➔ Regular Research Paper – NS

Classification for Breast Cancer Diagnosis

Sena GORAL
Mehmet Akif Ersoy University, Turkey
scelik@mehmetakif.edu.tr

Abstract

Breast cancer is one of the most common types of cancer in the world. It is the type of cancer with the highest death rate from cancer in women. As with all cancer types, early diagnosis is very important in breast cancer. Diagnosis of the disease and interpretation of the tests by experts can be a long process. Machine learning techniques have become an important aid in disease diagnosis. Machine learning can get very fast and successful results even in large and complex data sets. In this study, 4 different classification methods were examined to help in the early diagnosis of breast cancer. these four methods; logistic regression, KNN, random forest and SVM. As a result of the examinations and studies, these methods were compared. As a result, the most successful results were achieved with logistic regression and SVM methods.

Keywords: *Breast Cancer, classification, classification algorithm, machine learning, artificial intellince.*

1. INTRODUCTION

Breast cancer is one of the serious death threats especially for women. Early diagnosis is the most important way to reduce the death rate. It can be seen as a small tumor or mass in the breast tissue. The severity of the condition is determined according to whether the mass is malignant or benign. Different machine learning algorithms are used for the diagnosis of breast cancer. Today, thanks to the advancing technology, much more accurate and faster analyzes have become possible by leaving the decision-making abilities to the computer. Machine learning techniques provide great convenience to experts thanks to their successful classification and diagnostic capabilities.

Douangnolulack et al. aimed to produce the best results with minimal classification using PCA. It has been seen that the J48 decision tree classification method produces the best results[1]. Amrane et al. used two different classifications. These are NB AND KNN. KNN gave higher accuracy when comparing the two classifiers[2]. Yang et al. presented an effective approach for breast cancer diagnosis using the SVM method. They found an accuracy rate of 98.22% with this approach[3]. Bazazeh and Shubair compared 3 different machine learning techniques in their study. The classification accuracy rate of the SVM method was calculated as 96.60%, the accuracy rate of the Random Forest method was 99.90% and the accuracy rate of the Bayesian Networks method was calculated as 99.10%[4]. Delen et al. developed a model on the breast cancer dataset using two data mining algorithms. In the study, the decision trees were 93.6% and the artificial neural network model 91.2% accuracy[5]. Khan et al. classified breast cancer data with fuzzy decision trees and obtained more successful results than independent classification methods[6].

In this study, it is to analyze for the diagnosis of breast cancer disease. Thus, it will be beneficial to eliminate the loss of time in the vitally important diagnosis process. Providing a faster transition to the treatment phase will be a very important development for the patient.

2. MATERIAL METHOD

With the data set used in the study, 4 different classification methods are tested. First of all, the data set was examined. In the data set used, there are features belonging to 569 different people. In addition, it has been determined that 32 of the feature values in the data set are missing values. These values have been adjusted so that they do not affect the accuracy of the classification methods. In addition, the benign or malignant status of the individuals is also specified in the data set. This situation is shown in Figure 1 within the data set to be used in classification algorithms.

