Breast Cancer Tumor Classification

Guillermo Carsolio González A01700041
ITESM CQ, Intelligent Systems Gpo 1

Introduction:

We as humans in order to keep on living we constantly replace our cells by a process called mitosis, in which when a cell grows it splits into 2 cells in order to reproduce. This process happens millions of times every second around our body that is composed of trillions of cells. But sometimes this process is disrupted when corrupted or damaged cells split when they are not supposed too and create tumors. Depending on the nature of this tumors they can be benign (they do not spread all over your body) or malignant (they spread invasively in the rest of the individual's body). So, it is of great importance to have methods to be able to distinguish if a tumor is benign or malignant.

So, with the new advances in technology and the development of AI, models have been trained in order to give prognosis of patience that could have cancer. AI is a powerful tool that can give an edge to the health industry to improve the efficacy of their diagnostics.

Breast Cancer

One of the most common cancers that women develop is breast cancer coming second to skin cancer. Breast cancer happens when malignant tumors form in the breasts of a women. But it has different areas where it can develop. The most common area is the duct that delivers the flow of breast milk to the nipples, and it is named ductal cancer. Second to that would be in the glands called lobules that make the breast milk called lobular cancer. [1] Figure 1 shows the anatomy of the breasts, and we can appreciate the areas of origin were the cancer can develop

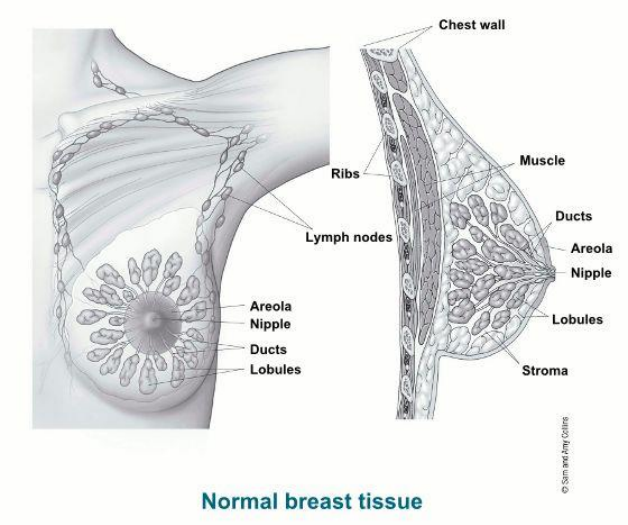

Figure 1

The way the cancer spreads is that the tumors extend to the lymph nodes in order to give the cancer cells access to the rest of your body.

Dataset

the data retrieved is of the distinct measurements of the tumors of patients retrieved by the University of Wisconsin Hospitals from 1989 to 1991. It is composed of 699 instances of patients with their tumor's measurements and diagnosis (benign or malignant). The data has a class distribution of 358 benign tumors (65.5%) and 241 malignant tumors (34.5%). It was necessary to remove 16 instances because they had some unknown values in either one or multiple attributes.

The instances had a total of 9 attributes that embodied different measurements of the tumors in the patients. it is necessary to give a brief understanding on these measurements and in order to get a better grasp of the findings. Some of the measurements that are self-explanatory are clump thickness (tumor size), uniformity of cells size, uniformity of cell shape, bare nuclei

(how exposed are the tumor cell's nucleuses), normal nucleoli (how abnormal the nucleolus [the housing of the nucleus] is) and Mitosis (how abnormal the splitting of cells is). Others need to be explained further and so we will start with marginal adhesion which describes how much cells stick to each other. Because normally cancer cells tend to not be as sticky as normal cells. Another measurement is single epithelial cell sized which just loos at how abnormal your surface cells are on size. Then you have bland chromatin, chromatin is the source material of chromosomes that make the DNA living in the nucleus of the cells, and in this case, it means how uniform this is.

All these attributes have a value in range of 1-10 1 being the most normal and 10 the most abnormal. One can conclude that if all the attributes are in an abnormal range, we can assume that the tumor is going to be malign, and in some sense that could be correct. But what matters most is what is the relation of these attributes and which of them have the most impact.

Models

To be able to analyze the data in this project I concluded that giving the user an option of using multiple models in order to get to a conclusion was necessary to give not only more input in the analysis itself but also the models.

Logistic Regression

This model is a machine learning technique that originated from the field of statistics. But in AI it is mostly used in binary classification problems. The method uses at its core the logistic function that can be seen in Figure 2.

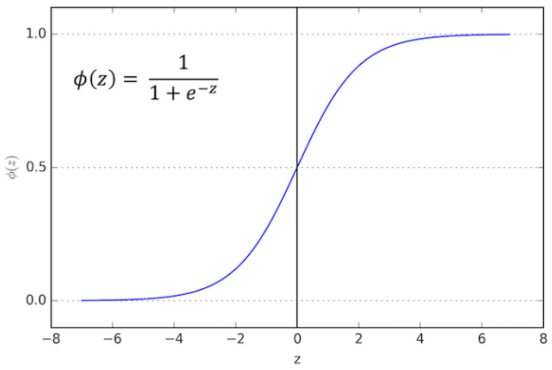

Figure 2

So, as we can see the S shape like curve can take a value between 0-1. Now the way it is used in machine learning is that the x values or attributes are combined with coefficient values to predict the y value or class. In our problem this would be if the cancer tumor is benign or malignant.

