Udacity Machine Learning Engineer Nanodegree

Predictive Analysis of Mammographic Masses Data

Henrique Dal Mora Rosendo da Silva
February 14, 2019

Capstone Proposal

Domain Background

Breast cancer is the most common among women. Statistics appoint that about 1 in 8 U.S. women, representing about 12.4%, will develop invasive breast cancer over the course of her lifetime. Sadly, about 41760 women in the U.S are expected to die in 2019 from breast cancer [1].

Over the years Computer Aided Detection Systems were developed to help the analysis of mammograms. However, new technologies integrated in Machine Learning and Deep Learning fields are being more widely used to develop systems focused in the prediction of Mammographic Masses. This is happening because these applications can have a better performance in the identification of breast cancer, or the classification of Masses as benign or malignant cases.

Studies and applications are being developed during the course of the last decade. An example of research using a Supervised Learning approach is applied for mass detection in digital mammograms base on Support Vector Machines Algorithm [2]. This capstone project will follow a similar path for mammographic mass detection, but with more focus in the classification of masses data as benign or malignant based on a series of evaluation of Supervised Learning algorithms and Artificial Neural Networks performances.

Problem Statement

One of the most difficult and important job of a doctor can be informing a cancer diagnosis to a specific patient. There's an instant emotional, mental and physical reaction from the patient that represents the abrupt changes that will occur in his life. The most important aspect of a tumor can be its level of destruction, whether it is benign or malignant. This classification can inflect directly in the future health and medical difficulties the individual will face.

Therefore, the resulting application of this project can be used to automatically classify a Mammogram Mass with more efficiency in terms of time and accuracy, based on its clinical data.

Datasets and Inputs

The chosen public dataset is provided by the UCI Machine Learning Repository. This data contains 961 instances of masses detected in mammograms, and contains the following attributes [3]:

1. **BI-RADS**** assessment: 1 to 5 (ordinal)
2. **Age**: patient's age in years (integer)
3. **Shape**: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
4. **Margin**: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal)
5. **Density**: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)
6. **Severity**: benign=0 or malignant=1 (binominal)

Following are the files description that will be used in this project, and the respective URL where they can be found:

1. **mammographic_masses.data** - The dataset containing case values for all the attributes above.
2. **mammographic_masses.names** - Relevant informations about the dataset.

https://archive.ics.uci.edu/ml/machine-learning-databases/mammographic-masses/

****BI-RADS** is an acronym for Breast Imaging-Reporting and Data System

This dataset can be treated as a binary problem, where we only have two target outputs, 1 or 0 representing a malignant and benign case, respectively. We have a total of 961 input instances. The Class Distribution of our data and the other variables are described in the following table.

Mass classification	Output target	Class Distribution	Volume percentage	Total of instances
Malignant	1	445	46,3%	961
Benign	0	516	53,7%	

According to the table above, and comparing the volume percentage that each target represents in the total instance, we can assure that this dataset is very well balance in terms of distribution of its classes. So, one relevant measure that can confirm the effectiveness of our predictions is the accuracy of the model. Since the data is balanced, the results won't be overfitted to a specific class.

Solution Statement

.The best supervised algorithm that will have the best accuracy and F-score for this prediction and classification problem is, at first, unknown. With this in mind, i decided to pick the following algorithms to evaluate

- Random Forest
- Support Vector Machines (SVM)
- Decision Trees
- Logistic Regression
- Naive Bayes
- K-Nearest-Neighbors (KNN)
- Artificial Neural Networks

These algorithms and techniques can be developed using TensorFlow and Keras Libraries. However, since the dataset is very simple and contains a low dimensionality (few number of features), this project will focus on the implementation of scikit-learn algorithms for classification and regression. The data will be splitted in a training and test sets, and the model will be trained based on these inputs.

Benchmark Model

For comparisons, the following research will be used as benchmark model:

- **“Predicting the Severity of Breast Masses with Data Mining Methods ”** https://arxiv.org/ftp/arxiv/papers/1305/1305.7057.pdf

The results in this paper show us that a accuracy of about 81.25% where achieved in a test partition of the data when applying an SVM model. The dataset source for this research is the same from UCI Machine Learning Repository. My attempt with this project is to achieve an score of, at least, 75%.

