Application of Machine Learning to Breast Cancer Prediction

Catarina Pimenta[1]
University of Coimbra
[1]Student, Department of Informatics Engineering,
Faculty of Sciences and Technology

Abstract

Breast cancer continues to be among the leading causes of death for women and is increasing in developing countries where most cases are diagnosed in late stages. The detection and classification of breast cancer in the early stages of its development may allow patients to have proper treatment. This paper analyzes the performance of seven Machine Learning (ML) algorithms: Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN) on the Wisconsin Breast Cancer Dataset (WBCD) from the University of California at Irvine (UCI) ML repository. For the implementation of the ML algorithms, the dataset was partitioned into the training phase and the testing phase. Experimental results show that the algorithms and techniques that we use achieved high processing performances in the breast cancer classification.

Keywords: Breast cancer classification, Machine Learning, Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Tree, Random Forest, Artificial Neural Network

1 Introduction

Breast cancer is the most diagnosed cancer, in women, and accounts for 1 in 4 annual cancer cases worldwide [1]. In 2020 the incidence in the WHO Europe region was estimated to be 576,300 [2] and, in the EU-27 was 355,500 [3]. An estimated 21% of breast cancer cases in Europe occur in women when they are younger than 50 years old and 35% occur at age 50–64 and the remaining cases in women above this age [2]. 1 in 11 women in the EU-27 will develop breast cancer before the age of 74 [3].

Breast cancer is the leading cause of cancer death in women worldwide accounting for 1 in 6 cancer deaths [1] and claims the lives of more European women than any other cancer [2]. In 2020 the estimated number of women who died from breast cancer in the WHO Europe region was 157,100 [2] and, in the EU-27 was 91,826 [3].

The increasing number of breast cancer cases may be due to changes in lifestyle habits such as sedentary lifestyle, weight gain, obesity and alcohol consumption. Reproductive factors like having children at a younger age (under 30), having several children, and breast-feeding for long periods of time reduces breast cancer risk. The use of contraceptives like combined estrogen-progestogen oral contraceptives are associated with an increased risk of breast cancer, notably among young women. Women who began menstruating before ages 11 or 12 or went through menopause after age 55 have a somewhat higher risk of breast cancer [4].

This paper proposes a classification method for breast tumor classification using ML classification methods. We evaluated the performance of the following classification algorithms: Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN). These algorithms provide different accuracy values. Our data source is the Wisconsin Breast Cancer Dataset (WBCD) taken from the University of California at Irvine (UCI) machine learning repository.

The objective of this study is to detect breast cancer from a dataset from a machine learning Python program that assists

The rest of the paper is arranged as follows: Section 2 reviews existing work on breast cancer tumor detection. Section 3 the material used is presented. Section 4 is about correlation of the data. Section 5 the proposed methodology is explained and the experiments. Section 6 presents and discusses the results obtained from the experiments. Section 7 concludes the paper and suggests some future direction of research for breast cancer tumor classification.

2 Related Work

The research using machine learning classifiers in the medical domain has been prevalent for a long time, especially in the diagnosis of breast cancer. Classification task is one of the popular types of machine learning tasks. Several research studies have been conducted by applying ML classifiers on different medical data, one of which is the WBCD.

Reference [5] examined the robustness of the least square Support Vector Machine (SVM) by using classification accuracy, analysis of sensitivity and specificity, k-fold cross

validation method, and confusion matrix. They obtained classification accuracy of 98.53%.

Reference [6] proposed the use of Feed Forward Artificial Neural Networks and back propagation algorithm to train the network. The performance of the network is evaluated using WBDS for various training algorithms. The highest accuracy of 99.28% is achieved when using Levenberg Marquardt algorithm.

Reference [7] investigated a new classification approach for detection of breast abnormalities in digital mammograms using Particle Swarm Optimized Wavelet Neural Network (PSOWNN). The proposed abnormality detection algorithm is based on extracting Laws texture energy measures from mammograms and classifying the suspicious regions by applying a pattern classifier. They achieved 93.671%, 92.105% and 94.167% for accuracy, specificity, and sensitivity, respectively.

Reference [8] used a rough set indiscernibility relation method with back propagation neural network (RS-BPNN). This work has two stages. The first stage handles missing values to obtain a smooth data set and to select appropriate attributes from the clinical dataset by indiscernibility relation method. The second stage is classification using back propagation neural network. The accuracy obtained from the proposed method was 98.6% on breast cancer dataset.

