Classification of tumors into 2 categories (benign or malignant) based on features of tumor cell nuclei

Author: Rehan Ali

Capstone project 1 for SpringBoard Data Science Career Track

Mentor: Ryan Herr

Table of Contents

[Motivation for this study 3](#_Toc511848968)

[Problem and the need for this analysis 3](#_Toc511848969)

[The Client 3](#_Toc511848970)

[Data used in this study 4](#_Toc511848971)

[Methodology 5](#_Toc511848972)

[Description of ‘CapstoneProject1\_EDA.ipynb’ notebook 5](#_Toc511848973)

[Description of ‘CapstoneProject1\_EDA.ipynb’ notebook 7](#_Toc511848974)

[Results and model 8](#_Toc511848975)

[Results 8](#_Toc511848976)

[Model 15](#_Toc511848977)

# **Motivation for this study**

## Problem and the need for this analysis

A tumor can be malignant (cancerous) or benign (non-cancerous). This classification is made by a pathologist who examines the tumor cells obtained from a biopsy under a microscope. One method employed by pathologists is to examine the shape of a cell nucleus and make a judgement based on it.

The problem that is encountered in this regard, is that the shape is defined by many parameters such as the area of a nuclei, it’s perimeter and numerous other numerical parameters. Making a judgement on changes in which of these parameters are more important relative to changes in other parameters requires a quantitative analysis. For example, one simple question to ask is if there is an increase in area of a cell nuclei but it is still circular in shape than what decision should a pathologist make?

## The Client

The analysis described in this study is targeted towards cancer pathologists and oncologists. A pathologist’s decision on tumor classification determines the type of treatment that an oncologist will undertake. The described analysis will aid a pathologist in deciding the type of tumor. Therefore, it will also increase the confidence of an oncologist who relies on the report of a pathologist.

## Data used in this study

The data for this study comes from UCI machine learning repository and is created by Dr. William H. Wolberg of University of Wisconsin Hospitals. It is obtained from the following Kaggle dataset link:

<https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data>

The data is a comma separated file with rows labelled with patient identification number and the first column, labeled diagnosis, contains a series of ‘M’ (malignant) or ‘B’(benign) strings. The rest of the columns contain values for 30 different parameters describing the shape of the cell nucleus.

# **Methodology**

## Description of ‘CapstoneProject1\_EDA.ipynb’ notebook

The methodology used in this study is described here. It is suggested to read this along with the ‘CapstoneProject1\_EDA.ipynb’ as the methodology describes this notebook file in detail. The following points describe the data analysis in detail:

1. Firstly, the required libraries were imported. These include pandas, matplotlib.pyplot, numpy, defaultdict, and scipy.stats.
2. The data file named ‘Cancer.csv’ was read using pandas ‘read\_csv’ method and the index column was set to the patient identification number. Next, the head of the data was explored.
3. ‘pandas.info’ method was called on the data frame to check for missing values and the data types of different columns. No missing values were found and the data types for all the columns were what they should have been except the diagnosis column which was of data type object.
4. The meta data for the csv file suggested that only the first 9 columns of the data frame contained useful values whereas, the rest of the columns were statistical values computed form the first 9 columns. Therefore, the data frame was sliced such that only the first 10 columns (the first column contained the diagnosis values) were kept.
5. To ensure that no patient data was duplicated, the maximum of counts of the index column was calculated and assert was used to ensure that this was equal to 1.
6. As mentioned in point 3, the data of the diagnosis column was of type object, it was changed to categorical variable to save memory.
7. The data frame was melted and only the diagnosis column was kept. This was done to make visualization of the data frame with seaborn easier.
8. Box plots overlaid with strip plots were created for the melted data frame using seaborn. The hue column used was the diagnosis column.
9. The plot showed that the area\_mean column values were much greater than the values of all the other columns due to which changes in those columns couldn’t be visualized.
10. To solve this problem a new data frame was created by excluding the area\_mean parameter and the new data frame was plotted as before.
11. This plot showed that the texture\_mean, radius\_mean and perimeter\_mean column values were much greater than the rest of the column values.
12. To remedy this, three data frames were created: 1- for area\_mean, 2- for texture\_mean, perimeter\_mean and radius\_mean, and 3- for the remaining columns.
13. These three data frames were plotted as before on three different axes. The plots showed that changes in all the columns could now be visualized.
14. scipy.stats library was used to test if the columns followed a normal distribution and then a significance change test was used to check if changes in different columns were significant or not.
15. Lastly, percentage change in median values of malignant nuclei relative to benign nuclei were measured for every column and plotted to check which of the parameters showed the greatest changes.

