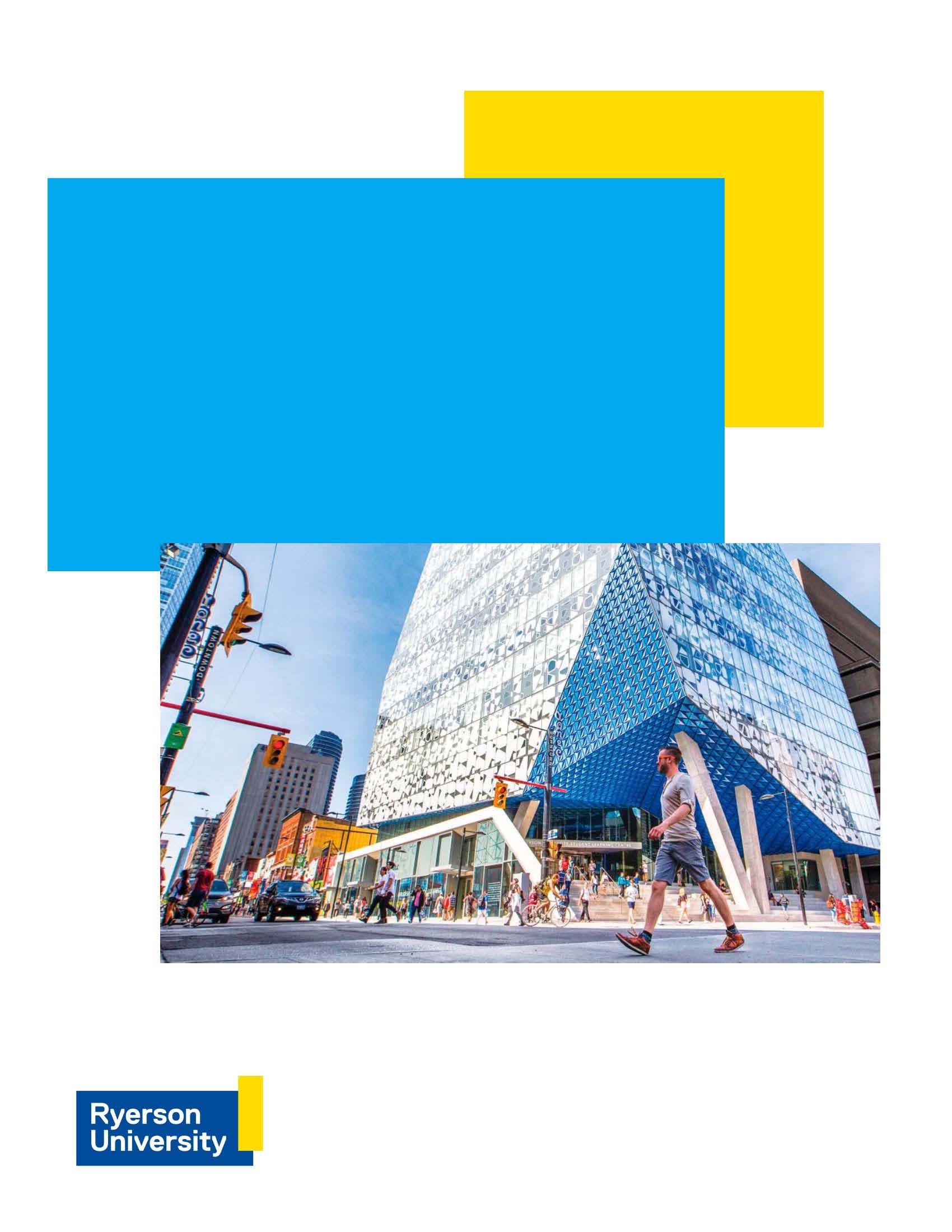
Predicting Student Dropout and Academic Success 

Literature Review

By: Evgeny Lipikhin

ID: 500157500

Supervisor’s name: Tamer Abdou, PhD

Submission date: February 20, 2023

**Table of Contents**

[**Introduction**](#_heading=h.30j0zll) **2**

[**Background**](#_heading=h.z2sm343iiq5d) **3**

[Research question](#_heading=h.l4qefggew7xg) 3

[Dataset](#_heading=h.pwsobkb7afjf) 3

[Data Profiling](#_heading=h.1xa8nyi3if7c) 4

[**Literature Search Strategy**](#_heading=h.k3mk395io70a) **5**

[**Literature Synthesis**](#_heading=h.2bbg2pjw5air) **6**

[**Evaluation Metrics**](#_heading=h.5hge3nn41q5c) **6**

[**References**](#_heading=h.nxb12tzdw4k4) **7**

# 

# Introduction

For the capstone project I have selected “Predict students' dropout and academic success” dataset from The UC Irvine Machine Learning Repository and the theme for the project is Supervised Learning - Classification and Regression.

