Application for predicting breast cancer through *Google Prediction API*.

Andrio Rodrigo Corrêa da Silva[1]

[1]Universidade Federal do Ceará (UFC) - *Campus* Sobral

andrio.rodrigo.silva@hotmail.com

Abstract. *Breast cancer is a disease that has been affecting thousands of women around the world. Detection of this disease in the initial stage is very important, since a treatment is initiated thus increasing the survival rate. In this paper we discuss the elaboration of an application using a machine learning model trained using the Google Prediction API. This application is able to predict if the tumor of a given patient is classified as benign or malignant.*

1. Introduction

Breast cancer is the second disease that has been caused death among women around the world, there is also a very small number of men who are susceptible to this disease [Parkin 1998]. According to [M. Nounou 2015] approximately 1.7 million new cases were diagnosed and 521.900 deaths were recorded in 2012, representing approximately 30% of breast cancer cases.

Breast cancer begins when there is uncontrolled growth of the cells in the breasts. This disease can manifest in different parts of the breast, most of them begin in the ducts (it is responsible for carrying the milk to the nipple). Some symptoms should be taken into account at the time of diagnosis, such as: some mass in the breast, changes in the shape and size of the breast, differences in the skin color of the breast [Osareh and Shadgar 2010].

Currently there are several methods for the detection of breast cancer, such as: biopsy, mammography and ultrasound [Gayathri and Sumathi 2016]. When the tumor is detected it can be classified into two types, the first is benign, when there is no risk of death, the second is malignant when there is a risk of death [Gayathri and Sumathi 2016].

This paper aims to create an application that can assist health professionals in the prediction and complement the detection of breast cancer. This prediction will be accomplished by training a model using Google Prediction API.

2. Materials and Methods

2.1. Dataset

This paper is based on the dataset Breast Cancer Original available in the repositories of UCI Machine Learning [William 1992]. The dataset consists of 699 instances, which were collected from 1989 to 1991, where 458 (65.5 %) of these samples are of the benign type and 241 (34.5%) are of the malignant type.

The dataset provides ten attributes for these instances, such as:

- *Clump Thickness*.
- *Uniformity of Cell Size*.
- *Uniformity of Cell Shape*.
- *Marginal Adhesion*.
- *Single Epithelial Cell Size*.
- *Bare Nuclei*.
- *Bland Chromatin*.
- *Normal Nucleoli*.
- *Mitoses*.
- *Class*.

The first nine attributes are of the integer type and range between 1 and 10, however the class attribute, which is the variable to be predicted, can only be 2, for benign or 4 for malignant.

2.2. Google Prediction API

Google Prediction APi provides the ability to use machine learning. After learning from the data that was provided it is able to predict a numerical value, regression, or a category, classification [Google 2018].

After upload the dataset to the Google storage platform, it's time to training. In order to train the model Google Prediction API divides the data into 90/10, where 90% will be used for training and 10% will be used for testing. In this paper will be used regression to obtain the prediction.

Google Prediction API is built based on HTTP and JSON, which facilitates sending data to an external application. The API has integration with several programming languages, such as: Java, GO, Javascript, Python, PHP and Ruby. In this paper, the proposed application will connect to the Google Prediction API through a library that can be imported into the Android Studio, IDE used to create applications for Android. Through the library it will be possible to send requests to the Google Prediction API and after processing there will be a response that will be sent back informing which class those values belongs to.

3. Methodology

This paper uses a dataset with 699 samples collected over 3 years. From these data Google Prediction API was used, it provides machine learning algorithms for data analysis and consequently to predict the results. In order to use prediction service it is necessary that the dataset is in CSV format and the first column of it must be the attribute which will be predicted.

Google Prediction API works on an approach called black box which there is no control on the part of the user in choosing the machine learning algorithm or in separating the data for training and testing, there is only the choice whether the model will be regression or classification.

In order to train the model it is necessary use the insert method inside the API panel, where it will be necessary to fill the id, store location and model type fields as parameter.

After execution there will be a return of code 200 stating that the model was inserted successfully and that the training was started. After the model is inserted the get method was used to check the training progress. After the training was completed, the creation of the graphical user interface was started, a simple interface composed only by the fields necessary for the prediction. The communication with Google Prediction API is done through an HTTP request which send the 9 attributes to the machine learning model and a response is sent back which it presents the class that those attributes belong.

