Summer 2022

Artifact Development for the Prediction of Stress Levels on Higher Education Students using Machine Learning

Valentina Quiroga
CCT College Dublin

Alejandra Hurtado
CCT College Dublin

José Rojas
CCT College Dublin

Artifact Development for the Prediction of Stress Levels on Higher Education Students using Machine Learning

Problem-Solving for Industry

Nora Valentina Quiroga – 2018318

Alejandra Daniela Quintanilla Hurtado – 2018289

José Virgilio Nájera Rojas – 2018261

Supervisor: Dr Muhammad Iqbal

PSI_22_12

Information and Communications Technology (ICT) Faculty
CCT College

Dublin

May 2022

Table of Contents

Table of Figures

Introduction

Stress is an adaptative reaction of an organism, human or not, to the demands of fitting in an environment (Kav Vedhara, 1996). When stress originates in an educational context, it is common to refer to it as a student and their mechanisms to adapt and cope with the academic demand.

All humans experience stress during their lifetime, but when this overwhelmed feeling is prolonged can affect human behaviour and the ability to deal with physical and emotional pressure, having, as a result, a different range of problems.

It is important for higher-level educations institutions, such as colleges and universities, to be aware and have a deep knowledge of the levels of academic stress in their students. This is one of the main factors that affect student performance and academic failure, as well as being associated with depression, chronic diseases, and malfunction of the immune system (Kav Vedhara, 1996).

Considering the susceptibility of third-level students to suffer long periods of pressure and anxiety, the purpose of this research is to predict the level of stress in college students enrolled in an Irish Higher Education course, with the purpose of helping to identify areas that require intervention within the institution and design preventive strategies.

Analysis Report

1. Data Mining

The data mining process starts with a collection of data. Once the relevant data is organized and clean the mining or machine learning tasks can begin. This process intent and goal is to discover valid, novel, and understandable patterns within the data.

To guide the data mining efforts, frameworks like CRISP-DM are used to create strategies that will lead to the completion of a project. It will help in the selection of right tools and approaches and to ensure that every stakeholder has a real understanding of the core values of the company or the business.

1.1. Tasks

A task in data mining can be defined as a type of problem that can be solved with a Machine Learning algorithm. Each task has its own requirements, and it can vary according to the final results that can be obtained.

The tasks can be dived in Predictive or Descriptive

1.1.1. Predictive

The objective of a predictive tasks is to estimate future or unknown values of a specific variable.

- Classification: A set of objects are joint by a specific attribute or characteristic. This label is a discrete value, and it is known for every object. The goal of this task is to assign the correct label to new and unlabelled items.
  This process is the most suitable to for the development of the project laid out in this report.

1.1.2. Descriptive

The objective of descriptive task is to identify patterns in the data that can describe, explain, or summarize the data points.

1.2. Data Mining Models

Techniques, algorithms, and methods are needed to solve the predictive tasks mentioned above. Some of those techniques are:

- Techniques based on decision tress: based on algorithms such as ‘divide and conquer’.
- Techniques based on artificial neural networks: based on weight and a set of nodes (neurons).
- Techniques based on density or distance: based on distances between elements such as K-Nearest Neighbors.

2. CRISP-DM Framework

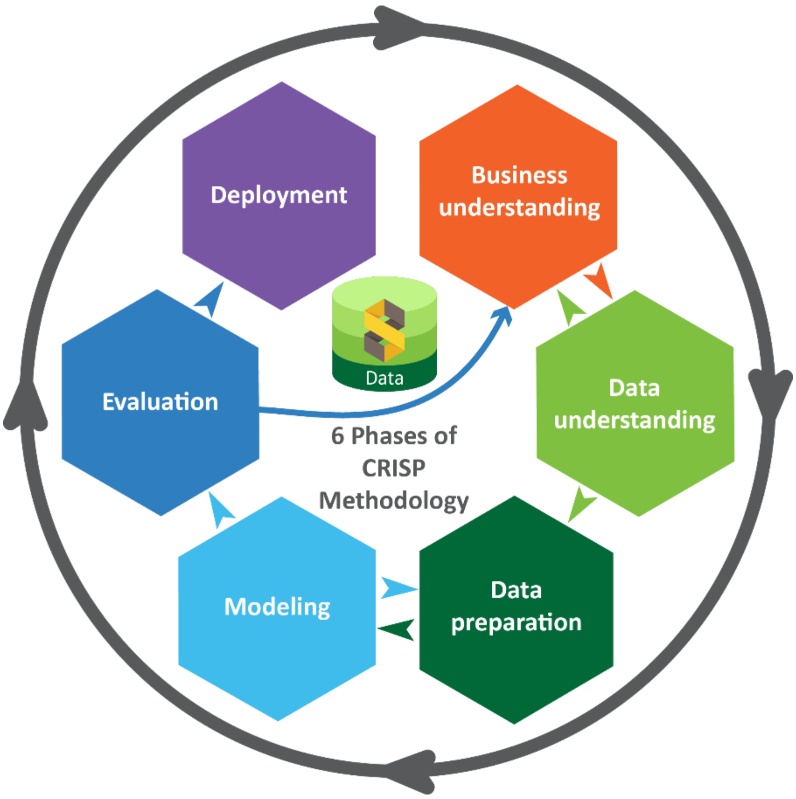


Figure 1 - CRISP-DM Framework

Introduction

Each human has a distinctive brain due to the fact it can be influenced over its lifetime by environmental or genetic factors, different and unique neural wiring, and individual cognitive development (Simon Neubauer, 2018).

The brain is responsible to detect stressful situations. Once detected will react releasing stress hormones and it will determine the consequences of the stress. This could lead to a diverse of psychological, medical, and behavioural problems (Jonathan D Quick MD, 1987)

Project Concept

Develop an artifact to predict stress levels on students enrolled in an Irish Higher Education course using Machine Learning and the CRISP-DM framework.

2.1. STAGE ONE: Business Understanding

Identifying the Business opportunity

Nowadays, life rhythm is more demanding than in the previous years. People's expectations are higher; meaning it could be more difficult to succeed in finishing a degree, especially during the last year of college or university. This project is based on the idea of developing an artifact to predict stress levels in students enrolled in the last year of an IT course. .

2.1.1. Determine business objectives

Aim and Objective

This project aims to be used to help and improve educational institutions. With the results and findings of this research, colleges and universities can be favoured and guided into developing changes in the design of the academic curriculum, planning of assignment submissions deadlines, final exam dates organization, and many other factors that can affect mental health in students.

This data mining project aims to make reliable predictions of the stress levels of the alumni of a specific university or college. The data is based on and provided by students at any particular university and/or specific course.

In the long run, the more significant objective is to enhance the teaching and learning experience and, consequently, attract more students to the institution.

2.1.1.1. Set objectives

For this project, the following objectives have been defined:

- This project aims to be used as a helpful tool to improve higher education services.
- This project aims to guide colleges and universities to develop changes in the academic curriculum design to favour student success.
- Identify and detect student stress levels around assignment submission deadlines, final exam dates, and other factors that can affect mental health in students to improve the organization and planning of those periods.

2.1.1.2. Produce project plan

Gantt chart: In the following chart, the six stages of the CRISP-DM framework were scheduled chronologically. The primary tasks and outputs of the project were assigned a time range to be completed by the team. A larger version of this Gantt chart may be found in Appendix A: Project management.

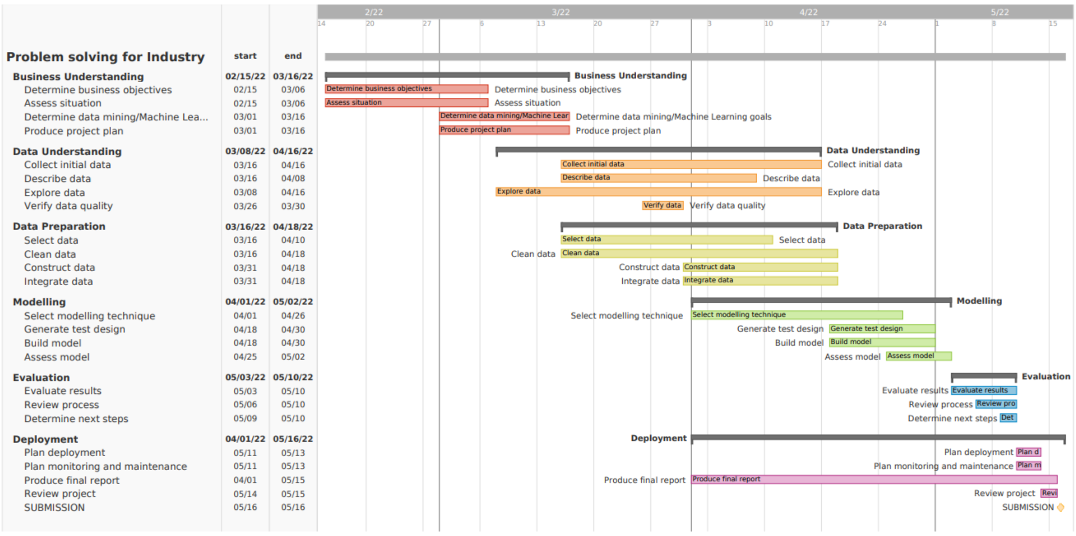


Figure 2 - Gantt chart

2.1.1.3. Business success criteria

Making accurate predictions of stress levels in students in the following academic year would help develop and structure new student success strategies.

Furthermore, the business success criteria would be linked to increasing students' academic achievements. Decreasing the pressure and anxiety level would lead to a higher number of students who pass a module, get higher grades, as well as reducing the number of student dropouts.

2.1.2. Assess situation

The team has access to the questions and raw data collected from a survey on stress in postgraduate university students in the United Kingdom. Students came from a wide variety of fields and disciplines, as well as different personal backgrounds. The survey's main goal is to inquire about how stressed students are, the sources of stress, as e well as how students deal with it. (Rolfe, 2020)

2.1.2.1. Inventory of resources

Technologies

Machine Learning (ML) comes with a varied collection of solutions, platforms, and software available in the market today. Furthermore, Machine Learning technology is constantly evolving. This section highlights each of the proposed technologies for this project and the justification for their consideration.

Technologies related to Machine Learning work by gathering information from the available data and creating logical models based on this specific knowledge (data). Therefore, they optimize and simplify the entire Machine Learning workflow.

Nowadays, ML technologies are in charge of training the model as well as evaluation, deployment, and production. At the end of this process, these trained models can be applied to automate future procedures.

Proposed Technologies

RapidMiner

RapidMiner is a comprehensive data science platform with a visual workflow design and full automation. It means that the user does not have to do the coding for data mining tasks. RapidMiner is one of the most popular data science tools (Arnaldo, 2021).

RapidMiner is a software platform that provides an integrated environment for ML, data mining, text mining, predictive analytics, and business analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process, including data preparation, results' visualization, validation, and optimization (Software ARGE Inc., 2022).

Jupyter Notebook

Jupyter Notebook is a web application (open source) that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, other multimedia resources, and explanatory text in a single document.

Jupyter Notebook can be used for numerous data science tasks, including data cleaning and transformation, numerical simulation, exploratory data analysis, visualization, statistical modelling, machine learning, and deep learning. Jupyter is an acronym for Julia, Python, and R, the three programming languages that Jupyter started with (Jupyter, 2022).

Azure Machine Learning

Azure Machine Learning is a cloud service for accelerating and managing the machine learning project lifecycle. Machine learning professionals, data scientists, and engineers can use it in their day-to-day workflows, training and deploying models and managing MLOps (Machine Learning Model Operationalization Management).

The user can create a model in Azure Machine Learning or use a model built from an open-source platform, such as Pytorch, TensorFlow, or sci-kit-learn. MLOps tools help monitor, retrain, and redeploy models (Microsoft, 2021).

TensorFlow

TensorFlow is an end-to-end open-source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML, and developers easily build and deploy MLpowered applications (Google Open Source, 2022).

2.1.2.2. Requirements, assumptions, and constraints

Requirements:

- *Technologies*: ML Software, Computer, Schedule of completion (refer to Figure 2 - Gantt chart)
- *Data collection and data security*: Dataset 'Student Stress Survey Jan2020 OPENDATA.xlsx', Survey (refer to Appendix B: Survey).

Assumptions:

- The size of the dataset is big enough to go ahead with a mining data project.
- Students will fill out the survey consciously and truthfully.

Constraints:

- For data privacy reasons, the survey must contain a disclaimer outlining the implications of volunteering information for the survey.

2.1.2.3. Risks and contingencies

The project will be at risk if the data is insufficient to train an ML model. In this case, the team will contemplate a simpler classifier model to avoid overfitting. If this action does not solve the problem, techniques like Data Augmentation and Synthetic Data will be considered. (Gonfalonieri, 2019)

Additionally, if a team member becomes unfit to work, the team will reassess the progress and move back or forward, adhering to Agile principles and practices.

SWOT Analysis

For the Business Analysis tool, the SWOT framework is used to identify the Strengths, Weaknesses, Opportunities, and Threats involved in this project. This strategic planning will help to come up with clearer objectives and have a better understanding of the capabilities, function, and dimensions of the research concept.

Figure 3 - SWOT Analysis

2.1.2.4. Costs and benefits

The data collected and used for this project do not imply any monetary cost per se.

