Scope of Work for Group 4 | ISC4242

Andres Machado, Bret Geyer, Jackson Small, Jadan Colon, and Thomas Tibbetts

Project Statement:
This project will work with a dataset consisting of 4000+ observations, each representing a student at the Polytechnic Institute of Portalegre, Portugal. Each observation records a number of variables such as the student's previous qualifications, admission score, number of enrolled curricular units, and early grades.

We aim to use classification algorithms to predict a student's performance in their program, based on independent variables related to their background, academic path, and early performance. The target variable to be predicted is a label designating a student as:

- **Dropout**; Indicating that the student dropped out by the end of the program's normal term.
- **Enrolled**; Indicating that the student was still enrolled (not graduated) at the end of the normal term of the program.
- **Graduated**; Indicating the student graduated by the end of the normal term of the program.

In this project, it is critical for the prediction model to achieve high recall, especially on "Dropout" students, so that high-risk students can be reliably identified. On the other hand, the precision for each target class is also very important, to ensure that the students identified are truly at risk of failure. By identifying key indicators of a student's future performance, we can single out students who are likely to drop out or fail to graduate during the normal term of the program. These students can be targeted with advising, class balancing, or assistive technologies.

Outline:
The scope of this project is roughly divided into the following phases:
- Data cleaning to handle an excess of categorical features with many levels
- EDA to identify preliminary trends or significant features
- Training baseline models and further transforming the dataset
- Advanced model tuning (hyperparameter optimization, ensembling, etc…)

File Management:
We plan to use a GitHub repository for version control and file sharing.

Timeline: (VERY IMPORTANT)
- Milestone 2: Due **March 9th**
- Milestone 3: Due **April 6th** (That means 4 weeks to complete EDA)
- Final Submission: Due **April 27th**

EDA Timeline:

- Research each variable thoroughly and reduce the complexity of the categorical variables by binning small categories together: *Thomas Tibbetts* **[March 16th]**
  - Make note of which small categories were reduced to "Other" and which ones remain.
- Outlier Detection and Checking for Multicollinearity with Correlation Analysis (VIF or Correlation Matrix): *Andres Machado*. **[March 24th]**
  - For nominal qualitative variables, we can use Cramer's V as a quasi-measure of correlation.
  - For all variables (categorical and numerical), use VIF (Variance Inflation Factor) to test for collinearity.
  - This will thin out the unnecessary features.
- Data visualization: - *Jackson Small* **[April 2nd]**
  - Univariate Analysis: Histograms, Boxplots
  - Bivariate & Multivariate: scatterplots, correlation heatmaps, pairwise plots.
  - Identifying patterns and clustering through K-Means and Hierarchical (if possible)
- Train baseline models, preliminary data transformation until [**April 6th]** deadline
  - Classification Models:
    - Logistic Regression (*Thomas Tibbetts*)
    - KNN (*Jadan Colon*)
    - SVM (*Bret Geyer)*
    - Perceptron (*Jadan Colon*)
    - Naive Bayes (*Andres Machado*)