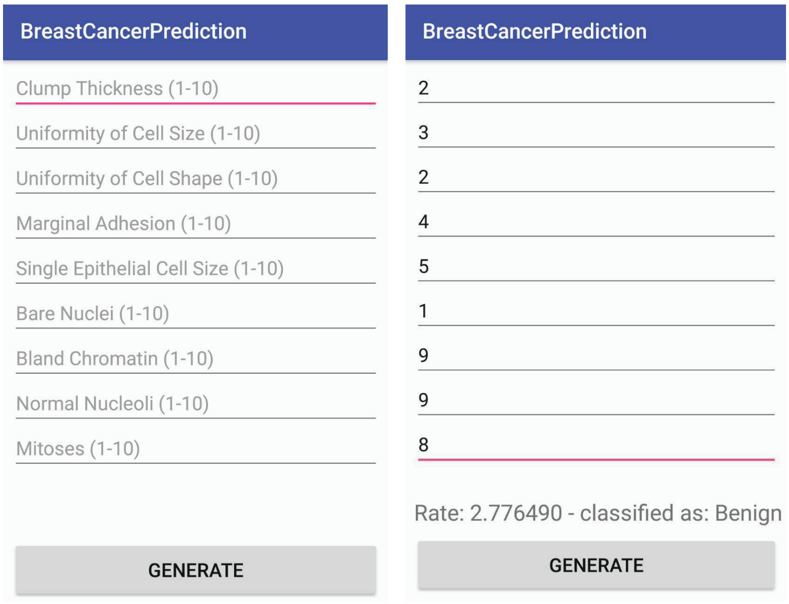
4. Results

For the training of the model using Google Prediction API, the value of the MSE (Mean Squared Error) was taken into account, which evaluates the prediction accuracy of the model, the smaller MSE means a more accurate prediction. The MSE can be calculated through the given equation:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Yi - f(xi))^2$$

In the equation Y is the response variable and f(x) is the prediction variable, in this way the regression model used can be evaluated, in the training the value of MSE was 0.16, a value which can be considered as good.

With the complete training the next step was to finish the application so that it could be connect with Google Prediction APi and be able to return the prediction.



BreastCancerPrediction

Clump Thickness (1-10)
Uniformity of Cell Size (1-10)
Uniformity of Cell Shape (1-10)
Marginal Adhesion (1-10)
Single Epithelial Cell Size (1-10)
Bare Nuclei (1-10)
Bland Chromatin (1-10)
Normal Nucleoli (1-10)
Mitoses (1-10)

BreastCancerPrediction

2
3
2
4
5
1
9
9
8

Rate: 2.776490 - classified as: Benign

(a) Home screen (b) Prediction

Figura 1. Application

6. Conclusion and Future Work

The use of machine learning through Google Prediction API has helped to create an application that can be used as a complement, not as substitute for other detection methods, for predicting breast cancer. The MSE value of 0.16 makes the result more reliable. This API presented a great performance of processing, it shows compatible with several programming languages and therefore with external applications, that helps to create application for different platforms.

For future work there is a need to improve the application graphic interface to become more useful to the final user. There is a need to create and use other machine learning algorithms as well for better prediction and reliability performance of the result.

Referências

Gayathri, B. M. and Sumathi, C. P. (2016). Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. Chennai, India.

Google (2018). Google prediction api: Developer's guide. available at: https://cloud.google.com/prediction/docs/developer-guide. accessed: 15th january 2018.

M. Nounou, F. ElAmrawy, N. A. K. A. S. G. H. S.-S.-Q. (2015). Breast cancer: Conventional diagnosis and treatment modalities and recent patents and technologies. In *Targeted Therapies in Breast Cancer Treatment*.

Osareh, A. and Shadgar, B. (2010). Machine learning techniques to diagnose breast cancer. In *2010 5th International Symposium on Health Informatics and Bioinformatics*. Antalya, Turkey.

Parkin, D. (1998). Epidemiology of cancer: global patterns and trends. In *Toxicology Letters*.

William, H. (1992). Breast cancer wisconsin (original) data set. available at: http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29. accessed: 15th january 2018.