The project will not generate any economic profit directly. However, it will impact the institution's revenue incidentally since the objective of this project is to improve the quality of services offered to students and, therefore, their satisfaction. This can translate into growing prestige and reputation, making more students consider enrolling in the institution.

2.1.3. Determine Data Mining/Machine Learning goals

The objectives in terms of Machine Learning are:

1) Predict the stress level of a higher education student in a specific time frame of the academic year based on the data obtained in a survey containing demographics, physical and mental states, and coping mechanisms, among other queries.

2) Predict the average stress level among students of a specific module at a specific academic period.
3) Identify factors that influence students' stress levels the most.

2.1.3.1. Business success criteria

Refer to Determine business objectives (2.1.1.3 Business success criteria)

2.1.3.2. Data mining / Machine Learning success criteria

As a business success criterion, the team has established that a reliable and accurate prediction should reach a percentage above or equal to 85 percent. The predictive level of accuracy will be determined by a specific algorithm; hence this subject is addressed further in the report (STAGE FIVE: Evaluation)

2.1.4. Produce project plan

2.1.4.1. Project plan

To facilitate the team organization, the project will be divided into the following phases:

- Phase 1: Study of the situation and analysis of the structure of the data (dataset examination)
- Phase 2: Execution of code for data representation (data exploration and visualization)
- Phase 3: Data preparation (selection, cleaning, formatting, and any other necessary actions)
- Phase 4 : Choice of modelling techniques. Model building
- Phase 5: Analysis of results obtained in previous phase. Repeat phase 4 and 5 if necessary.
- Phase 6: Production of report with results obtained.
- Phase 7: Presentation of final results.

Business Analysis Canvas

The planning tool for the development of the artifact is depicted in the image below. The ideas depicted in the figure were thought to design and find an effective and successful business approach.

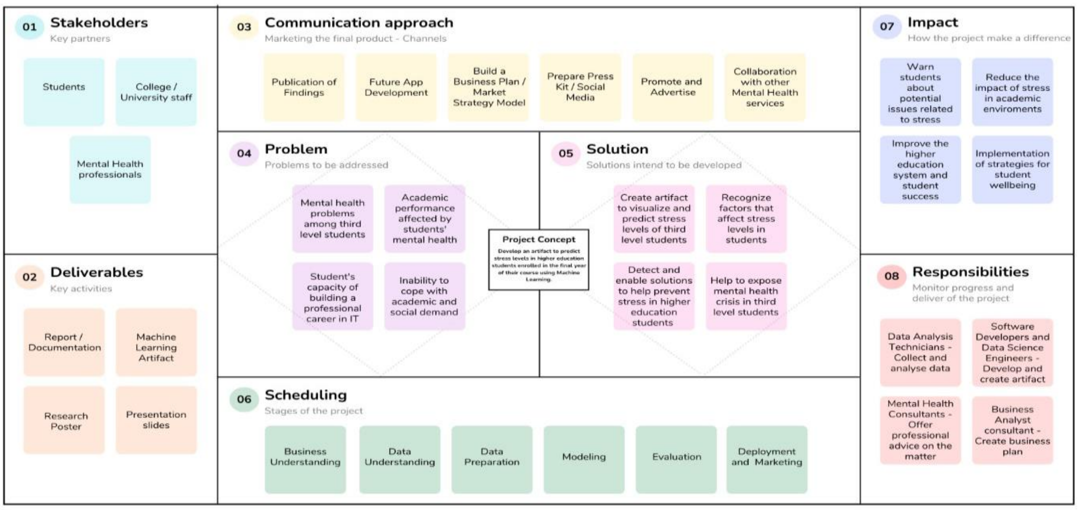


Figure 4 - Business Analysis Canvas

For a detailed breakdown of plan refer to Appendix A: Project management.

2.1.4.2. Initial assessment of tools and techniques

The following evaluation was done to determine the differences between the technologies mentioned previously and decide which is the best option for the project regarding the amount of data to be processed and analysed in real-time.

Additionally, it is essential to highlight the programming languages learned by the data analyst in previous courses; R and Python are considered.

RapidMiner works with Machine Learning but is a no-code development platform and enables the user to perform data mining tasks with a drag-and-drop feature. As a result, programming languages such as R and Python are not accepted.

Azure ML is cloud-based; the user must always have a reliable internet connection to use it. In addition, their development and maintenance are both expensive.

ASPECTS TO EVALUATE	Rapidminer	Jupyter Notebook	Azure Machine Learning	TensorFlow
Performance	✓	✓	✓	✓
Code presentation	✗	✓	✗	✓
User interface	✓	✓	✓	✓
Multiplataform	✓	✓	✓	✓
Languages supported (R / Python)	✗	✓	✓	✓
Data Size	✗	✓	✓	✓
Open-source	✗	✓	✗	✓
	43%	100%	71%	100%

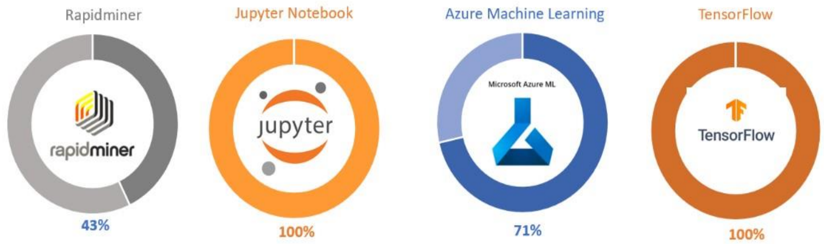


Figure 5 - Comparative table between technologies

Google Trends shows in the following figure how the popularity and interest for Jupyter and TensorFlow are more prevalent than RapidMiner and Azure.

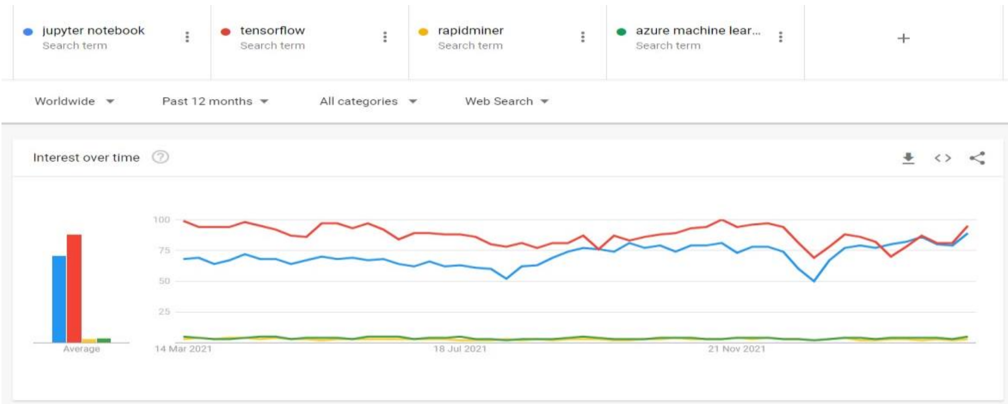


Figure 6 - Google trends on Jupyter, TensorFlow, RapidMiner and Azure (Google Trends, 2021)

Technologies in use

In conclusion, Jupyter Notebook is the best option for this project. It is compatible with real-time code, analytical models, visualization, and Markdown. Its functions include data cleaning and transformation, simulation analysis, statistical modelling, and deep learning.

Jupyter is one of the most widely used machine learning platforms. It is a straightforward task editor as well as an efficient platform.

In addition, it works with R or Python and allows the user to save and share live code in the notebooks. It is possible to obtain access via a graphical user interface (GUI) like Winpython navigator and Anaconda Navigator, which is the one that will be used for this project.

Key Features:

- Economical, it is open source with no cost.
- Friendly user interface.
- Works in the browser.
- Live code.
- Coding and error correction line by line.
- Easy to use for visualization and presentation code.
- There are many options for exporting and sharing results.
- Version control.
- Allows collaboration (JupyterHub).
- Supports more than 50 programming languages such as Python, Ruby, R, among others.

GitHub is a technology that allows users to host code in different languages. The team used GitHub to have better version control and collaborate with the different members of the group. Another advantage is that is free when the project is open source and public.

Furthermore, Anaconda Spyder is used as a scientific Python development environment for the editing, testing, and debugging of the web application.

Finally, Google Cloud Platform is used for the deployment of the web application, and it uses, as well, extra features in relation to maintenance, storage and version control.

2.2. STAGE TWO: Data Understanding

This step requires a further and detailed analysis of the data, this is because it needs to be avoided any conflicts with the future phase, data preparation. This stage is where exploring, creating new tables, and making visualizations help decide the quality of the dataset. (IBM, 2021)

2.2.1. Collect initial data

At this stage a compilation of the sources it is put together. The methods on how these data points were obtained are described below. This step also keeps track of any issues faced and the solutions found.

For the purpose of this project the dataset is a form of an existing questionnaire, specifically a survey.

2.2.1.1. Initial data collection report

The different data collected for this project was found online from third party websites. It was decided to choose an existing dataset; a survey made in 2019. This questionnaire it is part of a research of the University of Bristol in partnership with Pukka Herbs on stress levels on UK students. Even though other datasets were found, the information was not a good fit for the development of this project. In some cases, there were not enough entries or the contents of the dataset it was not in accordance with the objectives of this project.

2.2.2. Describe data

A description of the data, the number of records and an analysis of the dataset is performed in the step below. **Invalid source specified.**

2.2.2.1. Data description report

The dataset contains 218 entries and 36 columns. The greatest number of columns have a categorical object data type. The only numerical attribute is column number 3, which depicts the age of the participants. For this reason, using method describe() in Python to have a look at the statistics of the data, it can only get information from column Question 3.

Statistics of the features ¶

```
In [15]: dataset_1.describe()
```

Out[15]:

	Q3
count	218.000000
mean	36.550459
std	132.862771
min	21.000000
25%	24.000000
50%	26.000000
75%	29.000000
max	1987.000000

Only shows the statistical data of the attribute Q3 that is the only numeric one so far.

Figure 7 - Statistics Features

```
In [9]: dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 218 entries, 0 to 217
Data columns (total 36 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Q1      218 non-null    object
 1   Q2      218 non-null    object
 2   Q3      218 non-null    int64
 3   Q5      218 non-null    object
 4   Q6      218 non-null    object
 5   Q7      218 non-null    object
 6   Q8      218 non-null    object
 7   Q9      218 non-null    object
 8   Q10_1   218 non-null    object
 9   Q10_2   218 non-null    object
 10  Q10_3   218 non-null    object
 11  Q10_4   218 non-null    object
 12  Q10_5   218 non-null    object
 13  Q10_6   218 non-null    object
 14  Q10_7   218 non-null    object
 15  Q10_8   218 non-null    object
 16  Q10_9   218 non-null    object
 17  Q10_10  218 non-null    object
 18  Q10_11  218 non-null    object
 19  Q10_12  218 non-null    object
 20  Q11     218 non-null    object
 21  Q12     218 non-null    object
 22  Q13     218 non-null    object
 23  Q17_1   218 non-null    object
 24  Q17_2   218 non-null    object
 25  Q17_3   218 non-null    object
 26  Q17_4   218 non-null    object
 27  Q17_5   218 non-null    object
 28  Q17_6   218 non-null    object
 29  Q17_7   218 non-null    object
 30  Q17_8   218 non-null    object
 31  Q17_9   218 non-null    object
 32  Q17_10  218 non-null    object
 33  Q17_11  218 non-null    object
 34  Q17_12  218 non-null    object
 35  Q18     144 non-null    object
dtypes: int64(1), object(35)
memory usage: 61.4+ KB
```

Figure 8 - Data Types and Null Values on the dataset

2.2.3. Explore data

Exploring the data in this step is about understanding the usage of tables and producing visualizations to have a suitable approach to the storyline of the case and a better understanding of the data.

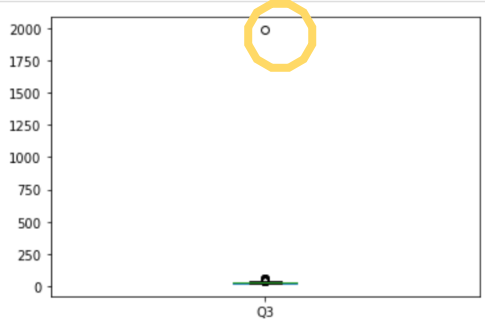
VISUALIZATIONS WITH RATIONALE can be found in the Jupyter Notebook File

2.2.3.1. Data exploration report

After exploring the data in early stages, a new table was created, called dataset_1. This new data frame has fewer columns because Question 11 and Question 18 are removed (further explanation of this matter is on the data preparation phase).

One technique used to check and visualize outliers is a box plot. As mentioned before, Question 3 is the only numerical attribute so far. In the figure below it is possible to observe an outlier.

In [16]:
```
dataset_1.plot(kind='box', subplots=True, layout=(1,1), sharex=False, sharey=False)
plt.show()
```



The data presents outliers in the Q3 feature corresponding to age.

Figure 9 - Outlier visualization

Further investigation of data in column 3 is performed. Using the method0 value_counts() the number 1987 highlights. It can be assumed that is an input error where a year was introduced instead of age. Using the method replace() the value was substituted for an accurate observation.

