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 3 Capstone Project |
| *Assessment Due Date* | 10th November 2024 23:59 |
| *Date of Submission* | 15th November 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](#_Toc182582366)

[Objectives 3](#_Toc182582367)

[Problem Definition 3](#_Toc182582368)

[Project management and planning 4](#_Toc182582369)

[Project plan, timeline, and execution 4](#_Toc182582370)

[Project Plan 5](#_Toc182582371)

[Project Timeline 5](#_Toc182582372)

[Challenges and solutions 7](#_Toc182582373)

[Data Analysis and Preprocessing 7](#_Toc182582374)

[Exploratory Data Analysis (EDA) 7](#_Toc182582375)

[Visualizations 9](#_Toc182582376)

[Machine Learning 13](#_Toc182582377)

[K-NN Regression Model. 13](#_Toc182582378)

[Splitting Data into Training and Testing 14](#_Toc182582379)

[K-NN regressor Hyperparameter Tunning (GridSearchCV) 14](#_Toc182582380)

[K-NN Regressor scores and results 15](#_Toc182582381)

[Cross Validation for KNN regression. 15](#_Toc182582382)

[Linear Regression Hyperparameter Tunning (GridSearchCV) 16](#_Toc182582383)

[Linear Regression Scores and Results. 16](#_Toc182582384)

[Cross Validation for Linear regression. 17](#_Toc182582385)

[Random Forest Hyperparameter Tunning (GridSearchCV) 18](#_Toc182582386)

[Random Forest Regressor Scores and Results 19](#_Toc182582387)

[Cross Validation for Random Fores Regressor 20](#_Toc182582388)

[Conclusions, Recommendations, and future work 21](#_Toc182582389)

[Final conclusions and future work 24](#_Toc182582390)

[References 24](#_Toc182582391)

# 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 educational context of schools, predictor variables (independent) were chosen to train 3 different models that predicts students’ grades, so then is possible to classify the expected academic outcome into different categories like high, medium, low.

The main idea is helping teacher, administrative and directors of school to apply early assistance to those students that may need additional support or early intervention for those with high grades to enhance their goals.

## Objectives

1. Develop an Academic Predictive model based on academic, social and contextual database from educational institutions that could forecast their grades future semesters.
2. Explore and examine different variables involved in students' academic performance, such as a family context, socioeconomic situation, distance from the school and non-academic activities.
3. Compare the performance of 3 different machine learning models and select the one with highest accuracy predicting academic grades according to the selected variables.
4. Present the results clearly and accessible for educator and administrators that are not familiar with data analytics.

## 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 model 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 or school student databases from a c countries members of the Organization for Economic Co-operation and Development (OECD), the options until this moment are Portugal, Ireland, or Chile, 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), Ireland [[6]](#footnote-6) 0,945 and Portugal[[7]](#footnote-7) 0.866. Working with data from these countries will let to have an open view and avoid bias that might be associated with the European context, including the South American context.

## Project Plan

The project plan will be managed with 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 [[8]](#footnote-8), 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.

## Project Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| Step | Date | Activity | progress |
| 1. Business Understanding | 9.03.2024 until 29.03.2024 | Business understanding and objectives:   * Identification of factors that may affect academic outcomes of students in some the these 3 countries, Portugal, Ireland, or Chile | Done |
| Step 2. Data Understanding | 15.04.2024 until 06.05.2024 | Identified and collected accurate data from at least one educative institution in Ireland, Chile or Portugal. | Collected data from a School in Portugal, visualization, and EDA are done |
| Step 3. Data Preparation | 06.05.2024 until 12.05.2024 | Data cleaning and feature engineering.  Detecting, removal, or imputation of missing values and duplicate values.  Encoding of Categorical data | All the categorical data was encoded.  Data cleaning done. |
| Step 4. Models | 06.05.2024 until 12.05.2024 | Choose at least one machine-learning model.  Split data in training and testing.  Apply the model.  Use cross-validation to check the performance of the model. | k-NN regressor  Linear Regressor  Done. |
| Step 5. Evaluation | 12.05.2024 until 19.05.2024 | Evaluation of the performance of the models.  Interpretation of the results.  Look for possible improvements.  Settle of possible new objectives | Done |

