Balancing Act: Sensitive Data and Accuracy in University Dropout Prediction

Bryan Chi Fai Pang
Student ID: 50120081
CIND820 Project Presentation
Supervisor: Dr Ceni BABAOGLU

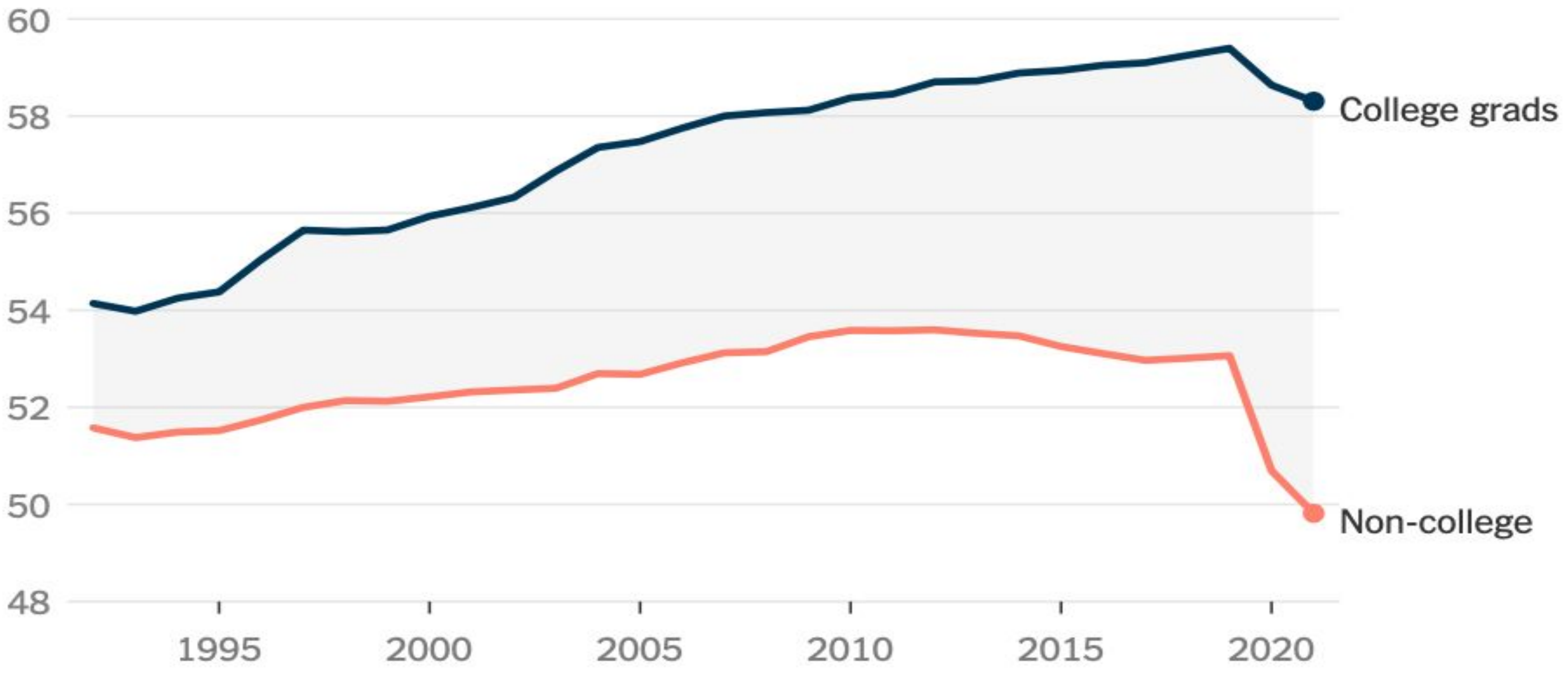
University Student Dropout

Global dropout rate ranging from 30% OECD countries to 50.9% in Costa Rica

- **North American Context:**
- **Canada: Up to 20% of students quit, 20%-50% of students change initial program.**
- **Canadian workers in their 40s with degree earn 53% (Cdn $13) per hour more**
- **In US, eight year difference in term of life expectancy between degree holders and non degree holders**

The mortality gap between Americans with and without four-year degrees is widening

Average years of life remaining for 25-year-old Americans



Source: Anne Case and Angus Deaton, Princeton University • By The New York Times

States and shortcomings of University Student Dropout Prediction in Machine Learning

While all the studies confirm the effectiveness of data mining and ML approach in predicting dropout...

