Homework 2

In this homework, you will work with a dataset that contains information about students taking different online courses. The dataset has a lot of features, including demographic details like age and gender, behavior on the platform like time spent and activities completed, and past performance. The main goal is to predict the target column "Completed", which tells us if a student finished the course or dropped out. One important thing to notice is that a large part of the data does not have the "Completed" label, so you can't just train the model directly on everything. You will need to handle these missing labels. A simple approach is to focus first on training your model using only the labeled data to establish a reliable baseline. Then, you can explore ways to make use of the unlabeled data without introducing too much noise. For example, you might assign tentative labels to some of these unlabeled examples based on predictions from your model or by using unsupervised approaches.

To complete this task, you must follow these steps:

1) Exploratory Data Analysis (EDA) – 2p
    a. Describe the structure of the data: number of rows, columns, and data types.
    b. Identify correlations between data
    c. Make plots to show data distributions
    d. Identify target and feature variables
    e. Highlight potential issues in the Data
2) Data preprocessing – 2 p
    a. Identify outliers
    b. Deal with missing values
    c. Deal with categorical data
    d. Standardize or normalize numerical features as needed
    e. Use principal component analysis (PCA)
3) Propose and test 2 different methods that deal with data imbalance (e.g. SMOTE, ADASYN, undersampling, etc.) – 1 p
4) Test out 5 different ML methods and print out all performance metrics (Recall, Precision, Accuracy, F1). Test the performance of Voting Classifier which combines predictions from multiple models – 4p
5) Document in a report everything that you experimented, from data preprocessing to different parameters used for the ML models - Homework won't be graded without the report
6) Use k-fold cross validation to prove robustness - 1p

Structure of the Report:

Introduction: Problem definition, objectives, and dataset description
EDA: Insights from data exploration, including plots and correlations
Data Preprocessing: Techniques applied and justification of your choice
Handling Imbalance: Methods tested and their impact on results
ML Models: Model descriptions, parameters tested, and performance results
Conclusion: Summary of findings, best-performing model, and future ways of improving

Visualizations and Reproducibility:

Ensure all plots are clear, labeled, and easy to interpret.
Use tables to summarize metrics, feature importance, and preprocessing effects.
Ensure that results are reproducible by others using your methodology.