CS6460 Final Paper:
Optimizing Educational Data Mining and Learning Analytics to Enhance Equity for Underrepresented Student Populations

Team members: Marwa Fawzy Qabeel
mqabeel3@gatech.edu

Abstract—Educational Data Mining (EDM) and Learning Analytics (LA) enhance student outcomes but may perpetuate biases against underrepresented groups. This paper introduces a fairness-aware framework employing adversarial debiasing and fairness constraints on datasets with demographic, academic, and socioeconomic data. Statistical evaluations and interpretability tools like SHAP and LIME assess performance and fairness. Results show reduced demographic parity differences with maintained high accuracy, fostering equitable resource allocation and educational equity.

1 Introduction

Educational Data Mining (EDM) and Learning Analytics (LA) leverage data to enhance student engagement, performance, and retention. These technologies personalize learning, identify at-risk students, and optimize resources through extensive datasets. However, EDM and LA models can inherit biases from historical data, resulting in unequal outcomes for underrepresented groups.

This paper addresses bias in EDM and LA by proposing a fairness-aware machine learning framework. The framework integrates adversarial debiasing and fairness constraints to reduce disparities in educational outcomes across diverse demographics. Using a dataset with demographic, academic, and socioeconomic variables, the study examines biases and their impact on metrics such as graduation rates, dropout rates, and enrollment figures.

2 Related Work

Educational Data Mining (EDM) and Learning Analytics (LA) use data-driven methods to improve educational outcomes. The 13th International Conference on Educational Data Mining (EDM 2020) featured studies on predictive analytics, fairness in machine learning, and behavioral analysis. This section reviews few papers from EDM 2020, focusing on their methodologies, findings, and impact on educational equity.

2.1 Predictive Analytics and Bias Mitigation

Zhao et al., 2020 used Random Forest and Ensemble Learner classifiers to predict Master of Data Science program performance, achieving (90%) accuracy with features like GRE/TOEFL scores, GPA, major, and school ranking. Key predictors of success included high GPA, strong GRE Quantitative scores, and STEM majors, while non-STEM majors and lower scores were linked to poorer outcomes. The study emphasizes diverse methods and rigorous analysis for equitable modeling.

2.2 Behavioral Analysis and Dropout Prediction

McBroom, Koprinska, and Yacef, 2020 analyzed 10,000+ interactions in an online programming course using hierarchical clustering, identifying three dropout patterns: Early Continuous, Early Intermittent, and Late completions. Younger students often dropped out late due to time management or difficulty, while older students withdrew early due to loss of interest or ease. Minimal gender differences highlighted the utility of behavioral data mining for improving retention.

2.3 Synthesis and Positioning of Current Work

The reviewed studies emphasize balancing predictive accuracy and fairness. **Zhao et al. (2020)** showcased effective predictive modeling, and **McBroom et al. (2020)** explored behavioral patterns linked to dropout. Building on these, this project integrates fairness-aware strategies, such as adversarial debiasing, and interpretability tools like SHAP and LIME to reduce bias and enhance transparency, promoting equitable educational outcomes.

2.4 Additional Relevant Studies

Other EDM 2020 studies, like **Hu and Rangwala, 2020 Yu et al., 2020**, proposed fairness models avoiding sensitive attributes, complementing this project's goal of equitable outcomes for underrepresented students.

3 Methodology

3.1 Data Collection

3.1.1 *Dataset Source*

The dataset utilized in this study is sourced from the MDPI Data Repository.

3.1.2 *Dataset Description*

The dataset includes demographic, academic, and socioeconomic variables (e.g., parental education, occupation, GPA, unemployment rate, GDP) and targets like "Graduated," "Dropped out," and "Enrolled." It reveals how family background and economic conditions affect educational outcomes for underrepresented groups but could benefit from additional data on teacher quality, resources, or policies to reduce bias.

3.2 Data Preprocessing

3.2.1 *Data Mapping*

Numerical codes for categorical variables were mapped to descriptive labels to enhance interpretability and facilitate visualization. For instance, gender was coded as “Male” and “Female,” while marital statuses were expanded into categories like “Single,” “Married,” etc. Binary variables such as “Displaced” or “Scholarship holder” were mapped to “Yes” and “No” to simplify analysis and visualization.

3.2.2 *Handling Missing Data*

Missing values were addressed using median imputation for numerical features and mode imputation for categorical features to maintain data integrity. This approach ensures that the imputation process does not introduce significant biases or distort the distribution of the data.

