1 Short Summary:

In our study, we developed machine learning models, including SVM, Random Forest, and Logistic Regression (with L1 and L2 regularization), to predict the likelihood of student academic dropout in higher education. Due to a high number of features and their inter-correlations, our models initially exhibited overfitting. To address this, we implemented feature selection techniques (PCA and feature importance analysis) along with model's parameter optimization. The refined models demonstrated improved performance, evidenced by a narrow gap between training and testing accuracy. Among the three, SVM marginally outperformed the others, achieving an accuracy of 80% and an AUC score of 0.89. Nonetheless, there is potential for further enhancement in model performance through additional feature engineering and more extensive parameter tuning.

2: How to run the code

Create the Conda environment using the provided YAML file:

Bash:

conda env create -f environment.yaml

In the src directory, you will find four Jupyter notebooks: data\_analysis\_final\_report.ipynb, data\_analysis\_model.ipynb, data\_analysis\_EDA.ipynb, and data\_analysis\_parameter\_optimization.ipynb. For a comprehensive view of the analysis, execute data\_analysis\_final\_report.ipynb, which integrates all individual parts. If you're interested in the specifics of each analytical segment, the other notebooks can be run separately to explore each in more detail.

3: Modify dependencies.

- Matplotlib >= 3.5.1

- Seaborn >= 0.11.2

- NumPy >= 1.22.3

- Pandas >= 1.4.2

- Altair >= 4.2.0

- scikit-learn >= 1.0.2

- SciPy >= 1.8.0

4. Reference:

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2. (Publisher link).

Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. DOI: 10.1023/A:1010933404324

Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), pp.2825–2830.