Breast Cancer Prediction

Members: Nilakshi Pokharkar, Shefali Gokarn, Mithika Pawar

Role of each team member:

1. Nilakshi Pokharkar:
   - To understand and compare the MCC with respect to other evaluation metrics.
   - Implementing MCC on each model present in the program.
   - To study GridSearchCV and its significance in the non-parametric algorithms.
   - Implementing GridSearchCV on the kNN classifier and predicts results along-with the study of accuracy output.
   - Implement Multilayer Perceptron on the dataset and train an predict its results.

The following research papers were studies by me to evaluate the metrices.

[1] Davide Chicco & Giuseppe Jurman, 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation'.

[2] Davide Chicco, Niklas Tötsch & Giuseppe Jurman, 'The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation'.

2. Shefali Gokarn:
   - Implementation of KNN model creation fitting
   - Implementation of Random Forest with K-Fold model creation fitting
   - Evaluation metrices: To evaluate all the built modules and to check their performance and make a comparison between the following metrices were used.
     i) Confusion matrix
     ii) Precision
     iii) Recall
     iv) Accuracy
     v) f1-score
     vi) Support

The following research papers were studies by me to build and check the model's performance.

[1] Mohammad H., Alshayeji, HanemEllethy, Sa'edAbed, RenuGupta, 'Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach'.

[2] KNN Classifier in Sklearn using GridSearchCV with Example

3. Mithika Pawar:
   - To understand and perform the pre-processing on data
   - Cleaning unnecessary rows or columns
   - Checking whether the data is biased or no to provide a nearly balanced or balanced data to get more accurate model and performing exploratory data analysis to make the data more precise

- Also, performing correlation matrix to determine the correlation between different columns. And performing the diagnosis on data.

The following research papers were studies by me to build and check the model's performance.

[1] Breast Cancer Wisconsin (Diagnostic) Data Set

[2] Ramik Rawal, 'BREAST CANCER PREDICTION USING MACHINE LEARNING'.

Summary:
Our project aims to study and implement different machine learning algorithms and techniques on the breast cancer dataset and perform comparative study on different evaluation metrics like Accuracy, Precision, F1-Score, Support. Moreover, we added Matthews Correlation Coefficient to analyze the models we implemented since the data is biased. In addition, we have also explored different techniques of optimal hyperparameter tuning and searching like GridSearchCV, Optuna and applied cross validation across the models to test the improvement in its accuracy. We have implemented kNN, kNN with GridSearch, MLP and Random Forest with K-Fold models to train and classify the breast cancer images as Malignant and Benign.

Resource Links:

1. **GitHub: https://github.com/nilpokharkar/ece_5831_Project.git (https://github.com/nilpokharkar/ece_5831_Project/blob/main/BreastCancerPrediction%20(2).ipynb)**

2. **Dataset: https://drive.google.com/file/d/1I0RcqzqB-56XJt6JEvLr61bIoOXJdSyY/view?usp=share_link**

3. **Presentation: https://youtu.be/E9IY0nTCJPE**