**Project\_4 Proposal**

**Title:** Comparing Supervised Learning Algorithms for Breast Cancer Diagnosis

Topics to be covered:

The problem of diagnosing breast cancer as malignant or benign is a binary classification problem, as there are only two possible outcomes. Therefore, we can build a binary classification model using supervised learning algorithms.

Some algorithms that could be used are:

* Logistic Regression
* Decision Tree
* Random Forest
* Support Vector Machine (SVM)
* Neural Network

**Research Question**

Which supervised learning algorithm performs the best in diagnosing breast cancer as malignant or benign using the Breast Cancer Wisconsin (Diagnostic) dataset?

**Purpose**

The purpose of this project is to compare the performance of different supervised learning algorithms in classifying breast cancer as malignant or benign based on the features present in the dataset. The results of this study can help in identifying the most effective algorithm for diagnosing breast cancer, which can ultimately contribute to improving the accuracy and efficiency of breast cancer diagnosis.

**Scope**

The project will focus on the Breast Cancer Wisconsin (Diagnostic) dataset, which includes measurements from digitized images of fine needle aspirates of breast masses. The study will use supervised learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Neural Network, to classify the breast cancer diagnosis. The project will evaluate the performance of each algorithm using metrics such as accuracy, precision, recall and F1 score.

**Limitations**

The study will be limited to the Breast Cancer Wisconsin (Diagnostic) dataset, which includes only ten features of cell nuclei present in fine needle aspirate images of breast masses. The dataset does not include other potential risk factors or clinical information that may affect breast cancer diagnosis. Additionally, the study will be limited to the performance of the selected supervised learning algorithms, and other algorithms or techniques may be more effective for diagnosing breast cancer.

**Overview**

The project will involve the following steps:

Data preparation: The Breast Cancer Wisconsin (Diagnostic) dataset will be downloaded from the kaggle.com and preprocessed, including data cleaning, feature selection, and normalization.

Model building: Five different supervised learning algorithms (Logistic Regression, Decision Tree, Random Forest, SVM, and Neural Network) will be implemented using Python libraries such as Scikit-learn, Tensorflow, and Keras.

Model evaluation: The performance of each algorithm will be evaluated using metrics such as accuracy, precision, recall and F1 score. The results will be compared to identify the most effective algorithm for diagnosing breast cancer.

Visualization: Logistic Regression Confusion Matrix, Decision Tree Confusion Matrix, Random Forest Feature Importances and Confusion Matrix for SVM Model will be presented, Comparisons of the accuracy results

**Analysis**

The analysis will focus on comparing the performance of different supervised learning algorithms in diagnosing breast cancer. The results will be presented using visualizations such as confusion matrices, and precision-recall curves. The findings will be discussed, and recommendations for the most effective algorithm will be made based on the evaluation metrics.