Prediction of breast cancer using machine learning algorithms on different datasets

Predicción del cáncer de mama utilizando algoritmos de aprendizaje automático en diferentes conjuntos de datos

Previsão do câncer de mama usando algoritmos de aprendizado automático em diferentes conjuntos de dados

Ömer Çağri Yavuz[1]
M. Hanefi Calp[2]
Hazel Ceren Erkengel[3]

Received: October 5th, 2022
Accepted: December 17th, 2022
Available: January 22th, 2023

How to cite this article:
Ö. Çağri Yavuz, M. Hanefi Calp, H. Ceren Erkengel, "Using Machine Learning Algorithms on Different Datasets," *Revista Ingeniería Solidaria,* vol. 19, no. 1, 2023.
doi: https://doi.org/10.16925/2357-6014.2023.01.08

Research article. https://doi.org/10.16925/2357-6014.2023.01.08

[1] Department of Management Information Systems. Karadeniz Technical University.
Email: omercagriyavuz@ktu.edu.tr
ORCID: https://orcid.org/0000-0002-6655-3754

[2] Department of Management Information Systems. Ankara Hacı Bayram Veli University.
Email: hanefi.calp@hbv.edu.tr
ORCID: https://orcid.org/0000-0001-7991-438X

[3] Department of Management Information Systems. Karadeniz Technical University.
Email: hazelceren@ktu.edu.tr
ORCID: https://orcid.org/0000-0002-7153-9375

Abstract

Introduction: The research paper "Prediction of Breast Cancer using Machine Learning Algorithms on Different Datasets", was developed at Karadeniz Technical University in the year 2022.

Problem: Breast cancer is a disease that is becoming more and more common, day by day, causing emotional and behavioral reactions and having fatal consequences if not detected early. At this point, traditional methods are insufficient, especially in early diagnosis. This study aims to predict breast cancer by using machine learning (ML) algorithms on different datasets and demonstrates the applicability of these algorithms.

Methodology: Algorithm performances were compared on balanced and unbalanced datasets, taking into account the performance metrics obtained in applications on different datasets. In addition, a model based on the Borda Voting method was developed by including the results obtained from four different algorithms (NB, KNN, DT, and RF) in the process.

Originality and Limitations of the Research: In the model developed within the scope of the study, the result values obtained from different algorithms such as NB, KNN, DT and RF were combined; the objective being to increase the performance of the model with this process, which is based on the Borda Voting method.

Results: The prediction values obtained from each algorithm were written in different columns on the same spreadsheet and the most repetitive value was accepted as the final result value. The developed model was tested on real data consisting of 60 records and the results were analyzed.

Conclusion: When the results were examined, it was seen that greater performance was obtained with the proposed RF model compared to similar studies in the literature. Finally, the prediction results obtained with the developed model revealed the applicability of ML algorithms in the diagnosis of breast cancer.

Keywords: Breast Cancer, Classification Algorithms, Machine Learning, Unbalanced Dataset

Resumen

Introducción: El trabajo de investigación "Predicción del cáncer de mama utilizando algoritmos de aprendizaje automático en diferentes conjuntos de datos", se desarrolló en la Universidad Técnica de Karadeniz en el año 2022.

Problema: El cáncer de mama es una enfermedad cada vez más común, día a día, provocando reacciones emocionales y conductuales y con consecuencias fatales si no se detecta a tiempo. En este punto, los métodos tradicionales son insuficientes, sobre todo en el diagnóstico precoz. Este estudio tiene como objetivo predecir el cáncer de mama mediante el uso de algoritmos de aprendizaje automático (ML) en diferentes conjuntos de datos y demuestra la aplicabilidad de estos algoritmos.

Metodología: se compararon los rendimientos de los algoritmos en conjuntos de datos equilibrados y no equilibrados, teniendo en cuenta las métricas de rendimiento obtenidas en aplicaciones en diferentes conjuntos de datos. Además, se desarrolló un modelo basado en el método Borda Voting al incluir en el proceso los resultados obtenidos de cuatro algoritmos diferentes (NB, KNN, DT y RF).

Originalidad y Limitaciones de la Investigación: En el modelo desarrollado en el marco del estudio se combinaron los valores de los resultados obtenidos de diferentes algoritmos como NB, KNN, DT y RF; el objetivo es aumentar el rendimiento del modelo con este proceso, que se basa en el método Borda Voting.

Resultados: Los valores de predicción obtenidos de cada algoritmo se escribieron en diferentes columnas en la misma hoja de cálculo y se aceptó el valor más repetitivo como valor final del resultado. El modelo desarrollado se probó en datos reales que constaban de 60 registros y se analizaron los resultados.

Conclusión: Cuando se examinaron los resultados, se observó que se obtuvo un mayor rendimiento con el modelo de RF propuesto en comparación con estudios similares en la literatura. Finalmente, los resultados de predicción obtenidos con el modelo desarrollado revelaron la aplicabilidad de los algoritmos de ML en el diagnóstico del cáncer de mama.

Palabras clave: cáncer de mama, algoritmos de clasificación, aprendizaje automático, conjunto de datos no balanceado.

Resumo

Introdução: O trabalho de pesquisa "Previsão de câncer de mama usando algoritmos de aprendizado de máquina em diferentes conjuntos de dados", foi desenvolvido na Universidade Técnica de Karadeniz no ano de 2022.

Problema: O câncer de mama é uma doença cada vez mais comum, dia após dia, causando reações emocionais e comportamentais e com consequências fatais se não for detectado a tempo. Nesse ponto, os métodos tradicionais são insuficientes, principalmente no diagnóstico precoce. Este estudo visa prever o câncer de mama usando algoritmos de aprendizado de máquina (ML) em diferentes conjuntos de dados e demonstra a aplicabilidade desses algoritmos.

Metodologia: Os desempenhos dos algoritmos em conjuntos de dados balanceados e não balanceados foram comparados, levando em consideração as métricas de desempenho obtidas em aplicações em diferentes conjuntos de dados. Além disso, foi desenvolvido um modelo baseado no método Borda Voting, incluindo no processo os resultados obtidos a partir de quatro algoritmos diferentes (NB, KNN, DT e RF).

Originalidade e Limitações da Pesquisa: No modelo desenvolvido no âmbito do estudo, foram combinados os valores dos resultados obtidos de diferentes algoritmos como NB, KNN, DT e RF; o objetivo é aumentar a performance do modelo com este processo, que é baseado no método Borda Voting.

Resultados: Os valores de previsão obtidos de cada algoritmo foram escritos em diferentes colunas na mesma planilha e o valor mais repetitivo foi aceito como valor final do resultado. O modelo desenvolvido foi testado em dados reais constituídos por 60 registros e os resultados foram analisados.

Conclusão: Ao examinar os resultados, observou-se que o modelo de RF proposto apresentou desempenho superior a estudos semelhantes na literatura. Por fim, os resultados de predição obtidos com o modelo desenvolvido revelaram a aplicabilidade dos algoritmos de ML no diagnóstico do câncer de mama.

Palavras-chave: câncer de mama, algoritmos de classificação, aprendizado de máquina, conjunto de dados desbalanceado.

1. INTRODUCTION

The data transferred to the digital environment increases with the development of technology and the widespread use of social media. This data, called garbage, is processed and used to solve various problems for some users. The data mining methods used to make sense of these data demonstrate the importance of data processing and pattern recognition processes. Based on data mining and high success rate estimates using experience, machine learning (ML) algorithms can adapt to changes and solve complex problems, which have been developed day by day and used in many fields [1]. In medical informatics, ML algorithms are used in many studies, especially in which predictions for disease diagnosis are made. ML algorithms are often used to predict diseases that are especially important for the early diagnosis of breast cancer [2]. Based on the World Health Organization (WHO), breast cancer is the world's most dominant kind of cancer, with a startling 2.26 million patients diagnosed in 2020. It is one of the most frequently diagnosed cancer types in 2020 and ranks fifth in cancer-related deaths [3]. Early detection of breast cancer is vital to the survival of patients.

The incidence of breast cancer is more common in western countries such as the USA than in African and Asian countries. This deadly disease is growing at an annual rate of 0.5% worldwide. This increase is more in Asian countries, around 3-4%. Also, many researchers reported that Indian women suffer from this disease at a very young age compared to other developed countries. It is claimed that Indian women are more likely to have a larger tumor size and be malignant [4].

The development of artificial intelligence technologies brings many solutions and alternatives to health. Thanks to ML and artificial neural networks, it can categorize patient data and develop predictive models that can determine the probability of patients getting any disease. The application of artificial intelligence in health enables diseases to be detected simultaneously and at a low cost more quickly [5]. In this context, K Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT) and Naive Bayes (NB) algorithms were applied on different datasets to demonstrate the applicability of ML algorithms in the diagnosis of breast cancer.

The use of ML algorithms, in line with the performance metrics obtained in the applications using three different datasets, will contribute to the decision process of the healthcare personnel in solving the problems related to the diagnosis of the disease. In addition, the utilization of ML algorithms to solve similar problems in medical informatics will contribute to the field, together with the elimination of the need for datasets in the field of health.

This study aims to demonstrate the applicability of ML algorithms in diagnosing breast cancer by using three different datasets. In this context, the existing literature is presented in the second chapter. In the third chapter, ML algorithms and the developed model are explained. Performance metrics and research findings are given in the fourth chapter. In the fifth chapter, experimental results are given and compared with similar studies in the literature. Finally, the conclusions and recommendations are given in the sixth chapter.

1.1. Literature review or research background

It is challenging to detect breast cancer in the early stages as the disease is asymptomatic. The disease becomes complex and costly to treat in later stages, leading to increased mortality rates [6]. The most important part of cancer detection is distinguishing between benign and malignant tumors. Thanks to ML, it is possible to make this distinction in a short time and accurately. In this direction, studies in which ML algorithms are used to diagnose the disease are frequently encountered. This part of the study includes studies in which ML algorithms are used to diagnose breast cancer.

