Classification of breast cancer using Machine Learning Algorithms

Houssam Benbrahim

Engineering Sciences Laboratory,
National School of Applied Sciences,
Ibn Tofail University, Kenitra,
Morocco

houssam.benbrahim@uit.ac.ma

Hanaâ HACHIMI

Systems Engineering Laboratory,
Sultan Moulay Slimane University,
Beni Mellal, Morocco

hanaa.hachimi@usms.ac.ma

Aouatif AMINE

Engineering Sciences Laboratory,
National School of Applied Sciences,
Ibn Tofail University, Kenitra,
Morocco

aouatif.amine@uit.ac.ma

***Abstract*—In Morocco, the number of cases of breast cancer is increasing, and this can become dangerous due to a delay in the diagnosis phase or failure in the prediction of the disease. We have built a solution capable of classifying breast cancer in order to help doctors, especially in Morocco, to better diagnose, detect, and quickly identify patients attacked by cancer to speed up the workflow. This work consists to produce a comparative study between the Logistic Regression algorithm and 10 machine learning algorithms using a breast cancer dataset. The results of the experimentation show that the Logistic Regression was achieved 94.74% of accuracy, which proves the capacity, performance, and efficiency of this algorithm.**

Keywords—Breast Cancer, Logistic Regression, Machine Learning , Accuracy

I. INTRODUCTION

Cancer disease is a major health problem requiring a comprehensive [1] care policy. Breast cancer is the number one cancer in women in the world, affecting 2.1 million women each year, and specifically targeting women between the ages of 40 and 70 [2]. This disease is considered to be an important subject of public health, however, most cases are preventable if detected early with better prediction [3]. The situation in Morocco is very worrying, breast cancer is not only the most frequent cancer in Moroccan women (36.1% for breast cancer) but causes a significant number of deaths due to error or diagnostic delay [4]. For a better medical examination of cancer, there are several techniques and tools for forecasting and decision support, among them we find Machine Learning (ML) algorithms.

ML is a way to model phenomena to make strategic decisions. It uses a variety of algorithms that iteratively learn from data to improve, describe inputs, and predict outcomes. ML problems can be differentiated into two categories, supervised and unsupervised [5]. It is a technology that can revolutionize the healthcare industry. ML algorithms can comprehensively assess cancer disease, achieving very high performance. In Morocco, the use of new technologies is very low in the health sector [6], which negatively affects the detection of cancer in general and breast cancer in particular. Morocco needs a national electronic health system that will allow doctors to better analyze and diagnose all diseases [7].

In this article, we create a comparative study between Logistic Regression and 10 ML algorithms using a database containing information on patients affected by breast cancer and examining the index of accuracy.

II. LOGISTIC REGRESSION

Logistic regression, logit model or binomial regression is a predictive technique [8]. It is a supervised classification algorithm popular in ML. LR a statistical approach that can be used to evaluate and characterize the relationships between a response variable of binary type, Y, and one, or more, explanatory variables, which can be categorical or continuous numeric type, X [9]. The formulation of the model is:

Either the dataset D made up of n pairs (x, y), with the description of an individual according to d descriptors, in the form of a real vector of size d, and y the membership class of this individual among 2 possible classes:

$$D = \{(x\ ,y\ ), (x_2, y_2), \dots, (x_n, y_n)\}\ \forall (x_i, y_i)\ _{\leq i \leq n} : x_i \in R^d, y_i \in \{0; 1\} \quad (1)$$

Logistic regression modeling the conditional probability $p(Y = 1|X = x)$ as follows:

$$p(Y = 1|X = x) = \frac{1}{1 + e^{-f(x)}} \quad (2)$$

We recognize the sigmoid function, S,also called logistic function:

$$S: \mathbb{R} \to [0; 1] \quad (3)$$

$$\alpha \mapsto \frac{1}{1 + e^{-\alpha}}$$

With $\alpha = f(x)$, which is a linear function of x:

(4)

$$f: \mathbb{R}^d \rightarrow \mathbb{R}$$

$$x \mapsto \omega_0 + \omega\ x\ + \omega_2 x_2 + \cdots + \omega_d x_d$$

Logistic regression is a powerful multivariate analysis method. It makes it possible to control for possible confusion bias. Its use is made easy by the use of statistical software [10].

