ISSN 1870-4069

Disease Prediction Applying Machine Learning: Case Study of Breast Cancer and Diabetes

Edgar Gonzalo Cossio Franco[1], María de los Ángeles Núñez Herrera[2], Keuri Adilene Machain Tarula[3], Carlos Daniel Robles Ontiveros[4], Javier Agustín Ramírez Martínez[3], Marco Julio Franco Mora[5], Manuel Iván Estrada Chávez[5]

[1] Instituto de Información Estadística y Geográfica de Jalisco, Mexico

[2] Instituto Tecnológico de la Piedad, Mexico

3 Instituto Tecnológico del Sur de Nayarit, Mexico

[4] Instituto Politécnico Nacional, Mexico

[5] Instituto Tecnológico Superior de Ciudad Hidalgo, Mexico

Abstract. Breast cancer and diabetes are part of the leading causes of death in the world [1]. According to the World Health Organization, Cancer ranks second while diabetes ranked seventh in 2016 [2]. Through the application of artificial intelligence (AI) it is possible to train a machine to predict future scenarios in terms of determining a positive or negative diagnosis of cancer or diabetes based on historical data. The objective of this research is to implement AI through Machine Learning (ML) for predictive purposes regarding breast cancer and diabetes and thereby diagnose in time. For the present study, linear regression and the J48 algorithm were used.

Keywords: Cancer, diabetes, artificial intelligence, machine learning, linear regression, j48 algorithm.

1 Introduction

Machine learning is an artificial intelligence technique that combines a set of algorithms with the purpose of training machines. This training allows them to learn and based on it discover patterns that allow you to predict scenarios based on historical data.

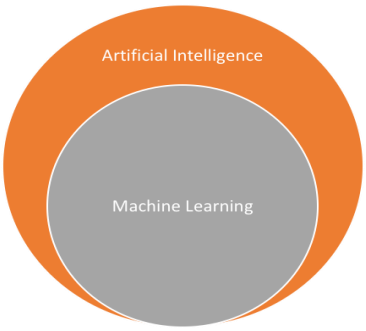

Fig. 1. Machine Learning and artificial intelligence.

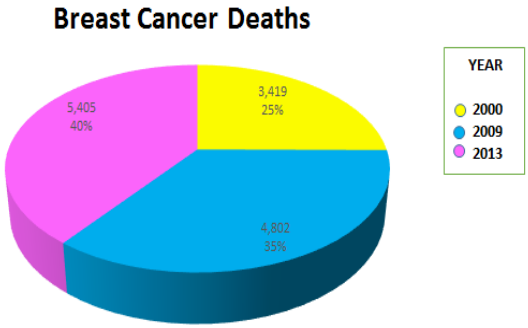

Fig. 2. Increase in deaths with breast cancer.

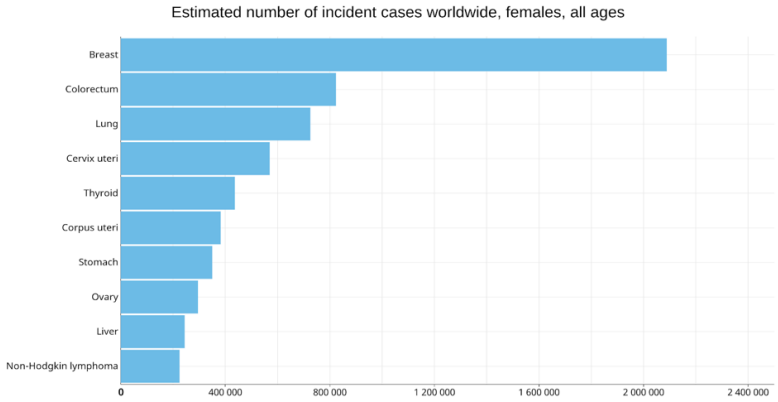

Fig. 3. Deaths worldwide according to type of cancer [5].

