Breast Cancer Prediction

Problem Statement

Breast cancer is one of the most common cancers among women in the world. Early detection of breast cancer is essential in reducing their life losses. Build a predictive model using machine learning algorithms to predict whether the tumor is benign or malignant.

Data Description

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. 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.

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)

Evaluation

Evaluation will be based on:

- Data Preparation
- Model Selection

Data Preparation

Check the data distribution of variables and perform transformation if a variable's distribution is skewed. Perform label encoding on categorical variables.

Model Selection

Implement KNN algorithm on training data, predicting labels for dataset and printing the accuracy of the model for different values of K.

Expected Outcome

Find the optimal k value having the lowest misclassification error, highest accuracy and highest AUC.