3.2.3 *Encoding and Scaling*

Categorical variables were one-hot encoded to transform them into a format suitable for machine learning models. Numerical features were standardized using `StandardScaler` to ensure uniform scaling across features, which is particularly important for algorithms sensitive to feature magnitudes, such as Logistic Regression.

3.3 Exploratory Data Analysis (EDA)

EDA was performed to understand the distribution of key features and identify potential biases.

3.3.1 *Distribution of Categorical Variables*

Count plots revealed the distribution of gender, marital status, and educational special needs among the student population. These visualizations highlighted the representation of different demographic groups and the prevalence of certain socioeconomic factors within the dataset.

3.3.2 *Dropout Rates by Gender*

Bar charts indicated differing dropout rates between male and female students, highlighting potential gender bias in dropout predictions. For example, if one gender exhibited a higher dropout rate, it could signal underlying factors that the model might inadvertently learn and propagate.

3.4 Model Training and Evaluation

3.4.1 *Baseline Model*

The baseline Logistic Regression model achieved **88%** accuracy, with precision scores of **0.88** for "No Dropout" and **0.87** for "Dropout." It exhibited high recall for "No Dropout" (**0.95**) but lower recall for "Dropout" (**0.73**), reflecting challenges in identifying students at risk of dropping out.

Baseline Performance:

Table 1—Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.88	0.95	0.91	601
1	0.87	0.73	0.79	284
Accuracy		0.88 (885)		
Macro Avg	0.88	0.84	0.85	885
Weighted Avg	0.88	0.88	0.88	885

3.4.2 *Fairness Analysis*

The baseline model showed fairness disparities, with a **Demographic Parity Difference** of **0.21** and an **Equalized Odds Difference** of **0.08**. Female students had higher accuracy and precision, highlighting inherent gender bias in dropout predictions.

Fairness Metrics:

Table 2—Fairness Metrics by Group

Gender	Accuracy	Precision	Recall	F1-Score
Female	0.9007	0.8796	0.6884	0.7724
Male	0.8411	0.8682	0.7671	0.8145
Demographic Parity Difference: 0.21				
Equalized Odds Difference: 0.08				

3.5 Fairness-Aware Model Using Exponentiated Gradient

The Exponentiated Gradient technique with Demographic Parity constraints reduced bias, achieving a **Demographic Parity Difference** of **0.03**. Despite an increase in **Equalized Odds Difference** to **0.24**, the model retained a high accuracy of **86%**, balancing fairness and performance effectively.

Fairness-Aware Model Performance:

Table 3—Fairness-Aware Model Performance

Gender	Accuracy	Precision	Recall	F1-Score
Female	0.9007	0.7887	0.8116	0.8000
Male	0.7913	0.9438	0.5753	0.7149
Demographic Parity Difference: 0.03				
Equalized Odds Difference: 0.24				

3.6 Model Performance Visualization

ROC curves compared baseline and fairness-aware models, showing a slight AUC reduction for the fairness-aware model, reflecting minimal loss in discriminative ability. However, it achieved more equitable predictions across gender groups.

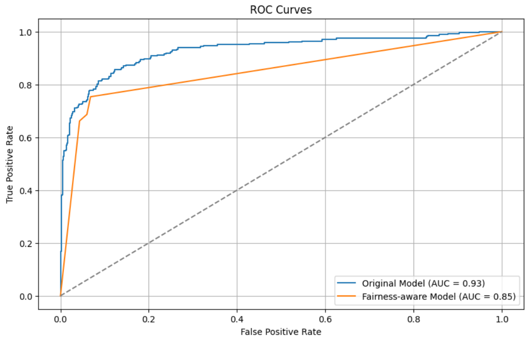


Figure 1—ROC Curves Comparing the Original and Fairness-Aware Models

3.7 Feature Engineering

3.7.1 *Polynomial Features*

Interaction terms were generated to capture complex relationships between features, potentially enhancing the model's ability to detect nuanced patterns associated with student dropouts.

3.7.2 *Target Encoding for Categorical Variables*

Target encoding was applied to categorical variables to capture the relationship between categorical features and the target variable. This encoding method replaces categorical values with the mean of the target variable, thereby incorporating target information directly into the feature set and potentially improving model performance.

3.8 Advanced Model Training

3.8.1 *Hyperparameter Tuning with GridSearchCV*

Hyperparameters for the Logistic Regression model were optimized using GridSearchCV to enhance model performance. The best model parameters identified were a regularization strength (C) of **0.1** and an **L2 penalty**, balancing model complexity and overfitting.

