- **DSC 630 Project Presentation**

- **Aditya Sumbaraju**
  DSC 630 Predictive Analytics
  Professor Fadi Alsaleem
  Oct 30, 2021

Table of Contents

Topic

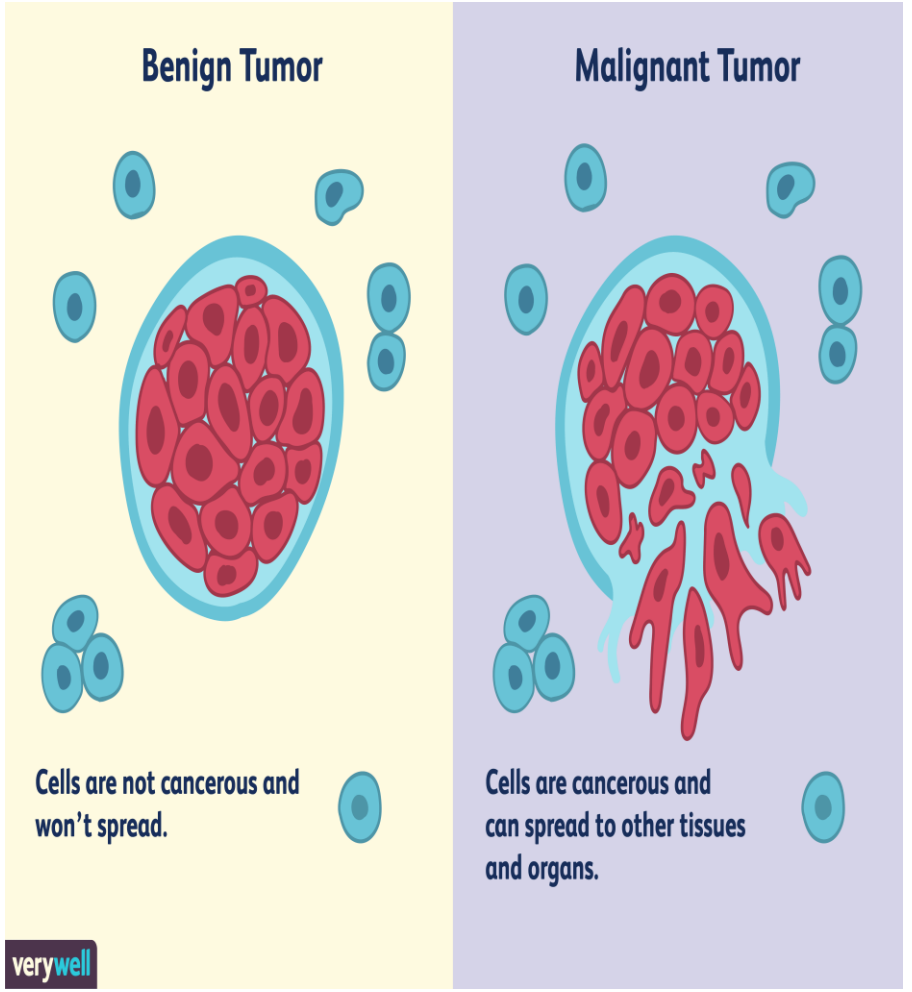
BREAST CANCER PREDICTION

Problem Understanding

- Breast cancer, the most common cancer among women worldwide accounting for 25 percent of all cancer cases and affected 2.1 million people in 2015. Early diagnosis significantly increases the chances of survival.
- The key challenge in cancer detection is how to classify tumours into malignant or benign. Machine learning techniques can dramatically improve the accuracy of diagnosis. Research indicates that most experienced physicians can diagnose cancer with 79 percent accuracy while 91 percent correct diagnosis is achieved using machine learning techniques

Tumour Classification

- In this case study our task is to classify tumours into malignant or benign tumours using features obtained from several cell images
- The first step in the cancer diagnosis process is to do what we call it fine needle aspirate or FNA process which is simply extracting some of the cells out of the tumour. And at that stage we do not know if that tumour is malignant or benign

