Balancing Act: Sensitive Data and Accuracy in University Dropout Prediction

Final Results and Project Report

Bryan Chi Fai Pang

Student ID: 501210081

TMU: The Chang School of Continuing Education

CIND 820 Big Data Analytics Project

Dr Ceni BABAOGLU

27 November, 2023

Introduction

University dropout has long been a matter of concern for educational institutions and policymakers worldwide. Dropping out not only hinders students' academic and career prospects but high dropout rates also have broader implications for society. With the emergence of machine learning and learning analytics, early intervention is now possible by identifying a student's likelihood to drop out. Nevertheless, the extensive use of personal data, particularly protected features like gender, age, and ethnic origin, has raised questions regarding algorithm fairness. While there is extensive research in university student dropout prediction, there is relatively little research done on the relationship between demographic, socioeconomic data and the performance of prediction model.

This project aims to explore the influence of including specific categories of student data – demographic, socioeconomic, macroeconomic and academic– on the performance of machine learning models in predicting university dropout.

To accomplish this goal, we initiate the process by establishing a baseline model that exclusively relies on macroeconomic and academic data. Following this, we proceed to create additional models that incorporate both demographic and socioeconomic data. Our research has found that using only student’s academic and macroeconomic data alone, we have developed a model that achieves 75% accuracy without using any demographic and socioeconomic data that can be regarded as sensitive and their uses should be restricted.

We anticipate that the findings of this research will make a valuable contribution to the ongoing discourse surrounding the balance between algorithm fairness and predictive accuracy in the context of university student dropout prediction.

Literature Review

I. Landscape of University Student Dropout Prediction in Machine Learning

[1] Lorenz Kemper et al., (2020) ***Predicting student dropout: A machine learning approach***

Kemper's study endeavors to illustrate that high predictive accuracy can be achieved through the utilization of logistic regression and classification. Notably, Kemper's focus lies in the efficacy of the model rather than algorithm fairness. His dataset and models employ just four basic demographic attributes – gender, origin, age, and date of enrollment – while deliberately omitting other students' socioeconomic data. The rest of the dataset is composed of attributes related to student academic career.

Kemper's observations offer valuable insights that are relevant to our study:

1. The four personal demographic features he employs are in alignment with the information routinely gathered by universities.
2. By intentionally reducing the reliance on demographic data, Kemper claims that his models sidestep the potential for pre-existing discrimination based on criteria unrelated to academic achievement.

Kemper's models yield remarkable results, achieving an 88% accuracy rate. Additionally, they accurately predict student dropout in 62% of cases, starting as early as the first semester. When considering all three semesters' marks, the models' accuracy soars to an impressive 92%.

Kemper's research serves as compelling evidence that it is entirely feasible to achieve a high degree of predictive accuracy in student dropout prediction, even when utilizing a minimal set of demographic data as long as extensive academic performance data is used for prediction.

[2] Bernes et al., *Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods*

Much like the Kemper dataset, Bernes' dataset encompasses only 4 demographic attributes: gender, place of birth, nationality, and immigration background (achieved through name-based imputation), with the remaining dataset features focused on academic performance. An intriguing observation from Bernes' study, relevant to our research inquiry, underscores the predictive capabilities of the models. When exclusively leveraging academic data, these models demonstrate a 71% accuracy rate in forecasting student dropout. However, when incorporating demographic data into the equation, the accuracy climbs modestly by 3%, reaching 74%, a point explicitly stated by Bernes.This marginal increase suggests that demographic information may only contribute minimally to the overall predictive performance and warrants further study.

[3] Francesco et al., *Deep learning approach for predicting university dropout: a case study at Roma Tre University*

Building upon the insights gleaned from the prior two studies, which underscore the critical role of academic data in predicting student dropout, Francesco's research delves into the possibility of achieving strong model performance using solely administrative (personal and socioeconomic) data. The findings of this investigation clearly indicate that models relying exclusively on administrative attributes yield subpar results, falling short when compared to models trained with a combination of administrative and academic performance attributes.

While the primary focus of these three papers isn't algorithm fairness, their collective findings converge to the same conclusion: demographic data may make only a modest contribution to the predictive capacity of models in the context of student dropout. Conversely, academic data, such as course load and examination results, emerges as the pivotal factor in accurately forecasting student attrition.

II. Algorithm Fairness and Discrimination-aware Data Mining Practice in General

[4] Žliobaitė, I. ***Measuring discrimination in algorithmic decision making***

This paper serves as a foundational introduction to the potential for machine learning-based decisions to inadvertently perpetuate discrimination against specific demographic groups. It elucidates the various metrics and methodologies available to gauge the fairness in machine learning. Notably, despite government regulations aimed at safeguarding individuals from differential treatment on the basis of gender and race etc, machine learning models can paradoxically utilize these "protected features" to formulate decision criteria. This leads to two individuals, sharing otherwise identical profiles and characteristics in all other aspects, can receive disparate predictions solely due to difference in a protective feature, such as gender.

Furthermore, the paper delves into the critical concept of the trade-off between fairness and accuracy and the metrics that measure them. While the omission of sensitive attributes can render machine models more equitable, this might come at the expense of reduced predictive accuracy.

[5] Kelley et al. ***Removing Demographic Data Can Make AI Discrimination Worse***

Kelley contends that refraining from collecting and utilizing sensitive features may appear to be a viable means to attain algorithm fairness. However, it is precisely the inclusion of these features that enables the assessment of bias and fairness. According to Kelly et al, a more nuanced and constructive approach would entail proactively establishing guidelines pertaining to the acquisition and utilization of sensitive features, as well as actively monitoring and assessing algorithmic outcomes for any signs of discrimination.

These two papers provide basic understandings surrounding algorithm fairness and the fairness-accuracy trade-off. Familiarity with these issues is essential for the two next two papers as they deal directly with university dropout prediction and algorithm fairness.

III. Research on Algorithm Fairness in Student Dropout Prediction

A limited number of studies are dedicated exclusively to the intersection of algorithm fairness and student dropout prediction. Intriguingly, two consulted studies on the topic employ very similar methodology and share a common finding: the incorporation of protected features has a negligible impact on the predictive performance of models in forecasting student dropout in university.

[6] Deho, O. et al. (2023). Should Learning Analytics Models Include Sensitive Attributes? Explaining the Why. ***IEEE Transactions on Learning Technologies***

[7] Yu, Renzhe, et al. (2021). Should College Dropout Prediction Models Include Protected Attributes?

These two studies, although different in terms of dataset size and the specific protected attributes involved, employ a shared methodology: utilizing their respective source datasets, they create multiple subsets: one set including protected attributes and another without them. Deho's study includes protected features such as gender, age, disability, and home language, while Yu's study covers gender, first-generation college status, minority membership, and high financial need. Subsequently, models are trained using these subsets, and their performance is assessed using standard performance metrics.

Both studies' findings indicate that the inclusion of protected attributes does not significantly impact the models' performance in predicting student dropout. Instead, the most predictive power in student dropout prediction stems from academic performance data.

