Group Assignment

Methods For Responsible AI.

Contents

Assignment Overview 1

Dataset Description: AI in University Admissions 1

Report Guidelines 3

The report must contain the following: 3

Code Guidelines 4

Submission Guidelines 4

Rubric 5

Template 7

Assignment Overview

In this assignment, you will work with a given dataset to explore challenges in responsible AI. The primary objective is to identify AI-related risks in model development, evaluate the impact of irresponsible AI use, and propose strategies to mitigate these risks. You will document your findings in a clear and structured report while providing a well-documented code submission that adheres to the specified requirements.

Dataset Description: AI in University Admissions

The dataset provided is centered around predicting university admissions. The target variable is a Boolean, **ADMISSION**, which indicates whether a student is expected to perform well and should be admitted. The dataset includes features that provide insight into the student's academic performance, demographics, family background, and school characteristics.

Data Creation:

The **ADMISSION** variable is derived from data about previously admitted students and their performance at the university. This implies that the target variable is designed to predict students' potential for success based on a model that integrates academic and personal factors. The dataset was created using national test results from the final year of high school. It includes information about the students and their family situations. The use of these features attempts to capture a holistic view of the factors that influence academic success.

Features:

The features are categorized into academic performance, demographic information, family background, and school-related characteristics.

1. **Academic Performance**: These features represent students' academic results and form the core of evaluating their preparedness for university.
    - **NATURAL_SCIENCE_SCORE**: The student's score in natural sciences.
    - **READING_SCORE**: A measure of the student's reading proficiency.
    - **MATH_SCORE**: The score in mathematics.
    - **SOCIAL_SCIENCE_SCORE**: The score in social science subjects.
    - **ENGLISH_SCORE**: The score in English.
2. **Demographic Information**: These features provide personal characteristics about the student that may influence the prediction.
    - **GENDER**: The student's gender.
    - **AGE**: The age of the student when taking the exam.
3. **Family Background**: Socioeconomic factors that may influence a student's educational experience and opportunities.
    - **PARENTS_EDUCATION**: The highest education level of the student's parents.
    - **FAMILY_INCOME_LEVEL**: The financial situation of the student's family.
    - **PEOPLE_AT_HOME**: The number of people living in the student's household.
    - **FAMILY_HAS_CAR**: A Boolean feature indicating whether the family owns a car, which may be an indicator of wealth or mobility.
    - **HAS_INTERNET_ACCESS**: A Boolean indicating whether the student has access to the internet at home, which may reflect access to resources for education.
4. **School-Related Characteristics**: These features relate to the type and resources available at the student's school.
    - **SCHOOL_CALENDAR**: The schedule or structure of the school's academic calendar (e.g., Academic year starting in February, in August, other arrangements).
    - **PUBLIC_SCHOOL**: A Boolean indicating whether the student attended a public school (True) or a private one (False).
    - **BILINGUAL_SCHOOL**: A Boolean indicating whether the student attended a bilingual school.

Report Guidelines

Your report must be clear, concise, and adhere to the specified length limits. Any content beyond the limits will not be graded. All figures and tables must be readable, concise, and informative.

The report must be at most **two** pages, including all sections and figures, and **800 words** maximum. Anything beyond the limits will not be graded (except for the team contribution table).

The report must contain the following:

1. Team Contributions Table

Provide a table detailing the contributions of each group member to the different components of the project. Each member should have substantial involvement in at least one key area. All members must participate in the activities and adhere to the responsibilities as stated in the Group Assignments and Project Contract.

2. Identifying AI Challenges

Focus on two specific challenges: **explainability** and **fairness**. Define these challenges in the context of the dataset and models. For each, provide a concise explanation of how the challenge is relevant.

- **Explainability**: Discuss the difficulty of interpreting the decisions made by models. Consider why explainability is critical in ensuring users, stakeholders, and regulators can trust AI systems.
- **Fairness**: Analyze issues related to bias in the data or model, especially where predictions may unfairly disadvantage certain groups. Clearly define what fairness means in your context and why it is important to address.

3. Consequences of Irresponsible AI Usage

For each identified challenge clarify the potential negative consequences of irresponsible AI usage. Illustrate these consequences with examples drawn from your dataset or model outputs.

4. Mitigation Strategies

Propose appropriate mitigation strategies for explainability and fairness challenges. These strategies should be well-motivated and supported by relevant literature or theory.

6. Experimental Setup

Provide a concise description of your experimental setup that ensures reproducibility. Describe relevant aspects of the dataset, preprocessing steps, model architecture, training configuration, and evaluation metrics used to assess explainability and fairness.

