Science, Engineering and Technology *Volume 1, Issue 2, Pages 16-23*
www.setjournal.com *https://doi.org/10.54327/set2021/v1.i2.16*

A practical implementation of machine learning in predicting breast cancer

Ajna Fetić, Adnan Dželihodžić

Polytechnic Faculty at University of Zenica, Zenica 72000, Bosnia and Herzegovina

Abstract

Cancer is the leading disease in the world by the increasing number of new patients and deaths every year. Hence, it is the most feared disease of our time. It is believed that lung cancer and breast cancer are the most common types of cancer and they both are subtypes of the same group of cancer – carcinoma. With this type of cancer, early detection is of great importance for patient survival. As it is a disease that has unfortunately been around for many years, today we have datasets with all necessary information for diagnosing and predicting cancer. Predicting cancer means deciding if the cancer is malignant or benign. The key to this answer lays in different values of parameters that have been stored when the disease was discovered. Machine learning plays a crucial role in predicting cancer, given the fact that algorithms such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), etc. are designed to find the pattern that occurs in large sets of data and based on that makes a decision. In this paper, the author's goal is to see how machine learning and its practical implementation on public datasets can help with an early breast cancer diagnosis and hopefully help save more lives.

Keywords*: breast cancer, machine learning, logistic regression, support vector machine, decision tree, random forest, regression decision tree*

1 Introduction

Breast cancer is statistically the lead cause of many deaths of women around the world. Despite the popular belief that only women can get breast cancer, scientific evidence shows otherwise. Even though there are small numbers of men who are diagnosed with this disease, they go through the same procedures and risks as women. With breast cancer, as well as any other cancer, early diagnosis is crucial, first and foremost, for patient survival. Cancer is a disease that has been around for many years, with growing number of patients. Fifty years ago, diagnosing and predicting cancer was done by one expert in medical field who would do tests like mammogram, ultrasound and MRI.

The ultrasound is a technique in which the sound waves are sent through the skin inside the body to be able to observe the condition inside. "*A transducer that emits sound waves is positioned on the skin and the echoes of the tissues of the body are captured with the bounce of sound waves*" [1]. These results are then converted into grey scale image by binary values. Positron emission tomography (PET) imaging scan is used to detect early signs of cancer, heart disease or brain disorder by injecting radioactive tracer to mark diseased cells. F-fuorodeoxyglucose, when injected, enables doctors to realize the position of a tumour [3]. Another widely used technique is dynamic MRI which enables analysis of blood vessels, although it is usually associated with metastasis. Mammography uses X-ray scanning in order to get tissue status. However, as years went by, more parameters for diagnosing cancer were required and the whole process became complex. Now, doctors have to go through numerous tests and make connections between parameters in order to predict the cancer – decide if the cancer is benign or malignant and try to not be bias. Due to imperfections in human mind and lack of objectivity, necessity for automatization arose. To illustrate, "*in mammography, the doctors should read a high volume of imaging data which reduces the accuracy"* [1].

Over the past years, there have been written many papers on how machine learning can help in diagnosing and predicting cancer. Reason why application of machine learning in this field has been thought of is because of its core principals and wide usage. Machine learning is a branch of Artificial intelligence which uses different algorithms and goes through large data sets in order to "learn" and give desired outputs. There are three different types of machine

Published: 30 October 2021
Corresponding author: *Adnan Dželihodžić (adnandz@gmail.com)*

learning which will be discussed in later chapters. As a deduction, machine learning represents a big step forward in diagnostic medicine, especially in cancer prognosis where large datasets are present. The goal of implementing machine learning in diagnostic medicine is not to replace human doctors, but to reduce inevitable human error and help doctors with quicker diagnosis.

"*There are several machine learning techniques that can help to predict whether the person affected is having a benign or malignant cancer, and this process would be efficient and without any errors"* [2]. As we all know, disease does not spare anyone and time is of the essence, particularly for oncological patients to whom a day early diagnosis can literally save life.

