Analysis of Breast Cancer Wisconsin Original Data Set

A data science approach to classifying breast cancer using machine learning models

# Data Science Team

The team is made up by:

* Lucas Withmire
* Edgardo Panza
* Federico Rueda L.

# Objective

Apply several concepts learned in the machine learning II lecture to the selected dataset and use different models to make an effective approach to the task.

# Data description

For this project we choose a dataset related to breast cancer detection. The breast cancer represented the 12.3% of the new cases diagnosed worldwide in 2018, according with the American Institute for Cancer Research. Designing new procedures or methods to detect early this disease will allow patients to increase the probability of survive.

The main characteristics provided by the dataset creator are shown in the figure below:

Graphical user interface, text, application

Description automatically generated

Figure. <http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29>

# Exploratory Data Analysis (EDA)

The first step is performing an EDA to the dataset, this will give us the general idea of what amount of data we have and its main characteristics. Also, we can decide if the dataset has the right information to approach the task.

The dataset consists of 569 observations and 31 variables. We imported the column names into R package and construct the dataset with the right attributes:

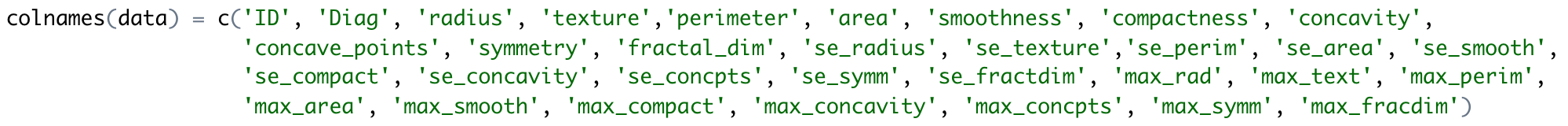


Figure. text

## Classification Task

As we can see, the first one is the unique ID from the patient and the second one “Diag” corresponds to the diagnosis (benign or malignant). So initially we can count on 29 variables in order to predict the diagnosis of the patient related with breast cancer.

**ML task:** binary classification

**Target Variable:** “Diag”

**Values:** “Benign” or “Malignant”

So, we need to perform at first a label encoding to map these two possible values into integers for the ML model to work well. In some cases, this is done by the model function itself in the R package.

## Missing values

We need to check if there are missing values and how we will handle them.

> table(is.na(data))

FALSE

17639

## Duplicates

It is important to identify if there are duplicates in the dataset.

> table(duplicated(data))

FALSE

569

## Imbalance

In a classification task, it is important to identify if we have an imbalance problem with the data. Much greater number of observations for one class will lead into a model bias, so it could minimize the loss function predicting the class with the most of observations. We need to train a model with balanced information and picking the right metrics to assess the two models.

> table(data['Diag'])

B M

357 212

There is a ratio of almost 6:10 “Malignant” to “Benign” diagnoses ratio in the dataset. We can say there is a slight imbalance between these two classes, but we assume it will not be a concern.

## Sensitivity vs. Specificity

It is also important to understand that in some cases, i.e., in medical diagnostics, there is a class that is more important to identify that the other in terms of the safety of the patient. In this case, it is more important to successfully identify the malignant cases, so if there is a false positive it is less dangerous that if there is a false negative that therefore put in risk the patient.

**Sensitivity:** ability of the model to correctly identify patients with malignant diagnostic

**Specificity:** ability of the model to correctly identify patients with benign diagnostic

In this case we need a model that better Sensitivity than Specificity preferably.

## Relations between target variable and other variables

Because our target variable is categorical, one way of checking the relations with other is to plot distributions of predictor variables and target variables. One example is the following:

Chart, histogram

Description automatically generated

Figure. Text

Correlation between “Diag” and the 2-11 variables:

A picture containing text

Description automatically generated

Correlation between “diagnosis” and the 12-21 variables:

A picture containing text

Description automatically generated

Correlation between “diagnosis” and the 22-31 variables:

A picture containing text

Description automatically generated

# Mathematical overview of models

Note: This is the mathematical overview of the two ML methods used

For our dataset, we decided to train two models that allow us to classify between “Benign” and “Malignant” diagnostic of breast cancer:

* Support Vector Machine (SVM)
* Artificial Neural Network

## Support Vector Machine

(Note: we can specify that we will use a SVC)

SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform the data and then based on these transformations it finds an optimal boundary between the possible outputs using the concept of separating hyperplane. It should not be a perfect one for the training dataset (overfitting) but one with a better generalization capability. This is achieved by selecting the hyperplane which distance to the datapoints is the largest (James et. al).

For our SVM we are going to use a “linear kernel” because it is recommended to be used when we have large number of features. Specifically, we will use an specific model called Support Vector Classifier (SVC).

The SVC allows us to select some data points in the wrong side of the boundary (misclassified) to let the model performs better with less influence from individual data points. For this purpose, the model uses the slack variables to adjust the amount of misclassified data points. In our case, we will use the variable C (cost) as the budget that will penalize the model.

## NEURAL NETWORK

Text

# Fitting process

This is a description of your fitting process including, a summary of how you arrived at your final model, the choice of hyper-parameters and how made this choice.

First, we separate our data into training and test, 80% and 20%. For this type of dataset, we are going to create 2 models, a SVM Classifier and a Neural Network.

## SVM

The accuracy of the model for training data is 99.6 % and for the test data is 98.2 %.

Even we had good results we wanted to confirm that were some options to improve the model:

Diagram

Description automatically generated

Figure X.

The previous process selected 16 out of the 30 features:

A picture containing text, receipt

Description automatically generated

With this 16 features we got an accuracy of 98.6% in the training set (1% less that our previous model) but we got the same accuracy of 98.2 % in our test set.

## Artificial Neural Network

Text

# Comparison of two methods

Text.

# Include Graphical presentation (included)

Include

# Conclusions

These are…

# Citation Request

This breast cancer databases were obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. If you publish results when using this database, then please include this information in your acknowledgements. Also, please cite one or more of:

1. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

2. William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

3. O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

4. K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).