III. Materials and Methods

In this article, first of all, we developed a comparative study between the Logistic regression algorithm and 10 ML algorithms using an open access Data Base "Breast Cancer Wisconsin (Diagnostic) Data Set" [11]. An ML algorithm is a method by which models appropriate for the application will have learned from the example data [12] There are many algorithms, we will choose a particular type of algorithm depending on the type of task we want to perform and the type of data available. Basically, in this study we have to use 30 different columns and 569 rows of a 144.5KB CSV file. We will predict the health status of patients, cases affected by breast cancer or not. To do this, we used the following list of algorithms: Logistic Regression (LR), Gaussian Naive Bayes (GNB), K Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), Linear Support Vector Classifier (SVC(linear)), Stochastic Gradient Descent (SGD), Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA), Neural Network (NN), and Extra Tree (TE). These algorithms are the most frequently used for this kind of problem. The choice of the latter comes after a long study and a very precise filtering. To analyze the data, we used the Python 3.7 programming language. The overall process of the experiment is illustrated in Figure 1.

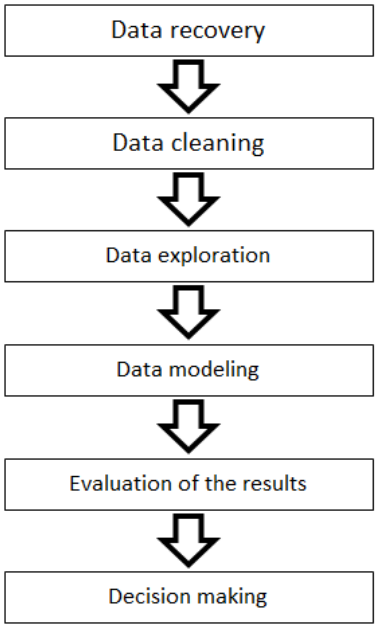

Fig. 1. The overall test process.

IV. Results

A. Data Explore

Regarding the database of our study, after importing the data, we noticed that the first column is the ID of type int64, the second is the diagnosis which means the health status of the patient after a medical examination: M (a malignant breast) and B (a benign breast) of object type, the last is NaN and the rest are different characteristics (30 columns) of type float64 which are: radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, and fractal_dimension_worst. Then we cleaned up the data, we removed the ID which represents the patient identifier, and Unnamed (NaN) which represents nothing.

First, we want to know the number of patients who are benign and malignant in the database. According to the results obtained, there are 357 benign cases with a percentage of 62.8% and 212 malignant cases with a percentage of 37.2%. The figure 2 corresponds to the results of the distribution of cases B/M.

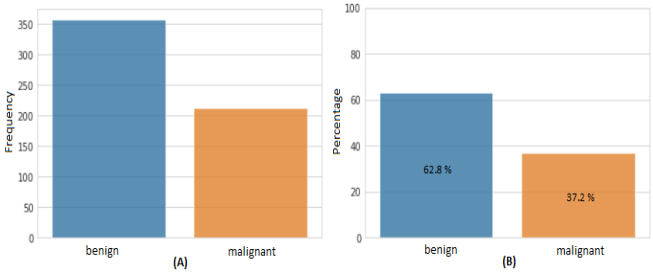

Fig. 2. Histogram of B / M cases. (A): frequency of diagnostic results; (B): percentage of diagnostic results.

Second, we have represented the different properties of the database. For this reason, we used multiple visualization using the density plot, we deployed both the data distribution and the degree of separation of the two sets of malignant and mild cases, in each entity direction. Figures 3, 4, and 5 visualize the characteristics of the tumor for positive and negative diagnosis.

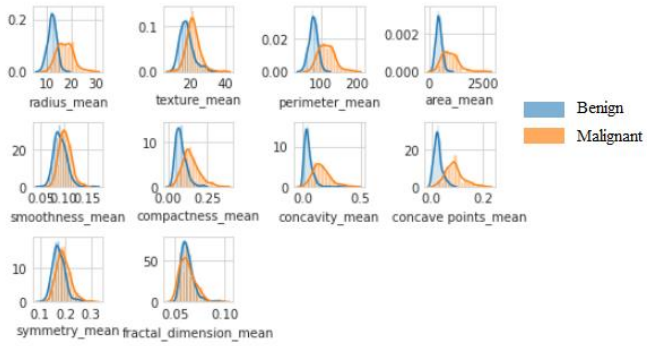

Fig. 3. Distribution of benign and malignant tumors of the "\ _mean" data group.

