PATIENTS SURVIVABILITY PREDICTION AFTER BREAST CANCER SURGERY USING SUPERVISED MACHINE LEARNING TECHNIQUES

Patrick Tshiaba Tshiaba
Student Number
1744742

Godsgift Uzor Ozioma
Student Number
1654209

Abstract

The goal of this paper is to predict the survival of breast cancer patients, who had undergone surgery through machine learning techniques that will determine the chances of living or not living above five years post-surgery.

1. Introduction

Symbol of femininity, the breast is an organ like any other, with its operational problems (breast pain, nipple discharge ...). In some cases, more serious problems may appear, such as the presence of cancer. Breast cancer affects one in eight women, especially from the age of 50. 5 to 10% of breast cancer are from an inherited genetic origin; 85 to 90% of cases (so-called sporadic or non-hereditary form) have environmental or unknown origins. A significant proportion of sporadic breast cancers are induced by the use of hormones, estrogen and progesterone contained in contraceptives or menopausal treatments. Risk factors are also the consumption of alcohol, sugar, dairy products, hydrogenated fats, obesity, lack of physical activity. A first late pregnancy and no breastfeeding would also promote this cancer.

According to medical protocols, the treatment may use surgery, radiotherapy, chemotherapy. The combination of these care can be supplemented by the hormone therapy.

1.1. Cancer

A cancer is a disease characterized by the presence of abnormal cells that multiply uncontrollably. The chances of recovery depend on the type of cancer and its stage of evolution at the time of treatment.

"Cancer" is a general term referring to several diseases that share the fact that certain cells of an organism adopt an abnormal behavior characterized by :

- Independence from signals that normally stimulate cell multiplication;
- Insensitivity to anti-proliferative signals and mechanisms;
- A proliferative capacity that is no longer limited (growth to infinity);
- The disappearance of the phenomenon of apoptosis;
- An abnormal ability to induce angiogenesis;
- And the acquisition of invasiveness and metastasis.

The resulting new so-called "cancerous" or "tumor" cells can form a malignant tumor (a neoplasm) or spread through the body.

1.2. Breast cancer

In the case of breast cancer, cancerous cells can remain in the breast or spread through the body via the blood or lymphatic vessels. Most of the time, the progression of breast cancer takes several months and can even take few years.

Breast cancer is the most diagnosed cancer in women worldwide, both before and after menopause1. Most often, breast cancer occurs after 50 years. The 5-year survival rate after diagnosis ranges from 80% to 90%, depending on age and type of cancer. One in eight women will develop breast cancer in her lifetime and 1 in 24 women will die from it. The number of people affected has increased slightly but steadily over the last 3 decades. On the other hand, the mortality rate has steadily declined over the same period, thanks to advances in screening, diagnosis and treatment.

Men also have breasts that are less developed than those of women. Breast cancer in humans is rare. Less than 1% of all breast cancers affect men. However, it is important for men to know that they may be affected by this cancer, especially in order not to overlook the symptoms.

2. Our approach

In this paper we use machine learning techniques over patients datasets to predict howmany of them can survive less or more than 5 years after surgery treatment.

Accurate prediction of breast cancer survivability can scientifically enable healthcare providers to make more informed decisions about a patients treatment.

Several data-driven machine learning methods have been used in for cancer prediction and prognosis.These methods learn patterns or statistical regularities from historic data in order to make predictions on new data. Specifically for breast cancer, researchers have used a wide variety of machine learning methods for predicting susceptibility [57], diagnosis [814],recurrence [1521] and survivability. For this task, some of the researchers who developed machine learning models had access to patients genomic and detailed clinical data from medical centers on which they trained their methods [2224,29,30]. Although smaller in size (in the order of a few hundred cancer incidences), these datasets were more detailed in patient information. But most other researchers with no access to such detailed patient data used the publicly available SEER cancer dataset [31] for training their methods [2528]. We have used Haberman's Survival dataset from UCI Machine Learning repository `https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival` for this paper. Although this dataset does not include genomic or detailed clinical information, but is still suitable for building accurate models for survivability.

Several methods have been used by researchers for predicting breast cancer survivability; the most common are Artificial neural networks (ANN) [32], support vector machines(SVM) [33], decision Trees [34] and logistic regression [36].

3. Material and method

The following list of material was used to realize the task on this paper:

- Haberman's Dataset;
- jupyter notebook with Python3;
- Previous work (see references);
- Stage-specific predictive models for breast cancer survivability;
- Predicting breast cancer survivability: a comparison of three data mining methods.

The dataset contains information about the survival status of the patients who had undergone breast cancer surgery. There are 306 samples with 3 attributes for each.

The task in this paper was materialized using machine learning methods such as K-NN, logistic regressiont andrandom fores.