In our study, we compared seven ML algorithms: LR, NB, SVM, KNN, DT, RF, ANN and their correlations also comparing with the predicted attribute (output). Our goal is providing best set of features in order to increase the precision and performance of the breast cancer diagnosing.

3 Materials

3.1 Dataset

In this study, I used the public Wisconsin Breast Cancer Dataset (WBCD) from the UCI Machine Learning repository, which provides very solid and useful information for the sole purpose of building models to predict whether the cancer is benign or malignant.

The dataset I'll be working with has 32 attributes (attribute 'ID' is omitted), and 568 instances. From all the attributes only 1 is categorical, as shown in the following figure.

ATRIBUTE NAME	DESCRIPTION	TYPE	DOMAIN
DIAGNOSIS	The diagnosis of breast tissues	Categorical	M or B (Malignant or Benign)
RADIUS_MEAN	Mean of distances from center to points on the perimeter	Numeric	From 6.98 to 28.1
TEXTURE_MEAN	Standard deviation of gray-scale values	Numeric	From 9.71 to 39.3
PERIMETER_MEAN	Mean size of the core tumor	Numeric	From 43.8 to 189
AREA_MEAN	Mean area size for the core tumor	Numeric	From 144 to 2500
SMOOTHNESS_MEAN	Mean of local variation in radius lengths	Numeric	From 0.05 to 0.16
COMPACTNESS_MEAN	Mean of perimeter^2 / area - 1.0	Numeric	From 0.02 to 0.35
CONCAVITY_MEAN	Mean of severity of concave portions of the contour	Numeric	From 0 to 0.43
CONCAVE POINTS_MEAN	Mean for the number of concave portions of the contour	Numeric	From 0 to 0.2
SIMETRY_MEAN	Mean symmetry for the core tumor	Numeric	From 0.11 to 0.3
FRACTAL_DIMENSION_MEAN	Mean for "coastline approximation" – 1.0	Numeric	From 0.05 to 0.1
RADIUS_SE	Standard error for the mean of distances from center to points on the perimeter	Numeric	From 0.11 to 2.87
TEXTURE_SE	Standard error for the standard deviation of gray-scale values	Numeric	From 0.36 to 4.88
PERIMETER_SE	Standard error for the size of the core tumor	Numeric	From 0.76 to 22
AREA_SE	Standard error for the area size for the core tumor	Numeric	From 6.8 to 542
SMOOTHNESS_SE	Standard error for local variation in radius lengths	Numeric	From 0 to 0.03
COMPACTNESS_SE	Standard error for perimeter^2 / area – 1.0	Numeric	From 0 to 0.14
CONCAVITY_SE	Standard error for severity of concave portions of the contour	Numeric	From 0 to 0.4
CONCAVE POINTS_SE	Standard error for the number of concave portions of the contour	Numeric	From 0 to 0.05
SYMMETRY_SE	Standard error for the symmetry size for the core tumor	Numeric	From 0.01 to 0.08
FRACTAL_DIMENSION_SE	Standard error for "coastline approximation" – 1	Numeric	From 0 to 0.03
RADIUS_WORST	Worst (or largest) value for mean of distances from center to points on the perimeter	Numeric	From.93 to 36
TEXTURE_WORST	Worst (or largest) value for standard deviation of gray-scale values	Numeric	From 12 to 49.5
PERIMETER_WORST	Worst (or largest) value for the size of the core tumor	Numeric	From 50.4 to 251
AREA_WORST	Worst (or largest) value for the area size for the core tumor	Numeric	From 185 to 4250
SMOOTHNESS_WORST	Worst (or largest) value for local variation in radius lengths	Numeric	From 0.07 to 0.22
COMPACTNESS_WORST	Worst (or largest) value for perimeter^2 / area – 1.0	Numeric	From 0.03 to 1.06
CONCAVITY_WORST	Worst (or largest) value for severity of concave portions of the contour	Numeric	From 0 to 1.25
CONCAVE POINTS_WORST	Worst (or largest) value for number of concave portions of the contour	Numeric	From 0 to 0.29
SYMETRY_WORST	Worst (or largest) value for the symmetry size for the core tumor	Numeric	From 0.16 to 0.66
FRACTAL_DIMENSION_WORST	Worst (or largest) value for "coastline approximation" –1	Numeric	From 0.06 to 0.21

Figure 1. Dataset
Note: It is necessary to enlarge the image to visualize

I also plotted the diagnosis data in order to get a good grip of how much of these results were Malignant or Benign.