Fig. 2

After the first circle applying the CRISP-DM model, it is expected to obtain outcomes of academic success prediction in at least one education institution of one of these 3 countries, in this case Portugal, with an acceptable level of accuracy, using some of 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

|  |  |  |
| --- | --- | --- |
| Challenges and solutions | | |
| Step | Challenge | Solution |
| 2 | Find complete datasets with social and educational information, most of them are large datasets with general worldwide information, like OCDE or PISA. | Start with small datasets but more comprehensive datasets. |
| 2 | The kind of information per country differs, so how do we standardize the information? | Maybe different models will be necessary for other countries, therefore could be better to start analysing one educative institution and expand the project to other countries. |
|  |  |  |
|  |  |  |
|  |  |  |

Fig. 3

# Data Analysis and Preprocessing

### Exploratory Data Analysis (EDA)

Dataset source[[9]](#footnote-9), The data set selected for this project includes socio-educational features as well as grades of a group of 649 high school students from two different schools in Portugal in the subject of the Portuguese language. The data set contains 31 features and 649 observations, some of which are: Categorical, binary, Categorical nominal, Categorical nominal, Numerical Features (Fig 4).

Texto

Descripción generada automáticamente con confianza baja

Fig 4

Some relevant appreciation about the kind of features: there an important diverse on the features characteristics, for examples there is the feature Commute time which is the travel time from home to school (time intervals: Up to 15 min, 15 to 30 min, 30 min to 1h, More than 1h,) this information tell us about the geographical context and the effort that some students needs to do in order to arrive to the school, then other feature which is School absences: Number of school absences (numeric: from 0 to 93) tells about responsibility , there are others related with the familiar context like the feature “Good Family Relationship”: Quality of family relationships (categorical: from 1 - very bad to 5 - excellent) and “Family size”: Family size (char: 'up to 3' - less than or equal to 3 or 'Above 3' - greater than 3).

In the other hand there 2 variables which are strict academic which are G1 and G2, this are the G1 - first period grade (numeric: from 0 to 20) and G2 - second period grade (numeric: from 0 to 20) in the subject of Portuguese, G2 will be the dependant variable to be predict with the 3 different models.

Initially, the composition of the columns of the dataset was made of 17 features classified as objects and 14 as entire numbers; to prepare the data for the machine learning step, it was necessary to transform the categorical features into numerical ones (Fig5)

Texto

Descripción generada automáticamente

Fig 5

Afterward, the statistics of some features were checked, this point is relevant to mention the mean of grades of the first and second semesters, 11.39 and 11.57 (Fig 6).

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Fig 6

Fig 6 shows several important statistics from numerical features of the dataset, for example median of grades in first and second semester are quite similar 11,4 and 11,6 each one, but it is also noticeable that the range of grades is wide, for example G1 has grades with a maximum of 20 and with minimum of 0, it implies that there is a huge academic performance variation.

Other features like the free time and absence shows high variation as well, This is relevant in the way that it that the student population, behaviours and lifestyle are highly diverse.

## Visualizations

One of the features that I consider relevant to visualize due to its potential as a predictor is the grades of the first semester, a boxplot (Fig 7) is useful to check outliers.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Fig 7. First Semester Grades Boxplot.

In this figure, it is possible to appreciate the presence of outliers, while the median is 11 and the mean is 11,40, there are some values over 17.5, probably the students with a better score, and some of them have 0, although these values are outliers, are also important information as the objective is the prediction of scores and the early detection of students in risk of dropout and students with high probability of success.

Then I did a histogram to visualize the frequency of these scores (Fig 8)

Gráfico, Histograma

Descripción generada automáticamente

Fig 8. First semester scores Histogram.