Fixing outliers

In [18]:
```
dataset_1['Q3'].value_counts()
```

Out[18]:
```
26      32
24      29
25      27
22      24
29      20
23      15
27      12
28      11
30       7
31       7
36       4
34       3
33       3
32       3
21       3
35       3
39       3
43       2
48       2
63       1
38       1
40       1
50       1
53       1
56       1
58       1
1987     1
Name: Q3, dtype: int64
```

Figure 10 - Fixing outliers

2.2.4. Verify data quality

In this step some of this actions were performed: verifying data quality, looking for errors in the dataset, examining inconsistencies, looking out for missing data or measurement errors. (IBM, 2021)

2.2.4.1. Data quality report

After fixing the issue mentioned above, the quality of the dataset is adequate. It is possible to observe using dataset_1.info(), that there are no null objects in the dataset and each column has the correct data type.

2.3. STAGE THREE: Data Preparation

One of the most consuming parts of Machine Learning and Data mining is the data preparation phase, in some cases this stage could take up to 50 percent of the time involved in Machine Learning process (IBM, 2021).

In this stage, the project is not only using the dataset downloaded from a third party website, but the team also has created a new survey. This is done with the aim of training the model with more specific and relevant data, as well as incrementing the data volume with newer entries and more significant to the objectives of this project. Some of the tasks required in this stage are:

2.3.1. Select data

Choosing the sample and labels from the dataset. Selecting rows and attributes accordingly.

2.3.1.1. Rationale for inclusion/exclusion

Once the dataset was comprehensible and relevant a new data set was created. There were two columns that had to be removed, question 11 and question 18. As mentioned before in the data understanding phase, this action had to be performed because of the type of answer these questions required.

This two columns contained text, and this type of data needs another type of analysis, a sentiment analysis. This method studies the emotions, opinions, and any type of expression in a text. At the time the team does not feel suited to perform this task, but it is revised in further stages.

Finally, the new data set contains 34 columns.

2.3.2. Clean data

The actions performed in this steps are related to changes to the dataset such as adding records or creating new attributes as well as cleaning the data if needed.

2.3.2.1. Data cleaning report

The new dataset created to work in the project is done for research and training purposes. In this data frame all the columns are converted to numerical data types using a coding scheme.

I.e.: Question 3 prompt the user with choosing a gender, the possible answers are Female, Male, and Prefer not to say. This options are changed to 1, 2, 3 respectively.

2.3.3. Construct data

With some algorithms, the data should be sorted before running the model. This will result in a better performance of the model and it saves in processing time.

To sort the data for the project, the first step was to reindex columns. Question 9 is the dependent variable, this question asks if the students felt any stress during the last 3 months, so it will be the label attribute to train the models.

To achieve a better performance, standardization and normalization are implemented to the dataset. Variables that have different scales do not contribute the same way to the model fitting. With Standardization those values can be re-sized to the same scales, and it is a good method to handle outliers as it will change the value distribution between 0 and 1. This is a great technique for machine learning that weight inputs, like regression and algorithms that require distance measurements like K-Nearest-Neighbours. (Liu, 2020).

Standardization Dataset

```
In [168]: scaling = StandardScaler()

In [169]: data_stand = scaling.fit_transform(dataNorm)

In [170]: data_stand = pd.DataFrame(data_stand)

In [171]: data_stand.sample(5)
```

Out[171]:

5	6	7	8	9	...	24	25	26	27	28	29	30	31	32	33
88	0.633372	1.481635	-0.984200	1.791878	...	-1.175078	-0.342243	-0.391713	1.426329	1.423861	0.826620	1.082873	1.357344	1.851998	1.592151
83	-1.145946	-1.081829	0.004535	-0.033493	...	-1.175078	-1.274855	-0.391713	1.426329	-0.240491	-1.513681	-0.852097	-1.068074	-0.961490	0.209327
35	-1.145946	1.481635	0.993271	1.791878	...	-1.175078	-1.274855	-1.373247	-2.448309	-1.072668	-1.513681	-1.819582	0.548871	-0.258118	0.209327
83	0.277508	1.481635	0.993271	1.791878	...	1.447792	1.522982	2.552889	0.457669	0.591685	0.046520	1.082873	1.357344	-0.258118	0.209327
88	1.345099	-1.081829	0.004535	-0.033493	...	2.322083	1.522982	-1.373247	-1.479649	-0.240491	0.826620	-2.787068	-0.259601	0.445254	-1.173498

Figure 11 - Standardization Process

Normalization Min- Max was also used, here the features are re-scaled to guarantee that the mean and standard deviation are both 0 and 1. This normalization is useful when comparing a dataset that has different factors. (Morrow, 2020) This process scales into small intervals so features will have the same scale, being very sensitive to the presence of outliers.

Normalize Dataset

```
In [186]: scalingMinMax=MinMaxScaler()

In [187]: data_scaled = scalingMinMax.fit_transform(dataNorm)

In [188]: data_scaled = pd.DataFrame(data_scaled)

In [189]: data_scaled.head(5)
```

Out[189]:

	0	1	2	3	4	5	6	7	8	9	...	24	25	26	27	28	29	30	31	32	33
0	0.0	0.00	0.25	1.0	0.333333	0.000000	0.000000	0.666667	0.50	0.25	...	0.00	0.00	0.00	0.00	0.5	0.0	0.00	1.00	0.25	0.00
1	0.0	0.00	0.25	0.0	0.333333	0.333333	0.222222	0.333333	0.75	0.50	...	0.25	0.00	0.50	0.50	0.5	0.0	0.50	0.50	0.50	0.00
2	0.0	0.00	1.00	0.0	0.333333	1.000000	0.000000	0.666667	0.75	0.25	...	0.75	0.25	0.50	0.25	1.0	0.0	1.00	0.25	0.50	0.75
3	0.0	0.25	0.25	0.0	0.333333	1.000000	0.000000	0.333333	0.50	0.50	...	0.00	0.25	0.00	0.25	0.5	0.5	0.25	0.50	0.50	0.25
4	0.0	0.00	0.00	0.0	0.666667	0.000000	0.555556	0.333333	0.50	0.25	...	0.25	0.25	0.25	0.50	0.5	0.0	0.00	0.50	0.75	0.50

5 rows × 34 columns

Figure 12 - Normalizing Data

The last normalization done was with the Lambda function. Lambda functions are anonymous functions that do not require any name, this is great for little tasks with less code. “Transforms features by scaling each feature to a given range”. (Scikit Learn, 2022) It is used for one-line expressions, in the code, we have a conditional statement, where the aim is to categorize using the formula

$$\frac{x - x.\min\,(axis = 0)}{x.\max(axis = 0) - x.\min\,(axis = 0)}$$

Normalize Dataset with MinMax

In [176]:
```
dataset_minmax = dataNorm.apply(lambda x: (x - x.min(axis = 0))/ (x.max(axis=0)- x.min(axis=0)))
```

In [180]:
```
dataset_minmax.sample(5)
```

Out[180]:

	Q1	Q2	Q3	Q5	Q6	Q7	Q8	Q10_1	Q10_2	Q10_3	...	Q17_4	Q17_5	Q17_6	Q17_7	Q17_8	Q17_9	Q17_10	Q17_11	Q17_12
2	0.0	0.0	1.0	0.0	0.333333	1.000000	0.000000	0.75	0.25	0.50	...	0.25	0.50	0.25	1.00	0.0	1.00	0.25	0.5	0.75
169	0.5	0.0	0.0	0.0	0.333333	0.000000	0.555556	0.25	0.00	0.00	...	0.25	0.00	0.00	0.25	0.0	0.50	0.25	0.0	0.00
216	1.0	0.0	0.0	0.0	1.000000	0.333333	0.444444	1.00	0.25	0.50	...	0.25	0.25	0.50	0.50	0.0	0.75	0.75	1.0	0.75
83	0.0	0.0	0.0	0.0	0.333333	0.000000	0.666667	0.50	0.50	0.25	...	0.50	0.25	0.25	0.50	0.5	0.50	0.25	0.5	0.50
123	0.0	0.0	0.5	0.0	0.333333	0.000000	0.555556	1.00	0.50	0.50	...	0.75	0.00	0.25	0.25	0.0	0.00	0.50	0.5	0.00

5 rows × 34 columns

Figure 13 - Normalization with Lambda Function

The transformation is given with this part of the formula (axis = 0). After performing the Standardization process, and two different types of Normalization that obtained the same results reassuring that the process is correct, the team moved to the next step.

Feature Selection and Feature Importance

Two types of feature selection methods are used to be able to create a relevant dataset for the ML process. This processes are Univariate Selection and Feature Importance. Both techniques search for the features with the strongest relationships for the prediction process.

Univariate Selection:

```
In [184]:  #Selection of independent and dependent features

X = dataset_1.iloc[:,0:33]  #independent columns
y = dataset_1.iloc[:,-1]    #target column i.e Stress Level dependent columns
```

```
In [185]:  #apply SelectKBest class to extract top 10 best features
bestfeatures = SelectKBest(score_func=chi2, k=10)
fit = bestfeatures.fit(X,y)
```

```
In [186]:  dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)
```

```
In [187]:  #concat two dataframes for better visualization
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Specs','Score']  #naming the dataframe columns
```

Figure 14 - Univariate selection

```
In [189]:  print(featureScores.nlargest(10,'Score'))  #print 10 best features

     Specs      Score
15   Q10_9  23.127301
10   Q10_4  17.983804
12   Q10_6  17.478934
18  Q10_12  14.662134
17  Q10_11  12.237136
11   Q10_5  12.046921
27   Q17_7  10.596965
23   Q17_3   9.987831
28   Q17_8   9.456110
30  Q17_10   8.721160
```

Figure 15 - Results Univariate Selection

The first step is to segregate the column with the dependant value. A separation is made between independent variables and the label. X represents the independent features and Y represents the dependent variable which is Question 9.

Then the method SelectKbest is run to select the top 10 features. By using the Feature importance technique is possible to observe the score for each relevant feature. This is an inbuilt class of the Tree Based Classifiers

```
Out[190]: ExtraTreesClassifier()

In [191]: print(model.feature_importances_) #use inbuilt class feature_importances

[0.01497599 0.01818264 0.02487601 0.01206925 0.02553202 0.02288846
 0.02633289 0.03468751 0.04040561 0.026373   0.08943186 0.03286455
 0.03684708 0.02528173 0.02785802 0.05763402 0.02193563 0.02768228
 0.03061792 0.03388047 0.02708419 0.03208554 0.02872929 0.02849931
 0.02412579 0.02827601 0.02946458 0.03565971 0.02696711 0.02156563
 0.03214467 0.02715454 0.0278867 ]

In [192]: #plot graph of feature importances for better visualization
feat_importances = pd.Series(model.feature_importances_, index=X.columns)
feat_importances.nlargest(10).plot(kind='barh')
plt.show()
```

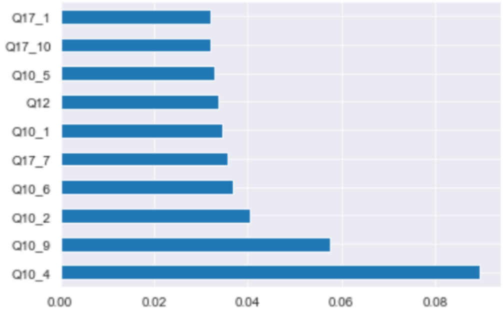


Figure 16 - Feature Importance

Lastly, after finalizing the process of Data Preparation, the dataset is ready for the modelling stage. For this, the dataset is divided into two subsets for training and testing the models.

```
dataset_sel.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 218 entries, 0 to 217
Data columns (total 11 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Q10_1   218 non-null    int64
 1   Q10_4   218 non-null    int64
 2   Q10_5   218 non-null    int64
 3   Q10_6   218 non-null    int64
 4   Q10_7   218 non-null    int64
 5   Q10_9   218 non-null    int64
 6   Q10_12  218 non-null    int64
 7   Q17_1   218 non-null    int64
 8   Q17_7   218 non-null    int64
 9   Q17_10  218 non-null    int64
 10  Q9      218 non-null    int64
dtypes: int64(11)
memory usage: 18.9 KB
```

```
array = dataset_sel.values
X = array[:,0:10]
y = array[:,10]
X_train, X_validation, Y_train, Y_validation = train_test_split(X, y, test_size=0.20, random_state=1)
```

Figure 17 - Train and Test splitting

2.3.3.1. Derived attributes

Not performed – Not needed

2.3.3.2. Generated records

Not performed – Not needed

2.3.4. Integrate data

2.3.4.1. Merged data

Not performed – Not needed

2.3.4.2. Aggregations

Not performed – Not needed

2.4. STAGE FOUR: Modelling

The ML model or models more appropriate to reach the objectives set in the previous phases are chosen correctly. After testing the model's performance, the techniques will be applied to pertinent data and evaluated accordingly with the business and ML success criteria.

2.4.1. Select modelling technique

In the early stages of the project, the team applied an ML Pipeline[1]. This principle was run with the aim of having an overall vision of the different ML model performances with the available data. A sequence of the following ML algorithms was created:

1) Logistic Regression (LR)
2) Linear Discriminant Analysis (LDA)
3) K-Nearest Neighbors (KNN).
4) Classification and Regression Trees (CART).
5) Gaussian Naive Bayes (NB).
6) Support Vector Machines (SVM).