While selecting a subject for my capstone project, I aimed to pick a topic that I could personally connect with. In my view, forecasting student dropout is a topic of great significance and interest, mainly because I have witnessed many students excel in their studies while others struggle. Although a considerable number of the students who underperformed in their academics had the potential, intelligence, and brightness, certain other circumstances caused them to either abandon their education or graduate with below-average GPAs.

Thus, I am curious to identify the primary factors that lead to student learning setbacks. I want to determine whether socioeconomic, demographic, macroeconomic, or academic factors are responsible for these issues.

In this literature review I am going to:

* Provide the description of the data
* Define my research question
* Plan my approach
* Search the literature
* Analyze material
* Manage results
* List several ML algorithms which were used in other studies
* Discuss evaluation metrics

# Background

This section provides an overview of the research question and brief description of the dataset to be used.

## Research question

Dropout prediction is an important task in education because it can help schools and educators identify students who are at risk of dropping out and provide them with the necessary support to improve their academic performance and prevent them from leaving school. Dropout rates are high in some countries, which leads to various problems. Machine learning can help to predict which students are more likely to drop out and provide early interventions to prevent this.

Following the introduction, my goal is to create multiple models that can assist me in pinpointing the specific areas where educational institutions should focus their efforts to enhance students' achievements. To achieve this, I plan to utilize various machine learning techniques and compare their accuracy based on evaluation metrics.

## Dataset

The dataset was downloaded from The UC Irvine Machine Learning Repository (<https://archive-beta.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success>).

After a quick analysis of the data:

* Data is in CSV format.
* Data is structured in a tabular format.
* Data contains 37 attributes (one of which is the target).
* Data contains 4424 rows.
* Most of the attributes are categorical however were converted to numeric (discrete or binary).
* Target has three categories: dropout, enrolled, and graduate.

The dataset includes information known at the time of student enrollment – academic path, demographic, socioeconomic, macroeconomic factors. The problem is formulated as a three category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course.

The Target variable in the dataset is heavily skewed towards one of the classes, with the 'Graduate' class accounting for around 50% of the records, while 'Dropout' and 'Enrolled' represent 32% and 18% respectively. This imbalance could lead to a situation where the prediction accuracy is heavily influenced by the majority class, resulting in poor performance for the minority classes.

### Data Profiling

As part of the Exploratory Data Analysis, I have performed data profiling using the “ydata\_profiling” library in python to create a profiling report. The profiling report is in \*.html format and due to GitHub’s upload size limitation had been compressed to \*.zip format.

The report is saved in the “Predicting-Student-Dropout-and-Academic-Success” repository at <https://github.com/eugene-kbl/Predicting-Student-Dropout-and-Academic-Success>.

Prior to running the report, I had converted the following variables to categorical data type:

| Marital status | Nacionality | Educational special needs |
| --- | --- | --- |
| Application mode | Mother's qualification | Debtor |
| Application order | Father's qualification | Tuition fees up to date |
| Course | Mother's occupation | Gender |
| Daytime/evening attendance | Father's occupation | Scholarship holder |
| Previous qualification | Displaced | International |

Description of all features in the dataset can be found on The UC Irvine Machine Learning Repository by following the link provided earlier under the Dataset section.

# Literature Search Strategy

This section will contain a description of the search strategy and the criteria used to select applicable studies.

To conduct this literature review, I have conducted a comprehensive search of academic databases such as IEEE Xplore, SpringerOpen, MDPI, ScienceDirect, Google Scholar to select studies applicable to my research question. I focused on articles that utilized machine learning techniques to predict student dropout and academic success.

The following keywords were used in the search:

* predicting student success
* predicting student dropout
* predicting academic performance
* academic success
* academic success factors
* external factors
* machine learning
* classification algorithms

Based on the search and the keywords mentioned above, six studies were selected, namely:

* Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). “Predicting Student Dropout and Academic Success”.
* Albreiki, Zaki, and Alashwal (2022), "A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques".
* Sokkhey and Okazaki (2021), "Hybrid Machine Learning Algorithms for Predicting Academic Performance".
* Malini and Kalpana (2016), "Investigation of factors affecting student performance evaluation using education materials data mining technique".
* Van Hoek, Portzky, and Franck (2019), "The influence of socio-demographic factors, resilience and stress reducing activities on academic outcomes of undergraduate nursing students: A cross-sectional research study".
* Silva, Vautero, and Ussene (2021), "The influence of family on academic performance of Mozambican university students".

