**BREAST CANCER PREDICTION USING MACHINE LEARNING ALGORITHMS**

Group Members:

Laxma Reddy Nalla

Vinay Rao Avirneni

Gayathri Keshamoni

Uday Kumar

**Motivation:**

Breast Cancer become most common type of cancer in humans. The doctors classified cancer into 4 stages like stage 1 to 4. If cancers are recognized in early stages, there is a high possibility to recover from it.

There are new technologies where we can detect the type of cancer and stage of it, but there is likely to be false positives and true negative results these FP and TN can end up with cost of human life. With the technology we can assess the FP,TN using Machine Learning Classification algorithms.

In this Machine Learning Project, we are proposing comparison between ML classification algorithms which can predict the type of cancer accurately.

**Significance:**

Is it very important to get high prediction accuracy since the false prediction can result in the cost of human life.

**Objectives:**

Compare a different classification algorithm which gives high accuracy for the breast cancer prediction problem.

Implementation of different ML classification algorithms against breast cancer dataset.

**Features:**

**Increment/ Implementation:**

**Dataset:**

[Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle](https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data)

**Dataset Description:**

Breast cancer is the most common cancer amongst women in the world. It accounts for 25% of all cancer cases and affected over 2.1 million people in 2015 alone. It starts when cells in the breast begin to grow out of control. These cells usually form tumors that can be seen via X-ray or felt as lumps in the breast area.

The key challenge against its detection is how to classify tumors into malignant (cancerous) or benign(non-cancerous). We ask you to complete the analysis of classifying these tumors using machine learning (with SVMs) and the Breast Cancer Wisconsin (Diagnostic) Dataset.

**Detailed Description of Features:**

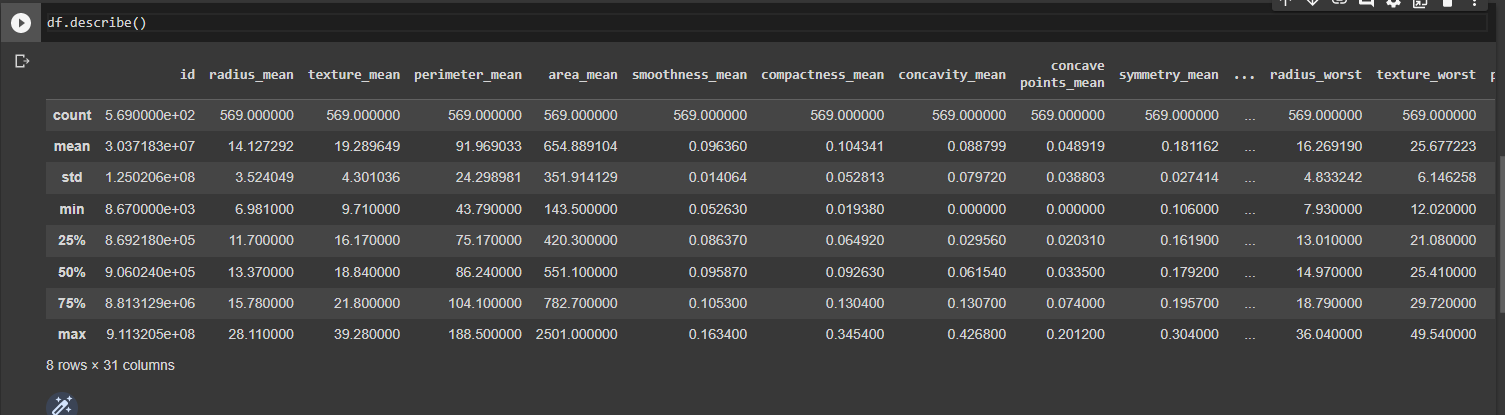
1) ID number  
2) Diagnosis (M = malignant, B = benign)  
3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)  
b) texture (standard deviation of gray-scale values)  
c) perimeter  
d) area  
e) smoothness (local variation in radius lengths)  
f) compactness (perimeter^2 / area - 1.0)  
g) concavity (severity of concave portions of the contour)  
h) concave points (number of concave portions of the contour)  
i) symmetry  
j) fractal dimension ("coastline approximation" - 1)