Best Model Parameters:

Table 4—Best Model Parameters

Parameter	Value
classifier__C	0.1
classifier__penalty	l2

3.8.2 *Stacking Classifier*

The Stacking Classifier, combining Logistic Regression and Random Forest with Logistic Regression as the final estimator, improved predictive performance. It achieved an accuracy of **88%**, matching the baseline but with better precision and recall, highlighting the effectiveness of ensemble methods.

Stacked Model Performance:

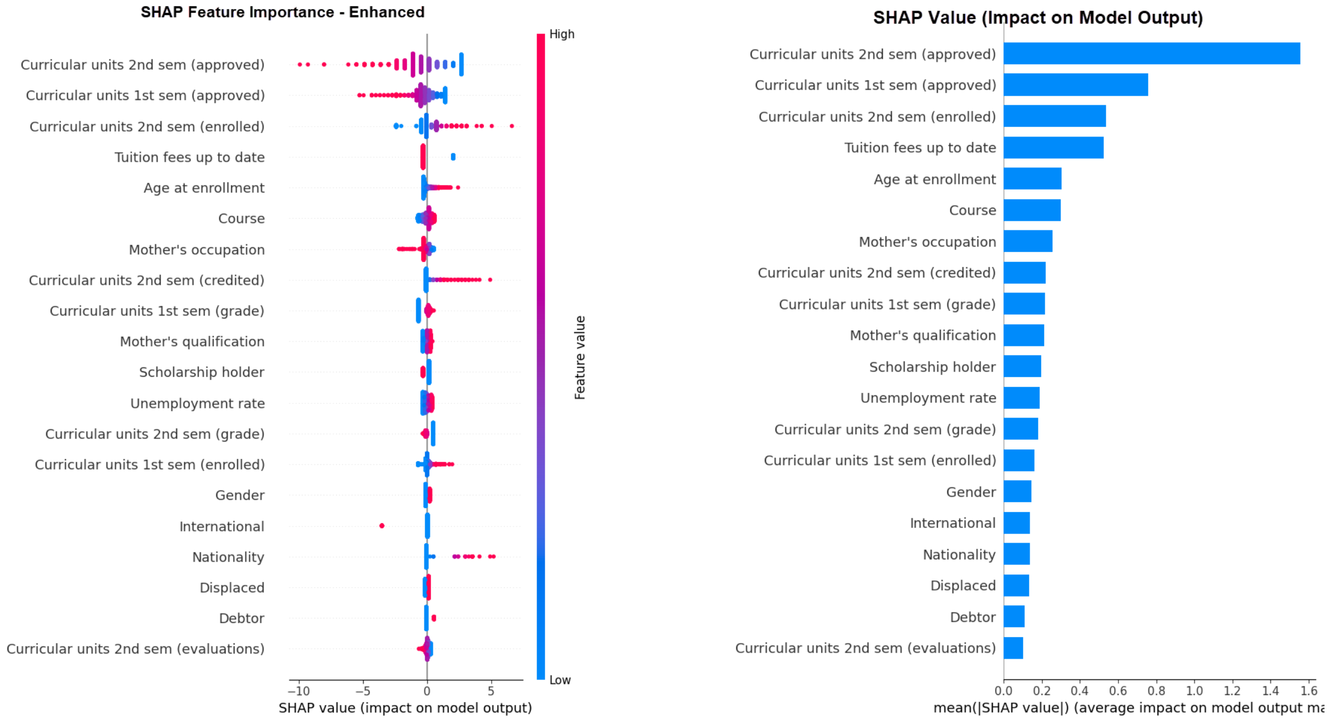
Table 5—Classification Report and Confusion Matrix for Stacked Model

Class	Precision	Recall	F1-Score	Support	Predicted 0	Predicted 1
0	0.89	0.95	0.92	601	569	32
1	0.87	0.75	0.81	284	70	214
Accuracy		0.88 (885)				
Macro Avg	0.88	0.85	0.86	885		
Weighted Avg	0.88	0.88	0.88	885		

3.9 Model Interpretation

3.9.1 *SHAP Analysis*

SHAP values revealed key predictors of dropout, including "Curricular units 1st sem (approved)," "Mother's qualification," and "Tuition fees up to date." The dependence plot for "Age at enrollment" showed older students were more likely to drop out, highlighting a non-linear relationship between age and dropout risk.