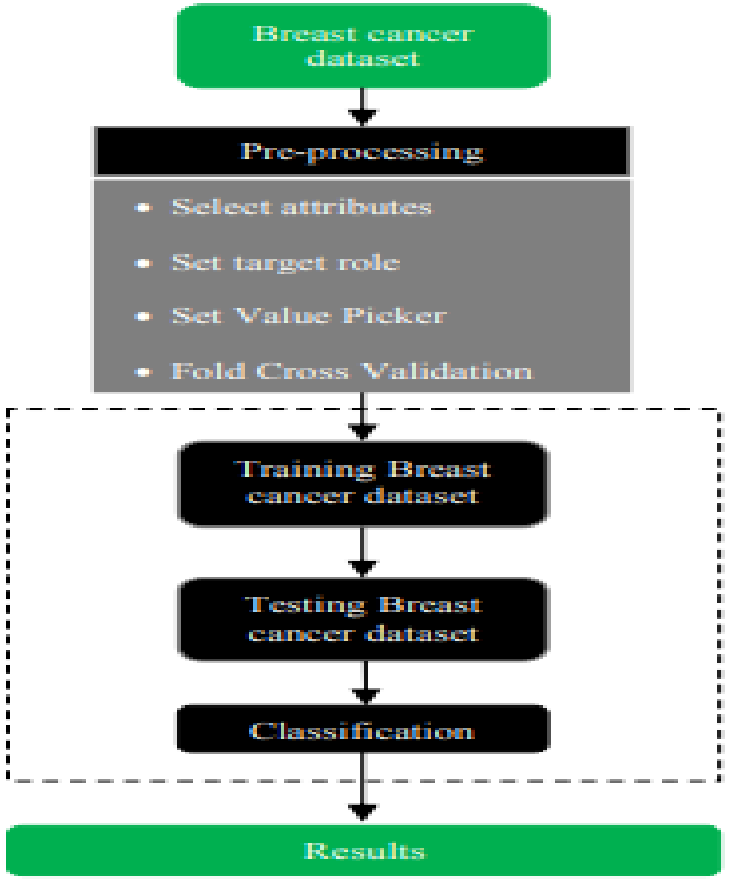


Background and History

- Machine learning may now be included in the armory of a predictive analyst, which may help them perform better at their jobs. Machine learning includes predictive analytics as one of its components. Machine learning algorithms may be capable of reliably detecting the development of breast cancer soon (Bakr et al.,2020).
- When medical practitioners have a more accurate prediction of their patients' breast cancer survival, they can make more informed treatment planning decisions, avoid unnecessary therapy and thus lower economic costs, more effectively enroll and exclude patients from randomized trials, and develop palliative care and hospice.

Machine Learning Approach

- As a part of this case study; I have considered the dataset that contains the features computed from digitized image of fine needle aspirate of a breast mass.
- When we say features that mean some characteristics out of the image such as radius for example of the cells such as texture, perimeter, area, smoothness and so on.
- We will be applying data wrangling and data transformations to feed all these features into our machine learning model. And the model will learn the data and predicts the occurrences of cancer.
- The dataset: Publicly available (created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA (Wolberg and Mangasarian 1990). It consists of
  - ✓ 699 instances (Benign: 458 Malignant: 241) a
  - ✓ 2 classes (65.5% malignant and 34.5% benign)
  - ✓ 11 integer-valued attributes.