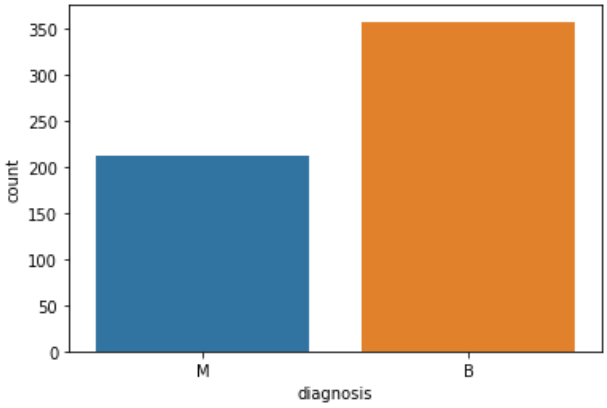

Figure 2. Number of Malignant and Benign results.

3.2 Frameworks

For the code implemented in this study there were applied multiple libraries and methods in order to test various alternatives to solve the problem. For most of these I used the library known as Sklearn which features several classification, regression and clustering algorithms, among which I have K-Nearest Neighbor, Gaussian's Naive Bayes, Support Vector Classifier, Logistic Regression, Decision Tree, Random Forest. It was also used Neural Network which required using the library Keras. The other libraries used such as seaborn, matplotlib are used to analyze data in graphics and plots and the library Pandas was used to extract the dataset.

4 Correlation of the data

4.1 Visualize Variables Dispersion and Relation

It is important to know how dispersed the diagnostic values are for each variable. For this, I used a seaborn pairplot with the middle columns and the diagnostic column to better understand and visualize the relation between the variables and the dispersion of the diagnosis in each relation. I did not use all variables because it would be difficult to visualize, therefore, I chose to plot the mean values.

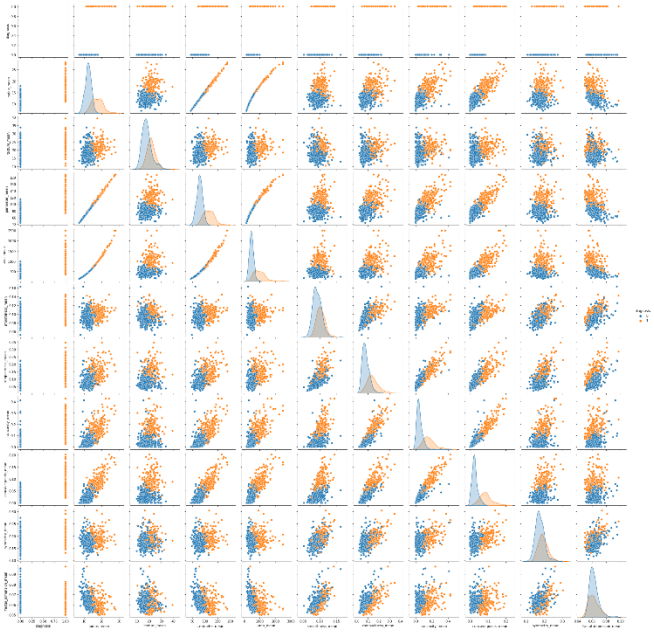

Figure 3. Relation and dispersion plot
Note: It is necessary to enlarge the image to visualize

In the plot above we can see that there are relations variables which diagnosis is not disperse like in *texture_mean* and *fractal_dimension_mean* relation and are others that in the opposite is very disperse like *texture_mean* and *radius_mean* relation.

4.2 Feature Analysis (Heat Map)

The heat map represents the correlations between each variable through a color scheme, the darker the color the stronger the correlation between each variable and the lighter the weaker.