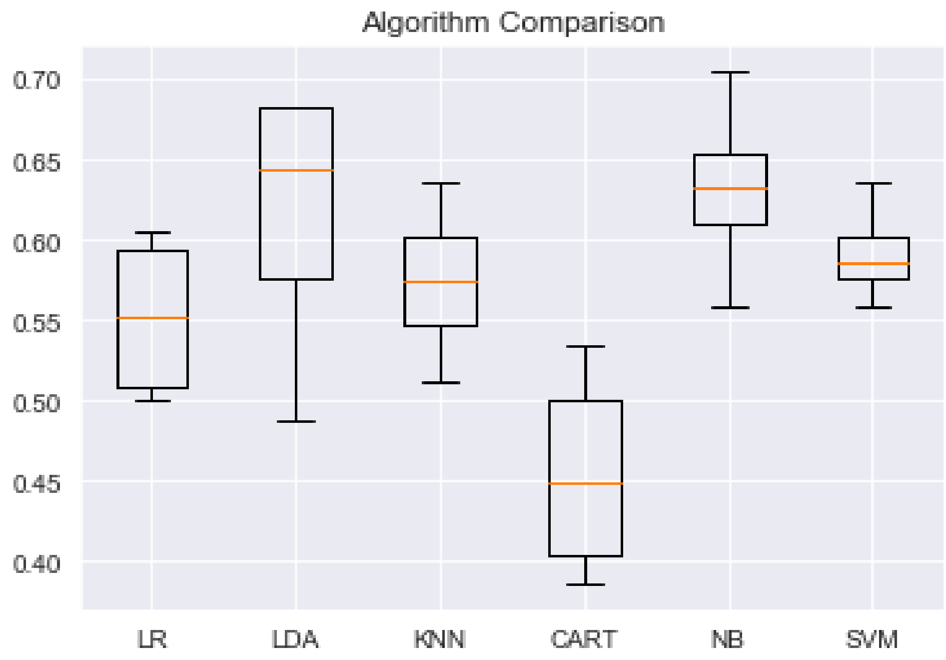


Figure 18 - Pipeline algorithms comparison

This ML Pipeline was not optimized or tuned; hence the accuracy rate is very low and the performance does not meet business or ML success criteria objectives set for this project.

[1] A pipeline is a linear sequence of data preparation options, modelling operations, and prediction transform operations. It allows the sequence of steps to be specified, evaluated, and used as an atomic unit. (Brownlee, 2021)

2.4.1.1. Modelling technique

The modelling techniques that better adapt to the problem are related to the supervised ML category. The dataset has labelled data that help the algorithm with the learning task, in this case, a predictive classification task between stressed and non-stressed students.

The model chosen for this task are:

1) Random Forest (RF)
2) K-Nearest Neighbors (KKN)
3) Decision Tree (DT)
4) Neural Network (NN)
5) Stacked Generalization (SG)

2.4.1.2. Modelling assumptions

- There are samples representing all values.
- The models work well with a relatively small dataset.
- Data preparation stage is completed.

2.4.2. Generate test design

2.4.2.1. Test design

To avoid overfitting, the data is split into two groups before working with ML models. The first set of data is used to generate the model; this works as the training data. The second set of data is used to evaluate and measure the model's quality; this works as the testing data.

Dataset splitting: Train = 174 data points. Test = 44

The approach for this project will be based on the technical report 'Relation Between Training and Testing Sets' that states that the ratio of 20:80 is empirically the best splitting approach for the training and the testing sets. (Afshin Gholamy, 2018)

2.4.3. Build model

Each ML model is described and executed on the training data. The model parameters are chosen with the aim of reaching the ML objectives.

2.4.3.1. Parameter settings

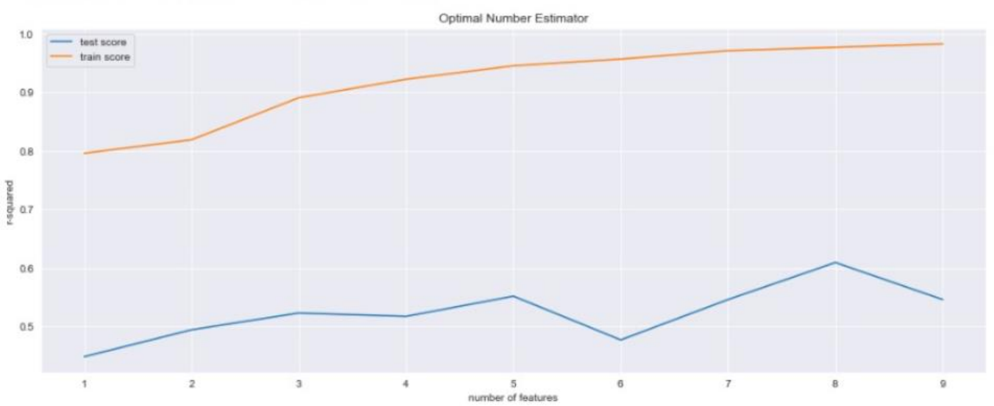
Hyperparameters

As seen previously, the models need to be optimized to reach an acceptable level of accuracy as well as get reliable predictions. The hyperparameters are chosen and set to control the learning process before the training phase begins. (Nyuytiymbiy, 2020). By using this kind of method, the possibilities of overfitting are reduced.

With the help of GridSearchCV, a library function from the package sklearn model_selection, the best hyperparameters are selected. This technique search for the best value to fit the parameter, after looping over values, the best is then extracted for implementation on the model. (Kotak, 2021)

1) Random Forest (RF) ⟶ n_estimators

This value relates to the higher number of trees created before the algorithm calculates the prediction averages. A greater number of trees improves performance, but it slows down the code. (Tavish, 2015)




```
pprint(modelrf_cv.best_params_)
{'n_estimators': 8}
```

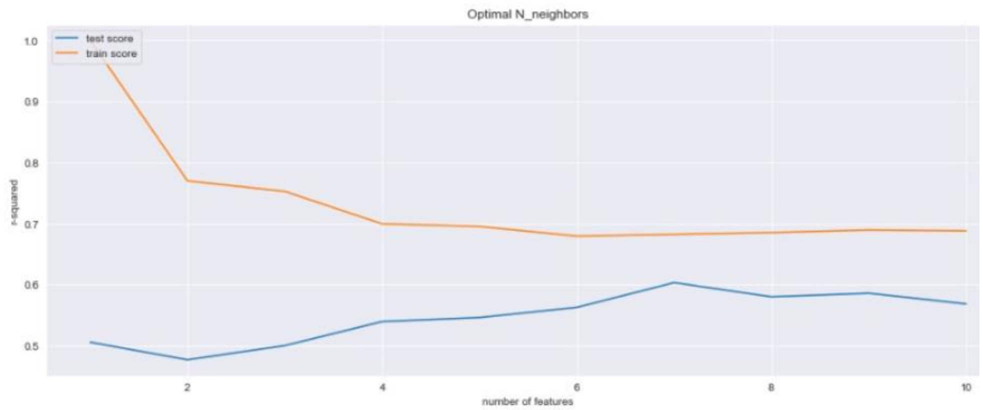
Figure 19 - Optimal Number Estimator

	n_estimators	Accuracy
Before GridSearchCV	10	98%
After GridSearchCV	8	95%

Table 1 - Random Forest Before and After Tuning

2) K-Nearest Neighbors (KKN) ⟶ n_neighbors

The decision of the number of neighbours for the K algorithm is based on the concept that values closer to zero (number of neighbours) will risk the prediction to be very volatile and overfit. A higher number of data points risks the algorithm to generalize and lose variance (curse of dimensionality). Which is what happens in the table below.




```
pprint(modelknn_cv.best_params_)
{'n_neighbors': 7}
```

Figure 20 - Optimal Number of Neighbors

	n_neighbors	Accuracy
Before GridSearchCV	3	77%
After GridSearchCV	7	66%

Table 2 - K-Nearest Neighbors Before and After Tuning

3) Decision Tree (DT) ⟶ max_depth

The maxium depth of the tree refers to the number of nodes to be consider from the root and down. Generally, deeper trees can reach higher levels of accuracy. This has to be consider as it can affect overfitting. (H2O.ai, 2022)

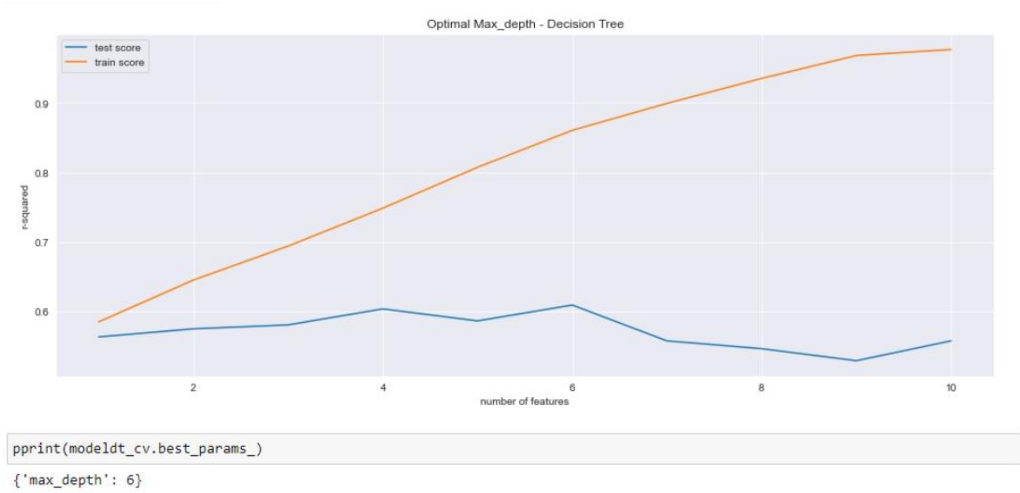


Figure 21 - Optimal Max_Depth - Decision Tree

	max_depth	Accuracy
Before GridSearchCV	5	78%
After GridSearchCV	6	86%

Table 3 - Optimal Max_Depth

4) Neural Network (NN) ⟶ alpha and max_iter

   Alpha is used as a regularization parameter. It controls the over and underfitting of the model. The highest is the alpha value is, the higher are the possibilities of overfitting.

   Max_iter is used to set the number of times the process of input and output it is going to be performed between the layers and its neurons.

5) Stacked Generalization (SG) ⟶ The models used on the Stack are already tunned with the hyperparameters mention above.

It was necessary to transform the data and drop the features less important for the predictions of stress. The models worked specifically with the following attributes:

Independent variables:

1) Low energy
2) Anxiety or tension
3) Sleeping problems
4) Rapid heartbeat or palpitations
5) Irritability
6) Sadness or tearfulness
7) Loneliness
8) Feeling overloaded with university work
9) Lack of time for relaxation
10) Lack of confidence with academic performance

Dependant variable - Label:

1) Stressed

2.4.3.2. Models

All five models were trained on a 20:80 ratio test and training data respectively. Below are the figures containing the details of each algorithm's parameters when performed.

```python
print('Random Forest - Parameters')
print('----------------------------------')
pprint(rf.get_params())
```

```
Random Forest - Parameters
----------------------------------
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 8,
 'n_jobs': None,
 'oob_score': False,
 'random_state': None,
 'verbose': 0,
 'warm_start': False}
```

Figure 22 - Random Forest Parameters

```python
print('K nearest neighbors - Parameters')
print('----------------------------------')
pprint(knn.get_params())
```

```
K nearest neighbors - Parameters
----------------------------------
{'algorithm': 'auto',
 'leaf_size': 30,
 'metric': 'minkowski',
 'metric_params': None,
 'n_jobs': None,
 'n_neighbors': 7,
 'p': 2,
 'weights': 'uniform'}
```

Figure 23 - K-Nearest Neighbors Parameters

```python
print('Decision tree - Parameters')
print('----------------------------------')
pprint(dt.get_params())
```

```
Decision tree - Parameters
----------------------------------
{'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': 6,
 'max_features': None,
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'presort': 'deprecated',
 'random_state': None,
 'splitter': 'best'}
```

Figure 24 - Decision Tree Parameters

```python
print('Neural network - Parameters')
print('----------------------------------')
pprint(mlp.get_params())
```

```
Neural network - Parameters
----------------------------------
{'activation': 'relu',
 'alpha': 1,
 'batch_size': 'auto',
 'beta_1': 0.9,
 'beta_2': 0.999,
 'early_stopping': False,
 'epsilon': 1e-08,
 'hidden_layer_sizes': (100,),
 'learning_rate': 'constant',
 'learning_rate_init': 0.001,
 'max_fun': 15000,
 'max_iter': 1000,
 'momentum': 0.9,
 'n_iter_no_change': 10,
 'nesterovs_momentum': True,
 'power_t': 0.5,
 'random_state': None,
 'shuffle': True,
 'solver': 'adam',
 'tol': 0.0001,
 'validation_fraction': 0.1,
 'verbose': False,
 'warm_start': False}
```

Figure 25 - Neural Network Parameters

The Model Stacked Generalization used a combination of all the models mentioned above.

2.4.3.3. Model descriptions

The results of the performance of each ML Model are depicted in the figures below.