Clearly state and motivate your choices for model parameters and hyperparameters, particularly in relation to explainability and fairness. For example, consider how choosing simpler models might trade off performance for better explainability, or how fairness constraints might affect model optimization.

7. Performance Comparison

Report and compare model performances before and after applying your mitigation strategies for explainability and fairness. Use tables to present these results clearly, highlighting how each strategy improved interpretability or reduced bias.

8. Figures and Tables

Include at least one relevant figure to effectively present results or analyses that complement the tables. Ensure the figure adds new insights without repeating information from the tables. All tables and figures should be clear and easy to interpret.

9. Discussion of Results

Discuss the results of your experiments, connecting them to the explainability and fairness challenges. For example, how did implementing explainability methods impact model performance, and how well did fairness metrics improve after mitigation strategies? Relate your findings to existing literature, highlighting both the strengths and limitations of your approach.

Weaknesses: Discuss the potential weaknesses of your mitigation strategies and suggest possible improvements. For instance, if improving fairness reduces model accuracy, what trade-offs must be considered? How could future work address these weaknesses?

10. Conclusions

Draw clear, concise conclusions from your findings. Summarize the key takeaways from your work on explainability and fairness, providing practical insights into how responsible AI can be implemented in real-world models.

Code Guidelines

1. **Code Format and Submission**:
   Submit the code through Canvas along with your report. Ensure that the code is well-organized and follows the project requirements.

2. **Code Functionality**:
   The code should run without modification and meet all project specifications, including implementing mitigation strategies for both explainability and fairness. For this, assume that the dataset is placed in the same folder as your code files.

3. **Code Coherence**:
   Ensure the code aligns with the methods described in your report. The implementations for explainability and fairness strategies should be reflected both in the code and in the report's discussions.

4. **Code Documentation**:
   Thoroughly comment your code to explain each step and decision. Comments should describe key functions, variable roles, and the rationale behind important choices, particularly in relation to explainability and fairness strategies.

Submission Guidelines

Your report and code must be submitted together in a single zip file through Canvas by the specified deadline.

Rubric

This is the rubric that will be used to grade the final report and code.

Component		Criteria	Possible points	Total per section
Report	Format	The format is correct: The length and sections are adequate, and the title includes group numbers, names, and student numbers. All the content is within the limit (anything beyond the limit will not be graded).	5	15
		The writing is adequate, and the figures and tables are clear and readable	5	
		A table specifying sufficient contributions of each member and component is provided	5	
	Problem definition	Valid, relevant, and clearly defined AI challenges are identified and concisely explained	5	10
		The consequences of irresponsible AI usage regarding the identified challenges are clarified with illustrative examples stemming from the data or models	5	
	Methods	An appropriate strategy is clearly defined for each identified challenge	5	20
		Each strategy is well-motivated and supported by relevant literature and theory	5	
		Choices of parameters and hyperparameters are clearly motivated and explicitly stated	5	
		The experimental setup is clearly and concisely described, guaranteeing reproducibility	5	
	Results	A comparison of model performances before and after applying the mitigation	5	20

		strategies is clearly reported for each strategy		
		The results of different mitigation strategies are clearly compared	5	
		Results are reported in clear, concise, and informative tables, which present the results of different strategies	5	
		At least one relevant and informative figure is effectively used to present results or analyses that complement the information in tables and text (not repeat)	5	
	Discussion and conclusion	The results and findings are discussed with clear connections to relevant literature and theory	5	15
		The discussion highlights insights from the results, connects with theory, and highlights the possible weaknesses of the proposed mitigation strategies and potential solutions	5	
		Clear and concise conclusions are drawn from the discussion, providing a clear takeaway for the reader	5	
Code	Submitted	The code has been submitted through Canvas along with the report and follows the correct format	5	20
	Working	The code runs flawlessly and meets all project requirements (it should run without modification by lecturers)	5	
	Adequate	The code is coherent with the methods and approaches proposed	5	
	Clarity	The code is thoroughly commented, making it easy to understand	5	
Total			**100**	

Template

Title

Group number

Member 1 (student number 1), Member 2 (student number 2),...

The report must be at most **two** pages, including all sections and figures, and **800 words** maximum. Anything beyond the limits will not be graded.

The only element that can be outside the two content pages is the Team Contributions Table.

The content must follow the report guidelines.

The report mut include:

- AI Challenges
- Consequences of Irresponsible AI Usage
- Mitigation Strategies
- Experimental Setup
- Performance Comparison
- Figures and Tables
- Discussion of Results
- Conclusions
- Team Contributions Table

Team Contributions Table: The table should follow the format below and, for each student, indicate whether they successfully participated in each of the three activities.

Student	Research and Analysis	Algorithm Development	Documentation and Reporting
Student 1 (student number)			
Student 2 (student number)			