Most researchers have addressed this issue. For example, Sahan et al. (2007) used the Wisconsin Breast Cancer Dataset, and their method is a hybrid of the KNN algorithm and a data reduction stage. They reached 99.14% classification accuracy. The results of the application made with the Wisconsin Breast Cancer Dataset, which is frequently used in the literature, were compared with other studies. It was emphasized that the accuracy rate obtained is the highest one in the literature, to the date of publication [7].

Eleyan (2012) compared two different ML classifiers: NB and KNN, classifying breast cancer and estimating their accuracy using cross-validation. Four new features were extracted from nine features in the dataset used within the scope of the study. It was stated that up to a 5% performance increase has been achieved with this process. The results showed that KNN had the highest accuracy, about 97.51%, with the lowest error rate resulting from the NB classifier [8, 9].

Ahmad et al. (2013) used data mining techniques to investigate risk factors to predict breast cancer. They used DT, SVM, and ANN ML algorithms to predict breast cancer recurrence to find which method performs better. The results displayed that SVM is the best predictive classifier, with 95% accuracy, while the ANN and DT (C4.5) had 94% and 93% accuracy, respectively [10].

Williams et al. (2015) used NB and DT (J8) techniques for the diagnosis of breast cancer. In the applications made, the DT algorithm showed higher performance than NB. As a result of the study, it was stated that the DT algorithm is a more effective classification algorithm in the diagnosis of breast cancer [11].

Mejia et al. (2015) used Thermogram images to detect breast cancer. It has been stated that this method is cheaper and safer than other methods. In addition, it was stated that detection was made at an earlier stage compared to other methods. Within the scope of the study, 18 cases were selected, including nine abnormal and nine normal cases. The KNN algorithm was used in the applications. As a result of the study, a performance of 88.88% for abnormal cases and 94.44% for normal cases was obtained [12, 13].

Zand (2015) integrated NB, ANN, and C4.5 algorithms to diagnose and indicate breast cancer and the survivability rate of breast cancer patients. The SEER Public Use

Database dataset was used, and C4.5 gave the highest accuracy with 86.7%. As a result of the study, it was stated that the performances obtained from the classification algorithms were quite acceptable. In this direction, it has been emphasized that with classification algorithms, experts can be assisted in early diagnosis [14].

Bevilacqua et al. (2016) presented a CAD system to support radiologists in classifying chest lesions from MR images. The materials of the study consist of the features revealed by processing the MR images. Within the scope of the study, various ANN topologies were tested using 100 random permutations. The average accuracy of 89.77% was obtained through an ANN optimized for GA. As a result of the study, the success in the classification of malignant lesions was emphasized [15].

Hussain et al. (2018) tried to distinguish between healthy tissue and cancerous tissue mammograms and applied the approach of DT, Support Vector Machine (SVM), and Bayes to do this. While doing this, they used statistical parameters such as positive predictive value, negative predictive value, sensitivity, and specificity. More than 90% of success was achieved in the study. As a result of the achievements, the applicability of ML algorithms has been demonstrated in similar studies [16].

Amrane et al. (2018) used KNN and NB classification algorithms to measure the classification performance of breast cancer. This model was trained using 683 breast cancer dataset samples. The maximum accuracy rate (97.51%) obtained from the study was obtained with the KNN algorithm. As a result of the study, it is stated that the working time for the KNN algorithm will increase if the dataset is large [17].

Bayrak et al. (2019) compared two ML methods (ANN and SVM) for breast cancer detection in their study. Based on the performance metrics of the applied ML techniques, they achieved an accuracy rate of 96.9957%, in another study where different algorithms and breast cancer prediction models were developed using RB, NB, and DT algorithms. It was emphasized that the most efficient model with an accuracy rate of 97.36% was NB [18].

Temesgen Abera Asfaw (2019) analyzed the performance of DT, Logistic Regression (LR), NB, and KNN for detecting breast cancer using the UCI Wisconsin breast cancer dataset. In applications without using standardization, low performance was obtained for LR and NB. However, it was stated that LR showed higher performance when using standardization. It showed LR to provide the best classification accuracy of 96.93 % [19, 20].

Yadav et al. (2019) compared the accuracy, precision, and recall of six ML techniques (ANN, NB, RF, DT, SVM, and KNN). They used the Wisconsin dataset and noticed that SVM and RF achieved the highest accuracy of 97.2%, NB Classifier had the highest precision and recall of 97.2% & 97.1% [21].

Shamrat et al. (2020) used six algorithms for cancer disease prediction. The results demonstrated that the DT and LR reached the highest precision (97%). NB achieved the most heightened sensitivity, and it is 100%. Moreover, NB also achieved the worst specificity (92%). Considering the F1 measure, all classifiers show the same performance, above 95% [22].

Assegie (2021) used the grid search method to determine the best k-nearest neighbor (KNN) settings. Their research demonstrated that parameter adjustment had a significant result on the model's performance. They displayed that it is conceivable to get 94.35% accuracy by fine-tuning the settings, whereas the default KNN reached approximately 90% accuracy [23, 24].

Muhtadi (2022) used three algorithms to classify breast tumors, malignant or benign, KNN with SVM, and RF. They obtained a maximum classification accuracy of 93.01%, a sensitivity of 94.62%, and a specificity of 91.4%. To reduce values, they applied SMOTE and hybrid SMOTE-Tomek sampling techniques on the dataset. To provide a completely balanced scenario, SMOTE increased the number of malignant samples from 26 to 104. The SMOTE-Tomek procedure reduced the number of benign samples from 104 to 93 and increased the number of positive samples from 26 to 93, again providing a completely balanced scenario [25].

2. MATERIALS AND METHODS

In the scope of the study, NB, KNN, DT, and RF algorithms were applied to three different datasets that included breast cancer findings and were classified as benign-malignant. The original dataset handled in the first application consisted of 569 records. Data shared by William H. Wolberg were collected at the Wisconsin hospital [26]. Different applications/trials were realized with datasets consisting of 30 inputs and 1 output value. An unbalanced dataset was created by randomly selecting records in various classes from this dataset. One of the resampling methods, SMOTE, was applied to this dataset where the class distributions were unbalanced. The aim was to eliminate the imbalance in the dataset and examine the effect of the imbalance in the dataset on performance. Performance metrics obtained from applications made with three different datasets were compared. The flow chart for the applications made are given in Fig. 1.

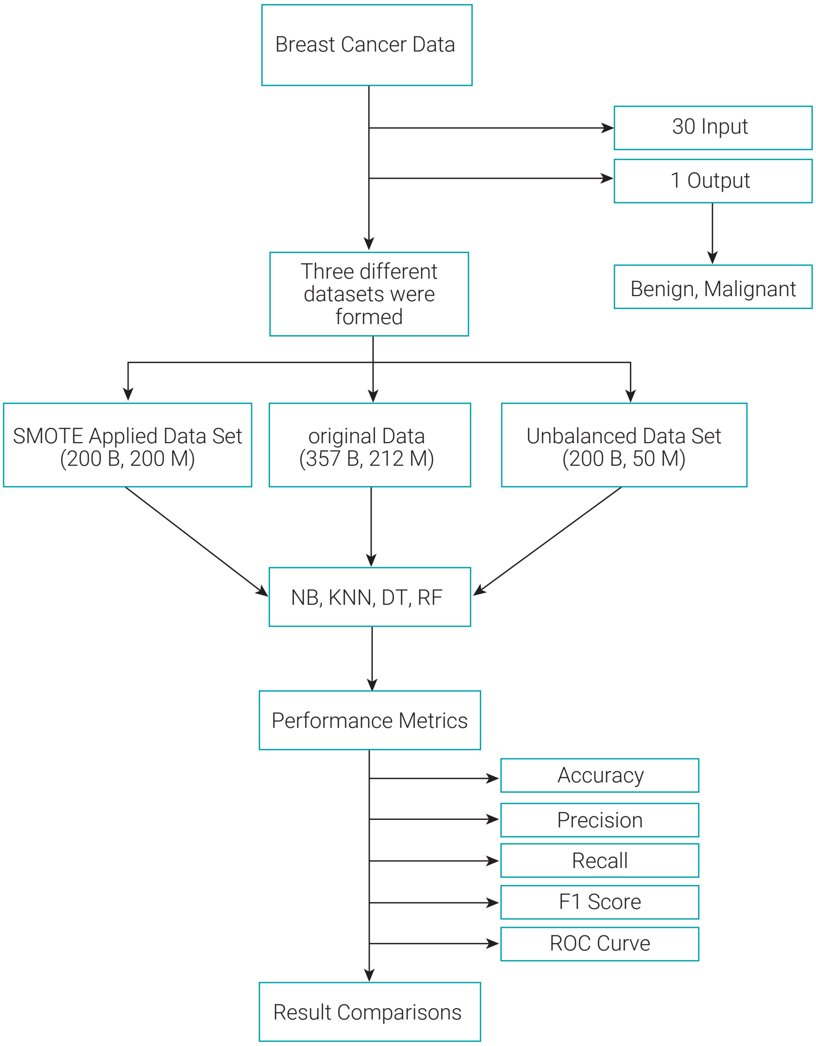
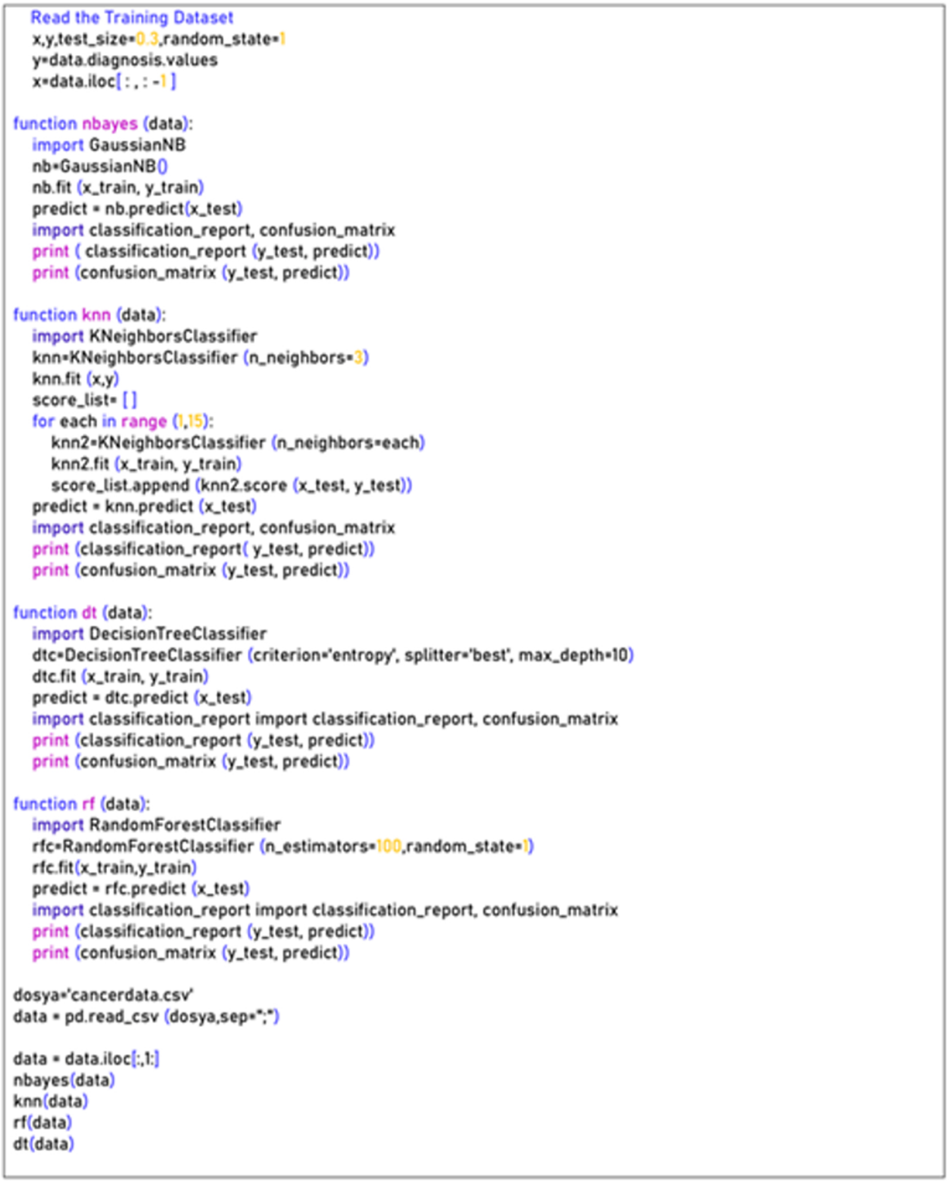