Each figure contains the values regarding the following:

- Accuracy (Training and Testing): Percentage of number of correct predictions in relation to number of entries. Expected for success ≥ 85
- MCC (Training and Testing): The Matthews correlation coefficient measures the values between the prediction made and the real values. Expected for success ≥ 70
- F1 Score (Training and Testing): It measures exactness. Low levels of exactness mean a high level of false positives. Expected for success ≥ 85
- Mean Absolute Error (Algorithm): the difference scale of the prediction and the real value of the observation. ≤ 45
- Mean Squared Error: it takes the Mean Absolute Error; it squares it and makes and average of the whole dataset ≤ 45
- Root Mean Squared Error: Is the squared root of the Mean Squared Error (negatives to positives for comparison)
  (Brownlee, 2016)

```
Random Forest
----------------------------------
Model performance for Training set
- Accuracy: 0.9885057471264368
- MCC: 0.9808557113364228
- F1 score: 0.988484399051696
----------------------------------
Model performance for Test set
- Accuracy: 0.6590909090909091
- MCC: 0.43656904896382104
- F1 score: 0.6473808010171648
----------------------------------
Evaluating the Algorithm
Mean Absolute Error: 0.3409090909090909
Mean Squared Error: 0.3409090909090909
Root Mean Squared Error: 0.5838742081211422
```

Figure 26 - Random Forest Description

```
K nearest neighbors
----------------------------------
Model performance for Training set
- Accuracy: 0.6666666666666666
- MCC: 0.42636031866445345
- F1 score: 0.6564464316964302
----------------------------------
Model performance for Test set
- Accuracy: 0.6136363636363636
- MCC: 0.31546542357865043
- F1 score: 0.5659536541889483
----------------------------------
Evaluating the Algorithm
Mean Absolute Error: 0.38636363636363635
Mean Squared Error: 0.38636363636363635
Root Mean Squared Error: 0.621581560508061
```

Figure 27 - K-Nearest Neighbors Description

```
Decision tree
----------------------------------
Model performance for Training set
- Accuracy: 0.8563218390804598
- MCC: 0.7566163748422178
- F1 score: 0.8539471297349891
----------------------------------
Model performance for Test set
- Accuracy: 0.6136363636363636
- MCC: 0.30633122542052443
- F1 score: 0.5632481924854806
----------------------------------
Evaluating the Algorithm
Mean Absolute Error: 0.38636363636363635
Mean Squared Error: 0.38636363636363635
Root Mean Squared Error: 0.621581560508061
```

Figure 28 - Decision Tree Description

```
Neural network
----------------------------------
Model performance for Training set
- Accuracy: 0.8505747126436781
- MCC: 0.7453326822782701
- F1 score: 0.8455288079230826
----------------------------------
Model performance for Test set
- Accuracy: 0.5454545454545454
- MCC: 0.1740242113090467
- F1 score: 0.49925239234449764
----------------------------------
Evaluating the Algorithm
Mean Absolute Error: 0.45454545454545453
Mean Squared Error: 0.45454545454545453
Root Mean Squared Error: 0.674199862463242
```

Figure 29 - Neural Network Description

```
Stacking Model
----------------------------------
Model performance for Training set
- Accuracy: 0.8448275862068966
- MCC: 0.7349141289288403
- F1 score: 0.8312315301746283
----------------------------------
Model performance for Test set
- Accuracy: 0.6590909090909091
- MCC: 0.4012214601128163
- F1 score: 0.5939787485242031
----------------------------------
Evaluating the Algorithm
Mean Absolute Error: 0.3409090909090909
Mean Squared Error: 0.3409090909090909
Root Mean Squared Error: 0.5838742081211422
```

Figure 30 - Stacking Generalization Description

2.4.4. Assess models

In the subsequent phase (Evaluation), the models are analysed more deeply, yet, at this point, an assessment oriented to the models' accomplishment of the Machine Learning Objectives is carried out.

2.4.4.1. Model assessment

The values of each Model can be observed in the table below

	Accuracy	MCC	F1
knn	0.666667	0.426360	0.656446
dt	0.856322	0.756616	0.853947
rf	0.988506	0.980856	0.988484
nn	0.850575	0.745333	0.845529
stack	0.844828	0.734914	0.831232

Table 4 - Models Assessment

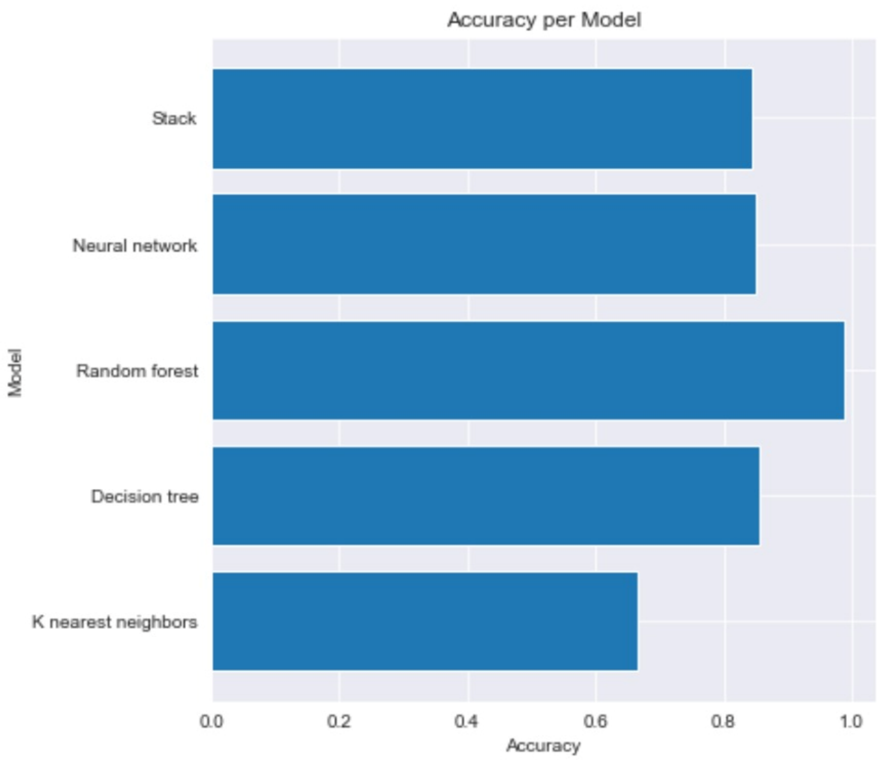


Figure 31 - Accuracy per Model

The models Random Forest (98%), Decision Tree (86%) and Neural Network (85%) accomplish the objective number one. The performance of this models ensures the predictions are reliable and accurate.

For the accomplishment of objective number two, extra data is needed. At the moment there is no information about specific modules, although there is data stating the area of the studies people are enrolled, there is no precise figures indicating academic modules or periods. At this point, this objective is not met.

2.4.4.2. Revised parameter settings

In the case of the Model Random Forest, a revision of the parameter regarding the nº of estimators, can be revised. Even though the accuracy increases when considering the results of the GridSearchCV, it should be analysed if there is a case of overfitting the model.

2.5. STAGE FIVE: Evaluation

2.5.1. Evaluate results

The models for Machine Learning are built and running with the training dataset. **Invalid source specified.** To evaluate the results is necessary to see if they meet the goals of the business criteria. It is important to focus if the results are clear to understand, once the results are out seen if they are any main discoveries to be mentioned. Finally, if the results brought more questions and the impact for the problem or business. **Invalid source specified.**. Another way to evaluate the models is using test applications. This stage is good to see if there are any other findings that could bring new challenges or redirect the project into a new horizon.

2.5.1.1. Assessment of data mining results

In order to be able to evaluate the results of the data mining process and the six ML models, an application was created with Anaconda Spyder.

This user interface (web application) calculates the stress levels according to the trained model the user chooses from. There are ten questions about different symptoms that are common when feeling stressed, once answered all the ten questions, the level of stress is calculated in a range from 1 to 5, being 1 not stressed at all and 5 very stressed.

As the business objective states, the project aims to enhance the teaching and learning experience. The web application can predict the levels of stress in the student using the interface, hence educational institutions can avail the help of this artifact to improve students learning outcomes. Therefore, it can be concluded that the results meet the business objectives and the ML and business success criteria.

2.5.1.2. Approved models

For this project five different models were implemented: K-Nearest Neighbors, Decision Tree, Random Forest, Neural Network and Stacking Generalization .

The best model for the prediction of stress levels in students is Random Forest, with an accuracy of 98%. This model is a combination of multiple Decision Trees, reaching a single result. This supervised algorithm can be prone to overfitting and bias, for this reason further analysis needs to be done in the future to corroborate the reliability of the model.

On the other hand, multiple Decision Trees like Random Forest can predict precise results by selecting a subset of features , like in this case the Feature Selection performed in the Data Preparation stage, this advantage can give the team confidence of keep working with this specific model that had the best performance. Later stages can include making predictions more specific and targeted to smaller groups or situation according to the institution needs.

Some challenges in this model are that demands more resources, is a complex model to interpret and when using a big data set it can be a time-consuming process. (IBM Cloud Education , 2020). This project did not encounter this issue since the dataset is relatively small, but it could be a problem in the future if the dataset grows.

The next approved model is Decision Tree with 85% accuracy in the training set. The last approved model is Neural Networks with 85% accuracy as well. This later model structure is based on the human brain, imitating the way neurons signal each other. (IBM Cloud Education , 2020) The structure consists of three layers

- Input. Once the input layer is set the weights can help calculate the importance of a feature.
- Hidden layers
- Output layer

Neural networks are adaptative, this means that they can change themselves as they learn from the training. (Ed Burns, n.d.) This is why this models can be taking into consideration for further studies and the later development of this project.

2.5.2. Review process

After acknowledging the approved models regarding their accuracy level, is important to see if the models have reached the level of the business needs. The following review is to analyse if there are any tasks that were overlooked and to check for quality assurance.

2.5.2.1. Review of process

The team has achieved the main goals set for this data mining project.

As with any project there were some complications, where at times, some task needed dedicated attention and to look more into detail. One subject of this matter was hyperparameters. At a stage the process of GridSearchCV was difficult to understand and at other times it was given results that were too high and made the models perform worse.

At the beginning of the project, it was difficult for the team to have a clear idea of how to develop the data mining project using CRISP-DM, but as the project move along each phase of the framework came smoothly and it prepared the team for the next challenge.

At a performance level, another element that could help improve the results of the models is to introduce new columns to the dataset, for example time or date or any other variable that the business case considers relevant. This will allow the expansion and flexibility of the final product as well as going deeper into the study of stress levels in academic environments.

2.5.3. Determine next steps

Depending on the results and the review process a decision must be made in order to decide if the project is ready for deployment or if more iterations must be made, or if it is necessary to start a new project.

2.5.3.1. List of possible actions

The project is ready to be deployed, some errors may be found in the project, but the team will only be able to see them and understand them in future stages as well as when more experience is gained in the ML area.

2.6. STAGE SIX: Deployment

2.6.1. Plan deployment

In the last phase of the CRISP-DM methodology, the ML artifact gets launched and the result are exposed to the relevant public. A strategy for the maintenance of the application and possible improvements are included on the report.

2.6.1.1. Deployment plan

To be able to implement the artifact in the real world, a web application is developed and hosted in a PaaS (Platform as a Service), in this case, the cloud platform service offered by Google .

Is necessary for the use of the application to have access to real data provided by students. By answering the online questionnaire, the student can observe the prediction of the personal level of stress at the specific moment of filling the form.

2.6.2. Plan monitoring and maintenance

2.6.2.1. Monitoring and maintenance plan

As a monitoring and maintenance plan, the following process is stablished.

- Semestral extraction and storage of the data obtained by the questionnaire (spreadsheet format).
- The data collected should get storage in a cloud platform alongside the web application. Automation of process using Software as a Service.
- Annual reassessment of queries and attributes used for the prediction of stress levels.

- Annual report containing visualizations, facts, and inferences regarding stress level in students. This report will be handed to the pertinent stakeholders to improve student experience.

2.6.3. Produce final report

2.6.3.1. Final report

Submitted on 16 of May 2022

2.6.3.2. Final presentation

Q&A Session on the 24 of May 2022

2.6.4. Review project

2.6.4.1. Problem encountered

As a first option, the deployment of the web application was going to be on Heroku. In later stages the team found out that the platform has been hacked and it does not allow uploads from GitHub. The solution was swapping for Google Cloud Platform.

Conclusion

The use of CRISP-DM methodology on this project made possible to achieve the main goal proposed at the start. The predictions of stress levels in students enrolled in a higher education course have been predicted accurately by three different Machine Learning Models.