2 What is breast cancer from medical point of view?

„Cancer is a generic term for a large group of diseases that can affect any part of the body" [3]. Apart from term cancer, this group of diseases is also referred to as malignant tumour or neoplasms. What makes a disease a cancer is rapid creation of abnormal cells that grow beyond their usual size or boundaries and over time invade other parts of the body – other organs. The state where abnormal cells reach other organs is called metastasis and are the leading cause of death in oncological patients.

There are different types of cancer: carcinoma, melanoma, sarcoma, leukemia, lymphoma, myeloma and brain and spinal cord tumours. Carcinoma is the most common type of cancer and is a cancer of epithelial cells which cover inside and outside surfaces of human body.

But how does one get a cancer? Cancer development starts from normal cells in human body that started their transition into tumour cells by undergoing "*multi-stage process that generally progresses from a pre-cancerous lesion to a malignant tumour*" [3]. Transition from pre-cancerous lesion to malignant tumour is greatly affected by persons genetic factors and three groups of external agents:

- physical carcinogens, such as ultraviolet and ionizing radiation.
- chemical carcinogens, such as asbestos, components of tobacco smoke, aflatoxin (a food contaminant), and arsenic (a drinking water contaminant); and
- biological carcinogens, such as infections from certain viruses, bacteria, or parasites [3].

The Figure 1 illustrates high numbers of new cases diagnosed with breast cancer in the year 2020. The next Figure 2 shows the world map with mortality rates of breast cancer in the same year.

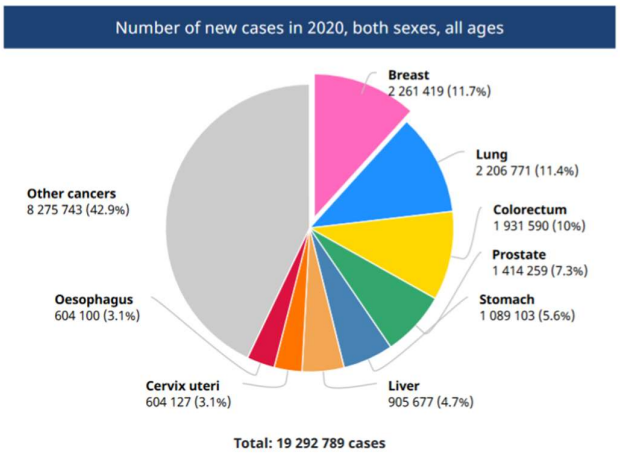


Figure 1. Number of cases in the world [4]

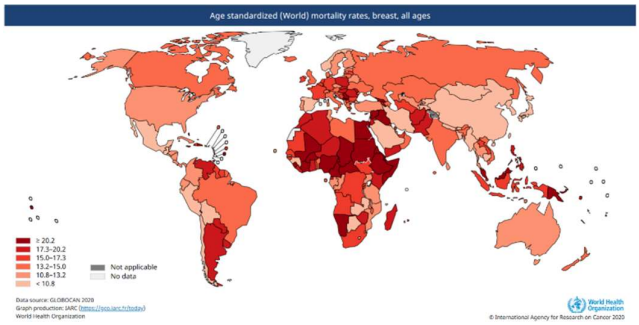


Figure 2. Mortality rate [4]

3 What is machine learning?

Machine learning is subset of artificial intelligence that uses sample data, known as „training data ", to build mathematical models that can make a prediction or decision without being explicitly programmed to do so. Definition on what learning means from mathematical point of view is the following: "*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T , as measured by P, improves with experience E*" [5].

The basic concept of machine learning lays in statistical learning and optimisation models that allow computer to analyse large datasets and find patterns. The aim of machine learning is to establish a regressor or classifier by using training dataset and then evaluate the performance of the regressor or classifier through test set.

