

Assignment Report

Student Dropout and Academic Success Prediction using Machine Learning

Module Code:

Convenor Name:

Student Name:

Student Number:

Date: 20 Feb, 2024

Actual Hours Spent: 30+ hrs (I work in 2 hour sessions, and take 5 min break every 30 mins. I needed 14 such sessions, and some extra work in between.)

Abstract

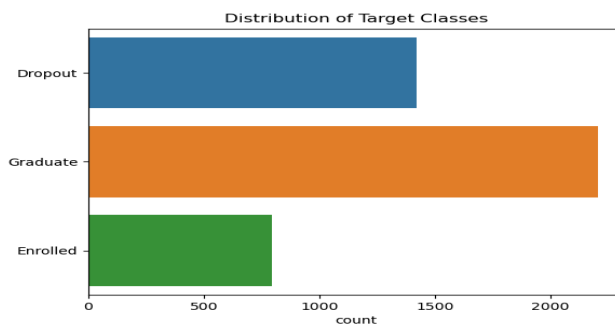
This report presents a predictive comprehensive analysis of student dropouts and academic success in higher education institutions via machine learning (ML). Utilizing a dataset from the UCI Machine Learning Repository, I applied three distinct machine learning models: Logistic Regression, Random Forest Classifier, and a Deep Learning model implemented with TensorFlow. The study aims to proactively identify students at risk of dropping out, allowing for timely intervention strategies.

Background and Problem Addressed

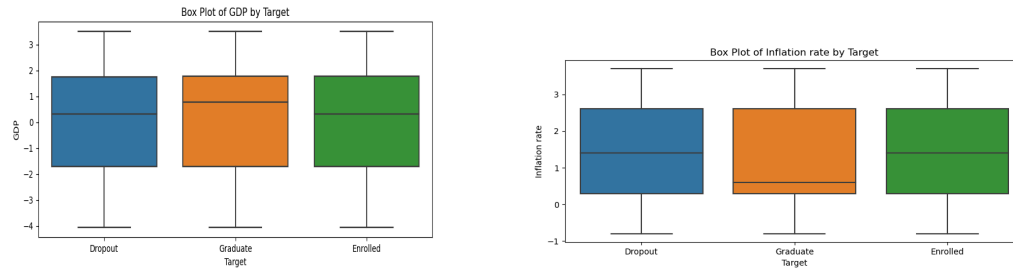
The challenge in the educational sector to predict student dropouts and successes is critical. Proactive identification of at-risk students can lead to timely interventions, improving educational outcomes. At risk students can be offered special lessons or support to improve their chances of getting Graduated. This study leverages machine learning technologies to address these challenges, contributing to the burgeoning field of educational data mining.

Exploratory Data Analysis

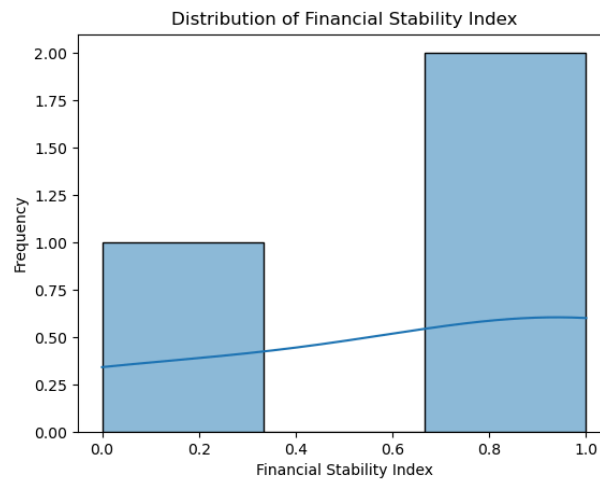
The dataset comprises 4424 instances with 36 features, including demographic, academic, and socio-economic factors. Initial explorations reveal a balanced representation of features like Marital Status, Gender, and others, with insights gained from distribution plots and correlation analyses.



This figure shows the imbalanced nature of our dataset, and as we can see, the Enrolled class has the least example, and our models have a harder time correctly finding this class.