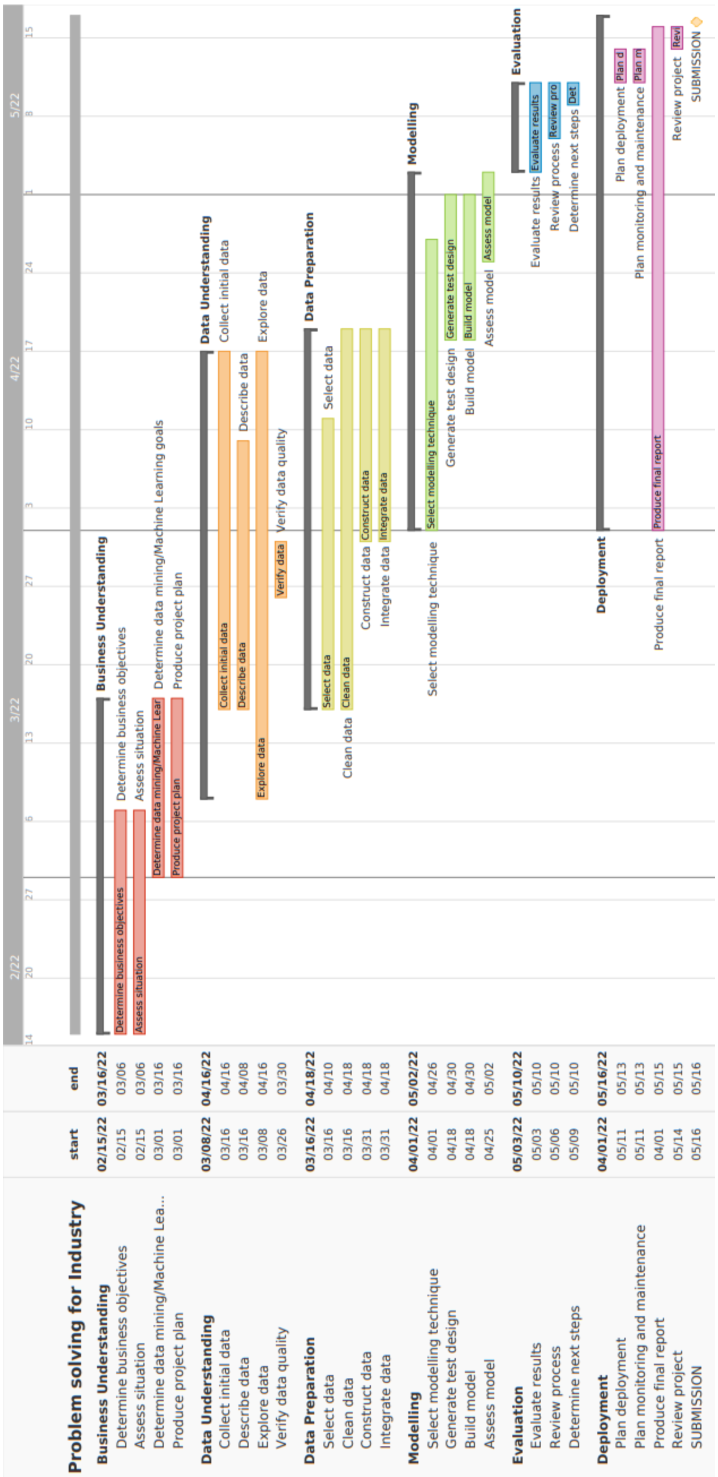
The results and findings of this research point out the need of providing students with alternative tools and solutions for managing academic stress. The prediction of stress levels can help universities and colleges to identify and carry-out measures to a problem that has increased dramatically over time.

The findings of the prediction also aim to reduce withdrawal from institutions due to academic stress. It is important to help students to stay motivated and keep working on their course, without putting to waste the progress already obtained.

The end result may allow universities and colleges to identify student profiles that are at risk and serve as a starting point to the reduction of academic stress within the institution.

Appendices

Appendix A: Project management



Appendix B: Survey

Timestamp	Gender	Ethnicity	How old are you?	What type of student are y	What qualification are you	What is your subject of stu	How often have you felt st	Low energy
3/30/2022 17:18:19	Female	Latino	35 - 39	International	Undergraduate	Computer science	Completely	Sometimes
3/30/2022 17:26:26	Female	Latino	30 - 34	International	Undergraduate	Computer science	Completely	Fairly often
5/11/2022 21:06:40	Male	Latino	30 - 34	International	Undergraduate	Computer science	To a large extent	Sometimes
5/11/2022 21:16:23	Male	Latino	25 - 29	International	PhD	Other	Somewhat	Sometimes
5/11/2022 21:28:28	Female	Latino	25 - 29	International	Undergraduate	Other	To a large extent	Sometimes
5/11/2022 21:33:28	Female	Latino	35 - 39	International	PhD	Other	Completely	Sometimes
5/11/2022 21:36:14	Female	Latino	25 - 29	International	Master's	Other	Completely	Sometimes
5/11/2022 21:42:14	Female	Latino	40 and over	International	Master's	Eduation	To a large extent	Fairly often
5/11/2022 21:46:55	Male	Latino	30 - 34	International	Master's	Other	To a large extent	Very often
5/11/2022 21:51:50	Male	White	30 - 34	EU/UK	Undergraduate	Computer science	Somewhat	Sometimes
5/11/2022 21:56:33	Male	Latino	25 - 29	International	Undergraduate	Other	To a small extent	Almost never
5/11/2022 22:06:47	Female	White	30 - 34	EU/UK	Master's	Eduation	To a large extent	Very often
5/11/2022 22:14:37	Female	Latino	30 - 34	International	Undergraduate	Arts and humanities	To a large extent	Almost never
5/11/2022 22:20:18	Male	Latino	30 - 34	International	Master's	Other	To a large extent	Very often
5/11/2022 22:21:15	Male	Latino	35 - 39	International	Undergraduate	Psychology	To a large extent	Sometimes
5/11/2022 22:24:47	Male	White	20 - 24	EU/UK	Undergraduate	Other	Somewhat	Fairly often
5/11/2022 22:27:52	Male	Latino	30 - 34	International	Master's	Other	To a large extent	Sometimes
5/11/2022 22:34:49	Male	Latino	35 - 39	International	Master's	Engineering and technolo	To a small extent	Sometimes
5/11/2022 23:12:55	Female	Latino	30 - 34	International	Master's	Psychology	Somewhat	Sometimes
5/11/2022 23:18:41	Male	Latino	25 - 29	International	Master's	Engineering and technolo	To a large extent	Sometimes
5/11/2022 23:29:31	Male	White	Under 21	International	Master's	Eduation	To a large extent	Fairly often
5/11/2022 23:32:34	Male	Latino	25 - 29	International	Undergraduate	Law	To a large extent	Fairly often
5/12/2022 0:03:20	Male	Latino	30 - 34	International	Master's	Psychology	To a large extent	Very often
5/12/2022 0:53:06	Female	Latino	30 - 34	International	Undergraduate	Other	Somewhat	Sometimes
5/12/2022 7:28:35	Female	Latino	20 - 24	International	Undergraduate	Other	To a large extent	Sometimes
5/12/2022 11:51:29	Female	Latino	30 - 34	International	Master's	Other	To a large extent	Very often
5/12/2022 14:23:48	Female	Latino	30 - 34	International	Master's	Social sciences	Completely	Very often
5/12/2022 14:29:37	Male	Latino	25 - 29	International	Undergraduate	Arts and humanities	To a large extent	Sometimes
5/12/2022 15:34:06	Female	Latino	30 - 34	International	Undergraduate	Psychology	Completely	Sometimes
5/12/2022 19:11:21	Female	Latino	25 - 29	International	Undergraduate	Other	To a large extent	Sometimes
5/13/2022 0:18:49	Female	Latino	35 - 39	International	Undergraduate	Computer science	Completely	Very often
5/13/2022 2:57:40	Female	Latino	40 and over	International	Masters	Social sciences	Somewhat	Fairly often
5/13/2022 3:04:18	Female	Latino	30 - 34	International	Undergraduate	Other	Somewhat	Sometimes
5/13/2022 10:12:27	Female	White	30 - 34	International	Undergraduate	Engineering and technolo	Somewhat	Sometimes
5/13/2022 10:45:30	Female	White	20 - 24	EU/UK	Undergraduate	Physical science	To a large extent	Sometimes
5/13/2022 10:54:40	Female	White	30 - 34	EU/UK	Undergraduate	Clinical, pre-clinical and h	Completely	Very often
5/13/2022 11:05:48	Female	White	20 - 24	EU/UK	Undergraduate	Education	To a large extent	Very often
5/13/2022 11:33:30	Male	Latino	40 and over	International	Undergraduate	Computer science	To a large extent	Very often
5/13/2022 12:14:40	Female	Asian	30 - 34	International	Undergraduate	Computer science	To a large extent	Sometimes
5/13/2022 12:35:26	Male	Latino	25 - 29	International	Undergraduate	Computer science	Completely	Sometimes
5/13/2022 14:55:17	Female	Latino	35 - 39	International	Undergraduate	Computer science	Completely	Very often
5/13/2022 15:03:11	Female	White	35 - 39	International	Undergraduate	Engineering and technolo	To a large extent	Always
5/13/2022 15:20:14	Male	Latino	40 and over	International	PhD	Computer science	Somewhat	Sometimes
5/13/2022 16:15:17	Female	Latino	35 - 39	International	Masters	Computer science	To a large extent	Sometimes
5/13/2022 22:45:42	Male	Black	30 - 34	International	Undergraduate	Computer science	To a large extent	Very often
5/16/2022 13:36:38	Female	Latino	30 - 34	International	Masters	Social sciences	Somewhat	Sometimes
5/16/2022 13:40:08	Other	White	25 - 29	EU/UK	PhD	Engineering and technolo	Completely	Sometimes

Headaches	Digestive problems	Anxiety or tension	Sleep problems	Rapid heartbeat or palpita	Irritability	Concentration problems	Sadness or tearfulness	Illness
Sometimes	Sometimes	Very often	Very often	Very often	Very often	Very often	Very often	Fairly often
Almost never	Very often	Very often	Sometimes	Almost never	Very often	Very often	Very often	Never
Almost never	Almost never	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Fairly often	Almost never
Fairly often	Fairly often	Sometimes	Very often	Almost never	Sometimes	Very often	Fairly often	Almost never
Very often	Sometimes	Very often	Sometimes	Almost never	Very often	Very often	Sometimes	Sometimes
Very often	Very often	Very often	Sometimes	Fairly often	Sometimes	Very often	Sometimes	Sometimes
Very often	Fairly often	Sometimes	Very often	Sometimes	Very often	Very often	Sometimes	Almost never
Fairly often	Very often	Very often	Very often	Sometimes	Very often	Fairly often	Fairly often	Sometimes
Almost never	Very often	Very often	Sometimes	Very often	Very often	Very often	Very often	Sometimes
Sometimes	Almost never	Sometimes	Sometimes	Never	Almost never	Sometimes	Sometimes	Sometimes
Never	Almost never	Almost never	Almost never	Almost never	Almost never	Fairly often	Almost never	Never
Sometimes	Fairly often	Sometimes	Very often	Sometimes	Fairly often	Very often	Sometimes	Almost never
Fairly often	Fairly often	Never	Sometimes	Never	Almost never	Sometimes	Fairly often	Very often
Fairly often	Almost never	Fairly often	Never	Never	Fairly often	Fairly often	Fairly often	Fairly often
Almost never	Almost never	Sometimes	Sometimes	Almost never	Sometimes	Fairly often	Sometimes	Fairly often
Almost never	Almost never	Fairly often	Sometimes	Sometimes	Very often	Very often	Never	Almost never
Fairly often	Almost never	Very often	Almost never	Fairly often	Sometimes	Very often	Fairly often	Fairly often
Almost never	Never	Almost never	Never	Never	Never	Fairly often	Never	Almost never
Sometimes	Almost never	Sometimes	Very often	Fairly often	Almost never	Sometimes	Almost never	Almost never
Fairly often	Sometimes	Fairly often	Fairly often	Almost never	Fairly often	Very often	Almost never	Sometimes
Sometimes	Sometimes	Sometimes	Fairly often	Very often	Fairly often	Very often	Very often	Fairly often
Almost never	Almost never	Sometimes	Sometimes	Almost never	Sometimes	Very often	Sometimes	Fairly often
Sometimes	Sometimes	Very often	Sometimes	Almost never	Sometimes	Very often	Sometimes	Fairly often
Almost never	Almost never	Fairly often	Fairly often	Almost never	Fairly often	Fairly often	Almost never	Almost never
Very often	Sometimes	Very often	Very often	Sometimes	Sometimes	Very often	Sometimes	Sometimes
Sometimes	Fairly often	Sometimes	Very often	Sometimes	Very often	Very often	Very often	Sometimes
Sometimes	Very often	Very often	Sometimes	Sometimes	Very often	Very often	Very often	Fairly often
Sometimes	Almost never	Fairly often	Fairly often	Fairly often	Sometimes	Sometimes	Fairly often	Almost never
Almost never	Very often	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes
Almost never	Never	Very often	Fairly often	Never	Sometimes	Sometimes	Almost never	Never
Almost never	Very often	Very often	Sometimes	Fairly often	Sometimes	Very often	Fairly often	Sometimes
Almost never	Almost never	Fairly often	Sometimes	Almost never	Fairly often	Fairly often	Almost never	Almost never
Sometimes	Sometimes	Sometimes	Fairly often	Fairly often	Sometimes	Very often	Sometimes	Fairly often
Never	Almost never	Sometimes	Almost never	Never	Fairly often	Very often	Fairly often	Sometimes
Almost never	Fairly often	Very often	Sometimes	Almost never	Almost never	Very often	Fairly often	Almost never
Sometimes	Very often	Very often	Very often	Very often	Very often	Sometimes	Almost never	Almost never
Sometimes	Almost never	Very often	Very often	Almost never	Very often	Very often	Very often	Sometimes
Very often	Very often	Very often	Very often	Very often	Very often	Very often	Very often	Sometimes
Never	Never	Very often	Sometimes	Sometimes	Almost never	Almost never	Sometimes	Sometimes
Never	Almost never	Very often	Sometimes	Almost never	Very often	Very often	Very often	Never
Sometimes	Sometimes	Always	Sometimes	Sometimes	Very often	Always	Sometimes	Almost never
Always	Always	Always	Always	Always	Always	Always	Always	Sometimes
Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes
Almost never	Sometimes	Very often	Very often	Almost never	Very often	Very often	Very often	Sometimes
Very often	Very often	Sometimes	Very often	Very often	Sometimes	Very often	Very often	Sometimes
Sometimes	Never	Almost never	Sometimes	Almost never	Sometimes	Very often	Sometimes	Very often
Almost never	Almost never	Always	Very often	Very often	Very often	Almost never	Almost never	Very often