Based on the nature of training data, machine learning is classified as:

- Regular or Euclidean structured data learning
- Unsupervised learning – uses training set of unlabeled data
- Semi-supervised learning – uses training data that has small number of labelled data and large amount of unlabeled data
- Reinforcement learning
- Transfer learning

3.1 Machine learning methods

Based on different machine learning methods for modelling the data for underlying problems, there are three types of machine learning:

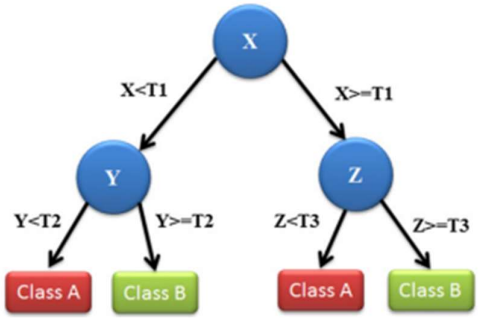
- Machine Learning Methods based on Network Structure
- Machine Learning Methods based on Statistical Analysis
- Machine Learning Methods based on Evolution

3.2 A preview of Machine learning algorithms

First algorithm that will be talked about is Linear regression. Linear regression is a regression algorithm that follows under algorithms of supervised learning and is mostly used for modelling continuous variables and for predictions. In Regression, output data values are determined by the values of the input data. The simplest form of Regression is Linear Regression where the input data is attempted to fit a straight line and that is only possible when the relationship between the data is linear. "*It shifts focus from statistical modelling to data analysis and pre-processing*" [6]. The biggest disadvantage of Linear regression is that oversimplifies the real-world problems. Linear regression is used for: prediction of price of real-estate, forecasting of sales, prediction of students' exam scores, forecasting of movements in the price of stock in stock exchange.

Second algorithm that is going to be covered is Logistic regression. It deals with the problems of classification. "*It gives the binomial outcome as it gives the probability if an event will occur or not (in terms of 0 and 1) based on values of input variables*" [6]. The advantages of Logistic regression are: efficiency – computational and from training perspective, simplicity of implementation and ease of regularization. The output of Logistic regression is probability score, and based on that score, we should determine some metrics depending on our business, so we could obtain the cut-off which could be used for classification. The disadvantages of Logistic regression are: "*inability to solve non-linear problem as its decision surface is linear, prone to over fitting, will not work out well unless all independent variables are identified"* [6]. Logistic regression is used for predicting the risk of developing a certain disease, cancer diagnosis, predicting mortality and in engineering for prediction the success of a certain system, process or product.

Another very popular algorithm is Decision Tree. It is a Supervised Machine Learning algorithm to solve both classification and regression problems. It is designed in a way where data is being continuously split based on a certain parameter and decisions are placed in the leaves while the data is being split in the nodes. Regression tree uses continuous data and Classification tree uses data which are categorical – the output is Yes or No. Advantages of using Decision Tree are: easy to interpret, easy to handle categorical and quantitative data, can be used both for classification and regression, has best capability for filling missing data with most probable value and has high efficiency due to traversal algorithm. Disadvantages of this algorithm are difficulty to control the tree size, it can be unstable, prone to error sampling and gives local optimal solutions instead of global. Decision Tree is used for predicting cancer and predicting future use of library books. Figure 3 presents an example of Decision Tree.



An illustration of a DT showing the tree structure. Each variable (X, Y, Z) is represented by a circle and the decision outcomes by square (Class A, Class B). T (1-3) represents the thresholds (classification rules) in order to successfully classify each variable to a class label.

Figure 3. Example of Decision Tree [7]

4 Implementation

For this paper we decided to use the public dataset from UCI (University of California Irvine) Machine Learning Repository called *Breast Cancer Wisconsin (Diagnostic) Data Set* [8] and ML.NET technologies to create a program that predicts breast cancer.

ML.NET is a free, open-source and cross platform machine learning framework that can easily be integrated in .NET applications in both online and offline scenarios. The core of ML.NET is a machine learning model which specifies steps that are necessary to transform the input

data into machine learning prediction. It is a framework that runs on Windows, Linux and MacOS using .NET Framework. For this paper, we used a Windows 10 Pro operating system on 11th generation i5 processor with iRISX^e intel graphics. The ML.NET framework works by the following code workflow:

- *Collect and load training data into an IDataView object*
- *Specify a pipeline of operations to extract features and apply a machine learning algorithm*
- *Train a model by calling Fit() on the pipeline*
- *Evaluate the model and iterate to improve*
- *Save the model into binary format, for use in an application*
- *Load the model back into an ITransformer object*
- *Make predictions by calling CreatePredictionEngine.Predict()* [9]

The workflow is shown in the Figure 4. The first thing we need to do is prepare our data, because there might be missing data in one or more columns or some data in given columns might have undesirable values. In our case, in .csv (comma separated value) file of dataset that we downloaded, there were three columns which were poorly named: *concave_points_mean, concave_points_se* and *concave_points_worst,* that had a space in the naming which would cause certain problems in our machine learning algorithms. After renaming, we looked at our data – we had 569 records of which 357 were Benign and 212 Malignant (Figure 5).

The dataset had ten real-valued features that were computed for each cell nucleus:

a) radius (mean of distances from centre to points on the perimeter)
b) texture (standard deviation of grey-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1).

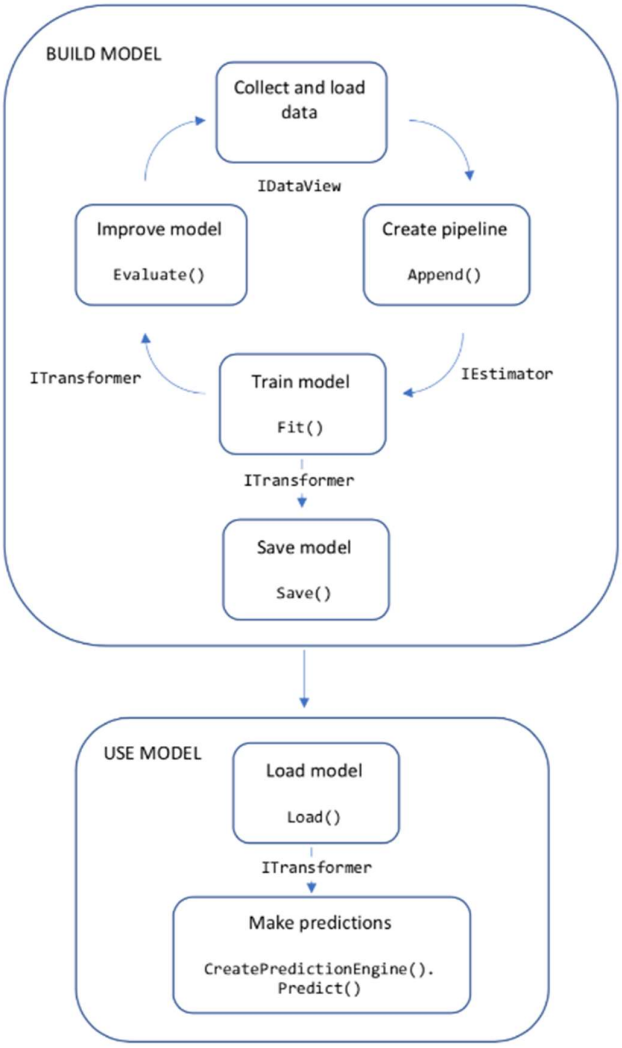


Figure 4. Diagram of the ML.NET code workflow [9]

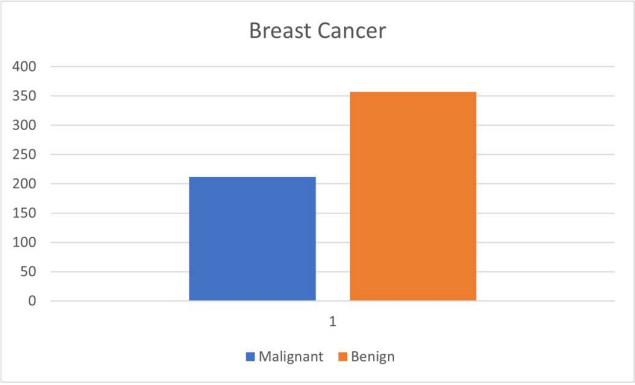


Figure 5. Graphic representation of dataset in Excel

After examining our data, we proceeded with implementation. First step was creating standard .NET web application with home controller and index view and then we added C# console application. The .NET web application will be used for testing our trained algorithm, while the console application is used to test the accuracy of algorithms on our data and choosing the right one for our web application. When web application is installed, we need to install NuGet package ML.NET and ML.NET.FastTree. ML.NET package enables us to use libraries necessary for machine learning, while ML.NET.FastTree enables us to use Decision Tree algorithms such as Binary Tree and Random Forest.