While these results are surprising, it's essential to note that both studies construct and tailor data features to address their specific research questions. Moreover, the comparison centers on the isolation of individual protected features, such as gender, rather than considering the entirety of demographic or socioeconomic features as a whole.

The summary of literature will follow, as it is incorporated with the research question.

Research Questions

Our study aims to contribute to the understanding of algorithm fairness and university dropout prediction in machine learning by asking these research questions:

Attribute Types and Prediction Performance in University Dropout

How do different types of attributes impact the performance of student dropout prediction models?

What are the consequences of including or excluding specific classes of features on the accuracy of student dropout predictions?

The existing body of literature suggests that demographic features have only a marginal impact on model performance in predicting student dropout. However, this conclusion is often drawn from experiments involving the removal of a single feature or a small subset of demographic attributes (up to four), especially when they only make up a very small portion of the total features. Given the increasing awareness and sensitivity surrounding the collection and utilization of individual demographic and socioeconomic data, further research in this area holds significant potential. This question is pivotal because if accurate dropout prediction models can be developed without relying on such data, it can provide reassurance that model fairness is not compromised for the sake of accuracy.

Dataset, Methodology and Codes Repository

The dataset used in this project is introduced by Valentim Realinho in a data descriptor paper named *Predicting Student Dropout and Academic Success*, with 4424 records and 35 attributes featuring 4424 students of an anonymized university, created by Polytechnic Institute of Portalegre. The dataset is also available in the UC Irvine Machine Learning Repository.

Github repository of this project: https://github.com/bryantoca/capstone_project

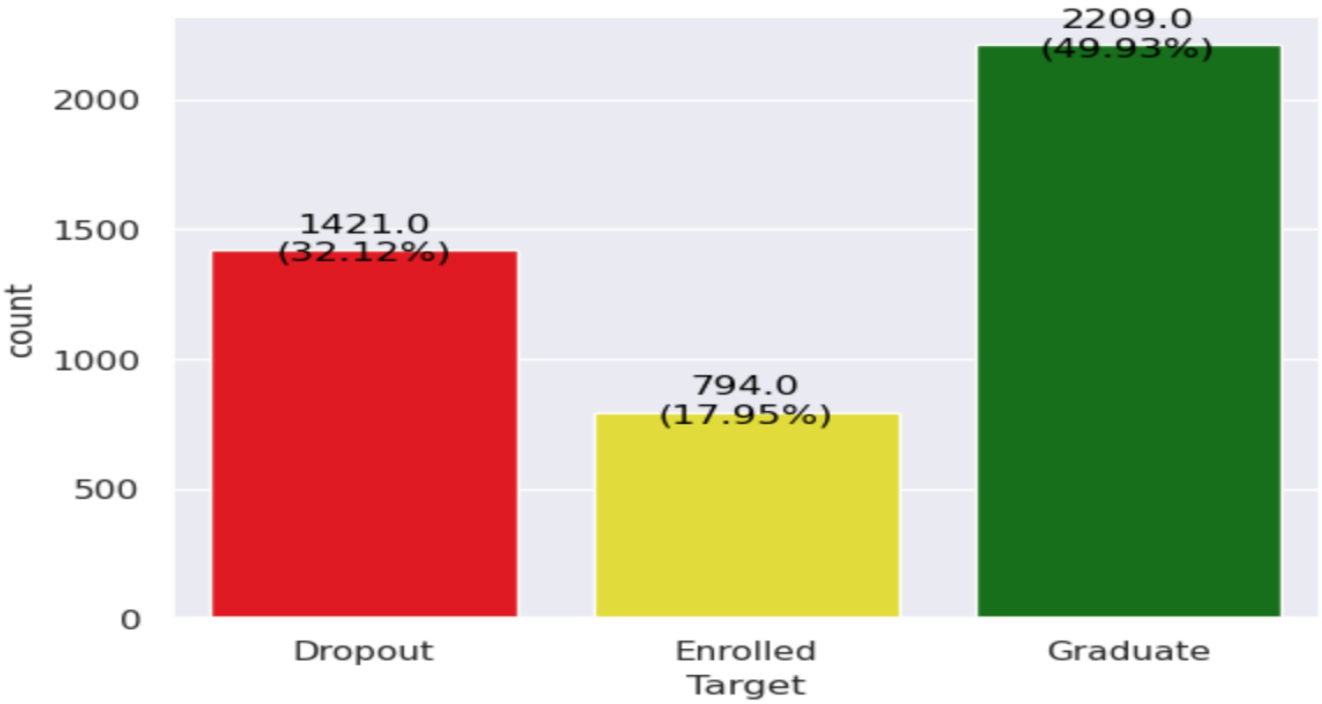
- Original Dataset: You can access the original dataset here.
- Reverse Encoded Version: If you need the dataset with nominal/categorical features reversed from numeric encoding, accessible here.
- Capstone_module_02.ipynb Notebook: Basic Exploratory Data Analysis (EDA) and summary statistics, accessible here.
- Capstone_module_03.ipynb Notebook: Initial Results and code, accessible here.
- Capstone_module_04.ipynb Notebook: Final Results and Project Report, accessible here.
- Final Results Only accessible here.

This research adopts Deho's methodology and calculation metrics for performance and fairness; modifications are made in order to serve our research questions.

Python, as well as Pandas, Seaborn, Matplotlib, Numpy libraries, are used in this project. The codes are saved in jupyter notebook format.

The dataset has a number of unique characteristics:

1. Target has three classes: Graduate, Enrolled, Dropout. (Status at the end of expected degree completion time frame)



2. Rich in demographic (6 features), socioeconomic features (8 features).
3. Inclusion of macroeconomic data at the time of student enrollment (3 features).
4. Academic features can be separated into three subcategories: pre-university information (5 features), academic at end of 1st semester (6 features), academic at the end of 2nd semester (6 features).
5. The dataset is prepared and shared for general machine learning.
6. Nominal features, such as job, qualification etc, are encoded as numerics (categorical data)
7. No missing or null values.

