UDC 57.088

Sh. Shamiluulu[1], U. Djakbarova[2]
[1]Department of Computer Science,
SuleymanDemirel University, Kaskelen, Kazakhstan, 040900
[2]Department of Molecular Biology and Medical Genetics,
Kazakh National Medicine University named after Asfendiyarov, Almaty, 035000
Department of Molecular Biology and Medical Genetics

IDENTIFICATION OF PATIENTS WITH BREAST CANCER BY USING MACHINE LEARNING ALGORITHMS OVER SCIKIT-LEARN ML FRAMEWORK

In this research study the effect of normalization techniques is examined. The five different supervised machine learning algorithms i.e., KNN, Decision tree, Naïve-base, Logistic regression and ANN are used on breast cancer dataset obtained from UCI machine learning repository and their performances are compared. The study reveal that different preprocessing techniques can increase the classification accuracy over 90% where high performance is given to Logistic regression and ANN. The proposed approach can be implemented in a well-known benchmark medical problem with real clinical data forbreast cancer disease diagnosis.
Keywords: *Breast cancer, Machine Learning Algorithms, Data Classification, Computer Aided Prognosis and Diagnosis*

I – ıntroductıon.

Presently, the use of artificial intelligence (AI) has become widely accepted in medical applications. This is manifested by an increasing number of medical devices currently available on the market with embedded AI algorithms [1]. Such devices are being used cancer diagnosis areas where prognosis and diagnosis of breast cancer plays an important role. Breast cancer is the most common cancer among women, except for skin cancers. According to CDC statistics 1 in 8 (12%) women in the US will develop invasive breast cancer during their lifetime. Breast cancer starts when cells in the breast begin to grow out of control [8]. These cells usually form a tumor that can often be seen on an x-ray or felt as a lump. The tumor is malignant (cancerous) if the cells can grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. Breast cancer occurs almost entirely in women, but also possible to occur in men.It's also important to understand that most breast lumps are not cancer, they are benign. Benign breast tumors are abnormal growths, but they do not spread outside of the breast and they are not life threatening. But some benign breast lumps can increase a woman's risk of getting breast cancer. Any breast lump or change needs to be checked by a health care provider to determine whether it is benign or cancer, and whether it might impact your future cancer risk [4].
The goal of a study is to reveal the presence of tumor and classify into two classes benign or malignant.During the analysis we studied the effect of preprocessing and normalization techniques on classification model. The published literature suggests that machine learning (ML) algorithms have been shown to be valuable tools in reducing the workload on the clinicians by detecting artefact and providing decision support, potentially with the ability to automatically re-estimate the prediction or classification model in real-time.

II - materials and methods.

2.1 Machine Learning Algorithms.

The scikit-learn machine learning framework with five algorithms has been used to evaluate the classification performance on breast cancer dataset. The brief explanations for algorithms are provided below.
i. K Nearest Neighbors (KNN) algorithm is one of the first simple supervised learning machine learning algorithms. The logic behind this method is to find a predefined number of training samples closest in distance to the new point, and predict the label from these given data-points. Despite its simplicity, nearest neighbors has been successful in a large number of classification and regression problems. As a distance metric generally the Euclidean distance measure is used. For detailed information refer [1].
ii. Decision Trees (D-Tree)is a supervised learning method that is used for classification and regression. The feature is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. This method has some advantages like being simple to understand and easy to interpret and also trees can be visualized and requires little data preparation. The method is based on information theory paradigm. The more information can be obtained [1].
iii. Gaussian Naïve Bayes (NB) is a classification technique based on Bayes' Theorem. In general, the Naïve Bayes classifier assumes the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an orange if it is orange, round, and about 10 cm in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'. This method's advantage is that Naive Bayes model is easy to build and particularly useful for very large data sets. For details refer [1].
iv. Logistic regression (Logit)is a part of regression models where the output value is binary or dichotomous. The prediction curve is S-shaped and based on a sigmoid function [1]. Because of non-linear nature this algorithm shows one of the best results on getting the classification model for the data, for details refer results and discussion section.
v. Artificial Neural Network (ANN)is a new alternative to Logit, the statistical technique with which they share the most similarities.Neural networks are algorithms that are patterned after the structure of the human brain [1]. They contain a series of mathematical equations that are used to simulate the biological processes such as learning and memory.In a ANNs, one has the same goal as in Logit modeling, predicting an outcome based on the values of some predictor variables.