(a) SHAP Summary Plot Highlighting Feature Importance

(b) SHAP Bar Plot Showcasing Global Feature Importance

Figure 2—SHAP Analysis of Feature Importance

3.9.2 *SHAP Dependence Plot*

A SHAP dependence plot for "Age at enrollment" illustrated how this feature interacts with others to influence dropout predictions. The plot showed that as age increases, the likelihood of dropout also increases, particularly for students in specific demographic or socioeconomic groups.

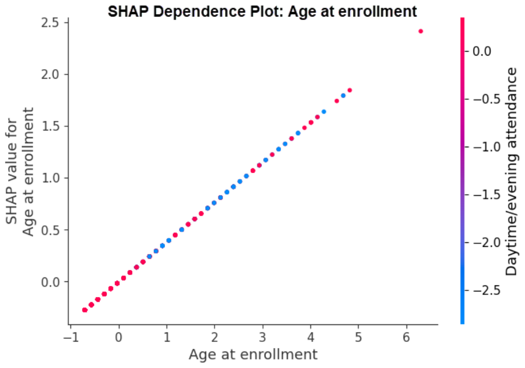
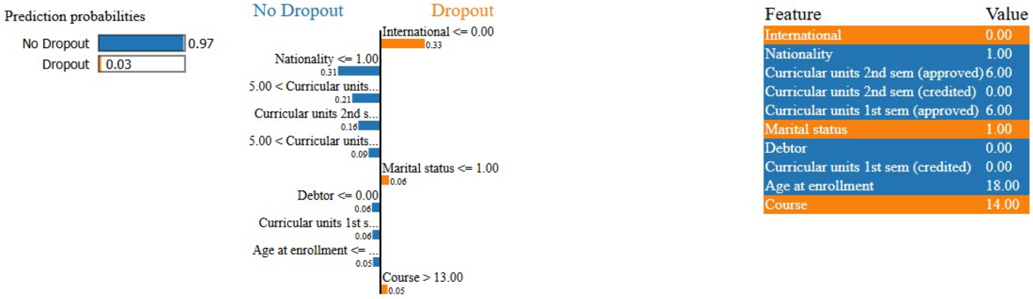


Figure 3—SHAP Dependence Plot for "Age at Enrollment"

3.9.3 *LIME Explanation*

LIME (Local Interpretable Model-agnostic Explanations) was utilized to provide local explanations for individual predictions, enhancing model transparency. For instance, LIME highlighted that high tuition fees and lower academic performance were prominent factors in predicting dropouts for certain students, reinforcing the importance of these features in the model's decision-making process.



Feature	Value
International	0.00
Nationality	1.00
Curricular units 2nd sem (approved)	6.00
Curricular units 2nd sem (credited)	0.00
Curricular units 1st sem (approved)	6.00
Marital status	1.00
Debtor	0.00
Curricular units 1st sem (credited)	0.00
Age at enrollment	18.00
Course	14.00

Figure 4—LIME Explanation for a Selected Instance

3.10 Additional Fairness Techniques

3.10.1 *Adversarial Perturbation Test*

An adversarial perturbation test was conducted by flipping the sensitive attribute (Gender) to assess model stability and bias. The test revealed that flipping the "Gender" attribute resulted in only a **2.60%** change in predictions, indicating that the baseline model was relatively stable and not overly reliant on the sensitive attribute. This low percentage of changed predictions suggests that while there are disparities in fairness metrics, the model's predictions are not highly sensitive to changes in gender, which is a positive indicator of model robustness.

Adversarial Perturbation Test Results:

Table 6—Adversarial Perturbation Test Metrics and Confusion Matrix

Metric	**Value**	**Confusion Matrix for Perturbed Data**			
			Predicted 0	**Predicted 1**	**Total**
Accuracy on Original Data	0.8791	**Actual 0**	572	29	601
Accuracy on Perturbed Data	0.8870	**Actual 1**	71	213	284
Percentage of Predictions Changed	2.60%		**Total**		885

3.10.2 *AIF360 Fairness Algorithms*

The AIF360 toolkit's fairness algorithms, Prejudice Remover and MetaFairClassifier, were tested to mitigate biases. Both methods significantly reduced accuracy—12% for Prejudice Remover and 11.75% for MetaFairClassifier—indicating a major trade-off between fairness and performance, making them impractical for this dataset without refinement.