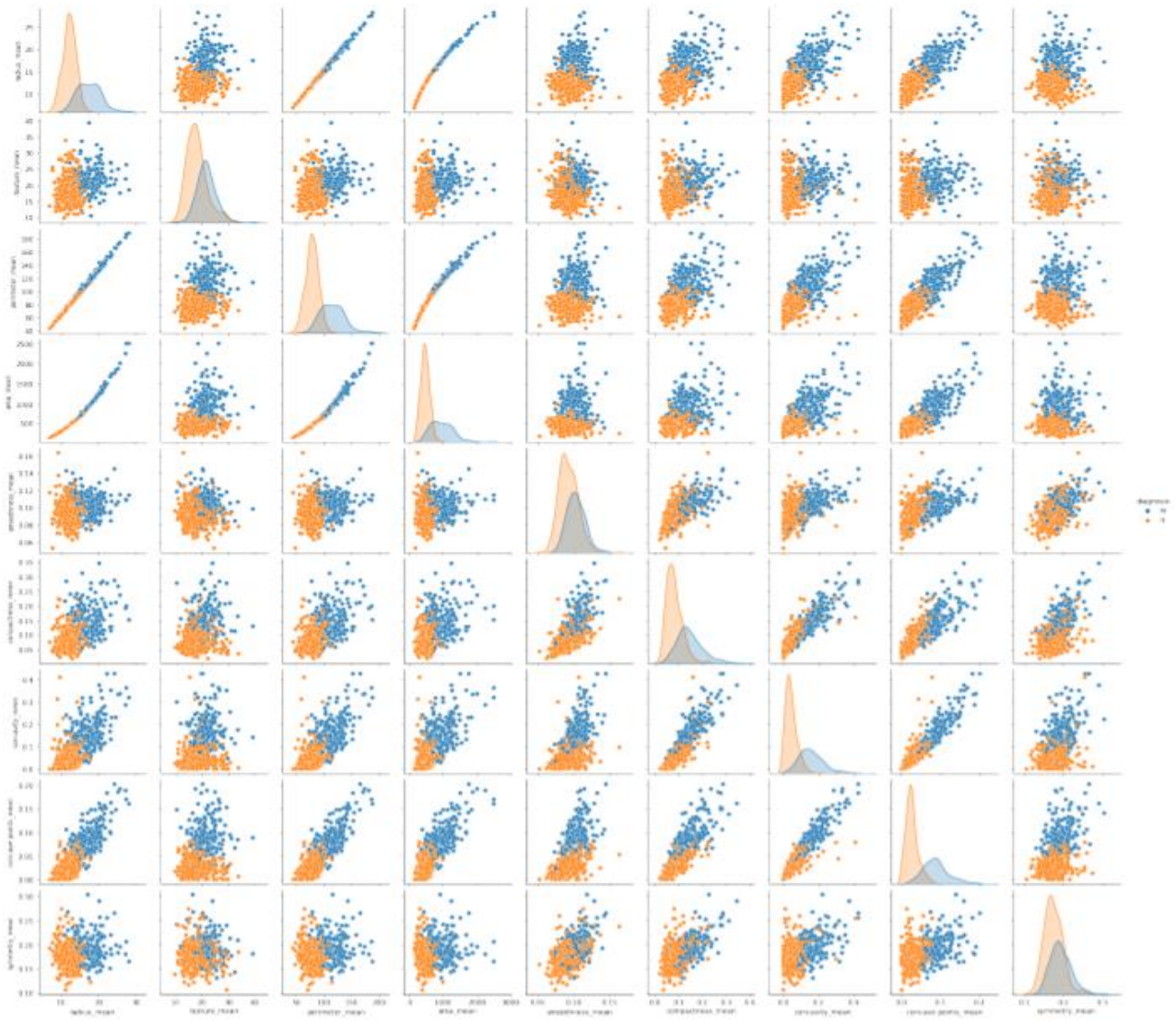


Preliminary Analysis

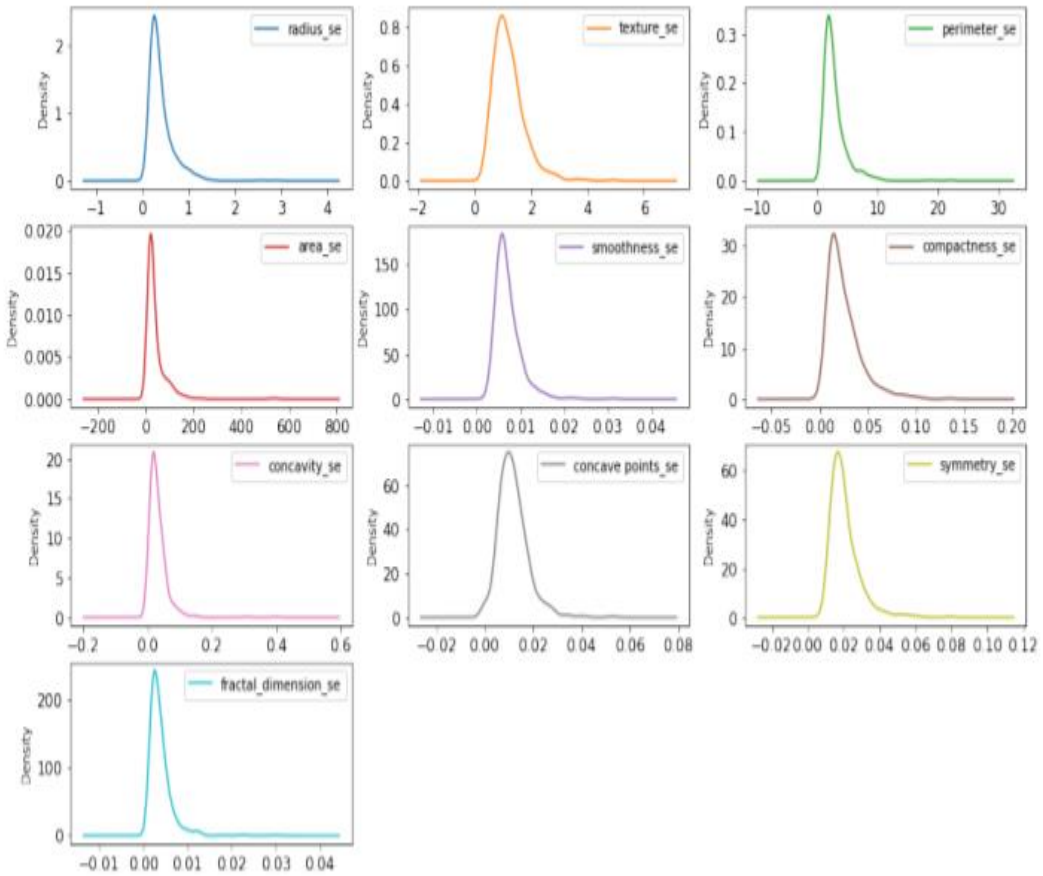
- The figure at right represents **all features** from our data and its relationship.
- Categorized ‘0’ for Benign tumour and ’1’ is for malignant tumour.