Among the machine learning algorithms it is possible to find: decision trees, Naive Bayes, logistic regression, K-means algorithm, linear discriminant analysis, support Vector Machines, Isotonic Separation, Random Forests, Neural Networks, Genetic Algorithms, among others [3].

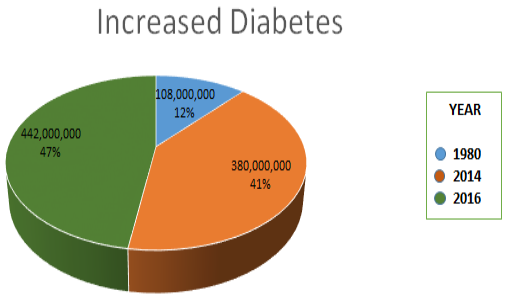

Fig. 4. Increase in deaths due to diabetes worldwide [7].

In recent years, machine learning has been applied with the objective of predicting scenarios in the health, banking, financial, educational and in general all those fields where there are large volumes of data (Big data) [4]. In figure 1 it is possible to see the location of machine learning with respect to artificial intelligence.

In the field of health, artificial intelligence, particularly machine learning, plays an important role because every day unstructured data is generated that allow the identification of patterns through grouping, analysis, segmentation, disease processing and prediction; In the case of the present investigation, we work with two particular cases: cancer and diabetes.

Cancer ranks second worldwide worldwide while diabetes ranks number seven in 2016, according to WHO data [1, 2]. Under this scenario it is important to work on strategies that contribute to the early prevention strategy for possible eradication.

1.1 Problem

When a cancer is detected after its onset, most of the time there is little to do. In figure 2 it is possible to see deaths worldwide according to the types of cancer.

Regarding diabetes, WHO reported that the disease was ranked 7 worldwide in terms of deaths, which represents 422 million people, corresponding to one in 11 people [6].

The reality is that it is an alarming fact and demands that the strategy be taken more seriously. The cause of diabetes occurs when there is a high level of blood sugar and the pancreas is not able to produce insulin or there is but in minimal amounts [7]. The increase in deaths from diabetes was 62 million in just two years, as shown in figure 3.

1.2 Related Work

To carry out this research, work was consulted that has to do with the prediction of cancer and diabetes through machine learning.

In [8] an investigation was carried out with different Machine Learning methods, the research was carried out with an integrated trend of mixed data, such as clinical and genomic, ensuring that the application of such data in machine learning methods could improve the accuracy of Cancer sustainability, recurrence and survival prediction.

Fig. 5. General process.

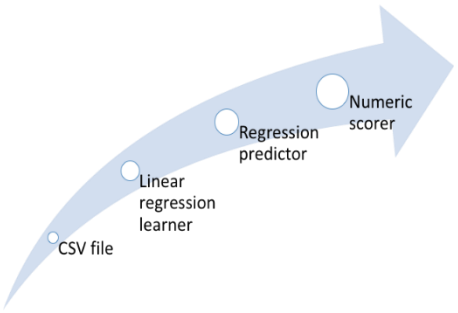

Fig. 6. Dataset treatment.

Table 1. Variable cancer.

mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
17.99	10.38	122.8	1001.0	1.184	0
20.57	17.77	132.9	1326.0	8.474	0
19.69	21.25	130.0	1203.0	1.096	0
11.42	20.38	77.58	386.1	1.425	0
20.29	14.34	135.1	1297.0	1.003	0
12.45	15.7	82.57	477.1	1.278	0
18.25	19.98	119.6	1040.0	9.463	0
13.71	20.83	90.2	577.9	1.189	0
13.0	21.82	87.5	519.8	1.273	0
12.46	24.04	83.97	475.9	1.186	0
16.02	23.24	102.7	797.8	8.206	0
15.78	17.89	103.6	781.0	971	0
19.17	24.8	132.4	1123.0	974	0

On the other hand, in [9], applying different supervised and unsupervised learning methods, the regression is highlighted as the most accurate method for a more accurate result using the Gradient Descent algorithm, which consists in that the cost function is reduced when the model adjusts its parameters. Another method used is neural networks, which are inspired by the biological neuronal system, although they do not work in the same way. This method receives data in a layer called input layer. The data

will be purchased and the model automatically identifies the characteristics of the data and labels or names them. This is the first model he selects for his research for the discrimination of malignant or benign tumors among patients with breast cancer, built with many hidden layers to better generalize the data.