2.2 Data collections.

The dataset obtained from UCI machine learning repository. There are 31features and over 600 instances.Table 1 shows details of attributes with correlation coefficients. The target attribute provides 4 categories where first three are related to heart diseases and last one to healthy state. The hold-out method used for training and testing the models, where 70% for training set and 30% for testing set.

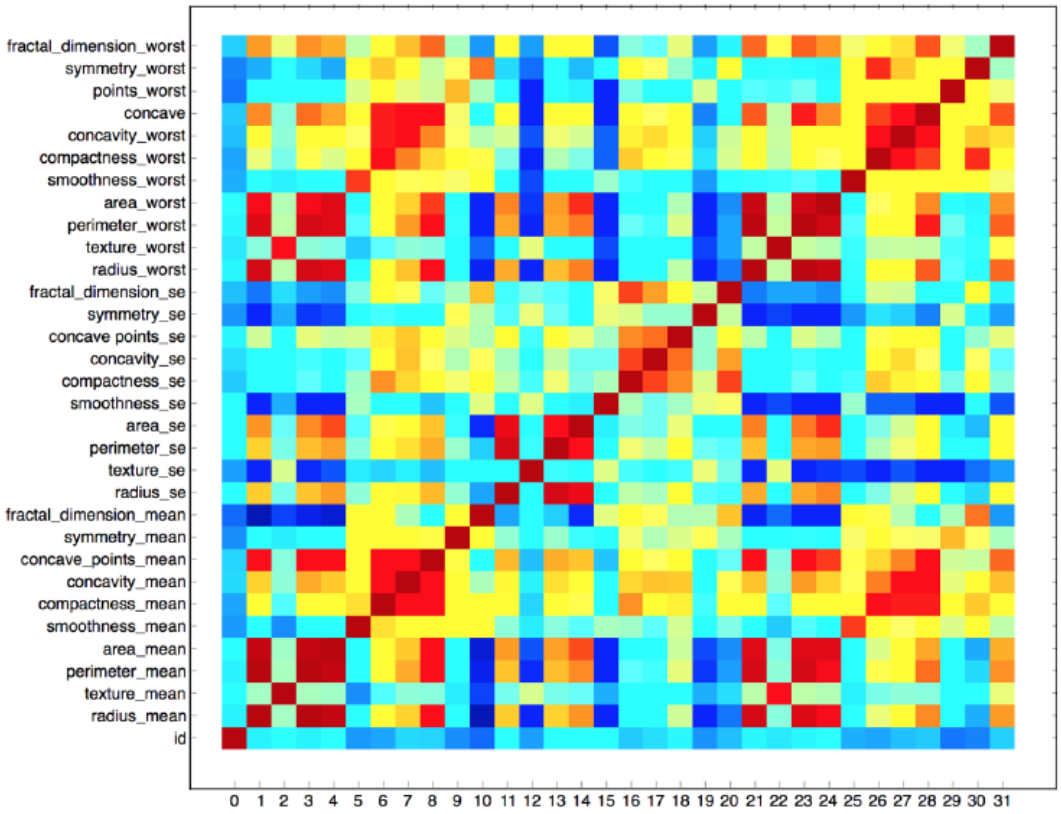

Figure 1 - Correlational plot for all features

In the Figure 1, we can see a correlational plot for 31 features. The red square cells indicate the high correlation whereas the blue dots low correlation. The dataset has been divided into two parts with highly correlated features and low ones. The goal was to study the effect of correlation and preprocessing techniques on classification performance of algorithms. The correlation was found by using spearman method, because it will be more precise for non-linear dataset. The features in highly correlated dataset is between $\pm 0.5 \leq r \leq \pm 1$ whereas in low correlated is $-0.49 \leq r \leq +0.49$.