Fig 1. Flowchart
Source: own work

A model was developed using the Python sklearn library to evaluate breast cancer in an actual situation. The developed model consists of five code blocks and three different files. As shown in Fig 2, performance metrics for four different algorithms are calculated with the help of the performance.py code block. The pseudocode for this code block is given in Fig. 2.



```
  Read the Training Dataset
  x,y,test_size=0.3,random_state=1
  y=data.diagnosis.values
  x=data.iloc[ : , : -1 ]

function nbayes (data):
  import GaussianNB
  nb=GaussianNB()
  nb.fit (x_train, y_train)
  predict = nb.predict(x_test)
  import classification_report, confusion_matrix
  print ( classification_report (y_test, predict))
  print (confusion_matrix (y_test, predict))

function knn (data):
  import KNeighborsClassifier
  knn=KNeighborsClassifier (n_neighbors=3)
  knn.fit (x,y)
  score_list= [ ]
  for each in range (1,15):
    knn2=KNeighborsClassifier (n_neighbors=each)
    knn2.fit (x_train, y_train)
    score_list.append (knn2.score (x_test, y_test))
  predict = knn.predict (x_test)
  import classification_report, confusion_matrix
  print (classification_report( y_test, predict))
  print (confusion_matrix (y_test, predict))

function dt (data):
  import DecisionTreeClassifier
  dtc=DecisionTreeClassifier (criterion='entropy', splitter='best', max_depth=10)
  dtc.fit (x_train, y_train)
  predict = dtc.predict (x_test)
  import classification_report import classification_report, confusion_matrix
  print (classification_report (y_test, predict))
  print (confusion_matrix (y_test, predict))

function rf (data):
  import RandomForestClassifier
  rfc=RandomForestClassifier (n_estimators=100,random_state=1)
  rfc.fit(x_train,y_train)
  predict = rfc.predict (x_test)
  import classification_report import classification_report, confusion_matrix
  print (classification_report (y_test, predict))
  print (confusion_matrix (y_test, predict))

dosya='cancerdata.csv'
data = pd.read_csv (dosya,sep=";")

data = data.iloc[:,1:]
nbayes(data)
knn(data)
rf(data)
dt(data)
```

Fig 2. Pseudocode for performance.py
Source: own work

The features were included in the study prior to the selection of the application. Each of the other "py" extension files ran the corresponding algorithm. The result from each algorithm was printed in the outcome.csv file. Afterward, the final result value was assigned by taking the mode of each row based on the Borda Voting method. Thus, it was aimed to include each algorithm in the process. The developed model is given in Fig. 3.

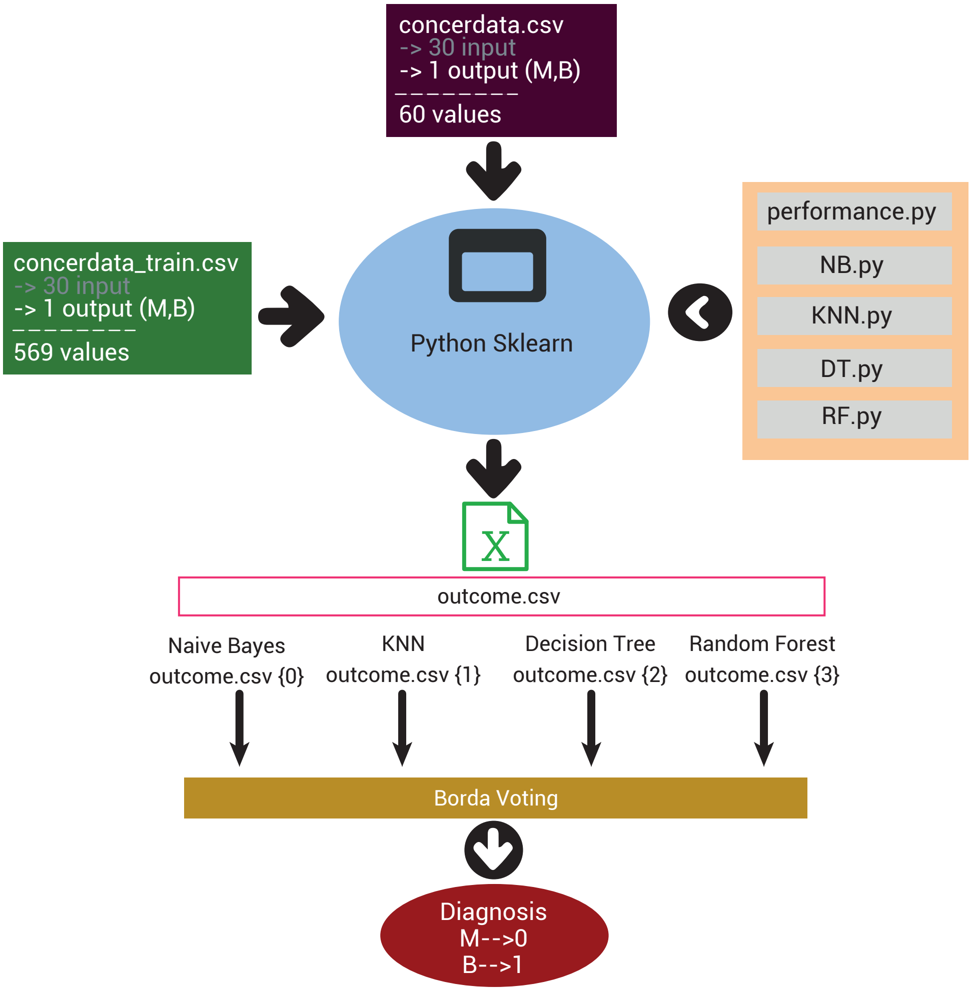


Fig 3. Proposal Model
Source: own work

The 60 records were randomly selected over the original dataset to test the developed model. Output values (B and M) of the records in the dataset were evenly distributed. The remaining records were used in the training dataset. The estimation results obtained for 60 records were compared with the actual values. The sequence diagram of the model was given in Fig. 4.

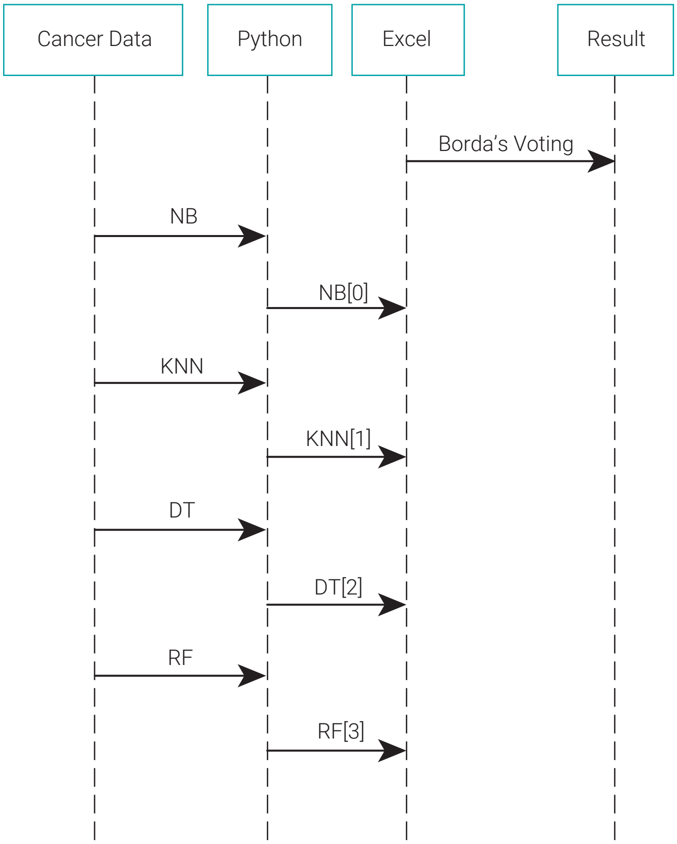


Fig 4. Sequence Diagram
Source: own work

2.1. ML Algorithms

Four different algorithms used in the applications were explained below. In the NB algorithm, the learning process was carried out over the training dataset. Following the calculation made after learning, the value with the highest probability was assigned to the relevant class. The said probability calculation was made with the help of the formula given in Eq. (1) [27].