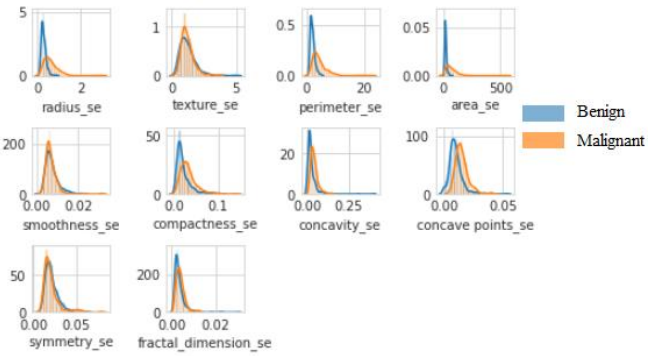

Fig. 4. Distribution of benign and malignant tumors of the "_se" data group.

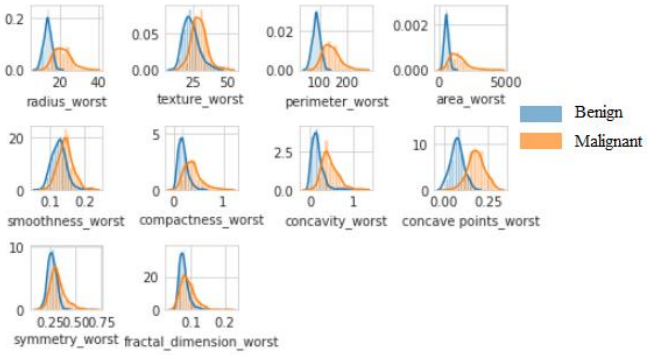

Fig. 5. Distribution of benign and malignant tumors of the "_worst" data group.

The values of radius_, perimeter_, and area_ can be used in the classification of cancer. Higher values of these parameters tend to show a correlation with malignant tumors. Values of smoothness_, symmetry_, or factual_ do not show particular preference of one diagnosis over the other. In any of the density plots there are no noticeable large outliers that warrant additional cleaning. Third, to perform a correct and very successful analysis, it is necessary to determine the correlation matrix between the different instances of the database. The latter designates the proximity between two variables and establishes a linear relationship between them. This correlation is represented in the figure 6.

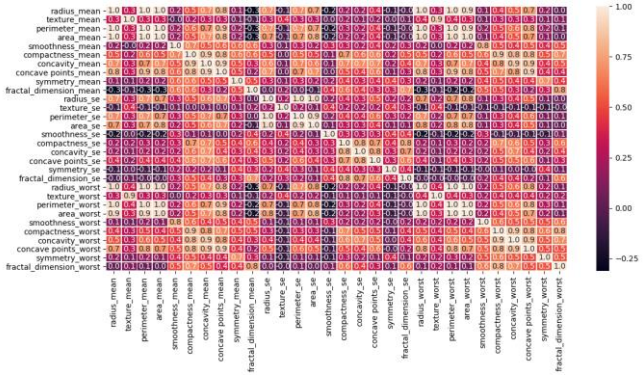

Fig. 6. The correlation between all features of the database.

From the correlation matrix shown in figure 6, we can deduce that there is a linear relationship between several variables. In addition, there are several parameters which are negatively correlated and others are not correlated. Figures 7, 8, and 9 show an example for each correlation case.

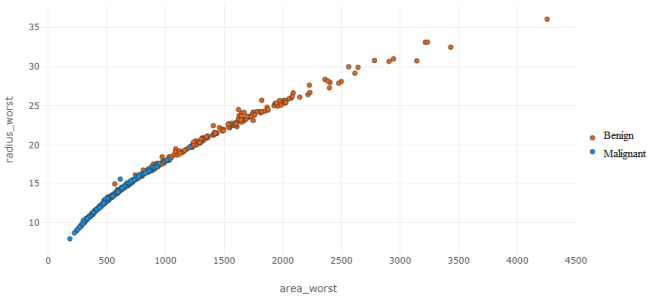

Fig. 7. Example of positive correlation.