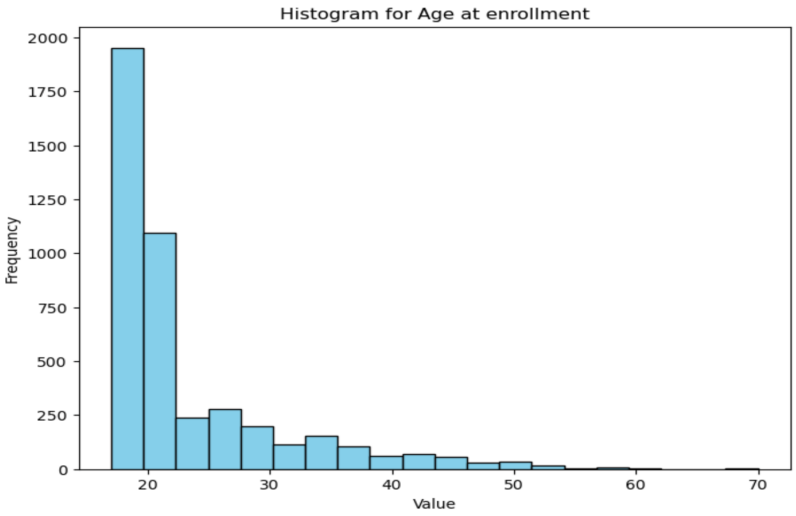
Class of Attribute	Attribute	Type
Demographic data	Marital status	Categorical
	Nationality	Categorical
	Displaced	Binary
	Gender	Categorical
	Age at enrollment	Numeric / discrete
	International	Binary
Socioeconomic data	Mother's qualification	Categorical
	Father's qualification	Categorical
	Mother's occupation	Categorical
	Father's occupation	Categorical
	Educational special needs	Categorical
	Debtor	Binary
	Tuition fees up to date	Binary
	Scholarship holder	Binary
Macroeconomic data	Unemployment rate	Numeric, Continuous
	Inflation rate	Numeric, Continuous
	GDP	Numeric, Continuous
Academic data at enrollment	Application mode	Categorical
	Application order	Ordinal
	Course	Categorical
	Daytime / evening attendance	Categorical
	Previous qualification	Categorical
Academic data at the end of 1st semester	CU 1st sem (credited)	Numeric, Discrete
	CU 1st sem (enrolled)	Numeric, Discrete
	CU 1st sem (evaluations)	Numeric, Discrete
	CU 1st sem (approved)	Numeric, Discrete
	CU 1st sem (grade)	Numeric, Continuous
	CU 1st sem (without evaluations)	Numeric, Discrete
Academic data at the end of 2nd semester	CU 2nd sem (credited)	Numeric, Discrete
	CU 2nd sem (enrolled)	Numeric, Discrete
	CU 2nd sem (evaluations)	Numeric, Discrete
	CU 2nd sem (approved)	Numeric, Discrete
	CU 2nd sem (grade)	Numeric, Continuous
	CU 2nd sem (without evaluations)	Numeric, Discrete
Target	Target	Categorical

Seventy-five records have been identified as anomalies and are identified by using

```
anomalies = data[(data['Target'] == 'Graduate') & (df.iloc[:, 21:33].eq(0).all(axis=1))]
```

These 75 students graduated with zero as value in number of courses taken and grades; they are removed as anomalies.

Attributes “Age at enrollment” and “Grades” have unique features:

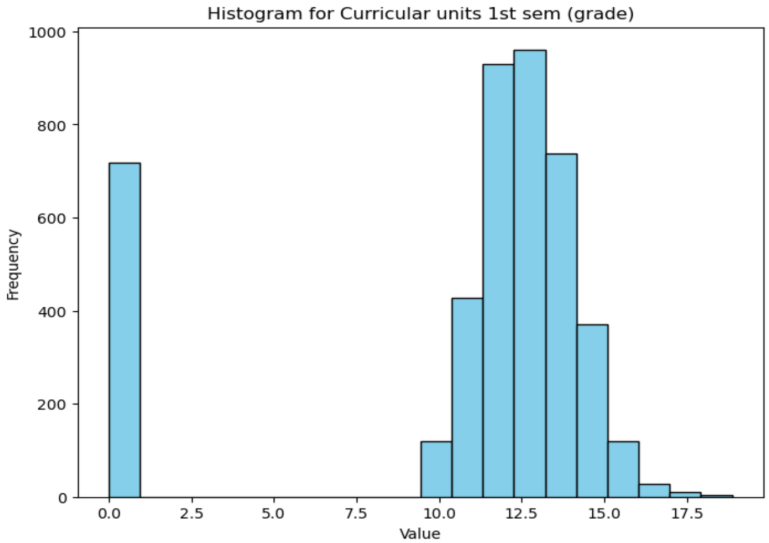


Mean age: 23.26

SD: 7.59

Range: 17-70

Age is not normally distributed; most of the students are in their 20’s.



Mean grade : 10.64

SD: 4.84

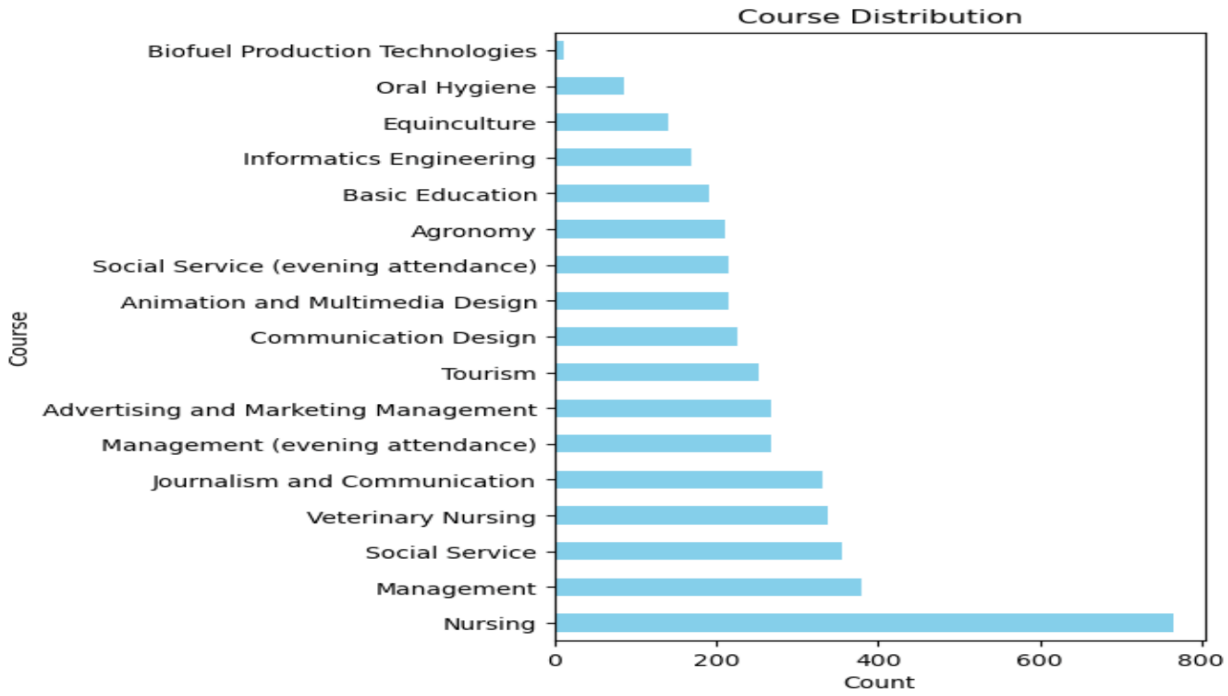
Range: 0 - 18.88

Grades below 10 are considered as fail and are reported as 0.