Before creating our first algorithm and pipeline, we need to create a second class that will store all the data that is loaded – "*CancerData.cs*". This class has only two properties: label and features.

After creating console application and "*CancerData.cs*" class, we added a class called "*MachineLearningAlgorithms.cs*" where all the methods with machine learning algorithm are implemented. The following part of the code shows the properties and constructor of the class. Our class has seven properties that are common for all algorithms. `trainData`, `testData` and `data` have type *IDataView* which is an interface for data that is used for machine learning. `trainData` is used to store data that is used for training the algorithm after splitting initial dataset and the `testData` is similarly used to store data meant for testing the algorithm after splitting the dataset. Property `data` is where the data set is being stored. MlContext class enables us to use all functions related to preparing data, creating pipeline etc when implemented. Since the column we are trying to predict is the type of *String* in our dataset, we need to map it into respectful values. For this, we used dictionaries. One is mapping values "M" from the *diagnosis* column to Boolean value true, and "B" value to the Boolean false. It is the `targetMapB` dictionary and is used for BinaryClassification algorithms since the value that these algoritms predict can be true or false. The other dictionary, `targetMapR`, maps values "M" to number 1 and values "B" to 0 and is used for Regression. The `MODEL_FILEPATH` property is used specify the location in which our algorithm will be stored.

In the constructor of this class, we firstly initialise MlContext class and then load our data from the .csv local file. The following line tells the ML.NET that it should use our "CancerData.cs" class while loading the data: `context.Data.LoadFromTextFile<CancerData>.` After creating the MlContext and loading the data, with the usage of local variable `splitTrainTestData` to store it and the argument `testFraction: 0.2` of `TrainTestSplit` function that is a part of Data class accessible through our context property, we are splitting the data set into two groups 80% is for training and 20% is for testing. The argument `seed: 0` sets randomness to zero. After that, we simply set these groups of data to respectful properties.

4.1 The implementation of algorithms

First algorithm that was implemented was Logistic Regression. First thing we need to do for all the implemented algorithms is create a pipeline. The pipeline is a set of APIs that allow the developer to pre-process data, model training and deployment. Pipeline consists of two parts: transformers and estimators. A transformer is an abstraction that changes data form one DataFrame to the other by calling `transform()` method. DataFrame is an API developed by Spark SQL that supports many basic and structured data types, since ML is working with data of different types. DataFrame labels data into columns that are typically named "text", "label" and "feature". The Transformer works by appending values to columns usually as follows:

- *A feature transformer might take a DataFrame, read a column (e.g., text), map it into a new column (e.g., feature vectors), and output a new DataFrame with the mapped column appended.*
- *A learning model might take a DataFrame, read the column containing feature vectors, predict the label for each feature vector, and output a new DataFrame with predicted labels appended as a column.* [10]

An Estimator is an abstract that performs learning algorithms and fits or trains on data. The standard for the way an Estimator works is calling the Fit() method that accepts imported DataFrame and produces a model from the data. The model is Transformer. In this case, a Logistic Regression algorithm is an Estimator and when we call Fit() method, it creates LogisticRegressionModel which is a Transformer. This can be seen in the following line of code:

```
ITransformer model = pipeline.Fit(trainData);
```

In general, pipeline represents stages our data goes through and each stage has its own Transformer and Estimator. Stages are run in order and initial DataFrame is modified with each stage. After our initial DataFrame goes through pipeline, we get PipelineModel that is used for training data. This is shown in the following Figure 6:

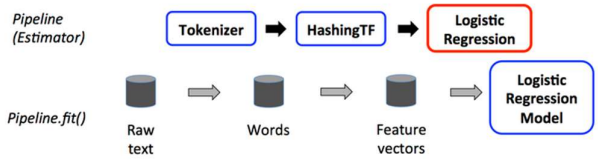


Figure 6. Initial DataFrame goes through pipeline [10]

The Transformers are highlighted with blue border and Estimators with red. After creating the pipeline, pipelineModel has the same number of stages, but all Estimators became Transformers. This is shown in the Figure 7, where we can see that all the stages are highlighted with blue border.