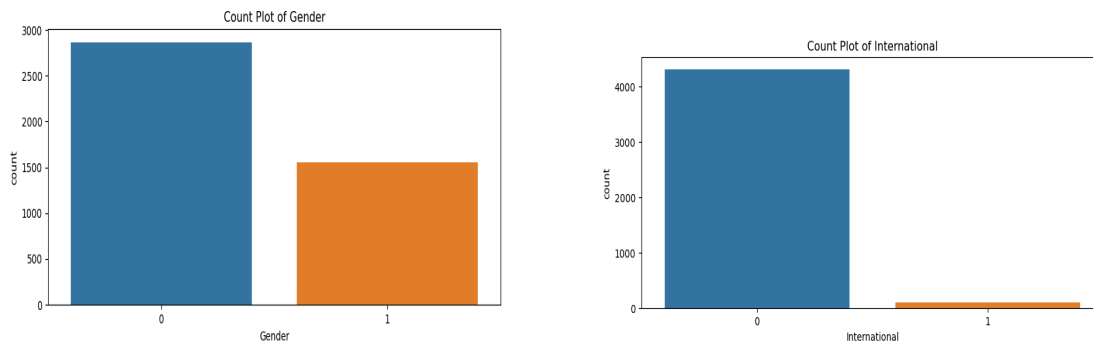


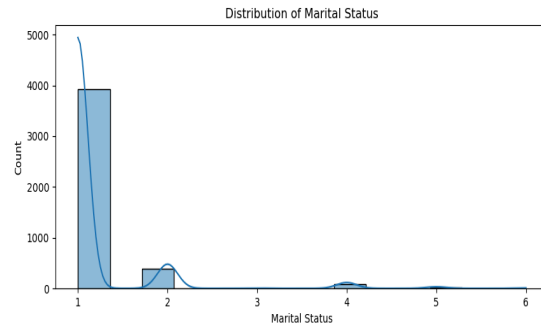
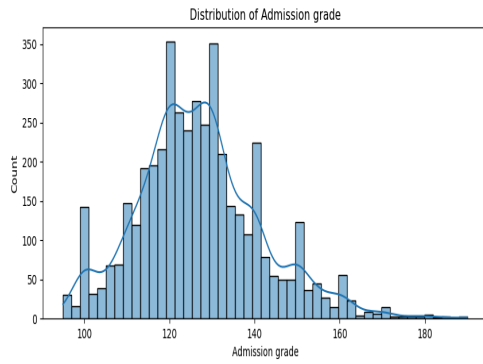
These figures show that current economic factors (GDP, Inflation) affect students' decisions. Then, we developed a feature that condenses these factors. We called it the Financial



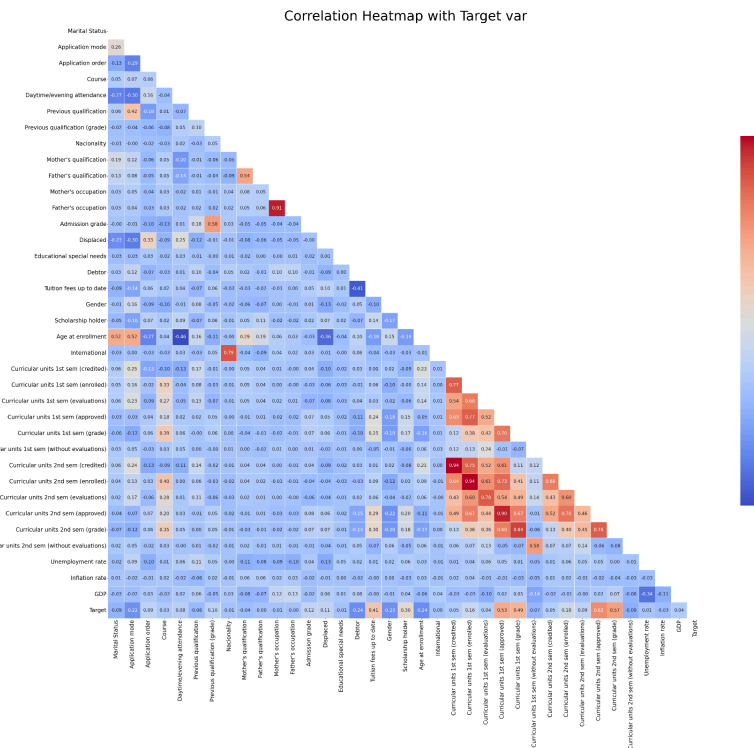
Stability Index.

