Ian Guintu, Sheng Lu, Matthew Rivera, Anthony Rosario

CIS 3120 ETRA [34795]

Group E: Yellowstone

12/12/2021

Final Project Report

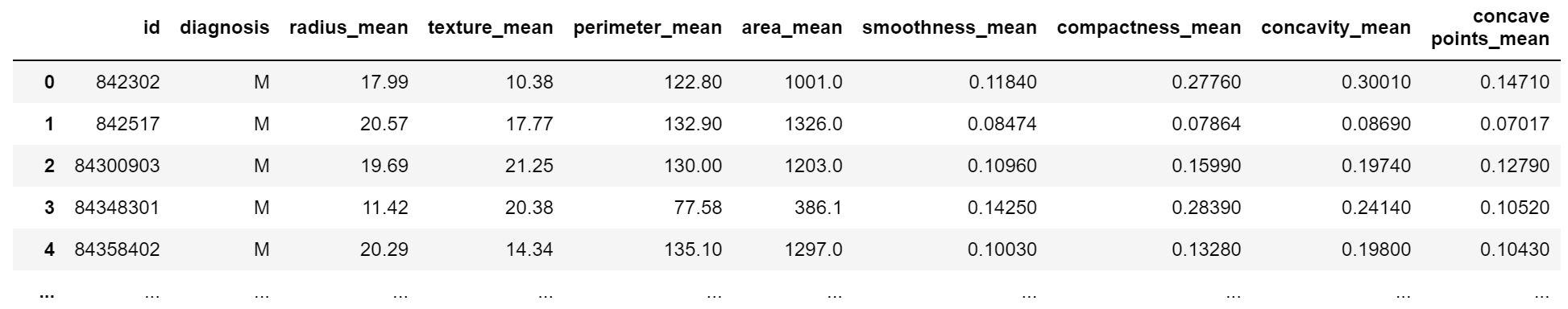
Introduction

The topic area our group chose focused solely on the healthcare industry and, more specifically, on breast cancer diagnosis. Our main objective was to analyze healthcare data to determine both the drivers of breast cancer and what specific variables define it as being either malignant or benign.

We wanted to take a closer look into what features make a cell malignant or benign as well as what factors can attribute a direct correlation to one determination over the other. We thought this would be important due to the fact that breast cancer is a serious issue that has or can have a significant effect on someone’s life. This can be considered a common occurrence and it is good to analyze breast cancer tumor data to both understand what the dominant features are that determine whether it is malignant or benign what is happening in a situation like this.

Description of Dataset

Group E thoroughly examined breast cancer data to determine the drivers of breast cancer. This dataset holds 32 columns of information on 568 different drivers of breast cancer. Initially, we viewed the first 5 rows of our dataset using the **.head()** function to glance through the data within the 32 different columns:

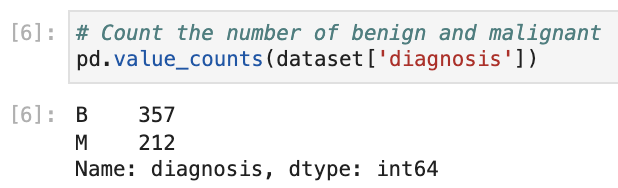


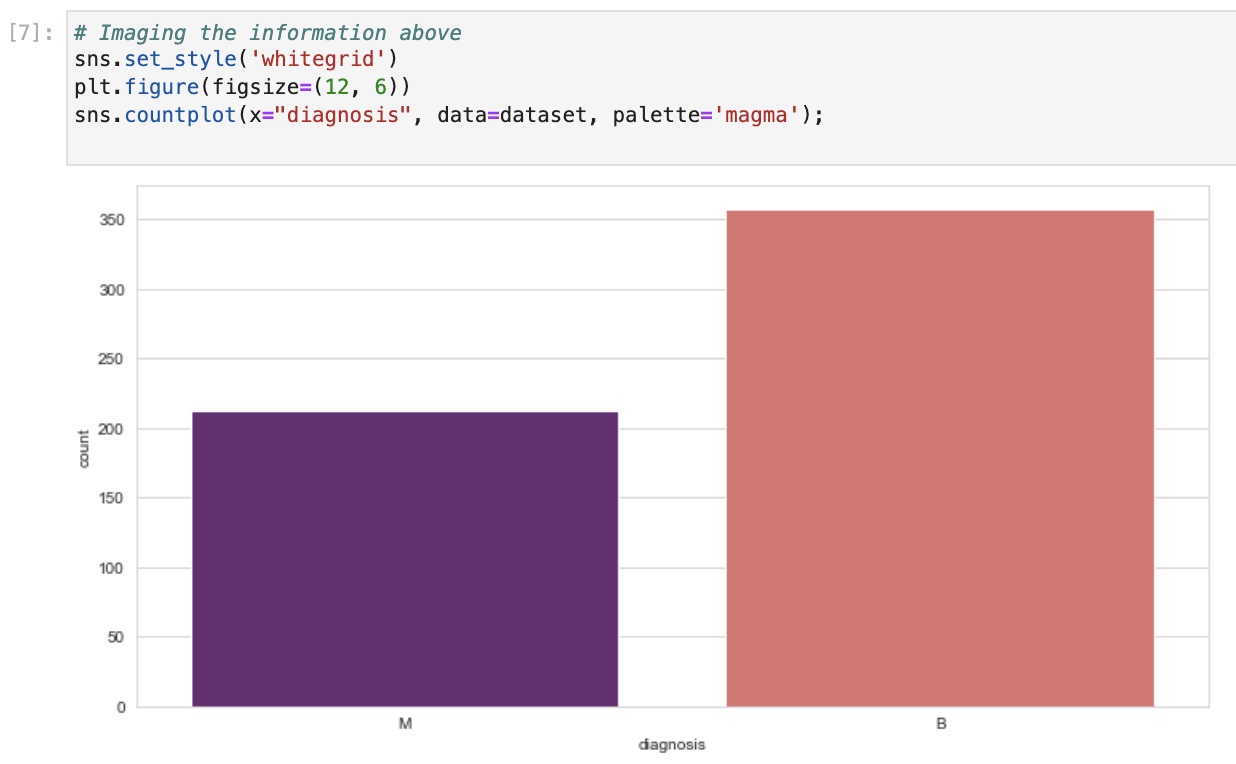
From the **.head()** function, we were able to define these variables: **radius** – the distance from core to points, **concavity** – the depth of concave sections, **concave points** – the number of concave sections within a core, **mean** – the average of the specific features within a core, and the **worst** – the mean of the 3 largest values in a specific core. Additionally, we narrowed down the 32 variables within this dataset to an even 10. From this, we can break them down even further into just 5 variables if we set aside their “\_mean” and “\_worst” attributes. These variables are radius, perimeter, area, concavity, and concave points.

We were able to use the **.info()** function in order to better gather information about the data frame. This function contains the total number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values). We were able to find no null values, 30 float values, 1 object value, and 1 integer value. Furthermore, we dropped the unnamed column (column 32) since it does not provide any usefulness in determining the drivers of breast cancer (see **Exhibit 1**).

Data Analysis

Before continuing with our analysis of the data, we defined the difference between malignant and benign. If a tumor is considered to be **malignant**, it is cancerous and can potentially recur even after its removal. A **benign** tumor is the opposite of malignant where it is not as harmful, and, once removed, it most likely will not recur. To ensure that there are only two results from diagnosis: Malignant (M) or Benign (B), we counted the total number of Malignant (212) and Benign (357) cases using the **.value\_counts** function and created a bar graph to display the total count of malignant and benign cases clearly:





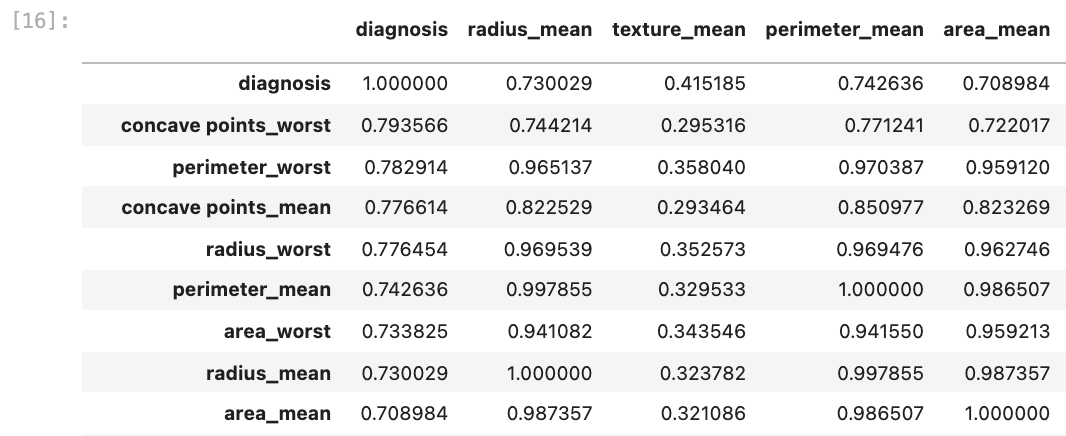
We moved to fractal analysis of images of breast tissue specimens in order to provide a numeric description of tumor growth patterns. We used fractal analysis scatterplots to show the relationships between two variables and visualize where the growth would be in relation to each other. From these scatterplots, we can get a description of the tumors and their growth patterns. (see **Exhibit 2**).

