CLARK UNIVERSITY

Machine Learning Algorithms to predict & diagnose breast cancer

Nisarg Thakkar
Vikas Jangra
Dhiraj Sharma

CLARK UNIVERSITY

Table of Contents

Significance of Study

CLARK UNIVERSITY

2,300,000

New cases of breast cancer were diagnosed last year.

High Death Rate

Than those for any other cancer, besides lung cancer.

Dataset Description

All things data!

The project is carried out on a single data set, which is provide by the UCI Machine Learning Repository and created by the faculty of the University of Wisconsin. The data in the dataset are collected by the digital images of fine needle aspirants (FNA) of the breast mass, those images describe the characteristics of the cell nuclei present in the mass.

There are 32 attributes present in the dataset including the ID Number and Diagnosis (Malignant and Benign).

Dataset Link: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Data Dictionary

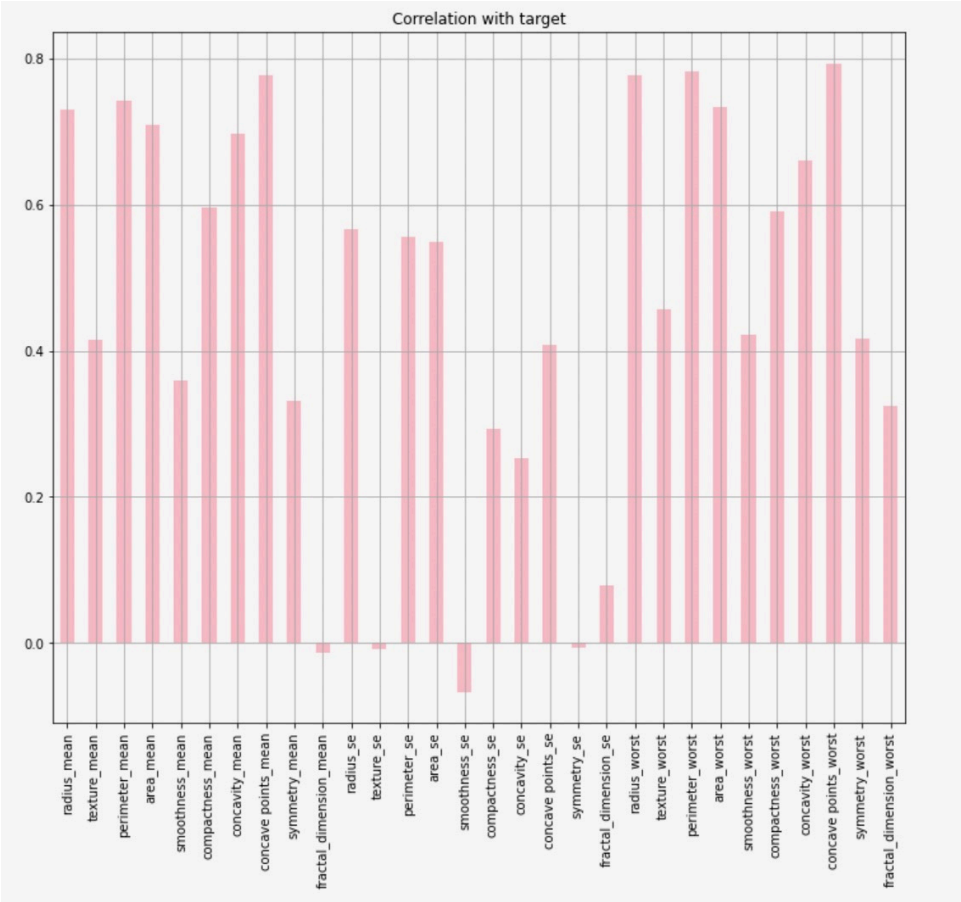
1. Radius – mean of the distance from the center to the points on the perimeter.
2. Texture – standard deviation of the gray-scale values.
3. Perimeter
4. Area
5. Smoothness – local variation in radius.
6. Compactness – perimeter2 / area — 1.0
7. Concavity – severity of the concave portions of the contour
8. Concave points – number of concave portions of the contour.
9. Symmetry
10. Fractal Dimension – Coastline Approximation – 1

Exploring Variables

Identifying Key Variables

Amongst the real value features, the mean and worst features of cell nucleus were highly correlated with our target variable (Malignant and Benign).

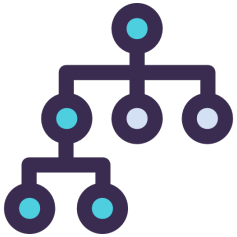
Most observations are over 60% correlated which helped us in finding the accurate predictions.