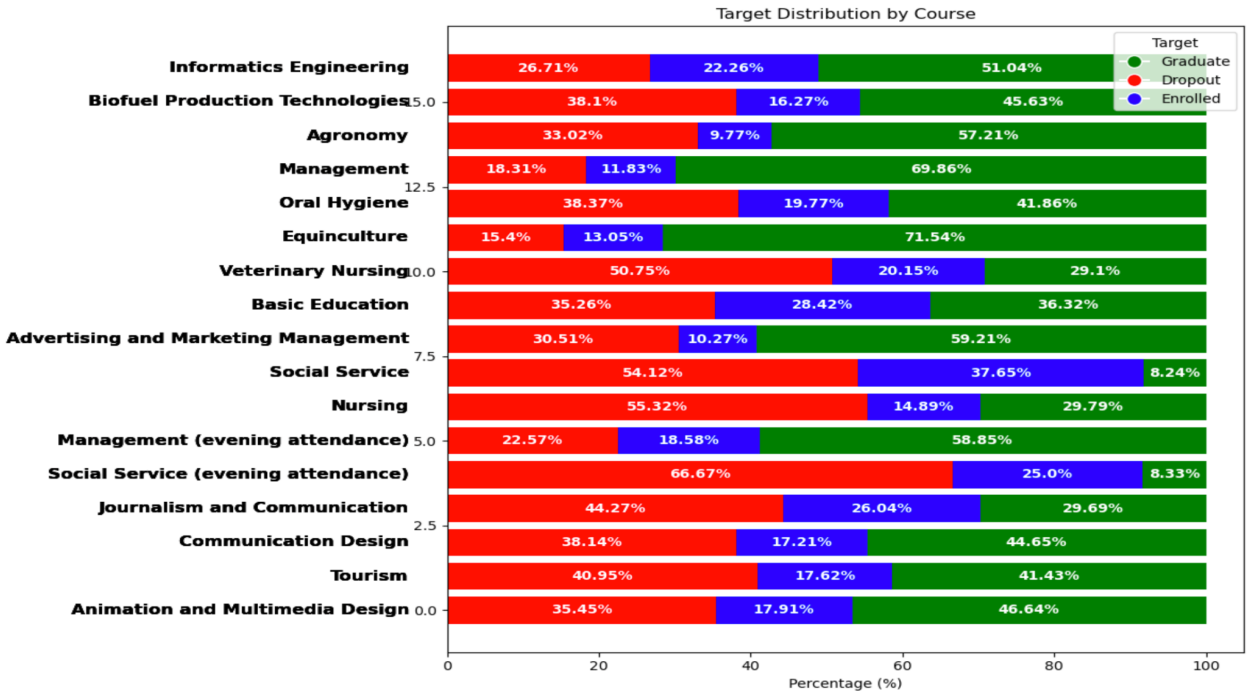
Grades other zero proximate normal distribution.

Most of the features are categorical in nature with high cardinality. However, their distributions are not even, as few values make up most of the distribution. Below is some examples:

Programmes Taken



Target distribution in each programmes



Methodology and Measurement Metric

The literature review underscores the significance of academic performance data in predicting student dropout. While demographic and socioeconomic are routinely collected and used in constructing machine learning models, their actual contribution as a cluster to a model performance is an area of research that remains underexplored.

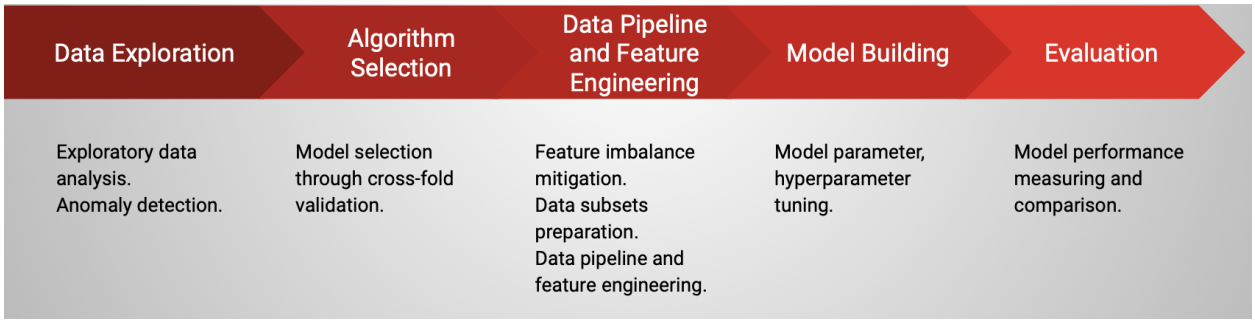
In order to address this research gap, this study sets out to construct five datasets, containing the same records: their differences lie in the classes of attributes they contain. Each set is used to train the same machine learning algorithms and their performances are recorded and compared..

		Set 1	***Set 2***	***Set 3***	***Set 4***	***Set 5***
Class of Attribute						
Demographic			x		x	x
Socioeconomic				x	x	x
Macroeconomic		x	x	x	x	
Academic		x	x	x	x	

Five initial algorithms are selected. Using Cross validation and the train-validation set, three final models are selected and these will be trained using the train-validation set and tested test set. The performance metrics used to measure and compare are accuracy, precision, recall and F1-score of each class as follows.

	Subset	Attribute Groups	Accuracy	Precision (Dropout)	Precision (Enrolled)	Precision (Graduate)	Recall (Dropout)	Recall (Enrolled)	Recall (Graduate)	F1-Score (Dropout)	F1-Score (Enrolled)	F1-Score (Graduate)
0	s1	Academic, Macroeconomic	0.7632	0.7393	0.5169	0.8204	0.7782	0.2644	0.9558	0.7582	0.3498	0.8829
1	s2	Academic, Macroeconomics, Demographic	0.7517	0.7546	0.4787	0.8008	0.7632	0.2586	0.9442	0.7589	0.3358	0.8666
2	s3	Academic, Macroeconomics, Socioeconomic	0.7828	0.7754	0.6344	0.8144	0.8045	0.3391	0.9488	0.7897	0.4419	0.8765
3	s4	Academic, Macroeconomic, Demographic, Socioeco...	0.7805	0.7802	0.6237	0.8095	0.8008	0.3333	0.9488	0.7904	0.4345	0.8737
4	s5	Demographic, Socioeconomic	0.5483	0.5292	0.2472	0.6143	0.5789	0.1264	0.7000	0.5530	0.1673	0.6543