Table 1 - Highly correlated featuresset

Feature ID	Name	Correlationcoefficient	Description
F11	radius_mean	0.730	Mean of distances from center to points on the perimeter
F12	perimeter_mean	0.743	Mean of perimeter
F13	area_mean	0.709	Mean of area
F14	compactness_mean	0.597	Mean of compactness, perimeter^2 / area – 1.0
F15	concavity_mean	0.696	Mean of concavity, severity of concave portions of the contour
F16	concave_points_mean	0.777	Mean of concave points, number of concave portions of the contour
F17	radius_se	0.567	Standard error of distances from center to points on the perimeter
F18	perimeter_se	0.556	Standard error of perimeter
F19	area_se	0.548	Standard error of area
F110	radius_worst	0.776	"worst" or largest (mean of the three largest values) of distances from center to points on the perimeter
F111	perimeter_worst	0.783	Mean of the three largest values of perimeter
F112	area_worst	0.734	Largest (mean of the three largest values) of area
F113	compactness_worst	0.591	Compactness's mean of the three largest values
F114	concavity_worst	0.660	Concavity's largest
F115	concave_points_worst	0.794	Worst of concave points

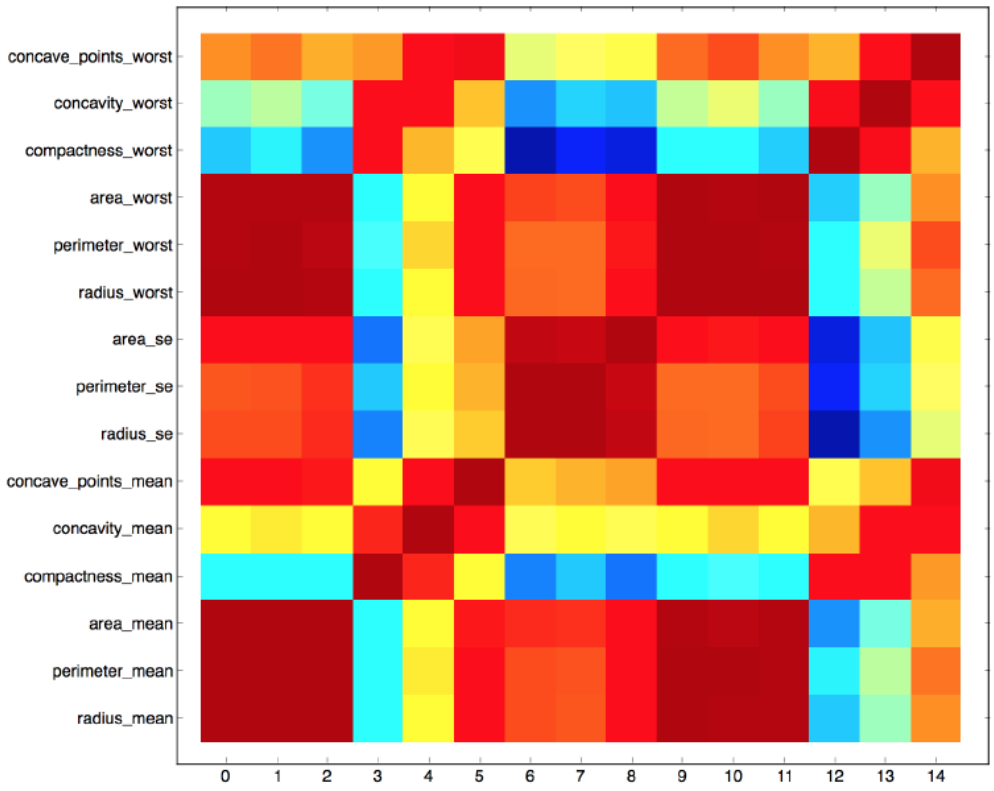

Figure 2 - Correlational plot for highlyrelated features

Table 2 - Low correlated features set

Feature ID	Feature name	Correlation coefficient	Description