These figures show the imbalance of gender, international/national students, marital status and the admission grades.





The Heatmap Below shows which factors are Correlated to the target.



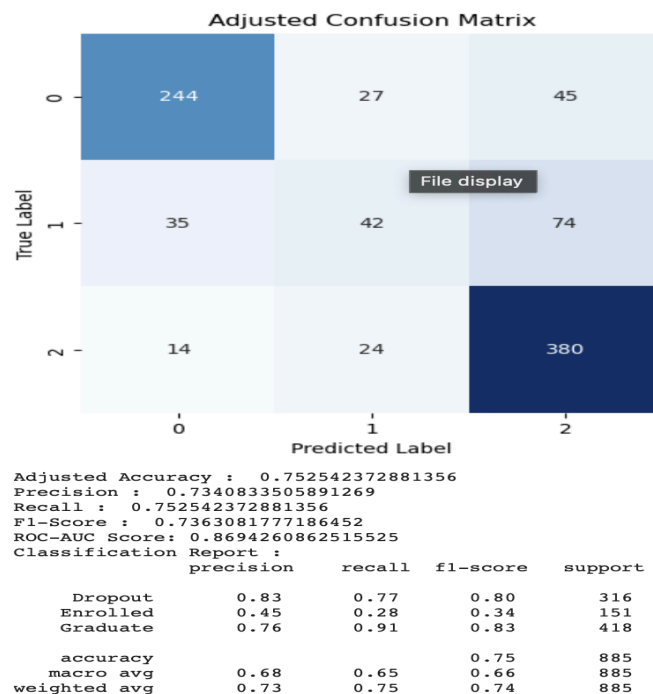
Data Pre-processing and Feature Selection

Data preprocessing involved normalization, checking for missing values (there were none in the dataset), and encoding categorical variables. Feature engineering enhanced the dataset, introducing new features like 'Approval Rate' and 'Financial Stability Index' while ensuring the relevance and impact of each feature on the model.

Machine learning models

Model 1: Logistic Regression

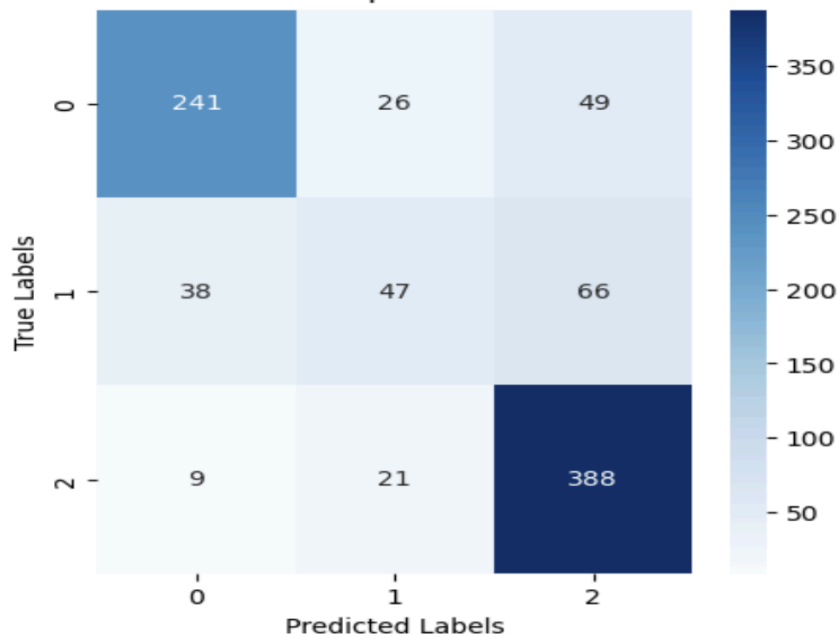
- **Summary:** Employed for its interpretability and efficiency, serving as a baseline model.
- **Parameters:** max_iter=500.
- **Rationale for using the Model:** Selected for its simplicity and interpretability.
- **Training and Evaluation:** Achieved an accuracy of 75.25%, with a weighted precision of 73%.
- **Analysis:** Exhibits balanced performance but slightly lower compared to other models.