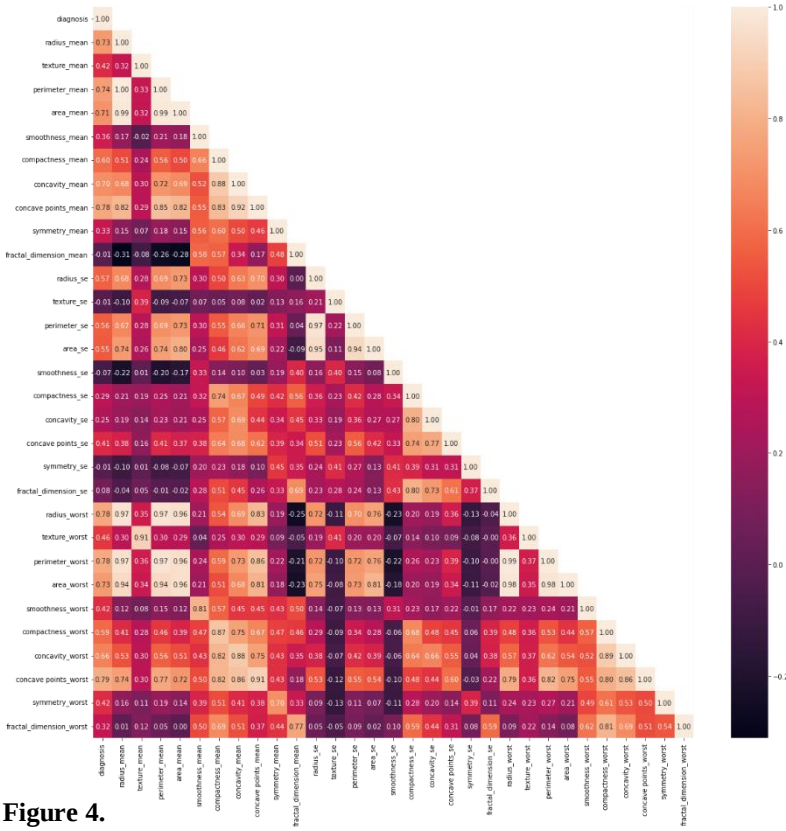

Figure 4.
Heat Map of all the correlations between columns.
Note: It is necessary to enlarge the image to visualize the values

For the prediction of breast cancer, only the correlation of the decision column with the other columns is relevant for the result, thus the correlation between each variable is ignored.

If we analyze deeply, it is possible to conclude that the highest correlation is 0.79 (between concave point worst and diagnosis), being a strong correlation. With that, we can conclude that the remaining variables will have moderate, weak, or null correlations.

Being these results and the objective of this project considered, none of the correlations in the heat map can be excluded from the research, because despite being small, the correlations end up affecting, even if minimally, the results.

4.3 Correlation's conclusion

After the study and analysis of the heat map, in the project I will use the following variables with the corresponding correlations (between the variable and diagnosis):

Radius mean: 0,73
Texture_mean: 0,42
Perimeter_mean: 0,74
Area_mean: 0,71
Smoothness_mean: 0,36
Compactness_mean: 0,60
Concavity_mean: 0,70
Concave_points_mean: 0,78
Symmetry_mean: 0,33
Fractal_dimension_mean: -0,01
Radius_se: 0,57
Texture_se: -0,01
Perimeter_se: 0,56
Area_se: 0,55
Smoothness_se: -0,07
Compactness_se: 0,29
Concavity_se: 0,25
Concave_points_se: 0,41
Symmetry_se: -0,01
Fractal_dimension_se: 0,08
Radius_worst: 0,78
Texture_worst: 0,46
Perimeter_worst: 0,78
Area_worst: 0,73
Smoothness_worst: 0,42
Compactness_worst: 0,59
Concavity worst: 0,66
Concave_points_worst: 0,79
Symmetry_worst: 0,42
Fractal_dimension_worst: 0,32

5 Methods

5.1 Methodology

The purpose of this study is to identify, within the dataset, which factors are the most important and the correlations between each other in order to obtain a model that precisely predicts if someone has the disease or not with pin point accuracy.

I started by importing/extracting our dataset with 33 attributes/columns and 569 instances/rows of information, including the column about Diagnosis column. Then, we proceeded to clean the data by checking for null values and removing them and checking the type of the variables to verify that we're working with what was expected. From pandas I was able to have a better visualization and insight about each attribute and their values.

I also used a seaborn *pairplot* with the *mean* columns and the diagnosis column to better understand and visualize the relation of each variable with diagnosis. I did not use every variable because it would make it hard to visualize, therefore, I opted to plot the mean values if each variable only.

