Objective:
This project aims to help students apply three fundamental data science techniques: Naive Bayes for classification, Market Basket Analysis for association rule learning, and Text Analysis for text classification. The project outcomes, including code and reports, will be hosted on GitHub. Students will work with a variety of datasets to demonstrate their ability to extract insights, make predictions, and perform text classification.

Tasks:

1. Naive Bayes Classifier
Objective: Build a Naive Bayes classifier to predict a target variable.

Deliverables:
- Python code implementing the Naive Bayes classifier.
- A short report explaining your data processing steps, model choice, performance evaluation, and insights gained from the model.

2. Market Basket Analysis
Objective: Use association rule learning techniques to perform Market Basket Analysis.

Deliverables:
- Python code implementing Market Basket Analysis.
- A short report summarizing the rules discovered and their potential real-world applications.

3. Text Analysis (Extra Model)
Objective: Build a Naive Bayes text classification model to categorize text data.

Deliverables:
- Python code implementing the Naive Bayes text classifier.
- A short report explaining the text preprocessing, model performance, and insights gained from the analysis.

Datasets:

1. Datasets for Naive Bayes Classification:
Dataset 1: Iris Flower Classification Dataset (https://archive.ics.uci.edu/ml/datasets/iris)

Dataset 2: Wine Quality Dataset (https://archive.ics.uci.edu/ml/datasets/wine)

Dataset 3: Titanic Dataset (https://www.kaggle.com/c/titanic/data)

Dataset 4: Pima Indians Diabetes Dataset (https://www.kaggle.com/uciml/pima-indians-diabetes-database)

Dataset 5: Spam Email Dataset (https://archive.ics.uci.edu/ml/datasets/spambase)

Dataset 6: Heart Disease Dataset (https://www.kaggle.com/ronitf/heart-disease-uci)

Dataset 7: Breast Cancer Wisconsin Dataset (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))

2. Datasets for Market Basket Analysis:

Dataset 8: Online Retail Dataset (https://archive.ics.uci.edu/ml/datasets/online+retail)

Dataset 9: Groceries Dataset (https://www.kaggle.com/heeraldedhia/groceries-dataset)

Dataset 10: Retail Dataset (https://www.kaggle.com/saurabhshahane/retaildataset)

Dataset 11: Instacart Market Basket Analysis Dataset (https://www.kaggle.com/c/instacart-market-basket-analysis/data)

Dataset 12: E-Commerce Retail Dataset (https://www.kaggle.com/carrie1/ecommerce-data)

Dataset 13: Supermarket Dataset (https://www.kaggle.com/abdulhamitcelik/supermarket-dataset)

Dataset 14: Brazilian E-Commerce Dataset (https://www.kaggle.com/olistbr/brazilian-ecommerce)

3. Datasets for Text Analysis:

Dataset 15: Sentiment140 Dataset for Sentiment Analysis (https://www.kaggle.com/kazanova/sentiment140)

Dataset 16: Spam Email Dataset (https://archive.ics.uci.edu/ml/datasets/spambase)

Dataset 17: Amazon Fine Food Reviews Dataset (https://www.kaggle.com/snap/amazon-fine-food-reviews)

Dataset 18: IMDb Movie Reviews Dataset (https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews)

Dataset 19: Twitter Airline Sentiment Dataset (https://www.kaggle.com/crowdflower/twitter-airline-sentiment)

Dataset 20: Yelp Reviews Dataset (https://www.kaggle.com/yelp-dataset/yelp-dataset)

Dataset 21: Fake News Detection Dataset (https://www.kaggle.com/c/fake-news/data)

Outcome and Submission:

GitHub Repository: Each team must create a public GitHub repository for the project. The repository should include:

- A well-structured folder hierarchy (e.g., Naive_Bayes/, Market_Basket_Analysis/, Text_Analysis/).
- Clear documentation (README file) explaining the project, how to run the code, and how the datasets were used.
- Python code for each part of the project, properly commented and organized.
- The reports for each task (Naive Bayes, Market Basket Analysis, and Text Analysis).

Grading Criteria:

- Code Functionality: The implementation of Naive Bayes, Market Basket Analysis, and Text Analysis models, the use of proper libraries, and the handling of datasets.
- Clarity of Reports: Quality of the reports, explanation of steps, and the insights drawn from the analysis.
- GitHub Presentation: Repository structure, clarity of instructions, and ease of navigation.