Model 2: Random Forest Classifier

- **Summary:** Chosen for its robustness against overfitting and ability to handle non-linear relationships.
- **Parameters:** default parameters.
- **Rationale for using the Model:** Chosen for its ability to handle complex datasets and robustness.
- **Training and Evaluation:** Recorded the highest ROC-AUC score of 0.8714, indicating superior class differentiation.
- **Analysis:** Shows the best overall performance, slightly excelling in handling the 'Enrolled' category.

Confusion Matrix Heatmap for RandomForestClassifier



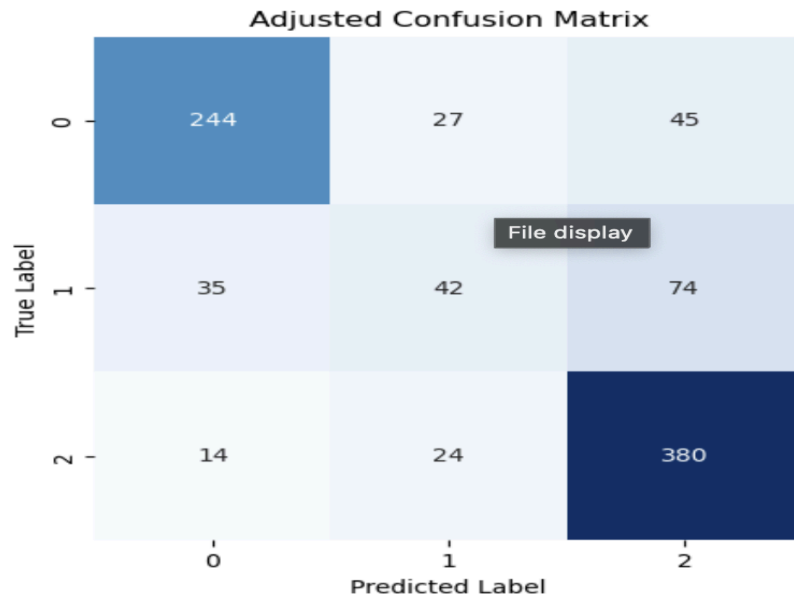
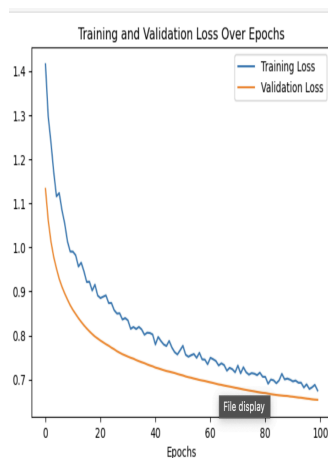
```
Random Forest Accuracy: 0.7638418079096045
Random Forest Precision: 0.7484338476360919
Random Forest Recall: 0.7638418079096045
Random Forest F1-Score: 0.7483592948498395
Random Forest ROC-AUC Score: 0.8712373920524056
Random Forest Classification Report:
      precision    recall  f1-score   support

Dropout      0.84      0.76      0.80       316
Enrolled     0.50      0.31      0.38       151
Graduate     0.77      0.93      0.84       418

 accuracy      0.76      0.76      0.76      885
 macro avg     0.70      0.67      0.67      885
 weighted avg  0.75      0.76      0.75      885
```

Model 3: Deep Learning with TensorFlow

- **Summary:** Implemented to capture complex relationships in data, offering scalability.
- **Parameters:** Sequential model with Dense layers, relu activation, L2 regularization, and dropout rate of 0.5, Learning rate = 0.0001
- **Rationale for using the Model:** Adopted for its capability to model complex relationships.
- **Training and Evaluation:** Showed comparable accuracy and precision to Random Forest, with a ROC-AUC of 0.8663.
- **Analysis:** Demonstrates potential in handling diverse data types but requires further tuning.