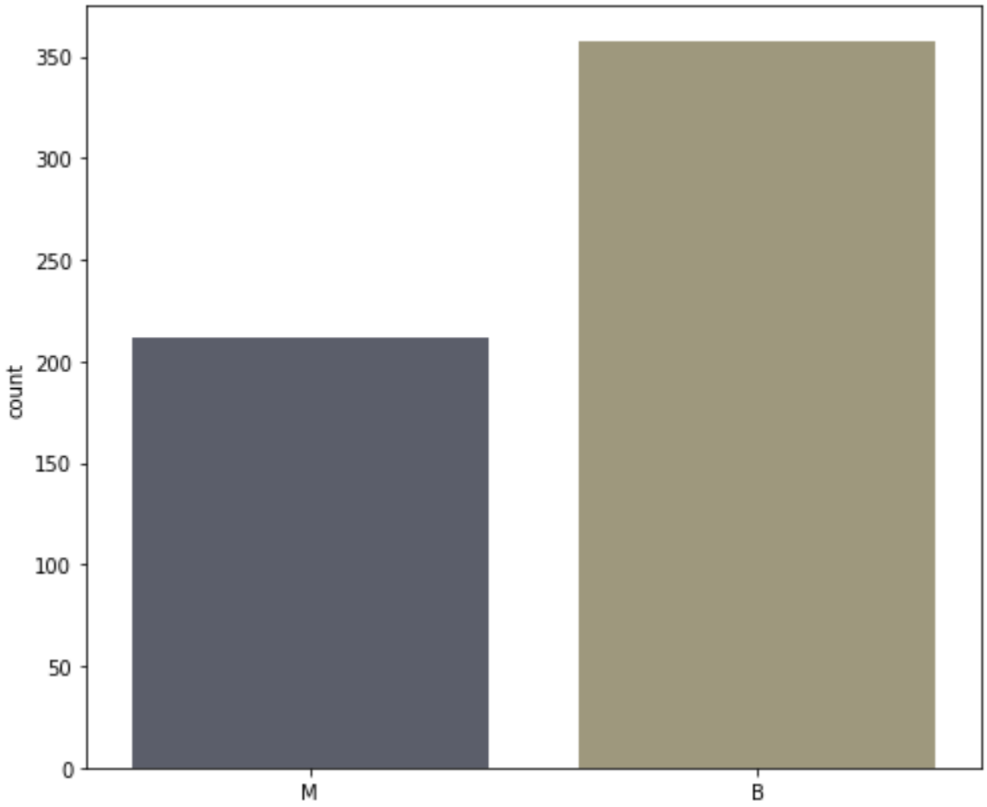

Figure 1. Diagnosis

The heat map results were evaluated for data set analysis. When the heat map was examined, it was seen that there were many negative correlations. 30 features were identified in the heat map. Calculations were made for each of these 30 features. As a result, the three features with the highest standard error rate and less useful for study are radius_mean, radius_se, and Radius_worst. The heatmap is shown in Figure 2.

