Final Project Proposal

Title: “Predicting Breast Cancer Patient Outcomes Using Machine Learning”

Problem Statement: Breast cancer is the major cause for mortality among women globally. Early prediction of survival outcomes can help in proactive treatment planning and reduce the risk of fatality. Manual prognosis is often time-consuming and subjecting. Leveraging machine learning models offers a scalable and data-driven solution to predict survival outcomes based on clinical features.

Objective: To develop a Machine Learning model that predicts the survival status (Alive or Dead) of breast cancer patients using hormonal, tumor-specific, age and demographic clinical features.

Dataset Description:

The dataset is Breast Cancer Patient Diagnosis with 4024 rows × 16 columns including Age, Tumor Size (mm), Regional Node Examined, T Stage and target column ‘Status’.

The dataset is taken from Kaggle. For reference, the data is available at this link: https://www.kaggle.com/datasets/reihanenamdari/breast-cancer

Methodology: Pipeline involves in

- Handling missing value and encoding the categorical values

-Splitting the dataset into training and testing datasets

-Standardizing features and train by using multiple classification models and comparing the performance.

Models: The models that will be used -Logistic Regression, Decision Tree Classifier, Random Forest and XGboost.

Evaluation Metrics: The models will evaluate using Accuracy, Precision, Recall, F1-Sore, and ROC-AUC Score.

Tools: This project will be done in python with data analysis, manipulation, model building and visualization libraries such as pandas, matplotlib, seaborn and scikit-learn.

The aim is to apply these models on breast cancer diagnosis data to find the cancer survival outcomes. These models can assist healthcare professionals in making effective decisions early in the treatment process.