Delving into the Key Characteristics of Procrastination

Using Exploratory Data Analysis & Machine Learning Algorithms

By
Dominic Lai

Student ID: 2379299

A thesis submitted to the University of Birmingham for the degree of MSc in Data Science

Supervisor: Russell Beale

School of Computer Science
University of Birmingham
Birmingham, UK

September 2022

Abstract

Procrastination, the action of delaying or postponing something, is a daily neglection for many people. Ubiquitous, it has always been an issue in the past, stemming all the way up to the world we know today. As with most studies, Harriot & Ferrari (1996) can support this, finding that the phenomenon that is procrastination affects most of the general population, while tending to linger amongst significant numbers of adults [1]. They also found that in another study, this time by McCown and Johnson (1989), of the adults they surveyed, over 25% reported that the effects of procrastination caused substantial problems for their daily lives. Adding to this, near 40% personally disclosed financial losses as a consequence of procrastinating [1].

We conducted an investigation into the background of this strange behavioural phenomena. To inform our understanding, we have explored the different varieties of reasonings and causations through past research and works.

Then using datasets of procrastination, stress, and motivation (from open sources), we performed a number of data cleansing techniques. This was followed by an approach, applying exploratory data analysis (EDA) to each one of the datasets, in order to delve into initial investigations inspecting the distinct features. Through exploratory data analysis, we are able to create a firm data foundation, backed by creating insightful visualisations using an assortment of plots and graphs; here discovering patterns, trends and even probable correlations that show up. This can also further aid us in being cautious with spotting potential anomalies.

Finally, by using supervised learning models, we made use of algorithms, more specifically regarding regression models, such, linear regression, logistic regression, decision trees and random forests. By helping us define relationships between the variables and features, these techniques allowed us to conclude the most significant features within each respective dataset, identifying which ones played the greatest roles and which ones could potentially not be needed (or irrelevant). Having found the importance of each feature, we are then able to turn back to the literatures and past studies, making comparative suggestions, seeing how our findings may match against what other people have done before, whether this be aligning with what they have previously found, or whether we have indeed discovered something different.

Using this research, data exploration and findings, the overall goal is to provide useful and actionable insights. With a sparse amount of work into procrastination, especially from a programming approach, insights which will contribute to understanding the phenomenon, using the information we may suggest future work, aiding the battle for procrastination avoidance, all of which is backed-up by the data.

Acknowledgments

I would like to take this time to wholeheartdly express my gratitude and appreciation to my supervisor Dr Russell Beale, for his exceptional guidence, patience and support for the entirety of the duration for this project. His advice and mentorship have been second to none, and I would not have been able to complete this research without it.

I would also like to give a generous amount of recognition and thanks to my second marker Dr Peter Hancox, who also provided me with invaluable advice and insights, which helped to visualise and shape the direction of my project.

Table of Contents

List of figures

1. Introduction

Everyone procrastinates. You do it, I do it, and it is quite evident that we all struggle with avoiding the effects of this strange behavioural and psychological phenomenon. It seems only natural for all humans to want to avoid both off-putting major and minor tasks at hand, especially if they are ones that cause us substantial distress.

Many past researchers have addressed the effects of procrastination claim it can be explained by the theory of basic behaviourism, and Lieberman (2019) is an example of one study that supports this [2]. The idea of basic behaviourism as the name suggests, is that all behaviours we know and act upon, are taught through environments, including an array of interactions with people and situations, acquired by conditioning [3].

On one hand it seems to make a lot of sense, that we would engage with these effects of procrastination, similarly to that of a 'fight or flight' response.

However, we could argue when we procrastinate, we are essentially engaging in active avoidance in order to 'feel better', to escape all this anguish we feel – but in fact, these problems that we try to evade have not really gone anywhere.

Even when we don't want to, we can still find ourselves falling prey to the effects of procrastination. Zarick & Stonebraker (2009) recognise this, saying that: "Although countless students repeatedly swear they will never procrastinate again, they inevitably do" [5].

Essentially all previous work has looked into many reasonings and causations for the effects of procrastination. With most of them bearing similar resemblance, it is especially apparent why this problem is not only very real but serious.
Tice & Baumeister (1997) study focuses to the nature of procrastination towards performance, stress, and health – particularly looking at what kinds of effects procrastination has on our health, varyingly described as "harmful, innocuous, or even beneficial" [4].

Overall, we believe strongly that there is still more work that needs to be done in understanding procrastination, moreover, that it is a topic that deserves greater attention worth solving.

1.1 – Outline and Structure of the Report

The outline of this report will be structured as follows:

> Section 2 we will go over the all the necessary background needed, providing context for the topic of this project. Looking into the core aims and objectives of the project that were set out and we wish to be accomplished; and any initial brief problems that we may have come across along the way. Furthermore, going through the methodology and how we plan to approach the problem at hand.
> Section 3 we will produce a literature review, discussing the summary and giving evaluation to all the current and past research and works linking to the topic.
> Section 4 we will give general background looking at the model selections and the reasoning behind the potential choices; in brief going over the mathematic metrics and their respective formulae.
> Section 5 we discuss the data of the project, its origin and how we came across it; why we chose those datasets and the rationale behind them; any challenges that came along with it; the preparation necessary to go further with it for exploration.
> Section 6 we examine the exploratory data analysis for the datasets found, delving into the initial investigations, establishing potential patterns and trends, and also do our best to discover correlations between certain features.
> Section 7 we practice and test our models, using a variety of different machine learning algorithms (which will be dependent on the knowledge learnt from exploratory data analysis), aiming to model relationships and dependencies, in order to depict the most important features that manage to predict and preserve the most data.

Section 8 we will discuss through the results from the models, their findings, and accuracies, interpreting what they mean and whether these lead towards beneficial insights.
Section 9 we will summarise and discuss comparisons between our own findings and those in the past literature, whether that be that the results yield the same, or if in fact we have found something different.

2. Background and Methodology

In this section, we will discuss over the necessary background in order to better understand as you read through the report. Starting with the core aims and objectives, through to the methodology and process in which we went through to achieve our results.

2.1 – Aims and objectives

Bearing the topic and main ideas in mind, to ensure the project progresses smoothly it was imperative to make sure the aims and objectives where clearly defined.

The aim of the work is to provide insights and understanding to better provide guidance for future work; all in the name of avoiding the effects of procrastination.

The main objectives are:

- to ascertain which variables and related features may influence procrastination and stress.
- to be establish the importance of each feature and how much of an encouraging they play in the majority when comparing results.
- to recognize relationships and correlations embedded within the data.
- finally, as remarked with the main aim, providing insights that will be key messages and takeaways found via an exploration of, and backed-up by the data.

2.2 – Initial Problems

As for any project, it would be uncommon not to encounter problems throughout the process. Even with the right amount of appropriate preparation beforehand and a substantial amount of time spent of research for gaining an understanding to the matter, it is inevitable that we would find ourselves with a few issues regardless.

Having said that, we believe that regardless of these problems, being able to get through them and still succeed to making your objectives is what ultimately produces the quality of a project.
Fortunately for this project the issues faced were not substantial in size and to a certain extent would not hinder us from continuing to make progress.

The only real problem of any significance during my project, was mainly to do with datasets.

Procrastination is indeed a widely known habit, we all know this. However, the topic of procrastination is surprisingly not often considered experimental wise. There are papers that describe experiments, backed-up with their respective data, but this remains a small amount.

With this in mind, finding datasets for procrastination or ones that are linked was more difficult than anticipated.

