Project Proposal

Problem Statement and Motivation:

Breast cancer (BC) is one of the most common cancers among women in the world today.

An early diagnosis of BC can greatly improve the prognosis and chance of survival for patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of tumors into malignant or benign groups is a subject of much ongoing research.

Machine Learning is a widely recognized methodology in this regard. With the use of advanced machine learning algorithms, I plan to build a model which accurately classifies tumors as Benign or Malignant based on certain features.

Data:

The dataset has been obtained from Kaggle. It contains 596 rows and 32 columns of tumor shape and specifications. The tumor is ultimately classified as benign or malignant based on its geometry and shape.

The features of the dataset include tumor radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, 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 and fractal dimension.

Brief outline of the project:

EDA and Visualization:

The dataset will be examined for any missing or null values and outliers. Violin plots will be plotted to visualize the 32 features which will throw some insights into the distribution pattern of the features, their mean, std or variance. A correlation matrix will also be used to check for multicollinearity between the features.

Statistical Testing:

This step might include building a logistic regression model using statsmodels. It might help us decide which features are most important in predicting the result.

Machine Learning:

Decision Tree based feature selection or Recursive Feature Elimination will be used to decide the feature importance, followed by scaling and training testing the data in different machine learning algorithms like KNearest Neighbor, SVM, GaussianNB, Decision Tree and Random Forest.

Deliverables:

Deliverables include report, source code and presentation slides.