NOVA IMS
Information Management School

SEPTEMBER 2024

Data-Driven Pathways for Change

Group Project Data Mining I 2024/25

I. INTRODUCTION

Data Mining is a powerful tool for uncovering patterns and insights that can drive business strategies and improve everyday life. In this project, you'll apply various data mining techniques to extract valuable information from real-world scenarios. This hands-on experience will reinforce the concepts learned throughout the course and enhance your skills in data preparation, analysis, and interpretation.

II. PROJECT GOALS

Success in data mining extends beyond technical skills; it also requires sharp critical thinking and effective problem-solving abilities.

In this course, you have the option to choose one of two projects to apply these skills:

Project 1: Students Mental Health
Project 2: Shopping Mall Credit

The ultimate aim of both projects is to leverage data mining techniques to deeply analyze datasets, revealing patterns that can inform precise strategies for advancing businesses or improving societal outcomes.

You are expected to:

1. Explore the data (dataset description, correlations, identify missing values, etc.).
2. Pre-process the data (incoherences, feature engineering, outlier treatment, etc.).
3. Employ quality visualizations to effectively present and analyze the data.
4. Identify meaningful segments in the data and thoroughly profile each one.
5. Suggest practical applications for the findings and recommend strategies for each cluster.

Contributions based on self-study and creativity will be valued, comprising alternative ways, for example regarding outlier treatment and the fill of missing values that have not been covered in class and the use of different types of visualizations.

Note: Thoroughly assess your preprocessing pipeline, considering clustering approaches and the pros and cons of different decisions. The quality of your conclusions and reasoning throughout the process will be evaluated.

III. PROJECT 1: Students’ Mental Health

The dean of a technical university abroad has noticed that many students seem unhappy and are struggling with mental health challenges. Concerned about their well-being and academic performance, he has turned to you, the experts from the Mental Health Department, for help.
After collecting comprehensive, anonymized data on student mental health, ranging from stress levels and academic pressures to social connections, you are now ready to take action. Using data mining techniques, your goal is to analyze this data and uncover patterns that reveal potential strategies to improve the student’s mental health and grades.

III. 1. PROJECT OBJECTIVES

Your objective is to employ data mining techniques to thoroughly analyze the dataset and uncover patterns that can inform targeted strategies for enhancing both students' mental health and academic performance. By leveraging advanced visualizations and data mining methods, you are expected to generate valuable insights into the underlying issues affecting the students.

In addition to these general findings, you will also perform clustering analysis to identify distinct groups of students who share similar characteristics. This segmentation will allow for more nuanced and tailored interventions, providing specific recommendations to address the mental health challenges and academic needs of each group. Through this refined approach, you will be able to deliver data-driven strategies that not only improve well-being but also foster academic success across the campus.

III. 2. PROJECT DATA

Variable	**Description and Data spectrum (if necessary)**
id	Index
gender	Gender of the student
age	Age of the student
date of birth	Date of birth of the student
university	Name of the university

degree_level	Under- or Postgraduate (you’ve conducted data from both)
degree_major	Topic subject of the major
academic_year	[1,2,3,4]
grade	Average grade ([50-100])
residential_status	Student lives Off- or On-Campus
campus discrimination	Student experiences discrimination/bullying by other students ([0,1])
sports_engagement	Times of sports a week
average_sleep	Average sleeping hours a night
study_satisfaction	felt level of the variable ([0-100]) by the student
academic_workload	felt level of the variable ([0-100]) by the student
academic pressure	felt level of the variable ([0-100]) by the student
finacial_concers	felt level of the variable ([0-100]) by the student
social_relationships	felt level of the variable ([0-100]) by the student
depression	felt level of the variable ([0-100]) by the student
anxiety	felt level of the variable ([0-100]) by the student
isolation	felt level of the variable ([0-100]) by the student
future_insecurity	felt level of the variable ([0-100]) by the student
sleep	the student uses the variable as a stress resolving strategy ([0,1])

outdoor_activities	the student uses the variable as a stress resolving strategy ([0,1])
religious_activities	the student uses the variable as a stress resolving strategy ([0,1])
sports	the student uses the variable as a stress resolving strategy ([0,1])
consume_food	the student uses the variable as a stress resolving strategy ([0,1])
creative_activities	the student uses the variable as a stress resolving strategy ([0,1])
social_activities	the student uses the variable as a stress resolving strategy ([0,1])
online_entertainment	the student uses the variable as a stress resolving strategy ([0,1])

IV. PROJECT 2: Shopping Mall Credit

Shopping Mall XYZ offers in-house credit services to enhance the shopping experience, providing customers with the flexibility to make purchases more easily. To improve their offerings and better meet customer needs, the mall has tasked your team with finding ways to gain a deeper understanding of its diverse customer base.