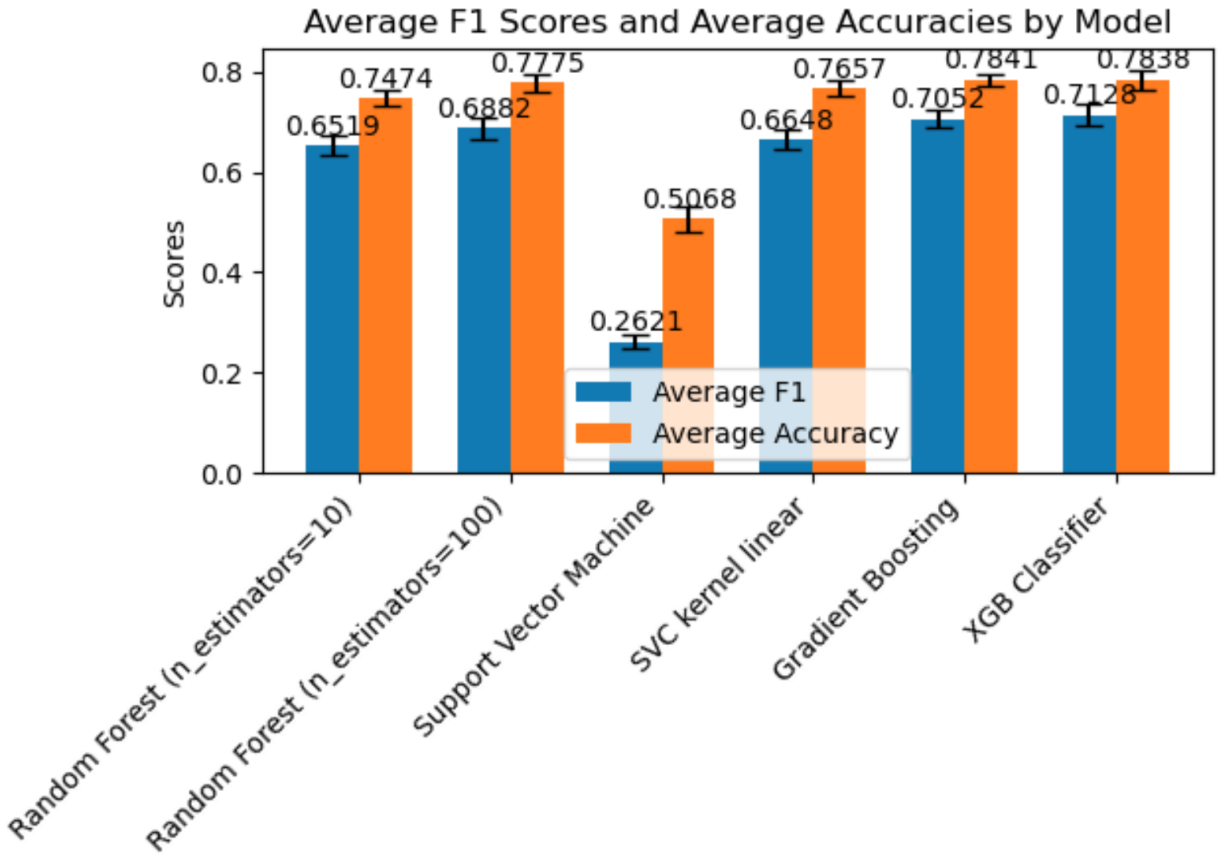
Project Progression



1. Data Exploration
    a. EDA to gain insights into the dataset and calculate general summary statistics.
    b. Review EDA findings which can be accessed at the following link https://nbviewer.org/github/bryantoca/capstone_project/blob/59ae45b4d4dfe117d56f179768cb4a20a5cb9d6c/Capstone_module_02.ipynb
    c. Anomaly detection and correction.

2. Algorithm Selection
    a. Train Validation set and Test set are created. (80/20 Split).
    b. Train Validation set is used for algorithm selection.
    c. 6 classification algorithms are selected Random Forest (n_estimators=10, n_estimatos=100), Support Vector Machine, SVC Kernel linear, Gradient Boosting and XGB Classifier
    d. K-fold cross-validation for each algorithm and Average F1 and Average Accuracy are used as selection criteria.
    e. The three best performing models, Random Forest, Gradient Boosting, and XGB Classifier, are selected for the final algorithm / model building.

RESULTS OF MODEL SELECTION

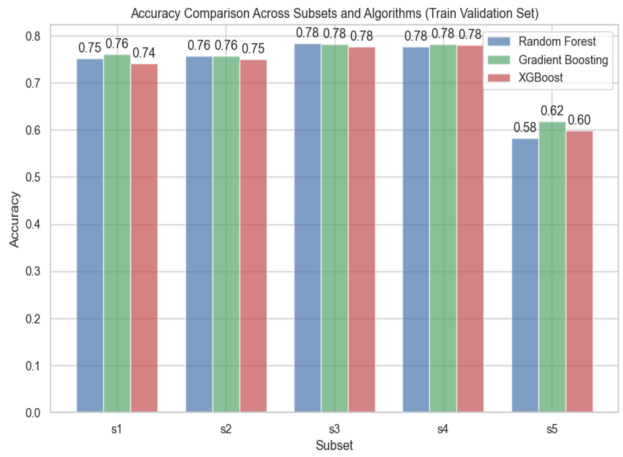


3. Model Building
    a. Three models are trained and tested twice.
        i. Using 10-fold cross validation with the train-validation set.
        ii. Using train-validation to train and test set to test(final test).
4. Results
    a. The codes used to arrive at the final results can be viewed here.
    b. For the final report, test validation set results and final results are reported.
    c. The full results and report jupyter notebook contains additional information such as classification report, feature importance, confusion matrix and normalized confusion matrix for cross-validation and final test. They are not included in this report for brevity.

FINAL RESULTS

Accuary of the 3 models (train-validate and test sets)

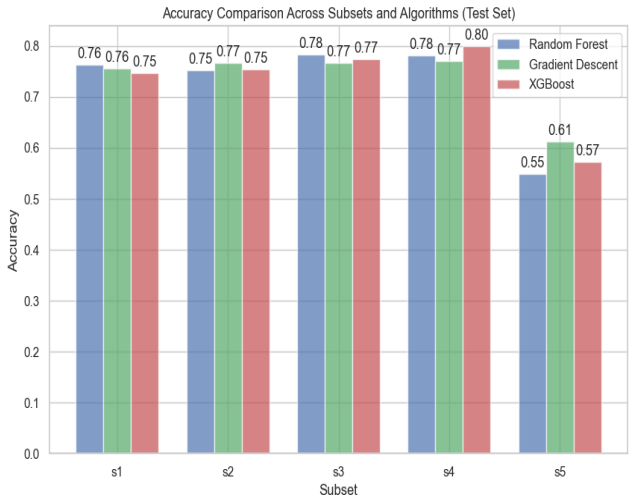
Subset	Attribute_groups	Ave Accuracy RF (validate)	Accuracy RF (test)	Ave Accuracy GB (valide)	Accuracy GD (test)	Ave Accuracy XGB (validate)	Accuracy XGB (test)
s1	Academic, Macroeconomic	0.752514	0.763218	0.759988	0.755172	0.741303	0.747126
s2	Academic, Macroeconomics, Demographic	0.757112	0.751724	0.756249	0.766667	0.749641	0.754023
s3	Academic, Macroeconomics, Socioeconomic	0.783268	0.782759	0.781255	0.766667	0.776088	0.773563
s4	Academic, Macroeconomic, Demographic, Socioeconomic	0.776655	0.78046	0.781258	0.770115	0.780968	0.798851
s5	Demographic, Socioeconomic	0.582357	0.548276	0.618277	0.611494	0.597879	0.572414



The accuracy scores on the validation and test sets exhibit similar patterns and demonstrate comparable scores.