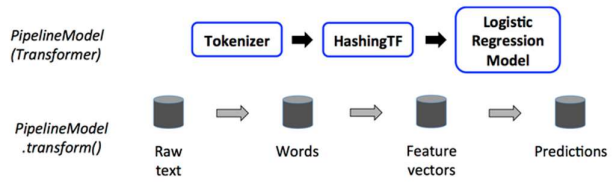


Figure 7. DataFrame goes through pipelineModel [10]

This is done in the following line of code:

```
var metrics =
context.BinaryClassification.Evaluate(model.T
ransform(testData));
```

Here we also created variable metrics which contains the object that is returned when calling the function `Evaluate()`. Metrics is very important in choosing the right algorithm and will be discussed in the next chapter.

After writing all metrics evaluation parameters in our console window, we called the function `CrossValidate()` and forwarded our data, created pipeline and set the `numberOfFolds` to 5. What this function does is it splits our data in random sets of test and train data in the proportion of 20 to 80 and it does it 5 times – specified in the argument `numberOfFolds`. This is done since accuracy of our algorithms can depend on how the data is split into test and train sets and it gives us more credible results.

The second algorithm was Support Vector Machine, and it was implemented in the same way. The only difference between the SVM and Fast Tree algorithm and Logistic Regression and Decision Tree is that the first two algorithms mentioned require some value for probability for the function Evaluate so we decided to use function `EvaluateNonCalibrated()` which does not require any probability.

In our implementation we demonstrate why we do not use the regression algorithms for binary classification. Same as classification, regression algorithms require creating the pipeline. This pipeline also has Transformer and Estimator, but instead of calling the `BinaryClassification` class, we call `Regression` class. Here we decided to use Decision Tree algorithm to better see the difference. After creating methods that implement the algorithms, we need to save the model.

5 Results and discussion

In order to choose the right algorithm for our data set, we need to look at the accuracy of every algorithm that we used. As previously stated, every algorithm has access to properties that contain their accuracy, and these accuracies were written in the console window. The property accuracy *"is the ratio of number of correct predictions to the total number of input samples"*, the F1 score is "*the harmonic mean of the precision and recall"* and LogLoss "*measures the performance of a classification model where the prediction input is a probability value between 0.00 and 1.00"* [9]. Entropy is a measurement of uncertainty or disorder and the closer the number is to one, there is higher level of disorder (meaning low level of purity). The values of accuracy and F1 score are desired to be as close to one as possible and the value of LogLoss should be as closer to 0.00 as possible. In the chapters below we will show each algorithm's accuracy.

Logistic regression was the first algorithm that we used, and it had the accuracy of Cross Validation of 97.28%:

```
----------Training Model Logistic Regression----------

----------Testing model----------

----------Score----------
Accuracy: 0.9743589743589743
F1 Score: 0.9620253164556963
Log Loss: 0.152663794897633
Entropy: 0.9344491365829437
----------Cross-Validation----------
Mean cross-validated accuracy score: 97.29%
```

Figure 8. Accuracy of Logistic Regression

In Figure 8, we can see that this algorithm had high values of F1 score and Accuracy – 0.96 and 0.97, respectfully. The value of LogLoss is also desirable, since it has the value of approximately 0.15. Another parameter that is used to determine how good a machine learning algorithm is performing is Confusion matrix. Confusion matrix has two rows: positive and negative and two columns: positive and negative. The way this table is used is that it tells us how many patients were misdiagnosed. In Confusion matrix we have the following terms:

- True positive – measures the proportion of actual positives that are correctly identified as such
- True negative - measures the proportion of actual negatives that are correctly identified as such
- False positive - test result which incorrectly indicates that a particular condition or attribute is present
- False negative - test result that indicates that a condition does not hold, while in fact it does.

