Predicting Breast Cancer in a patient

Abstract:

Breast cancer represents one of the diseases that make a high number of deaths every year. It is the most common type of all cancers and the main cause of women's deaths worldwide. Classification and data mining methods are an effective way to classify data. Especially in the medical field, where those methods are widely used in diagnosis and analysis to make decisions.

Problem Statement:

Given the details of cell nuclei taken from breast mass, predict whether or not a patient has breast cancer using the Ensembling Techniques. Perform necessary exploratory data analysis before building the model and evaluate the model based on performance metrics other than model accuracy.

Dataset Information:

The dataset consists of several predictor variables and one target variable, Diagnosis. The target variable has values 'Benign' and 'Malignant', where 'Benign' means that the cells are not harmful or there is no cancer and 'Malignant' means that the patient has cancer and the cells have a harmful effect

Variable Description:

Column	Description
radius	Mean of distances from center to points on the perimeter
texture	Standard deviation of gray-scale values
perimeter	Observed perimeter of the lump
area	Observed area of lump
smoothness	Local variation in radius lengths
compactness	perimeter^2 / area - 1.0
concavity	Severity of concave portions of the contour

concave points	number of concave portions of the contour
symmetry	Lump symmetry
fractal dimension	"coastline approximation" - 1
Diagnosis	Whether the patient has cancer or not? ('Malignant','Benign')

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

Scope:

- Analysing the available data and exploring relationships among given variables
- Data Pre-processing
- Training SVM classifier to predict whether the patient has cancer or not
- Assess the correctness in classifying data with respect to efficiency and effectiveness of the SVM classifier in terms of accuracy, precision, sensitivity, specificity and AUC ROC
- Tuning the hyperparameters of SVM Classifier provided by the scikit-learn library

Learning Outcome:

The students will get a better understanding of how the variables are linked to each other and build an SVM model. Apart from various performance measures, they will also learn about hyperparameter tuning with cross-validation to improve these scores.