Models trained solely with demographic and socioeconomic data (s5) perform the worst, with accuracy ranging from 55% to 62% in both validation and test sets across all three algorithms.

Models trained exclusively with academic and macroeconomic data (s1) achieve accuracy rates ranging from 74% to 76%. S1 serves as the baseline for comparisons with s2, s3 and s4.



The inclusion of additional demographic data (s2) results in accuracy fluctuations of ±1% when compared to s1.

Models trained with baseline and additional socioeconomic data (s3) likewise exhibit slight fluctuation, up to 3% compared with baseline.

Models trained with all features (s4) show the most substantial increase, ranging from a 1% to 5%. The test set accuracy is highest for models using XGBoost, reaching 80%.

Because the validation set and test set perform similarly and exhibit similar patterns, the following discussion only uses test set results. The results of the three algorithms are presented. But discussions followed are based on results and finding of models using XGBoost.

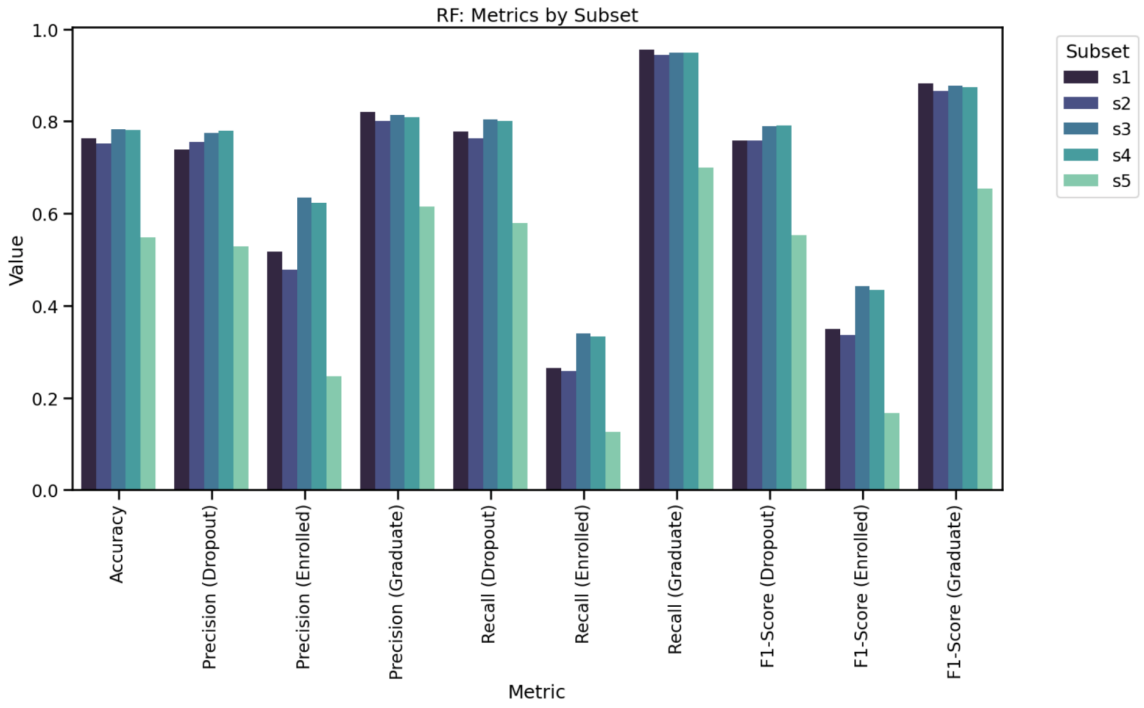
1. Random Forest

Test Set Results

```
In [26]: RF_test_df
```

26]:

	Subset	Attribute Groups	Accuracy	Precision (Dropout)	Precision (Enrolled)	Precision (Graduate)	Recall (Dropout)	Recall (Enrolled)	Recall (Graduate)	F1-Score (Dropout)	F1-Score (Enrolled)	F1-Score (Graduate)
0	s1	Academic, Macroeconomic	0.7632	0.7393	0.5169	0.8204	0.7782	0.2644	0.9558	0.7582	0.3498	0.8829
1	s2	Academic, Macroeconomics, Demographic	0.7517	0.7546	0.4787	0.8008	0.7632	0.2586	0.9442	0.7589	0.3358	0.8666
2	s3	Academic, Macroeconomics, Socioeconomic	0.7828	0.7754	0.6344	0.8144	0.8045	0.3391	0.9488	0.7897	0.4419	0.8765
3	s4	Academic, Macroeconomic, Demographic, Socioeco...	0.7805	0.7802	0.6237	0.8095	0.8008	0.3333	0.9488	0.7904	0.4345	0.8737
4	s5	Demographic, Socioeconomic	0.5483	0.5292	0.2472	0.6143	0.5789	0.1264	0.7000	0.5530	0.1673	0.6543