Evaluation Metrics

The performance of our final model, and consequently the best supervised algorithm to fit the mammographic mass data classification problem, is evaluated through the following statistical measures:

1. **Accuracy**

   The accuracy of the model can be acquired by the following relation:

$$Acc = \frac{TP + TN}{TN + FN + TP + FP}$$

   The above variables represent:

   TP: True Positive
   TN: True Negative
   FN: False Negative
   FP: False Positive

   The accuracy value obtained represents the number of correctly classified cases in our model output.

2. **Precision**

   For the models precision, we have the equation:

$$Pcs = \frac{TP + TN}{TN + FN + TP + FP}$$

3. **Recall**

   Finally, the Recall of the model is given by:

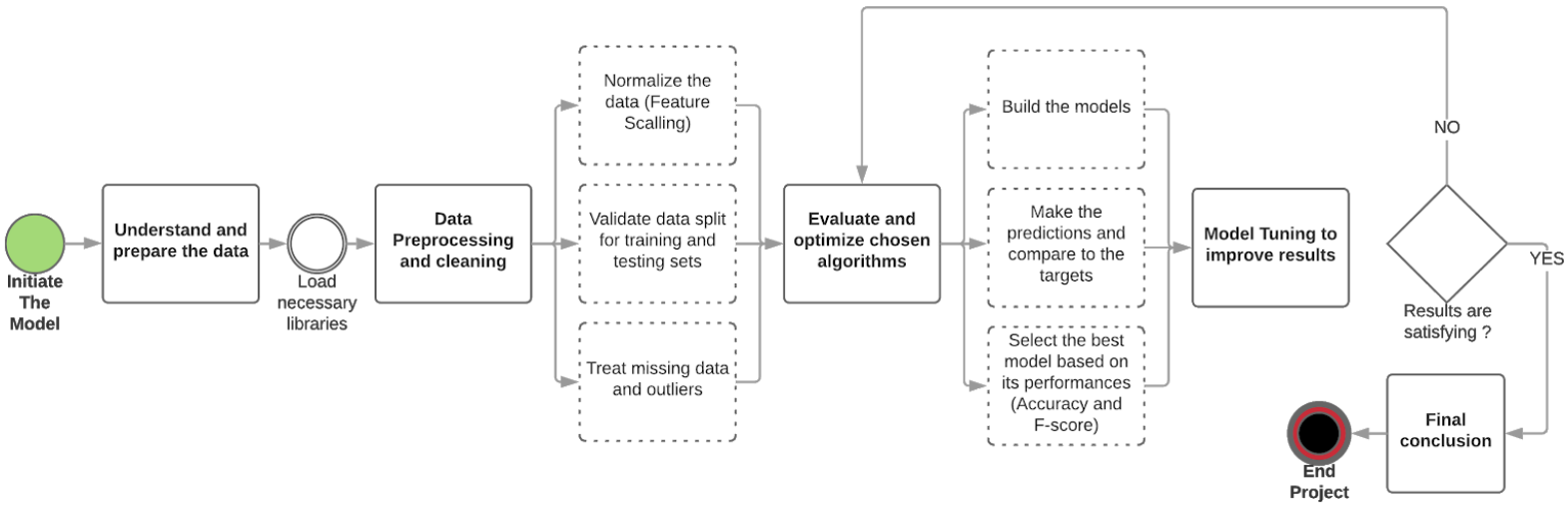
$$Rec = \frac{TP + TN}{TN + FN + TP + FP}$$

Project Design

To understand the dataset labels and values, an initial data exploration will be performed. This will be valuable to preprocess the data to achieve a better model predictions (e.g. Identify Features and targets and Training and Validation data split).

After this actions, the step of evaluating the algorithms chosen in the “Solution Statement” section will take place. This process consist in building the models, make the respective predictions and, finally, select the best model. For last, for a final conclusion, the chosen model hyperparameters and variables will be tuned to achieve a better result.

The Project Design workflow is structured in the fluxogram bellow:



References

[1] https://www.breastcancer.org/symptoms/understand_bc/statistics
[2] https://iopscience.iop.org/article/10.1088/0031-9155/49/6/007/meta
[3] https://helda.helsinki.fi/bitstream/handle/10138/20368/mammogra.pdf?sequence=1