This algorithm had 38 True positives and 76 True negatives, while it had 0 False positives and 3 False negatives. This means that 0 patients were falsely diagnosed as having malignant tumour and 3 were falsely diagnosed as not having cancer or that their tumour was benign.

5.1 Support Vector Machine

Support Vector Machine in our implementation had the score of 92.21%. Since this algorithm was implemented without the values for probability, we only have values for accuracy and F1 score that are, compared to Logistic Regression, not as good. SVM has values of 0.88 and 0.84, and these are lower by 0.08 and 0.12, respectfully. This algorithm misdiagnosed 12 patients as cancer patients and 2 patients as having benign tumour. Based on this we can conclude that this algorithm is not the right solution for our dataset.

5.2 Decision tree

Just like Logistic Regression, Decision Tree had high accuracy score of 96.57%. This algorithm had very high values of accuracy and F1 score, but unlike Logistic Regression, it had less desirable values for LogLoss. From the Matrix, we can see that this algorithm misdiagnosed only 4 patients – one as having cancer and tree as not having cancer.

5.3 Random Forest

As previously stated, Random Forest is the algorithm is like Decision Tree since it works by implementing several Decision Trees. So, it is a no surprise that this algorithm has an accuracy score of 95.43%.

Since this algorithm works with probabilities, just like SVM, we only have accuracy and F1 score. Compared to the Decision Tree algorithm and Logistic Regression, F1 score of 0.93 and accuracy of 0.95 are not the desirable values. This is better seen in the Confusion matrix. This algorithm misdiagnosed 3 people of not having cancer and 2 of having it.

5.4 Regression Decision Tree

We have already mentioned that Regression is never used for Binary Classification, but we wanted to graphically show the difference in accuracy. The Cross Validation accuracy for this algorithm was 85.30%. This score is very much undesirable. Regression algorithms do not have Confusion matrix.

6 Conclusion

To deduct, cancer is a disease that is threatening all humankind. Since there is no cure for cancer, the best way to beat the disease is early detection. This is done by multiple medical tests that give us information about the tumour. As previously stated, not every tumour is cancer, but many have tendencies to become one. The best way to examine test results is proven to be Machine Learning. Depending on the type of data that are provided, different algorithms are desirable. For our dataset, the Logistic Regression algorithm proved to have best accuracy score – it had the lowest number of wrongly diagnosed patients. This is especially important when human lives are in question, since every misdiagnosis poses a risk for patients' life. The goal of Machine Learning is not to replace the expertise of medical doctors and medical practitioners, but to give them a helping hand in setting a diagnosis when there are too many parameters that need to be considered and thus reduce the error or misdiagnosis.

With the rate at which the technology is developing and integrating into our lives, using Machine Learning in medical diagnosis seem like a natural thing to do. It poses a revolution in medicine that is lonely awaited for, just like industrial revolution. Human beings are constantly evolving in every aspect of their life so it is only natural to believe that the aspects of human doing will evolve too.

References

[1] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan and M. N. Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques," *Springer Nature Singapore,* p. 14, 1 September 2020.

[2] T. Thomas, N. Pradhan and V. S. Dhaka, "Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey," *Proceedings of the Fifth International Conference on Inventive Computation Technologies,* p. 5, 2020.

[3] WHO, September 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer.

[4] "Global Cancer Observatory," September 2021. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf.

[5] X.-D. Zhang, "A Matrix Algebra Approach to Artificial Intelligence," p. 803, 23 May 2020.

[6] S. Ray, "A Quick Review of Machine Learning Algorithms," p. 5, September 2019.

[7] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal 13,* p. 10, 2015.

[8] "UCI Machine Learning Repository," September 2021. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

[9] Microsoft, September 2021. [Online]. Available: https://docs.microsoft.com/en-us/dotnet/machine-learning/.

[10] "Apache Spark," September 2021. [Online]. Available: https://spark.apache.org/docs/latest/ml-pipeline.html.