**Project proposal**

**Enhancing Steel Plate Defect Prediction with Advanced Feature Engineering**

## **Overview**

*The project aims to develop a sophisticated machine learning model to predict the probability of various defects on steel plates. Leveraging both the competition dataset and the original Steel Plates Faults dataset from UCI, the project will focus on extensive feature engineering using Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. This approach is expected to improve model performance by simplifying the data while retaining its essential characteristics. The final submission will be a CSV file with predicted probabilities for each defect category for each id in the test set, evaluated using the area under the ROC curve (AUC) for each category*

| **Milestone** | **Task** | **PACE Stage** |
| --- | --- | --- |
| Project Planning | Identify audience, problem, and impact | Plan |
| Data Exploration | Familiarize with datasets, EDA | Plan |
| Statistical Analysis | Formulate research question, hypotheses | Plan |
| Model Construction | Identify stakeholders, objectives | Plan |
| Data Integration | Assess data reliability, ethical considerations | Plan |
| Feature Engineering | Perform extensive feature engineering using PCA | Construct |
| Data Analysis | Evaluate dataset sufficiency, plan EDA | Analyze |
| Statistical Insights | Understand descriptive statistics, null/alternative hypotheses | Analyze |
| Model Evaluation | Evaluate model assumptions, ethical considerations | Analyze |
| Model Construction | Address identified problems, assess model fit | Construct |
| Data Visualization | Determine necessary visualizations, machine learning algorithms | Construct |
| Statistical Formulation | Formulate hypotheses, draw conclusions | Construct |
| Model Improvement | Evaluate model performance, consider improvements | Construct |
| Model Execution | Recommend initial investigations, share insights | Execute |
| Business Recommendations | Propose business recommendations based on insights | Execute |
| Ethical Considerations | Address ethical implications, model ethics | Execute |

**DATA PROJECT WORKFLOW**

**PACE: Plan Stage**

* **Project Planning:** Define the project's scope, objectives, and expected outcomes.
* **Data Exploration:** Familiarize with the datasets, conduct initial exploratory data analysis (EDA) to understand the data structure and distribution.
* **Statistical Analysis:** Formulate the research question and hypotheses, considering the impact of the project on business needs.
* **Model Construction:** Identify stakeholders, define the project's objectives, and outline the model's architecture.
* **Data Integration:** Assess the reliability of the data, consider ethical implications, and plan for data integration.

**PACE: Analyze Stage**

* **Data Analysis:** Evaluate the sufficiency of the datasets for the project goal, plan EDA to effectively achieve project goals.
* **Statistical Insights:** Understand descriptive statistics, formulate null and alternative hypotheses.
* **Model Evaluation:** Evaluate model assumptions, consider ethical implications

**PACE: Construct Stage**

* **Feature Engineering:** Perform extensive feature engineering using PCA to reduce the dimensionality of the dataset.
* **Model Construction:** Address identified problems, assess model fit, consider improvements.
* **Data Visualization:** Determine necessary visualizations, select machine learning algorithms.
* **Statistical Formulation:** Formulate hypotheses, draw conclusions.
* **Model Improvement:** Evaluate model performance, consider improvement.

**PACE: Execute Stage**

* **Model Execution:** Recommend initial investigations, share insights.
* **Business Recommendations:** Propose business recommendations based on insights.
* **Ethical Considerations:** Address ethical implications, consider model ethics.

**Timeline**

Start Date: March 1, 2024

Entry Deadline: March 31, 2024

Team Merger Deadline: March 31, 2024

Final Submission Deadline: March 31, 2024

**Evaluation**

Submission File Format: The submission file should contain a header and the predicted probabilities for each defect category for each id in the test set.

Evaluation Criteria: The final score is the average of the individual AUC of each predicted column.