The issue with the topic of procrastination is that is it a very difficult thing to quantify, measure and evaluate. Because of this, there is indeed a very limited amount of data available from open sources to be able to use.

We carried out a literature review to identify what recurring features linked or resembled procrastination the greatest – to seek and explore those with data for procrastination.

2.3 – Methodology

The bulk of the research body can be detailed into a few main sections.

The first consists of in-depth research into the topic of procrastination. As with all projects, this seems a sensible beginning step. For our project, it was the start of understanding the ins and outs of the topic and better educating ourselves in order to truly be able to help in aiding to a solution.

Producing a literature review we scoured the web, looking at many search engines and sources for papers and works, both from past and present, all related to the topic of procrastination, as well as other associated features. The literature review has helped us in being able to analyse the relevant existing research and sets a foundation to be able to build upon.
Discovering the purpose of any project is already difficult as it is, this has helped to acquire a sense of direction to the development of the project and bring clarity to the goal for what we are trying to achieve.

For our literature review we believed we had looked into all the most important areas to summarise the topic, covering a wide breadth. In the end, we examined: 'How & Why, We Procrastinate', 'Who procrastinates?', 'Are There Potential Positives To Procrastinating?' and 'What Have People Found With Technology & Machine Learning So Far?'.

We introduced four different relevant datasets, and the next step of the project was to conduct an approach of exploratory data analysis (EDA) through the datasets we acquired. Exploratory Data Analysis, or EDA, is the crucial measure at the start of using any dataset where perform initial investigations.

For each survey dataset we selected an appropriate supervised learning (modelling) method seeing, how well certain features and variables were able to predict target ones and preserve the data simultaneously.
We identified (and tested the power of) the key factors in explaining each respectively.

Finally using just those key features, we will see how well they can, by themselves, do the job, comparing that to the performance of the original – resulting in findings of which features bear the greatest significances.

3. Literature Review

3.1 – Introduction

In this chapter, we will discuss various research conducted on our topic about Procrastination Avoidance. Using the information and knowledge gathered from exploring a variety of papers, we investigated both the past and current theories and experiments. In doing this we have gained a greater understanding of the topic, allowing us to identify the major issues at hand. Being able to review the array of approaches of what has already been discovered we can compare those to more current theories and papers, seeing how they may agree or perhaps differ to one another.

To be more precise, we aim to be delving into:

- the how's and why's of procrastination to present a core understanding of the topic as well as the theorised situations, reasonings and causations.
- whether procrastinating can be seen as a good or bad phenomenon, taking a step back and seeing both sides of the argument so we can give a fair view of the situation.
- what people's concerns are, as the overall goal is to help avoid the effect procrastination has on people, delving into the main variables that are affecting people internally, such stress on mental health; and finally,
- what other people have done, as well as, with technology, taking a look into the existing work exploring what methods and results they have come up with.

3.2 – How & Why, We Procrastinate

With Procrastination being such a hot topic, there are many theories and explanations that researchers have proposed. Some do tend to agree and present similar findings in their works, whereas some tend to be somewhat contradictory.

Past researchers, such as Burka & Yuen, would describe the effects on people with procrastination to be more closely related to that of a personality trait [5] ([7]). A few suggested that the reasons accounting for this were character faults, that people who suffered from procrastination were lazy or perhaps showed lack of control, and as a result experienced feelings of low self-esteem [5] ([7]).
Zarick & Stonebraker have found that in recent studies, where they have made use of rational choice models, the effects of procrastination can indeed be predictable, even if on the surface of things, the decisions may seem

irrational [5]. They say that "Within a rational choice framework, procrastination is not an irrational personality disorder; it is logical", and though procrastination itself seems "potentially inefficient behaviour driven by a reasoned comparison of perceived costs and benefits" [5].
In theory then, we should be able to predict the people and the situations that are most likely to fall to procrastination, evaluating relative cost/benefits they encounter.

Zarick & Stonebraker review past research of procrastination causations and based on that theorise and hypothesise about three main issues: who procrastinates, when procrastination will occur, how we might limit procrastination. In their study, they use data which was aggregated from the use of questionnaires, distributed to undergraduate students and faculty members at a university [5].

On the other hand, [2] has a particularly interesting approach to the situation, and their paper suggests a different argument, titled: "Why You Procrastinate (It Has Nothing To Do With Self-Control)". As the title suggests [2] claims that those who suffer from the effects of procrastination are not as much in control of it as they may believe to be. They propose that the 'self-awareness' we experience during procrastinating is a key aspect, that we are not only very knowingly aware that we are avoiding the tasks or goals, but also that we know it will probably be burdened with negative feelings. Nevertheless, we continue to do it effortlessly. Lieberman has found that Dr Sirois says, "People engage in this irrational cycle of chronic procrastination because of an inability to manage negative moods around a task" [2].

And from this, Lieberman starts to question "Do we procrastinate because of bad moods?" [2]. In short they say, yes. The effects of procrastination are not always a mystery, there are reasons behind why just about everyone could be affected by it. One plausible reason is that procrastinating is just coping mechanism we suddenly put into effect to challenge the complex emotions and negative induced moods from particular tasks and goals we want to reach [2].
This is supported by a study by Pychyl & Sirois, who say that the developments of procrastination can be expressed by, "the primacy of short-term mood repair over long-term pursuit … a primacy of present self over the needs of the future self", further backing the theory that Lieberman believes [8].

As mentioned before, past researchers have generalized the effects of procrastination, however, as we go deeper to the reasonings of 'why?', the nature of it can be circumstantial within particular situations, leading to our unwillingness to attend or perform these tasks. Lieberman roughly separates this into two main reasons; with short-term procrastination related tasks, it could be that the task itself just presents as naturally unpleasant, for example, cleaning/tidying up your bedroom or perhaps finally started revising for a test that is coming up [2]. But, for greater tasks with perhaps higher risks, Lieberman says it may stem from much deeper reasonings, feelings related closely to the task itself, such as stress, anxiety, self-doubt, low self-esteem and so on [2].

Similarly, Zarick & Stonebraker discuss some similar interpretations for the causes of procrastination. As in line with Lieberman, Zarick & Stonebraker support the theory about an unwillingness towards certain tasks [5]. They say that for theorists of rational choice, the usual speculations are for the initial costs, which daunt us from beginning tasks (e.g., a student during term time who may have a large pile of work who must now use large upfront costs just to begin hacking away at it, they are likely to drift elsewhere to alleviate themselves) [5]. Zarick & Stonebraker found a study by Akerlof, who uses salience; the things that we perceive to be the most noticeable or important, to explain this mysterious reasoning [5] ([9]). This same concept can be used to explain procrastination (likewise to Lieberman in relation to the idea of emotion playing a role) as the events and emotions of day to day are immediate, pressing … whereas future ones are vague and less vivid [5]. This implies that we as people, in our minds, frequently disproportionate the weight of salient events with those less alarming to our present selves. "The salience of 'today's' pressures drives a wedge between current and future costs and creates a systematic behaviour that damages our long-run best interests" [5].

Xia, Sun & Zhang have also found some research linking to this. They discovered that some studies suggest that despite the awareness of the consequences, fully knowing that deadline is approaching, the gratification of immediate emotional improvements outweighs the important task at hand and so causing delays [10]. Just as Zarick & Stonebraker imply about ones future self, Xia, Sun & Zhang add to this saying, people attain this false belief that their future selves will be "better and more emotionally equipped" [10].