The research of [10] applies methods of support vector machines and generalized discriminant analysis for classification of information with an accuracy ranging from 78.21 to 82.05 percent in the prediction of diabetes disease.

In [11], an algorithm based on neural networks is applied for the detection of breast cancer, inspired by social search algorithms of animals which offers promising results even above the particle accumulation algorithm. The k-means method is applied with modifications in [12], as well as the Support Vector Machine method is applied for the prediction of diabetes with an accuracy of 99.64%.

2 Methodology

For the present investigation, we worked with the methodology shown in Figure 4 where the process of obtaining, analyzing, training and results of the data set is described. The linear regression method and decision trees were used.

2.1 Process

As part of the process that guided the present investigation, data exploration was necessary, for which the dataset of Pima indigenous women of at least 21 years of age is used, in the case of diabetes. In the case of cancer, the data were obtained from the Hospitals of the University of Wisconsin, Madison, from Dr. William H. Wolberg. The data set is from the National Institute of Diabetes and Digestive and Kidney Diseases.

In this process, the data were obtained for analysis and processing. In the case of diabetes, the variables are composed as:

Diabetes: Number of pregnancies, Age, Pedigree, Plasma, Blood Pressure, Insulin in the body, Body mass, Skin thickness. In the case of cancer, the following variables are available.

2.1.1 Data

Cancer: Radius of the tumor, Texture of the tumor, Perimeter of the tumor, Area of the tumor and softness of the tumor.

Both datasets were processed as shown in Figure 5 where four strategic moments are identified: the dataset reader (CSV file), linear regression learner, regression predictor and numeric scorer.

2.1.1.1 CSV

Cancer: In the cancer dataset the variables to contribute are very specific since the data obtained are from tumors extracted from the body of people. Table 1 shows an extract of the variables.

Mean Radius: it is the measure of the radius of the extracted tumor that is given in millimeters, the radius can be large or small and it is not determined that it is benign or malignant until it is in conjunction with the other variables.

Mean Texture: It is also known as the average grayscale on radiographs or tomographs performed on the tumor. Grayscale depends a lot on the texture of the tumor.

Mean Perimeter: this measure is the perimeter of the tumor, as it is well known that the tumors are neither round nor oval at all but it is a deformed fat pellet therefore when it has a perimeter greater than 2 cm it is said to be accurate malignant, but that does not always turn out that way until they are related to the other variables and have a more accurate prognosis.

Mean Area: this variable is the most important in a medical analysis and diagnosis of the tumors in conjunction with the perimeter and texture of the tumor, since if its area exceeds 5 cm or the size of a lemon it is probably that you have cancer. In most cases, where the area is 5 cm it is not a single tumor anymore, if they are not several together as a bunch of grapes, diagnosed as advanced cancer where despite the treatment there is no longer a cure.
Mean Smoothness: Average variation in radio lengths, as mentioned above, the tumors have an irregular shape, which takes an average of variations between the different possible radii of these.

Diabetes: The diabetes dataset shows variables that have to do with the factors that can determine it. Table 2 shows an extract.

Preg: pregnancy is the number of pregnancies that women have had regardless of whether or not they had abortions. The number of pregnancies in every woman is important because they create new substances during and after pregnancy, which if they reproduce to a greater extent or do not reproduce cause serious side effects on health.

Plas: is the plasma glucose concentration at 2 hours in an oral glucose tolerance test. When the plasma glucose is above 11.1 mmol/dl it is diabetes.

