CST 8502 Final Project (Mandatory)

(Part 1 due: Nov 7, Part 2 Due: Nov 21, Presentations: Nov 24 – Dec 5)

The goal of this project is to apply the algorithms we learned this term on a real dataset. You must follow **CRISP-DM** to complete this project. You must prepare your data to perform:

1. Classification by Decision Tree (DT)
2. Clustering using kMeans & Outlier detection by clustering approach.
3. Outlier detection using LOF & distances method.

You must do this project using **RapidMiner or Python**. Any other tool or language is not acceptable. As a team, you must decide either RM or Python, but not both. If all the team members are not ready to use Python, then it is advisable to use RapidMiner.

Before starting the project report, complete your workload distribution table, which is attached along with the project instructions on Brightspace. You must mention the selected tool in the Workload distribution table.

Dataset List

Dallas Police Incidents	1, 2, 3, 4
Dallas 911 Calls Burglary	5, 6, 7
Austin Crash Report Data	8, 9, 10
Austin Crime Reports 2018	14
New York Motor Vehicle Collisions	15, 16
Montgomery Traffic violations	11, 12, 13

Groups are published in Brightspace. You can see your group number in the announcement and Content/Presentation Schedule. The report should be typed in OneDrive Word Doc and should be shared with me (thomasa@algonquincollege.com). I should be able to see the version-history.

First Step: Create a OneDrive Word Doc and share with me: thomasa@algonquincollege.com.

Part 1 – Business Understanding, Data Understanding & Data Preparation

This project should be done in 2 parts. As Part 1, you will be working on Business Understanding, Data Understanding and Data Preparation. You need to work on the given dataset (check the above table to find your dataset based on your project group number) and propose a data science project that can be done for the given data. For the classification task, you should frame a question that you want to answer by your analysis. This question should not be something that can be easily answered using Excel. You have 3 main tasks –classification by DT, clustering, and outlier detection. As part of the business understanding, you must provide a project plan that explains how you are going to compete the project – each task should be done by **one** student.

Detailed workload distribution table should be included in the project plan. You must then perform Data Understanding and Data Preparation phases of the project. Use Association Rule Mining and Correlation Matrix Analysis to identify relationships between attributes. These methods help to uncover hidden patterns and dependencies within the dataset, aiding in better feature selection and model performance. As you must use distance-based methods and tree-based methods, you must prepare data accordingly.

As we have seen in labs and assignments, there will be some general preparation steps (like removing duplicates, generating or assigning ID, setting the correct data types for the attributes based on the meaning of the attributes, setting the role for class attribute etc) and then there will be model-specific preparations (selection of attributes that can contribute to your task – attributes which are good for clustering may not be good for outlier detection, binning, scaling, type conversions, handling missing data etc.). If you have latitude and longitude columns, make sure to do a clustering only for those columns to create regions out of it. When you create bins, make sure to create less than 10 bins – 5 to 8 should be good enough in most cases.

A template for the final report along with the table of contents (page numbers should be updated) is also provided. The report should be written in a professional report style.
Font: Times New Roman size 12 with **1.5** line spacing, justified

If you have a lot of data from multiple years, you can consider data from the latest year (don't filter for 2025 as we don't have the data for the full year). Even after filtering, if you have a lot of data, you can apply sampling techniques (stratified) to get a sample of 10,000 instances. When you prepare data, make sure to have **at least** 10 relevant attributes (this is a minimum requirement, more relevant attributes lead to meaningful results). You can create new attributes too. For example, if you have a date-time column, you can create date, time, month, time of the day (morning, afternoon, evening, night, late night etc.), weekday/weekend, day etc. appropriately. Also, make sure that you don't have redundant data. **More attributes will give you better results.**

You must choose a title for your project that reflects its goal. Be creative and ensure the title effectively conveys the purpose of your analysis. The document should have a cover page (should include student names and numbers, project title etc.). You should follow CRISP-DM when you do this task. Every step of each phase must be documented in the project report.

Sample project: For example, if we have a crime dataset that has information about the victim (age, sex, race), offender (age, sex, race), time, location, whether the person died or not, number of people involved, number of officers involved, weapons involved etc., the question may "what are the contribution factors for a crime to end up in fatality". To answer this question, you may create a decision tree with these factors by considering the fatality column as the class. You may have some outliers in this dataset (victim is a child etc.) which you will detect using outlier detection methods. You will be able to cluster those instances too using clustering techniques. Similar crimes based on type, location, time, season etc. will be grouped together and dissimilar ones will be in different clusters.

Part 2 – Modeling & Evaluation

You must now apply different modeling techniques for outlier detection (LOF, distances), clustering (kMeans to cluster and find outliers from the clusters, elbow method to find best k), and classification (DT). Use cross validation for DT classification. All these steps must be reported in a professional style. Also, when you build some prediction model, give a detailed screenshot of the results. Describe the accuracy of your prediction by presenting confusion matrices, R^2 values, etc. Also, interpret your classification, clustering and outlier detection results (rules of DT, why those instances are clustered together – any patterns in clusters, the reason why those instances are outliers etc.) You must provide interpretation for at least a few clusters and at least a few outlier instances.

If outlier detection by distance takes a long time to finish, you can choose any other outlier detection approach available in RM. Either way, you must use 2 approaches and combine the results to get the common outliers.

