Python Final Group Project

“Breast Cancer Detection”

Class: Python Programming (CS-661)

Professor: Agar John Richard

Group 8 Members:

Lolyna de la Fuente Ordaz

Ibrahim Mohammed Hamed

Abhayrajsinh Rana

Chukwuebuka Ozoh

Niyati Patil

Sunny Ma

1. Topic

: “Breast Cancer Detection”

2. Summary

- Dataset: Kaggle

  Dataset : https://www.kaggle.com/code/faresmohammad/breast-cancer-detection

- Purpose:

  The purpose of our project is to predict with accuracy if a tumor cell is malignant or benign by utilizing machine learning algorithms with various predictive models for breast cancer detection. By employing different classification algorithms, we aim to explore and compare their performance in accurately identifying whether a tumor is malignant or benign based on the provided attributes.

  -Independent Variable (X) = All the columns except for the ‘Diagnosis’ variable

  -Dependent Variable(y) = ‘Diagnosis’ (Predicting if a patient has cancer (1=M) or not (0=B))

  The analysis involved data preprocessing, exploratory analysis, outlier detection, model building, evaluation, and comparison. Each step contributes to understanding the dataset and evaluating the performance of various machine learning algorithms in breast cancer detection.

  1. Exploratory Analysis: Conduct exploratory data analysis (EDA) to understand the distribution and relationships between different features within the dataset. This will involve visualizations and statistical summaries to gain insights into the data.
  2. Feature Importance: Evaluate the importance of features in determining the nature of the tumor (malignant or benign) using various algorithms such as Decision Trees, Random Forests, and others to identify which attributes have the most significant impact on the classification.
  3. Model Development: Implement different classification algorithms including Logistic Regression, Naive Bayes, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, and Random Forests. Compare and evaluate their performance in terms of accuracy, precision, recall, F1-score, and area under the ROC curve.
  4. Model Comparison: Analyze and compare the performance metrics of each model to identify the most effective algorithm for breast cancer detection based on the dataset.
  5. Optimization and Validation: Optimize the selected models by tuning hyperparameters and conduct cross-validation to ensure robustness and reliability of the chosen algorithm(s) for accurate prediction.

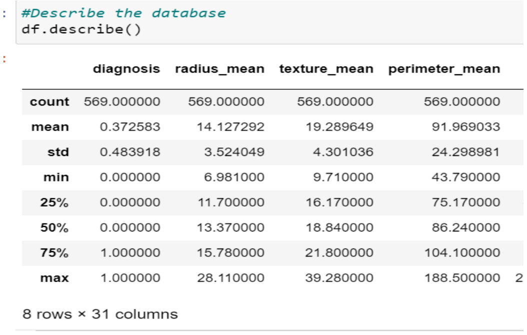
3. Predictive Models

1. EDA - Lolyna / 2. Logistic Regression (Scikit-learn) - Ebukka/

3. Support Vector Machine (Linear) (Scikit-learn) - Abhay / 4. K-Nearest Neighbours (Scikit-learn) - Sunny

5. Decision Tree (Scikit-learn) - Ibrahim/ 6. Random Forest (Scikit-learn) - Niyati

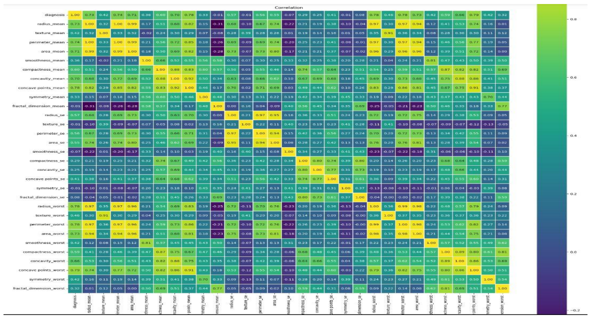
4. Data Visualization

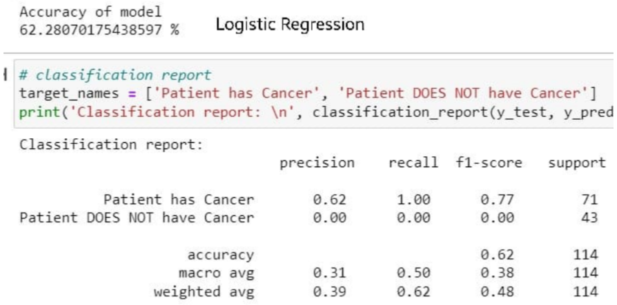


```
#Describe the database
df.describe()
```