F21	id	0.040	ID number
F22	texture_mean	0.415	Standard deviation of gray-scale values
F23	smoothness_mean	0.359	Mean of local variation in radius lengths
F24	symmetry_mean	0.330	Mean of symmetry
F25	fractal_dimension_mean	-0.013	Mean of coastline approximation" – 1
F26	texture_se	-0.008	Standard error of texture (standard deviation of gray-scale values)
F27	smoothness_se	-0.067	Standard error of smoothness (local variation in radius lengths)
F28	compactness_se	0.293	Standard error of compactness (perimeter^2 / area - 1.0)
F29	concavity_se	0.254	Standard error of concavity (severity of concave portions of the contour)
F210	concave points_se	0.408	Standard error of concave points (number of concave portions of the contour)
F211	symmetry_se	-0.007	Standard error of symmetry
F212	fractal_dimension_se	0.078	Standard error of fractal dimension ("coastline approximation" - 1)
F213	texture_worst	0.457	Largest (mean of the three largest values) of texture
F214	smoothness_worst	0.421	Worst local variation in radius lengths
F215	symmetry_worst	0.416	Mean of the three largest values of symmetry
F216	fractal_dimension_worst	0.324	Largest of "coastline approximation" - 1

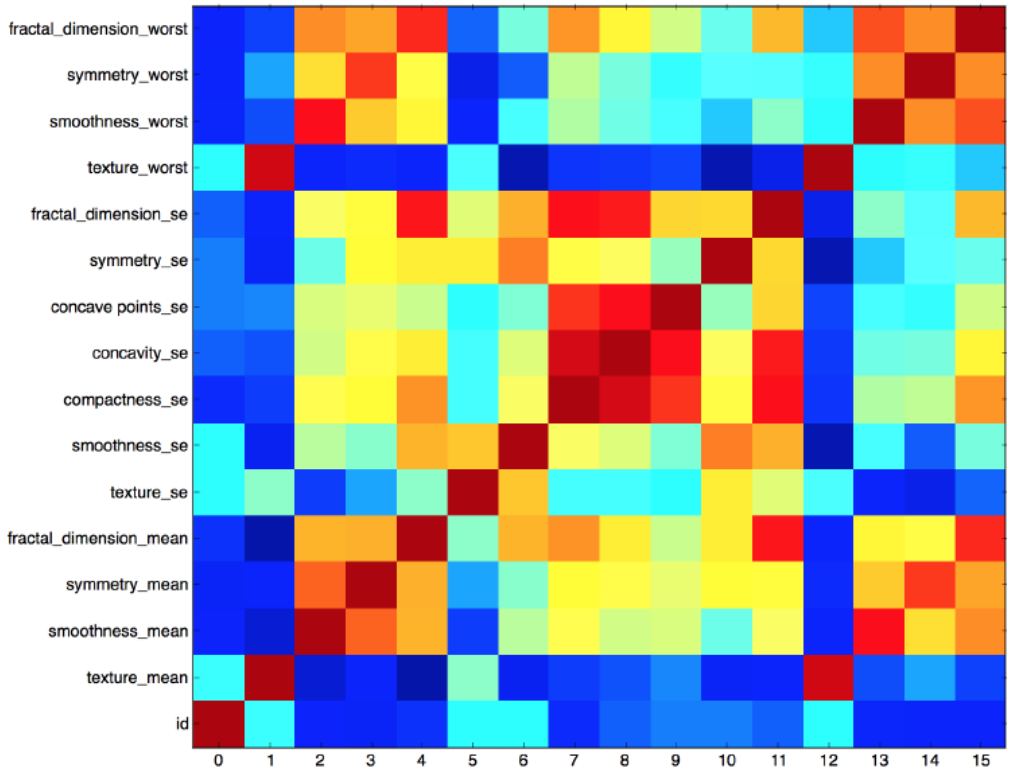

Figure 3 - Correlational plot for lowrelated features

III - literature review.

