DATA MINING ASSIGNMENT (2026-2027)

Scenario and objective

A wearable sensor company is developing a system for recognizing everyday and sports activities. As the data analysis team, you are asked to use real sensor measurements to identify the user's activity and discover groups of observations with similar characteristics. The aim is to understand the data, investigate their quality, and compare alternative choices based on measurable results.

Data and access

For this assignment, you will use the PAMAP2 dataset, which contains 3,850,505 records from 9 participants and 18 activities in total. The measurements come from three inertial sensors and one heart rate monitor. The file contains real-valued data.

Dataset page, download and documentation — UCI Machine Learning Repository

Use the files in the Protocol folder, covering the 12 specified activities.

Although the records are time series, the files are already in tabular form. For the basic approach, you may use each row as a separate observation, using the sensor values at that particular point in time.

Each row has 54 fields: a timestamp, an activity label, and 52 sensor-related variables. Of these, 40 are measurement values and 12 describe orientation and may be ignored. Records with activityID = 0 represent transitional activities and are excluded from the study.

As an extension, you may convert the file into time windows and extract features such as the mean or standard deviation for each window.

Study the accompanying documentation files before processing the data.

TASKS

A) Preparation and initial analysis

A1. Loading and quality checks

- Create a procedure for loading the files. Record the names and types of the variables you will ultimately use, as well as the number of observations before and after each preprocessing step.
- Present statistics such as activity frequencies, coverage per participant, and the distribution of missing values per variable.
- Check for duplicate records, constant variables, and suspicious values.
- Investigate noise and outliers. Explain when a value is likely to be a measurement error and when it may reflect a valid, intense activity.

A2. Exploratory analysis

- Examine the distributions of key features overall and by activity. Use focused plots with clear axes, units, and a brief interpretation.
- Identify strongly correlated or redundant features. Present a readable correlation table or heatmap and analyze selected pairs.
- Propose the most important features for activity recognition. Justify the ranking using an appropriate method and discuss how correlations between features affect the interpretation.

A3. Preprocessing and dimensionality reduction

- Design a procedure for handling missing values, different scales, and problematic measurements. Record the alternatives you considered and justify your final choices.
- Compare the original valid representation with at least one reduced-dimensional representation, e.g. PCA. Justify the method and the number of dimensions. Evaluate the new representation in both classification and clustering.
- You may use subsets for computationally demanding analyses. Describe the sampling method and examine how representative the subsets are in terms of activities and data composition.

B) Categorization/classification

B1. Experimental design

Define and justify your data splitting strategy, parameter selection, and final evaluation procedure. Report the size and composition of the sets you use. Choose a consistent procedure to ensure meaningful model comparisons.

B2. Models and metrics

- Implement a baseline that predicts the most frequent activity and a simple classifier with basic preprocessing.
- Compare at least five different classifiers. Select different families and justify their usefulness.
- Explore classifier hyperparameter selection within a reasonable computational budget. Report the values you considered and how you selected the final configuration.
- Use macro-F1 as the primary metric. Also present balanced accuracy, precision and recall per activity, and a confusion matrix for the final model. Include training and prediction times.

B3. What improved the solution?

- For at least one classifier, compare the original and reduced representations in terms of performance and time (report the number of features).
- Construct a learning curve using three training data sizes, for example 10%, 30%, and 100% of your defined training set. Describe the sampling procedure and compare performance and time using a common evaluation set.
- Identify and analyze at least two common class confusions. Suggest possible explanations and discuss performance on the less frequent classes.

C) Clustering and interpretation

C1. Representations and algorithms

- Cluster the observations using the input features without taking the activity label into account. Choose an appropriate scale and similarity or distance measure, and justify your choices.
- Compare at least three different approaches from different algorithm families (e.g. centroid-based, density-based, hierarchical clustering).
- Use both the original and the reduced representation.

C2. Configuration selection and evaluation

- Explore different numbers of clusters or other key parameters. The number of clusters does not have to match the number of activities.
- Select configurations based on internal metrics (e.g. silhouette score), and check stability with at least three runs or resampling iterations. State any sampling used. Explain how you compare the results.
- After selecting the configurations, compare the clusters with the known activities: calculate the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), and present a cluster–activity table.
- When an algorithm labels observations as noise, report their percentage and explain how they are handled in the metrics.

C3. Interpreting the findings

Which activities coexist in the same clusters? Which are split across several clusters? Describe the profiles of selected clusters through their features. Examine whether the structure relates to differences between participants, movement intensity, or sensor location. Compare the findings with the classification errors and explain the effect of dimensionality reduction.

SUBMISSION

The submission deadline is 15/1/2027 on eclass.

Each team must submit a single package linking the questions, experiments, and conclusions. The package must cover Sections A, B, and C.

Section A deliverable

In the report: a summary of the working dataset, a data quality table, 4–6 focused plots, documentation of correlations and important features, a table of preprocessing decisions, and a description of the reduced representation. In the code: an executable loading and transformation procedure.

Section B deliverable

In the report: a description of the experimental protocol, a consolidated comparison table of classifiers and configurations, a learning curve, a confusion matrix, and error analysis. In the code: a training and evaluation procedure and a CSV results file.

Section C deliverable

In the report: a table of algorithms, representations, parameters, metrics, and times; stability documentation; a cluster–activity table; 1–2 visualizations; and interpretation of at least three clusters. In the code: reproducible runs and export of the corresponding results to CSV.

Final submission package

Deliverable	Contents
report.pdf	An 8–12-page technical report, excluding the cover page, references, and optional appendices. Separate Sections A, B, and C, conclusions, limitations, and team member contributions.
src/ or notebooks/	Executable code for loading, preprocessing, EDA, classification, and clustering. Notebooks must include the main outputs.
README.md and environment	Instructions for downloading/placing the data and running the code, library versions, random seeds, hardware requirements, and estimated runtime. requirements.txt or environment.yml.
results/	CSV experiment tables, key plots, and configurations that link each result to its corresponding run.

Submit one compressed archive per team. Do not include the original data; provide clear access instructions.

Final presentation

A presentation of 5–7 slides is recommended, focusing on decisions and findings. All members participate in preparation and presentation. Each member must be able to explain the team's overall approach.

Suggested content

- Problem, data, and research questions
- Key EDA findings and preprocessing decisions
- Classification, comparisons, and error analysis
- Clustering, evaluation, and interpretation
- What helped, key limitations, and final conclusion

Prefer readable plots and comparisons to lengthy code excerpts. Be ready for brief questions about your choices and results.

Assessment

The quality of documentation and interpretation is a key part of the assessment. Grades are not based solely on the highest performance. A well-supported comparison, an insightful failure analysis, or a negative result can make a meaningful contribution. Not every technique is required to improve the metrics; you are required to examine and explain its effect.

GRADING BREAKDOWN

Criterion	Weight	What is assessed
Preprocessing and EDA	20%	Quality of checks, correlations, important features, and justification of decisions.
Dimensionality reduction and sampling	15%	Meaningful comparisons of representations and data sizes, performance, and cost.
Classification	25%	Experimental design, baselines, different families, metrics, and errors.
Clustering	25%	Comparison of methods, configuration selection, stability, and cluster interpretation.
Reproducibility and report	5%	Executable code, complete instructions, and a clear link between results and conclusions.
Oral presentation	10%	Clarity, selection of findings, time management, participation and understanding of the solution.

The two assignments with the best documentation will receive a 10% bonus.