CST 8506 Final Project (Mandatory)

(Part 1 due: Jul 18,

Project Report due: Jul 28,
Presentations: July 30 – August 06

The goal of this final project is to apply the techniques we learned in this course to a real dataset. For the presentation, every project will get 20 minutes, if you are doing the project alone, you can take the full 20 minutes, otherwise, the time must be divided equally among team members.

Workload Distribution

Task	Subtask	Student Name
Introduction		
Business Understanding		
Data Understanding		
Preparation, **classification**, evaluation and discussion of results of both Image and feature dataset		
Preparation, **clustering**, evaluation and discussion of results of both Image and feature dataset		
Preparation, **outlier detection**, evaluation and discussion of results of both image and feature dataset		
Discussion of results		
Conclusion		

This table should be included in the project plan in Business understanding phase.

You should follow CRISP-DM for this work. You will be working either with Durum Wheat dataset (https://www.muratkoklu.com/datasets/.) or with the Skin Cancer MNIST: HAM10000 (https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000?resource=download). For both options, you have to work with the feature dataset and the image dataset.

The report should be typed in OneDrive Word Doc and should be shared with me (adhadur@algonquincollege.com). This should be the first activity that you do in the project. I should be able to see the version history. A template for the final report along with the table of contents (page numbers should be updated) is also provided. You must use this template to create your report.
You can see your project and the presentation schedule in the following table.

Group No	Student 1	Student 2	Student 3	Project Name	Presentation Date
1	Gauma, Priya (41129554)	Rani, Sonam (41126501)	Steffi Michael (41175645)	Durum Wheat Dataset	30-Jul
2	El Amine, Dalal (41181941)	Gu, Tianyi (40736589)	Tyagi, Shelly (41180437)	Skin Cancer MNIST	30-Jul
3	Rawat, Ayush (41082910)	Zhang, Golden (41051971)		Skin Cancer MNIST	30-Jul
4	Chowdhury, Md. Irteza (41126343)	Dave, Vasant D. (41154429)	Yosufi, Hasibullah (41012318)	Skin Cancer MNIST	06-Aug
5	Bouadjimi, Mohammed (41179141)	Dinh, Derron (41072352)	Wang, Judy (41177656)	Skin Cancer MNIST	06-Aug
6	Patel, Arpitkumar (41159097)	Savaliya, Mithil (41163550)	Wembonyama, Exalte (41165956)	Durum Wheat Dataset	06-Aug
7	Desai, Keyur M. (41090494)	Dua, Vishu (41169294)	Patel, Parth V. (41128658)	Durum Wheat Dataset	30-Jul
8	Mohammadian, Reihane (41129772)	Rahmouni, Yazid (41174752)	Saleh, Khaled (41174829)	Durum Wheat Dataset	06-Aug
9	Kumar, Romit (41174857)	Rungpholsatit, Thep (41066248)		Durum Wheat Dataset	30-Jul

Part 1

This project should be done in 2 parts. As Part 1, you will be working on Business Understanding, Data Understanding and Data Preparation.

Data Understanding
For the feature dataset, you should describe and explore data and verify its quality. You must include nice visualizations using Tableau/PowerBI. You should understand the corresponding image dataset too (number of images, dimensions etc).

For the image dataset, you should check the basic stats and explore images to get a basic idea of the dataset.

Data Preparation
Once you understand your data, you can select, clean, construct, integrate, and format data. Data has to be prepared for every task. As clustering aims to find patterns and groupings within the data while outlier detection looks for rare anomalous points that deviate from the norm, attributes which are good for clustering may not be good for outlier detection. Also, the effect of dimensionality reduction may also be different for these tasks. So, each student must prepare the dataset for their task. After data preparation, you must have at least 5 sets of data – original data, normalized data, standardized data, standardized dataset reduced by PCA, and standardized dataset reduced by LDA. You must prepare for distance based and tree-based approaches.

Apply imbalance technique SMOTE.

For the image datasets, you must have a CNN model to extract features (this model is just to extract features using convolution, pooling and dropouts). Other required prep steps should also be done.

Plot images (first 5 from each class, one class in one row) before and after convolution operations.

Part 2

Modeling for the feature dataset
Classification: Naïve Bayes, SVM, MLP, Stacking (with NB, SVM and MLP as base learners and logistic regression as a the stacking model)
Outlier detection: LOF, ISF and OCSVM
Clustering: kMeans, DBSCAN, EM.

Modeling for the image dataset
Classification: Naïve Bayes, SVM, MLP, CNN.
Outlier detection (must use features extracted by CNN model): LOF, ISF and OCSVM methods
Clustering (must use features extracted by CNN model): kMeans, DBSCAN, EM.

As you have repetitive tasks, you must write functions that takes the datasets, models and model parameters as parameters to avoid repetitions.

You must document all results.

Discussion of Results
Wherever possible, combine your models and results. Also, in detail, discuss your findings and results. When you report classification results, include accuracy, precision, recall and F1-score. For clustering of the feature dataset, find and discuss the best k. Also explain why those instances are clustered together (similarities). For outlier detection, explain why the outlier instances are flagged as outliers.
For image dataset, explain all results. For clustering, increase the number of clusters until you get a few small clusters (less than 10 instances). Plot the images in the small clusters. For outlier detection, plot the images of outlier instances flagged by all methods. When you plot images, plot the original images (not after convolution).

In this section, you must talk about the need and efficiency of traditional methods and deep learning methods on feature dataset and image datasets.

A project template is provided. You must use this template to complete your report. The report should have a cover page (with names of both students and student numbers), table of contents, tables, pictures, etc. and references.

It should be written in a professional report style.
Font: Times New Roman size 12 with 1.5 line spacing, justified

Project Presentation

Give a short 30-minute presentation (use PowerPoint slides, time should be divided equally among team members) which summarizes the steps in your report. Briefly describe your dataset, various data understanding and preparation steps, etc. Describe your analysis and the results by mentioning the algorithms. Also include an analysis of the accuracy of your results and their importance. Each student must talk about their task.
The presentations will be during the last two weeks of school, either in the lecture or in the lab. **This should be from the perspective of you providing a report 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).

Submission:

Part 1: You should submit your report with the sections Introduction, Business Understanding, Data Understanding and Preparation along with the Colab/Jupyter notebook. I assume that you use the file as is and do all changes in Python. You should not ZIP files, zipped files will not be graded.

Part 2: You should submit your final report and PowerPoint presentation along with the completed Colab/Jupyter notebook through Brightspace. Again, you should not ZIP files, zipped files will not be graded.

To get grades, BOTH submission (report & slides before due date) AND presentation as per the schedule are required. Successful completion of the project is mandatory for this course.

General Expectation

Each student's marks will be based primarily on their individual contributions, even though this is a group project. Every student must contribute in Data Understanding phase for one third of the columns and individually perform all required steps in the Data Preparation phase based on their chosen task. Attributes selected for classification may not be suitable for clustering or outlier detection and vice versa. Choose your attributes based on your task. 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. Every student will be evaluated for only one task from the given tasks - classification, clustering, or outlier detection.