

**Dilkaran Dhaliwal**

**500971670**

**Supervisor’s Name: Tamer Abdou**

**CIND 820**

**Date: June 23, 2025**

**Literature Review, Data Description, and Approach**

Table of Contents

**Problem Statement and Context......................................................................................3**

**Research Questions and Justification............................................................................3**

**Dataset Selection and Rationale......................................................................................4**

**Proposed Methodology and Tools...................................................................................4**

**Literature Review...............................................................................................................5**

**Data Description................................................................................................................6**

**Project Approach..............................................................................................................9**

**Visual Representation......................................................................................................10**

**GitHub Repository............................................................................................................10**

**References Page...............................................................................................................11**

**Problem Statement and Context**

It is common to assume that post-secondary students around the globe can have vastly different experiences attending school from one another. It is sometimes forgotten that this can also be the case when students are attending the same school as one another. Factors inside and outside of school hours can shape a student's future and their young adulthood. The problem here is how post-secondary institutions can help stop students from dropping out of their studies so that they can transition to their young adult lives successfully. In Canada, the OECD has seen a trend of mathematics and reading scores going down in 15-year-olds in 2022 when compared to 2018, and we scored the lowest averages in math, reading and science than in all previous assessments (OECD 2023). This downward trend will follow these students into post-secondary education, where the dropout rate is turning into a concerning issue. The theme of classification and regression is chosen, and I believe this field is very important for people to understand, as these years in post-secondary education can be the stepping stone for a student's successful future, and we need to have a better understanding of how to improve their time in university.

**Research Questions and Justification**

The first research question would be related to the usage of evening courses compared to daytime courses. Do Evening courses have a considerable difference in the dropout rate compared to daytime courses in higher education? This would be an interesting question to answer, as it would give universities an insight into the possibility of adding more evening courses to help their students complete their education. Currently, daytime courses are offered at a much higher rate than evening courses, but that is not set in stone and can be changed semester to semester. This research question can help universities with getting the right ratio of daytime to evening courses. The second research question would be looking at the age of admission and its correlation to the dropout rate. Looking deeper into this correlation could help the universities when it comes to supporting the age groups that need the most help. This question can also be linked back to our first research question, as older students who are returning to university studies would prefer evening courses due to them working during the day. These two might be correlated, and performing an analysis may help see if that is true. The third research question would be looking at the parents of the students and their education level. If a student's parents are educated, would that help the student when it comes to staying in university? We would also look to see if the father's or mother's education has more of an impact on the dropout rate of the student.

**Dataset Selection and Rationale**

Dataset: <https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success>

DOI: 10.24432/C5MC89

The dataset being used for this study is from a Portuguese university. The target attribute being assessed is whether the student graduated, dropped out or is currently enrolled. Other attributes include the program they enrolled in and at what age, the occupation and education level of their parents, whether they are enrolled in evening or daytime courses and more. This data set may not be set in North America, but it can still give good insight into the post-secondary experience that these students had. Portugal has a culture of caring about their education as they have invested heavily in their public schooling system. In 2014, 5.1% of its gross domestic product (GDP), was devoted to financing pre-primary, primary and secondary education. This is 1 percentage point higher than its southern European peers in Italy and Spain (Liebowitz, David, David, et al). This data set was chosen as there are 36 features/factors, giving us plenty of information to look at. A limitation of this dataset is the location where it was collected. Portugal may have a different university climate from Canada, but the findings of this research could help Canadian universities lower their dropout rate and increase retention.

**Proposed Methodology and Tools**

The tool that we will be using is Python. Using Python libraries, we can make classification models and predictive models to see if we can predict a student's performance based on listed factors, both in and out of school hours. These models can be used to answer the research questions, like what factor plays the biggest role when it comes to a student's grade. This can help give the post-secondary institutions more information on what areas to focus on.