Aches and pains not due	Loneliness	Please choose your copin	Please choose your copin	Please choose your copin	Please choose your copin	Please choose your copin	Please choose your copin	Please choose your copin
Sometimes	Sometimes					Options	Options	
Fairly often	Never	Options	Options	Options	Options	Options	Options	Options
Fairly often	Fairly often	Practicing art or hobbies, Eatiing, Playing games / watching Tv, Socializing, Taking drugs, Talking about it(with friends, family or professionals)						
Fairly often	Sometimes	Practicing art or hobbies, Drinking alcohol, Eatiing, Meditating, Talking about it(with friends, family or professionals)						
Sometimes	Sometimes	Practicing art or hobbies, Excercising / sports, Eatiing, Playing games / watching Tv, Meditating, Spending time with pets, Socializing, Talking about it(with friends, family or professional						
Fairly often	Sometimes	Taking antidepressants, Practicing art or hobbies, Drinking alcohol, Eatiing, Talking about it(with friends, family or professionals)						
Sometimes	Almost never	Practicing art or hobbies, Excercising / sports, Playing games / watching Tv, Spending time with pets, Taking drugs						
Sometimes	Almost never	Eatiing, Religion, Nothing						
Very often	Almost never	Taking antidepressants, Playing games / watching Tv						
Sometimes	Never	Practicing art or hobbies, Drinking alcohol, Excercising / sports, Playing games / watching Tv, Meditating, Spending time with pets, Socializing, Talking about it(with friends, family or pro						
Almost never	Never	Practicing art or hobbies, Excercising / sports, Playing games / watching Tv, Socializing, Talking about it(with friends, family or professionals)						
Sometimes	Sometimes	Drinking alcohol, Playing games / watching Tv, Spending time with pets, Socializing, Talking about it(with friends, family or professionals)						
Sometimes	Sometimes	Practicing art or hobbies, Drinking alcohol, Excercising / sports, Eating, Playing games / watching Tv, Meditating, Spending time with pets, Socializing, Talking about it(with friends, famil						
Fairly often	Never	Practicing art or hobbies						
Almost never	Very often	Playing games / watching Tv						
Almost never	Sometimes	Practicing art or hobbies, Drinking alcohol, Excercising / sports, Eating, Playing games / watching Tv, Socializing, Talking about it(with friends, family or professionals)						
Fairly often	Sometimes	Practicing art or hobbies, Excercising / sports, Eating, Playing games / watching Tv, Spending time with pets, Socializing						
Never	Never	Practicing art or hobbies, Excercising / sports, Playing games / watching Tv						
Fairly often	Almost never	Practicing art or hobbies, Eating, Playing games / watching Tv, Spending time with pets, Talking about it(with friends, family or professionals)						
Sometimes	Very often	Practicing art or hobbies, Drinking alcohol, Eating, Playing games / watching Tv, Socializing						
Fairly often	Sometimes	Excercising / sports, Eating, Playing games / watching Tv						
Fairly often	Sometimes	Practicing art or hobbies, Excercising / sports, Playing games / watching Tv, Spending time with pets, Religion, Socializing, Taking drugs, Talking about it(with friends, family or professic						
Almost never	Fairly often	Eating, Talking about it(with friends, family or professionals)						
Never	Almost never	Practicing art or hobbies, Excercising / sports, Eating, Playing games / watching Tv, Spending time with pets, Religion, Socializing						
Fairly often	Sometimes	Practicing art or hobbies, Excercising / sports, Playing games / watching Tv, Meditating, Socializing						
Sometimes	Almost never	Eating, Socializing, Talking about it(with friends, family or professionals)						
Fairly often	Very often	Drinking alcohol, Excercising / sports, Eating, Meditating, Spending time with pets, Socializing, Taking drugs, Talking about it(with friends, family or professionals)						
Almost never	Almost never	Practicing art or hobbies, Drinking alcohol, Excercising / sports, Eating, Meditating, Taking drugs						
Sometimes	Sometimes	Drinking alcohol, Meditating, Talking about it(with friends, family or professionals)						
Never	Sometimes	Excercising / sports, Eating, Meditating, Socializing, Talking about it(with friends, family or professionals)						
Almost never	Never	Taking antidepressants, Practicing art or hobbies, Drinking alcohol, Eating, Playing games / watching Tv, Meditating, Spending time with pets, Socializing, Talking about it (with friends, f						
Sometimes	Never	Eating, Playing games / watching Tv, Socializing, Talking about it (with friends, family or professionals)						
Sometimes	Sometimes	Exercising / sports, Playing games / watching Tv						
Never	Sometimes	Practicing art or hobbies, Meditating, Spending time with pets, Taking drugs						
Almost never	Almost never	Eating, Meditating, Spending time with pets, Talking about it (with friends, family or professionals)						
Never	Sometimes	Practicing art or hobbies, Exercising / sports, Eating, Spending time with pets, Talking about it (with friends, family or professionals)						
Almost never	Very often	Exercising / sports, Eating, Playing games / watching Tv, Spending time with pets, Socializing, Talking about it (with friends, family or professionals)						
Sometimes	Very often	Practicing art or hobbies, Eating, Playing games / watching Tv, Religion						
Almost never	Sometimes	Exercising / sports, Eating, Playing games / watching Tv, Meditating, Socializing, Talking about it (with friends, family or professionals)						
Never	Never	Nothing						
Almost never	Sometimes	Exercising / sports, Eating, Playing games / watching Tv						
Sometimes	Always	Taking antidepressants, Exercising / sports, Eating, Playing games / watching Tv, Talking about it (with friends, family or professionals)						
Sometimes	Sometimes	Taking antidepressants, Drinking alcohol, Eating, Playing games / watching Tv, Meditating, Spending time with pets, Religion, Socializing						
Almost never	Fairly often	Drinking alcohol, Exercising / sports, Spending time with pets, Taking drugs, Talking about it (with friends, family or professionals)						
Sometimes	Fairly often	Exercising / sports, Eating, Playing games / watching Tv						
Very often	Never	Practicing art or hobbies, Exercising / sports, Playing games / watching Tv						
Very often	Sometimes	Taking antidepressants, Eating, Playing games / watching Tv						

Please choose your copir	Please choose your copin	Please choose your copin	Please choose your copin	Please choose your copin	Please choose your copin	Do you feel that your copi	Do you feel that yo are ge	Feeling overloaded with u
					Options	Not sure	Not sure	Very often
Options		Options	Options	Options		Not sure	Yes	Very often
						Yes	No	Sometimes
						Yes	No	Sometimes
s)						Yes	Not sure	Sometimes
						Not sure	No	Sometimes
						Yes	No	Sometimes
						Not sure	Not sure	Sometimes
						Yes	No	Almost never
fessionals)						Yes	Not sure	Sometimes
						Yes	Yes	Almost never
						Not sure	No	Fairly often
ly or professionals)						Yes	Yes	Sometimes
						Not sure	No	Sometimes
						Yes	No	Very often
						Yes	No	Fairly often
						Not sure	No	Sometimes
						Yes	Yes	Almost never
						Yes	Not sure	Sometimes
						No	No	Fairly often
						Yes	No	Sometimes
onals)						Yes	No	Sometimes
						Not sure	No	Sometimes
						Yes	Not sure	Fairly often
						Yes	Not sure	Sometimes
						Yes	No	Very often
						Not sure	Yes	Very often
						Yes	No	Sometimes
						Yes	Not sure	Sometimes
						Yes	No	Sometimes
family or professionals)						Yes	Yes	Sometimes
						Yes	Not sure	Fairly often
						Not sure	Not sure	Sometimes
						Yes	Not sure	Very often
						Yes	Yes	Very often
						Yes	No	Very often
						No	No	Very often
						Yes	Not sure	Very often
						Yes	No	Very often
						Not sure	No	Always
						Not sure	No	Very often
						Yes	Not sure	Always
						Not sure	Not sure	Sometimes
						Not sure	Not sure	Very often
						No	Yes	Always
						Not sure	Yes	Very often
						No	Not sure	Always

Spending too much time (	Competition with peers	Difficulties with supervisor	Unpleasant working envir	Criticism about work	Lack of time for relaxation	Difficult home environmer	Financial issues	Lack of confidence with a
Almost never	Fairly often	Never	Very often	Sometimes	Very often	Never	Fairly often	Fairly often
Never	Sometimes	Sometimes	Never	Fairly often	Almost never	Never	Sometimes	Almost never
Almost never	Fairly often	Almost never	Fairly often	Very often	Very often	Sometimes	Very often	Fairly often
Sometimes	Fairly often	Almost never	Fairly often	Sometimes	Very often	Almost never	Sometimes	Very often
Fairly often	Fairly often	Sometimes	Sometimes	Sometimes	Fairly often	Fairly often	Very often	Sometimes
Almost never	Sometimes	Fairly often	Almost never	Fairly often	Sometimes	Sometimes	Very often	Sometimes
Sometimes	Almost never	Sometimes	Very often	Very often	Very often	Fairly often	Sometimes	Sometimes
Fairly often	Sometimes	Almost never	Very often	Sometimes	Sometimes	Almost never	Very often	Very often
Never	Sometimes	Never	Very often	Sometimes	Almost never	Almost never	Never	Sometimes
Never	Almost never	Sometimes	Almost never	Never	Almost never	Never	Almost never	Sometimes
Almost never	Almost never	Fairly often	Fairly often	Fairly often	Almost never	Never	Almost never	Almost never
Sometimes	Sometimes	Fairly often	Never	Never	Fairly often	Never	Very often	Sometimes
Sometimes	Almost never	Almost never	Almost never	Fairly often	Never	Sometimes	Sometimes	Sometimes
Fairly often	Fairly often	Almost never	Almost never	Almost never	Sometimes	Almost never	Fairly often	Sometimes
Sometimes	Sometimes	Very often	Very often	Very often	Sometimes	Very often	Very often	Sometimes
Almost never	Sometimes	Fairly often	Sometimes	Almost never	Very often	Almost never	Very often	Sometimes
Fairly often	Fairly often	Fairly often	Fairly often	Sometimes	Sometimes	Almost never	Fairly often	Almost never
Never	Sometimes	Fairly often	Sometimes	Sometimes	Almost never	Almost never	Fairly often	Sometimes
Almost never	Never	Never	Never	Never	Fairly often	Almost never	Fairly often	Fairly often
Almost never	Sometimes	Almost never	Very often	Almost never	Fairly often	Very often	Very often	Very often
Sometimes	Very often	Fairly often	Sometimes	Fairly often	Fairly often	Fairly often	Almost never	Almost never
Fairly often	Sometimes	Very often	Very often	Fairly often	Fairly often	Very often	Very often	Sometimes
Fairly often	Very often	Sometimes	Almost never	Almost never	Very often	Never	Very often	Sometimes
Fairly often	Fairly often	Almost never	Fairly often	Almost never	Fairly often	Almost never	Almost never	Fairly often
Sometimes	Sometimes	Almost never	Sometimes	Fairly often	Almost never	Fairly often	Sometimes	Sometimes
Very often	Very often	Almost never	Almost never	Fairly often	Fairly often	Sometimes	Very often	Very often
Sometimes	Very often	Sometimes	Sometimes	Very often	Very often	Fairly often	Very often	Sometimes
Sometimes	Fairly often	Fairly often	Fairly often	Fairly often	Sometimes	Sometimes	Almost never	Fairly often
Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes
Fairly often	Sometimes	Sometimes	Almost never	Fairly often	Fairly often	Sometimes	Sometimes	Sometimes
Almost never	Never	Almost never	Never	Almost never	Very often	Never	Sometimes	Sometimes
Almost never	Almost never	Almost never	Almost never	Almost never	Almost never	Almost never	Fairly often	Almost never
Fairly often	Almost never	Fairly often	Fairly often	Fairly often	Sometimes	Sometimes	Fairly often	Sometimes
Never	Sometimes	Almost never	Sometimes	Fairly often	Very often	Almost never	Very often	Fairly often
Sometimes	Sometimes	Fairly often	Almost never	Almost never	Fairly often	Never	Almost never	Fairly often
Very often	Very often	Very often	Sometimes	Sometimes	Sometimes	Sometimes	Very often	Very often
Very often	Very often	Very often	Sometimes	Sometimes	Very often	Sometimes	Sometimes	Very often
Sometimes	Almost never	Almost never	Sometimes	Sometimes	Very often	Sometimes	Sometimes	Very often
Very often	Sometimes	Sometimes	Sometimes	Almost never	Sometimes	Sometimes	Almost never	Always
Sometimes	Very often	Sometimes	Very often	Sometimes	Always	Sometimes	Very often	Always
Almost never	Almost never	Sometimes	Sometimes	Almost never	Sometimes	Sometimes	Always	Almost never
Almost never	Almost never	Very often	Never	Never	Sometimes	Very often	Never	Sometimes
Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	Never
Very often	Sometimes	Sometimes	Almost never	Sometimes	Always	Almost never	Sometimes	Sometimes
Very often	Almost never	Sometimes	Very often	Sometimes	Sometimes	Sometimes	Very often	Very often
Very often	Almost never	Sometimes	Sometimes	Sometimes	Very often	Sometimes	Very often	Sometimes
Always	Very often	Sometimes	Very often	Always	Always	Almost never	Never	Very often

