NOVA IMS INFORMATION MANAGEMENT SCHOOL

SEPTEMBER 2025

“Why are they not here today?”

A Clustering-Based Exploration of Workplace Absence Patterns

Group Project
Data Mining I 2025/26

I. “Why are they not here today?”

Data mining is a powerful tool for discovering patterns and structures that help organizations make informed decisions. In this project, you will apply clustering techniques to analyze employee absenteeism data. This hands-on experience will reinforce your skills in data preprocessing, unsupervised learning, and interpretation, while providing you with the opportunity to see how data-driven insights can address practical challenges in the workplace.

A courier company in Brazil has recently grown concerned about rising absenteeism among its employees. Absences not only affect productivity but also impact team collaboration and overall morale. To better understand the situation, the Human Resources department has collected detailed, anonymized data on employee demographics, work conditions, and absence records. They have now turned to you, the data mining experts, to help uncover meaningful patterns hidden in this information.

II. PROJECT OBJECTIVES

The primary components and goals for the project are:

- **Exploratory Data Analysis (EDA) & Preprocessing:** The first step is to **explore and preprocess the dataset** fully. You are expected to carefully examine the data, addressing missing values, outliers, and any other issues encountered, and prepare the features for clustering through appropriate transformations. During this stage, you should also extract meaningful insights from the dataset, such as distributions, correlations, and engineered variables, which will help you better understand the employees’ profiles and absenteeism behaviors before moving to clustering.

- **Worker Segmentation:** Once the data is ready, your task is to apply clustering techniques to segment employees into meaningful groups. You should experiment with different clustering approaches, evaluate the quality of your results, and provide an interpretation of the clusters. The ultimate aim is to **uncover whether deeper patterns exist** that explain differences in absenteeism and propose **practical solutions** for the findings, recommending strategies for each cluster.

- **Knowledge in Action:** Finally, your goal is to translate the knowledge gained from the analysis into **something actionable for the company**. This means going beyond describing the clusters and providing practical insights or tools that could guide Human Resources in addressing absenteeism. For example, you might create a dashboard (visualizations) that allows HR to easily explore and understand employee clusters, develop a predictive model that assigns new employees to clusters, or analyse and discuss the importance of the features to differentiate workers. This part of the project is **open-ended**, and you are encouraged to be creative and propose solutions that you believe would benefit the company.

III. PROJECT DATA

The file ***employees.csv*** contains information about employees and their absences.

Variable	**Description**	
ID	Unique identifier for each employee	ID
Reason for absence	Registered reason for absence	Nom
Month of absence	Month in which the absence occurred	Ord
Day of the week	Weekday of the absence (Monday - Friday)	Ord
Seasons	Season in which the absence occurred	Ord
Days since previous absence	Days since the employee's previous absence	Num
Transportation expense	Employee's monthly commuting expenses (in BRL)	Num
Distance from Residence to Work	Distance from employee's residence to workplace	Num
Estimated commute time	Estimation of the employee's commute duration	Num
Service time	Years the employee has worked at the company	Num
Years until retirement	Number of years until retirement eligibility	Num
Date of Birth	Employee's date of birth	Num
Disciplinary failure	Whether the employee's absence violated workplace policies	Bin?
Education	Highest level of education (1-*High School*, 2-*Graduate*, 3-*Postgraduate*, 4-*Master's or PhD*)	Ord
Number of children	Employee's number of children	Num/Ord
Social smoker	Whether the employee socially smokes	Bin
Social drinker	Whether the employee socially drinks	Bin
Number of pets	Employee's number of pets	Num/Ord

Weight	Employee’s weight (Kg) Num
Height	Employee’s height (cm) Num
Body mass index	Employee’s body mass index Num
Absenteeism time in hours	Registered duration of absence Num

IV. DELIVERABLES

- A **.ipynb notebook** (or zip of multiple notebooks) featuring all the code you used throughout the project to:
  a. Decide on your final solutions for the problem at hand.
  b. Obtain your final results (code that helped you make decisions, but does not directly contribute to reaching the final solution, should be included but commented).
  c. File naming format: **GroupXX_DMI_2526.ipynb**.
- A structured report that summarises the analytical processes and the main conclusions obtained, with a maximum of **15 pages** (excluding cover and abstract, but including annexes).
  File naming format: **GroupXX_DMI_2526_report.pdf**.

V. EVALUATION

Your final project will be graded out of **20v**, according to the following criteria:

Criteria	Percentage (%)	Max Grade (out of 20)
Notebook Explanation	10	2
Data Exploration	20	4
Data Preprocessing	20	4
Clustering	25	5
Knowledge in Action	15	3
Report Quality & Storytelling	10	2

- Students can submit all deliverables with a **maximum delay of 3 days**, incurring a penalty of **1 point per day**. Beyond these three days, submissions will not be accepted.
- Deliveries made before the deadline will receive a bonus of **0.15 points per day** of delivery in advance (up to a maximum of 1 point).