This section reviews several studies related to applications of machine learning algorithms for working with medical data especially related to cancer. It can be seen that a great variety of methods were used which reached high prediction and classification accuracies using the datasets generally taken from UCI-ML repositories.Zhongyu Pang and Lloyd, S.R (2008) developed an innovative signal classification method that is capable of differentiating subjects with sleep disorders which cause excessive daytime sleepiness (EDS) from normal control subjects who do not have a sleep disorder based on EEG and pupil size [2]. In another study, Kiyan et. al., trained Neural Network using back propagation and achieved an accuracy level on the test data of approximately 94% on breast cancer data [3]. The authors in this research study [4] presented BP-ANN attempt where they used 47 input features and achieved an accuracy of 95%.Moris et al. used logistic regression algorithm on heart diseases dataset. By applying various preprocessing techniques, he achieved in obtaining 77.0% of classification accuracy [5]. Further, Kamruzzaman et al. proposed a neural network ensemble based methodology for diagnosing of the heart disease diagnosis and achieved prediction accuracy over 80% [6]. Moreover, Das et al.[7] in 2008 applied genetic algorithm (GA) based Neuro Fuzzy Techniques for breast cancer identification and adaptive neuro fuzzy classifier has been introduced to classify the tumor mass in breast. So from the research studies above it can be seen the ml algorithms can be successfully applied in medical field.

IV-implementation, results and discussions.

Models simulations performed over scikit-learn ml framework for 5 different algorithms explained in Section 2.1 over breast cancer dataset. In order to reveal the true potential of algorithms the dataset has been divided into two parts shown in Table 1 and 2. The first part contains the highly correlated features where $r >= 0.5$ and second part is lower than $r < 0.5$.

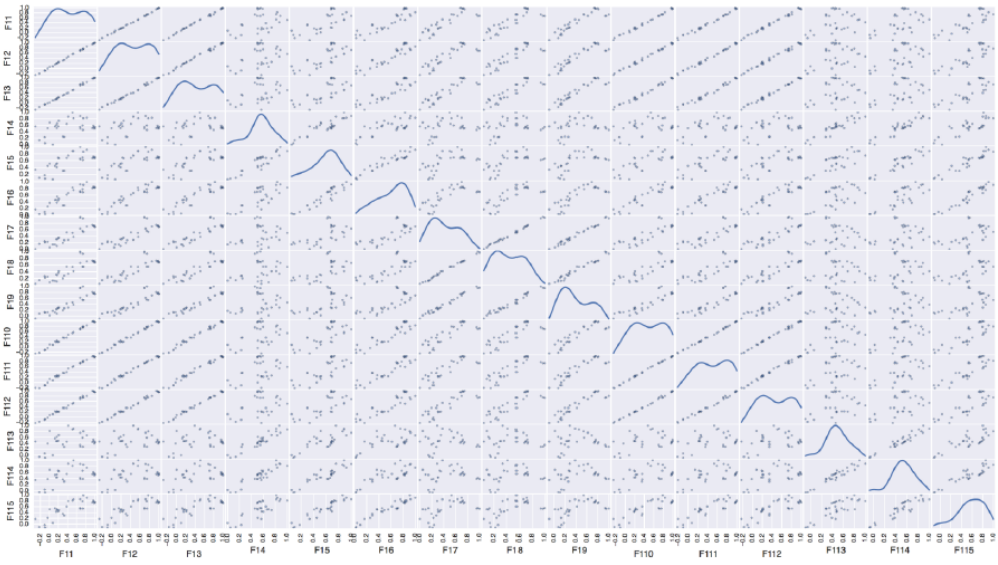

Figure 4 - Descriptive analysis for highly correlated features

The descriptive analysis studies performed for two sets and show in Figure 4 and 5. Based on this analysis the one of the features with correlation of 0.85 were removed from dataset.