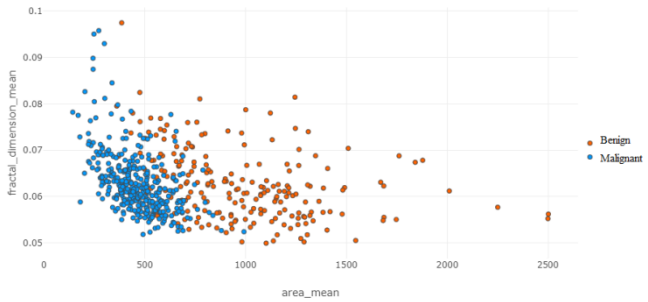

Fig. 8. Example of negative correlation.

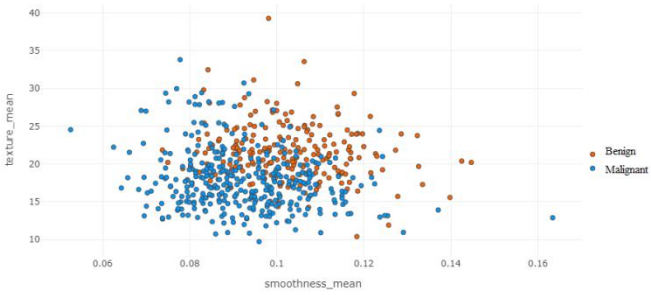

Fig. 9. Example uncorrelated attributes.

The highest correlations are between:

perimeter_mean and radius_worst, perimeter_mean and area_mean, radius_mean and area_mean, radius_mean and perimeter_mean, radius_worst and area_mean, radius_worst and radius_mean, area_worst and radius_worst, area_worst and are _mean, perimeter_worst and radius_mean, perimeter_worst and perimeter_mean, perimeter _worst and area_mean, perimeter_worst and radius_worst, perimeter_worst and area_worst, perimeter_se and radius_se, and area_se and radius_se.

B. Experiment

For a better classification of breast cancer, the database has been divided into two phases: the first is the training phase with a percentage of 80%, and the second is the test phase with a percentage. by 20%. For our experiment we only used the best 6 characteristics, which are: area \ _mean, perimeter \ _mean, radius \ _mean, radius \ _worst, perimeter \ _worst, area \ _worst. We calculated the performance index of the Logistic Regression and 10 ML algorithms for our experiment by measuring the accuracy. Accuracy is the percentage of correct predictions for the test data. The figure 10 visualizes the variation of the score of the accuracy of the LR and 10 algorithms for 6 characteristics.

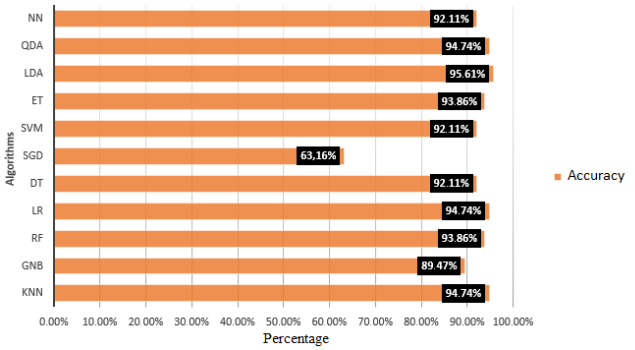

Fig. 10. The index of the accuracy of the Logistic Regression and 10 ML algorithms for 6 attributes.

The results of the first experiment show that most of the algorithms tested achieved high performance over 92% except GNB which scored 89.47% and SGD which achieved only 63.16%. What is important that the LR is scored an important ranking 94.74% of the accuracy for the study experience.