Project Presentation

During the last 2 weeks of the term, you will be presenting your final project. The schedule for presentation is available under Content → Presentation Schedule. Using PowerPoint slides, give a **short 30-minute** presentation that summarizes the steps in your project. **Every student should talk for around 10 minutes; team with 2 members will get 20 minutes and solo project will get 10 minutes. If a student uses more than 10 minutes, you are using the time of other students and because of that, you will be penalized**. Briefly describe your dataset, the question that you answer by your analysis, various data understanding and preparation steps, etc. Describe your analysis and the results by mentioning the algorithms. Briefly explain whether the analysis confirmed or denied your expectations and explain any surprises that you found. Also include an analysis of the accuracy of your results and their importance along with an interpretation of your results (DT rules, why those instances are clustered together, the reason why those instances are outliers etc.) You should have these main titles (along with related slides) –Introduction, Business Understanding, Data Understanding, Data Preparation, Modeling, Discussion of results, and Conclusion.

General Expectation

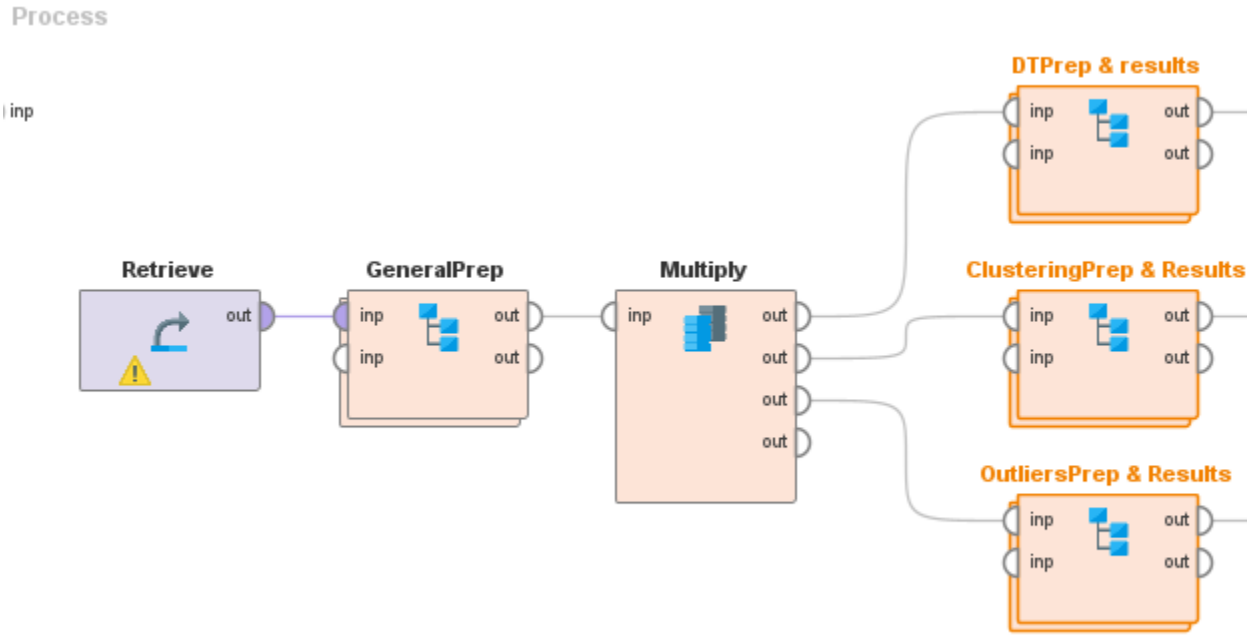
Each student's marks will be based primarily on their **individual contributions**, even though this is a group project. Every student must **independently complete all sub-steps** of the **Data Understanding** phase for one third of the columns and **perform all required steps** in the **Data Preparation** phase based on their chosen model. Attributes selected for clustering may not be suitable for outlier detection and vice versa. Choose your attributes based on your model. Students must **document** their entire process, including all steps, assumptions, approaches, challenges, solutions, and results in the report. Each student is responsible for writing their individual contributions in the report. Each student must perform their **own modeling**, **tuning**

parameters to achieve optimal performance, validate and evaluate their model & results. The final submission must include a **presentation PPT, a consolidated RMP file (or py file if the entire team choose to use python), and the final report**.

Every student will be evaluated for only one task from the given tasks - classification, clustering, outlier detection.

In the report, make sure to include the screenshot of corresponding subprocesses in the corresponding sections.

The template for the entire process MUST look as follows (if you are using python, make sure to follow the same approach).



Submission:

The deliverables should be from the perspective of providing a report and presenting it to a company or job interview where they aren't sure what data science is about (Just creating some tables and pictures is not enough).

Sections 1-3, 4.1, 5.1, and 6.1 should be completed and submitted (rmp or py & report – **DO NOT Zip files, zipped files will not be graded**) as Part 1 on Nov 7th, 2025.

Remaining sections should be completed as a continuation of Part 1 files and submitted as Part 2, including the RMP, report, and PPT files (**DO NOT zip files**) by Nov 21. This will be followed by a group presentation in week 11 & 12 (Nov 24-Dec 5). Please check Brightspace to see the presentation schedule.

To get grades, BOTH submission AND presentation are required. Successful completion of the project is mandatory to pass this course.