University at Buffalo
Department of Computer Science & Engineering
CSE 4/587 - Data Intensive Computing

Fall-2025

Project Phase II

Release Date: October 20, 2025

Submission Date: November 24, 2025, by 11:59 PM

1 Introduction

This document specifies the requirements for Phase II of the CSE 4/587 course project. In Phase I, you worked in groups to choose your dataset, used the provided cluster setup for data ingestion, performed local EDA, and formulated *N* ML problems suitable for your dataset and use case. You also defined your *2N* objectives that you aimed to pursue regarding those ML problems. In Phase II, you will continue working in the same group and with the same dataset. In this phase, you will implement your ML algorithms to find solutions for your defined ML problems and explore answers to your objectives. You will not use the Hadoop cluster from the Phase I; instead, you will use a local instance of Spark (PySpark).

2 Tasks

2.1 Data cleaning using PySpark

All data cleaning operations conducted in Phase I shall be implemented using PySpark. This task focuses on preparing the dataset for subsequent exploration and analytical tasks. The cleaning process includes handling missing values, detecting and treating outliers, converting data types, and performing data aggregation and grouping operations.

2.2 Exploratory data analysis (EDA) using PySpark

Perform exploratory data analysis (EDA) on the selected dataset using PySpark within a Jupyter Notebook or Python script. This task involves generating summary statistics and creating preliminary visualizations to understand data distributions, identify patterns, and detect anomalies, while developing skills in PySpark.

2.3 Machine Learning Algorithms using PySpark

For this task, you will implement different Machine Learning Algorithms to solve N problem statements defined in Phase I. This should be done using PySpark, and more specifically, you can use MLLib. You may include the following steps:

1. **Data Preprocessing:** Perform comprehensive data preprocessing by integrating data cleaning, missing value imputation, and feature extraction. Use the VectorAssembler to combine multiple feature columns into a single feature vector as part of a structured ML pipeline.
2. **Model Training and Evaluation:** Split the dataset into training and testing subsets. Employ cross-validation or a train-validation split to identify the best-performing model. Evaluate models using appropriate metrics — such as RMSE for regression tasks and accuracy or F1-score for classification tasks.
3. **Hyperparameter Tuning:** If applicable/suitable, implement grid search or random search techniques to optimize model hyperparameters and enhance overall model performance.

2.4 Report

This report should be comprehensive, covering both Phase I and Phase II. It should include the following components:

1. **Title of the Project:** Provide the complete title of your project along with the names of all group members.
2. **Data Cleaning:** Clearly describe the dataset used, including the total number of records and features considered. Explain how the data was analyzed, cleaned, and preprocessed — including handling of missing values, removal of outliers, and feature selection. Include visualizations that illustrate the data quality before and after cleaning.
3. **Exploratory Data Analysis (EDA):** Include at least *2N* meaningful EDA steps performed on the cleaned dataset. Use appropriate visualizations and statistical summaries to uncover trends, correlations, and patterns relevant to your problem statements.
4. **Data Analysis Objectives:** Clearly state the N problem statements defined in your project. Each problem should be well-defined, justified based on the dataset's characteristics, distinct from others, and provide meaningful insights for analysis.
5. **Problem Statements:** Clearly describe the *2N* goals corresponding to the defined problem statements. Each goal should be specific, measurable, and directly linked to your EDA findings.
6. **Machine Learning Models:** List and describe all ML models you implemented. Justify your choice of each model and provide details about training, testing, and hyperparameter tuning. Include plots showing model fitting (e.g., learning curves, confusion matrices, ROC curves) and summarize key performance metrics such as accuracy, precision, recall, and F1-score.
7. **Key Findings and Recommendations:** Summarize the main insights and conclusions derived from your analysis and model results. Discuss what the outcomes imply for the dataset or use case, and provide recommendations for potential improvements or future work. Essentially, you would detail how and what objectives have been achieved. You should also provide further comments and discussion for the shortcomings or any possible improvements.