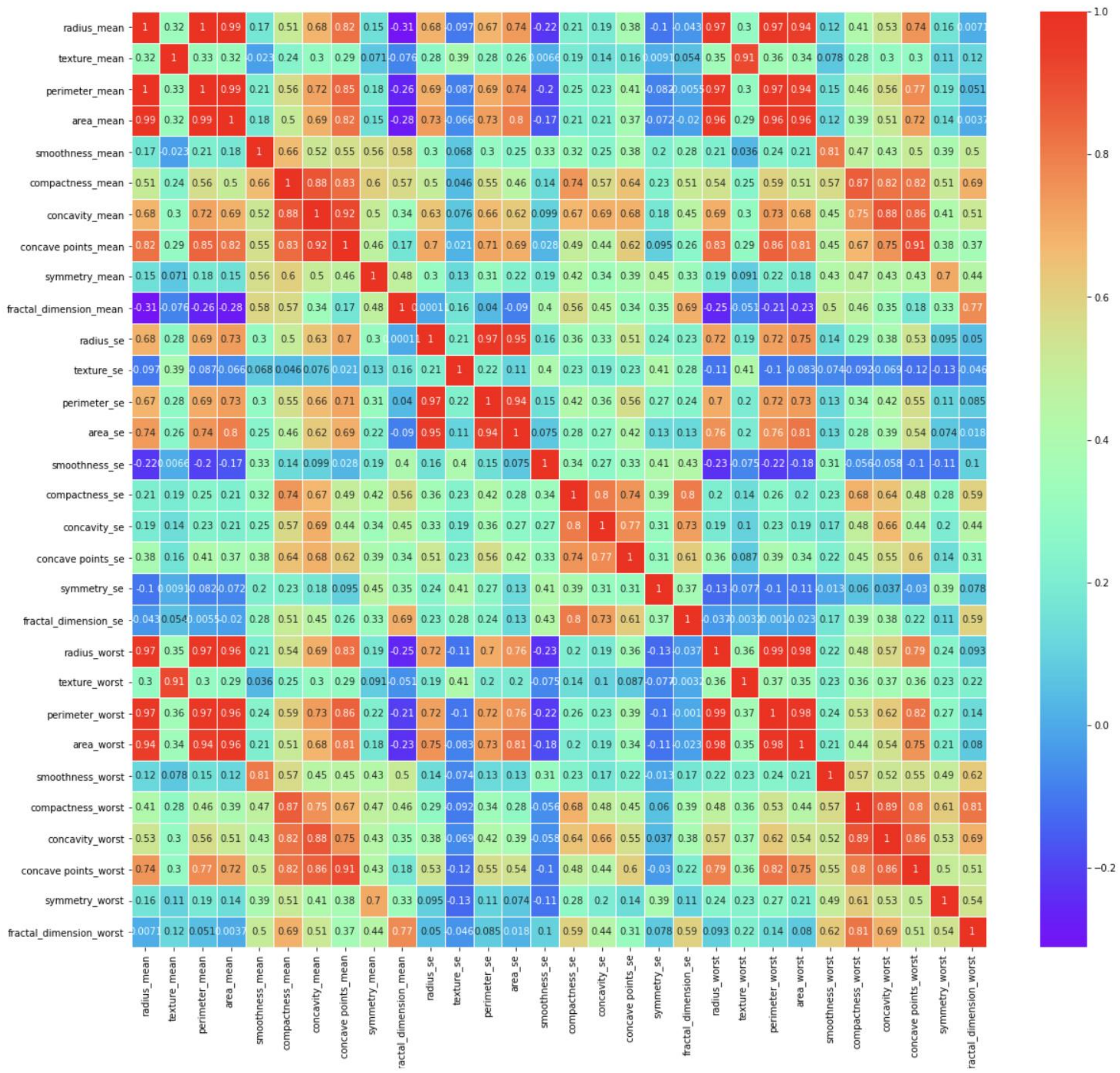

Figure 2. HeatMap

Considering the values in the data set, the number of malignant was determined as 357 and benign 212. For this data set consisting of 569 people, splitting was done for train and test operations. The data set was split as 455 people for the train set and 114 people for the test set. Scaling was done with StandartScaler. This was a convenient way for the distribution to approach normal. After the data set preparation was completed, the classification algorithm results were calculated. The necessary definitions and abbreviations for results for logistic regression, k nearest neigbors , random forest and support vector machines(SVM) are as follows:

Accuracy=(TP+TN)/(TP+FP+FN+TN)

Precision=TP/(TP+FP)

Recall=TP/(TP+FN)

F1-Score=2X(PrecisionXRecall)/ (Precision+Recall)

Information:

TP = True positive;

TN = True Negative;

FP = False Positive;

FN = False Negative.

Confusion matrix were determined for all models and calculations were made. The values for the logistic regression are shown in Table 1. Accuracy of the logistic regression model is 0.9824561403508771.

Table 1. Classification Report(Logistic Regression)

	Precision	Recall	F1-Score	Support
0	0.97	1.00	0.99	71
1	1.00	0.95	0.98	43
Accuracy			0.98	114
Macro Avg	0.99	0.98	0.98	114
Weighted Avg	0.98	0.98	0.98	114

Values between 1 and 41 were checked for K nearest neigbors classification. It was determined which values showed the lowest mean error among these values. The error rates of the values are shown in Figure 3. From this graph, k value of 9, 34, 35,36, 40 and 41 seem to Show the lowest mean error. So using one of these values.