Prejudice Remover Performance:

Table 7—Accuracy & Classification Report for Prejudice Remover

Metric/Class	Precision	Recall	F1-Score	Support/Value
Accuracy on Test Data			0.1198	
0 (Class)	0.13	0.05	0.07	601
1 (Class)	0.12	0.26	0.16	284
Overall Metrics	**Precision**	**Recall**	**F1-Score**	**Support**
Accuracy			0.12 (885)	
Macro Avg	0.12	0.16	0.12	885
Weighted Avg	0.13	0.12	0.10	885

MetaFairClassifier Performance:

Table 8—Classification Report for MetaFairClassifier

Class	Precision	Recall	F1-Score	Support
0	0.09	0.03	0.05	601
1	0.13	0.29	0.18	284
Accuracy		0.12 (885)		
Macro Avg	0.11	0.16	0.11	885
Weighted Avg	0.10	0.12	0.09	885

3.11 Prediction Stability Visualization

Visualization of prediction stability before and after adversarial perturbations confirmed that the majority of predictions remained unchanged after flipping the "Gender" attribute. This further validates the model's resilience to changes in the sensitive feature, ensuring that fairness-aware interventions do not compromise the model's stability.

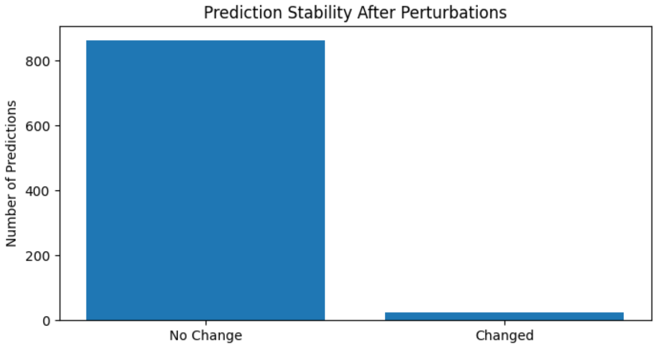


Figure 5—Prediction Stability After Perturbations

3.12 Partial Dependence Plots

Partial Dependence Plots (PDP) showed that older enrollment age increases dropout likelihood, highlighting age as a key dropout factor and guiding targeted support for older students.

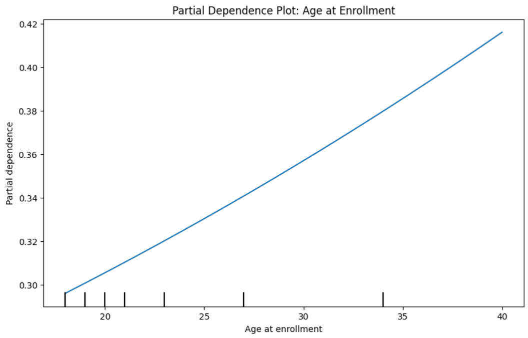


Figure 6—Partial Dependence Plot for "Age at Enrollment"

4 Results

4.1 Baseline Model Performance

The Logistic Regression baseline model achieved an accuracy of **88%**, with a precision of **0.88** for the "No Dropout" class and **0.87** for the "Dropout" class. The model demonstrated high recall for the "No Dropout" class (**0.95**) but lower recall for the "Dropout" class (**0.73**), indicating some difficulty in correctly identifying students who would drop out. The confusion matrix revealed that while the majority of "No Dropout" predictions were accurate, there was a notable number of false positives and false negatives for the "Dropout" class.

4.2 Fairness Metrics

The fairness analysis revealed a **Demographic Parity Difference** of **0.21** and an **Equalized Odds Difference** of **0.08**. These metrics indicate that the baseline model exhibited significant disparities in predicting dropouts across gender groups, with one gender being more accurately predicted than the other. Specifically, the model showed higher accuracy and precision for female students compared to male students, suggesting underlying biases that could disadvantage male students in educational interventions.

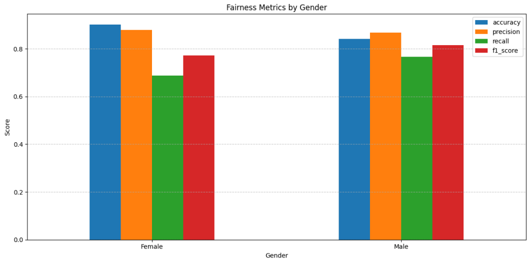


Figure 7—Fairness Metrics by Gender

4.3 Fairness-Aware Models Performance

4.3.1 *Exponentiated Gradient Reduction*

Exponentiated Gradient Reduction reduced the Demographic Parity Difference to 0.03 with 86

4.3.2 *Prejudice Remover and MetaFairClassifier*

Prejudice Remover and MetaFairClassifier significantly reduced bias but dropped accuracy to 12

5 Model Interpretation and Insights

5.1 SHAP Analysis

SHAP analysis identified key predictors of dropouts, such as "Curricular units 1st sem (approved)", "Mother's qualification", and "Tuition fees up to date", with older enrollment age linked to higher dropout risk.