Your grade will reflect our assessment of the quality, correctness, clarity, and efficiency of your work. Below is a description of what is expected in each evaluation component:

- **Notebook Explanation:** We will assess the clarity and readability of your notebook. This includes whether the goals are clearly stated, the code is properly commented, the chosen techniques are appropriate, and the insights and conclusions are clearly highlighted. We will also consider how well you connect your findings to potential next steps.
- **Data Exploration**: You should provide a clear description of the dataset and extract meaningful insights relevant to the problem. Avoid unnecessary visualizations or elements that do not contribute to your analysis.
- **Data Preprocessing:** We will evaluate the completeness and correctness of your preprocessing pipeline. This includes how you handle missing values, outliers, feature encoding, normalization, and any other transformations necessary to prepare the data for clustering.
- **Clustering:** This component assesses how you approach the clustering task. It is divided into several parts:
    - Feature selection: Rationale for which variables were included or excluded.
    - Modeling approach: The strategy adopted and the algorithms explored.
    - Evaluation: Use of internal validation metrics and justification for your final choice of clusters.
    - Cluster analysis: Interpretation of the identified clusters.
    - Strategies and recommendations: Practical implications and suggestions for the company based on your findings.
- **Knowledge in Action:** This section evaluates your ability to go beyond the clustering and generate additional value. It is assessed across the following dimensions:
    - Formulation and adequacy: Clarity and relevance of the idea.
    - Difficulty: The level of challenge and innovation.
    - Correctness and efficiency: How well your solution is implemented.
    - Explanation and discussion: Quality of your interpretation and connection to the company's needs.
- **Report Quality and Storytelling:** A strong report should provide a clear understanding of the problem, the methodology, and the rationale behind your decisions, all linked to your main results and insights. When referencing figures or tables, highlight the key message they convey. This section also considers the overall quality of your introduction and conclusions, as well as the coherence and flow of your storytelling.

VI. OUTLINE

Your project report, written in English, should respect the following outline and format:

Abstract

Provide a small overview of your work (200 to 300 words): What is the context? What are your goals? What did you do? What were your main results, and what conclusions did you draw from them?

I. Introduction

- Overview of the project
- Main goals of the project

II. Data Exploration

- Description of the data received
- Usage of visualisations, Statistical and Data Mining methods to uncover valuable insights before clustering
- Key Insights from the data

III. Data Preprocessing

- Description of the data received
- Steps taken to clean the data (handling inconsistencies, outliers, missing values, etc.) and prepare it for further analysis
- Justification of steps taken.

IV. Clustering

- Description and justification of the clustering process
- Description and comparison of found clusters
- Discussion of possible strategies for each cluster

IV. Knowledge in Action

- Objectives for the section
- Description of the actions taken
- Results and discussion of main findings

V. Conclusion

- Summary of objectives and findings
- Do the findings match what you initially expected? How?
- Discussion of the limitations of your work
- Suggestions for possible work to follow up on your work

Report settings:

• Heading: Calibri, size 14 pt, in bold
• Text: Calibri, size 11 pt, line spacing 1.15 pt, and paragraph spacing of 6 pt

VII. FINAL NOTES

- The report will be the primary method of evaluating your work. When preparing it, remember that a reader should be able to understand your work without needing to check your notebook. We won't be able to consider any steps or results not mentioned in your report.
- Ensure your report is concise, focused and based on reliable sources. You should look to source information provided from peer-reviewed journals (thus, avoid citing Medium, TowardsDataScience and similar sources). Avoid irrelevant, unimportant, or redundant information. Don't provide theoretical explanations of topics covered in class.
- Your submitted notebook should include all the unneeded code you used to obtain your final solution, but it should also be commented.
- We will run your Jupyter Notebooks if we have any doubts. So, please, make sure we can run the notebook from start to finish in one go. Notebooks that do not fulfil this condition will be penalized.
- **The report and code will pass through a process of plagiarism and AI generation checking**.
- When determining the grade for your work, there will be a **comparative component** between it and the work presented by your peers.

Friendly Reminders:

- Do not include techniques/algorithms/steps you cannot explain in your report: we will ask about them in the defense.
- If a team member is not contributing to the project, you must report it at least one month before the submission date.
- **Attendance at the defense is mandatory for approval of the project**. The discussion has a group component and an individual component. Considering this, a student’s final grade can change during the defense depending on their performance, **without any limitations.**

Project Delivery: 19th December, 18h00