BREAST CANCER DETECTION

SMRUTHI GOWTHAM

SNIGDHA SINHA

VRIDHI GOYAL

YASHI CHAWLA

PROBLEM STATEMENT

- Breast Cancer detection aims at extracting features from the given samples of the Wisconsin Breast Cancer Dataset and predicting it as Benign or Malignant.
- Targets at coming up with an improvised approach which helps in effective and accurate prediction of malignant and benign breast cancer.
- Aims to help detect the occurrence of cancer at an early stage to avoid preventable deaths.

DATASET

- Wisconsin Breast Cancer Dataset (WBCD), acquired from the repository of UCI Machine learning and is a benchmark dataset.
- It is distributed over 37.25% cancerous samples and 62.75% non-cancerous samples.
- It contains 569 instances and 32 features that provide accurate information regarding the diagnosis of breast cancer.
- The columns of the dataset represent the features of the cell nuclei found in the digitised image of fine needle aspirates (FNA) of breast mass.

USEFULNESS

- 2020 diagnosis: 2.3 million women diagnosed. 685,000 global deaths.
- 5-year survival rate: 7.8 million women as of 2020 => most prevalent cancer.
- Approximately 13% of the diagnoses missed Stage 1 breast cancer. Meanwhile, 48% failed to detect atypia hyperplasia, a precursor to breast cancer.
- The number of women who have died of breast cancer has decreased by 41% from 1989 to 2018 thanks to early detection and treatment improvements. As a result, more than 403,000 breast cancer deaths were prevented during that period.
- Diagnosis stage:
    - 64% of breast cancer patients have local-stage breast cancer
    - 27% have regional stage
    - 6% have distant (metastatic) disease.

.

APPROACH

- For the analysis of our problem statement, six ML models have been used which are, Logistic Regression, Support Vector Classifier (SVC), AdaBoost, Neural Network, K-Neighbours classifier and Random Forest.
- Grid Search was performed on all models to find out the best parameters, with various combinations of preprocessing.
- PCA was performed to reduce the dimensionality of the dataset.
- Various preprocessing combinations were tried out with every model to determine the best results.
- Logistic Regression with preprocessing as standardization and PCA gave the best results with F1 score as 98% on the dataset.

EVALUATION

- Considering the size of the dataset, we implemented cross validation on 15 splits.
- Grid search was implemented to get the best set of hyperparameters.
- F1 Non- stratified and Stratified scores were used as the base metrics for the models evaluation since it handles multi class and class imbalance.
- Accuracy, Roc-Auc score, Mean Squared Error were calculated based on the best parameters.

CONTRIBUTION

- Smruthi and Snigdha performed the EDA and visualisation in the first phase of the project, providing us with insights to the data and visualizations to have a better understanding of the data. Additionally worked on SVC model for prediction.
- Vridhi worked on literature review and provided us with the background information on various works done on breast cancer predictions. Also helped in calculating the performance metrics for the different models used.
- Yashi helped out with the literature review to provide background information for the already existing work for the problem and contributed towards working on providing different combinations of preprocessing techniques and by performing grid search to choose the best parameters for Logistic Regression, AdaBoost, Random Forest, KNN.

THANK YOU