Student’s Dropout and Academic Success Prediction

Brief by Jiaming Huang

***Introduction* – Student’s Dropout and Academic Success Prediction** is critical, as higher education plays a vital role in employment, social equity, and economic development. In the era of big data, the rapid increase in demographic, socioeconomic, and academic data enables us to predict student dropout and academic success.

***Methodology* – EDA and Data Preprocessing:** First, the distribution of categorical and numeric variables is examined after checking NAs, leading to the decision to use standardization for normalization. Second, ‘target’ variable is encoded using a label encoder, and due to its severely imbalanced distribution, SMOTE is applied to balance the training dataset. Third, with columns ordered by demographic, economic, and academic variables, a heatmap reveals multicollinearity within groups, which is addressed by applying LASSO to eliminate redundant columns. **Model Construction:** Five machine learning algorithms, including eXtreme Gradient Boosting (XGB), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Naïve Bayes (NB), as well as four deep learning algorithms, including Long Short-Term Memory (LSTM), Residual Neural Network (ResNet), and Transformer, are used for model construction. To further improve performance, ensemble models are evaluated using the top five most accurate models. **SHAP Interpretation:** To explain the high accuracy achieved by the final most accurate ensemble model, its composite individual models are interpreted using SHapley Additive exPlanations (SHAP), providing insights into the key features. An overall workflow including algorithms is demonstrated in **Figure 1**.

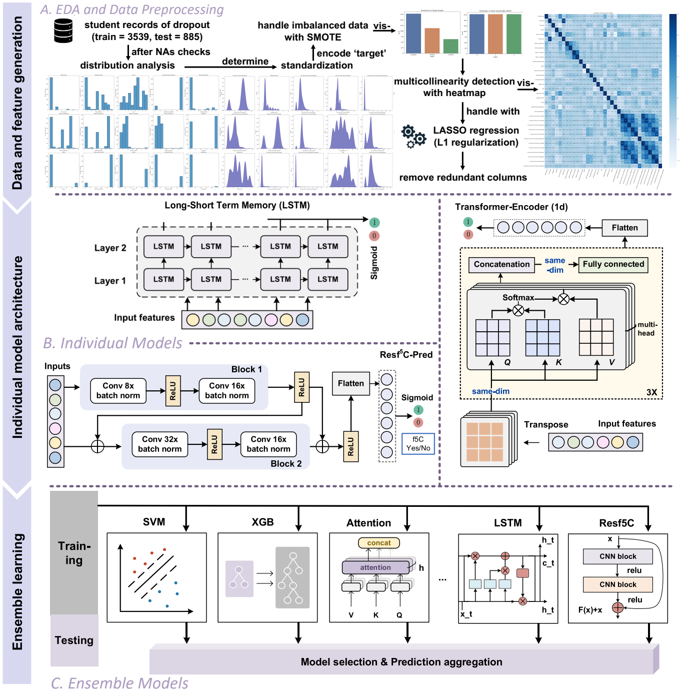

Figure 1: Overall workflow with algorithms and frameworks

***Results* – Model Construction:** Individual models’ performance, including Accuracy (ACC), Sensitivity (Sn), Specificity (Sp), F1 score, Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUROC) is presented in **Table 1.** From the table, XGB, SVM, RF, LSTM and Transformer are top 5 accurate models.

Table 1: Model performance of individual models in validation set

Model	ACC (%)	Sn (%)	Sp (%)	F1	MCC	AUC
XGB	**83.04**	82.76	91.44	0.8290	0.7468	0.9438
SVM	**82.47**	82.48	91.20	0.8254	0.7369	0.9375
RF	**85.57**	85.41	92.74	0.8551	0.7836	0.9574
LR	73.38	73.19	86.62	0.7336	0.6006	0.8882
NB	65.51	65.29	82.69	0.6506	0.4826	0.7819
CNN	80.29	80.25	89.63	0.8028	0.7012	0.9287
LSTM	**84.35**	84.54	91.66	0.8439	0.7513	0.9493
ResNet	80.70	80.59	90.31	0.8070	0.7208	0.9180
Transformer	**89.72**	89.68	94.32	0.8975	0.7923	0.9632

The performance is further enhanced through ensemble models, with the results presented in **Table 2**. From the table, the ensemble model of RF and Transformer has the highest accuracy of 92.82%. Applying this best model to the test set, we achieved a score of 0.95248 on Kaggle, outcompeting other competitors’ models.

Table 2: Model performance of ensemble models in validation set

Model	ACC (%)	Sn (%)	Sp (%)	F1	MCC	AUC
RF+Transformer	**92.82**	92.39	94.97	0.9292	0.7993	0.9890
RF+Transformer+XGB	87.82	87.39	93.97	0.8782	0.7893	0.9790
RF+Transformer+SVM	86.26	86.82	93.69	0.8626	0.7808	0.9696
RF+Transformer+XGB+SVM	82.38	82.98	91.25	0.8251	0.7375	0.9779
RF+Transformer+XGB+SVM+LSTM	84.25	84.78	92.20	0.8432	0.7651	0.9717

SHAP Explanation: SHAP offer feature contribution visualization for each composite individual model within the most accurate ensemble model. The result is shown in **Figure 2**. From the figure, several features, such as ‘Curricular units 2nd sem (grade)’ and ‘Curricular units 2nd sem (grade)’ shared across both models, indicating their importance in students’ dropout prediction.

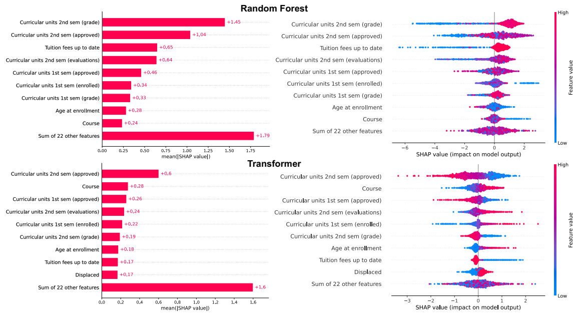

Figure 2: SHAP explanations of RF and Transformer

***Discussion* – Pro 1:** Final ensemble model incorporate “Transformer”, whose attention mechanism enable better focus on the input data. **Pro 2:** Achieves high accuracy with robustness. **Con 1:** Limited regional training data may lead to poor model generalizability. **Con 2:** Other SOTA models remain unevaluated and can be explored in future work.

***Conclusion* –** In this experiment, five machine learning and four deep learning algorithms were used for model construction. The top 5 models were then selected to build ensemble models, with the best achieving 92.82% accuracy.