To further support this, Zarick & Stonebraker also found that many other researchers that likewise establish the aversion of tasks as another causable variable. They say that we delay in acting with tasks and we seek for anything to do except the task itself, in which for the better we should be doing. Generally, we are less prone to acting where we do not want to [5].

Uncertainty is another cause, which has been identified in relation to those affect by procrastination. When we are uncertain of things, we tend have to be upfront with the task, planning well in advance, accompanied with the fear of the possible outcomes [5]. For example, in with students and researchers alike, planning ahead of writing a paper and conducting a project, what resources may be required and how much allocated time is needed, … etc. All these are examples that can reduce the self-confidence.
Researchers such as Onwuegbuzie & Wolters have also expanded on the belief that the fear of failure, hinders us, when we fear to act and make mistakes [5].
With the possibly densely burdened decisions that we make during life, we would ideally be able to resolve any uncertainty from the situation but making the most optimal choices comes with a certain pressure, one that potentially costs us in making any choices at all [5]. O'Donoghue & Rabin add to this, since it is common to have the tendency to procrastinate for more important goals, rather than less important ones [5] ([11]).

Results after procrastinating would however still result to the same outcome since these feelings will not have gone anywhere once we return to the task at hand [5]. If not, perhaps worse, because now all these negative associations are aggregated on, and reactions such as stress and anxiety will very likely have grown in relation to time [5].
This self-inflicting mental stress and further negative feelings we tend to cause to ourselves does not particularly make sense then.

Lieberman (2019) suggests that this short-lived and momentary alleviation is actually what traps us into the cycle [2]. The idea of procrastinating is to turn away from the problem at hand, to anything that outweighs our desire to work on a task, anything that appeals to us more.

Another perspective of why this is happening, is to gratify our minds, and reward ourselves by doing a task more pleasant. Lieberman (2019) links this towards basic behaviourism [2].
As before mentioned, a behavioural learning theory is the "idea that all behaviours are all acquired through conditioning", where our responses to certain situations can be developed or influenced [2]. Conditioning results in both positive and negative reinforcements. And if we keep putting off tasks in our head, we are subconsciously positively reinforcing our minds that this 'reward' is a good thing, and "we know that from basic behaviourism that when we're rewarded for something, we tend to do it again" [2].
This theory does seem to further convey precisely why the effects of procrastination leaves those affected in a chronic cycle, and unfortunately not just a one-off demeanour.

The longer those who are affected by procrastination, and the more they practice it (even if it is unwillingly), the worse the outcomes can be [2]. With continual procrastination smaller initial problems such as the cost to productivity can slowly exacerbate into more detrimental issues that affect our health, potentially mentally and physically [2]. These can include chronic stress, general psychological distress, symptoms of depression, anxiety, chronic illness and more [2].

Taking a step back however, we have talked about procrastination being rewarding for ourselves, in an attempt to momentarily make ourselves feel better. Yet, outcomes we have talked about post-procrastinating have seemed to produce negative feelings in both the short and long term, outweighing any of the 'good' by a substantial amount.
It is ironic that we tend to procrastinate to get away from and to avoid negativity, but consequently it is what leads us directly there [2].

Research such as Dr Hershfield's proposes that we tend to distinguish our future-selves as strangers. Their work has looked into differentiating between the present and future selves. They have said "If people tend to consider the future self as a stranger … that future self feels on an emotional level like another person" [2] ([12]).
Lieberman has built upon this, theorising that during the effects of procrastination, on a neural level, perhaps parts of our brain think we are putting these tasks off for another person and that it is their problem to deal with, that person being our future selves [2].

Lieberman believes to really break down to the core root of procrastinating in general, it is not a productivity defect, but it starts with gaining a control of our emotions [2]. Procrastination needs to be dealt with internally, and therefore one cannot be dependent on anything but ourselves [2].

So, who procrastinates?

Past researchers such as Ellis & Knaus showed statistics that up to 95% of the students that they overviewed at least procrastinated occasionally [5] ([13]). Zarick & Stonebraker wanted to test their hypotheses against this and found some similar answers, using collected data from a group of students and members of faculty from a university of mid-size. With the results of their questionnaires, answered by all subjects involved, they measured 3 different outcomes. They asked if: procrastination ever lowered the quality of a paper/project (low quality); caused students to turn in assignments late (turn in late); and ever resulted in lower exam scores (lower scores). Subjects would have 5 possible responses to this, being: always, usually, sometimes, rarely, and never. They found that that effects of procrastination can indeed be very spread, as they expected to be so, however, among students, they did find a pattern that generally all types of students, the results of procrastination remained strikingly consistent [5].

Many studies have discovered that there is an association between high levels of procrastination with poor performances in academic studies, such as Rothublum et al., Senecal et al. and Tuckman [5] ([14], [15], [16]). Another study had a supporting view of this. Orpen explained that some students who score lower GPAs may see their studies and classes as significantly less important, and thus view the costs of procrastinating as less serious. These students tend to be more susceptible in falling to task aversion and that fear factor of failing when facing tasks [5] ([17]).
Unsurprisingly, Zarick & Stonebraker found that their results were aligned with the results of previous studies. In their experiment, the students with lower GPAs showed that they were significantly more prone to report lower quality work, hand assignments in late and/or receive lower scores; all due to the result of procrastination effects on them [5].
Having said this, Zarick & Stonebraker believe that it may be that the level of GPA for student is not related to the tendency to procrastinate, as those individuals with a high GPA were just as likely to procrastinate, reporting that procrastination would also cause issues in the non-academic areas of their lives [5].
Zarick & Stonebraker ponder if perhaps the reason that students with high GPAs are succeeding in the academic aspects, not letting the effects of procrastination affect their level of performance, is not because they would procrastinate any less than those with lower GPAs, but because they are able to succeed in spite of it [5].

Schraw et al. found something interesting: they saw that those students who had planned to procrastinate willingly retain a more positive view of procrastination and its effects. These students believed that with the added pressure that comes after procrastinating, they were able to complete the tasks and works more effectively and with creative twists, that would not have arose had they not left acting at the last minute [5] ([18]). They would claim procrastinating as their most effective approach to studying, praising that it had a positive influence on the quality work they produced. [5] ([8]).
Zarick & Stonebraker saw that their student responses did admit that the effects of procrastination did negatively impact their personal academic performances, having said this, that does not prove procrastination needs to be fixed as such. They said that those students who openly invite procrastination must suppose that the benefits short-term, far outweigh the future costs [5].

So … do we say procrastinating is a good or bad thing?

3.3 – Are There Potential Positives To Procrastinating?

As we have been discussing through the how's and the why's of procrastination there have certainly been many theories and discussions, most of which have deemed the effects of procrastination a negative phenomenon. Whilst the amount of work and research seems to strong suggest this to be the correct perceived connotation of procrastination, as we got towards the end of the previous section, we briefly glanced that Schraw et al. potentially discovered a new side that students would view to the effects of procrastination, that being one in a more positive perspective [5] ([18]).

In 2005, Chu & Choi conducted a study: "Rethinking Procrastination: Positive Effects Of "Active" Procrastination Behaviour on Attitudes and Performance" [6]. Chu & Choi saw that albeit literature in bother practical and academic sense had already paired all these negative connotations to the effects of procrastination, they found that some researchers had also found some supporting information for the benefits of procrastination, most specifically inducing short-term benefits [6]. Encouraging research that Chu & Choi found was by Tice & Baumeister, and they reported those affected by procrastination generally had less experiences with stress and overall better physical health early in the timeline for deadlines [6] ([4]). Together they claim that in a way, using procrastination can be perceived as a means to regulate one's personal negative emotions [6] ([19]).