Adjusted Accuracy : 0.752542372881356
Precision : 0.7340833505891269
Recall : 0.752542372881356
F1-Score : 0.7363081777186452
ROC-AUC Score: 0.8694260862515525

Classification Report :				
	precision	recall	f1-score	support
Dropout	0.83	0.77	0.80	316
Enrolled	0.45	0.28	0.34	151
Graduate	0.76	0.91	0.83	418
accuracy			0.75	885
macro avg	0.68	0.65	0.66	885
weighted avg	0.73	0.75	0.74	885

Performance Measures and Evaluation Strategies

- **ROC-AUC Score:** Used to measure the models' ability to distinguish between classes.
- **Accuracy, Precision, Recall, and F1-Score:** These metrics provided a comprehensive view of model performance.
- **Confusion Matrix:** Offers insights into the classification accuracy across different categories.

Results Comparison and Analysis

To solve this classification Problem, 8 models were developed. The top 3 models were discussed in this study, these 3 models were even performing better on Unseen test data. The three models demonstrated close performance metrics. However, Random Forest outperformed others in ROC-AUC, accuracy, and other metrics. Deep Learning and Logistic Regression showed their respective strengths in specific areas.

- **Gradient Boosting Machines (GBM):** Known for their predictive power, GBMs performed well in terms of accuracy and ROC-AUC scores.
- **Support Vector Machines (SVM):** SVMs showed high accuracy levels, benefiting from their ability to find the optimal hyperplane for class separation.
- **Naive Bayes Classifier:** This model, with its simplicity and speed, demonstrated decent accuracy. Its performance on the ROC curve suggests its potential utility in certain scenarios within our dataset.
- **K-Nearest Neighbors (KNN):** KNN's performance was commendable in terms of accuracy. The model's simplicity and effectiveness in classification tasks were evident in our analysis.
- **Decision Trees:** These models were useful for their interpretability and ease of use. Their accuracy levels were competitive, offering a good baseline for understanding the dataset's structure.
- Also, the 3 models discussed. Among these,

Random Forest showed superior class differentiation.

Logistic Regression was effective but slightly lower in performance.

Deep Learning indicated potential but required further tuning.

Conclusion, Recommendations, and Future Work

Conclusion: The study highlights the effectiveness of machine learning in predicting student outcomes in educational settings, with each model presenting unique advantages. Random Forest Classifier emerging as the most effective overall. The insights gained from this study are invaluable for early identification and intervention for at-risk students, ultimately contributing to improved educational outcomes.

Recommendations:


- **Further Feature Engineering:** To enhance the models' ability to predict the 'Enrolled' category.
- **Model Tuning and Ensemble Methods:** Exploring hyperparameter tuning and ensemble methods for improved performance.
- **Continuous Monitoring:** Regular updates to the models as more data becomes available.

Future Work:

- **In-depth Analysis:** A more detailed examination of feature impacts and model behaviors.
- **Advanced Deep Learning Techniques:** Experimentation with various architectures and methods in the Deep Learning model.
- **Handling Class Imbalance:** Strategies to address imbalances in data, especially for underrepresented categories.

References

1. Leo, Breiman. (2001). Random Forests. 45(1):5-32. doi: 10.1023/A:1010933404324
2. Lenis, Wong. (2023). Model for the Prediction of Dropout in Higher Education in Peru applying Machine Learning Algorithms: Random Forest, Decision Tree, Neural Network and Support Vector Machine. 116-124. doi: 10.23919/FRUCT58615.2023.10143068

- 
3. Ioanna, Lykourantzou., Ioannis, Giannoukos., Vassilis, Nikolopoulos., George, Mpardis., Vassili, Loumos. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. Computer Education, 53(3):950-965. doi: 10.1016/J.COMPEDU.2009.05.010
 4. Tahira, Alam., Chowdhury, Farhan, Ahmed., Sabit, Anwar, Zahin., Muhammad, Asif, Hossain, Khan., Maliha, Tashfia, Islam. (2019). An Effective Recursive Technique for Multi-Class Classification and Regression for Imbalanced Data. IEEE Access, 7:127615-127630. doi: 10.1109/ACCESS.2019.2939755
 5. https://typeset.io/papers/a-new-multi-layer-classification-method-based-on-logistic-3ox9pvgmxb?utm_source=chatgpt

Appendices

- 1.
- 2.

