REPORT ON CLASSIFICATION RESULTS

The goal of this evaluation was to create a machine learning (ML) model that can predict student performance using data collected from Turkish students at the Faculty of Engineering and the Faculty of Educational Sciences in 2019. The report presents the results of the evaluation process, including exploratory data analysis, data preparation, classification study, grid search for the best estimators, and comparison between procedures. The performance of different models was assessed using various metrics such as accuracy, recall, precision, and area under the curve (AUC). Finally, the best classifiers for grade prediction were identified based on their performance on the test set.Exploratory Data Analysis

The initial phase of the evaluation involved exploring the collected data, which included personal, family-related, and education habits-related questions. Further analysis and pre-processing were performed to prepare the data for the ML models.

# Data Preparation

During the data preparation phase, error handling, data transformation, and k-fold cross-validation were applied to ensure the reliability and generalizability of the ML models. The metrics used for evaluation were precision, recall, and accuracy, which provide insights into the model's performance for each class individually as well as an overall view of the correctness of the predictions.

Precision and recall provide insights into the model's performance for each class individually, while accuracy gives an overall view of the model's correctness. Since this is a multi-class classification problem, precision and recall are calculated individually for each class and then averaged.

**Precision**: measures the proportion of correctly predicted grades out of all grades predicted as a specific grade. In this case, when predicting an AA grade what proportion of all predicted AA grades where truly AA grades. The procedure is repeated for each individual grade. High precision indicates that the model is good at correctly identifying a specific grade without misclassifying with the other grades. However, it doesn't consider the case when a grade was not predicted as the real grade.

**Recall**: measures the proportion of correctly predicted grades out of all actual grades in the set. In this case, when predicting an AA grade what proportion of all AA grades were predicted as AA grades. The procedure is repeated for each individual grade. High recall indicates that the models good at predicting most of the grades from each category to its real category.

**Accuracy**: measures the overall correctness of the model's predictions across all grades. It calculates the proportion of correctly predicted grades out of the total number of grades. It provides an overall assessment of the model's performance, considering both correct predictions for identifying the real and false grade category. However, it may not be the most informative metric when dealing with imbalanced datasets, where the number of instances in each class varies significantly.

# Classification Study

The classification study included the evaluation of various estimators with different parameters to identify the best possible model. Several classifiers were tested, including logistic regression, SGD classifier, MLP classifier, decision tree classifier, random forest classifier, extra tree classifier, adaboost classifier, Gaussian process classifier, gradient boosting classifier, and support vector classifier (SVC).

## Standard Estimators

The initial evaluation of the standard estimators showed varying levels of performance. The logistic regression model achieved an accuracy of 22.22% and had the highest precision and recall for the AA grade. The best-performing models in terms of accuracy were the decision tree classifier (33.33%) and the random forest classifier (33.33%).

## Standard Estimators & Feature Selection with Variance Threshold of 0.10

Applying feature selection with a variance threshold of 0.10 resulted in changes in the performance of the models. The logistic regression model showed an accuracy of 30.56%, making it the best-performing model. However, overall performance did not improve significantly compared to the previous evaluation.

## Standard Estimators & Feature Selection with Variance Threshold of 0.20

Further increasing the variance threshold to 0.20 did not lead to significant improvements in the model's performance. The logistic regression model remained the best-performing model with an accuracy of 22.22%.

## Grid Search for Best Estimators

A grid search was performed to identify the best estimators and their corresponding parameters. The models were evaluated based on accuracy, recall, precision, and AUC. The best-performing models were the SGD classifier, logistic regression with elastic net penalty, and MLP classifier.

## Comparison Between Procedures

A comparison between the different procedures revealed that the SGD classifier had the highest accuracy and AUC among the evaluated models.

# Test Set Performance

The performance of the best classifiers on the test set was evaluated. The SGD classifier achieved an accuracy of 44.83% and demonstrated reasonable recall and precision for most grades. The Extra Tree classifier achieved an accuracy of 34.48% but showed lower recall and precision values.

# Best Classifiers for Grades Prediction

The Logistic Regression with l1 penalization and the Random Forest classifier were identified as the best classifiers for grade prediction based on their performance on the test set. The logistic regression classifier demonstrated higher accuracy, recall, and precision compared to the rest of the studied estimators.

# Conclusion

In conclusion, the evaluation of various models reveals their performance on the classification task. The results demonstrate the impact of feature selection and hyperparameter optimization on model performance. The best-performing model, the Logistic Regression with l1 penalization, shows promising results in terms of accuracy, recall, precision, and AUC in comparison to the other classifiers.

Nonetheless, the performance of such model is still poor – given the fact that the tunning process is made for just one model applied to each grade leaving the rest out. If a model per grade is developed and fine-tuned better classification performances can be achieved.