Similarly, another study found by Chu & Choi was by Knaus, 2000; who further supported the notion that, not all aspects of procrastination, such as delays, lead to outcomes of a negative nature. An example they gave of this was: intentionally implemented the use of delays to give additional time that can be spent for planning and gathering additional information that could prove vital towards the end goals [6] ([20]).

Many people believe that under the effects of procrastination, they can still complete the work on time and in fact, it is due to the last-minute starts that they are able to accomplish their works more efficiently, additionally developing more creative ideas within pressured situations [6].

From this Chu & Choi establish a theory: "This line of thought on procrastination suggests that there might be more than one kind of procrastinator and that in some cases procrastination behaviour might lead to positive outcomes" [6].
This interesting approach expands as Chu & Choi identify within the study two types of procrastinators and further explore comparisons between the two potential attributes they may differ in. In the study they review over each type's characteristics such as use and perception of time, stress-coping strategies, personal outcomes etc. [6].

With this, they aim to try focus on the prospect that not all those affected by the effects of procrastination lead to negative connotations. Chu & Choi arrive to differentiating two types of procrastinators: Passive VS Active [6].
'Passive' procrastinators have been described by Chu & Choi as the type we know in a conventional sense, this type of procrastinator does not consciously expect to procrastinate, but due to their impotence to take effective action and make decisions, they will frequently delay tasks [6].
'Active' procrastinators on the other hand, are shown are more proficient, capable upon acting on their choices punctually. This type of procrastinator consciously chooses to procrastinate, knowingly, they suspend their actions to aim their attention more at other important tasks [6].
From this Chu & Choi conclude that on cognitive, affective, and behavioural levels, passive procrastinators must differ from active ones [6].

Chu & Choi found some further research that lines with they have hypothesised affectively. Those who are passive procrastinators tend to experience an unfavourable pressure, direct greater pessimism towards tasks, and this is especially apparent in terms of their self-efficacy to produce satisfactory results [6] ([21]).
Active procrastinators do not so much worry about this issue. In contrast to passive procrastinators, they tend to gravitate towards working under pressure; when encountering tasks at the last minute, they are surged with motivation and challenged positively and this feeling shields them from negative suffering that induces passive procrastinators undergoing similar situations, feelings such as, guilt and depression [6].
With this Chu & Choi present that behavioural patterns are a result of different interactions or responses both cognitively and affectively. In terms their two types of procrastinators, passive procrastination is most likely to lead to fail of completion of tasks with their sufferers prone to giving up; whereas active procrastinators are able to find a level of persistence leading them to succeed in completing tasks on time, even at the last minute [6].

In Chu & Choi's findings they present the possibility that are active procrastinators do indeed share many similar characteristics and attributes with non-procrastinators, both being significantly contrasting to traits of a passive procrastinator. Both non-procrastinators and active procrastinators alike generally manage more effective use of time and attain higher levels of self-belief. Their results show that the patterns demonstrate that non-procrastinators along with active procrastinators are expected to encounter greater positive outcomes. This supports Chu & Choi's hypothesis, that there is a chance that not all procrastination is bad, and that indeed there can be some potential positive practical outcomes from the effects of procrastination [6].

3.4 – Solutions with Technology, Applications & Machine Learning So Far

Lukas & Berkling conducted an analysis with the goal to reduce the tendency for procrastination for those involved, using a smartphone-based treatment program [22].
With investigation, they found that apps that are based around mental health and computer-based therapy carry many benefits since: (a) they are relevant in today's world and generally continually available to most people [22] ([23]), (b) have little maintenance expenses [22] ([24]), (c) are already owned by the majority of the population and is therefore efficient and effective to disperse [22] ([25]), (d) interactive using user inputs [22] ([26]) and (e) designed for ease-of-use for efficacy [22] ([27]).
Due to these positives, we are starting to see that smartphone-based treatments and interventions are becoming progressively favoured for supporting people with mental health issues [22].

Lukas & Berkling found that for leaning principles, modelling unconditioned responses by the user, this in the course of time, progressively improves by introducing and maintaining high user motivation – they use this idea with operant conditioning [22].

The app Lukas & Berkling created is MT-PRO and its main goal is to reduce procrastination by addressing systematically the users' approach/avoid motivations towards behaviours based on the effects of procrastination [22]. The app prompts the users to either actively void with dysfunctional stimuli or actively approach functional material [22]. MT-PRO is an app that trains attitude-contrary behaviour and encourages the development in relevant attitudes – users are concentrated to wiping away pictures showing engagement in typical alternate activities and negative statements, as well as more carefully wipe pictures of areas of study and positive statements related to the effects of procrastination - fostering avoidance and approach and reinforcing potential effects in with training using immediate feedbacks from operant conditioning [22].
They hypothesised that the app would be able to reduce the effects of procrastination within an academic field, and the effects from the study are expected to include a stable follow-up, which is indeed what pilot study result analyses showed. They discovered that their app was not only successful is positively affecting general procrastination, but also, the effects were continuous at the 4-week follow-up assessment [22].
The app is further accomplished at "generating, rewarding and maintaining strong positive habits…" [22].

Xia, Sun & Zhang have also created something similar with similar interests to the study above. In their project they designed a mobile application, iProgress, and the main aim again is to focus and target on the problem of the effects of procrastination, by concatenating the "useful strategies" [10]. The app they designed is purposed to track habits and goals, using machine learning, with additional unique features that composed to reduce the tendency that its users are affected by procrastination, whilst simultaneously increasing productivity [10].
Xia, Sun & Zhang say that whist simple routines that break down tasks into smaller segments have shown to be a favoured technique, such as a timer… without the procrastinator addressing external demands and requirements, it can be troublesome for them to commit, without the distractions and temptations, to completing said tasks, rendering the technique unreliable [10].

Features included within this app include: a customisable schedule, which is used as a managerial aspect, setting, and keeping track of tasks, in both daily and long-time goals; encouraging the users to consciously identify the advantages of not delaying particular goals, the app monitors the user's emotions and reward seeking behaviour, and prompting them to sway away from brief distractions; inspired from video games, a ranking-based system, purposed with encouraging and motivating users to be competitive against other users, completing more goals such that they are less inclined to procrastinate [10].
Xia, Sun & Zhang hypothesis that these with efficient methods implements into their app, they will be able to assist people to effectively tackle the effects of procrastination, whilst encouraging and forming a beneficial lifestyle behaviour, progressing healthy habits and better work ethics [10].

With the help of evaluations and feedback from users experience with the app, Xia, Sun & Zhang concluded that their app had an overall beneficial effect on those people with are affected and struggle with the effects of procrastination [10]. The majority of the users expressed that they had positive experiences with the app in terms of decreasing their impulses and tendency to procrastinate [10].

4. Related Work and Further Background

In this section, we give the general background for the methods and techniques used in this project, further discussing intricate details and definitions of the models, algorithms, and mathematical metrics.

4.1 – General Background to the Research

Machine Learning is vastly growing towards the technical field in the today's world. It is becoming increasingly popular for uses to addressing problems and improving performance.
Mitchell et al. (1990) have said that "the goal of machine learning research is to produce a domain-independent enabling technology for a broad range of computer applications" [28].
The idea of machine learning is that we are using computer algorithms and teaching them to improve automatically, through 'learnt' experience [29].
Machine Learning has been designed to mimic the intelligence of humans, to learn like we do, it has branched off from Artificial Intelligence, a sub-field.