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right) * P(A)}{P(B)} \tag{1}$$

P(A/B) is the posterior probability of the class given predictor.
P(A) is the prior probability of class.
P(B/A) is the likelihood, which is the probability of the predictor given class.
P(B) is the prior probability of the predictor [28].

Eq. 1 is the formula for a single column property. If there is more than one column, the same operations must be done for each column. In other words, in NB, each feature (X1, X2,...Xn) is evaluated independently of each other. In this case, the formula in Eq. (2) is used [29].

$$P(y|x_{1,\ldots,}x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)} \quad (2)$$

It is based on the NB Bayes Rule, used to estimate the probability of each class "y" given "x" values. Since the amount of data in the training dataset is linear with respect to the number of features, the training time is not affected by the dataset size. Also, low variance has low performance. After the calculation made in the NB algorithm, the value with the highest probability is assigned to the relevant class [30].

The model for the NB algorithm is given in Fig. 5.

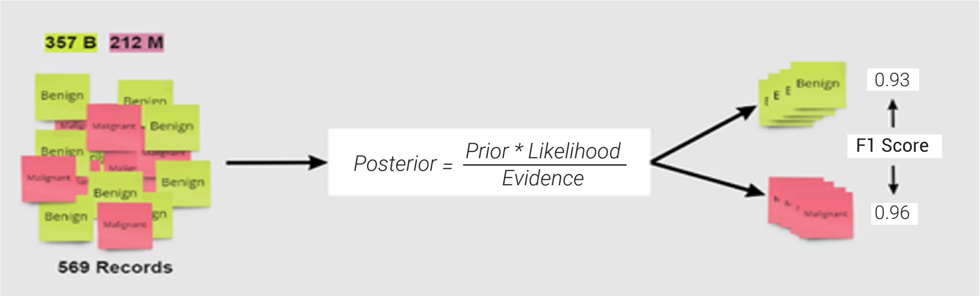


Fig 5. The model for the NB algorithm
Source: own work

KNN, which is one of the nonlinear classification algorithms, was developed by Cover and Hart in 1967 [31]. Since it is easily applicable and easily classifiable, the KNN algorithm is widely used in various fields. The most common areas in which KNN is used can be cited as pattern recognition, text classification, and object recognition [32]. This algorithm can be used for data in different fields and with different characteristics, showing that it has a strong learning ability [33]. This algorithm is the most widely used classical algorithm in cluster analysis. Many people are making improvements based on the K means algorithm [34]. While calculating the accuracy rate in the KNN algorithm, the test sample (X) is made to resemble the shape of the feature vector (X1, X2, ... Xm) to apply the previous KNN. Then, taking the Detroit sample, the similarity between the training sample and the X test sample (di1, di2, ...dim) should be calculated. The similarity is calculated by the formula in Eq. (3) [35].

$$Similarity\ (X, d_i) = \frac{\sum_{j=1}^{m} X_j.d_{ij}}{\sqrt{(\sum_{j=1}^{m} X_j)^2} + \sqrt{(\sum_{j=1}^{m} d_{ij})^2}} \qquad (3)$$

Next, select “k” samples with the greatest similarity and calculate the probability by the formula in Eq. (4) [36].

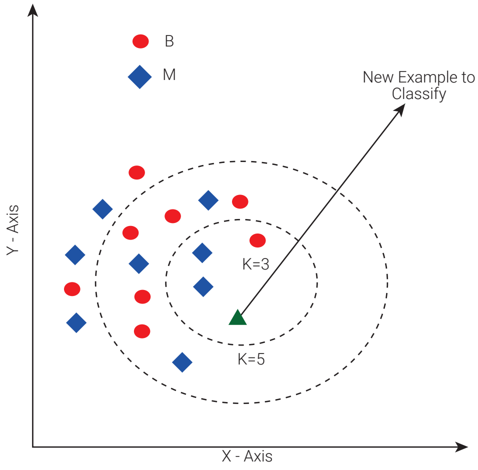
$$P\ (X, C_j) = \sum_{d}^{m} Similarity(X, d_i)\ .\ y(d_i, C_j) \qquad (4)$$

Where “y(di, Cj)” is an attribute of the category function with the following conditions [37]:

$$y(d_i, C_j)(\begin{cases} 1, d_i \ \in \ C_j \\ 0, d_i \ \notin \ C_j \end{cases}) \qquad (5)$$

Enter the instance X in the category with the largest P (X, Cj).

As can be seen in Fig. 6, the K value was determined first while finding the class to which a new value belongs, with the datasets consisting of blue and red dots. For each K value, the points closest to the new value were taken into account. The distance from each of these points to the new value was then calculated. The model of the KNN algorithm was given in Fig. 6.



	Accuracy
K=3	0.975
K=4	0.9725
K=5	0.9675
K=6	0.9625

Fig 6. The model for the KNN algorithm

Source: own work

DT, which is frequently used in solving classification problems, is a ML algorithm based on dividing input variables into more than one homogeneous cluster. The use of DT, which is very simple to use, does not require statistical information and analytical infrastructure and offers various advantages to users. In addition, data preprocessing, which is the basis of ML applications, facilitates data cleaning processes and minimizes the rate of being affected by missing values and outliers [38]. The model for the DT algorithm used in the study is given below.

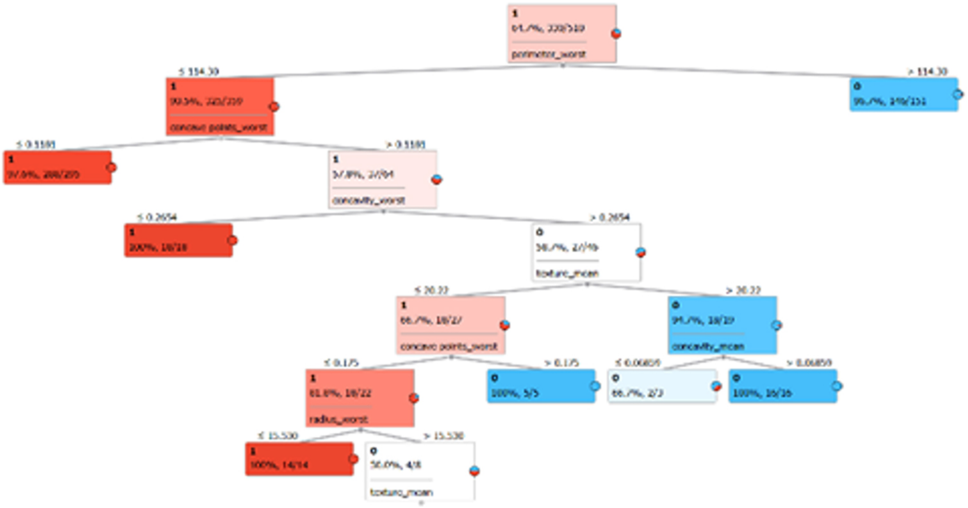


Fig 7. The model for the DT algorithm
Source: own work

Breiman (2001) RF land depends on the project of trees to be created depending on the values of the land [39]; The decision forest was formed by combining in this way. This was done for a tree that was thought over on the set. Its near decision value from a tree was assigned one vote and was consequently considered the most repeated decision value. It was hoped that this forest construction would achieve higher success in the problems carried out. The working diagram of the RF algorithm is given in Fig. 8.

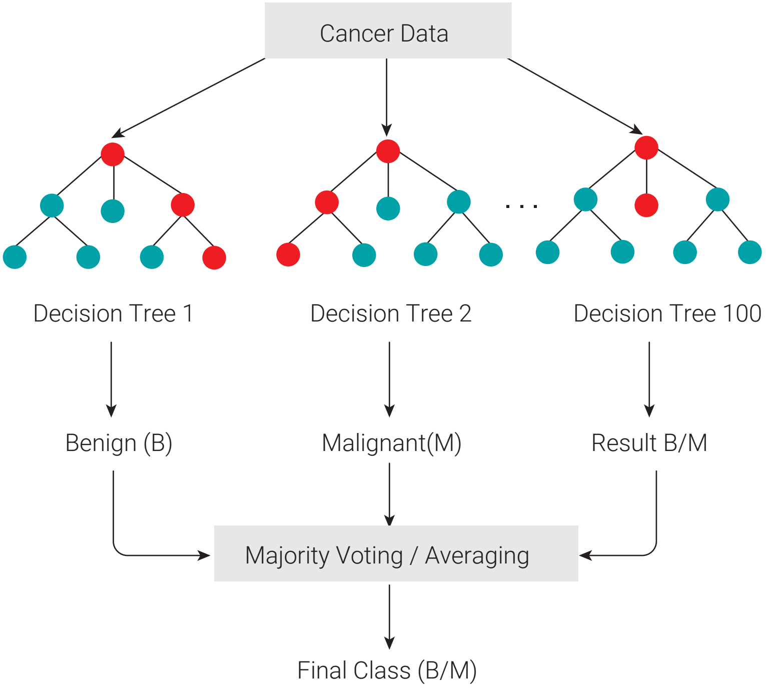


Fig 8. The model for the RF algorithm
Source: [40]

2.2. Performance Metrics

The performance in classification problems made with the help of ML algorithms can be expressed with the confusion matrix. The confusion matrix shows the number of predicted values for different outcome values [41]. A complexity matrix for a set of two classes is given in Fig 9.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	True Positive [1,1]	False Negative [1,0]
	Negative (0)	False Positive [0,1]	True Negative [0,0]

Fig 9. Confusion Matrix
Source: own work

According to the principles mentioned above;

- TP (True Positive), the number of correctly predicted data whose true value is 1,
- TN (True Negative), the number of correctly predicted data whose true value is 0,
- FN (False Negative), number of incorrectly predicted values whose true value is 1,
- FP (False Positive) represents the number of incorrectly predicted data with a true value of 0.