YET:

They focus on algorithms and not on feature space...

CONCERNS:

Use of sensitive features: Is it necessary ?

Two studies focused on Algorithm Fairness and Model Performance

Deho, O. et al. (2023). Should Learning Analytics Models Include Sensitive Attributes? Explaining the Why.

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

Aware Model	Blind Model
All the features	Excluding gender, age, disability, home language

Features are considered as isolation instances, rather than part of a cluster.

Up to 4 features are excluded in these two studies and both show no significant difference in model performance.

		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	

Research Question:

Attribute Types and Prediction Performance in University Dropout

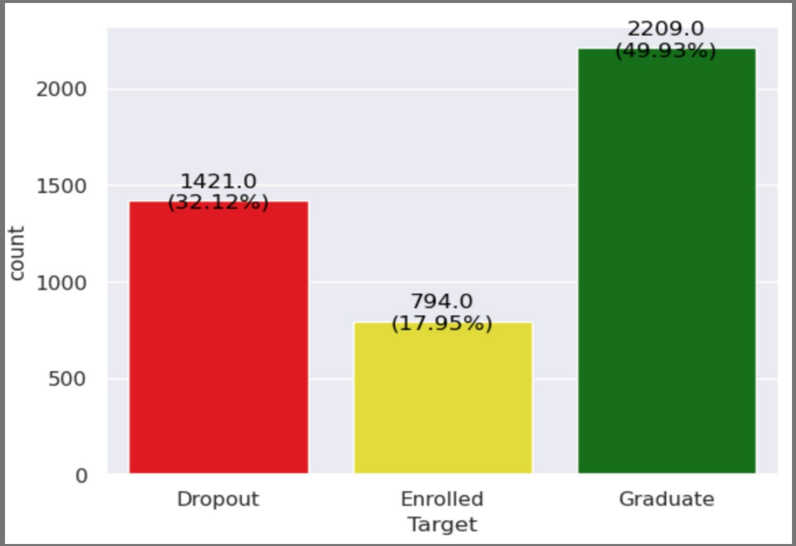
What are the consequences of including or excluding specific classes of features on the accuracy of student dropout predictions?

Valentim Realinho's dataset: 4424 students and 35 attributes, presented as a paper and available in UC Irvine Machine Learning Repository

