background of dataset

- This dataset provides many attributes about students as they are enrolled into various undergraduate degree programs published by Portuguese SATDAP program
- These attributes include demographic data such as age, nationality, marital status, etc.
- Socio-economic data such as mother and father occupation and educational background, as well as

scholarship and debt holders and macro-economic conditions

- Academic data, such as courses taken, application mode, attendance, curriculum units.

Dataset Analysis

- The dataset consists 35 columns and 4424 rows(records). No missing values.
- The Target column has three values: Graduate(2209), Dropout(1421), and Enrolled(794).
- All categorical feature column labels have been encoded for prediction, thus we are dealing with numerical features. (Feature Encoding)
- Density Distribution plot of all the features to understand the correlation and aid in feature selection using Pearson Correlation.

- **Correlation with the Target value.**
- **The number of “Enrolled” student is irrelevant for predicting whether a student will dropout or not. Thus, we plan to drop that and go forward with the “Graduate” and “Dropouts” values.**
- **Pie chart and density distribution plots of the 2 updated target values.**

Data Preparation

- We are going to prepare the dataset by finding individual relationships between each attribute and the target class
- We will investigate dimensionality reduction since we have irrelevant and redundant attributes.
- Not only can we eliminate irrelevant attributes, but we can merge trailing classifier values into one value for simpler analysis. For example, under "Marriage Status", the values "Widower", "Divorced", "Facto Union", and "Legally Separated" can be merge into one value of "Separated".
- We expect there to be many factors that go into if a student will dropout or not. The most likely causes may be socioeconomic status, followed by family history and personal predicament.

- From preliminary analysis, age, marital status, gender, course subject, parent's background, and money all show correlation to graduation status. Other factors like macro-economic indicators, nationality, and course mode appear to have little contribution.

Problem Statement: Predicting Dropouts from Student data

Approach:

1) **Supervised Learning: Since we have a labelled data (Graduate or Dropouts). We want to predict whether the student will tend to be a dropout based on historical data.**
2) **Standard scaling the features and making 80:20 Train-Test split.**
3) **Model Selection: Logistic Regression, Naïve Bayes, Random Forest , Support Vector Classifier,.**
4) **Training & Evaluation: After training all the above models, evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC will be used to assess model performance.**

5) **We are planning to use Minimum Absolute Error as an evaluation metric for KNN Classifier to identify the K value leading to the highest accuracy.**

6) **Finally, we will compare the accuracies of all the above-mentioned models.**

7) **We are expecting that Logistic Regression and SVM classifier results in the best prediction since they are the most adaptable and work well with high dimensional data and non-linear relationships.**

Resources

- **Link to data:** https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-ofstudent-retention/data

- Paper with classification labels in Appendix A: https://www.mdpi.com/2306-5729/7/11/146
- Alternate link to data with description and classification labels included: https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success