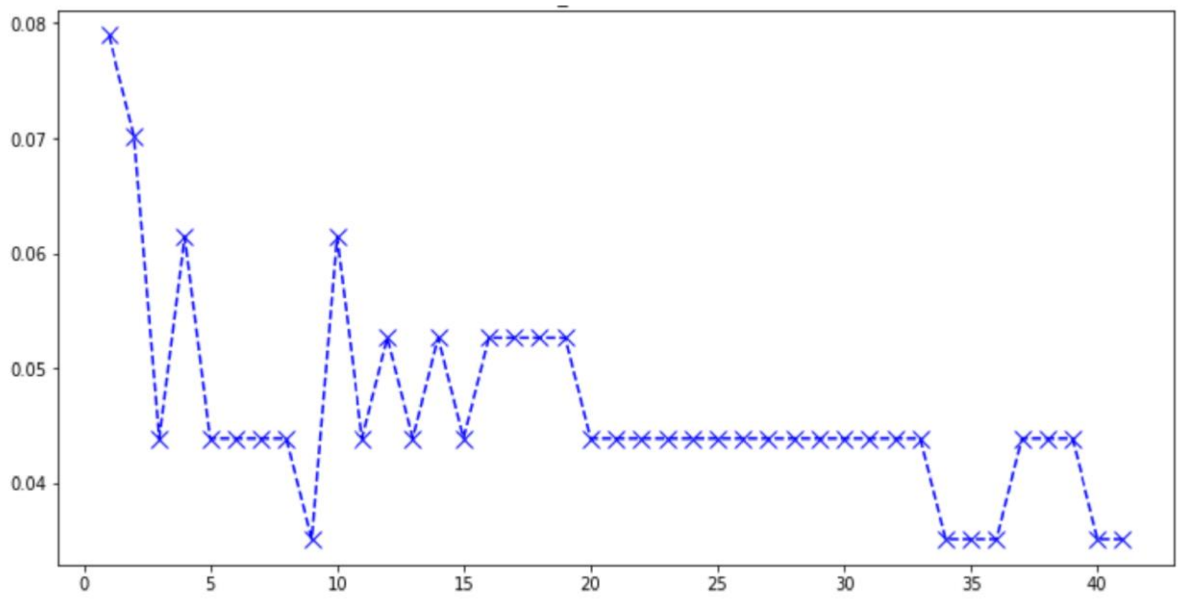

Figure 3. Error rate vs k value

The results for K nearest neigbors classification are shown in Figure 4. Accuracy of k neigbors classifier model is 0.9649122897917544.

Table 2. Classification Report(k Nearest Neigbors Classification)

	Precision	Recall	F1-Score	Support
0	0.97	0.97	0.97	71
1	0.95	0.95	0.95	43
Accuracy			0.96	114
Macro Avg	0.96	0.96	0.96	114
Weighted Avg	0.96	0.96	0.96	114

Random forest is a flexible, easy-to-use machine learning algorithm. It is one of the most commonly used algorithms. The results of the algorithm applied for the system are shown in Table 3. Accuracy of Random Forests Model is 0.9649122807017544.

Table 3. Classification Report(Random Forest)

	Precision	Recall	F1-Score	Support
0	0.96	0.99	0.97	71
1	0.98	0.93	0.95	43
Accuracy			0.96	114
Macro Avg	0.97	0.96	0.96	114
Weighted Avg	0.97	0.96	0.96	114

Finally, necessary procedures for SVM have been done. SVM is one of the supervised learning methods generally used for classification problems. The results are shown in Table 4. Accuracy of SVM model is 0.9824561403508771.

Table 4. Classification Report(SVM)

	Precision	Recall	F1-Score	Support
0	0.97	1.00	0.99	71
1	1.00	0.95	0.98	43
Accuracy			0.98	114
Macro Avg	0.99	0.98	0.98	114
Weighted Avg	0.98	0.98	0.98	114

3. CONCLUSION

Breast cancer is a disease that affects many people. As with almost all diseases, early diagnosis is of great importance. In this study, logistic regression, KNN, random forest and SVM models, which are important classification algorithms, were compared. In fact, all models produced successful results. However, the most suitable results belong to the logistic regression and SVM models. The successful results of the study will also facilitate comparison with other studies. Cancer is an important disease. Early intervention reduces the risk of death. In the future, the diagnostic results can be compared with the visual data sets of the study, and the accuracy can be checked. The results for the classification methods used in the study are shown in Figure 4.