Accuracy: One of the simplest methods used to measure the performance of ML algorithms, the accuracy rate expresses the ratio of the number of correctly predicted data to the total number of samples. The accuracy rate was determined by Eq. (6) [42].

$$Accuracy \ = \frac{(\mathrm{TP+TF})}{(\mathrm{TP+TF+FP+FN})} \tag{6}$$

Precision: The precision value, which expresses the ratio of correctly predicted 1 values to the total number of predicted values as 1, was calculated with Eq. (7) [43].

$$Precision \ = \frac{\mathrm{TP}}{(\mathrm{TP+FP})} \tag{7}$$

Recall: Sensitivity value, based on the ratio of correctly predicted 1 values to the total number of positive samples, expresses the ratio of positive values to correctly classified values. The recall value was calculated with Eq. (8) [44].

$$Recall \ = \frac{\mathrm{TP}}{(\mathrm{TP+FN})} \tag{8}$$

The F1 score is based on the harmonic mean of the precision and recall values [45]. The F1 score value was calculated with Eq. (9)

$$F1 \ Score \ = 2 * (\frac{Precision*Recall}{Precision*Recall}) \tag{9}$$

3. FINDINGS AND DISCUSSION

In addition to the original dataset consisting of 569 records, four different algorithms were applied to the imbalanced distributed dataset created by the random selection method. The average of the relevant columns was assigned to each empty cell before the application using the Average Assignment method, which was stated to be effective in estimation studies [46]. The input and output values in the dataset discussed in the applications are given in Table 1.

Table 1. Inputs and Output Values

Input	Min. Value	Max. Value
Radius Mean	6.981	17.99
Texture Mean	9.71	24.54
Perimeter Mean	43.79	122.8
Area Mean	143.5	1001
Smoothness Mean	0.05263	0.1184
Compactness Mean	0.01938	0.2776
Concavity Mean	0.000692	0.3001
Concave Points Mean	0.001852	0.1471
Symmetry Mean	0.106	0.2419
Fractal Dim. Mean	0.04996	0.07871
Radius Se	0.1115	1.095
Texture Se	0.3602	1.428
Perimeter Se	0.757	8.589
Area Se	6.802	153.4
Smoothness Se	0.001713	0.007189
Compactness Se	0.002252	0.04904
Concavity Se	0.000692	0.05373
Concave Points Se	0.001852	0.01587
Symmetry Se	0.007882	0.03003
Fractal Dimension Se	0.000895	0.006193
Radius Worst	7.93	25.38
Texture Worst	12.02	30.37
Perimeter Worst	50.41	184.6
Area Worst	185.2	2019
Smoothness Worst	0.07117	0.1622

(continúa)

(viene)

Input	**Min. Value**	**Max. Value**
Compactness Worst	0.02729	0.6656
Concavity Worst	0.001845	0.7119
Concave Points Worst	0.008772	0.2654
Symmetry Worst	0.1562	0.4601
Fractal Dim. Worst	0.05504	0.1189
	Output	
Benign (B)		Malignant (N)

Source: own work

30% of the records in the dataset were used for testing and 70% for training. The applications were made using the Python Sklearn library. In the first of the applications created with three different datasets, the dataset of 569 records was discussed. A value of 0 was assigned for Malignant and 1 for Benign in the dataset.

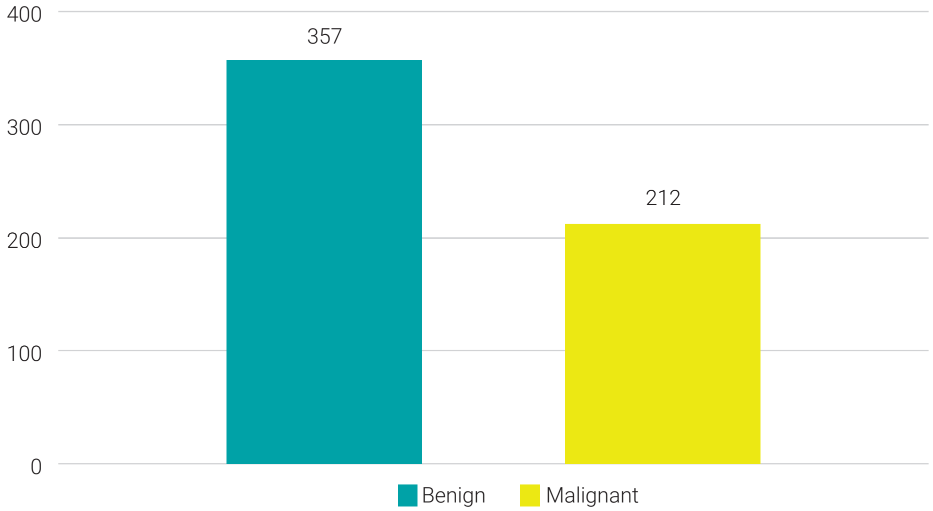


Fig 10. The distribution of the Classes
Source: own work

Performance metrics obtained in the application with NB, KNN, DT, and RF algorithms were given in Table 2. The highest accuracy rate was obtained with the KNN algorithm. At the same time, the 0.98 F1 Score obtained for the value 1 in the application with the KNN algorithm also revealed the algorithm's performance. In addition to the performance metrics, the ROC Curve was created for each application. The ROC Curve was defined as decision thresholds that give estimates of sensitivity and specificity. It was stated that the ROC curve provides several advantages in testing the accuracy of applications [47]. The ROC Curve created for each algorithm in the first application is given in Fig. 11.

Table 2. Performance Metrics for Breast Cancer Data

	Accuracy	Output	Precision	Recall	F1 Score
NB	0.947	0	0.94	0.92	0.93
		1	0.95	0.96	0.96
KNN	0.956	0	0.98	0.94	0.96
		1	0.96	0.99	0.98
DT	0.91	0	0.93	0.90	0.92
		1	0.95	0.96	0.95
RF	0.953	0	0.95	0.92	0.94
		1	0.95	0.97	0.96

Source: own work

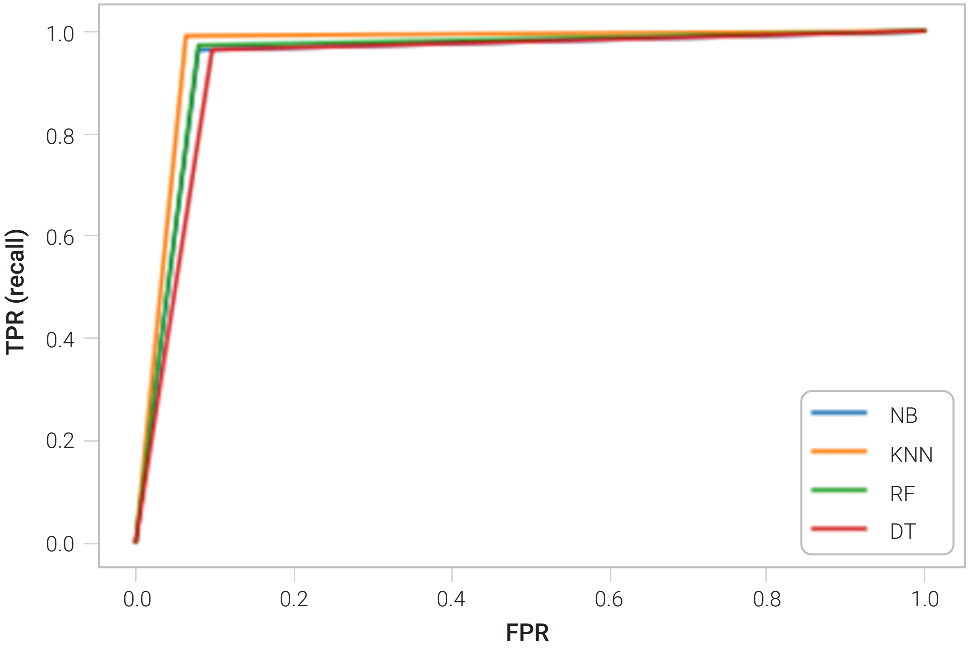


Fig 11. ROC Curve (Original Dataset)
Source: own work

In particular, unbalanced datasets used in the diagnosis of rare diseases can have a negative effect on the performance evaluations of algorithms. A new dataset consisting of 250 records was created using the random selection method to test this situation in breast cancer diagnosis. As seen in Fig 12, 50 of the records in the created dataset belong to M (Malignant) class, and 200 belong to B (Benign) class.

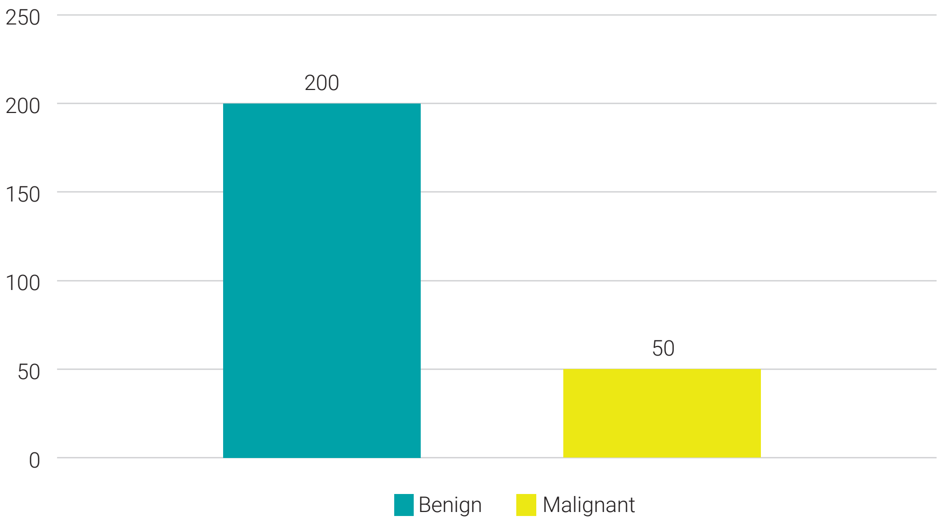


Fig 12. The distribution of the Classes (Unbalanced Dataset)
Source: own work

As in the previous application, the highest accuracy rate was obtained with the KNN algorithm. Although the Accuracy rate increased in the prior application, the F1 score for the value 1 decreased. However, a 0.99 F1 score was obtained for the 0 value. The performance metrics obtained from the RF algorithm of the created unbalanced dataset decreased. Performance metrics for the unbalanced dataset are given in Table 3. The ROC Curve created for each algorithm in applications with an unbalanced dataset is given in Fig. 13.

