A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Marcelo Fernando Sotomayor Palacios |
| *Student Number* | 2024043 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA 2 Capstone Report |
| *Assessment Due Date* | 19th May 2024 23:59 |
| *Date of Submission* | 19th May 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Content

[Introduction 3](#_Toc166894057)

[Objectives 3](#_Toc166894058)

[Problem Definition 3](#_Toc166894059)

[Project management and planning 4](#_Toc166894060)

[Project plan, timeline, and execution 4](#_Toc166894061)

[Data Resource 5](#_Toc166894062)

[Ethical considerations 6](#_Toc166894063)

[References 6](#_Toc166894064)

# Introduction

The educational context in schools and universities worldwide faces several challenges related to its intrinsic nature, working with people leading their complex learning process involves confronting an array of obstacles, one of the most important is the precise and timely detection of problems in the learning process.

The possibility of predicting academic outcomes in advances with the development and applications of data analytics and machine learning techniques is certainly an enormous contribution to making the work burden and academic load easier[[1]](#footnote-1), helping in this way to the big challenge of taking education to the next evident level, a level that dialogues with technological advance.

Finally, in this project, using databases from the educative context from universities of Chile and Ireland, predictor variables (independent) will be chosen to train a model and generate a student's academic success outcome as success (approval of the course), reprove (failure the course) or even drop out the course.

## Objectives

The hypothesis presented for this project is that an academic performance predictive and dynamic model based on data analytics and machine learning might help teachers and educational institutions to predict the academic success of students (approval or reproval), and in that way contribute to making more accurate and on time academic decisions.

1. Develop an Academic Performance Predictive and dynamic model based on academic and social databases from educational institutions that could predict students' academic success.
2. Explore and analyze different variables involved in students' academic performance.
3. Test different predictive models and select the one that best suits predicting academic success according to the selected variables.

## Problem Definition

According to the Organization for Economic Co-operation and Development (OECD), “There are fewer than 28 pupils per class in all the countries with available data, except Chile with 31 pupils. At the lower secondary level, the average class size in OECD countries is 23 students.”[[2]](#footnote-2) This means that teachers should supervise, manage, and guide the complex learning process of at least one entire course in an ideal case. Still, the reality shows a different scenario, where teachers are facing the educational pathway of at least four entire courses, this implies a huge amount of data coming from the student, like assessment marks, academic backgrounds, family context, socioeconomic context plus the inherent learning process of each student.

This reality creates a scenario where educational institutions can not properly follow the learning progress of the pupils, for example, by giving accurate feedback or preparing appropriate assessments and jobs for each student.

Therefore, the important ability to predict if the students are on the way to academic success or not [[3]](#footnote-3)on time becomes an important tool that is difficult to handle due to the circumstances of educational reality.

An academic performance predictive and dynamic model (APPM) based on data analytics and machine learning, which could incorporate the academic and social data from students and predict if the student is on the way to success (approval of the course), reprove (failure the course) or even drop out the course could be a strong tool for teachers and institutions on the mayor challenges of making early academic decisions [[4]](#footnote-4)

# Project management and planning

## Project plan, timeline, and execution

Firstly, this project aims to produce an academic performance predictive and dynamic model based on data analytics and machine learning, which means the selection of several social and educational features that may be connected or not with the academic success of university students, for instance; Age, academic background, previous Laboral experience, motivation, parental economic support, education level of parents. The boundaries of these features, which in this case will be the independent variable or predictor, will be given by the dataset that can be obtained, however, the more features can be analyzed the more possibilities to find out any relation with the target variable, which will be the academic success of the students, this means that the limits of this project will be given by de availability of data and the selection of relevant independent variables.

To achieve this objective, the study will be framed on university student databases from 2 countries members of the Organization for Economic Co-operation and Development (OECD), Ireland and Chile, two countries with a high development score according to the Human Development Index (HDI), which measures it in a scale from 0 to 1, Chile has the best HDI score in south America with 0,860 [[5]](#footnote-5), and Ireland [[6]](#footnote-6) 0,945. The idea of working with data from these two countries will let to have an open view and avoid bias that might be associated with the European context, including the South American context.

The methodology thought for this project is the implementation of the CRISP-DM. A model that has been used for more than twenty in a wide array of fields, but above all in data analysis and machine learning [[7]](#footnote-7), its flexibility and iteration will allow it to move from a basic project with a limited number of variables to another with larger perspectives.

The project plan will have the structure as follows:

Figure 1.

After a first circle applying the CRISP-DM model, it is expected to obtain outcomes of academic success prediction in university students of Ireland and Chile with an acceptable level of accuracy, using some the machine learning models like random forest and linear regression, which should be useful for prediction and classification of the target variable.

## Challenges and solutions

Challenge: Find complete datasets with social and educational information/ solution start with small datasets but more complete

The kind of information per country is different / how standardized the information, maybe the models are not comparables.

# Data Analysis and Preprocessing

• Exploratory Data Analysis (EDA)

• Descriptive statistics

• Visualisations

• Initial findings of possible patterns, anomalies or insights from the data.

• Data/data processing/feature engineering has been explained correctly with the aid of visualisation tools and metrics.