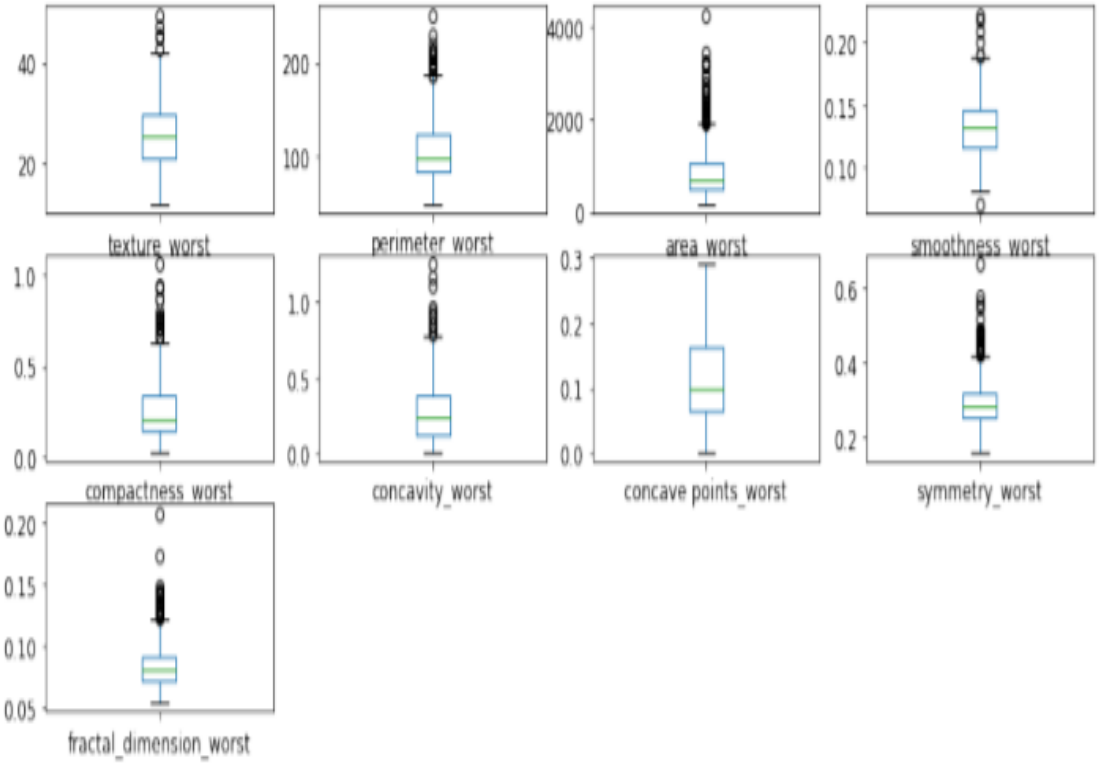
Observations:

- One of the many observations to see is that mean radius and mean perimeter have a linear relationship.
- Whenever the mean area increases w.r.t mean radius, the chances of a benign cell increases. This resembles that the patient can be safe.
- When the mean smoothness is at the middle, the mean area decides if the cell is benign or malignant. The higher the area, more are the chances of a benign tumour.