The data provided, which includes variables related to demographics, financial status, and credit behavior, offers an opportunity to categorize XYZ's customers into distinct groups. This will empower the mall's management to better understand customer profiles, enabling the delivery of more personalized credit services and targeted promotional activities.

IV. 1. PROJECT OBJECTIVES

The objective of your team in this project is to identify key segments within XYZ’s customer base and gain deeper insights into the business and its customers. The final output will be a report that not only identifies these segments but also provides an initial draft of suggested business applications based on the findings, along with general strategies tailored to each cluster. Each segment should be described in a way that is clear and actionable for the company.

By meeting these objectives, the project will empower Shopping Mall XYZ to better serve its customers, increase credit service adoption, and enhance overall financial performance.

IV. 2. PROJECT DATA

Variable	Description and Data spectrum (if necessary)
Client_ID	Unique identifier of client
Gender	Gender of client
Owns_Car	The client owns a car
Owns_Realty	The client owns property
Number_of_Children	The number of children the client has
Annual_Income	Client's annual income
Income_Type	The type of income the client has
Education_Level	Last educational level completed
Marital_Status	Client's marital status
Housing_Type	Client's residence place
Days_Birth	Number of days since birthday, count backwards from the current day (0), -1 means yesterday.
Days_Employed	Number of days the client has been employed, count backwards from the current day (0). If positive, it means the person is currently unemployed.
Has_Mobile_Phone	Client registered a mobile phone
Has_Work_Phone	Client registered a work phone
Has_Personal_Phone	Client registered personal home phone
Has_Email	Client registered email address
Occupation_Type	Client's Occupation
Family_Size	Client's family size
Credit_Status	Status of client's credit - 0: 1-29 days past due; 1: 30-59 days past due; 2: 60-89 days overdue; 3: 90-119 days overdue; 4: 120-149 days overdue; 5: Overdue or bad debts, write-offs for more than 150 days; C: paid off that month; X: No loan for the month

V. OUTLINE

Your project deliverables (especially the report) should respect the following outline:

Abstract
A brief summary of your segmentation project (200 to 300 words): What is the context for the segmentation? What methods did you employ to segment the data? What are the main findings and insights, and what strategies did you derive from the results?

I. Introduction

- Overview of the project
- Main objectives of the project

II. 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.

III. Data Visualisations & Analytics

- Usage of visualisations, Statistical and Data Mining methods to uncover valuable insights before clustering

IV. Clustering

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

V. Conclusion

- Discussion of your main insights
- Discussion of limitations of your work (e.g. what could have been done differently)
- Suggestions for strategies based on your work

VI. DELIVERABLES

- A .ipynb notebook (or zip of multiple notebooks) with all the needed code implemented to obtain the results presented in the report (File naming format: **GroupXX_DM1_2425.ipynb**).
- A structured report that summarizes the analytical processes and the main conclusions obtained, with a maximum of 10 pages (excluding cover, abstract and annexes) (File naming format: **GroupXX_DM1_2425_Report.pdf**).
- A presentation (meant for 5 minutes) describing the process that led to your final segmentation, highlighting key findings and recommendations (File naming format: **GroupXX_DM1_2425_Presentation.pdf**).

VII. EVALUATION

Criteria	Description	Max Grade (out of 20)
Data Exploration	Visualizations, statistical and DM methods (besides clustering) to uncover valuable insights	3.5 v
Data Preprocessing	Dealing with Incoherences, Outliers, Missing Values, Scaling, etc.	3.5 v
Clustering	Clustering process, algorithms, description and analysis of segments, conclusions and strategies, etc.	5 v
Report	Organization, Clarity, Storytelling, etc.	2 v
Creativity and Self-Study	Creative approaches, other data mining and clustering techniques	2 v
Presentation and Defense	Appearance, timing, originality, argumentation and transmission of knowledge	4 v

A project that focuses only on the techniques and methodologies approached during the practical classes will have at most 18 values. The remaining 2 values are possible to achieve if contributions based on self-study and creativity are clearly explained in the report (topic of “Creativity and Self-Study”).

VIII. FINAL NOTES

- The report and code will pass through a process of plagiarism checking.
- Please don’t provide long theoretical explanations of topics covered in class in your report.
- We will run your Jupyter Notebooks if we have any doubts. Ensure that the notebook runs smoothly from start to finish in one go. Notebooks that do not fulfil this will be penalized.
- Attendance at the presentation is mandatory for approval of the project. The presentation and discussion have a group component and an individual component.