Name: Hassan Fakih Osman

ID: 121200078

CMPE 407 Machine Learning Project

**Breast Cancer Prediction**

**Introduction:**

I was thinking of having a good prediction model for any sort of serious medical issue, so for a start my aim is to be able to efficiently predict whether a certain patient is facing breast cancer. To do that I will use the Wisconsin Breast Cancer Dataset from the UCI repository that I found while looking at a certain study conducted by Madu Kumari and Vijendra Singh from the NorthCap university (2018).

Being able to predict a disease at its early stage is quite important as it could be of help in avoiding cancer and stopping it from a possible spread across tissues.

Of course, to achieve this, we will depend on classification with its different algorithms, to classify the patient as either “Malignant” or “Benign”. However, before we go, bear in mind that malignancy is the “Cancer” however benign isn’t.

**Dataset:**

Again, our dataset is called “Breast Cancer Wisconsin (Original) Dataset” from the UCI machine learning repository, it is made by Dr. William H. Wolberg from University of Wisconsin Hospitals (Check first link in reference page). It is made up of 699 records and 11 attributes:

1. Id number (however, it won’t be used in our algorithms)
2. Clump Thickness: 1-10
3. Uniformity of Cell Size: 1-10
4. Uniformity of Cell Shape: 1-10
5. Marginal Adhesion: 1-10
6. Single Epithelial Cell Size: 1-10
7. Bare Nuclei: 1-10
8. Bland Chromatin: 1-10
9. Normal Nucleoli: 1-10
10. Mitoses: 1-10
11. Class (2 for benign, 4 for malignant)

In addition to this, this dataset has 16 null values all of which lie in the ‘Bare Nuclei’ column (details about this in the preprocessing section). The preceding attributes shall help us in determining the class which the patient belongs to (Malignant or Benign).

Most of the attributes seen here seem to have an increasing linear relationship with the class, let us see some plots showing this…

(Data Visualization next page)

**A picture containing screenshot, text, pattern, square

Description automatically generated**

You can see from the preceding heatmap (in the last row or last column), that all features have positive correlation with class - Uniformity of cell size for instance has 0.82, Bare Nuclei also has 0.82 and so on… - so this means that as the attributes increase from 1 to 10, so does the probability that the patient could be malignant (i.e., from 2 to 4).

Now let us see how uniformity of cells (Sizes and Shape) would affect the patients:

A picture containing text, plot, line, screenshot

Description automatically generatedA picture containing text, plot, line, screenshot

Description automatically generated

Why does its increase lead to malignancy? Cell sizes or shapes can differ greatly, however if it is within the same tissue it tends to be similar (Ginzberg et al., 2015), But their loss in proportionately within the same tissue - that is if within the same tissue one cell is different in size or shape than the others - the cells tend to malfunction, and when cells malfunction they undergo a process call “Phagocytosis” in other words “Cell suicide”; however when the phagocytosis process fail… the cell may become a cancer cell.

So, in general this explains why such increase as seen in the graph would lead to malignancy.

! Please note that phagocytosis in detail is out of the scope of this report, but rather I spoke generally of it.

Let us look at how ‘Mitoses’ affects Malignancy:

A picture containing screenshot, text, colorfulness, rectangle

Description automatically generated

When a person is facing malignancy, they tend to have a very high mitotic activity; for example, look at bar 1, a benign tumor tends to have mitosis normally, however, when a patient is facing malignancy mitoses can become 10 times faster. So, you can see the dangers of it all when the tumor becomes malignant, and it shows us the importance of having to find a tumor when it’s still benign to avoid such consequence. However, bear in mind that mitoses doesn’t have high correlation as other attributes.

Let us now see the effects of Bland Chromatin on malignancy:

A picture containing text, plot, screenshot, line

Description automatically generated

Bland Chromatin illustrates a texture of the nucleus in benign cell, where in malignant cells they tend to be harsh or “coarser”, so it seems that from the graph as it increases, becoming more coarser, chances are that the tumor is malignant.

Another, is the ‘Normal Nucleoli’:

A picture containing screenshot, text, colorfulness, rectangle

Description automatically generated

When the tumor is malignant, nucleoli becomes very prominent and increases in numbers; but in normal cells they tend to be very small, almost not visible at all. So, we could check here for the number of nucleoli, if large then it’s highly probably that the tumor is malignant.

A picture containing plot, line, text, diagram

Description automatically generated

In here, Bare nuclei is a term which means nuclei not surrounded by cytoplasm which is something very much shown a lot in malignant tumors, which could also be seen in the graph above.

A picture containing diagram, plot, line, text

Description automatically generated

Normal cells usually are stuck together but during malignancy it becomes more loose, so this is what marginal adhesion implies, that the more adhesion loss there is the more probable the tumor is malignant.

A picture containing text, screenshot, colorfulness, rectangle

Description automatically generated

In short words, Epithelial Cells tends to be large in malignant tumors.

A picture containing text, plot, line, screenshot

Description automatically generated

Cancerous cells are often grouped in multilayers, while benign tends to be ‘Mono-layered’. In general, from the plot the more layers the more the thickness which implies malignancy.

**Preprocessing:**

1. **Dropping Null Values:**

Before feeding the data to our models, some preprocessing needs to be done. I first had to deal with the null values present, currently there are 16 null values which are present in the ‘Bare Nuclei’ column (You may see that in the following figure). Dropping the null values is a good option, however I decided to put the mode value of the Bare Nuclei as a replacement to these null values.

A screenshot of a computer

Description automatically generated

Why use mode and not some other techniques like median or mean? Well in general 16 null values I believe won’t be as critical, however since we have a very small discrete number of possibilities (1 to 10), mode seemed the most suitable for this situation.

1. **Dropping the ID column:**

Of course, since it brings us no relevant information to predict the label it is dropped.

1. **Feature Selection:**