# Machine Learning Implementation and Evaluation

• Algorithm selection and application

• Model development and validation.

• Methodology is sensible and explained appropriately.

• Model has been discussed.

• Code works

# Data Resource

The main data sources selected for this project are:

1. Open data website of the educational ministry of Chile[[8]](#footnote-8), which contains a wide platform with datasets of teachers and educational assistants, students, and educational institutions, this source is a public access source.
2. Educational Dataset contained on the website of data.gov.ie operated by the Government Reform Unit in the Department of Public Expenditure and Reform. Which is free to use[[9]](#footnote-9)

# Ethical considerations

Firstly, is vital to know that our target group will be students, even if the focus group is going to be university students, that does not mean that some of them could be under 18 years old, then it is important to consider it in case needs any special permission.

Secondly, as the aim is to generate a prediction on academic success, it is relevant that this outcome works as early advice for teachers and students to generate improvements and assistance and not as a source of discrimination and selection for the institutions.

# References

1. Attaran, M., Stark, J. and Stotler, D. (2018). Opportunities and challenges for big data analytics in US higher education. Industry and Higher Education, 32(3), pp.169–182. doi: https: //doi.org/10.1177/0950422218770937.
2. OECD iLibrary (2024).| OECD iLibrary. [online] www.oecd-ilibrary.org. Available at: https://www.oecd-ilibrary.org/sites/e2f6a260-en/index.html?itemId=/content/component/e2f6a260-en. [Accessed 29 Mar. 2024].
3. Lye, C.-T., Ng, L.-N., Hassan, M.D., Goh, W.-W., Law, C.-Y. and Ismail, N. (2010). Predicting Pre-university Student’s Mathematics Achievement. Procedia - Social and Behavioral Sciences, 8, pp.299–306. doi: https://doi.org/10.1016/j.sbspro.2010.12.041.
4. Singh, H.P. and Alhulail, H.N. (2022). Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach. IEEE Access, 10, pp.6470–6482. doi: https://doi.org/10.1109/access.2022.3141992.
5. Nations, U. (2020). *Human Development Report 2023*. United Nations.
6. Nations, U. (2024). Data Center. [online] hdr.undp.org. Available at: https://hdr.undp.org/data-center/human-development-index1. [Accessed 29 Mar. 2024].
7. Schröer, C., Kruse, F. and Gómez, J.M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, [online] 181(1), pp.526–534. Available at: <https://www.sciencedirect.com/science/article/pii/S1877050921002416>.
8. Ministerio de educacion MINEDUC (n.d.). Datos Abiertos. [online] datosabiertos.mineduc.cl. Available at: <https://datosabiertos.mineduc.cl/>.
9. Government Reform Unit in the Department of Public Expenditure and Reform. (n.d.). Datasets - data.gov.ie. [online] data.gov.ie. Available at: https://data.gov.ie/dataset?tags=education [Accessed 29 Mar. 2024].

**GitHub link:** https://github.com/CCT-Dublin/capstone-project-feb-2024-ft-msotomayorpal/commits/main/

1. Attaran, M., Stark, J. and Stotler, D. (2018). Opportunities and challenges for big data analytics in US higher education. Industry and Higher Education, 32(3), pp.169–182. doi:https://doi.org/10.1177/0950422218770937. [↑](#footnote-ref-1)
2. OECD iLibrary (2024).| OECD iLibrary. [online] www.oecd-ilibrary.org. Available at: https://www.oecd-ilibrary.org/sites/e2f6a260-en/index.html?itemId=/content/component/e2f6a260-en. [Accessed 29 Mar. 2024]. [↑](#footnote-ref-2)
3. Lye, C.-T., Ng, L.-N., Hassan, M.D., Goh, W.-W., Law, C.-Y. and Ismail, N. (2010). Predicting Pre-university Student’s Mathematics Achievement. Procedia - Social and Behavioral Sciences, 8, pp.299–306. doi: https://doi.org/10.1016/j.sbspro.2010.12.041. [↑](#footnote-ref-3)
4. Singh, H. P. & Alhulail, H. N. (2022). Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach. *IEEE Access*, Digital Object Identifier: 10.1109/ACCESS.2022.3141992. [↑](#footnote-ref-4)
5. Nations, U. (2020). Human Development Report 2020. United Nations. [↑](#footnote-ref-5)
6. Nations, U. (2024). Data Center. [online] hdr.undp.org. Available at: https://hdr.undp.org/data-center/human-development-index1. [Accessed 29 Mar. 2024]. [↑](#footnote-ref-6)
7. Schröer, C., Kruse, F. and Gómez, J.M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, [online] 181(1), pp.526–534. Available at: https://www.sciencedirect.com/science/article/pii/S1877050921002416. [↑](#footnote-ref-7)
8. Ministerio de educacion MINEDUC (n.d.). Datos Abiertos. [online] datosabiertos.mineduc.cl. Available at: https://datosabiertos.mineduc.cl/. [↑](#footnote-ref-8)
9. Government Reform Unit in the Department of Public Expenditure and Reform. (n.d.). Datasets - data.gov.ie. [online] data.gov.ie. Available at: https://data.gov.ie/dataset?tags=education [Accessed 29 Mar. 2024]. [↑](#footnote-ref-9)