Lack of confidence with s	Conflicts between study a	Please describe anything else that has influenced your stress/anxiety levels over the last academic year
Fairly often	Sometimes	Being an Immigrant
Never	Very often	nothing else
Almost never	Very often	Pnademic, work, visa.
Very often	Very often	Work and College at the same time
Sometimes	Very often	Work and study
Sometimes	Sometimes	Modifications to restrictions regarding COVID-19, inability to see my family due to restricted travel.
Sometimes	Sometimes	Covid 19 and family health problems
Very often	Sometimes	Being away from home
Almost never	Never	I've had depression my entire life
Sometimes	Never	Exams
Almost never	Fairly often	Toxic people
Sometimes	Sometimes	Covid was quite stressful to deal with while trying to study for college. Distance learning is tough
Very often	Very often	I was working in a newspaper full time. It was scary
Very often	Never	Post pandemic Emotional effect
Very often	Very often	schedule conflict
Fairly often	Very often	Nothing
Almost never	Fairly often	
Almost never	Almost never	deadlines
Almost never	Fairly often	.
Very often	Very often	Parents
Almost never	Fairly often	Homework
Sometimes	Very often	The fact that my home enviourement requieres most of my Time because i live with elder people so there are some responsabilities exclusiblely mine to take care of
Almost never	Very often	Not getting enough sleep
Fairly often	Fairly often	Deadlines
Sometimes	Sometimes	Nothin else
Sometimes	Sometimes	social media
Sometimes	Sometimes	Having to work and study
Almost never	Fairly often	Problems with friends and Flatmates
Sometimes	Sometimes	Pandemic
Sometimes	Almost never	.
Never	Very often	Covid
Almost never	Fairly often	Income issues and economical situation.
Sometimes	Almost never	Need to get my degree to get a job
Never	Very often	nothing else
Fairly often	Almost never	Na
Very often	Very often	na
Very often	Very often	My learning difficulties not being properly handled
Sometimes	Sometimes	Pandemic
Always	Always	None
Always	Always	Covid
Sometimes	Almost never	Over thinking about my future visa situation, and how difficult is going to be to get work permit.
Never	Almost never	Pressure
Never	Sometimes	Problems on love relationship
Sometimes	Always	Covid
Sometimes	Very often	rental prices
Almost never	Sometimes	I dont know
Never	Never	I like study but i dont have time

Appendix C: Reflective Journal

This analysis is the opportunity to identify and reflect on the elements that help us to achieve the culmination of this project and to expand our knowledge of the subject.

At the beginning of this project finding the right idea to develop the business case and the machine learning artifact was challenging. Trying to find the balance between the skills we have in coding and different languages and a relevant topic that everybody liked was a difficult task.

Some of the ideas were:

- Predicting the number of houses needed to help with the housing crisis in Ireland.
- Predict the level of birth rates in Japan by 2030
- Predict the winning numbers for the lottery
- Predict the winning team for the World Cup.

These ideas weren't adequate, some of them were already being done by other groups in the class and some other were unfit for the project.

One of our group members, Valentina, concentrated on the concept of finding an idea which we can all relate.

The idea chosen was to predict levels of stress on IT students.

Finding a dataset related to it with a good enough number of entries was a little bit challenging. This needed to have a solid base, and it had to be big enough to train the model for the machine learning.

Going through each step of the CRISP-DM Framework helped the group to have a better understanding of the Data Mining Process and how to improve the approaches to come up with better results for the models.

The data preparation is evidently one of the longest steps in the machine learning and in our case some changes were necessary and a bit tedious, like the code scheme, to convert attributes into numerical data.

The modelling was also a long and arduous process that with the help of the lecturer Muhammad Iqbal we were able to know how to improve constantly, learning new concepts like underfitting or overfitting models, reaching better accuracy levels and more. We are really grateful for the guidance that the lecturer gave to us week by week.

The elements in the dataset were used as a guide to create and adapt a survey for the new dataset. With the survey created the group aimed to understand the different expressions of the human brain.

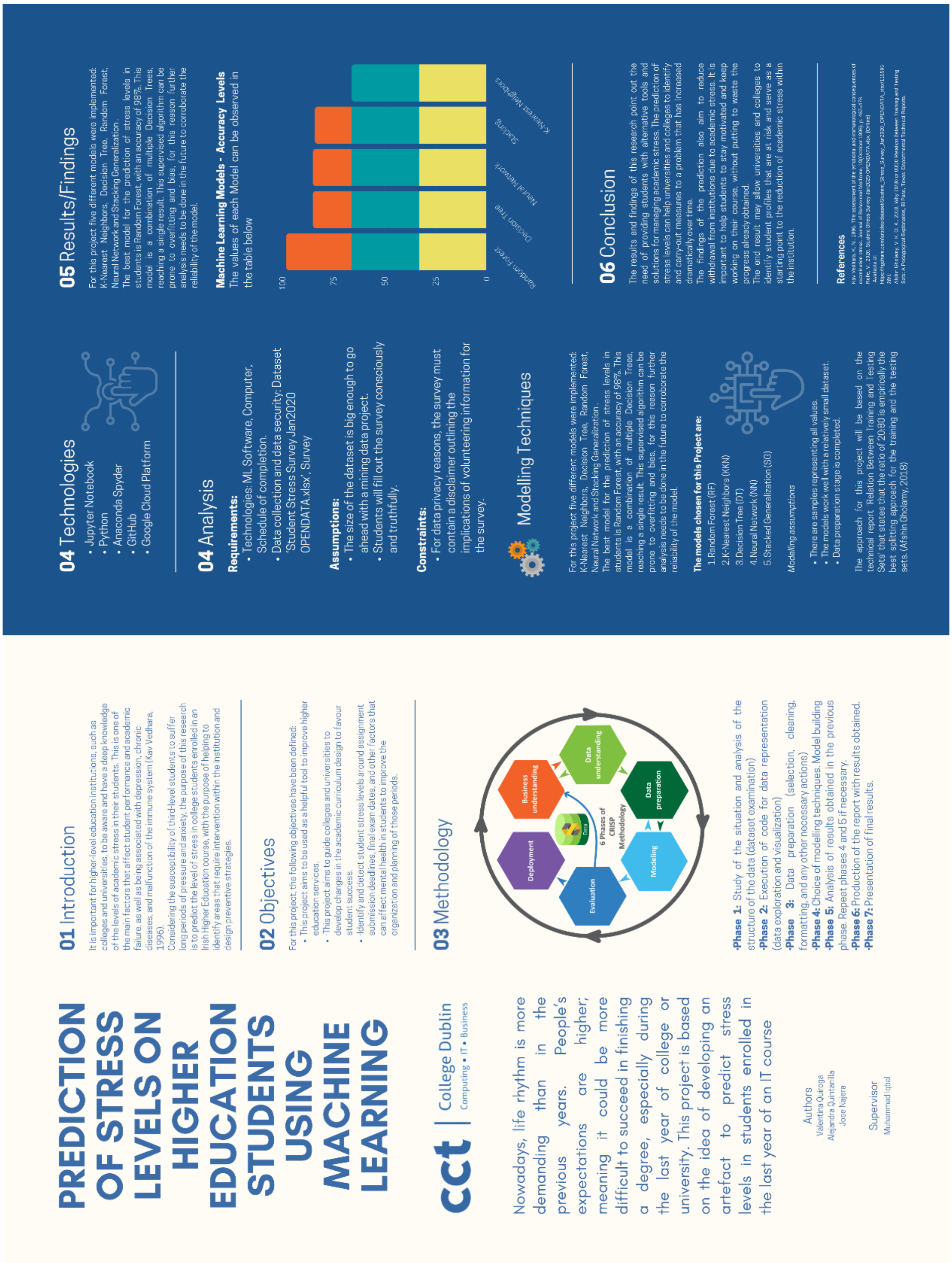
Working together as a team was an easy task and made the project run smoothly, each member contributed to different areas where each has the right skills to accomplish the task, finishing the project on time is due to all of us.

This was a reflective and learning journey that we all have enjoyed and also suffered. But we are sure that all is going to be worth it in the near future.

Appendix D: Extras

GitHub Link: https://github.com/Dani-elaqh/PiRates.Stress

Poster:



References

H2O.ai, 2022. *Description: max_depth.* [Online]
Available at: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/algo-params/max_depth.html
[Accessed 16 May 2022].

Afshin Gholamy, V. K. O. K., 2018. *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation,* El Paso, Texas: Departmental Technical Reports.

Arnaldo, M., 2021. *Intro to Rapidminer: A No-Code Development Platform for Data Mining (with Case Study).* [Online]
Available at: https://www.analyticsvidhya.com/blog/2021/10/intro-to-rapidminer-a-no-code-development-platform-for-data-mining-with-case-study/
[Accessed 06 March 2022].

Brownlee, J., 2016. *How To Implement Machine Learning Metrics From Scratch in Python.* [Online]
Available at: https://machinelearningmastery.com/implement-machine-learning-algorithm-performance-metrics-scratch-python/#:~:text=Root%20Mean%20Squared%20Error,-Another%20popular%20way&text=RMSE%20is%20calculated%20as%20the,the%20original%20units%20for%20comparison.
[Accessed 16 May 2022].

Brownlee, J., 2021. *A Gentle Introduction to Machine Learning Modeling Pipelines.* [Online]
Available at: https://machinelearningmastery.com/machine-learning-modeling-pipelines/
[Accessed 16 May 2022].

Georgiou, M., 2021. *Best Strategies to Conduct Market Research for Your Mobile Application Idea.* [Online]
Available at: https://imaginovation.net/blog/market-research-strategies-mobile-app-idea/
[Accessed 6 March 2022].

Gonfalonieri, A., 2019. *5 Ways to Deal with the Lack of Data in Machine Learning.* [Online]
Available at: https://www.kdnuggets.com/2019/06/5-ways-lack-data-machine-learning.html
[Accessed 16 May 2022].

Google Open Source, 2022. *TensorFlow.* [Online]
Available at: https://opensource.google/projects/tensorflow
[Accessed 10 March 2022].

Google Trends, 2021. *Data analyst jobs.* [Online]
Available at: https://trends.google.com/trends/explore?date=all&q=Data%20analyst%20jobs
[Accessed 14 December 2021].

IBM, 2021. *IBM.* [Online]
Available at: https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=preparation-data-overview
[Accessed 12 MAy 2022].

IBM, 2021. *IBM Documentation.* [Online]
Available at: https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=guide-data-understanding
[Accessed 12 May 2022].

Jonathan D Quick MD, R. S. H. &. J. C. Q. P., 1987. Health Consequences of Stress. *Journal of Organizational Behavior Management,* 8(2), pp. 19-36.

Jupyter, 2022. *Jupyter Notebook Documentation.* [Online]
Available at: https://buildmedia.readthedocs.org/media/pdf/jupyter-notebook/latest/jupyter-notebook.pdf
[Accessed 10 March 2022].

Kav Vedhara, K. N., 1996. The assessment of the emotional and immunological consequences of examination stress. *Journal of Behavioral Medicine,* 19(October 1996), p. 467–478.

Korstanje, J., 2021. *The k-Nearest Neighbors (kNN) Algorithm in Python.* [Online]
Available at: https://realpython.com/knn-python/#author
[Accessed 16 May 2022].

Kotak, P., 2021. *DaskGridSearchCV – A competitor for GridSearchCV.* [Online]
Available at: https://www.geeksforgeeks.org/daskgridsearchcv-a-competitor-for-gridsearchcv/
[Accessed 16 May 2022].

Microsoft, 2021. *What is Azure Machine Learning?.* [Online]
Available at: https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-machine-learning#:~:text=Azure%20Machine%20Learning%20is%20for,day%2Dto%2Dday%20workflows.
[Accessed 10 March 2022].

Nyuytiymbiy, K., 2020. *Parameters and Hyperparameters in Machine Learning and Deep Learning.* [Online]
Available at: https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac
[Accessed 16 May 2022].

Rolfe, V., 2020. *Student Stress Survey Jan2020 OPENDATA.xlsx.* [Online]
Available at:
https://figshare.com/articles/dataset/Student_Stress_Survey_Jan2020_OPENDATA_xlsx/11559528/1
[Accessed 16 May 2022].

Simon Neubauer, J.-J. H. P. G., 2018. *The evolution of modern human brain shape.* [Online]
Available at:
https://pubmed.ncbi.nlm.nih.gov/29376123/#:~:text=Our%20data%20show%20that%2C%20300%2C000,100%2C000%20and%2035%2C000%20years%20ago.
[Accessed 10 March 2022].

Software ARGE Inc., 2022. *Rapidminer.* [Online]
Available at: https://www.softwarearge.com/products/rapidminer/?lang=en
[Accessed 06 March 2022].

Tavish, 2015. *Tuning the parameters of your Random Forest model.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/#:~:text=n_estimators%20%3A,but%20makes%20your%20code%20slower.
[Accessed 15 May 2022].