## Description of ‘CapstoneProject1\_EDA.ipynb’ notebook

As the file ‘CapstoneProject1\_FinalCode.ipynb’, was used as a first attempt to explore the data a different notebook, ‘CapstoneProject1\_FinalCode.ipynb’, was created which excluded many lines of code. This notebook has the following major changes:

* Lines such as df.head(), df.info(), etc., have been removed.
* The three data frames were directly made without first visualizing the data.
* As creating boxplots with overlaid strip plots for each data frame was a repetitive process, a function called ‘box\_strip\_plot’ was created which takes a list of data frames as an argument and returns the required figure which can then be plotted.

# **Results and model**

## Results

A total of 569 patient’s data was analyzed with 212 malignant cases and 357 benign cases. The box plots and strip plots shown in figure 1 show that every parameter describing the nucleus increased in malignant cases compared to the benign cases. Certain parameters depend on other parameters, therefore, a change in the dependent parameters is to be expected. Area\_mean, perimeter\_mean and compactness\_mean all depend on the radius\_mean parameter. Figure 1b shows that the radius\_mean of malignant nuclei is greater than the radius\_mean of benign nuclei. As area and perimeter of a circle increase with an increase in the radius of the circle the increase in area\_mean (figure 1a) and perimeter\_mean (figure 1b) of malignant nuclei compared to benign nuclei was expected. Compactness can increase or decrease depending on the value of the radius. The results show that compactness\_mean also increases for malignant nuclei compared to benign nuclei.

A screenshot of a social media post

Description generated with very high confidence

A screenshot of a social media post

Description generated with very high confidence

A close up of a map

Description generated with high confidence

Figure : Box plots with over laid strip plots for different parameters of nuclei for malignant and benign cases.

Overall, the results shown in figure 1 show that the malignant nuclei increase in size compared to benign nuclei. Moreover, the increase in all the parameters is significant as determined by Mann-Whitney U test shown in table 1.



*Table 2: Table summarizing Mann-Whitney test values for all the parameters.*

Apart from the increase in the size of the malignant nuclei, the number of concave points (concave\_points\_mean) and the concavity of these points (concavity\_mean) also increases significantly for malignant nuclei (figure 1c, table1). Therefore, the local change in radius of a nucleus (smoothness\_mean) also increased, as expected, in malignant nuclei which is shown in figure 1c.

Lastly, the mean\_texture of malignant\_nuclei also increases compared to benign\_nuclei. This implies that the constituents of the membrane around the nuclei change in malignant nuclei. The specific change cannot be determined from this data. Also, there is also a change in symmetry\_mean parameter, but the meta data for this data set does not explain what this parameter means, therefore, no comments can be made on the change in symmetry\_mean.

To get more insights from the data the percentage increase in the median of all the parameters for malignant nuclei relative to benign nuclei was calculated and is shown in figure 2.

A screenshot of a cell phone

Description generated with very high confidence

Figure 2: Percentage increase in the median of all the parameters of malignant nuclei compared to benign nuclei.

Two important points need to be noted from figure 2, which are the following:

1. The radius of malignant nuclei almost doubles when compared to the benign nuclei and the perimeter and the area change as expected from the change in the radius.
2. The Parameters that show the largest changes are the number of concave points and their concavity. This implies that the greatest difference between the two nuclei is the distortion of the nuclei shape; a benign nucleus is relatively circular whereas, the malignant nucleus loses this circularity and its shape becomes distorted.

## Model

Based on the results obtained from the analysis, the following model is presented which describes the changes in the nucleus of a malignant tumor compared to a benign tumor.

**Malignant nucleus**

**Benign nucleus**

*Larger in size.*

*distortion in shape because of increase in the number of concave points and their concavity*

*Smaller in size.*

*No distortion in shape*