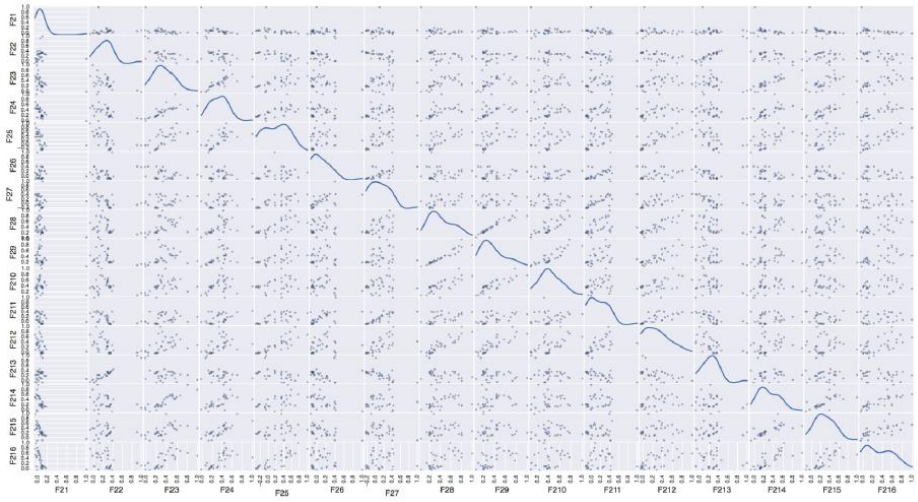

Figure 5 - Descriptive analysis for low correlated features

The accuracy scores before preprocessing are given in Figure 6. We can see that Logit and DT are performing the best. The standardization preprocessing technique gave the highest accuracy scores for ml algorithms.

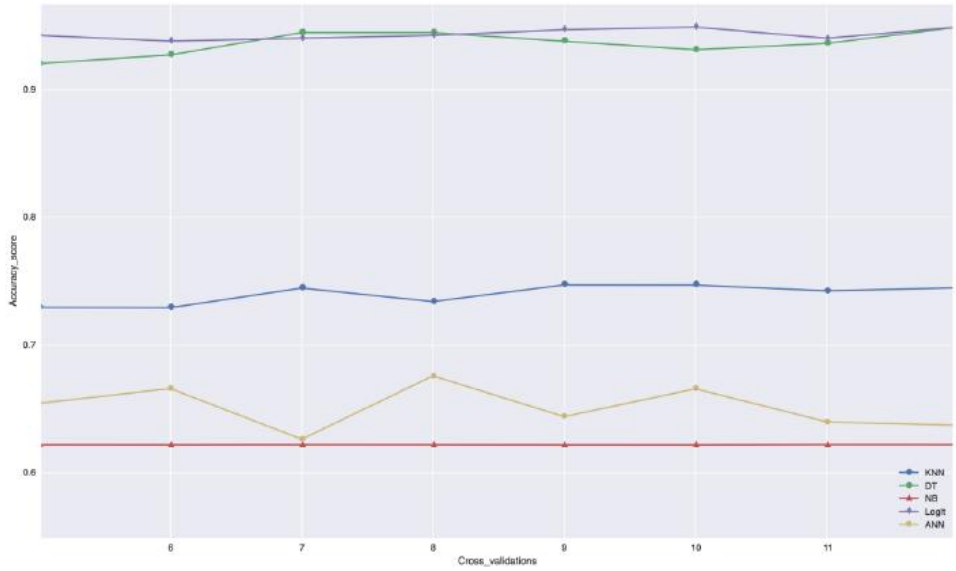

Figure 6 - Accuracy scores on not preprocessed data

There are other important concepts related to real-world applications where our data will not come naturally as a list of real-valued features. In these case, we will need to have methods to transform non real-valued features to real-valued ones. Besides, there are other steps related to feature standardization and normalization, are needed to avoid undesired effects regarding the different value ranges.