**Literature Review**

Research into the student population in universities is not a new field. There have been countless studies looking at the student body and how different factors affect a student's studies on campus. Education is a very important aspect of our lives as it helps mould us into the people we are today. To me, we can never stop improving our education systems, as the world is changing faster than we have seen before, and our education system needs to keep up to help prepare its students to take on the world. A study performed by Matti Vaarma and Hongxiu Li looked at student dropout rates in Finnish higher education. They looked at dropout rates within university degree programs, which is similar to the study that I will be conducting. Their study used three machine learning models, CatBoost (CAT), neural networks (NN) and Logistic regression (LR). What they found was that accumulated credits, the Moodle activity count and the number of failed courses were some of the most important features when it came to predicting if a student would drop out of their university program (Vaarma, Li 2024). A limitation of their study was that they only used demographic and transcript-related information. I plan to look at other non-academic information, such as the education of the student’s parent. That is where my study will differ. Out-of-school factors could have an impact on their chance to graduate from a university program, and I wish to look at that information. Another study by the Higher Education Quality Council of Ontario looked at University completion rates and the labour market outcomes of those who did not graduate. The province has a vested interest in their university graduation rates as graduates earn more on average, are less likely to be unemployed, and have lower student loan default rates. They found that non-completion rates were higher for men, domestic students and those who received federal financial aid (Colyar, Chatoor, Deakin 2023). Through a Regression analysis, they found that the earning gap was largest among the students who spent the least amount of time in post-secondary education. For example, a student who dropped out in their third year earned more than someone who dropped out in their first year. Ontario boasts about its higher post-secondary participation rates, but nearly one-quarter of students did not graduate after eight years, and further research is needed to help institutions, and the government improve those numbers. More sociodemographic factors associated with dropping out of university need to be studied, which is what my study will be looking at. A study in Greece was done to see to what extent student dropout is due to intrinsic (related to the student) factors and to what extent it is due to extrinsic (non-student) factors. This study was done at a Greek distance education university and focused on undergraduate and graduate students. This study was done through a voluntary survey, and the four main reasons for the students to drop out of their studies were professional, family/personal, health and academic. This university was comprised of older students who were coming back for education or continuing their education for postgraduate studies. The findings of this report may not directly apply to traditional universities that have mainly younger students. My research will look at a mix of older and younger students.

**Data Description**

The data set that will be used for this research consists of 36 features and 4424 instances, with each instance representing 1 student. The data was selected from a university in Portugal. This data set combines demographic data, socio-economic factors and academic performance information that will be used to help understand the factors that play a part in making a student drop out of university. 9 university courses are being looked at, and some are offered in the evening. Below is a breakdown of each feature that was recorded for each instance. There is no missing data.

Data EDA Report: <https://github.com/Dilkarand/CIND820-Big-Data-Analytics-Project/blob/266c7fa63ce4bb4de25ada19f579080c9812b5a7/Student_Analytics_Project_EDA.html>

|  |  |
| --- | --- |
| **Marital status** | The marital status of the student. (Categorical) |
| **Application mode** | The method of application used by the student. (Categorical) |
| **Application order** | The order in which the student applied. (Numerical) |
| **Course** | The course taken by the student. (Categorical) |
| **Daytime/evening attendance** | Whether the student attends classes during the day or in the evening. (Categorical) |
| **Previous qualification** | The qualification obtained by the student before enrolling in higher education. (Categorical) |
| **Nationality** | The nationality of the student. (Categorical) |
| **Mother's qualification** | The qualification of the student's mother. (Categorical) |
| **Father's qualification** | The qualification of the student's father. (Categorical) |
| **Mother's occupation** | The occupation of the student's mother. (Categorical) |
| **Father's occupation** | The occupation of the student's father. (Categorical) |
| **Displaced** | Whether the student is a displaced person. (Categorical) |
| **Educational special needs** | Whether the student has any special educational needs. (Categorical) |
| **Debtor** | Whether the student is a debtor. (Categorical) |
| **Tuition fees up to date** | Whether the student's tuition fees are up to date. (Categorical) |
| **Gender** | The gender of the student. (Categorical) |
| **Scholarship holder** | Whether the student is a scholarship holder. (Categorical) |
| **Age at enrollment** | The age of the student at the time of enrollment. (Numerical) |
| **International** | Whether the student is an international student. (Categorical) |
| **Curricular units 1st sem (credited)** | The number of curricular units credited by the student in the first semester. (Numerical) |
| **Curricular units 1st sem (enrolled)** | The number of curricular units enrolled by the student in the first semester. (Numerical) |
| **Curricular units 1st sem (evaluations)** | The number of curricular units evaluated by the student in the first semester. (Numerical) |
| **Curricular units 1st sem (approved)** | The number of curricular units approved by the student in the first semester. (Numerical) |
| **Curricular units 1st sem (grade)** | Grade average in the 1st semester (between 0 and 20) (Numerical) |
| **Curricular units 1st sem (without evaluations)** | Number of curricular units without evaluations in the 1st semester (Numerical) |
| **Curricular units 2nd sem (credited)** | Number of curricular units credited in the 2nd semester (Numerical) |
| **Curricular units 2nd sem (enrolled)** | Number of curricular units enrolled in the 2nd semester (Numerical) |
| **Curricular units 2nd sem (evaluations)** | Number of evaluations of curricular units in the 2nd semester (Numerical) |
| **Curricular units 2nd sem (approved)** | Number of curricular units approved in the 2nd semester (Numerical) |
| **Curricular units 2nd sem (grade)** | Grade average in the 2nd semester (between 0 and 20) (Numerical) |
| **Curricular units 2nd sem (without evaluations)** | Number of curricular units without evaluations in the 1st semester (Numerical) |
| **Unemployment rate** | Unemployment rate (%) (Numerical) |
| **Inflation rate** | Inflation rate (%) (Numerical) |
| **GDP** | GDP (Numerical) |
| **Target** | The problem is formulated as a three-category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course (Categorical) |

