Project Overview

Objective: To help students apply data mining techniques (Association Rule Mining, Classification, Clustering) on real datasets using only core libraries (NumPy, Pandas, and visualization).

Tools Allowed: NumPy, Pandas, Matplotlib / Seaborn only. **No** high-level ML libraries (e.g., scikit-learn, TensorFlow).

Learning Outcome: Understand the full lifecycle of data mining — from preprocessing and exploration to model building and evaluation.

Datasets Description

Sr.	Dataset	Link	Algorithm
1	Online Retail	https://www.kaggle.com/datasets/vijayuv/onlineretail	Apriori
2	Heart Disease	https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset	ID3
3	Credit Card Customers	https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers	K-Means

Timeline & Weekly Plan

Week	Date Range	Task	Description
1	16 Jun – 21 Jun	Data Preprocessing & Association Rule Mining – Part 1	Clean & explore retail data. Handle missing values, outliers, and perform data transformations.
2	23 Jun – 28 Jun	Data Preprocessing & Association Rule Mining – Part 2	Perform one-hot encoding and generate transaction format suitable for Apriori.
3	30 Jun – 05 Jul	Data Preprocessing & Classification – Part 1	Explore heart disease dataset, handle nulls, and encode categorical data.
4	07 Jul – 12 Jul	Data Preprocessing & Classification – Part 2	Normalize data, perform feature selection, and prepare target attribute.

5	14 Jul – 19 Jul	Data Preprocessing & Clustering – Part 1	Explore credit card dataset; handle scaling and outlier detection.
6	21 Jul – 26 Jul	Data Preprocessing & Clustering – Part 2	Finalize cluster features; decide number of clusters using Elbow method.
7	28 Jul – 02 Aug	Apply Apriori on Online Retail Dataset	Implement Apriori algorithm. Generate frequent itemsets and association rules.
8	04 Aug – 09 Aug	Evaluate Apriori Results	Use support, confidence, lift for evaluation.
9	11 Aug – 16 Aug	Apply ID3 on Heart Disease Dataset	Implement ID3 decision tree. Train on preprocessed data.
10	18 Aug – 23 Aug	Evaluate Classification Results	Evaluate using accuracy, precision, recall; create decision boundaries and visualize.
11	25 Aug – 30 Aug	Apply K-Means on Credit Card Dataset	Implement K-Means. Use preprocessed features to cluster customer types
12	01 Sep – 06 Sep	Evaluate Clustering Results	visualize clusters

Preprocessing Tasks (Weeks 1–6):
25 mini tasks or questions per week (e.g., handling missing values, outlier detection, scaling, encoding).
Algorithm Implementation (Weeks 7–12):
No scikit-learn or built-in models allowed.
Only use NumPy, Pandas, and Matplotlib/Seaborn.
Clear modular implementation and visualizations are expected.

Progress of student is evaluated on every week.