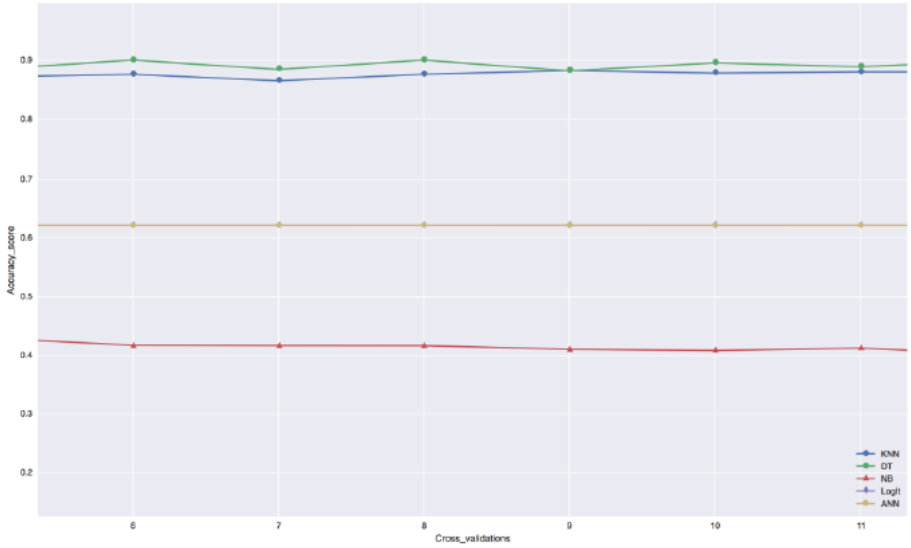

Figure 7 - Accuracy scores after normalization technique applied

After applying the normalization technique, we can see that there is an effect on accuracy scores. In this case KNN and DT shows the highest scores for over 90% but the accuracy scores for other drops shown in Figure 7.

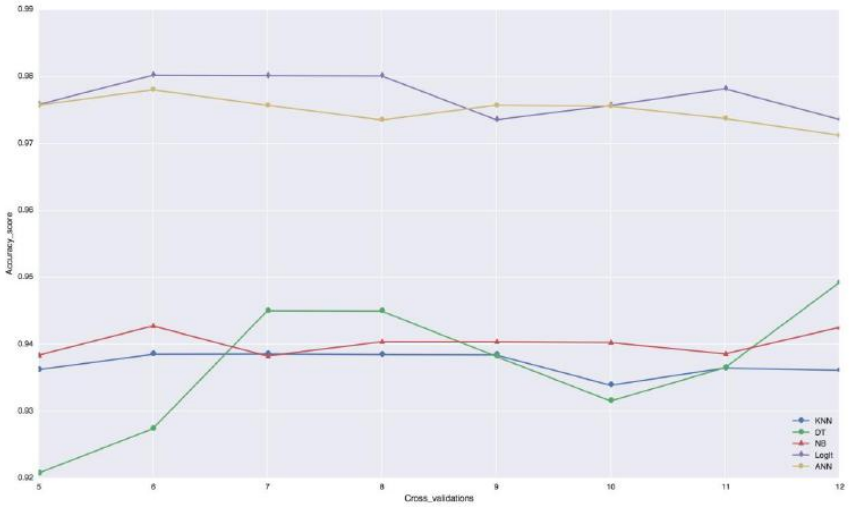

Figure 8 - Accuracy scores after standardization technique applied

The standardization technique best fits the Logit and ANN algorithms; we can see that the accuracy scores for them increases over 92% shown in Figure 8. So there is an effect of correlation and normalization techniques on accuracy scores.

Conclusions.

The breast cancer is one of the most common and deadly diseases in the world. The detection and diagnosis of breast cancer in its early stage is the key of its cure for women. In this research study we have analyzed the effect of different preprocessing techniques on ml algorithms accuracy scores. The study found that normalization increases scores for KNN and DT for over 90% but the accuracy scores for other drops. On the other hand, the standardization increases the scores for the Logit and ANN algorithms for over 92%. In conclusion we can say that before making any diagnostic assumptions several preprocessing techniques has to be applied and accuracy scores tested.