So how does it work? Machine Learning is built up by model-based approaches, using the data and passing through to the algorithms. The data is accumulated to be arranged and employed forming what we define as the training data for the model [30]. Training data is what will be fed into the models, it is the initial data from which the algorithms learn pick up and train with.
From here, the machine learning algorithms will educate themselves to be able to compose predictions or perform aspired tasks for which it was created for. From a linear perspective, the more data we have to use, the more that the models can use to learn from, and thus the better the predictions and accuracies of outcomes [30].

Yufeng (2017) nicely outlines the main steps that are involved with machine learning:

I. Defining the problem: the first step to the process, establishing an issue, hoping we can split the information into features. Making note of the problems helps to identify a solution, being the model and what you want to gain from it.
II. Data Gathering: Convened from one or many sources, it is an important step as the foundation of the process – both quality and quantity are big roles here for concluding how accurately your model can predict, being that this will be the source from which our machines learn from.
III. Data Preparation: Putting all the data found together, as well as randomising it to allow the learning process to remain unaffected, we also want to clean the data. This can include anything from removing missing values, missing rows, missing columns, general noise and duplicate values, or potential anomalies. It would also be common to manipulate the data, such that you restructure the dataset more appropriately to your needs.
IV. Model of choice: Based off initial investigations and nature of the data you've gathered; it is important to choose a model that best fits and models the data well and appropriately – making sure to stick to the relevance of the goal.
V. Training the model: Using a subset of the data to filter through to the model, this is one of the most important steps, as this subset of data is what is used for the model to learn, and improve it ability's to predict from, based off the nature and its patterns. The more and more data the model has to train from, the better it will perform.
VI. Evaluating the model: Post training the model, it is now reading for testing, and to see how well it can perform. Using unseen labelled data, for the model to predict, you can gauge its performance and accuracies. The general split for training and testing is usually 80/20.

[31]

From here, machine learning algorithms branch out even more, most commonly into two main categories; these are supervised and unsupervised learning. As sub-fields of machine learning, they both still follow the procedures general from however, with distinctive twists.

For this project, we used supervised learning. Supervised learning has the ability to split the data into training and testing data. As beforementioned, the training data is what the models will use to teach themselves. However, this time, the model is also fed what we called testing data or labelled data, this is too additionally fed into the model. Testing/labelled data are known pre-existing observations and outcomes (also referred to as the

dependent variables) [32], which the model can also use to learn and predict from, and this allows for the models to better learn and produce higher accuracies [30].
Thus, the more training and testing data a model is fed to educate themselves with, the better over time they are able to predict with minimal error.

As some of our project aims are looking to establish the most significant characteristics of procrastination and related features, and that we planned to make this happen by testing models. To see how well certain features are able to predict particular dependent variables, whilst retaining a good level of accuracy; it was clear a supervised learning approach was the appropriate choice for its uses in for predictive analysis.

Following from supervised learning, we come to different types. The one type that best fitted to our project was regression.
Regression is a form of supervised learning for machine learning algorithms, and there are many models that follow this, using its predictions based on continuous outcomes [32]. The regression technique helps by using its predictions to help establish connections between labels and features of data points. In more mathematical terms, we are using a range of features, represented as independent variables 'x', which are implemented to be able to predict and estimate the dependent variable of 'y' [33].

4.2 – Models and rationale

In this section, we will be discussing the variety of different models, introducing the algorithms, that were tried and tested in this research - what they're about and how they work.

With there being many types of different models, Vellido et al. (2012) have defined it as such, "a model is meant to capture some of the intrinsic regularities or patterns that might be present in the data" [34].

4.2.1 – Linear Regression

The linear regression model is one of, if not, the most popular supervised learning algorithm currently in today's society. Being one of the oldest models, there is good reason for why remains so relevant.
It was developed to aid an understanding between both input and output numerical features, and the relationships that can be defined.
Linear Regression boasts in simplicity, making it an easy interpretation for everyone to understand.
Typically, 'x' is used to define the features of the dataset, individually noted as instances; 'y' is the target/dependent variable [33].
Along with the features, are the parameters, also known as weights – they are often denoted using beta or theta values.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$$
[33]

Some of the great advantages to linear regression models compared to others, is that due to its straightforward nature, other than its east to understand and interprets, it can be conveniently trained, easily and efficiently meaning that computational power requires and run times remain low [35].

In linear regression, the unknown parameters, the β's, are found by maximising the likelihood of the observations (minimising the norm of the errors, the ε's), over a set of given observations (assumed independent), and usually implemented numerical liner algebra.

4.2.2 – Logistic Regression

Like linear regression, logistic regression is another popular supervised learning model, where it differs however, is, instead of predicting upon continuous numerical values, it makes better use towards categorical

predictors. This is usually in the form of binary classifiers, calculating output dependent variables, it has been especially shown in helping to reduce potential bias [36].

$$y = \frac{1}{1 + e^{-\beta x}}$$

[37]

As with before, the logistic regression estimates for the unknown parameters, the b's, are found by maximising the likelihood of the observations (usually implemented numerically), in a given set, which are assumed independent. The 'x' marks the features, 'y' the dependent/target variable and beta here the parameter. This is usually interpreted as the probability, observations (or instances) are a member of a desired set, S: it thus discriminates between the desired set and its complement. Observed values for y are thus binary, indicating observed membership of S, and f(x) denotes the probability that y=1, that the observation is in the set S, given its corresponding features value, x.

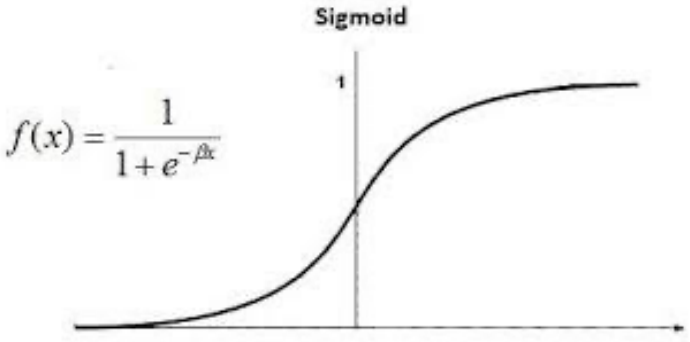


Fig. 1 – Sigmoid function [37]

Logistic Regression's can learn from the provided datasets linear relationships, represented in the form of Sigmoid functions as shown in figure 1.

Some of the great advantages to using a logistic regression model are its ease of implementation and uses. As opposed to linear regression, where the model depends on the variability of the outcomes being the same for all predictors, this does not align to match for binary level outcomes, thus logistic regression was formed to fill this gap [36]. It is also able the most important features, by determining coefficients, which is indeed very relevant and useful for the aims for our project [37].

4.2.3 – Decision Trees

A decision tree algorithm is again a supervised learning algorithm, one of the most powerful and able of classification and regression, more over being able to handle both categorical and numerical data. They are well-renowned for extracting patterns, classifying large and uncertain data [38].
The idea of the decision tree is to break down the dataset bit by bit, into smaller and smaller subsets. To do this it can be conveyed in the form of 'if/else' or 'true/false' statements in the form of the 'tree' structure – using these questions, starting with the root node, it will then branch off into smaller and smaller branches, producing both decision and leaf nodes [39].
Root nodes are the foundation and starting point for the decision tree and this corresponds to the greatest value predictor. Decision nodes are ones that typically hold two or more branches, a decisive turning point. Leaf nodes represents the decisions that have been made and lay at the bottom of the trees [39].