One of the facts that is possible to extract, the distribution is slightly skewed to the right. the number of students in the outlier’s area is not that much, however relevant information, at this point, a question appears, does this distribution become more skewed in the second semester? But why this point is important, the educational process is not instantly things, and tendencies might be more pronounced over time.

Gráfico, Histograma

Descripción generada automáticamente

Fig 9. Second-semester grades histogram.

As can be seen in Figure 9, the number of outlier students increased, it means that the academic differences are bigger, maybe the explanation for this is inside the other features, but with these first Initial findings and insights into the data, Second-semester scores can be defined as target variable for further analysis.

However, 31 features can be better visualized with a heatmap (Fig 10).

Gráfico, Gráfico de barras

Descripción generada automáticamente

Fig 10. Heatmap

According to the Heatmap, G1, and G2 are highly correlated, let's see these features in a scatterplot Fig 11.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Fig 11 Scatterplot

There is a linear correlation between first and second-semester grades.

To conclude this section, the insight extracted from the data suggests that a good model would be a regression model, because of the correlation between some of the features.

# Machine Learning

## K-NN Regression Model.

The target variable selected for this analysis is “Second-semester grades”, this is a numerical value correlated with some other features, and due to this reason, a regression model will be implemented.

The first model chosen to be evaluated is the k-NN regressor. The first step was to choose the best number of neighbours to get an accurate performance Fig 12.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

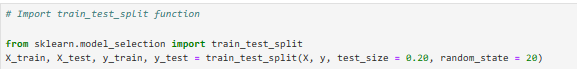
Fig 12. The best number of neighbours to get accurate results.

Fig 12 is showing the relation between accuracy against the number of neighbour for a K-NN regressor model, orange line which shows the training data, shows that when the model has few neighbours, the accuracy is high, this is because of the overfitting, but when the number of neighbours increases, accuracy decreases, because the model is getting more general as the it is considering more neighbourhoods.

Moreover, the blue line shows the testing accuracy of the model along the number of neighbours, this accuracy reaches a peak between 6 and 11 neighbours apparently, after this point the accuracy of both lines it just decreases.

According to Figure 12, to get a good performance in the model, approximately to 10 neighbours should be chosen.

## Splitting Data into Training and Testing

Fig 13. Splitting Data into Training and Testing

The data was training and testing (Fig 13), with a testing percentage size of 20%. This means the 80 percentage of the data will be used for training the machine learning models and the other 20 percent for testing it, is important to consider that this step is relevant because doing this, the models might avoid overfitting, in other word there is space for new data to be testing and the models can generalize the information, which exactly what it is needed If machine learning models wants to be used for predictions with new data.

Interfaz de usuario gráfica, Texto, Aplicación, Sitio web

Descripción generada automáticamenteFig 14. Shape of data after splitting.

## K-NN regressor Hyperparameter Tunning (GridSearchCV)

To find the exact parameters to have the best K-NN model, GridSearchCV function was used (Fig 14).

Three parameters were evaluated to tunning the model, first one is the number of neighbours, Grid Search examinate between 3 and 10 neighbours, even though this parameter was analysed with the previous K-NN accuracy graph, here a most precise number capture which is 7. This is the point where there is a balance between an overfitting model and a too generalized one.

The second parameter was weights, which determines how big is the influence of each neighbour for the model, the outcome was weights = Uniform, it means that the model treats each neighbour in the same way, each neighbour affects the same the prediction[[10]](#footnote-10).

The third parameter examined is the distance metric p, this metric is used to get nearest neighbours, the metric selected was p = 1, Manhattan distance which is different from Euclidean distance because it adds the absolute distance in each coordinate instead of the straight line distance that Euclidian gets.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamenteFig 14. Best parameters for K-NN Regressor after GridSearchCV applied.