Pres: diastolic blood pressure (mm Hg) commonly called high blood pressure or hypertensive people who suffer from it. High blood pressure shows a very high prevalence in type 2 diabetes mellitus and is a risk factor for the development of cardiovascular complications. Strict control of blood pressure to figures less than 130/80 mm Hg reduces cardiovascular and renal morbidity and mortality to a greater extent than the control of other complications.

Skin: Thickness of the triceps skin fold (mm).

Insu: 2-hour serum insulin (mu U / ml) is a hormone that takes glucose from the blood and transports it into the body's cells where it is used as energy. Diabetes occurs when the pancreas does not produce enough insulin or when the body does not use insulin properly.

Mass: Body mass index (weight in kg / (height in m) ^ 2). Diabetes favors the appearance of muscular atrophy. Given this scenario, researchers have discovered that

Table 2. Diabetes variables.

preg	plas	pres	skin	insu	mass	pedi	age	class
6	148	72	35	0	33.6	627	50	tested_positive
1	85	66	29	0	26.6	351	31	tested_negative
8	183	64	0	0	23.3	672	32	tested_positive
1	89	66	23	94	28.1	167	21	tested_negative
0	137	40	35	168	43.1	2.288	33	tested_positive
5	116	74	0	0	25.6	201	30	tested_negative
3	78	50	32	88	31	248	26	tested_positive
10	115	0	0	0	35.3	134	29	tested_negative
2	197	70	45	543	30.5	158	53	tested_positive
8	125	96	0	0	0	232	54	tested_positive
4	110	92	0	0	37.6	191	30	tested_negative
10	168	74	0	0	38	537	34	tested_positive
10	139	80	0	0	27.1	1.441	57	tested_negative
1	189	60	23	846	30.1	398	59	tested_positive
5	166	72	19	175	25.8	587	51	tested_positive
7	100	0	0	0	30	484	32	tested_positive

an increase in blood sugar levels triggers the decrease in muscle mass. In addition, they have observed that the abundance of the transcription factor KLF15 increased in the skeletal muscle of diabetic mice.

Pedi: Pedigree function of diabetes (genetics). There is a complex genetic transmission of diabetes. A predisposition to the disease is inherited to which different environmental triggers are added. This is true for both diabetes 2 and 1, although more marked in the first.

Age: Age of the people with whom the research was conducted in the indigenous Pima group.

2.1.1.2 Linear Regression Learner

Method used to train the machine with the dataset from the set of independent variables to a dependent. Perform a multivariate linear regression.

2.1.1.3 Regression Predictor

It is the node that predicts the response using the regression model. This node must be connected to CSV reader and the Linear Regression learner. It is executed as long as the test data contains the columns related to each other. The node adds at the end of the dataset a column to the input table that contains the prediction for each row; depending on the case of the dependent variable.

2.1.1.4 Numeric Scorer

This node statistically calculates the values of the numerical columns called r¡ and the predicted values known as p¡. The calculated variables are as follows:

R^2: It is a statistics used in the context of statistical models whose main purpose is the prediction of future results. In short, it can be defined as the precision in which the prediction is made according to the dependent and independent variable:

$$1\text{-SS res/SS tot} = 1\text{-}\sum(p\ i\text{-}r\ i)^2 / \sum (r\ i\text{-}1/n\ *\sum r\ i)^2 .$$

Error Absolute Medium: measurement of difference between two continuous variables:

$$(1/n * \sum \mid p\ i\text{-}r\ i).$$

Error Quadratic Medium: sum of the squares of the waste. It is a measure of the difference between the data and an estimation model:

$$(1/n * \sum (p\ i\text{-}r\ i)^2).$$

Error Root Mean Square: As used in the differences between the values predicted by a model or an estimate and the observed values. Represents the square root of the second sample of the differences between the predicted values and the observed values:

$$(\text{sqrt}\ (1/n * \sum (p\ i\text{-}r\ i)^2)).$$

Difference with middle sign: It shows that it summarizes how well a set of estimates equals the quantities, which they must estimate. It is a statistics that is used to evaluate an estimation procedure:

$$(1/n * \sum (p\ i\text{-}r\ i)).$$

2.1.2 T2 (Training and Test)

Before applying the Machine Learning method, a Data Science is performed that is responsible for cleaning the dataset to make a more accurate prediction and the nodes can be configured correctly.