Decision Trees

Another model that we will be using in this project will be DTs which are a supervised learning technique that can be used for both classification and regression. We use the data from the dataset to "grow" the tree using the attributes and certain values to choose the nodes and leaves of the tree. In figure 3 it is shown one tree that was grown with Sklearn and the dataset

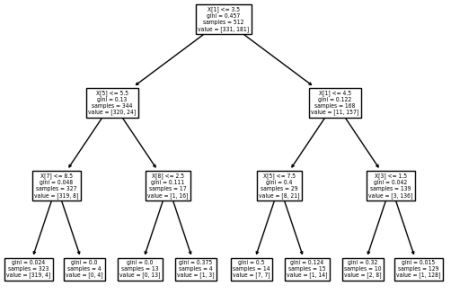

Figure 3

Random Forest

Finally in the last model just like in Decision Trees this is also a supervised learning technique that can be used to for either classification or regression problems. But this is not a single tree but mor so a “forest” or group of trees that are build using a method called bagging. Bagging is Bootstrap Aggregation that builds the trees using random sub-samples of data specifically the attribute or features. Then it uses these forests of trees to predict the outcome of either the classification issue or regression task. In Figure 4 we can visualize how the process works in order to build the random forest.

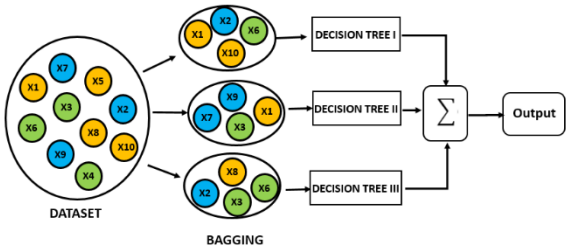

Figure 4

Program

The code that is allocated int the same repository as this documentation is designed for a user to interact with in order to predict an input that the user desires. The script splits into 3 different parts; models, dataset and user interface. First there are 3 functions that build and train the selected models (Logistic Regression, Decision Tree, Random Forest). Then in the main function I have a portion that cleans the dataset and labels it with the appropriate attribute and class names. Finally, I have multiple functions that work in conjunction that make up the interaction with the user in order to create predictions. All the code is documented with the specifics of parameters per function and outputs.

Findings

As I was developing the project, I realized that the prediction itself become intuitive because of the nature of the attributes. Because they can be classified in a range of normal (1) to totally abnormal (10) one could assume that the larger the value in all attributes the more likelihood of it being a malignant tumor. And that line of thought is not particularly wrong if you put all the attributes with a value of 4 and above most of the models (every time the models are trained it changes their accuracy because it selects a random part of the dataset to split) will give as a prediction a malignant tumor. What I found most interesting is what attributes are the ones that seem to have the most impact and that if there is a small range of abnormality it will always predict the tumors as being. It is necessary for multiple attributes to be at a larger range of 4 for the tumor to be malignant. The attributes that were spiking more impactful in the tests that I have made were both uniformity of cell shape and size and marginal adhesion. But as my tests progressed, I realized that singular attributes aren’t particularly impactful in the outcome but more so the combination of multiple.

Conclusion

This project found that there can be tools like AI in order to be applied in the health industry to try and have another perspective in what attributes impact more the conditions on any disease. I still believe that this tool is still far from what we can achieve with technology in order to improve diagnostics, but it is a start to analyze the information we have at our disposal and look at the issue at hand with a widen outlook.

References:

Admin. (2020, October 22). Nucleolus - Function, Difference Between Nucleus & Nucleolus. BYJUS. https://byjus.com/biology/nucleolus/.

Brownlee, J. (2021, January 4). Bagging and Random Forest for Imbalanced Classification. Machine Learning Mastery. https://machinelearningmastery.com/bagging-and-random-forest-for-imbalanced-classification/.

Brownlee, J. (2020, August 14). Logistic Regression for Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/logistic-regression-for-machine-learning/.

Chromatin. Genome.gov. (n.d.). https://www.genome.gov/genetics-glossary/Chromatin.

Farlex. (n.d.). Bare Nucleus. The Free Dictionary. https://medical-dictionary.thefreedictionary.com/Bare+Nucleus#:~:text=Naked%20Nucleus,typically%20seen%20in%20cell%20degeneration.

Gupta, P. (2017, November 12). Decision Trees in Machine Learning. Medium. https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052.

UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. (n.d.). https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

What Is Breast Cancer?: Breast Cancer Definition. American Cancer Society. (n.d.). https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html.

What Is Cancer? National Cancer Institute. (n.d.). https://www.cancer.gov/about-cancer/understanding/what-is-cancer.