Finally, and in combination with GridSearchCV I have integrated a Pipeline, this function helps to do the workflow more organized, also helps to avoid data missing and miss lectures.

## K-NN Regressor scores and results

Texto

Descripción generada automáticamenteIn the next Fig 15, training and testing scores are showing for the K-NN model, the r2 Score for the training set is 0.72 and for testing set is 0.82. This result means that the model generalizes well to data outside the training set, and also both scores are quite hight, it means that the model can generalize prediction, in conclusion the model is neither overfitting nor underfitting, therefore, the initial result is promising.

Fig 15. Training and testing scores.

We can see some results of the predicted values versus the actual values in the following Figure 16.

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamenteFigure 16. Actual values and predicted values.

## Cross Validation for KNN regression.

Continuing with the study with this model, a Cross Validation was done, the number of folds used was 10 this means that the data was split in 10 parts, so then from the data, training and testing scores were got with 9 of these 10 parts each time, each time with different parts doing 10 iterations of the data analysis.

The results (Fig 17) showed that the average of the accuracy for this model 0,66, which is the precision of the model to predict. However, some results during the iteration of the cross validations showed low scores values and other hight scores, this could imply that the model is not complete stable. In other words, the accuracy for K- NN is acceptable, but with many variations within the data.

Texto

Descripción generada automáticamenteFig 17. Cross Validation accuracy K-NN model.

## Linear Regression Hyperparameter Tunning (GridSearchCV)

Similarly with the k-NN model, GridSearchCV was used together with a pipeline to find the best model (Fig 18) In this case, the parameter being evaluated was whether there was an interception point

Gráfico

Descripción generada automáticamente

Fig 18, Best model Linear Regression.

## Linear Regression Scores and Results.

In the next Fig 19, training and testing scores are showing for the Linear Regression model, the r2 Score for the training set is 0.72 and for testing set is 0.84, These scores indicate that the model generalizes well to data outside the training set, the testing score is relatively high and better than the training score, this outcomes indicates that the model can make reliable predictions on new data without signs of overfitting or underfitting. Therefore, these initial results are promising and indicate a well-balanced model.



Fig 19. Training and testing scores.

Additionally, we can see some results of the predicted values versus the actual values in the following Figure 20.

Tabla

Descripción generada automáticamente con confianza mediaFig 20. Linear Regression actual values and predicted values.

## Cross Validation for Linear regression.

Texto

Descripción generada automáticamente

Fig 21. Cross Validation accuracy Linear Regression.

The results (Fig. 19) show that the average accuracy for this Linear Regression model is 0.6731, we can notice significant variation across the folds with minimum values of 0.1934 and maximum values of 0.9086, this means some inconsistency when the model is predicting grades.

Finally comparing cross validation average accuracy with R² of the training set, we can notice that the training set score is higher (0.72), the model performance is good enough for the educational context, but it cannot capture all the signal to get a better score.

## Random Forest Hyperparameter Tunning (GridSearchCV)

Texto

Descripción generada automáticamenteFinally, the last model to be applied was Random Forest, then, to find the exact parameters to have the best Random Forest, GridSearchCV function was used for (Fig 22).

Fig 22. Best parameters for Random Forest Regressor after GridSearchCV applied.

The parameters tested to optimize the performance of the model were 5.

Firstly, max depth was set on 5, it is related with the deep of each one of the trees, the meaning of this is parameter is the deeper the trees the more complex the system, and a system to complex could face risk of overfitting, so limiting the max depth overfitting is being avoiding.

The second parameter was Max Features, the options selected was ‘SQRT’ which means that each tree is going to use square root of total number of feature when it splits nodes.

The third parameter Minimum Samples per Split, the number chosen was 5 which is the limit of splits, therefore the model is not becoming overly complex, avoiding in this way overfitting.

The fourth parameter was Minimum Samples per Leaf, this parameter was set to 1, indicating that each leaf node can contain as few as one sample, his setting allows the model to capture finer patterns in the data, but it also increases the risk of overfitting if the data contains noise.

