Capstone Project Documentation

Customer Segmentation + Purchase Prediction using Supervised and Unsupervised Learning

1. Project Overview

This capstone project demonstrates both supervised and unsupervised machine learning techniques. A predictive classification model is built using Logistic Regression, and clustering is applied using KMeans to group similar samples. The project also includes an interactive Gradio app that runs inside Google Colab to allow users to test predictions.

2. Objectives

• Build a supervised learning model to predict tumor type (malignant vs benign).

• Apply unsupervised learning to cluster similar tumor samples.

• Deploy a simple interactive application for predictions and cluster assignment.

3. Dataset

The Breast Cancer Wisconsin dataset from Scikit-learn was used. It contains 569 samples with 30 numerical features describing tumor characteristics. The target labels are:

Target Value	Meaning
0	Malignant
1	Benign

4. Supervised Learning (Predictive Model)

A Logistic Regression classifier was trained using a pipeline consisting of StandardScaler and LogisticRegression. StandardScaler normalizes all feature values to improve model performance. The dataset was split into training (80%) and testing (20%) using stratified sampling.

Evaluation metrics included accuracy score, classification report (precision, recall, F1-score), and a confusion matrix.

5. Unsupervised Learning (Clustering)

KMeans clustering was applied after scaling the full dataset. The optimal number of clusters (K) was selected by computing silhouette scores for K values between 2 and 7. The K value with the highest silhouette score was chosen as the final clustering configuration.

Clusters were visualized in 2D using PCA (Principal Component Analysis), which reduces the feature space from 30 dimensions to 2 dimensions for plotting.

6. Application Development (Gradio App)

A Gradio interface was built to run inside Google Colab. The app allows users to enter tumor feature values and receive two outputs: (1) a tumor class prediction from the supervised model and (2) a cluster group assignment from the KMeans model.

7. Conclusion

The supervised model achieved high accuracy on the test set, showing strong predictive performance. The clustering model provided additional insight by grouping similar samples. The Gradio app provides a simple deployment-style interface for demonstrating machine learning results directly in Colab.

Appendix: Tools & Libraries Used

• Python (NumPy, Pandas) • Scikit-learn (datasets, preprocessing, model training, clustering) • Matplotlib & Seaborn (visualization) • Gradio (interactive app in Colab) • ReportLab (PDF generation)