To determine the splits of a dataset, we can calculate (an seek to maximus) a Gini impurity and it is defined as:

$$Gini(D) = 1 - \sum_{i=1}^{k} {P_i}^2$$

[39], [40]

Here, ‘D’ denotes the dataset, of which contains k classes of samples, ‘Pi’ represents the probability of the samples belonging to a class, at index ‘i’ [39], [40].

Some advantages of the decision tree are that it can capture non-linear relationships, which comes useful for when linear regression is not an option. It is very fast and efficient, at recognising the most important features and relations. It has also commonly been shown to be able to handle both numerical and categorical data types. Decision tress can however be prone to overfitting; however, this can be settled by tuning parameters. To use them effectively also, it is best to have a larger dataset, otherwise high levels of variances can lead to more errors.

4.2.4 – Random Forest

Random forest is a powerful supervised learning algorithm, and, as the name somewhat suggest, it is the combination of bountiful number of decision trees.

The idea of a random forest is to reduce the variance levels. With individual decision tress bearing high variances each, it is more optimal for us to combine them all together. When we do this, the resulting variance is much lower, and this is because each tree will be wholly trained to the precise dataset. Therefore, the subsequent output of the model is based on multiple decision tress, and not just one [41].

Each tree with random forests uses a subset of covariates, selected, and deployed at random This allows the model to affect some correlation reduction by further element of randomness [42].
Although a more complicated model compared to the ones we have already discussed, there are many advantages when it comes to using a random forest model.
A technique that comes with random forests, aiming to reduce the correlation of different trees, is called bagging. Increasing the diversity of different trees and enforcing them to grow form different training subsets, random forests create random resampling of the original data, i.e., when generating the next subset, possible data from input samples will not be erased [41].
This way, more reliable stability is established, simultaneously creating a greater robust model, with increased accuracies, even when faced with slightly variating datasets [41].

4.3 – Metrics and mathematical models

In this section, we will briefly cover the outlines of the metrics.

4.3.1 – R2

R2 or the ‘coefficient of determination’ is a metric evolution, that is commonly used along with machine learning algorithms to score the performances models, based of levels of accuracies.
It is the proportion of the variations of a response variable, which is interpreted based on explanatory ones [43].
Most popularly, the purpose of the R2 score is to test the strength of a regression model, in particularly, the relationship [43].

$$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}$$

[43]

Here, y_i are the data set values, $\widehat{y_i}$ are the predicted values and $\overline{y_i}$ denotes the mean of the observed data.

Essentially what we see in formula, is the (total variance explained by model) / (total variance).
As in discussed, for this project, we have made use of supervised learning algorithms, more specifically into regression models. Thus, is seems appropriate to use R2 as a means to test the accuracy of each model built.
An R2 score can vary from 0-1.0; if we get a score for 1.0, this implies that the variables for features and target are perfectly correlated, meaning that the accuracy for the model is perfect, and can predict and preserve the data with no issues [44]. On the other hand, a low R2 value, means there is low levels for correlation, meaning the model does not fit the data well, as it cannot predict or preserve the data with any decent level of accuracy [44].

4.3.2 – MSE

Mean Squared Error is another metric, in which is used to measure the accuracy and performance of a model. Most typically it is used as a performance benchmark purposed towards models with continuous data, [45], which will be appropriate for the nature of our datasets and our project.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

[46]

Likewise, to R2, y_i are the data set observed values and $\hat{y}_i$ are their respective predicted values.
As we can see for the mathematical formula for mean squared error, it takes the average squared distance between estimated values and the true ones [46]. Logically, as the results are squared, we cannot have a negative value for mean squared error.
Unlike R2, for mean squared error, we in fact favour a lower score. The aim of mean squared error is to be minimised, for example, a larger mean squared error would point towards significant error loss, whereas 0 would imply a perfect model [44].

5. Datasets

In this section we will go through everything to do with the datasets.

5.1 – Choice of datasets and their retrieval

As the topic is clear by now, we seek to establish characteristics effecting procrastination.

This will drive our efforts to examine the literature once again. Almost all research to date points to the same prominent feature, stress. The following studies strongly support this:

Tice & Baumeister (1997) conducted a study about the effects that procrastination on quality of performance, to stress and potential illness. They found significant correlations between procrastination and depression, low self-esteem, anxiety and more. All have well renowned linked to stress. Moreover, they discovered, that under stress, performance of tasks is substantially worse, especially when we are faced with looming deadlines, reaching the greatest level of stress at that point leading to further serious consequences [4].

Supporting this, Zarick & Stonebraker (2009) found that their conclusions post the effects of procrastination would still result in the same outcomes [5].

Finally, Lieberman (2019) found almost the same outcome, saying that we compound our negative associations towards certain tasks, and even avoiding them, those feelings will have remained regardless, with accumulated stress, anxiety, and feelings of low self-esteem [2].

We can hypothesise the link between stress and procrastination.
Consequently, we decided interest to focus on stress datasets, and their features and characteristics.

Each dataset used had a particular section for licencing, stating open access and granting us approval to be able to use and explore that data we wished to work with.
Reading through the these, we were acknowledged with the data's general code of conduct and that we thoroughly credited the work respectively throughout its uses within this project.

In this project we identified four datasets to explore:

- 1x Procrastination (procrastination dataset students and staff from a UK school, 2022),
- 2x stress (study of stress and wellbeing in final year medical student, 2020; student stress survey, 2020)
- 1x motivation (academic motivation scale and intrinsic motivation questionnaire and data, 2020).

Each dataset used can be found:

- https://repository.lboro.ac.uk/articles/dataset/Procrastination_dataset_155_design_students_and_staff_from_a_UK_school_of_design_and_creative_arts/19160666
- https://zenodo.org/record/4264764#.Yr2dDS8w1qs
- https://figshare.com/articles/dataset/Student_Stress_Survey_Jan2020_OPENDATA_xlsx/11559528
- https://zenodo.org/record/3866179#.Yu78ni8w1qs

We wanted to ensure that each dataset was produced recent. To compare between now and the past literature, in order to propose fair contrasts and findings.

5.2 – Data preparation

As each dataset was found from a different respective source, and by different researchers each conducting their own unique experiments and surveys, it was natural that the datasets were uniquely different.

Because of this, the datasets had their own formats, layouts, and shapes: even coming down to the data types for each entry being distinct. We investigated the datasets individually, yet thoroughly.

So, what does data preparation include?.
There are few things that are imperative before using datasets for machine learning and making sure your datasets have cleaned and checked are a couple of those things.
These include are obtaining datasets, importing libraries, importing your datasets, locating missing values/rows, finding anomalies, label encoding particular data types, then later splitting data appropriately [47].

We started by importing the necessary libraries into python; NumPy, Pandas, Matplotlib and Seaborn.
NumPy is used to perform most of the mathematical computations that you may need. Pandas, potentially the most useful and well known, is used to all things data frames, essential for importing, managing, and manipulating your data. Finally, Matplotlib and Seaborn are useful tools, where we can take advantage of visualising our data, calling functions to particular graphs, plots and maps that are most appropriate. [47]

From here, we will import out datasets, and take a peek at what the current situations are. For most datasets the natural state and condition is usually undesirable, and so work needs be performed such that we manipulate them in the shape that is best for further steps.

Firstly, we call the data frame of the dataset, and check for missing values. These NULL values can be a nuisance to have to work with as they can interfere and cause inaccuracies to your results. One method we found particularly useful, was the use of heatmaps, passing through the dataset, and it would show me the density of the data, including where all the missing values are located.