The fifth parameter evaluated was Number of Estimators, it was set in 100 trees. The more threes the les variance, but it increases other variable, which is the computational cost.

## Random Forest Regressor Scores and Results



Fig 23. Training and testing scores for Random Fores Regressor.

Additionally, we can observe in figure 23, training and testing scores for the Random Forest Regressor model. The R² score for the training set is 0.76, while for the testing set, it is 0.82. These scores indicate that the model generalizes well to data outside the training set, as the testing score is relatively high and slightly better than the training score. This outcome suggests that the model can make reliable predictions on new data without showing signs of overfitting or underfitting. Therefore, these initial results are promising and indicate a well-balanced model.

Additionally, we can see some results of the predicted values versus the actual values in the following Figure 24.

Tabla

Descripción generada automáticamente

Fig 24. Cross Validation accuracy Random Forest Regressor.

## Cross Validation for Random Fores Regressor

Imagen que contiene interior, tabla, computadora

Descripción generada automáticamente

Fig 25. Cross Validation accuracy Random Fores Regressor.

The results (Fig. 25) show that the average accuracy for this Random Forest model is 0.7078. There is some variation across the folds, with accuracy scores going from a minimum of 0.5452 to a maximum of 0.8946. This indicates a moderate level of inconsistency in the model’s performance when predicting on different subsets of the data, regardless it is the best score between the 3 models performed. Then comparing the cross-validation results with the training se score (0,76) we notice that both scores are quite hight, without showing overfitting signals, it means that the model is capturing the inner signals of the data well enough to generalized it.

# Conclusions, Recommendations, and future work

A dataset for the Portugal educative environment was analysed and performed with 3 machine learning models, the accuracy results after cross-validation for both models are shown in Table 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Machine learning model | Training set score | Cross-validation before Hyperparameter Tunning | Cross-validation after Hyperparameter Tunning | commentaries |
| K-NN regressor | 0.72 | 0.63 | 0.66 | It is possible to obtain a good prediction, but the model looks quite underfitting |
| Linear Regression | 0.72 | 0.65 | 0.67 | Similarly with K-NN regressor It is possible to obtain a good prediction, but the model looks quite underfitting |
| Random Forest Regressor | 0.76 | 0.68 | 0.71 | The performance between the 3 models, it does not show nether underfitting nor overfitting. |

Table 2. General Accuracy results for 3 Machine Learning Models

These results can be seen in the figure 26, as well. It is possible to notice that Random Forest Regressor has the higher score and the distance between training set score and Cross validation shorter than in the other models, it tells us that this model is the most appropriate to predict the grades of the students in this particular context

Gráfico, Gráfico de barras

Descripción generada automáticamente

Fig 26. Training set and cross validation scores

After modelling both machine learning models, the best accuracy score was obtained by the Random Forest Regressor model, so this would be the model to follow in the next step of the project further in the future. However, the R2 score can still be improved.

Finally, a SHAP (SHapley Additive exPlanations) was done (Figure 26)

Gráfico

Descripción generada automáticamente con confianza media

Figure 26. SHapley Additive exPlanations for Random Forest Regressor.

This graph shows the impact of some of the features on the predictive performance of Random Forest Regressor, hence is a measurement of the importance and impact. The possibility of obtaining information about these specific characteristics can be of great help to teachers and administrators, allowing them to focus their efforts when making decisions aimed at improving the academic performance of their students.

The result of this plot shows that the most impactful characteristics are sex, age and the school where students came from, probably this are features that cannot be controlled in any way, but in second place there are other features like level of absences that have a great impact on the prediction, hence this is a feature that might be focus with an specific strategy from the school staff. Afterwards

To get better accuracy scores it should be necessary to do more CRISP-DM cycles with the addition of new parameters could be useful, as it was said before, there is a context behind the grades, and unlike other predictions, here we are talking about students—developing human beings with all their complexities behind them.