Table 3. The Performance Metrics for Unbalanced Dataset

	Accuracy	Output	Precision	Recall	F1 Score
NB	0.96	0	0.95	1.00	0.98
		1	1.00	0.81	0.90
KNN	0.976	0	0.98	1.00	0.99
		1	1.00	0.94	0.97
DT	0.946	0	0.97	0.95	0.96
		1	0.81	0.87	0.84
RF	0.933	0	0.94	1.00	0.97
		1	1.00	0.75	0.86

Source: own work

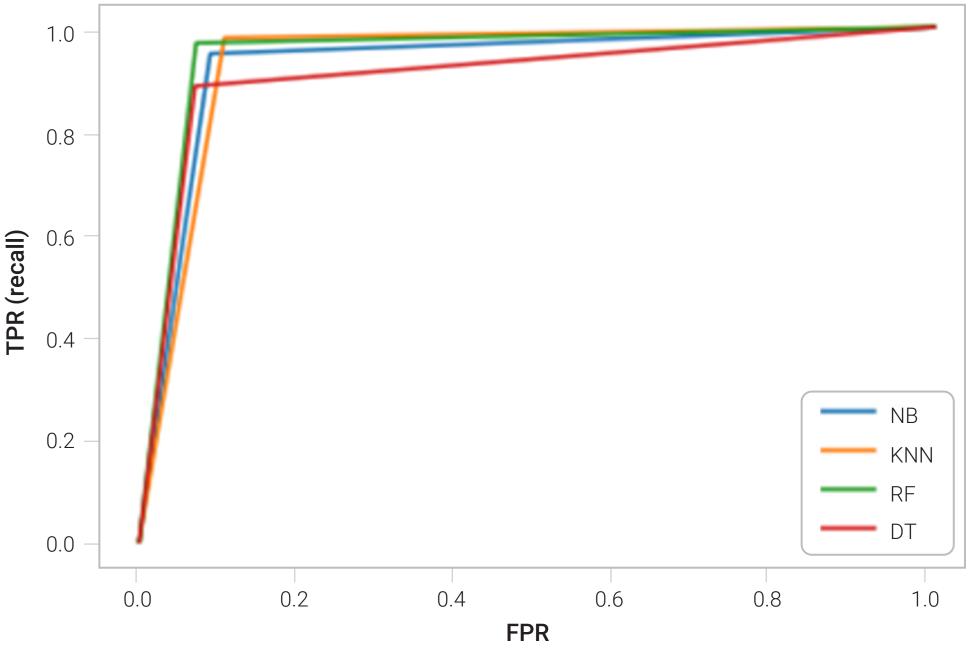


Fig 13. The ROC Curve (Unbalanced Dataset)
Source: own work

Performance metrics showed a decrease in some applications made with a dataset with a different number of classes, 50 and 200. In this context, the SMOTE algorithm, which was used to eliminate performance losses caused by skewed distribution, provided artificial reproduction of classes. In this way, the balancing of the dataset was realized. It was stated that the interpolation-based SMOTE algorithm was successful in various applications in different fields [48]. The performance metrics obtained in the applications with the balanced dataset obtained by applying SMOTE are given in Table 4. In applications with unbalanced datasets, the lowest accuracy was obtained in the application with the DF algorithm. However, the highest accuracy rate was obtained from the RF algorithm in the applications made with the dataset balanced by applying SMOTE. In addition, an F1 Score of 0.99 was obtained for both values with the KNN algorithm. The ROC Curve created for each algorithm in the applications made with the SMOTE applied dataset is given in Fig. 14.

Table 4. The Performance Metrics for SMOTE applied dataset

	Accuracy	Output	Precision	Recall	F1 Score
NB	0.96	0	0.94	0.98	0.96
		1	0.98	0.96	0.97
KNN	0.97	0	1.00	0.98	0.99
		1	0.99	1.00	0.99
DT	0.91	0	0.97	0.88	0.92
		1	0.87	0.96	0.91
RF	0.98	0	0.98	0.98	0.98
		1	0.99	0.99	0.99

Source: own work

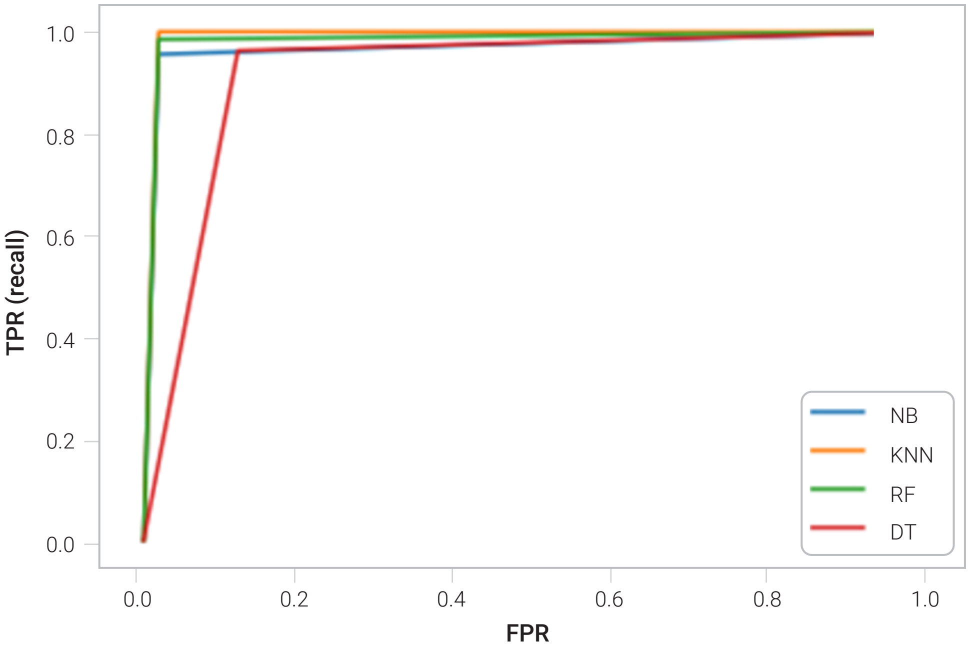


Fig 14. The ROC Curve (SMOTE Applied Dataset)
Source: own work

4. EXPERIMENTAL RESULTS

In order to test the proposed model, the 60 records were extracted from the dataset and tested. The estimation results obtained for these 60 records were compared with the actual values. All result values for these 60 records were found to be correct. It was seen that all four algorithms gave the same result for 52 records. For a value, NB and KNN algorithms gave the result 1, while DT and RF algorithms gave the result 0. As coefficients depending on the accuracy rate were examined, the value of 1 was assigned because the coefficients of the NB and KNN algorithms were more significant.

In the applications realized, the highest accuracy was obtained with the RF algorithm. The lowest performance was obtained with the DT algorithm based on the tree structure. In addition, as seen in Table 5, 29 out of 30 values for the NB algorithm, 25 of the 30 values for the KNN algorithm, and 29 of the 30 values for the DT algorithm were estimated correctly, depending on the Benign values. Finally, all the values were estimated correctly with the RF algorithm.

Table 5. The Performance Metrics for the test dataset

	TP	*TN*	*FN*	*FP*
NB	29	30	1	0
KNN	25	28	5	2
DT	29	30	1	0
RF	30	30	0	0

Source: own work

In studies using ML algorithms, the complexity matrix is used to reveal classification performance [49]. Actual values and predicted values were presented with the help of a two-dimensional matrix. As can be seen in Fig. 15, all the records of B values were estimated correctly with the NB, DT, and RF algorithms. Although the accuracy rate of the DT algorithm was low, high performance was achieved compared to the KNN algorithm in the application. B values were estimated correctly with the DT algorithm and it was observed that there was an error in only 1 record in the M values. The performance metrics obtained in the applications showed the applicability of ML algorithms in similar studies.