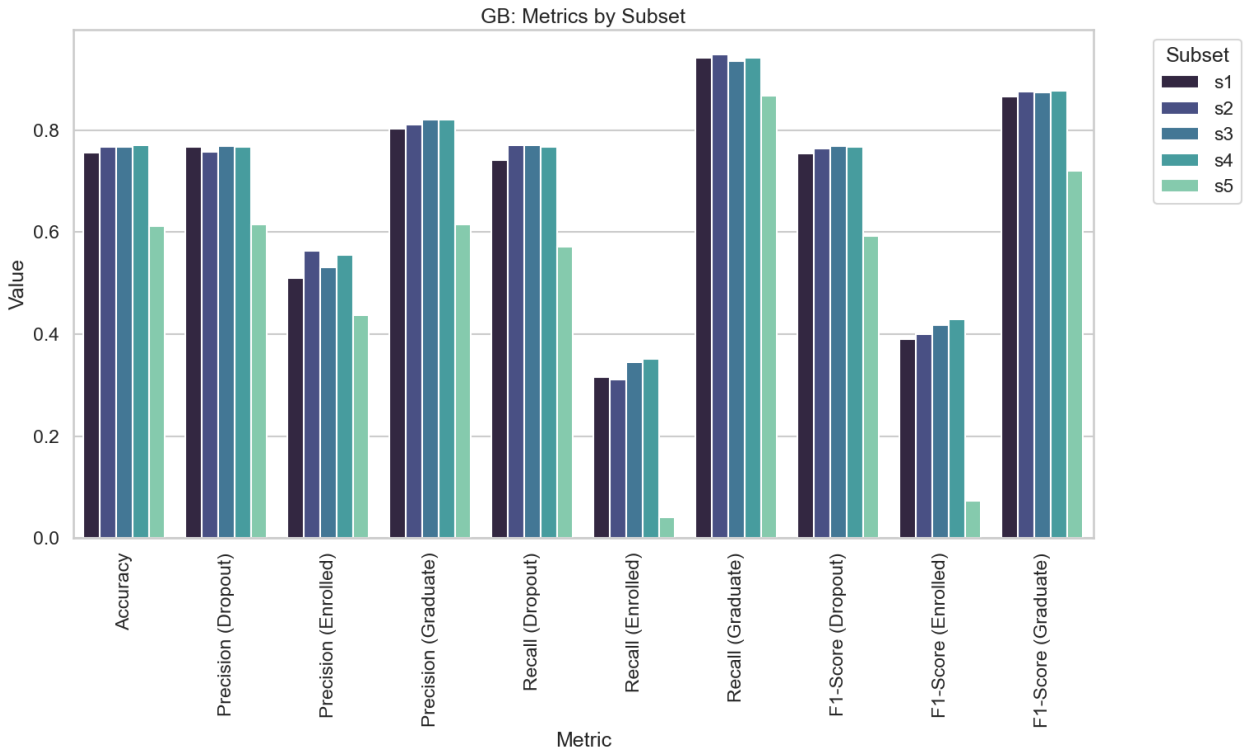


2. Gradient Boosting

```
In [28]: GB_test_df
```

28]:

	Subset	Attribute Groups	Accuracy	Precision (Dropout)	Precision (Enrolled)	Precision (Graduate)	Recall (Dropout)	Recall (Enrolled)	Recall (Graduate)	F1-Score (Dropout)	F1-Score (Enrolled)	F1-Score (Graduate)
0	s1	Academic, Macroeconomic	0.7552	0.7665	0.5093	0.8020	0.7406	0.3161	0.9419	0.7533	0.3901	0.8663
1	s2	Academic, Macroeconomics, Demographic	0.7667	0.7565	0.5625	0.8111	0.7707	0.3103	0.9488	0.7635	0.4000	0.8746
2	s3	Academic, Macroeconomics, Socioeconomic	0.7667	0.7678	0.5310	0.8204	0.7707	0.3448	0.9349	0.7692	0.4181	0.8739
3	s4	Academic, Macroeconomic, Demographic, Socioeco...	0.7701	0.7669	0.5545	0.8198	0.7669	0.3506	0.9419	0.7669	0.4296	0.8766
4	s5	Demographic, Socioeconomic	0.6115	0.6154	0.4375	0.6145	0.5714	0.0402	0.8674	0.5926	0.0737	0.7194

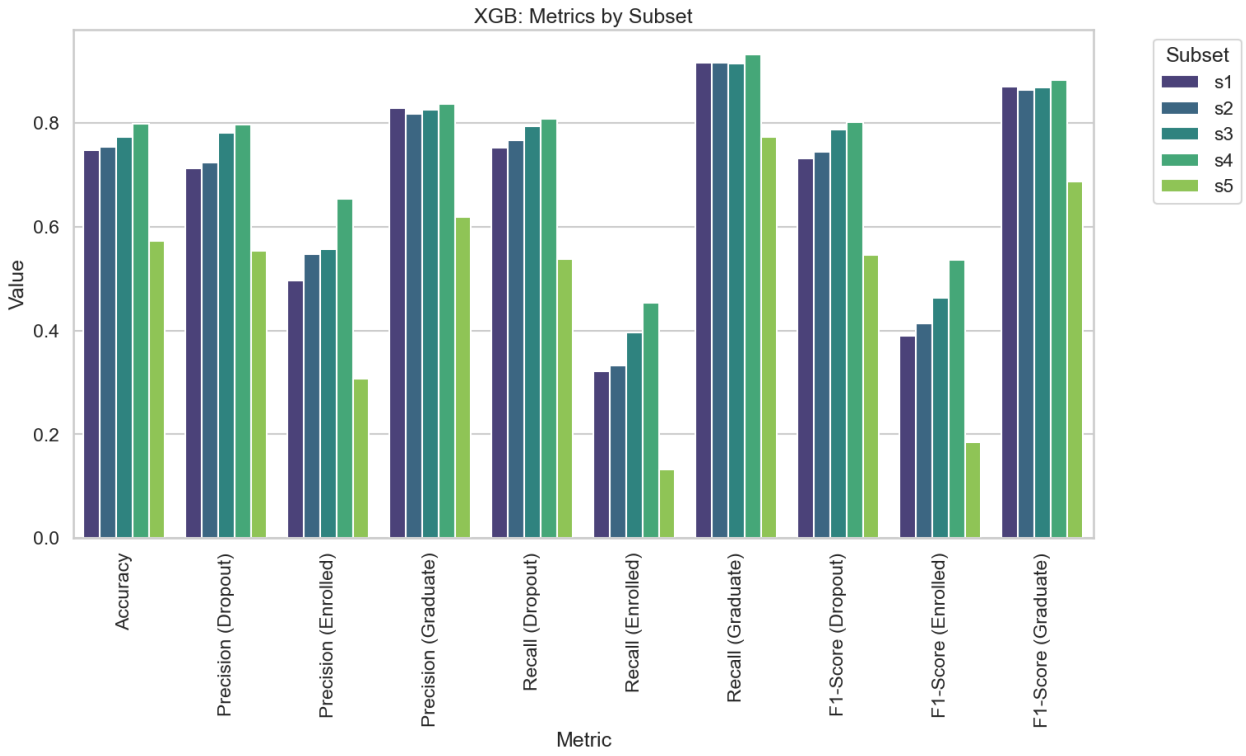


3. XGBoost

```
In [31]: XGB_test_df
```

[31]:

	Subset	Attribute Groups	Accuracy	Precision (Dropout)	Precision (Enrolled)	Precision (Graduate)	Recall (Dropout)	Recall (Enrolled)	Recall (Graduate)	F1-Score (Dropout)	F1-Score (Enrolled)	F1-Score (Graduate)
0	s1	Academic, Macroeconomic	0.7471	0.7117	0.4956	0.8277	0.7519	0.3218	0.9163	0.7313	0.3902	0.8698
1	s2	Academic, Macroeconomics, Demographic	0.7540	0.7234	0.5472	0.8174	0.7669	0.3333	0.9163	0.7445	0.4143	0.8640
2	s3	Academic, Macroeconomics, Socioeconomic	0.7736	0.7815	0.5565	0.8256	0.7932	0.3966	0.9140	0.7873	0.4631	0.8675
3	s4	Academic, Macroeconomic, Demographic, Socioeco...	0.7989	0.7963	0.6529	0.8372	0.8083	0.4540	0.9326	0.8022	0.5356	0.8823
4	s5	Demographic, Socioeconomic	0.5724	0.5543	0.3067	0.6182	0.5376	0.1322	0.7721	0.5458	0.1847	0.6867



Accuracy

Patterns and characteristics identified in XGBoost models (s1, s2, s3, s4, s5) are observed in Random Forest, and Gradient Descent models. The XGBoost models display diverse performance across attribute groups represented by subsets s1 to s5. Subset s1, incorporating academic and macroeconomic data, achieves an accuracy of 74.71% and serves as a reference point for comparison. This result is slightly higher than the accuracy reported in Bernes research (71%), while also using only academic data alone , used a different dataset and machine learning model.

