Course Project: Data Science Competition on ChallengeData

Overview
In this course project, you will participate in a data science competition hosted on the ChallengeData platform. Your goal is to develop and optimize predictive models to achieve high performance on the given dataset:

- **AssurPrime : Saurez-vous prédire la prime d'assurance ?**
- **Identification de gaz toxiques**
- **Maladie de Parkinson : Prédire et Corriger les Biais dans l'Évaluation du Score Moteur**

Beyond achieving strong results, you will also analyze and compare different methodological approaches, produce a detailed written report, and defend your work through an oral presentation.

Objectives

- Gain hands-on experience applying data science methodologies to a real-world problem.
- Explore a variety of modeling techniques, comparing their advantages and drawbacks.
- Demonstrate critical thinking in model selection, hyperparameter tuning, and validation strategies.
- Communicate your process, insights, and results effectively, both in writing and orally.

Project Components & Grading

Your final grade will be based on three components:

1. **Submission Score (30%)**
    1. This is based on your final leaderboard ranking on ChallengeData.
    2. The platform will provide a specific metric to evaluate model performance.
    3. Your score component will be assigned according to your final standing at the competition deadline.
2. **Technical Report (40%)**
    1. Length: Approximately 10-15 pages (excluding appendices).
    2. The report should be professional, clear, and well-structured.
3. **Report Structure and Guidelines:**
    1. **Introduction & Problem Understanding**
        - Clearly define the problem and objectives.
        - Identify data-specific challenges (e.g., missing data, imbalance).
    2. **Data Exploration & Preprocessing**
        - Summarize key findings from exploratory data analysis (EDA).
        - Discuss data cleaning steps, feature engineering, and any transformations.
    3. **Methodological Approaches with Pros & Cons**

        - Present at least three different modeling approaches (e.g., linear models, tree-based methods, neural networks, ensembles).
        - For each approach, detail advantages, disadvantages, and reasons to consider or discard it.
    4. **Model Selection, Tuning & Validation**
        - Describe your strategy for hyperparameter tuning (e.g., grid search, random search, Bayesian optimization).
        - Explain your validation framework and why it ensures reliable performance estimates.
    5. **Final Chosen Solution & In-Depth Analysis**
        - Identify the best-performing approach and explain why you selected it.
        - Interpret the model's predictions (e.g., feature importance, SHAP values) to understand key drivers of performance.
        - Discuss what worked well and what could be improved.
    6. **Conclusion & Lessons Learned**
        - Summarize key insights, challenges encountered, and lessons for future projects.
4. **Report Assessment Criteria:**
    1. Clarity, organization, and coherence of writing.
    2. Depth of EDA and justification for preprocessing steps.
    3. Range and depth of approaches tested, including thoughtful discussion of their pros and cons.
    4. Sound reasoning for model selection and tuning methods.
    5. Quality of analysis and critical reflection on results.
5. **Oral Defense (30%)**
    1. You will present your work in a 15-20 minute session (10-15 minutes for the presentation and 5-10 minutes for Q&A).
6. **Presentation Guidelines:**
    1. **Content & Structure**
        - Give a clear overview of the problem, methods, and results.
        - Use visuals (charts, plots) to illustrate key findings.
    2. **Technical Depth & Understanding**
        - Show that you understand why you chose certain methods and not others.
        - Be prepared to explain your model's behavior, strengths, and weaknesses.
    3. **Responses to Questions**
        - Be ready to answer questions on your approaches, tuning strategies, interpretation methods, and potential improvements.

Project Timeline

- The project will take place between **15/12 and 15/01**.

Academic Integrity

- A bibliography must be included with your report.
- Teams should consist of 2 to 4 students maximum.

Support and Resources

- Practical sessions will be used to start working on your projects.