**Breast Cancer tumor prediction**

**Description:**

Breast cancer is the frequent type of cancer in women worldwide. It is responsible for 25% of all cancer cases and afflicted approximately 2.1 million people in 2015. The proliferated cells typically become tumors, which can be seen on an X-ray or felt as lumps in the breast.The main obstacle to detection is determining whether tumors are malignant (cancerous) or benign (non-cancerous). I have used the Wisconsin breast cancer dataset for my analysis. Features describe the characteristics of the cell nuclei.My goal is to find whether the tumors are malignant or benign.

**The outcome of EDA:**

As part of my analysis, I have chosen the five variables, namely, radius\_mean, texture\_mean, perimeter\_mean, area\_mean, and smoothness\_mean. Those five variables are analyzed under two different scenarios malignant and benign. The outcome of the analysis is:

* The data set has no missing values
* The data frame does not have duplicates
* There are a total of **569 patient records**. 357 have benign, which means they are not cancer affected, and**212 are cancer,** i.e., **62.7%**  are not affected by cancer, and **37.3%**  are affected by cancer.
* Most of the tumors are benign and have low mean\_radius.
* Most of the tumors lie in size range of (12,15).
* Radius\_mean and texture mean for the malignant tumor are higher than for benign tumor.
* compactness\_mean and smoothness\_mean both are higher for Benign tumors.
* The median for fractal\_dimension\_mean ,smoothness\_se,texture\_se symmetry\_se are very close to each other.These features do not create too much impact on the target variable.
* The interquartile range for the malignant tumor is higher, and its median level is also on the higher side than a benign tumor.
* As the radius\_mean of the tumor increases, the chance of turning benign cancer cells into malignant cells is increasing.
* Most people are cumulated near 12, which is the mean of our normal distribution. Malignant tumors are distributed more on the right side, and benign tumors are distributed on the left side of the graph.
* When the radius increases, the area also increases. Therefore, they are positively correlated, and area\_mean is caused by radius\_mean.
* When radius increases, perimeter also increases. Therefore, they are positively correlated, and perimeter\_mean is caused by radius\_mean.
* Spearman's correlation is a little higher than Pearson's correlation.
* The logistic regression model accuracy is 97%.

**What do you feel was missed during the analysis?**

In the initial analysis, I have not entirely understood the dataset. I had gone through the dataset multiple times and understood the attributes' actual meaning. What are the actual attributes to be considered for the analysis? That part was missing in the starting. But later, I researched and came up with a good analysis.

**Were there any variables you felt could have helped in the analysis?**

I felt compactness\_mean, compactness\_se, and radius\_worst would also help analyze the tumors.

**Were there any assumptions made you felt were incorrect?**

I assumed texture\_mean would be high for the benign tumor. But after doing an analysis and plotting the graphs, I realized that it is high for malignant tumors.

**What challenges did you face, and what did you not fully understand?**

I felt challenged in framing the statistical questions for the datasets and which variables to choose for the analysis. After that, I spent a lot of time writing the model of the dataset and PMF plots. I got many errors. After a lot of effort, finally, I was able to complete them.

**References**

Learning, U. C. I. M. (2016, September 25). *Breast cancer Wisconsin (Diagnostic) data set*. Kaggle. Retrieved June 4, 2022, from <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>

kanncaa1. (2018, September 9). *Statistical learning tutorial for beginners*. Kaggle. Retrieved June 4, 2022, from <https://www.kaggle.com/code/kanncaa1/statistical-learning-tutorial-for-beginners/notebook>