In subsets s2 (baseline + demographic data) and s3 (baseline + socioeconomic data), the models show improvements compared to the baseline, with accuracy values of 75.40% and 77.36% respectively. These findings underscore the model's effectiveness in predicting student outcomes using academic and macroeconomic features alone. Notably, the introduction of additional features yields marginal improvements in accuracy.

When trained with all available features (subset s4), the model experiences a substantial increase in accuracy, reaching 79.89%. This represents a notable 5.18% improvement compared to the baseline model s1. In contrast, using only demographic and socioeconomic data (s5) results in a lower accuracy of 57.24%, indicating a significant 17.47% decrease compared to the baseline.

Comparison of Precision and Recall of Classes

For the graduate class, across subsets (s1-s4) the models maintain a high precision and recall for graduate instances, reaching 90% in recall values and 80% for precision.

Conversely, the dropout class displays lower precision and recall rates, consistently ranging between 70-80% across subsets (s1-s4).

The enrolled class demonstrates the widest range and fluctuation in precision rates (50%-65%) and recall rates (32% to 45%), showing an increase from s1 to s4. These observations show the varying performance of the models in predicting instances of different classes, highlighting the

need for nuanced interpretation of results and fine tuning the models so that they are optimized for the class of primacy interest.

Discussion and Conclusion

This research project aims to answer a simple question: does using academic data alone allow us to train machine learning models that are sufficiently accurate to predict university student's dropout. Unlike the findings of Yu and Deho, which state that excluding sensitive features does not significantly impact the machine learning model's performance, our research finds a significant difference in the machine learning model using just academic and macroeconomic data and the models using additional demographic and socioeconomic features. The additional features lead to an increase of 5% in accuracy (XGBoost model, from 75% to 80%).

While this research's findings do not agree with those of the two leading researchers, it is important to note that our research differs significantly from Yu's research and Deho's research in two important aspects. First of all, their studies only exclude and compare up to 4 features: Deho's study excludes protected features such as gender, age, disability, and home language, while Yu's study covers gender, first-generation college status, minority membership, and high financial need. This research involves the exclusion of 14 features (the differences between s1 and s4 set). Another major difference is that, unlike their binary classification (graduate and dropout) , our research involves three-class classification (graduate, enrolled, and dropout). So while the subject of the research remains the same, to predict university student's dropout, the paradigm of this research is different from those cited in literature review.

Another shortcoming of this present research pertains to the complexity of its methodology and experiment design. Using only one best machine model, together with binary classification (graduate/non graduate, rather than dropout / enrolled / graduate) and two subsets (one with all the features and one with only academic features) would have decreased the coding and programming complicity and made the finding more concise and easier to understand.

The precision and recall rates of the three classes in our research differ greatly. Graduate class has the highest precision and recall rate and less fluctuation, followed by the dropout class.

Enrolled has the lowest precision and recall as well as highest fluctuation. Two possible explanations for these differences: 1: There is a slight class imbalance - Graduate (49.93%), Dropout (32.12%) and Enrolled (17.75%). 2: the demarcation and boundary are less definite between dropout and enrolled classes, leading to lower precision and recall rates in these two classes. Further research with the same dataset can be done using a binary classification approach (Graduate / Non Graduate), which would also have a balanced dataset and be more comparable with existing research.

Because of the methodology of this research, feature selection is done as a cluster of features based on types of data and as subsets (s1, s2, s3, s4, s5), rather than the conventional method of selecting and deselecting individual features. This may have limited the performance of the models developed. Additionally, it is important to note that statistical tests should have been conducted to analyze the research data reported. Unfortunately, due to time constraints, this aspect was not thoroughly addressed in the current report.

Nevertheless, this present research shows that a reasonably accurate machine learning model can be developed using only academic data to predict university student dropout with 74% accuracy. While it is true that using additional features related to demographic and socioeconomic data leads to a higher accuracy in dropout prediction, the improvement is at the cost of using more sensitive and private features, many of them considered as protected. As concerns and questions surrounding fairness and potential discrimination in AI-driven decision making are on the rise, we believe our research contributes to informed decision making by assessing the costs and benefits associated with the inclusion or exclusion of protected and sensitive data in predicting university student dropout and paves the way for future studies into the interplay between predictive modeling, privacy and fairness concerns and the accuracy of outcomes.

References

Agrusti, F., Mezzini, M., & Bonavolontà, G. (2020). Deep learning approach for predicting university dropout: a case study at Roma Tre University . *Journal of E-Learning and Knowledge Society*, *16*(1), 44-54.https://doi.org/10.20368/1971-8829/1135192

Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining*, *11*(3), 1–41. https://doi.org/10.5281/zenodo.3594771

Deho, O. B., Joksimovic, S., Li, J., & Zhan, C. (2023). Should Learning Analytics Models Include Sensitive Attributes? Explaining the Why. *IEEE Transactions on Learning Technologies, 16*(4), 560-572. 10.1109/TLT.2022.3226474

Kapila Devi & Saroj Ratnoo (2022) Predicting student dropouts using random forest, Journal of Statistics and Management Systems, 25:7, 1579-1590, DOI: 10.1080/09720510.2022.2130570

Kelley, St et al. (2023, March 6).Removing Demographic Data Can Make AI Discrimination Worse. Harvard Business Review https://hbr.org/2023/03/removing-demographic-data-can-make-ai-discrimination-worse#:~:text=and%20machine%20learning-,Removing%20Demographic%20Data%20Can%20Make%20AI%20Discrimination%20Worse,race%20can%20produce%20fairer%20outcomes.

Kemper, L. et al.(2020) Predicting student dropout: A machine learning approach, European Journal of Higher Education, 10:1, 28-47, DOI: 10.1080/21568235.2020.1718520

Namoun A, Alshanqiti A. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. Applied Sciences. 2021; 11(1):237. https://doi.org/10.3390/app11010237

Jovial Niyogisubizo, Lyuchao Liao, Eric Nziyumva, Evariste Murwanashyaka, Pierre Claver Nshimyumukiza. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization, Computers and Education: Artificial Intelligence,Volume 3,2022,100066,ISSN 2666-920X. https://doi.org/10.1016/j.caeai.2022.100066.

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, *7*(11), 146. https://doi.org/10.3390/data7110146

Realinho,Valentim, Vieira Martins,Mónica, Machado,Jorge, and Baptista,Luís. (2021). Predict students' dropout and academic success. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89.
https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Yu, R., Lee, H., Kizilcec, R. F. (2021, June 8).Should College Dropout Prediction Models Include Protected Attributes? *L@S '21: Proceedings of the Eighth ACM Conference on Learning @ Scale (*2021)

Žliobaitė, I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 1060–1089 (2017). https://doi.org/10.1007/s10618-017-0506-1