**Machine Learning in Healthcare: Breast Cancer Prediction**

**Sahi, Ustatt Singh (MS Electrical Engineering)**

### Introduction

Machine Learning is a field of study that has emerged and being used in today's technical world. It is a booming technology of this age and has a lot of implementation being done in the field of healthcare as well. A lot of emerging new companies are implementing this technology into different parts of the healthcare and technical world. Startups are investing a lot of time and money to provide the best and precise healthcare solution to this struggling community with chronic diseases and disorders. A lot of companies are focusing on how it could also help in research and diagnosis of patients who are struggling with cancer. Companies are using machine learning to develop models that can detect malignant and benign tissues in mammograms.

Using ML, the wait time is reduced extensively. Different features used are patient information, medical history, discharge sheet, lab reports, charts, allergies and discomfort and chart even the layout of a hospital for waitlists. With the help of predictive model, we can manage patient admissions and discharge. This way we can keep the medical cost low and affordable for patients by decreasing the price, giving more accuracy, better management and much more.

### Methodology

**Model: Testing & Training**

Various institutes and companies develop their own model using different types of data. It is possible that there might exist more than one model and both could be equally good to work and predicting the expected results. The more the data we feed for training our model the better the prediction would be. For our prediction to be precise it is very important we do data preprocessing and feature extraction in the data available. We can use multiple models to predict different risk factors that a patient might have, and we can train our model accordingly.

**Methodology**

In healthcare domain when we implement AI, we aim to achieve best precision to treat the patients well. Basically, we achieve sensitivity over specificity. Thus, our algorithm should have minimum errors and well labelled. In this project we focus our healthcare platform to detect and predict breast cancer patients. It is a very challenging task as a minor error could vary and give us unexpected results, but it will assist physicians for better and precise analysis. Our dataset has 569 rows or data points and 33 columns or features/attributes. We have one null feature and we have to drop it. So, we are left with 32 features in total before we proceed towards working our model.

Using our matplotlib library we can have a boxplot view of our dataset. We observe two classifiers of the output which include malignant and benign tissues in the breast. Matplotlib help us to see the dataset how our output is varied in the given dataset.

The first and the foremost step is to upload the dataset. We use our pandas library to visualize the dataset. We get to see all our attributes, features and data points distributed. We store our dataset in a data frame. Data can be in any format, what's important is to represent and clean and preprocess the data well before we train it.

**Prediction Model: Naïve Bayes Classifier**

In this supervised machine learning technique predictors are independent of each other. So, it assumes that all features in the class are unrelated and different from each other. This model is used basically when we have large datasets. This model is easy to build and simple in form. It performs way better than any other classification technique, based on the type of dataset. It also performs good where prediction has to be done among multiple class.

Confusion matrix tables define how a classification model has performed on a test data. It shows how the model gets confused while making predictions can be a FN: false negative and FP: false positive. We have zero error predictions called TP: true positive and TN: true negative. Which means we predicted what was expected.

With confusion matrix we can calculate the accuracy, recall, precision and F-measure of the model based on the test data. Classification rate or accuracy

### Analysis and Results

Based on performance matrix and confusion matrix we can evaluate which model performed the best. We trained and tested models like Random classifier, decision tree classifier, Naive Bayes classifier, support vector classifier, K nearest neighbors and logistic regression. Based on these performance matrices we figured out Naive Bayes highly productive and precise.

Confusion matrix tables define how a classification model has performed on a test data. It shows how the model gets confused while making predictions can be a FN: false negative and FP: false positive. We have zero error predictions called TP: true positive and TN: true negative. Which means we predicted what was expected.

We also know it by other names such as simple Bayes or independence Bayes. Another advantage of this model is that it can perform very well for classification even if the size of the dataset is very small.

**Summary/Conclusions**

Improvements could be made if we add more values to our dataset that trigger the cancer cells. We have explored all risk factors to our implementation by choosing the right model. In our results we have also shown comparison of different model accuracies. Our Naive Bayes model has outperformed all other models in terms of accuracy and performance. We still have a lot of improvements and future work to be implemented given the right resources which is abundance of data.

### Key References


### Acknowledgements

It took 3 months of hard work to do some research and reading a lot of relevant textbooks and articles to finish up this master's project. It was a good journey overall. My seniors were also an important source of information for this project. Although I worked on this project without any partner, but I had strong support of my advisors Prof. Juizi Zhao and my co advisor Prof. Birsen Sirkeci.