# Literature Synthesis

This section will contain the key themes and patterns across the studies, which could be leveraged in answering the research question.

## Key themes

The reviewed studies revealed that machine learning techniques are effective in predicting student dropout and academic success. Albreiki, Zaki, and Alashwal (2022) conducted a systematic literature review of student performance prediction using machine learning techniques. The study analyzed a number of studies that used machine learning techniques to predict student performance. The study found that the most commonly used machine learning algorithms were decision tree, logistic regression, and SVM. The study also identified several factors that affect student performance, including demographics, prior knowledge, class attendance, and socio-economic status.

Realinho et al. (2022) found that prior academic performance, demographics, and socioeconomic factors were significant predictors of student dropout and academic success. Additionally, the study found that decision trees and random forests outperform logistic regression in predicting student dropout and academic success.

Van Hoek et al. (2019) conducted a cross-sectional research study to explore the influence of socio-demographic factors, resilience, and stress-reducing activities on the academic outcomes of Belgian nursing students. They found that resilience and stress-reducing activities were significant predictors of academic outcomes, while socio-demographic factors had less influence.

Silva et al. (2021) investigated the influence of family on academic performance of Mozambican university students. The study found family support and expectations were significant predictors of academic performance, while family income and status had little influence.

Malini and Kalpana (2016) investigated factors affecting student performance using data mining techniques. The dataset used in the study contained demographics, academic history, and performance in courses. The study found that the factors affecting student performance were class attendance, prior knowledge, and academic background.

Sokkhey and Okazaki (2021) investigated the performance of hybrid machine learning algorithms, namely decision tree, random forest, and support vector machine (SVM) in predicting academic performance. They used a dataset containing student data, including demographics, academic history, and performance in courses. The study found that the SVM algorithm outperformed the other two algorithms.

## Key patterns

Based on the identified studies, the most influential factors in predicting students’ performance are past and current academic performance, and socio-demographic data.

Additionally, the most commonly used machine learning algorithms which performed well are decision trees, random forests, and support vector machine (SVM) algorithms.

# Project approach

After reviewing the studies, my plan for the project is to:

1. Prepare the data:
   1. deal with data imbalance - under or over sampling methods.
2. Feature scaling:
   1. standardization.
   2. mean normalization.
   3. min-max scaling
   4. unit vector.
3. Feature selection analysis.
   1. this can help to reduce the complexity of the model and prevent overfitting.
4. ML models:
   1. decision trees.
   2. random forest.
   3. support vector machine.
5. Model evaluation:
   1. confusion matrix.
   2. accuracy
   3. precision and recall
   4. f1 - score
   5. ROC curve
   6. AUC

# Model Evaluation Metrics

Confusion Matrix: is a table that summarizes the performance of a classification model by comparing the predicted labels with the actual labels.

Accuracy: is the ratio of the number of correct predictions to the total number of predictions. It is one of the simplest and most commonly used metrics for evaluating classification models.

Precision and Recall: these metrics are used to evaluate the performance of a binary classification model. Precision measures the fraction of true positives out of all predicted positives, while recall measures the fraction of true positives out of all actual positives.

F1-score: is a weighted average of precision and recall, with equal weights assigned to both metrics.

ROC Curve: is a graphical representation of the performance of a binary classification model. It plots the true positive rate against the false positive rate at different classification thresholds.

AUC: is the area under the ROC curve and provides a single scalar value that summarizes the overall performance of a binary classification model.

# References

1. 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
2. Albreiki, B., Zaki, N., & Alashwal, H. (2022). A Systematic Literature Review of Student’ Performance Prediction Using Machine Learning Techniques. Education Sciences, 11(9), 552. https://doi.org/10.3390/educsci11090552.
3. Sokkhey, P., & Okazaki, T. (2021). Hybrid Machine Learning Algorithms for Predicting Academic Performance. Information, 12(2), 59. doi: 10.3390/info12020059.
4. Malini, J., & Kalpana, Y. (2016). Investigation of factors affecting student performance evaluation using education materials data mining technique. Journal of Intelligent Learning Systems and Applications, 8(4), 31-43. doi: 10.4236/jilsa.2016.84004.
5. Van Hoek, G., Portzky, M., & Franck, E. (2019). The influence of socio-demographic factors, resilience and stress reducing activities on academic outcomes of undergraduate nursing students: A cross-sectional research study. Nurse Education Today, 80, 68-75. doi: 10.1016/j.nedt.2019.06.019.
6. Silva, A. D., Vautero, J., & Ussene, C. (2021). The influence of family on academic performance of Mozambican university students. Journal of Further and Higher Education, 1-15. https://doi.org/10.1080/0309877X.2021.1924293