Data Mining Term Paper

Adel Ahmadi, Amirhossein Dashtban and Susan Kiani

March 2, 2025

1 Project Description

In this project, each team must select a **data mining-related topic** and implement it using the **BRFSS dataset**. The chosen topic must first be approved by the **teaching assistants** to ensure its relevance to the available data and course topics. The selected topic does not have to be new, but **innovative and creative topics** will receive higher scores. Additionally, the topic must be directly related to this dataset and the concepts covered in class.

The project must be done **in teams of four**. All team members must actively participate in the process. After selecting and implementing the topic, teams must analyze their results and present them in the form of a **scientific paper**. This report must have a structured and organized format, including sections such as introduction, methodology, results analysis, and conclusion.

In addition to submitting a written report, each team is required to present their project results in an **online presentation**. In this presen-

tation, all team members must participate and present different parts of the work.

2 Dataset Introduction

The **Behavioral Risk Factor Surveillance System (BRFSS)** dataset contains survey data, focusing on behavioral risk factors and public health status. This dataset includes variables such as health status, education level, income, medical conditions, and other demographic information.

The BRFSS dataset is one of the most reliable sources for health research, and numerous scientific papers have been published using it. You can utilize this dataset for various data mining analysis. The dataset download link will be shared in the group.

3 Guidelines for Selecting a Topic

To assist with topic selection, links to relevant scientific papers will be provided in the group so that teams can study them for ideas and inspiration. The choice of topic is entirely up to the team members, and each group can select a topic based on their interests and abilities. If guidance is needed, teams can consult the **teaching assistants**.

Existing research ideas can also be used. However, innovative and creative topics will receive extra credit. Therefore, teams that attempt to introduce new methods or different analyses will have a better chance of obtaining a higher score.

4 Grading Breakdown

The total score for the project is **4 points**, distributed as follows:

- **Code and Analysis Review (2 points)** – The technical correctness, efficiency, and robustness of the implemented code, as well as the depth and rigor of the data analysis, will be assessed.
- **Documentation and Scientific Paper (1 point)** – Students are required to write an **academic research paper** that adheres to academic standards. The paper must include essential sections such as:

  1. **Introduction** – Background information, problem statement, and motivation.
  2. **Literature Review (Optional)** – Discussion of previous studies related to the topic and identification of research gaps.
  3. **Methodology** – Explanation of the data preprocessing steps, models, and algorithms used.
  4. **Results** – Presentation of key findings, visualizations, and statistical analysis.
  5. **Discussion** – Interpretation of the results and comparison with previous studies.
  6. **Conclusion** – Summary of findings and potential future directions.

  The clarity, organization, and scientific rigor of the paper will be evaluated.

- **Online Presentation (1 point)** – The effectiveness, clarity, and organization of the online presentation will be reviewed. Each team must communicate their findings concisely and demonstrate a comprehensive understanding of their work in an **10-15-minute-online presentation**.

- **Bonus: Up to 1 Extra Point** – Teams that demonstrate an **exceptional presentation** and introduce **innovative analytical approaches or novel insights** may receive additional points.

Participation Requirements: All team members must actively contribute to both the project development and the online presentation to receive a score. Individual grades may vary within the same team based on each member's involvement in coding, data analysis, documentation, and presentation. Failure to participate in the development process or the virtual presentation may result in a lower individual score compared to other team members.

5 Final Guidelines

- **Project Deadline:** The deadline for this project is set to approximately **two weeks before the final exams** and **will not be extended**. Teams should plan their work accordingly.

- **Recommended Tools:** While not mandatory, it is highly recommended to use **Jupyter Notebook** for analysis and **LaTeX** for doc-

umentation to maintain a high standard of presentation and reproducibility.