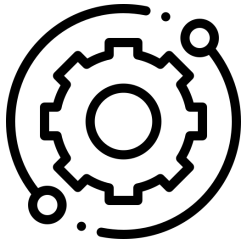
Modelling Techniques

Algorithms Trained

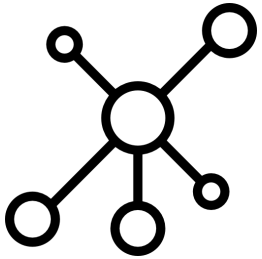
Logistic Regression



Decision Tree



Support Vector Machine



Random Forest Classifier

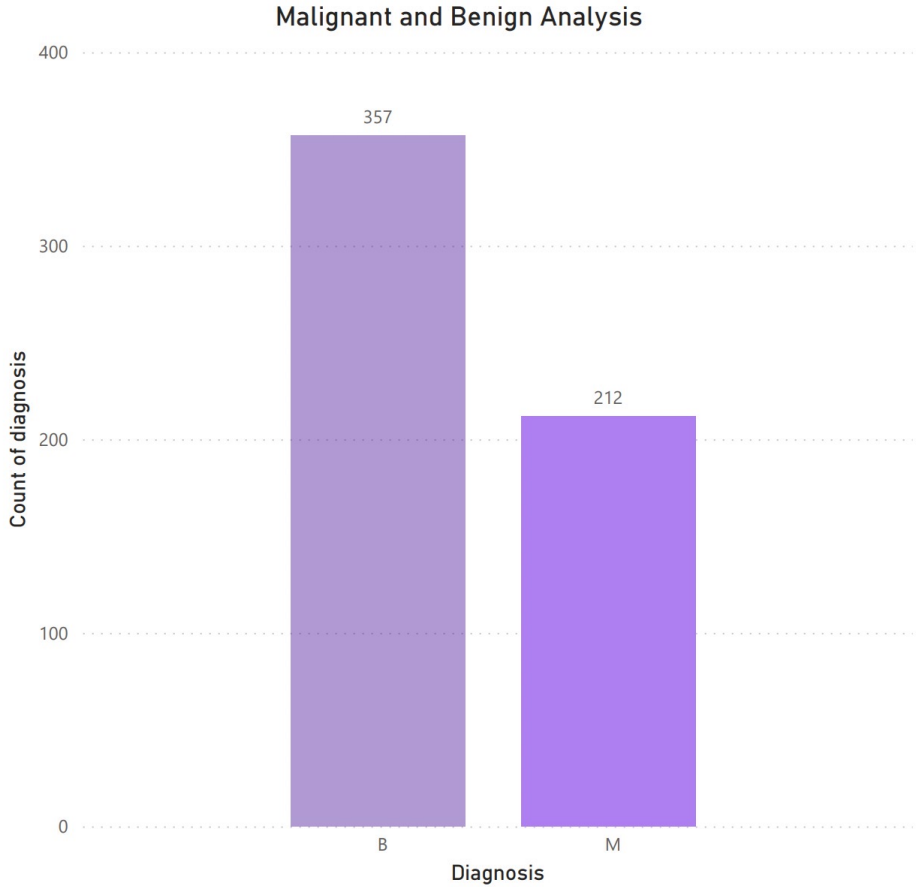
Results & Findings

Analysis

Among the dataset almost 63% were observed with a benign tumor type whereas the remaining 37% were diagnosed with malignant tumor.

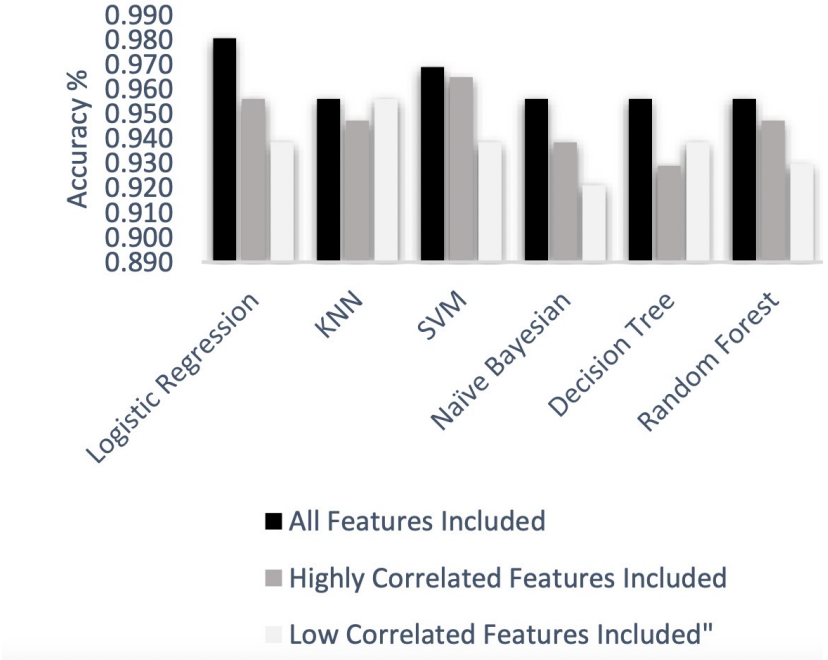
The results were achieved by testing the data on the clear and positive correlated features.

The model can be replicated to diagnose type of tumor provided the dataset has similar variables.



Analysis

When the models were run on highly correlated features, SVM saw the highest accuracy followed by Logistic Regression.



Thank you

References

- https://www.reuters.com/article/health-cancer-int/breast-cancer-overtakes-lung-as-most-common-cancer-who-idUSKBN2A219B
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7349542/pdf/healthcare-08-00111.pdf
- https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
- https://github.com/Prianca25/Machine-Learning/blob/master/Breast%20Cancer%20Dignostics.ipynb
- https://www.kaggle.com/code/vikasukani/breast-cancer-prediction-using-machine-learning#ML-Model-Selecting-and-Model-PredPrediction
- https://www.kaggle.com/code/rafetcan/breast-cancer-classification-99-acc