3.1. K- Nearest Neighbor Algorithm

K-NN is one of the most used learning algorithms. K-NN uses a database in which data points are seperated into several classes to predict the classification of a new sample point. K-NN is non-parametric given that it does not make any assumptions on the underlying data distribution. The data determines the model structure. Our choice of selecting K-NN as a classifier for this classification study is as a result of we not having prior knowledge about the distribution data especially as it is more related to the medical field. In K-NN there is explicitly little or no training phase and the learning phase is pretty fast. K-NN keeps most or all of the training data required during the testing phase. Hugely based on feature similarity which is a trait shared with fellow neighborhood scheme random forest. Thus, the closer the similarity of our out sample features are to our training set, determines how we classify a given data point. Objects are classified by a majority vote of its neighbors.

- Error = 0.21;
- Accuracy = 0.77;

3.2. Logistic Regression

is a probabilistic function being developed that can give us a chance for an input to belong to any of the classes we have.

Logistic regression looks at the elements in a training set with their corresponding labels and the sigmoid function which is the function of theta. We minimize the cost function to find out the values of theta for which we have our error function minimized. Once we have our probability function which is still our sigmoid function then we can correctly predict our target outcome;

We found out the following result in our logistic regression:

- MSE = 0.136;
- Accuracy = 0.728;

3.3. Random forest

similarly known as random decision forests are ensemble learning method for classification and regression, operates by the construction of multiple decision trees during training and giving an output which is the mode of the classes or mean prediction for classification and regression respectively of individual trees. Random forest tends to correct the over-fitting habit of decision trees on their training set. Decision trees is pretty popular for numerous machine learning tasks. When the trees are grown very deep they tend to learn highly irregular patterns subsequently leading to over-fitting their training sets meaning they tend to have very high variance with low bias. With random forest multiple deep decision trees are averaged by training on different part of the

training set and the goal is reducing variance. This is done by making a small increase in the bias which like in various cases leads to loss of interpretation but it does generically boosts the performance of the final model is a pretty decent way.

Random forest makes use of tree bagging algorithm or bootstrap aggregating to tree learners. Bagging repeatedly selects a random sample, with this procedure the model performs better as it ensures decrement in the variance of the model leaving the bias without increment. Random forest slightly differs from bagging and the general decision tree scheme because it uses a modified tree learning algorithm that selects at each candidate split in the learning process a random subset of features which is referred as feature bagging. If one or a few features are very strong predictors for the target output it will be selected in many of the B trees, making them become correlated.

Adding a step of randomization by adding another tree yields extremely randomized trees, which are trained using the bagging and random subspace method, also the top-down splitting in the tree learner is randomized.

Random forest builds several decision trees and merges them together to get a more accurate and stable prediction. Can be used for both regression and classification problems.

The random-forest algorithm brings extra randomness into the model, when it is growing the trees. Instead of searching for the best feature while splitting a node, it searches for the best feature among a random subset of features. This process creates a wide diversity, which generally results in a better model.;

- Error = 0.25;
- Accuracy = 0.75;

4. Challenges encountered

The challenges encountered came mostly as a result of no prior knowledge of the dataset as it is from the medical field, so we had to learn a few things about cancer, the causes and the chances of survival of cancer patients then narrowed our study down to breast cancer. There has not been a lot of work done with the Herberman dataset. We did have to make a few modifications to the dataset to enable it be readable and usable and predictable when looking at the target output. Another challenge of note is that as a group it was not so straight forward classifying the survival chances of the patients as a result of the randomness of the data as a result we got curious to find out how these models will fare with the data set.

5. Conclusion

This paper report is a research effort in health sector to find the outcome of post surgery survival of breast cancer patients. We worked on a dataset with 306 samples and 3 features. After we have processed over all the dataset, we come to the conclusion that our classifiers shows us that there are more patients living more than 5 years considering that the dataset came from a time where health care was not particularly great. In recent times there are more chances of survival post cancer survival. We can say that this prove enough that surgery is a good treatment for this disease.

References

[1] Rohit J. Katea, Ramya Nadig *Stage-specific predictive models for breast cancer survivability.*

[2] Dursun Delen, Glenn Walker, Amit Kadam. *Predicting breast cancer survivability: a comparison of three data mining methods.*

[3] Street, W. Nick, Olvi L. Mangasarian, and William H. Wolberg. *"An inductive learning approach to prognostic prediction." In Machine Learning Proceedings 1995, pp. 522-530. 1995.*

[4] Delen, Dursun, Glenn Walker, and Amit Kadam. *"Predicting breast cancer survivability: a comparison of three data mining methods." Artificial intelligence in medicine 34, no. 2 (2005): 113-127.*

[5] Kate, Rohit J., and Ramya Nadig. *"Stage-specific predictive models for breast cancer survivability." International journal of medical informatics 97 (2017): 304-311.*

[6] Ltsch, Jrn, Alfred Ultsch, and Eija Kalso. *"Prediction of persistent post-surgery pain by preoperative cold pain sensitivity: biomarker development with machine-learning-derived analysis." British Journal of Anaesthesia 119, no. 4 (2017): 821-829.*

[7] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. *"Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." In International Conference on Intelligent Computing, pp. 878-887. Springer, Berlin, Heidelberg, 2005.*