CENG 464 – Data Mining – Fall 2024-2025

Course Project

Deadline: 23 December 2024

In this **Course Project**, you are expected to **apply the supervised machine learning method of classification to a dataset from the agricultural domain**. You will do this project as a **group of 2-3 students.**

Register your group here:

https://docs.google.com/spreadsheets/d/1zbKPHGkCXAK13iWK_t90J2cH2uNAFQbz/edit?usp=sharing&ouid=110008666492784186395&rtpof=true&sd=true

---

1. Data: Download the Course Project folder which has all the files, and locate the file

- Data_processed**.xls** is the file that you will be analyzing using classification.

- The Variables_Details.xls **file gives info about the features.** You will use these descriptions in your report to describe the important attributes that have high ranking.
- The folder also contains CSISA_KVK_Wheat_DoS_Trial_Data_Original.xls which is the original dataset
- **GrainYield attribute** is the class attribute with categorical values.
- All the other attributes are either already numerical attributes or need to be converted into numerical attributes within your Python code.
- Missing values need to be imputed.
- In the "Data" section of your report, you will be describing where the data came from and what it is about.
- The data source paper and repository should be mentioned in your report.

---

2. Analysis: [50 points] Analyze the data using the techniques you learned in the course, earlier courses, or other courses you took, with the following **requirements**:

The fundamental research question is:

- **"How do different classification algorithms perform on this data?"**
- **"What insights can be obtained from the data and its analysis?"**

You will modify it based on the focus of your analysis.

Specifications:

- **You must apply**
    1. **Imputation** (replacement of missing data values)**,**

2. **Data:** Each group member should perform comparison of classification results with respect to **two** of the categorical attributes (State, LandType, SoilType, SowingSchedule CropEstablishment or VarietyClass)
3. **Feature Selection** (ranking the importance of predictor attributes through different methods), Each group member should implement a feature selection method (that results in the best classification performance) & compare the results of the analysis with and without feature selection.
4. **Test and Evaluate** (5-fold cross validation with stratification, with at least 10 different classification algorithms, reporting AUC, CA, F1, Precision, Recall, MCC-Matthews Correlation coefficient),
5. **ROC** (for all 10+ algorithms in one chart), **and**
6. **Confusion Matrix** (only for the best algorithm)
   - **and report their results.**

- You **must do any improvement you can within your Python code,** and try different algorithms and Python libraries so that **you must have CA>0.80 with 5-fold cross validation with stratification for at least one of the classification methods/algorithms**.
- Your analysis and the framework that you apply should be rich enough so that **your report, after some minor improvements, can be submitted at least to a conference paper**, and preferably to a journal paper.
- You can use **any Python library/ IDE of your choice** to conduct the analyses.
- In your submission, all **data and analysis results** should be within a single MS Excel file. In other words, there should be a single Excel file constructed.
- Name your sheets with the same numbers as the Figure numbers of the screenshots. For example, if Figure 1 is screenshot of your first analysis, the Excel sheet corresponding to that should be named Figure 1. If some figures are referring to analysis in other software, you can skip them in the Excel file.
- Similarly, name the sheets that are related to tables also in the same number as the table. For example, the data for Table 1 should be in the Excel sheet named Table 1, corresponding to that table.
- Make sure to submit your complete Python code as a zipped file (.py files).

---

3. Project Report & Presentation: [40 points = 30 for Report + 10 for Presentation]

3a. Project Report [30 points] Write a report in MS Word, where you include all the components of a proper **research paper,** including **title, authors, affiliations, e-mails, Abstract, Keywords, Introduction, Literature, Method, Data, Analysis, Conclusion,** and **References**.

If you have work that you would like to place in an internally embedded or external Appendix file, or an external Supplement file, you can also do so.

Here are some further **requirements**:

- **You must apply the methods described earlier and report all your results.**
- Describe the data scraping/acquisition/harnessing and data cleaning processes.

- Describe the characteristics of the data, including dimensions (ex: number of rows and columns for tabular data, number of arcs and nodes for graph data, etc.), attribute names, descriptions, and types.
- Provide a flowchart of the data analytics framework/workflow that you developed/applied, preferably drawn using a graph analytics software (such as yEd desktop software from https://yworks.com) to obtain the most reader-friendly layout and visualization.
- Provide screenshot of all of the analysis performed, enough such that the insights can be seen within the figure itself (you still need to explain the insights within the text),
- Number your figures (Figure 1, Figure 2, etc.) and tables (Table 1, Table 2, etc.),
- Make sure to put **captions *below* each *figure***:
    - Example: "**Figure 1.** The workflow implemented in the project."
- Provide multiple screenshots of the same analysis (ex: Figure 1.a, Figure 1.b, etc) if there are multiple screenshots,
- Make sure to put **captions *above* each *table***:
    - Example: "**Table 1.** The data attributes, descriptions, data types."
- **Cite your figures and tables in your report body**:
    - Example: "Figure 1 shows workflow implemented in the project."
- Describe the interesting and actionable insights.

3b. Project Presentation [10 points]

Create a MS PowerPoint presentation, that consists of a maximum of 20 slides in total, including the title, contents, thanks slides. Make sure that your presentation contains the tables/figures for the methodology, and all the important results.

- You should follow the guidelines in the files in the folder
  - Creating Great Presentations
- And especially in the files
  - How to Create Great Presentations – Presentation**.pptx**
  - How to Create Great Presentations – Checklist**.docx**

---

4. Formatting [10 points]:

The following are the **requirements** for formatting the report and naming the files:

Formatting the Report

With respect to formatting, in your report, you should use the

- LNCS Paper Template**.docx**

file as the template **if you are targeting** the conferences whose proceedings are published with the **Lecture Notes series from Springer**. If you use this template, you can enable the macros if you prefer so, yet you should be able to write your report without any macros enabled.

For your convenience, for this case study, the template file has been edited further, so you can **start editing not the LNCS template, but instead the following file:**

- Course Project - Report - v00g**.docx**

Naming the Files

After completing the project report, name the report file in the format

- Groupid_GroupName - CENG464 Project - Report**.docx**

where GroupName is the inspiring name of your group (ex: Superstars, Falcons).

Similarly, the PowerPoint presentation file should be in the format

- Groupid_Group Name - CENG464 Project - Presentation**.pptx**

The data files in Excel would be named as

- Groupid_GroupName - CENG464 Project - Data**.xlsx**

Naming for the Python zip file should follow similar formatting, as follows:

- Groupid_GroupName - CENG464 Project - Python Files**.zip**

For example, for a group named “Falcons”, the files would be named as follows:

- 44_Falcons – CENG464 Project - Data**.xlsx**

- 44_Falcons – CENG464 Project - Report**.docx**
- 44_44_Falcons – CENG464 Project - Presentation**.pptx**
- 44_Falcons – CENG464 Project - Python Files**.zip**

---

5. Submission : You will then submit all these files.

- **VERY IMPORTANT: Please make sure to always send the source MS Word files, rather than the pdf printouts.**
- The email should have as attachments all the source files (all the mentioned **xlsx** and **docx** files, py files, as well as **any other files** created using other software).
- Put all files in a folder titled "Groupid_GroupName - Course Project", and **zip that folder to obtain a zip file.**
- Please upload all files as a **single zip file under WebOnline system**.