V. DISCUSSION

During this study, we established a classification of breast cancer using ML algorithms. Regarding this classification, we have launched a comparative study between the Logistic Regression and 10 ML algorithms and we can deduce that LR is a reliable, powerful and efficient algorithm. It kept very high results for our experience exercised for this problematic, 94.74% of the accuracy. This method achieves a very advanced capability of predicting breast cancer in the database. in another work a comparison of six ML algorithms: LR, KNN, GRU-SVM, Multilayer Perceptron, Softmax regression, and SVM. The results of this comparison show that all the ML algorithms tested performed very well (all the algorithms exceeded 90% test precision) on the [13] classification task. In similar work [14], the percentage of accuracy of neural classifiers reaches almost 98% for the diagnosis of breast cancer. Also, in another comparison, the SVM and KNN classification techniques achieved an accuracy of 98.57% and 97.14% [15]. From this comparison of different experiments performed in the same database, it can be deduced that the LR algorithm has proven its prediction capacity and performance, and that the small differences that appear in the percentage of the accuracy between the different work is due to the way of selecting the characteristics which represent a strong and positive relationship.

VI. CONCLUSION

In this paper, we have developed a comparison between the Logistic Regression algorithm and 10 ML algorithms. The implementation of this list of algorithms was performed using the University of Wisconsin Breast Cancer (Diagnosis) Database. The aim is to classify patients who may be carriers of this disease between mild and malignant using diagnostic characteristics. In the experiment we chose 6 attributes. Firstly, the results show that the accuracy of all the algorithms tested was very important except that the SGD algorithm which achieved a low success rate. Secondly we saw that the LR algorithm scored important results for and that it achieved very high performance. We can conclude that the LR algorithm has a very advanced capacity for classifying patients between benign and malignant in relation to cancer disease.

REFERENCES

[1] Bekkali, R. LUTTE CONTRE LE CANCER AU MAROC L'APPORT DE LA FONDATION LALLA SALMA. International Journal of Medicine and Surgery, vol. 4, no 1, p. 55, 2017.

[2] World Health Organization: Breast Cancer. https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[3] International Agency for Research on Cancer. World cancer report: World Health Organization. IARC Press, 2003.

[4] Programme de détection précoce. http://www.contrelecancer.ma/en/detection_precoce_action.

[5] Hurwitz, J., & Kirsch, D. Machine learning for dummies. IBM Limited Edition, 75, 2018.

[6] Benbrahim, H., Hachimi, H., & Amine, A. Survey on the Use of Health Information Systems in Morocco: Current Situation, Obstacles and Proposed Solutions. In International Conference on Advanced Intelligent Systems for Sustainable Development, Springer, Cham, 2018. p. 197-204, 2018.

[7] Benbrahim, H., Hachimi, H., & Amine, A. Moroccan Electronic Health Record System. In Proceedings of the International Conference on Industrial Engineering and Operations Management Paris, France, 2018.

[8] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. An introduction to logistic regression analysis and reporting. The journal of educational research, vol. 96, no 1, p. 3-14, 2002.

[9] Hoffman, J. I. Biostatistics for medical and biomedical practitioners. Academic press. 2015.

[10] El Sanharawi, M., & Naudet, F. Comprendre la régression logistique. Journal français d'ophtalmologie, vol. 36, no 8, p. 710-715,2013.

[11] Dua, D., & Graff, C. UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California,

School of Information and Computer Science, zuletzt abgerufen am: 14.09. 2019.

[12] Lahmiri, S. On simulation performance of feedforward and NARX networks under different numerical training algorithms. In Handbook of research on computational simulation and modeling in engineering, IGI Global, p. 171-183, 2016.

[13] Agarap, A. F. M.: On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. In: Proceedings of the 2nd International Conference on Machine Learning and Soft Computing. ACM, Phu Quoc Island, Viet Nam, p. 5-9, 2018. doi: 10.1145/3184066.3184080

[14] Anagnostopoulos, I., Anagnostopoulos, c., Rouskas, A., Kormentzas, G.,Vergados, D.: The Wisconsin Breast Cancer Problem: Diagnosis and DFS time prognosis using probabilistic and generalised regression neural classifiers. Oncology Reports, Special Issue Computational Analysis and Decision Support Systems in Oncology. vol. 15, no 4, p. 975-981, 2006.

[15] Islam, M. M., Iqbal, H., Haque, M. R., Hasan, M. K.: Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors. In: IEEE Region 10 Humanitarian Technology Conference (R10-HTC). IEEE Press, Dhaka, Bangladesh, p. 226-229, 2017. doi: 10.1109/R10-HTC.2017.8288944