**Features analysis:**

Dataset statical information:



Type of cancers using count plot:

Benign: 357

Malignant: 212

Chart, bar chart

Description automatically generated

**Count Plot of Type of Cancers**

Chart, histogram

Description automatically generated

**Area mean of breast cancer tumour**

Chart, histogram

Description automatically generated

**Radius of Cancer tumour**

Correlation matrix:

The correlation matrix for the will shows the correlated features in dataset. The breast cancer dataset consists of 32 features, so it is difficult to see the correlation matrix so took the correlation matrix of features which has correlation thresholds above 0.75 with respect to Target.

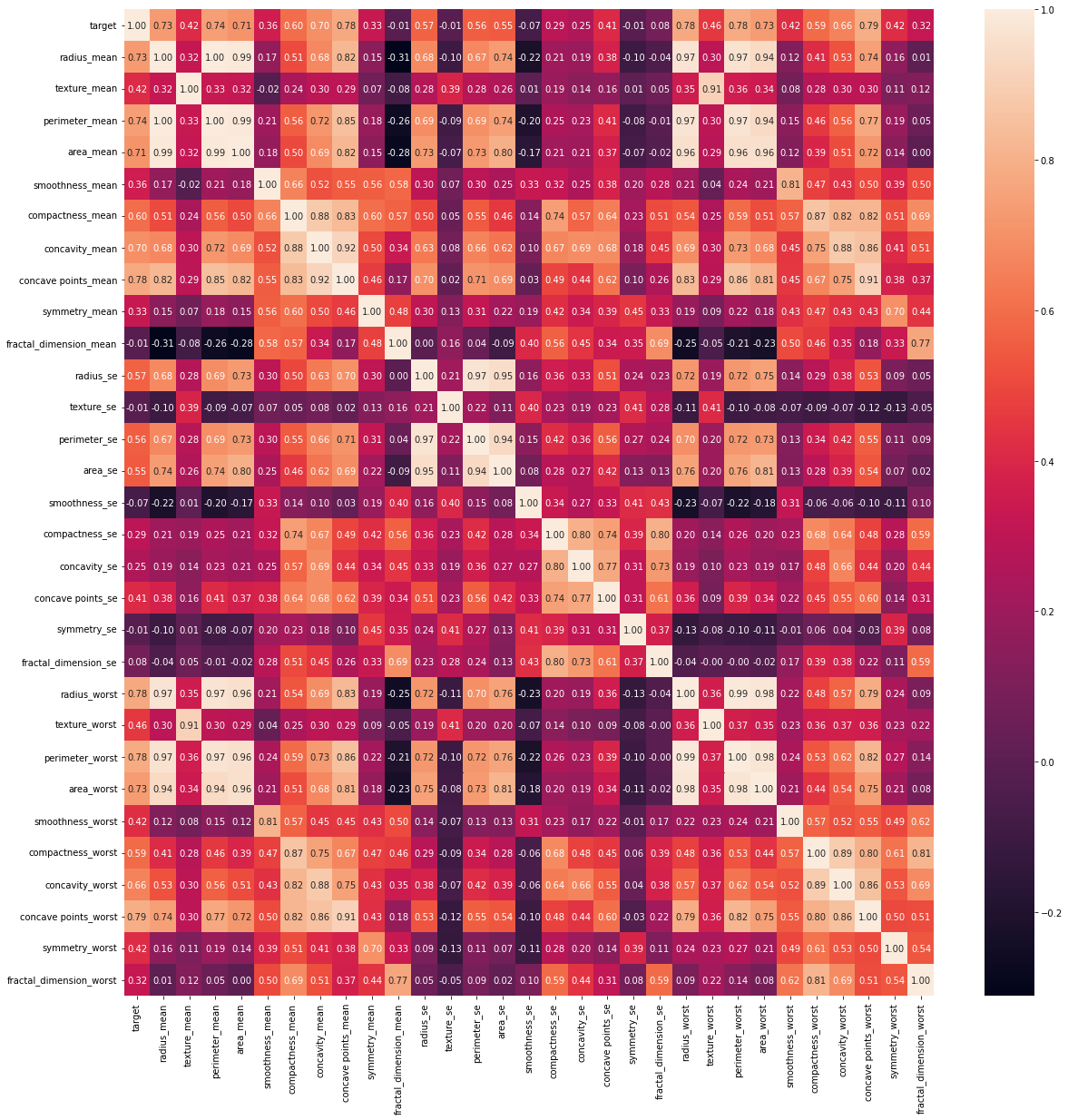
Text

Description automatically generated

Graphical user interface, application, Teams

Description automatically generated

Fig: Correlation matrix of features



There is a high correlation of 0.75 and above is between**radiusworst,perimeterworst,concave pointsmean, concave points worst**these features sowe removed some features such as

**Radiusworst,perimeterworst,concave points**mean these are highly related features. By removing we can reduce features in data with losing important features.

**Implementation of Models:**

For this breast cancer dataset we thought ofusing multiple machine learning algorithms such as

1. PCA-KNN with n\_negibours as 2.
2. NCA-KNN with n\_negibours as 2
3. Logistic Regression
4. Decision Tress
5. SVM
6. XGBoost

**KNN with PCA:**

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

Principal Component Analysis or PCA is a widely used technique for dimensionality reduction of the large data set. Reducing the number of components or features costs some accuracy and on the other hand, it makes the large data set simpler, easy to explore and visualize. Also, it reduces the computational complexity of the model which makes machine learning algorithms run faster.

Text

Description automatically generated

This is the sample implementation of KNN algorithm using 2 neighbors because we are classifying data between weather the cancer is benign or malignant.

Text

Description automatically generated

By applying PCA with KNN algorithm we got the

Accuracy: 0.9321794871794872

Test Scores: 0.947368421052631

Train Score: 0.9346733668341709

**KNN with NCA:**

Neighborhood component analysis (NCA) is a non-parametric method for selecting features with the goal of maximizing prediction accuracy of regression and classification algorithms.It learns a linear transformation in a supervised fashion to improve the classification accuracy of a stochastic nearest neighbor’s rule in the transformed space.