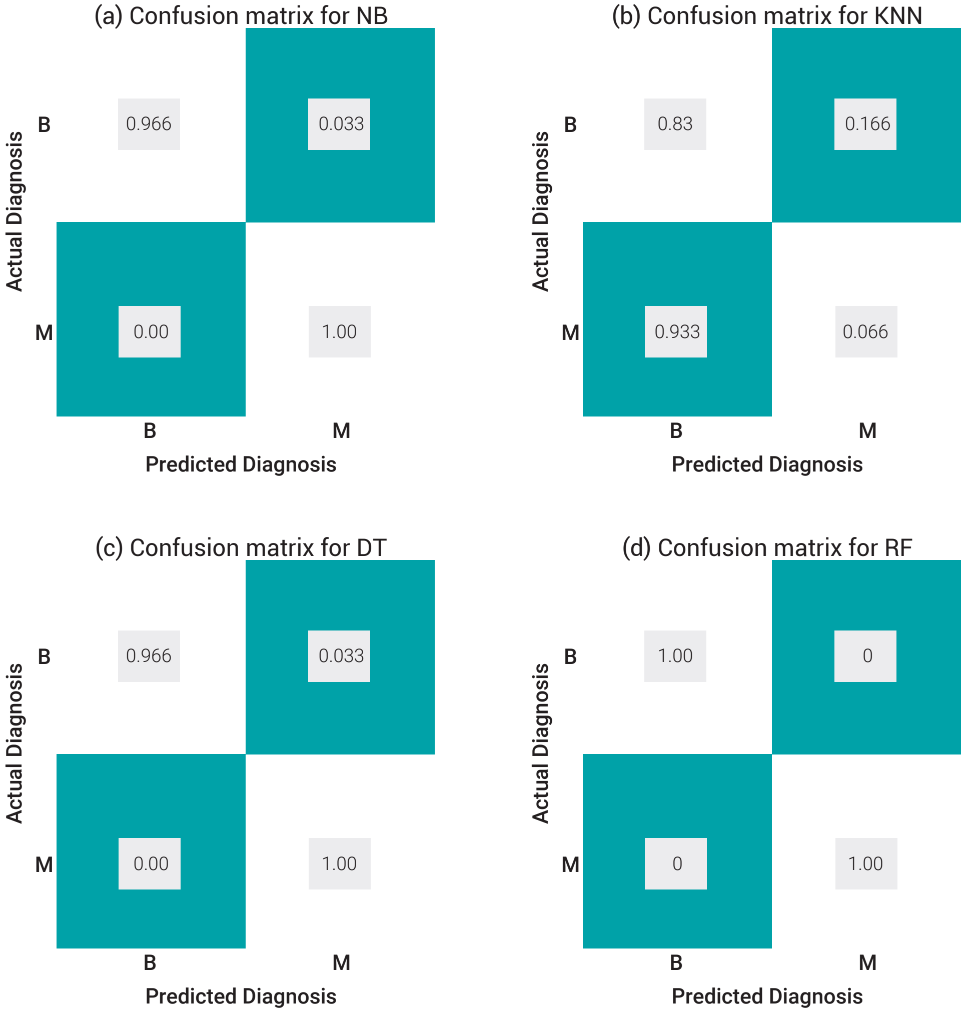


Fig 15. Confusion matrix for (a) NB; (b) KNN; (c) DT and (d) RF
Source: own work

As seen in the performance metrics and confusion matrix, the performances of the four algorithms differ from each other. The KNN algorithm showed low performance compared to other algorithms. 7 of the predicted values were estimated incorrectly. A value was predicted incorrectly with the NB and DT algorithms. Four algorithms were included in the process with the model developed within the scope of this study. In this way, although there were errors in the algorithms used alone, the performance of the system was increased by making use of the voting method. It was observed that almost all of the results obtained using four different algorithms were predicted correctly. In Table 6, the actual values (AV) of 20 randomly selected patients were compared with the final result (FR) of the model, and the predicted values.

Table 6. The comparison of estimated values with actual values

	NB	*KNN*	*DT*	*RF*	*FR*	*AV*
P1	M	B	M	M	M	M
P2	M	M	M	M	M	M
P3	M	M	B	M	M	M
P4	M	M	M	M	M	M
P5	M	B	M	M	M	M
P6	M	M	M	M	M	M
P7	M	M	M	M	M	M
P8	M	M	M	M	M	M
P9	B	B	M	M	B	M
P10	M	B	M	M	M	M
P11	B	B	B	B	B	B
P12	B	B	B	B	B	B
P13	B	B	B	B	B	B
P14	B	M	B	B	B	B
P15	B	B	B	B	B	B
P16	B	B	B	B	B	B
P17	B	B	B	B	B	B
P18	B	B	B	B	B	B
P19	B	B	B	B	B	B
P20	B	B	B	B	B	B

Source: own work

In Table 6, the algorithms were tested over the actual values and the results were analyzed. According to the analysis results, it was seen that the RF algorithm showed higher performance. The KNN algorithm, on the other hand, showed low performance compared to other algorithms. In addition, in the study of Bevilacqua et al. (2016), high performance was obtained in estimating the data belonging to the Malignant class. However, in this study, higher performance was obtained in the estimation of the data belonging to the Benign class. In addition, for a record, NB and KNN algorithms given the result B, while DT and RF algorithms given the result M. Since the accuracy rates of NB and KNN values were higher, the final result value B was assigned. However, it was observed that this value differed from the predicted value. Finally, it was seen that all values were estimated correctly with the RF algorithm.

Table 6. The Proposed Method comparison with literature

Research	Year	Max Accuracy Rate (%)	Algorithm
Eleyan	2012	97.51	KNN
Ahmad et al.	2013	95	SVM
Zand	2015	86.7	DT
Bevilacqua et al.	2016	89.77	ANN
Amrane et al.	2018	97.51	KNN
Bayrak et al.	2019	97.36	NB
Asfaw	2019	96.93	LR
Yadav et al.	2019	97.2	SVM-RF
Assegie	2021	94.35	KNN
Proposal Method	**2022**	**98**	**RF**

Source: own work

As can be seen in Table 7, similar studies in the literature were examined. As a result of the examination, the highest accuracy rates obtained in the applications made within the scope of the studies were given. Within the scope of this study, 98% accuracy was obtained for the RF algorithm in the application with the SMOTE applied dataset. It can be seen that this rate is the highest among other studies. In addition, with the Borda Voting method, four algorithms were included in the process and the final result value was assigned. In tests with actual values, it was seen that the correct final result was obtained for 59 of 60 records.

5. CONCLUSIONS AND RECOMMENDATIONS

This study aimed to demonstrate the applicability of ML algorithms in diagnosing breast cancer by using three different datasets. In this direction, the dataset consisting of 569 records was first discussed. Then, an unbalanced dataset was created by the random selection method. Finally, SMOTE was applied to compare the performance of the algorithms in the balanced-unbalanced dataset. Various performance metrics were used to evaluate the performance of algorithms in the applications. Although an accuracy rate of 0.933 was obtained for the RF algorithm in the application with the unbalanced dataset, 0.75 recall was obtained for B values. In this direction, it can be said that the use of different performance metrics in similar studies will contribute to the evaluation of performances.

In the studies of Chaurasia et al. (2018), the results were compared using different algorithms to diagnose breast cancer. The results of four different algorithms used in this study were printed in different columns on the same file, and Borda Voting Method was applied. The use of voting methods can contribute to performance order to include each of the algorithms used in similar studies in the process [50].

An increase was observed in the performance metrics obtained by the SMOTE method applied to the unbalanced dataset. Especially in datasets used to diagnose rare diseases, performance losses can be seen due to data imbalances. In this context, applying various resampling methods to unbalanced datasets will be beneficial in terms of performance.

Amrane et al. (2018) used NB and KNN algorithms in their study. Within the scope of this study, it was seen that the highest performance was obtained with the RF algorithm in the applications made with the SMOTE applied dataset. It can be said that increasing the number of algorithms used in similar studies will contribute to performance.

In medical informatics, the processing and interpretation of patient data are considered important in solving various problems. ML algorithms are frequently used, especially in studies where predictions for disease diagnosis are made. The performance metrics obtained in the applications made within the scope of this study reveal the applicability of ML algorithms in the diagnosis of breast cancer. The use of ML algorithms in solutions that include data-driven problems, such as breast cancer diagnosis, will contribute to the field.

6. REFERENCES

[1] B. Mahesh, “ML algorithms-a review,” *International Journal of Science and Research (IJSR)*, vol. 9, 381-386, 2020. doi: https://doi.org/10.21275/ART20203995

[2] R. D. Nindrea, T. Aryandono, L. Lazuardi and I. Dwiprahast, “Diagnostic accuracy of different ML algorithms for breast cancer risk calculation: a meta-analysis,” *Asian Pacific journal of cancer prevention: APJCP*, vol. 19, no. 7, 1747, 2018. doi: https://doi.org/10.22034/APJCP.2018.19.7.1747

[3] World Health Organization. Cancer, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer.

[4] S. Chaudhury, S. Mukhopadhyay and D. Kbah, “A Systematic Review of Cad System Based Approach in Diagnosing Breast Cancer and Analyze Effectiveness of ML and Deep Learning Algorithms in Early Detection,” *IJBPAS*, vol. 10, no. 11, pp. 804-827. doi: https://doi.org/10.31032/IJBPAS/2021/10.11.1069

[5] Z. Ahmed, K. Mohamed, S. Zeeshan and X. Dong, Artificial intelligence with multi-functional ML platform development for better healthcare and precision medicine, 2020. doi: https://doi.org/10.1093/database/baaa010

[6] S. Nanglia, M. Ahmad, F.A. Khan, N.Z. Jhanjhi, An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. Biomedical Signal Processing and Control, vol. 72, 103279, 2022. doi: https://doi.org/10.1016/j.bspc.2021.103279

[7] S. Sahan, K. Polat, H. Kodaz and S. Güneş, “A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis,” *Computers in Biology and Medicine,* vol. 37, no. 3, pp. 415-423, 2007. doi: https://doi.org/10.1016/j.compbiomed.2006.05.003

[8] A. Eleyan, “Breast cancer classification using moments”, *Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting*, 1–4, 2012. doi: https://doi.org/10.1109/SIU.2012.6204778

[9] M. A. Al-Hashem, A. M. Alqudah, and Q. Qananwah, “Performance Evaluation of Different ML Classification Algorithms for Disease Diagnosis,” *International Journal of E-Health and Medical Communications (IJEHMC),* vol. 12, no. 6, pp. 1-28, 2021. [Online]. Available: https://www.igi-global.com/gateway/article/full-text-pdf/278822

[10] L. G. Ahmad, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi and A. R. Razavi, “Using three ML techniques for predicting breast cancer recurrence,” *J Health Med Inform*, vol. 4, no. 124, p. 3. 2013. doi: https://dx.doi.org/10.4172/2157-7420.1000124

[11] K. Williams, P. A. Idowu, J. A. Balogun, and A. I. Oluwaranti, “Breast cancer risk prediction using data mining classification techniques,” *Transactions on Networks and Communications*, vol. 3, no. 2, pp. 1, 2015. doi: https://dx.doi.org/10.14738/tnc.32.662