To analyze which attributes I will use in the methods, I draw a heat map of the correlation between all columns, with only the correlation between columns and diagnosis being important. Because it is this relationship that will help achieve better results. The heat map made me conclude that all variables are important to obtain a better result regardless of the strength of the correlation.

The dataset was then divided into 55% and 45% for model training and testing purposes, respectively. The accuracy/efficiency of the model is obtained with the test dataset. To build this model I used several sklearn algorithms such as:

The algorithms:

1. Logistic Regression

The Logistic Regression model measures, through probabilities, the relationship between a categorical dependent variable and all the independent variables, in this case the dependent variable is the diagnosis, and all the others are the independent variables [9].

2. *Naïve Bayes*

Based on Bayes' theorem, Naïve Bayes calculates is a probabilistic machine learning algorithm that measures the probability of an event occurring given that another event has occurred, by assumption [10].

3. *Support Vector Machine (SVM)*

The purpose of a Support Vector Machine algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies data point [11].

4. *K-Nearest Neighbors*

KNN algorithm's purpose is to assess the correct class for the test data points, calculating the distance between these and the training data points [12].

5. *Decision Tree*

A decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, it's also widely used in machine learning [13].

6. *Random Forest*

A Forest Tree consists of many Decision Trees that operate as an ensemble, each individual decision tree gives an output and the output with the most votes become this model's prediction [14].

7. *Artificial Neural Network*

Its structure is formed by node layers, containing an input layer, one or more hidden layers, and an output layer. Each node connects to another node and has an associated weight

and threshold. If the output of any individual node is above the specified threshold value, the node sends data to the next layer of the network, otherwise, no data is passed on [15].

I measured the accuracy of each classification model to evaluate the performance of each one. The model's accuracy is the fraction of all the test dataset that the model predicted correctly.

5.2 Experiments

The experiments in this study are based in selecting features through the classification models implemented performances and I opted to give as the result the one with the best result. This evaluation, as said before, is based on each model's accuracy measured.

6 Results and Discussion

In the previous steps, multiple selections and choices were made on how to use the dataset to achieve a better result, with greater accuracy.

From the methods mentioned above and the variables selected from the correlations, it was possible to train the methods for certain random 55% of the data and achieve their *Training Accuracy:*

Models Accuracy:

[0] Neural Network Training Accuracy:
0.9968
[1]K Nearest Neighbor Training Accuracy:
0.9807692307692307
[2] Gaussian Naive Bayes Training Accuracy:
0.9519230769230769
[3]SVC Linear Training Accuracy:
0.9903846153846154
[4]SVC RBF Training Accuracy:
0.9807692307692307
[5] Logistic Regression Training Accuracy:
0.9903846153846154
[6] Decision Tree Classifier Training Accuracy:
1.0
[7] Random Forest Classifier Training Accuracy:
0.9967948717948718

After being trained, I was finally able to test the remaining 45% of the data in the methods and get the accuracy of each one shown below:

Models Accuracy:

[0] Neural Network Testing Accuracy:
0.9689
[1]K Nearest Neighbor Testing Accuracy:
0.9494163424124513
[2] Gaussian Naive Bayes Testing Accuracy:
0.9377431906614786
[3]SVC Linear Testing Accuracy:
0.9571984435797666

[4]SVC RBF Testing Accuracy:
0.980544747081712
[5] Logistic Regression Testing Accuracy:
0.9649805447470817
[6] Decision Tree Classifier Testing Accuracy:
0.914396887159533
[7] Random Forest Classifier Testing Accuracy:
0.9416342412451362

Through the calculated accuracy it is concluded that the best result is found using the SVC RBF method:

Model SVC rbf

	precision	recall	f1-score	support
0	0.98	0.99	0.98	160
1	0.99	0.96	0.97	97
accuracy			0.98	257
macro avg	0.98	0.98	0.98	257
weighted avg	0.98	0.98	0.98	257

Figure 5. Model SVC RBF

This method provides the best results of this work but still has flaws. If we compare the results predicted by the method with the true results, we can see the following graph:

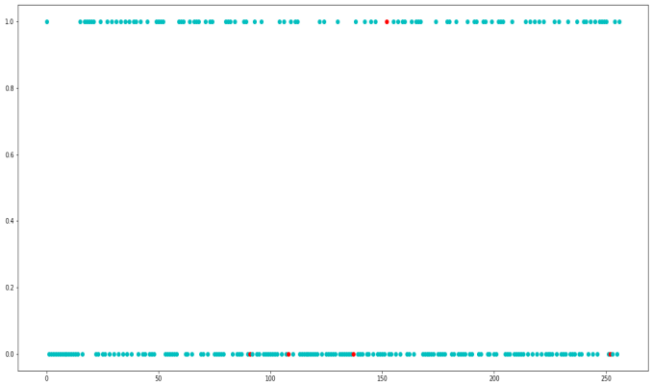

Figure 6. Prediction and True values
Note: It is necessary to enlarge the image to visualize the values

The graph shows that for the respective data the model fails 5 times, 4 of which represent false negatives. False negatives in breast cancer are at greater risk than false positives.

Although the method has an almost perfect precision and reaches almost all values, for medicine, it would be necessary

to have fewer false negatives for the project to be more reliable.

Even so, I can say that the results obtained were close to perfection, failing only 5 diagnoses out of 257.

7 Conclusions and Future Work

Breast cancer (BC) is a common cancer in women around the world, and early detection of BC can greatly improve prognosis and chances of survival by promoting early clinical treatment of patients.

In this project, I did a machine learning Python program to detect breast cancer from a dataset. The goal of this work is providing an early clinical prognosis that will be able to possibly help save lives just using data, Python, and machine learning.

In this paper I talked about the algorithms and techniques that I use to make this artificial intelligence and become it useful to predict cancer.

To conclude, if I had more knowledge about how to improve our results in the work context, maybe I could reduce the false negatives (even if in the worst case it induces more false positives) and if possible, improve the accuracy to reach a value close to 100%.

References

.

1. International Agency for Research on Cancer, Latest global cancer data: Cancer burden rises to 19.3 million new cases and 10.0 million cancer deaths in 2020, 15 December 2020 from https://www.iarc.who.int/wp-content/uploads/2020/12/pr292_E.pdf

2. Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer, from https://gco.iarc.fr/today/home

3. European Commission, Breast cancer burden in EU-27, European Cancer Information System, from https://ecis.jrc.ec.europa.eu/pdf/Breast_cancer_factsheet-Dec_2020.pdf

4. Cancer.net, Breast Cancer: Risk Factors and Prevention, July 2020, from https://www.cancer.net/cancer-types/breast-cancer/risk-factors-and-prevention

5. Polat, K., Günes S. (2006). Breast cancer diagnosis using least square support vector machine. Digital Signal Processing 17 (2007) 694–701. From https://reader.elsevier.com/reader/sd/pii/S1051200406001461?token=4628C65B06D2D675C621E5AA96F624445408EF01CB0535351515C66EF338B74E42B3DF976D94EECCECA90A18110CA3E8&originRegion=eu-west-1&originCreation=20211230190915

6. Paulin, F., Santhakumaran, A. (2011). Classification of Breast cancer by comparing Back propagation training algorithms. From http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.301.8837&rep=rep1&type=pdf

7. Dheeba, J., Singh, N., Selvi, S. (2014). Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. Journal of Biomedical Informatics 49 (2014) 45-52. From https://reader.elsevier.com/reader/sd/pii/S1532046414000124?token=C776CDAFB3458834333FD21CD7E3AA73B800BAE5A588D0B4C0A285684F1004432FBF812E53E1FCBE2F3AA02B5208C1B8&originRegion=eu-west-1&originCreation=20211230192914

8. Nahato, K., Harichandran, K., Arputharaj, k. (2015). Knowledge Mining from Clinical Datasets Using Rough Sets and Backpropagation Neural Network. From https://www.hindawi.com/journals/cmmm/2015/460189/

9. Logistic Regression - Detailed Overview, 15 March 2018, from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

10. Naïve Bayes Algorithm: Everything you need to know, from https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

11. Support Vector Machine, Introduction to Machine Learning Algorithms, 7 June 2018, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

12. *K*-Nearest Neighbors algorithm, from https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

13. Decision Trees in Machine Learning, 17 May 2018, from

https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

14. Understanding Random Forest How the Algorithm Works and Why it Is So Effective, 12 June 2019, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

15. Neural Networks by IBM Cloud Education, 17 August 2020, from https://www.ibm.com/cloud/learn/neural-networks