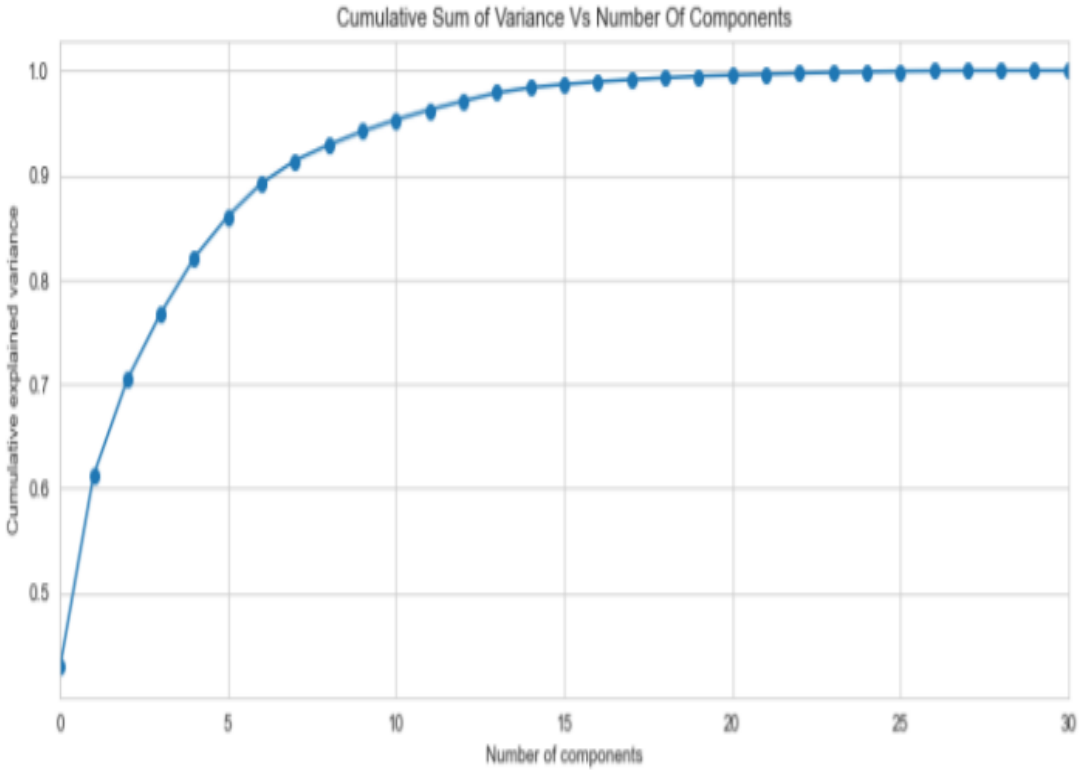
Contd.. Feature Visualizations



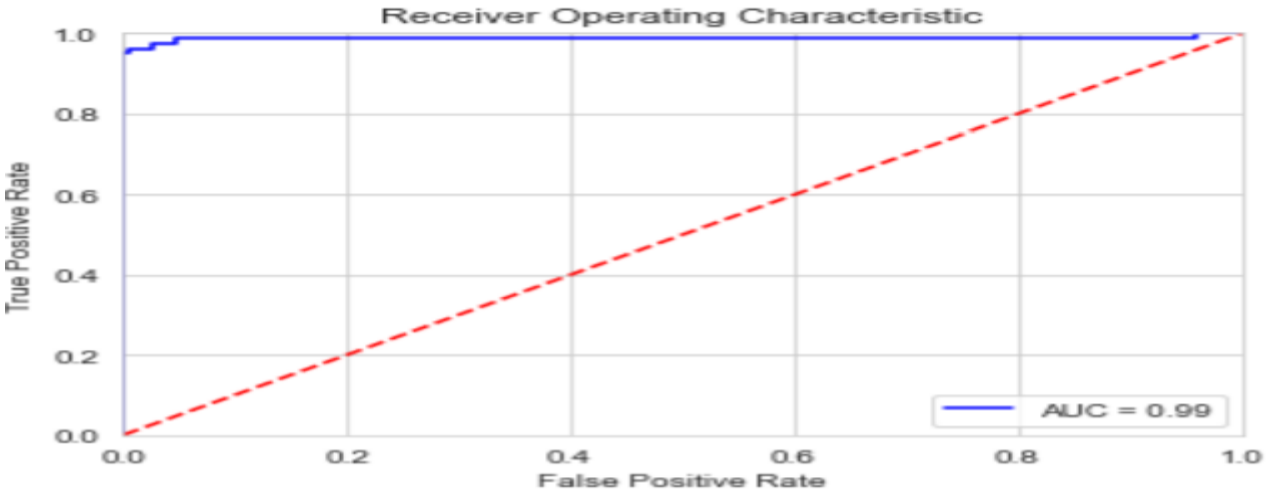


Contd.. Principal Component Analysis(PCA)

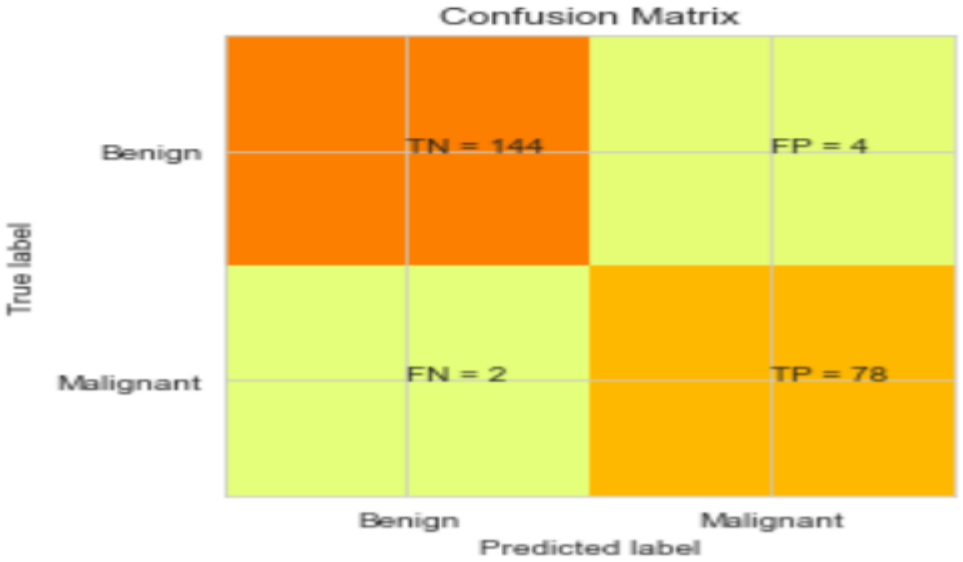
- In this project PCA was done on small dataset. The analysis helps to get better understanding of the data and dependencies between variables.
- The predictive model were built on original data but it is rather the fact of not high dimensional dataset. It is always good to check, whether such analysis can improve the mode.



