” *Ayon Dey / (IJCSIT) International Journal of Computer Science and Information Technologies*, vol. 7, no. 3, pp. 1174–1179, 2016,

[8]S. Khanna, “Machine Learning v/s Deep Learning,” 2008 International Research Journal of Engineering and Technology (IRJET) 6 (2)

[9]H. Wehle, "Machine Learning, Deep Learning, and AI: What’s the Difference?", in *Data Scientist Innovation Day*, 2017

[10]E. DiGangi and J. Hefner, "Ancestry Estimation", *Research Methods in Human Skeletal Biology*, pp. 117-149, 2013. Available: 10.1016/b978-0-12-385189-5.00005-4 [Accessed 27 April 2021].

[11]A. Chowdhury, "Breast Cancer Detection and Prediction using Machine Learning", 2021. Available: 10.13140/RG.2.2.23969.84320 [Accessed 27 April 2021].

[12]L. Khairunnahar, M. Hasib, R. Rezanur, M. Islam and M. Hosain, "Classification of malignant and benign tissue with logistic regression", *Informatics in Medicine Unlocked*, vol. 16, p. 100189, 2019. Available: 10.1016/j.imu.2019.100189 [Accessed 27 April 2021].

[13]J. Sultana and A. Jilani, "Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers", *International Journal of Engineering & Technology*, vol. 7, pp. 22-26, 2018. [Accessed 27 April 2021].

[14]H. Lin and Z. Ji, "Breast Cancer Prediction Based on K-Means and SOM Hybrid Algorithm", *Journal of Physics: Conference Series*, vol. 1624, p. 042012, 2020. Available: 10.1088/1742-6596/1624/4/042012 [Accessed 27 April 2021].

[15]N. Ayoob, "Breast Cancer Diagnosis Using K-means Methodology", *JOURNAL OF UNIVERSITY OF BABYLON for Pure and Applied Sciences*, vol. 26, no. 1, pp. 9-16, 2017. Available: 10.29196/jub.v26i1.348 [Accessed 27 April 2021].

[16]A. Dubey, U. Gupta, and S. Jain, "Analysis of k-means clustering approach on the breast cancer Wisconsin dataset", *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 11, pp. 2033-2047, 2016. Available: 10.1007/s11548-016-1437-9 [Accessed 27 April 2021].

[17]M. Peker, O. Özkaraca, and A. Şaşar, "Use of Orange Data Mining Toolbox for Data Analysis in Clinical Decision Making", *Expert System Techniques in Biomedical Science Practice*, pp. 143-167, 2018.