Once the dataset is cleaned of missing NULL values, the next step is the check the columns of the data frame. Common to all datasets, the columns represent the features of the data. We can take this opportunity to remove any unneeded or unnecessary features that we do not need.

As we are using machine learning algorithms, it is important all the datatypes to be of a numerical format, whether that be an integer (int) or float. This will really depend on how the dataset is released, however usually, the natural state would be an 'object' type.

In order to be able to convert individual values from object or string, into numerical form, we can make use of an effective tool, this is Label Encoding.

Finally for the sake of ease-of-use, it can also be favourable to rename certain column names, so that they are generally shorter and easier to call upon. Using pandas this is a very straightforward manipulation.

6. Exploratory Data Analysis

In this section, we will discuss through the stages for Exploratory Data Analysis with each of our datasets. As a quick reminder we will talk about the how the initial examinations of the data gives us the opportunity to take remarks of the initial patterns and trends of the dataset and to identify any obvious and unexpected data issues. We will look at how each step was decided and support probable claims using visualisations.

6.1 – Procrastination Dataset

The first dataset we explored was a procrastination one, is it was created using a questionnaire that asked about the frequency and forms of procrastination of students and staff from a school in the UK.

Available at: https://doi.org/10.17028/rd.lboro.19160666.v1

During the process of cleaning, I decided to remove all the columns that that were not needed for the purpose of my project and this exploration. This was achievable by creating a list of those columns, and passing them through a drop() function.

By using Label Encoding, and defining the unique values of each row, we can simply convert these data types into a numerical format, ready for visualisation and machine learning models.

```
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   age                           153 non-null    object
 1   put off reading               152 non-null    float64
 2   put off evaluation            152 non-null    float64
 3   put off review                152 non-null    float64
 4   put off decision-making       153 non-null    int64
 5   put off writing report        153 non-null    int64
 6   put off completing documents  152 non-null    float64
 7   do not do anything            143 non-null    float64
 8   go back to bed                149 non-null    float64
 9   watch TV/films                149 non-null    float64
 10  eat/drink                     151 non-null    float64
 11  talk with friends             147 non-null    float64
 12  socialise                     150 non-null    float64
 13  walk/exercise                 151 non-null    float64
 14  tidy room                     153 non-null    int64
 15  do other less important tasks 151 non-null    float64
 16  talk about what I should do   150 non-null    float64
 17  plan what I should do         152 non-null    float64
 18  distracted by entertainment   142 non-null    float64
 19  distracted by new projects    140 non-null    float64
 20  importance of task            130 non-null    float64
 21  impending deadline            124 non-null    float64
 22  too many concurrent tasks     137 non-null    float64
 23  dislike task                  141 non-null    float64
 24  no interest in task           138 non-null    float64
 25  bewildered to purpose of task 132 non-null    float64
 26  task difficulty               143 non-null    float64
 27  low self-efficacy             135 non-null    float64
```

Here, we have used the info() function which allows use to quickly glance at some useful information about the current dataset.
We can see that each datatype is in a numerical format successfully and thus is in a good state for further steps.

Looking closely towards the columns, we start to notice the nature of each question or feature.
For columns 1-6, the questionnaire would ask participants about how often they would procrastinate, more specifically about how often they would put off tasks.
For columns 7-17, about in what way participants would put off tasks and how often.
For columns 18-27, features that influence procrastination behaviour, and how participants would react to these.

With three distinct types of questions, we decided to would be a good idea to later group these together and explore them individually.

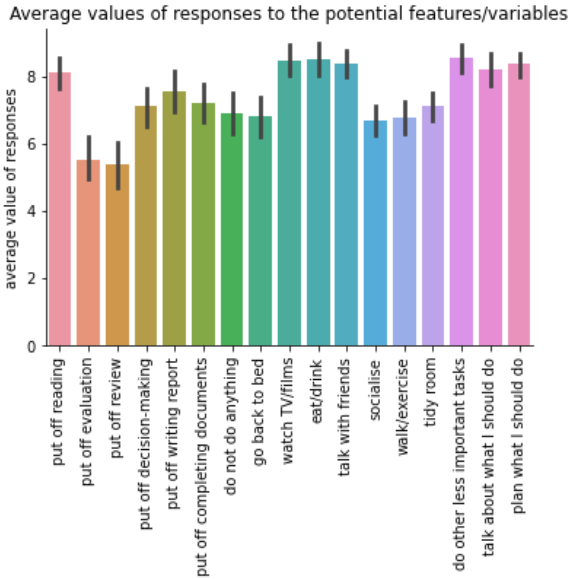


Fig. 2

The first thing we decided to see was taking a look at the first 17 columns, these are the columns generally ask the participants about features that cause procrastination, or the idea to put off tasks.

These are the columns of features/variables which are potential causes to procrastination
so it is interesting to see how participants responded to these in general.

To do this we have taken the sample mean (and estimated error on the sample mean) of all participants' responses to the feature and plotted together using a bar-plot.

Responses to each feature was based off the frequency the participant felt that they would take action towards each question.

Possible responses were: Not Applicable to my role == 0, Never == 1, Once a year == 2, More than once a year == 3, Once a month == 4, More than once a month == 5, Once a fortnight == 6, More than once a fortnight == 7, Once a week == 8, More than twice a week == 9, Once a day == 10, More than once a day == 11.

As we desired each response and entry to be of numerical value, we used label encoding to assign each response a number, similar to that of a 'Likertt scale', and this is shown above.
The higher the number response to a feature, the higher frequency that participant would feel towards a certain task or statement.

From bar-plot we have a clear visualisation of the average responses to features, we can see which variables that participants responded most closely and agreed to.
The feature and characteristics that initially emerge the most are: 'how often do you put off reading', 'I watch TV or films', 'I eat or drink', 'I talk with friends', 'I do other less important tasks', 'I talk about what I should do' and 'I plan what I should do'.

As these features are the highest scoring for average responses, we can initially see them as potentials for the most significant characteristics and distractions for the effects of procrastination. For these features in particular, the average scores point towards 8, which equates to responding to the feature/distraction at least 'once a week', if not more.

We planned to group the three distinct question types and use them to help in exploration.
So how can we group these? As above particular columns correspond to certain question types, so we can group particular columns and label them with the type of question we want.

- Question_type_1, will represent the columns 1-6, the questionnaire would ask participants about how often they would procrastinate, more specifically about how often they would put off tasks.
- Question_type_2, will represent the columns 7-17, about in what way participants would put off tasks and how often.
- Question_type_3, will represent columns 18-27, features that influence procrastination behaviour, and how participants would react to these.

