Final Project Proposal
Title: Breast Cancer Prediction

Contributors ➔

Full Name	Students ID
Reza Zare	**002287388**

1. **Project Objective ➔** to develop a model that accurately predicts the diagnosis of breast masses as either benign or malignant, based on various medical and demographic factors and available patient data.

2.1. Description ➔ this project aims to develop a model that can predict the likelihood of breast cancer in patients. The data is preprocessed and cleaned to remove missing values, outliers and normalize the data. Feature selection methods are then used to identify the most important predictors of breast cancer. A ML algorithm such as decision tree is then trained on the data to develop the model. The model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score.

2.2. Significance ➔ The breast cancer prediction project has significant implications for healthcare. Early detection of breast cancer is critical for successful treatment and improved patient outcomes. By accurately predicting the likelihood of breast cancer, healthcare professionals can implement preventative measures to reduce the risk of developing breast cancer.

2.3. Hypotheses ➔ The hypotheses, includes: "The features extracted from breast cancer cells can accurately predict whether the cells are malignant or benign."

2.4. Assumptions ➔ The data is representative, complete, independent, normally distributes, balanced and target variable is binary.

3. **High-Level Methodological Outline To Achieve Project Objectives ➔**
   - **Business / Problem understanding ➔** to develop a model that can accurately predict the likelihood of breast cancer in patients based on various medical factors.
   - **Data Understanding ➔** dataset contains information about various features of breast cancer cells and a corresponding classification of whether the cells are malignant or benign.
   - **Exploratory Data Analysis (EDA) ➔** involves descriptive statistics, correlation analysis, analyzing the various features of breast cancer cells and their classifications to gain insights into the data.
   - **Data Preparation ➔** involves data cleaning, feature selection and engineering, data splitting, handling class imbalance and encoding categorical variables which leads to more accurate predictions.
   - **Model Building ➔** create a model that can accurately classify breast cancer cells as malignant or benign based on their features, which is used to improve diagnosis and treatment of breast cancer.
   - **Model Evaluation ➔** involves measuring how accurately the model can predict the target on the testing data and using metrics such as accuracy, precision, recall, and F1-score to assess its predictive power.

4. Resources ➔

Dataset: The Breast Cancer Wisconsin dataset is accessible on UCI, Scikit-learn, R packages, Kaggle and other data science repositories.

Tools: We will need various data analysis and machine learning tools such as Python, Jupyter Notebook & VScode IDM, and any relevant libraries and packages such as Pandas, NumPy, Scikit-learn, Seaborn and Matplotlib, etc.

5. Deliverables ➔
   - A well-documented Jupyter Notebook contains the code used for data cleaning, exploration, feature engineering, model training, and evaluation, a README file for dependencies.
   - A final report summarizing the project, including data visualization, data exploration and analysis, the machine learning model used, the evaluation metrics, and the results obtained.
   - A presentation summarizing the key findings and insights from the project, involve presenting the characteristics of breast cancer cells and their classifications using a wide range of visualization dashboards.

6. References/Bibliography ➔

- Wisconsin Diagnostic Breast Cancer (WDBC) Dataset, UCI Machine Learning Repository: (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))
- Breast Cancer Wisconsin (Diagnostic) Data Set, Kaggle. (https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data)
- Wisconsin Diagnostic Breast Cancer (WDBC) Dataset, Scikit-Learn Machine Learning Repository: (https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html)