Chart, scatter chart

Description automatically generatedAs you can see above plt nca gives better solution than pca (dots are seperated).

Text

Description automatically generated

By Applying NCA with KNN we the

Accuracy: 0.99

Test Score: 0.99

Train Accuracy: 0.99

**SVM:**

A Support Vector Machine (SVM) is a binary linear classification whose decision boundary is explicitly constructed to minimize generalization error. It is a very powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification, regression and even outlier detection.

SVM is well suited for classification of complex but small or medium sized datasets.

Graphical user interface, text, application

Description automatically generated

A picture containing graphical user interface

Description automatically generated

By implementing svm we got

Accuracy: 0.8829787234042553

These are some of the advantages of SVM:

Effective in high dimensional spaces.

Still effective in cases where number of dimensions is greater than the number of samples.

Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

Versatile: different [Kernel](http://scikit-learn.org/stable/modules/svm.html#svm-kernels) functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

**WORK COMPLETED:**

As we proposed we decided to implement multiple machine learning algorithms we gathered the dataset from Kaggle and carried out some Exploratory data analysis on the dataset. After implementing EDA we started with implemented KNN algorithm in which we implemented KNN with PCA (Principal component analysis) and KNN with NCA (Neighborhood component analysis). We have also implemented the SVM (Support vector machines) algorithm on the breast cancer dataset.

Responsibilities:

We as a group we divided the tasks equally among us:

Each part of the project we divided the work as follows:

* **Uday and Gayathri** have invested their time into referring to different papers and articles to assess and decide the machine learning algorithms.
* **Laxma Reddy and Vinay** have handled the dataset collection, preprocessing, EDA and implementation of algorithms and ML models.

Contribution:

* We four members worked together on this use case by discussing on all the aspects that needs to be done.
* On the whole each person in the group contributed equally to the project i.e 25% of work per person.

**WORK TO BE COMPLETED:**

* At the end of part two of project we left few other Machine Learning Algorithms and the performance tuning part.