Model Evaluation: SVM

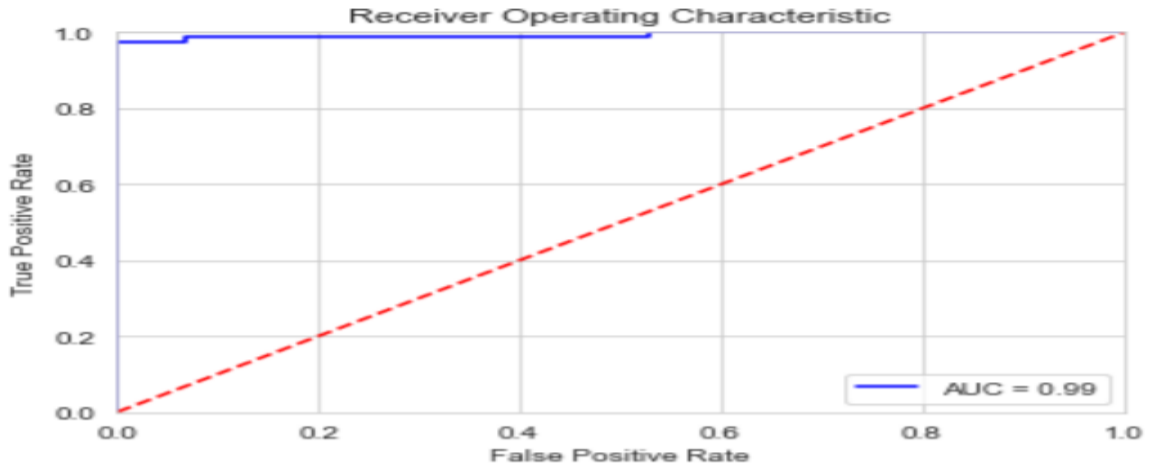


Classification Report:				
	precision	recall	f1-score	support
0	0.99	0.97	0.98	148
1	0.95	0.97	0.96	80
accuracy			0.97	228
macro avg	0.97	0.97	0.97	228
weighted avg	0.97	0.97	0.97	228

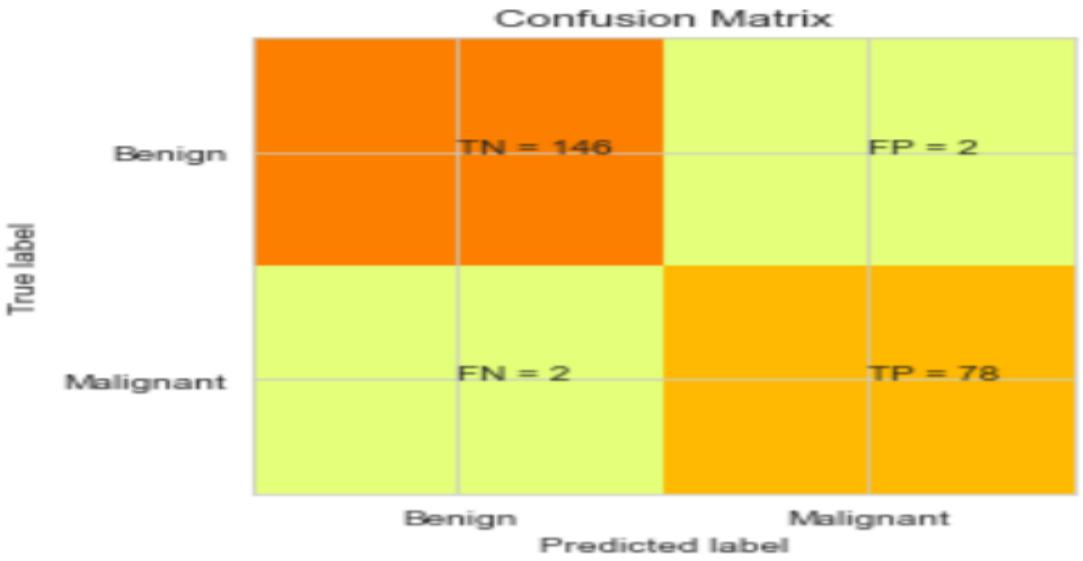


There are two possible predicted classes: "1" and "0".
Malignant = 1 (indicates presence of cancer cells) and Benign = 0 (indicates absence).
The classifier made a total of 228 predictions (i.e 228 patients were being tested for the presence breast cancer).
Out of those 228 cases, the classifier predicted "yes" 82 times, and "no" 146 times.
In reality, 80 patients in the sample have the disease, and 148 patients do not.

Model Evaluation: Logistic Regression

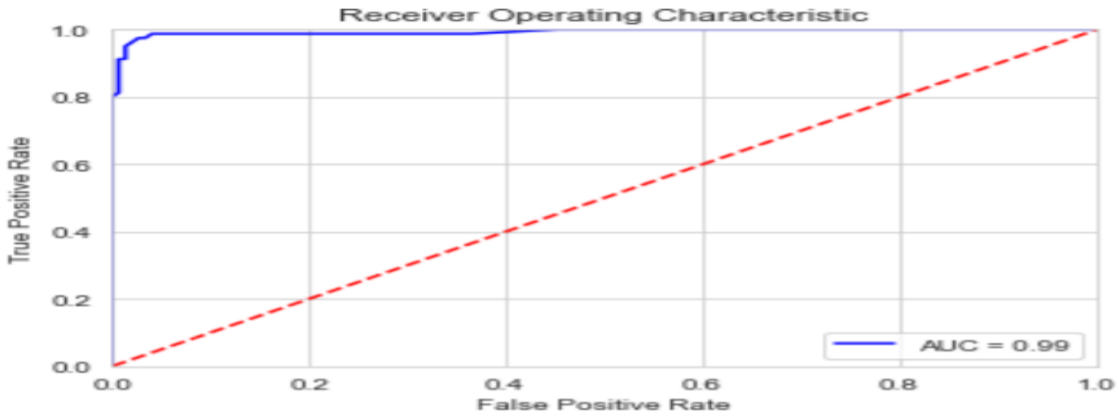


Classification Report:				
	precision	recall	f1-score	support
0	0.99	0.99	0.99	148
1	0.97	0.97	0.97	80
accuracy			0.98	228
macro avg	0.98	0.98	0.98	228
weighted avg	0.98	0.98	0.98	228