Rich in features

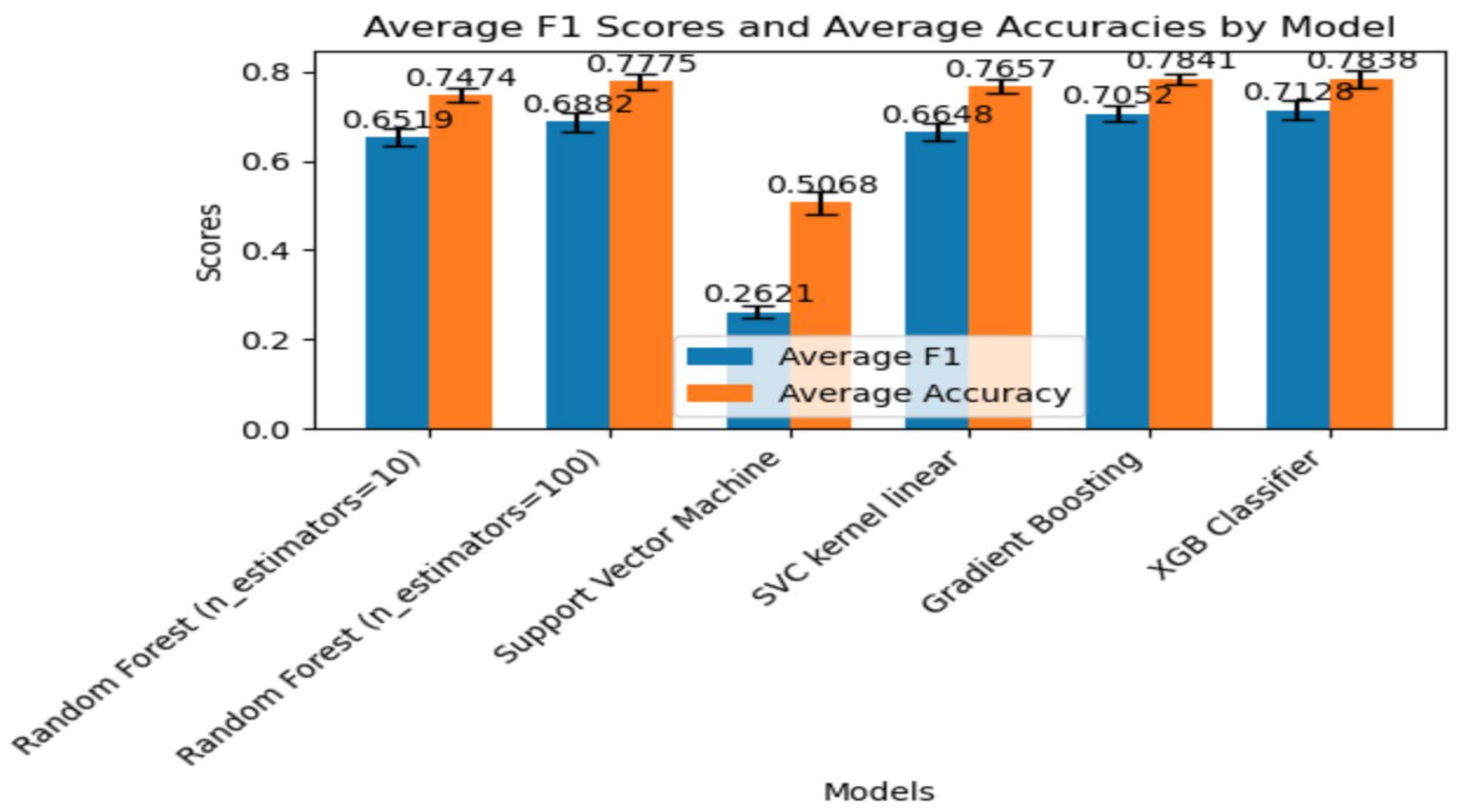
- Demographic (6 features)
- Socioeconomic (8 features)
- Macroeconomic (3 features)
- Academic (17 features)

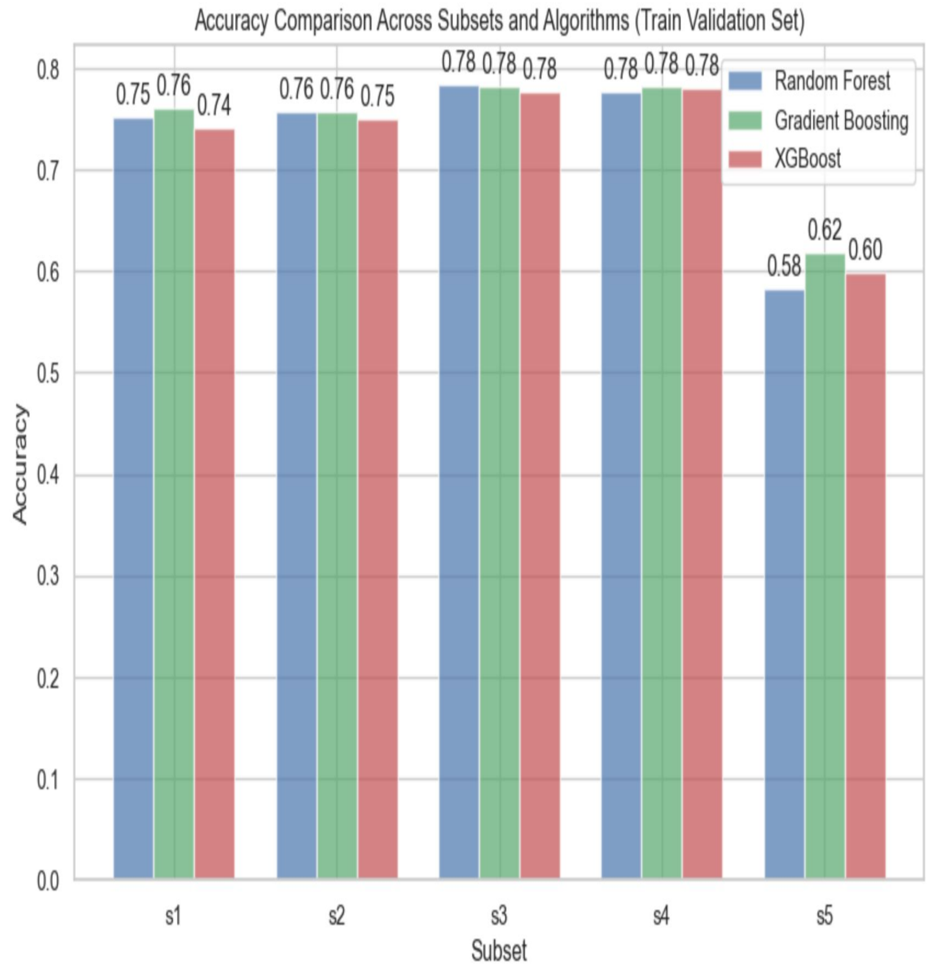
Target is three classes (Dropout, Enrolled, Graduate)