The accuracy of Logistic Regression Model is 98.24%,

The accuracy of KNN model is 96.49%

The accuracy of Random Forest Model is 96.49%

The accuracy of SVM Model is 98.24%.

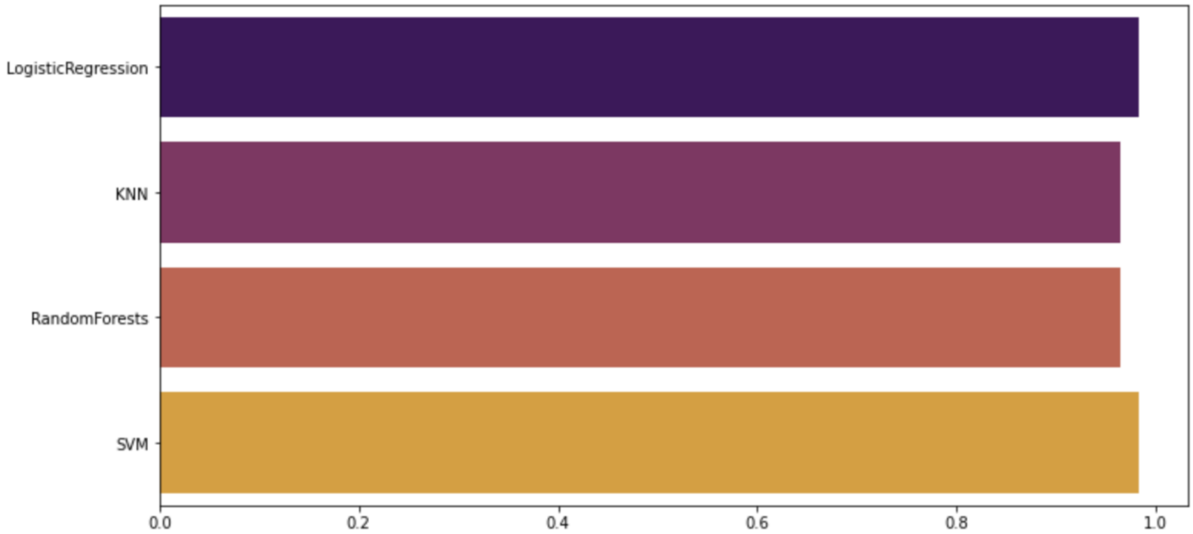

Figure 4. Final Results

REFERENCES

[1] Douangnoulack, P., & Boonjing, V. (2018). Building minimal classification rules for breast cancer diagnosis. In 2018 10th International Conference on Knowledge and Smart Technology (KST) (pp. 278-281). IEEE.

[2] Amrane M., Oukid S., Gagaoua I., Ensarİ T., Breast cancer classification using machine learning. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-4). IEEE. 2018.

[3] Yang, X., Peng, H., & Shi, M. (2013, August). SVM with multiple kernels based on manifold learning for breast cancer diagnosis. In *2013 IEEE International Conference on Information and Automation (ICIA)* (pp. 396-399). IEEE.

[4] Bazazeh, D., & Shubair, R. (2016, December). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In *2016 5th international conference on electronic devices, systems and applications (ICEDSA)* (pp. 1-4). IEEE.

[5] Delen, D., Walker, G., Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127, 2005.

[6] Khan, M. U., Choi, J. P., Shin, H., Kim, M., Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare, In Engineering in Medicine and Biology Society 30th Annual International Conference of the IEEE , 5148-5151, 2008.

JOMUDE
http://www.jomude.com