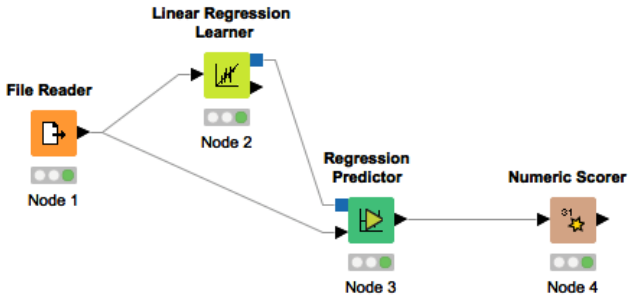

Fig. 6. Model for prediction.

Linear Regression

The Linear Regression method is the relationship between independent or explanatory variables and dependent or response variables. Which will allow obtaining a prediction of the dependent variable or response based on the given values of the independent variable.

The linear regression is expressed by the following expression:

$$Y=Mx + n.$$

J48 Algorithm

It is an induction algorithm, which generates a rule or tree structure from subsets or case windows extracted from the training dataset. Its form of processing is based on generating a structure of rules and assesses its goodness using criteria that measure the accuracy in the classification of cases using two main criteria to direct the process:

1. Calculates the value of the information provided by a candidate rule or branch of the tree, with a routine called info.
2. Calculate the overall improvement that a branch or rule provides using a routine called gain or benefit.

3 Results

The results obtained in the research were satisfactory for the topics discussed since they are the main causes of deaths worldwide in women, where statistics increase every day also spreading in men and children.

The predictions and classification of data with the Linear Regression, Decision Tree j48 methods have the following precision:

- Linear Regression:
- Cancer, R^2=0.93,
- Diabetes, R^2=0.95.

4 Conclusion

Through the application of machine learning, such as linear regression and J48 algorithm properly trained and tested, it was possible to establish a reliable model of prediction of breast cancer and diabetes in .93 and .95 percent, respectively, which is encouraging as it opens the possibility of application and testing with other methods that make the prediction stronger.

5 Future Work

It is contemplated to continue working with different tools offered by artificial intelligence such as:

Neural networks of single layer and multilayer for the prediction and prevention of high impact issues in society; criminal incidence and causes of maternal death during pregnancy.

The possibility of georeferencing the criminal incidence through heat maps is contemplated. Another objective is to create dashboards that show the real-time information of a fact linked to the dataset.

References

1. Who: Cancer key facts, from WHO (2019)
2. Who: Cancer key facts, from WHO (2019)
3. Abreu, P.H., Santos, M.S., Abreu, M.H., Andrade, B., Silva, D.C.: Predicting breast cancer recurrence using machine learning techniques: A systematic review. ACM Computing Surveys, 49(3), pp. 1–40 (2016)
4. Lugo-Reyes, S., Maldonado-Colín, G., Murata, C.: Artificial intelligence to assist the clinical diagnosis in medicine. Alergia México Magazine, 61(2), pp. 110–120 (2014)
5. Who: International Agency of Research on Cancer: Who (2019)
6. Who: International Agency of Research on Cancer: Who (2016)
7. Who: International Agency of Research on Cancer: Who (2014)
8. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Photoiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal (2015)
9. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I.: Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal (2016)
10. Polat, K., Güneş, S., Arslan, A.: A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. Expert Systems with Applications (2006)
11. He, S., Wu, Q.H, Saunders, J.R.: Breast cancer diagnosis using an artificial neural network trained by group search optimizer. Transactions of the Institute of Measurement and Control, 31(6), pp. 517–531 (2009)
12. Afzali, S., Yildiz, O.: An effective sample preparation method for diabetes prediction. International Arab Journal of Information Technology (IAJIT), 15(6), pp. 968–973 (2018)