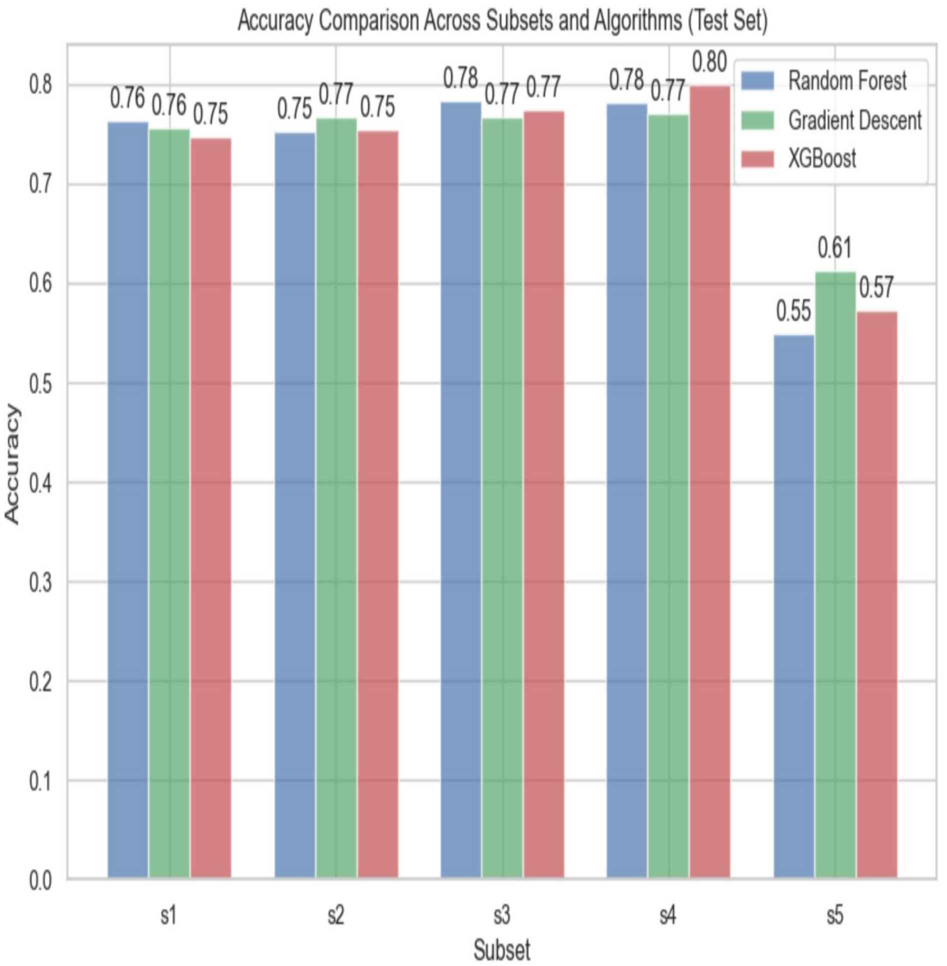


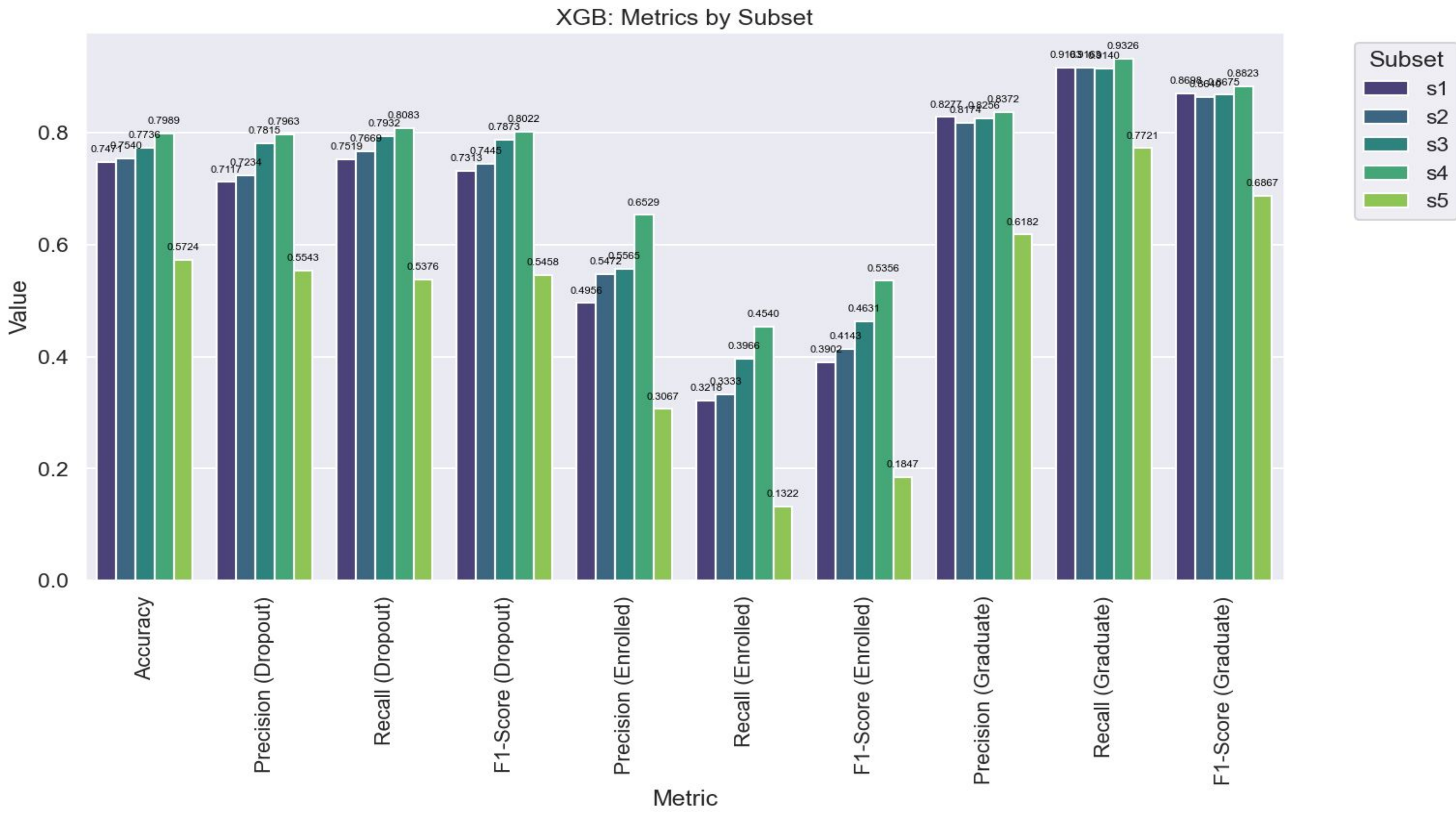
Methodology

- Dataset is split into Train/validation set (TV set) (80%) and Test Set (20%)
- Algorithm Selection: Cross validation with TV set (Random Forest n=10, RF n=100, Support Vector Machine, SVC kernel linear, Gradient Boosting, XGB Boost)
- Best three are used, based on Average F1 and Average Accuracy