**Project Approach**

To reiterate the three research questions, they are:

1. Do Evening courses have a considerable difference in the dropout rate compared to daytime courses in higher education?
2. Would the age of the student at enrolment have a considerable impact on the probability of dropping out of higher education? Are younger students or older students more at risk of dropping out?
3. How significant is the level of education for a student's parents in their chances of dropping out of higher education? Does the father’s or mother’s education have more of an impact?

When it comes to the models to use to help answer these research questions, I will look at other related studies to see what are the most common. Anaile Rabelo and Luis Zarate found 62 articles related to the problem of student dropouts. Out of these articles, they noted the machine learning techniques that were most used. In the table below, it can be seen that the most used techniques are Decision tree, Naive Bayes and k-nearest neighbour (Rabelo & Zárate, 2025). To answer my research questions, I would be using the Decision Tree and K-nearest Neighbour machine learning techniques.

|  |  |  |
| --- | --- | --- |
| **Techniques** | **Quantity** | **Percentage (%)** |
| Decision tree | 35 | 19.9 |
| Naïve Bayes | 30 | 17.0 |
| K-nearest neighbours | 17 | 9.7 |
| Logistic regression | 17 | 9.7 |
| Random forest | 17 | 9.7 |
| Neural networks | 16 | 9.1 |
| Support vector machine | 15 | 8.5 |
| Rules | 8 | 4.5 |
| Gradient boosting machine | 5 | 2.8 |
| Fuzzy | 3 | 1.7 |
| Others | 13 | 7.4 |
| Total | 176 | 100 |

**Visual Representation**

A computer screen with a diagram

AI-generated content may be incorrect.

**GitHub Repository**

<https://github.com/Dilkarand/CIND820-Big-Data-Analytics-Project>

**References**

Colyar, J., Chatoor, K., & Deakin, J. (2023). Linking postsecondary non-completion rates and

labour market outcomes. Higher Education Quality Council of Ontario. <https://heqco.ca/wp-content/uploads/2023/11/Linking-Postsecondary-Non-completion-Rates-and-Labour-Market-Outcomes-in-Ontario-FINAL.pdf>

Liebowitz, D., González, P., Hooge, E., & Lima, G. (2018). OECD reviews of school resources: Portugal 2018. In OECD Reviews of School Resources. OECD Publishing. <https://doi.org/10.1787/9789264308411-en>

Mitra, R., Habib, K. N., Siemiatycki, M., Keil, R. and Bowes, J. (2020) StudentMoveTO - From Insight to Action on Transportation for Post-Secondary Students in the GTHA: 2019 Transportation Survey Findings. Web portal: [www.StudentMoveTO.ca](http://www.studentmoveto.ca)

OECD (2023), PISA 2022 Results (Volume I and II) - Country Notes: Canada, OECD

Publishing, <https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/canada_901942bb-en.html>

OECD (2023), PISA 2022 Results (Volume I and II) - Country Notes: Portugal, OECD

Publishing, <https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/portugal_777942d5-en.html>

Pierrakeas , C., Xenos, M., Panagiotakopoulos , C., & Vergidis, D. (2004). A Comparative Study of Dropout Rates and Causes for Two Different Distance Education Courses. Hellenic Open University. <https://www.irrodl.org/index.php/irrodl/article/view/183/265>

Rabelo, A. M., & Zárate, L. E. (2025). A model for predicting dropout of higher education students. Data Science and Management, 8(1), 72–85. <https://doi.org/10.1016/j.dsm.2024.07.001>

Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict Students' Dropout and Academic Success [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5MC89>.

Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. Technology in Society, 76, 102474. <https://doi.org/10.1016/j.techsoc.2024.102474>