## Final conclusions and future work

The educational context is a multidisciplinary field which includes the most diverse aspects and areas, in one hand there the students each one with their owns challenges, each one with unique circumstances and life experiences, this complexity is a layer that is very difficult to unravel and is one of the most challenge every day.

In the other hand there is the teacher figure, humans as well, each of them carries their own personal stories which makes each teaching instance unique and hardly replicable.

These are just two examples of the layer and layer of complexity behind students, and this scenario is becoming even more complex in the context of exams and standardized evaluation.

For all these reasons and ideas, attempting to predict academic performance is quite demanding, and having achieved moderately high accuracy levels, I see it as an initial success.

This analysis about the complexity of education context and therefore the machine learning techniques applies make us understand one of the probable reason of what Random Forest Regressor had a better performance than Linear Regression and K-NN regressor, this is because Random Forest due to his structure is able to capture more complex data in contrast to Linear regression which suppose a lineal relation between the variables, it makes the model unprecise.

Random Forest which is based on decision trees can manage not lineal relation, or if the data has difficult patterns or interactions, Random Forest can manage it. Additionally, there is another characteristic of random forest that might be related with the better performance, and it is about the structure based on decision trees groups, it helps the model to decrease the impact of the noise of extreme value. We saw that between grades of first semester there were extreme values going from 0 o to 20, also due to the range of features of two different schools, noise is definitely something that the model needs to manage.

Expanding the analysis of this kind of dataset to other educative realities like Chile and Ireland, will be useful to contrast and understand which variables are the most important predictors of academic outcomes in each reality.

Certainly, the next steps are to improve the accuracy of this analysis and apply it to a bigger dataset of different countries.

# 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. United Nations Development Programme (UNDP). (2024) Human Development Report 2023-24: Breaking the Gridlock: Reimagining Cooperation in a Polarized World. Available at: https://hdr.undp.org/content/human-development-report-2023-24 (Accessed: 19 May 2024).
8. Schröer, C., Kruse, F. y Gómez, J.M. (2021) ‘A Systematic Literature Review on Applying CRISP-DM Process Model’, Procedia Computer Science, 181(1), pp. 526–534. Disponible en: <https://www.sciencedirect.com/science/article/pii/S1877050921002416>
9. Cortez, P. y Silva, A. (2008) ‘Usando a Mineração de Dados para Prever o Desempenho do Aluno do Ensino Médio’, en Brito, A. y Teixeira, J. (eds.) Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008). Porto, Portugal: EUROSIS, pp. 5-12. ISBN 978-9077381-39-7.
10. Scikit-learn (2023) sklearn.neighbors.KNeighborsRegressor, scikit-learn 1.5. Disponible en: https://scikit-learn.org/1.5/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

https://github.com/msotomayorpal/CA-2-Capstone-Report-Strategic-Thinking/blob/6616c84695d91af31fa5badafadd74271a0586f1/CA%202%20strategic%20thinking%20Capstone%20Report.ipynb

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. United Nations Development Programme (UNDP). (2024) Human Development Report 2023-24: Breaking the Gridlock: Reimagining Cooperation in a Polarized World. Available at: https://hdr.undp.org/content/human-development-report-2023-24 (Accessed: 19 May 2024). [↑](#footnote-ref-7)
8. 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-8)
9. 2. Cortez, P. y Silva, A. (2008) ‘Usando a Mineração de Dados para Prever o Desempenho do Aluno do Ensino Médio’, en Brito, A. y Teixeira, J. (eds.) Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008). Porto, Portugal: EUROSIS, pp. 5-12. ISBN 978-9077381-39-7.. [↑](#footnote-ref-9)
10. Scikit-learn (2023) sklearn.neighbors.KNeighborsRegressor, scikit-learn 1.5. Disponible en: https://scikit-learn.org/1.5/modules/generated/sklearn.neighbors.KNeighborsRegressor.html [↑](#footnote-ref-10)