Average F1 Scores and Average Accuracies by Model
Scores
0.0
0.2
0.4
0.6
0.8
0.6519
0.7474
0.6882
0.7775
0.2621
0.5068
0.6648
0.7657
0.7052
0.7841
0.7128
0.7838
Average F1
Average Accuracy
Random Forest (n_estimators=10)
Random Forest (n_estimators=100)
Support Vector Machine
SVC kernel linear
Gradient Boosting
XGB Classifier
Models


Accuracy Comparison Across Subsets and Algorithms (Train Validation Set)
Accuracy
Subset
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
s1
s2
s3
s4
s5
0.75 0.76 0.74
0.76 0.76 0.75
0.78 0.78 0.78
0.78 0.78 0.78
0.58 0.62 0.60
Random Forest
Gradient Boosting
XGBoost


Accuracy Comparison Across Subsets and Algorithms (Test Set)
Accuracy
Subset
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
s1
s2
s3
s4
s5
0.76 0.76 0.75
0.75 0.77 0.75
0.78 0.77 0.77
0.78 0.77 0.80
0.55 0.61 0.57
Random Forest
Gradient Descent
XGBoost



XGB: Metrics by Subset
Value
Metric
Subset
s1
s2
s3
s4
s5
Accuracy
Precision (Dropout)
Recall (Dropout)
F1-Score (Dropout)
Precision (Enrolled)
Recall (Enrolled)
F1-Score (Enrolled)
Precision (Graduate)
Recall (Graduate)
F1-Score (Graduate)
0.7471
0.7540
0.7736
0.7989
0.5724
0.7117
0.7234
0.7815
0.7963
0.5543
0.7519
0.7669
0.7932
0.8083
0.5376
0.7313
0.7445
0.7873
0.8022
0.5458
0.4956
0.5472
0.5565
0.6529
0.3067
0.3218
0.3333
0.3966
0.4540
0.1322
0.3902
0.4143
0.4631
0.5356
0.1847
0.8277
0.8174
0.8256
0.8372
0.6182
0.9326
0.9140
0.7721
0.8823
0.8675
0.6867

Results:

Model trained using Academic and Macroeconomic data (s1) performs well at 74.71 % accuracy.

Model with additional Demographic data (s2) shows slight improvement at 75.40% accuracy (0.69 %)

Model with baseline and additional socioeconomic data (s3) shows (2.65% increase) at 77.36% accuracy.

Model with all the data (s4) performs the best, with 79.89% accuracy (5.1% increase)

Model with just demographic and socioeconomic data shows the worst performance 57.24% accuracy (decrease in 17.47%)

Precision and Recall and Classes

- Graduate: 80% Precision and 90% in Recall throughout s1-s4
- Dropout: 70-80% both Precision and Recall s1-s4
- Enrolled: 50-65 % Precision, 32-45% in Recall s1-s4

Discussions

Use of demographic and socioeconomic data increases model accuracy by 5% in XGBoost Model, comparing to using only academic and macroeconomic data alone.

However, the increase of performance is achieved through addition of 14 features, many of them can be considered sensitive and unrelated to academic performance.

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

Achievement of the study

Refocuses on the feature space and features rather than algorithms used.

Demonstrates that reasonable good model can be developed with academic data alone.

Contributes to study of AI Fairness and ML-based decision making.

Limitations of study and scope for further research

The target of the dataset is three-class, whereas all of the studies reviewed are binary classification.

Complexity of the methodology: three models, 5 subsets and three classes. A simpler approach (two subset and a binary (graduate, non-graduate models) would be easier to understand and execute.

Thank you.