[12] T. M. Mejía, M. G. Pérez, V. H. Andaluz and A. Conci, “Automatic segmentation and analysis of thermograms using texture descriptors for breast cancer detection,” *Asia-Pacific Conference on Computer Aided System Engineering, IEEE*, 2015. doi: https://doi.org/10.1109/APCASE.2015.12

[13] M. Tahmooresi, A. Afshar, B. B. Rad, K. B. Nowshath, and M. A. Bamiah, “Early detection of breast cancer using ML techniques,” *Journal of Telecommunication, Electronic and Computer Engineering (JTEC),* vol. 10, no. 3-2, pp. 21-27, 2018. [Online]. Available: https://jtec.utem.edu.my/jtec/article/view/4706/3462

[14] H. K. K. Zand, "A comparative survey on data mining techniques for breast cancer diagnosis and prediction," *Indian Journal of Fundamental and Applied Life Sciences*, 4330-9, 2015. doi: https://dx.doi.org/10.26808/rs.re.v3i5.04

[15] V. Bevilacqua, A. Brunetti, M. Triggiani, D. Magaletti, M. Telegrafo, M., M. Moschetta, "An optimized feed-forward artificial neural network topology to support radiologists in breast lesions classification," In Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion, pp. 1385-1392, 2016. doi: https://doi.org/10.1145/2908961.2931733

[16] L. Hussain, W. Aziz, S. Saeed, S. Rathore and M. Rafique, "Automated breast cancer detection using ML techniques by extracting different feature extracting strategies," In 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE) pp. 327-331, IEEE, 2018. doi: https://dx.doi.org/10.1109/TrustCom/BigDataSE.2018.00057

[17] M. Amrane, S. Oukid, I. Gagaoua, T. Ensari, "Breast cancer classification using ML," In 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT) pp. 1-4, IEEE, 2018. doi: https://doi.org/10.1109/EBBT.2018.8391453

[18] E. A. Bayrak, P. Kırcı and T. Ensari, "Comparison of ML methods for breast cancer diagnosis," In 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT) pp. 1-3, IEEE, 2019. doi: https://doi.org/10.1109/EBBT.2019.8741990

[19] T. A. Asfaw, "Comparative Analysis of Classification Approaches For Breast Cancer," *International Journal of Computer Engineering and Technology (IJCET)*, vol. 10, no. 4, pp. 10-16, 2019. [Online]. Available: https://iaeme.com/Home/issue/IJCET?Volume=10&Issue=4

[20] V. P. C. Magboo and M. S. A. Magboo, "ML Classifiers on Breast Cancer Recurrences," *Procedia Computer Science,* vol. 192, pp. 2742-2752, 2021. doi: https://doi.org/10.1016/j.procs.2021.09.044

[21] A. Yadav, I. Jamir, R. R. Jain, M. Sohani, "Comparative study of ML algorithms for breast cancer prediction-a review," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, vol. 5, no. 2, pp. 979-985, 2019. doi: https://doi.org/10.32628/CSEIT1952278

[22] F. J. M. Shamrat, M. A. Raihan, A. S. Rahman, I. Mahmud, R. Akter, "An analysis on breast disease prediction using ML approaches," *International Journal of Scientific & Technology Research*, vol. 9, no. 02, pp. 2450-2455, 2020. [Online]. Available: http://www.ijstr.org/final-print/feb2020/An-Analysis-On-Breast-Disease-Prediction-Using-Machine-Learning-Approaches.pdf

[23] T. A. Assegie, “An optimized K-Nearest Neighbor based breast cancer detection,” *Journal of Robotics and Control (JRC)*, vol. 2, no. 3, pp. 115-118, 2021. doi: https://doi.org/10.18196/jrc.2363

[24] M. Manjurul Ahsan and Z. Siddique, Z, “ML based disease diagnosis: A comprehensive review,” arXiv:2201.02755v1, arXiv e-prints, arXiv-2112, pp. 1-5, 2022. doi: https://doi.org/10.48550/arXiv.2112.15538

[25] S. Muhtadi, “Breast Tumor Classification Using Intratumoral Quantitative Ultrasound Descriptors,” *Computational and Mathematical Methods in Medicine*, pp. 1-18, 2022. doi: https://doi.org/10.1155/2022/1633858

[26] D. Dua and C. Graff, UCI ML Repository. Irvine, CA: University of California, School of Information and Computer Science, 2019.

[27] B. Chandra, M. Gupta, “Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data,” *Expert Systems with Applications*, vol. 38, no. 3, pp.1293-1298, 2011. doi: https://doi.org/10.1016/j.eswa.2010.06.076

[28] S. Ray, 6 Easy Steps to Learn Naive Bayes Algorithm. 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained.

[29] I. Zobu, İ. Naive Bayes, Teorisi ve Python uygulaması. 2019. [Online]. Available: https://medium.com/kaveai/naive-bayes-ve-uygulamalar%C4%B1-d7d5a56c689b.

[30] G. I. Webb, Naive bayes, Encyclopedia of Machine Learning, C. Sammut and G. I. Webb, Eds., pp. 713–714, Springer, New York, NY, USA, 2010.

[31] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21-27, 1967. doi: https://doi.org/10.1109/TIT.1967.1053964

[32] S. Dhanabal and S. Chandramathi, “A review of various k-nearest neighbor query processing techniques,” *International Journal of Computer Applications*, vol. 31, no. 7, pp. 14-22, 2011.

[33] X. Huang, “An improved KNN algorithm and its application in real-time car-sharing prediction,” M.S. thesis, Dalian University of Technology, Daian, China, 2018.

[34] Z. Lv, K. Ota, J. Lloret, W. Xiang, P. Bellavista, “Complexity Problems Handled by Advanced Computer Simulation Technology in Smart Cities 2021,” Hindawi Complexity, Article ID 9847249, 2022. doi: https://doi.org/10.1155/2022/9847249

[35] R. I. Borman, R. Napianto, N. Nugroho, D.Pasha, Y. Rahmanto and Y. E. P. Yudoutomo, "Implementation of PCA and KNN Algorithms in the Classification of Indonesian Medicinal Plants," In 2021 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), pp. 46-50, IEEE, 2021. doi: https://doi.org/10.1109/ICOMITEE53461.2021.9650176

[36] F. Rossi, A. Aizzuddin and A. Rahni, A., "Joint Segmentation Methods of Tumor Delineation in PET – CT Images : A Review", 7, pp. 137–145, 2018. doi: https://dx.doi.org/10.14419/ijet.v7i3.32.18414

[37] P. Prasetyawan, I. Ahmad, R. I. Borman, Y. A. Pahlevi, D. E. Kurniawan, "Classification of the Period Undergraduate Study Using Back-propagation Neural Network," In 2018 International Conference on Applied Engineering (ICAE) pp. 1-5, IEEE, 2018. doi: https://doi.org/10.1109/INCAE.2018.8579389

[38] W. Sullivan, "ML For Beginners Guide Algorithms: Supervised & Unsupervised Learning, Decision Tree & Random Forest Introduction", CreateSpace Independent Publishing Platform, 2017.

[39] L. Breiman, Random forests, *ML*, vol. 45, no. 1, pp. 5-32, 2001. doi: https://doi.org/10.1023/A:1010933404324

[40] H. Ampadu, Random Forests, Understanding. 2021. [Online]. Available: https://aipool.com/a/s/random-forests-understanding. Accessed on: April 1, 2022.

[41] N. Seliya, T. M. Khoshgoftaar, J. Van Hulse, "A study on the relationships of classifier performance metrics," In 2009 21st IEEE international conference on tools with artificial intelligence, pp. 59-66, IEEE, 2009. doi: https://doi.org/10.1109/ICTAI.2009.25

[42] H. Nizam and S. S. Akın, "Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması", XIX. Türkiye'de İnternet Konferansı,pp. 1-6, 2014. [Online]. Available: https://inet-tr.org.tr/inetconf19/bildiri/10.pdf

[43] M. Almseidin, M. Alzubi, S. Kovacs, M. Alkasassbeh, "Evaluation of ML algorithms for intrusion detection system," In 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY) (pp. 000277-000282). IEEE, 2017. doi: https://doi.org/10.48550/arXiv.1801.02330

[44] M. Wu, X. Zhong, Q. Peng, M. Xu, S. Huang, S., Yuan, J., J. Ma, T. Tan, "Prediction of molecular subtypes of breast cancer using BI-RADS features based on a "white box" ML approach in a

multi-modal imaging setting," *European journal of radiology*, vol. 114, pp. 175-184. 2019. doi: https://doi.org/10.1016/j.ejrad.2019.03.015

[45] S. Balakrishna, M. Thirumaran, V. Solanki, "Machine Learning based Improved Gaussian Mixture Model for IoT Real-Time Data Analysis: Análisis de los datos," Revista Ingeniería Solidaria, vol. 16, no. 1, Jan. 2020. doi: https://doi.org/10.16925/2357-6014.2020.01.02

[46] H. M. Dodeen, "Effectiveness of valid mean substitution in treating missing data in attitude assessment," *Assessment & Evaluation in Higher Education*, vol. 28, no. 5, pp. 505-513, 2003. doi: https://doi.org/10.1080/02602930301674

[47] N. A. Obuchowski, "ROC analysis," *American Journal of Roentgenology*, vol. 184, no. 2, 364-372, 2005. doi: https://doi.org/10.2214/ajr.184.2.01840364

[48] A. Fernández, S. Garcia, F. Herrera,"SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*," vol. 61, pp. 863-905, 2018. doi: https://doi.org/10.1613/jair.1.11192

[49] O. I. Obaid, M. A. Mohammed, M. K. A. Ghani, A. Mostafa, F. Taha, "Evaluating the performance of ML techniques in the classification of Wisconsin Breast Cancer," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 160-166, 2018. [Online]. Available: http://185.104.157.219:8080/repoAnbar/bitstream/123456789/4488/1/IJET-23737.pdf

[50] V. Chaurasia, S. Pal, B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 119-126, 2018. doi: https://dx.doi.org/10.1177/1748301818756225