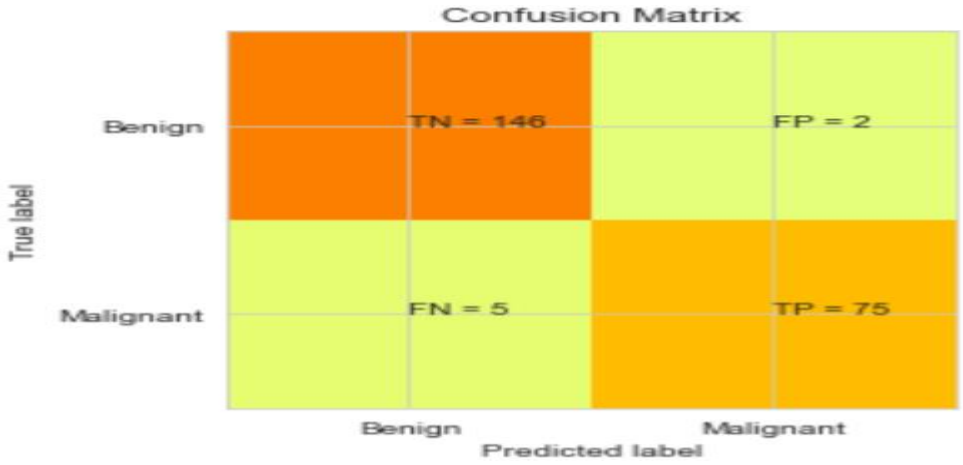


There are two possible predicted classes: "1" and "0".
Malignant = 1 (indicates presence of cancer cells) and Benign = 0 (indicates absence).
The classifier made a total of 228 predictions (i.e 228 patients were being tested for the presence breast cancer).
Out of those 228 cases, the classifier predicted "yes" 80 times, and "no" 148 times.
In reality, 80 patients in the sample have the disease, and 148 patients do not.

Model Evaluation: Random Forest

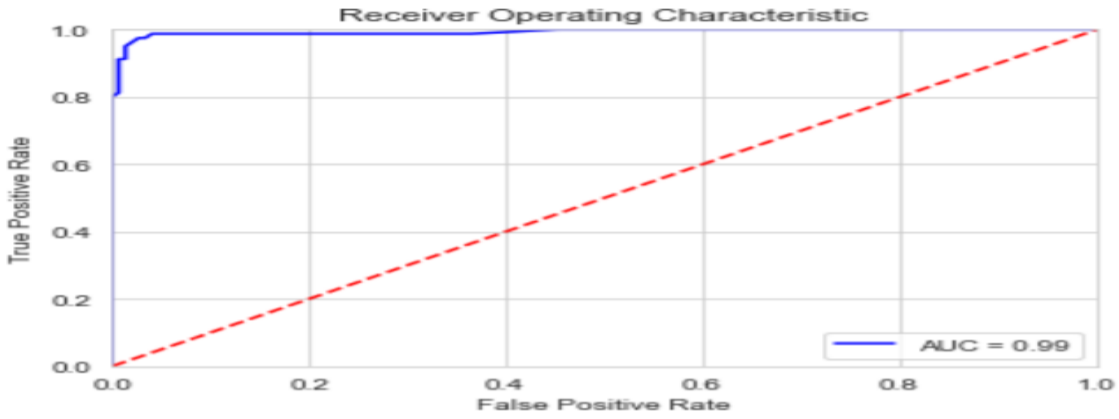


Classification Report:				
	precision	recall	f1-score	support
0	0.97	0.99	0.98	148
1	0.97	0.94	0.96	80
accuracy			0.97	228
macro avg	0.97	0.96	0.97	228
weighted avg	0.97	0.97	0.97	228