REFERENCES

1 Harrington, Peter. "Machine learning in action". - Greenwich, CT: Manning, 2012. - Vol. 5. – P. 88-96.
2 Derong Liu; Zhongyu Pang; Lloyd S.R, "A Neural Network Method for Detection of Obstructive Sleep Apnea and Narcolepsy" // Based on Pupil Sizeand EEG. – 2008. - V.19, I.2. – P. 126-169.
3 Kiyan, T., and Yildirim, T. (2003). Breast Cancer Diagnosis Using Statistical Neural Networks // International XII. Turkish Symposium on Artificial Intelligence and Neural Networks. University Besiktas, Istanbul. – Turkey: 2003. – P. 51-56.
4 Seker .H., Odetao M.,Petroric D. and Naguib R.N.G.(1994)- "A fuzzy logic based method for prognostic decision making in breast and prostrate cancers" // Biomedicine (IEEE transactions). - 2003. - №73. – P. 88-96.
5 S. Haykin. Neural Networks: A Comprehensive Foundation. - New York: 1994. – 523 p.
6 Morise, A. P., Detrano, R., Bobbio, M., & Diamond, G. A. (1992). Development and validation of a logistic regression-derived algorithm for estimating the incremental probability of coronary artery disease before and after exercise testing // Journal of the American College of Cardiology. – 1992. - №20(5). – P. 1187-1196.
7 S. M. Kamruzzaman, Ahmed Ryadh Hasan, Abu Bakar Siddiquee and Md. EhsanulHoqueMazumder // Medical diagnosis using neural network, ICECE 2004, 28-30 December 2004. - Dhaka, Bangladesh: 2004. – P. 28-34.
8 Arpita Das and Mahua Bhattacharya, GA based Neuro Fuzzy Techniques for breast cancer Identification // IEEE. – 2008. - №2. – P. 978-986
9 Hongmin Zhang, Xuefeng Dai, The Application of Fuzzy Neural Network in Medicine-A Survey // International Conference on Biological and Biomedical Sciences Advances in Biomedical Engineering. – 2012. - Vol.9. – P. 52-56.

Ш. Шамилулу[1], У. Диакбарова[2]
[1]Сулеймен Демирел университеті, Алматы, Қазақстан040900
Компьютерлік ғылым кафедрасы
[2]С.Ж. Асфендияров атындағы Қазақ ұлттық медицина университеті 035000
Молекулалық биология және медициналық генетика кафедрасы

SCIKIT-LEARNMLFRAMEWORK МАШИНАСЫНЫҢ АЛГОРИТМІН ҚОЛДАНУ АРҚЫЛЫ СҮТ БЕЗІ ҚАТЕРЛІ ІСІГІНЕ ШАЛДЫҚҚАН НАУҚАСТАРДЫ САРАЛАУ

Түйін: Мақаладақалыпқа келтіру әдістерінің әсері зерттелді.UCI оқу машинасырепозиториінен алынған бес түрлі басқарылатын оқыту машинасының алгоритмдерісүт безі қатерлі ісігінің деректер жинағы үшін пайдаланылады және алынған нәтижелерді салыстыру жүргізілді. Емдеудің алдын-алудың әртүрлі әдістері логистикалық регрессиясы және АНН жоғары өнімділігімен қамтамасыз етілген кезде жіктелудің90% - дан астам қатесіз болатындығы зерттеу жұмысының нәтижесі көрсетеді. Нақты клиникалық көрсеткіштері бар ұсынылып отырған әдіс сүт безі қатерлі ісігінің медициналық диагностика мәселелері үшін қолданылу мүмкін.
Түйінді сөздер: сүт безі қатерлі ісігі, компьютерлік оқыту алгоритмдері, деректерді жіктеу, компьютерлік болжау және диагностика

Ш. Шамилулу[1], У. Диакбарова[2]
[1]Университет Сулеймана Демиреля, Алматы, Казахстан040900
Кафедра компьютерных наук
[2]Казахский национальный медицинский университет имени С.Д.Асфендиярова035000
Кафедра молекулярной биологии и медицинской генетики

ИДЕНТИФИКАЦИЯ ПАЦИЕНТОВ С РАКОМ МОЛОЧНОЙ ЖЕЛЕЗЫ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ ОБУЧЕНИЯ МАШИНЫ SCIKIT-LEARN ML FRAMEWORK

Резюме: В данной статье исследуется пять различных контролируемых алгоритмов машинного обучениядля набора данных о раке молочной железы, и сравниваются полученные результаты.Исследование показывает, что различные методы предварительной обработки могут повысить точность диагностики более чем на 90%, когда высокая производительность предоставляется логистической регрессии и ANN.Предлагаемый методвместе с клиническими данными может быть использован для диагностики медицинских проблем рака молочной железы.
Ключевые слова: рак молочной железы, алгоритмы машинного обучения, классификация данных, компьютерный прогноз и диагностика