Note:

1. All plots included in the report should be properly labeled and clearly visualized. All steps, metrics, and analyses should be thoroughly documented.
2. The report format must strictly follow the LaTeX template provided and should be prepared using LaTeX: `https://piazza.com/class_profile/get_resource/meq4mkiwgl91b1/mgi0rmoio5328h`
3. Any additional visualizations and supporting materials should be documented in the report. You should also provide details on the problems you have faced and how you have opted to solve them.
4. You may include a section representing each group member's contribution on a scale of 1 to 10.
5. All references must be included in the report by adding an additional section.

2.5 Presentation Video

Record a presentation video of approximately 4 minutes that presents the complete project. Each member of the group should appear and describe their role. This video presentation should include your chosen dataset, your defined ML problems, your defined objectives, the ML algorithms that you have implemented, and the insights you have achieved.
Upload the video to UBBox, make it public, and include the link in the report under the “Video Presentation” section. Make sure it is accessible through that link.

2.6 Demo Day (Only for CSE587 students)

Prepare a poster on your project and participate in presenting it at the CSE Demo Day (early December). Details will be provided later. This task applies only to CSE587 students. You can find poster template at `https://www.buffalo.edu/brand/resources-tools/ub-templates-and-tools/research-poster.html`

3 Grading Criterion

3.1 For CSE487

1. Data cleaning with PySpark	**[15 Points]**
2. EDA with PySpark	**[15 Points]**
3. ML problems implementation	**[30 Points]**
4. Project Report	**[30 Points]**
5. Video presentation	**[10 Points]**

3.2 For CSE587

1. Data cleaning with PySpark	**[10 Points]**
2. EDA with PySpark	**[10 Points]**
3. ML problems implementation	**[30 Points]**
4. Project Report	**[30 Points]**
5. Video presentation	**[10 Points]**
6. Poster presentation	**[10 Points]**

Submission Guidelines:

You should submit a single `.zip` file with the following:

```
│ <ubitname1_ubitname2_ubitname3>_phase2.ipynb
└─ <ubitname1_ubitname2_ubitname3>_demo.mp4
└─ <ubitname1_ubitname2_ubitname3>_poster.pdf
└─ <ubitname1_ubitname2_ubitname3>_report.pdf
```

1. The report should be typed in format using this template: `https://piazza.com/class_profile/get_resource/meq4mkiwgl91b1/mgi0rmoio5328h`
2. Source code for the complete Phase II (using PySpark). The code should include all tasks as outlined in the steps, with proper headings in the .ipynb notebook. The code must be executable.
3. Please submit only one .zip, and ensure that the folder is properly named as per instructions before compressing it. Submissions in any other format or structure will not be accepted.

Notes

1. Document all your work and include explanations, observations, and justifications for your choices, wherever applicable.
2. Mention all of the issues you have faced and your approaches to solve them.
3. A good report is a well documented report, detailing each step, providing meaningful discussions, visualizations and results.
4. All submissions must include proper references for any external resources used. Please note that citing a source does not permit directly copying and submitting it as your own work. Simply modifying minor changes to existing code may still be considered plagiarism. Your submission must reflect your original understanding and contribution. Kindly review the Academic Integrity statement for further guidance.
5. No late submissions are accepted, start working on the project early on and submit as per the guidelines.
6. Only one submission per group is required.

Academic Integrity

Academic integrity is a fundamental part of the learning process. As a student, it is your responsibility to complete all work honestly and in accordance with the expectations set by your instructor. The goal is to ensure that you genuinely engage with and learn the course content, in alignment with UB's academic integrity principles.

This is an group assignment. Submitting code, report content, or any other assets (such as models, logs, or graphs) that are not entirely your own constitutes a violation of the academic integrity policy.

Thank you for upholding your personal integrity and contributing to UB's tradition of academic excellence.

For more information, please visit: `https://www.buffalo.edu/academic-integrity.html`