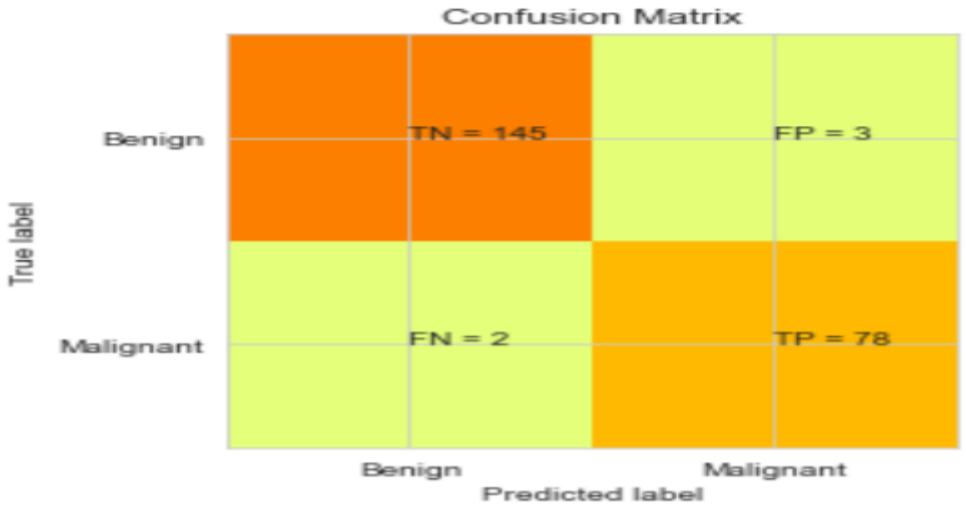


There are two possible predicted classes: "1" and "0". Malignant = 1 (indicates presence of cancer cells) and Benign = 0 (indicates absence).
The classifier made a total of 228 predictions (i.e 228 patients were being tested for the presence breast cancer).
Out of those 228 cases, the classifier predicted "yes" 77 times, and "no" 151 times.
In reality, 80 patients in the sample have the disease, and 148 patients do not.

Model Evaluation: KNN

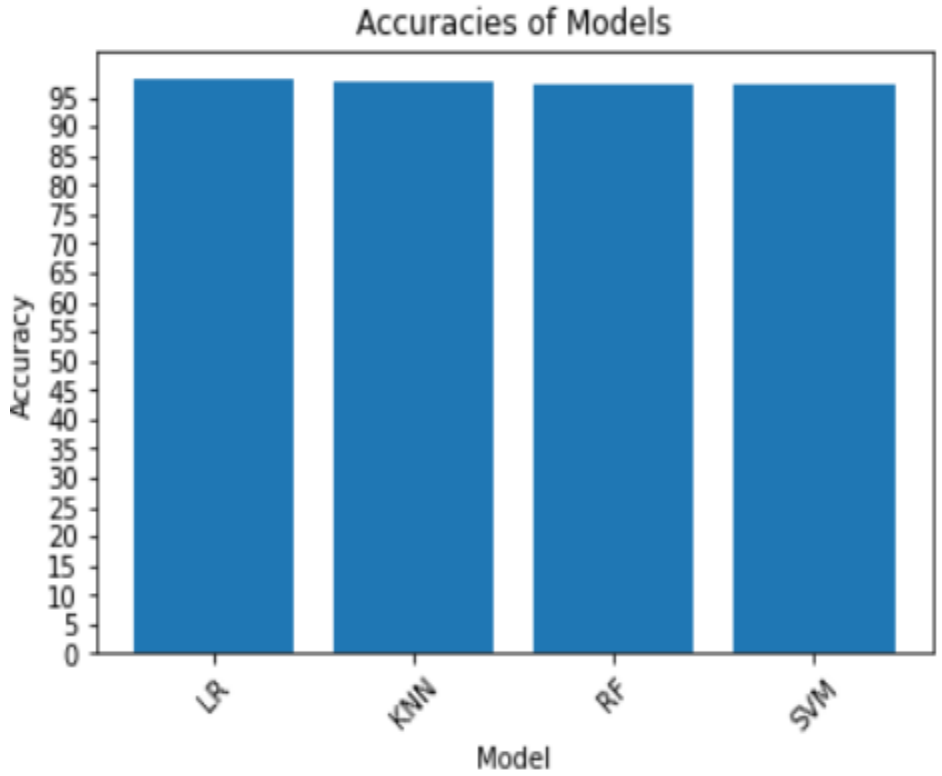


Classification Report:				
	precision	recall	f1-score	support
0	0.99	0.98	0.98	148
1	0.96	0.97	0.97	80
accuracy			0.98	228
macro avg	0.97	0.98	0.98	228
weighted avg	0.98	0.98	0.98	228



There are two possible predicted classes: "1" and "0". Malignant = 1 (indicates presence of cancer cells) and Benign = 0 (indicates absence).
The classifier made a total of 228 predictions (i.e 228 patients were being tested for the presence breast cancer).
Out of those 228 cases, the classifier predicted "yes" 81 times, and "no" 147 times.
** 80 patients in the sample have the disease, and 148 patients do not.

Model Comparison



Model	SVM	Logistic Regression	Random Forest	KNN
Prediction Yes	82	80	77	81
Prediction -No	146	148	151	147
Accuracy	97%	98%	97%	98%

In comparison between logistic regression, Random Forest, KNN and SVM, the logistic regression model was more accurate in predicting breast cancer's class. It seems that for classification of breast cancer's class, logistic regression method is appropriate to be used. The Logistic regression, correctly classifies patients with and without breast cancer 96% of the times. Its AUC of 99% indicates a great ability to distinguish between a benign lump and a malignant tumor

Conclusion

Breast cancer cannot be prevented. However, the deaths due to breast cancer can be reduced. And the reduction is achieved if and only if **'being aware'** of symptoms of breast cancer and get treated on time. A stitch in time saves nine.

References

- Bakr, M. A. H. A., Al-Attar, H. M., Mahra, N. K., & Abu-Naser, S. S. (2020). Breast Cancer Prediction using JNN. *International Journal of Academic Information Systems Research (IJAISR)*, *4*(10).

- **Dataset:**

http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.