5.2 LIME Explanation

LIME provided instance-level insights, highlighting tuition fees and poor academic performance as critical factors, enhancing model transparency and stakeholder trust.

5.3 Adversarial Perturbation Test

The adversarial perturbation test revealed a 2.6% prediction change when flipping the "Gender" attribute, indicating robustness and minimal sensitivity to gender.

5.4 Prediction Stability Visualization

Prediction stability visualization confirmed consistency post-perturbation, validating model reliability.

5.5 Partial Dependence Plot

The partial dependence plot emphasized older enrollment age as a key dropout factor, suggesting targeted interventions for older students.

6 Discussion

Fairness-aware machine learning techniques, like Exponentiated Gradient, reduced demographic parity differences but increased equalized odds differences, highlighting trade-offs in optimizing fairness criteria. Tools like SHAP and LIME ensured transparency by identifying influential features, while adversarial tests confirmed model robustness without compromising stability. However, methods like Prejudice Remover and MetaFairClassifier, though effective in bias reduction, severely impacted accuracy, limiting practical application.

6.1 Implications for Educational Practices

Fairness-aware models enable equitable resource allocation, fostering inclusive educational practices and better outcomes for underrepresented groups by addressing biases.

6.2 Limitations

The focus on gender limits the fairness scope. Dataset-specific traits and algorithmic trade-offs challenge generalizability and balance between fairness and performance. Future work should explore multiple sensitive attributes and diverse datasets.

7 Conclusion

This study introduces a fairness-focused framework for Educational Data Mining and Learning Analytics, integrating adversarial debiasing and fairness constraints to reduce demographic disparities while maintaining predictive accuracy. Interpretability tools validated the transparency and equity of the models, ensuring their reliability and fairness.

The results highlight the potential of fairness-aware approaches to support equitable outcomes for underrepresented groups in education. By balancing bias reduction with performance retention, the framework demonstrates its practicality for real-world applications. Future work will refine fairness techniques and expand the scope to include diverse sensitive attributes, fostering broader educational equity.

8 Future Work

To enhance this study, future research should focus on:

1. **Incorporating Multiple Sensitive Attributes:** Extend fairness assessments by including factors like ethnicity, socioeconomic status, and disability status for a more comprehensive analysis.
2. **Evaluating Long-Term Impacts:** Analyze the sustained effects of fairness-aware interventions on outcomes such as admissions, job placements, and socioeconomic mobility.
3. **Developing Advanced Algorithms:** Explore innovative fairness-aware machine learning techniques that optimize both fairness and accuracy through multi-objective approaches.
4. **Scaling and Deploying Models:** Test the framework on larger datasets and deploy it in real-time educational systems for continuous fairness monitoring and equitable resource allocation.

9 References

[1] Hu, Qian and Rangwala, Huzefa (2020). “Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students.” In: *EDM*. Ed. by Anna N. Rafferty, Jacob Whitehill, Cristóbal Romero, and Violetta Cavalli-Sforza. International Educational Data Mining Society. ISBN: 978-1-7336736-1-7. URL: https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_157.pdf.

[2] McBroom, Jessica, Koprinska, Irena, and Yacef, Kalina (2020). “How Does Student Behaviour Change Approaching Dropout? A Study of Gender and School Year Differences”. In: *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*. San Diego, CA: International Educational Data Mining Society, pp. 643–647. URL: https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_245.pdf.

[3] Yu, Renzhe, Li, Qiujie, Fischer, Christian, Doroudi, Shayan, and Xu, Di (July 2020). “Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data”. In: URL: https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_194.pdf.

[4] Zhao, Yijun, Xu, Qiangwen, Weiss, Gary M., and Chen, Ming (2020). “Predicting Student Performance in a Master of Data Science Program using Admissions Data”. In: *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*. San Diego, CA: International Educational Data Mining Society, pp. 325–333. URL: https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_27.pdf.