Analyzing our pairplot creation, we can see this on a much larger scale. The immediate difference here being we separated them by their “\_mean” and “\_worst” attributes. We can see that the plots in the “\_worst” attribute are more tightly grouped together than those in the “\_mean” attribute. This may be due to having a more focused average since the “\_worst” attribute utilizes the 3 highest values and takes their average:



We were able to ask the question: Does the average size of the tumor determine whether it is *malignant* or *benign*? Perhaps the tumor size gives an increased probability that the tumor is cancerous or non-cancerous. Referring to “\_mean”, this is simply the average of the specific features within a core. When we apply this attribute to a variable, we’re taking the average of the variable in mind. Similarly, this occurs with “\_worst” which, in this case, is the average of the 3 largest values in a specific core. We can use these numbers effectively because a malignant cell will often show more concavity as well as more concave points, so it is useful to take note of the averages and the highest recorded average. In order to compare the core average size between malignant and benign, we utilized the **.groupby()** function to split the diagnosis and the radius\_mean. We found the average malignant size to be 17.4628 and the benign size to be 12.1465 (see **Exhibit 3**). From the **.groupby()** function, we moved to creating bar graphs utilizing the **.pivot\_table()** function showing a comparison of benign and malignant tumor size (see **Exhibit 4**).

Using the **.corr()** function, we are able to see the specific features that affect the diagnosis of the tumor. The most important of these factors turned out to be the core average size, concave points, and concavity, which, after reviewing the data, makes a lot of sense. We can see that the larger and more rigid the tumor is, the more likely it will be malignant:



This data holds 32 columns and 568 rows of information regarding the drivers of breast cancer. This dataset was reduced to only look at those with a strong link to determine whether a cell is malignant or benign. We ran a correlation table to measure those top factors and decided that these were going to be our main variables. There were no null values but there were 30 float values as well as 1 object value and 1 integer value. Our float values were the variables themselves. The one object value was the diagnosis of “malignant” or “benign”. The integer value was the ID number that was given for each diagnosis. Does the average size of the tumor determine whether it is malignant or benign? We concluded that it ultimately does. It is essential within the group of our top ten variables. While it isn’t wise to just look at this variable alone to determine the final outcome, it does put a spotlight on the fact that the bigger the size of the cell, the more likely it is to become malignant. Logistic Regression determined that using our 10 variables, we could accurately determine the outcome for a tumor with 95% accuracy. This is a good indication that we have pinpointed the main features found within a malignant growth. We deployed a heatmap in order to see this correlation effectively as well as made a mask to convert our data to binary so the logistic regression could work properly.

^Not 100% complete… Will fix.

General Takeaways

The main takeaways we arrived at from this project have been: learning what variables lead to a malignant tumor, what those exact variables are that contribute to this diagnosis, and with what accuracy can we predict a growth’s outcome using these variables. We found that there were some key variables that significantly assist in determining the outcome of a growth. Precisely, variables like the perimeter, concave points, and concavity were the star features here. Following analysis of the data, we discovered that the growths that were determined as malignant shared a larger perimeter size as well as more concave bumps and ridges. These specific features provided a significant outlook into what makes a growth malignant or benign. The logistic regression heatmap was able to predict how likely a tumor was malignant or benign by using these top ten features, which it did so with a 95% accuracy. We concluded that this was as precise as we could have been regarding the number of variables included without lowering the accuracy.

One lesson that I, Anthony, learned from this is that Python is more extensive than I originally thought. I took a beginner Python course about a year ago and we covered very basic principles. I programmed some small, simple machines like converters and calculators. Python grew on me and easily became my favorite to use. I then had to work with some that I didn’t find as easy to understand and manipulate, specifically R. I learned R in a data analytics course and did not find myself getting accustomed to the program. I found it very interesting that a lot of the concepts there I could bring right back into Python. For example, using a dataset and creating a training and testing dataset respectively, so that we can find the logistic regression. I learned that Python is a lot more extensive and thorough than I thought it to be at first.

One lesson that I, Matthew, learned from CIS 3120 is some of the specific tools used for data analysis. I have been familiar with Python, SQL, and R languages utilized for data analysis. I took CIS 2300, a Python course, covering some simple tools used for python. In addition, I took CIS 3400, a database managment class, using SQL to build databases. I took CIS 3920 in order to learn more about R to create statistical analytics. Python is a coding language capable of doing all the functions of the other languages. With CIS 3120, I learned to aggregate, to analyze data utilizing the .groupby and .pivot\_table functions, and joining different datasets to create deeper insights. I learned how to use statistical analysis tools like a logistic regression in our group project. Python utilizes a lot of functions used in other languages, but having the benefit of it brings in a single language. I found it very interesting that a lot of the concepts there I could bring right back into Python that I learned at other classes. Python grew on me and easily became a favorite programming language to use in my data analytics career.