	diagnosis	radius_mean	texture_mean	perimeter_mean	
count	569.000000	569.000000	569.000000	569.000000	
mean	0.372583	14.127292	19.289649	91.969033	
std	0.483918	3.524049	4.301036	24.298981	
min	0.000000	6.981000	9.710000	43.790000	
25%	0.000000	11.700000	16.170000	75.170000	
50%	0.000000	13.370000	18.840000	86.240000	
75%	1.000000	15.780000	21.800000	104.100000	
max	1.000000	28.110000	39.280000	188.500000	2

8 rows × 31 columns



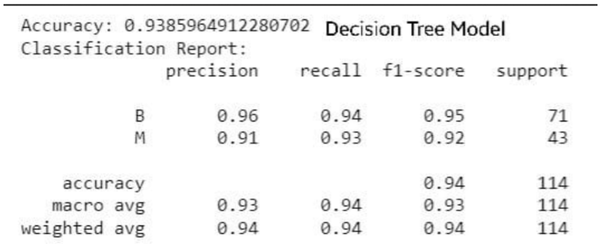


```
Accuracy of model
62.28070175438597 %
```

Logistic Regression

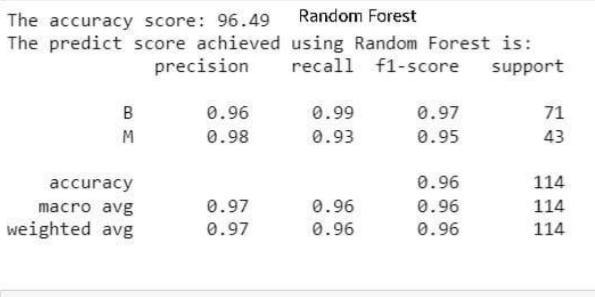
```
# classification report
target_names = ['Patient has Cancer', 'Patient DOES NOT have Cancer']
print('Classification report: \n', classification_report(y_test, y_pred
```

```
Classification report:
                               precision    recall  f1-score   support

          Patient has Cancer       0.62      1.00      0.77        71
Patient DOES NOT have Cancer       0.00      0.00      0.00        43

                    accuracy                           0.62       114
                   macro avg       0.31      0.50      0.38       114
                weighted avg       0.39      0.62      0.48       114
```



```
Accuracy: 0.9385964912280702
Classification Report:
              precision    recall  f1-score   support

           B       0.96      0.94      0.95        71
           M       0.91      0.93      0.92        43

    accuracy                           0.94       114
   macro avg       0.93      0.94      0.93       114
weighted avg       0.94      0.94      0.94       114
```

Decision Tree Model

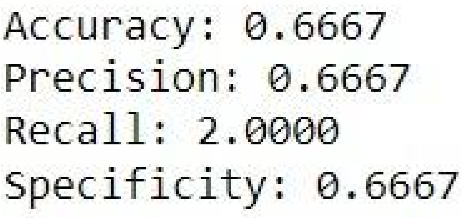


```
The accuracy score: 96.49
The predict score achieved using Random Forest is:
              precision    recall  f1-score   support

           B       0.96      0.99      0.97        71
           M       0.98      0.93      0.95        43

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114
```

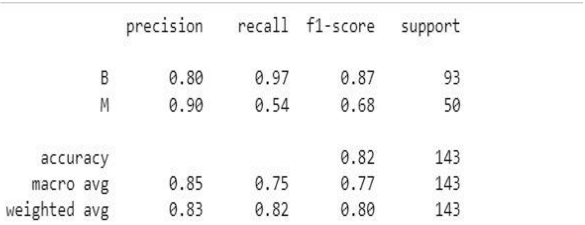
Random Forest

Support Vector Machine Algorithm

K-NN/ Accuracy: 81.82%



```
Accuracy: 0.6667
Precision: 0.6667
Recall: 2.0000
Specificity: 0.6667
```



```
              precision    recall  f1-score   support

           B       0.80      0.97      0.87        93
           M       0.90      0.54      0.68        50

    accuracy                           0.82       143
   macro avg       0.85      0.75      0.77       143
weighted avg       0.83      0.82      0.80       143
```