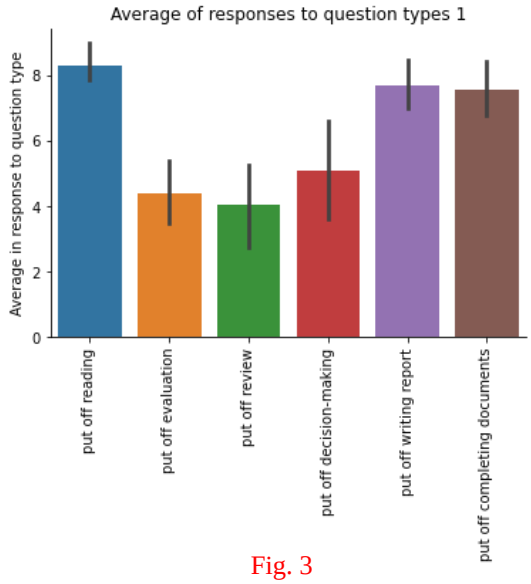


Fig. 3

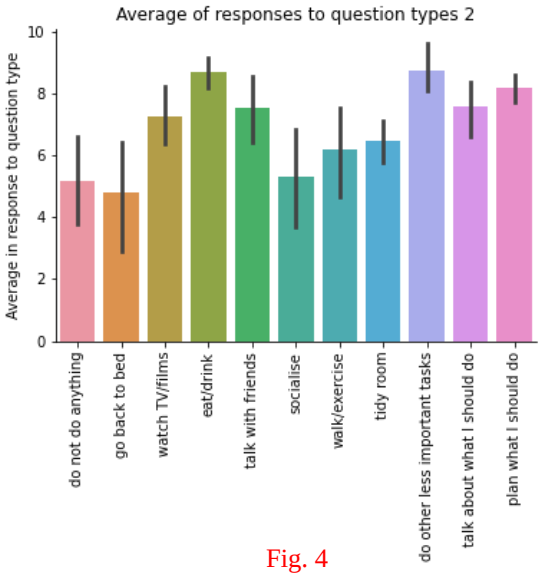


Fig. 4

Using the bar-plot visualisations once again, we can once again plot the average values of each response from the participants, however, now grouped by their question types. The advantage of doing this is that we are now able see more clearly the distinctions between the characteristic and distraction.
A great thing about grouping like this that we can make use of is being able to now combine different types of groups and make further deductions, as you will now see in this next part.

Another approach we thought would be interesting was that of age groups.
Each participant answered the questionnaire, inputting themselves in the appropriate age range. These groups were: ‘18-19yrs old’, ‘20-24yrs old’, ‘25-35yrs old’, ‘36-45yrs old’, ‘46-55yrs old’, ‘56-65yrs old’.

Using these age groups, we can group the participants respectively and explore how the different sub-groups differentiate with one another with regards to procrastination and the features. How each one may react differently, any patterns or correlations or similarities.
Merging both question types and age groups, we are able to pick out exactly how different ages would respond to different question types, more specifically how age groups would respond to features for how often they would procrastinate and in what ways they would do so.

As responses for question_type_1 and question_type_2 had the same list of possible responses for the participants to choose from, it made sense for us to plot some graphs to compare them to one another and perceive how differently participants grouped by ages, respond to features about how often they would procrastinate against in what ways they would.

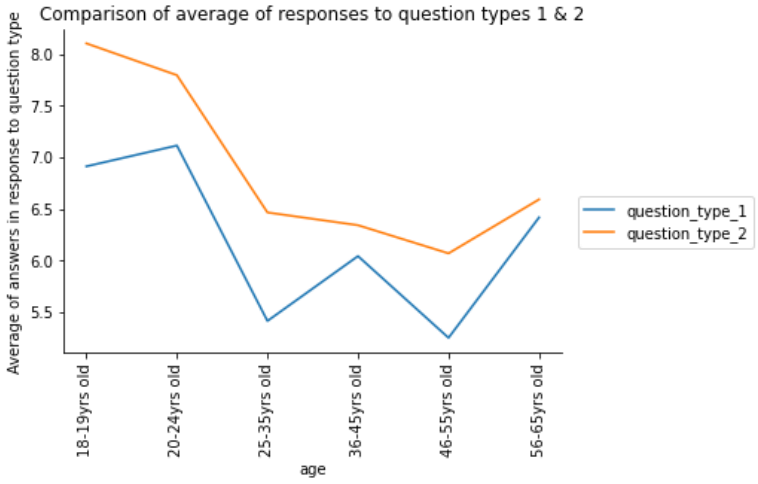


Fig. 5

For this next plot, we have made use of line plots. Grouped by age ranges, the lines flow to present the average of responses to the different question types.
From a quick glance at the plot, we can immediately see that for all ages, the participants responded more frequently to questions of type 2, being in what way would participants put off a task and how often.
One explanation for this could be explained, because features from question_type_2 are the types of ways participants would put off task. A few examples for context, are: 'I eat or drink', 'I talk to friends' or 'I do other less important tasks' – the characteristics are prevalent day-to-day in our lives, and extremely common, so perhaps when participants see these responses they can more greatly relate to them, resulting in a higher frequency.
Whereas with question_type_1, these are the features for which task participants put off and how often – these types of features would only be experienced at time of work or school, meaning the actual interactions with these types of procrastinating is a constrained amount of time during a day.

Another point that comes shouts immediately obvious to us are the significantly higher responses for the age ranges of '18-19' and '20-24' – the 'younger' age groups. In general, we can depict that the younger ages seem to be substantially more prone to the effects of procrastination.

This is evident by looking at both line plots, for each respective question type. As we know, this means that the general mean of responses used more severe levels of frequency to answering questions/statements.

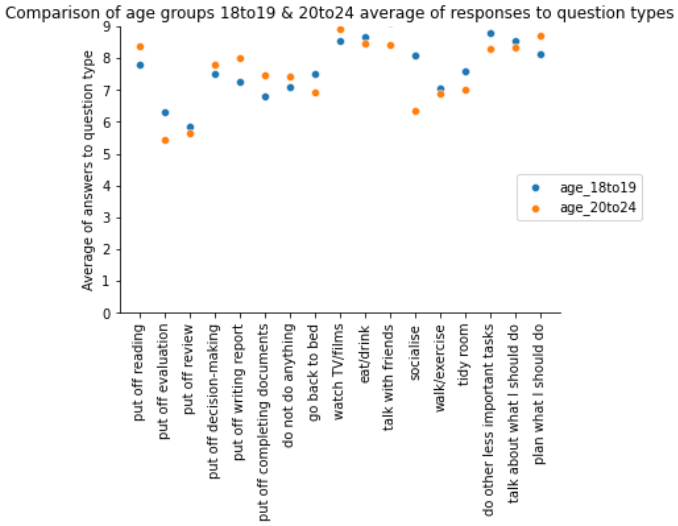


Fig. 6

Separating the two age groups for '18-19' and '20-24', and quickly plotting a scatterplot, as we look more specifically at the behaviour of responses.

Just by glancing at the plot, we can see that two groups would respond to the features in an extremely similar manner, suggesting that there isn't a significant difference between these two age groups – but that younger ages as a whole react alike.

This finding is interesting as we can use it to compare to what other studies have found in the past. Zarick & Stonebraker (2009) are one study, who have found that, students seem especially adept to procrastination [5]. As our dataset is based on students and staff, those who are age ranged from '18-24' are almost certainly students and so our results are just as Zarick & Stonebraker (2009) found.

These ages are most prone to procrastination - Zarick & Stonebreaker (2009) have said, students with lower GPA, tend to find most costs when it comes to studying, they believe it to be less beneficial and so report greater issues with procrastination [5].

Aljoscha & Berkling (2017) support this, finding similar outcomes, they conducted a study, showing that academic procrastination has been consistently connected to deeper emotions of severity, linking to anxiety, depression, and poor task performance [22].

As the last question type was slightly different due to the possible responses we deemed it better to evaluate and visualise it by itself.

The third type of question was about: influences on procrastination behaviour, and what ways the participant would respond to these.

There responses possible were only either one of 2 choices, these were: 'It makes me want to do other things (not this task)', 'It makes me give up and walk away'.

In order to be able to perform analysis with these responses, I had converted them into numerical data, thus we assigned: 'It makes me want to do other things (not this task)' == 0 and 'It makes me give up and walk away' == 1
(It is important to note that particularly for these columns and question type, we did not assign each number with the intent that one response is greater than the other, or that one is leads to more severity as a response.)

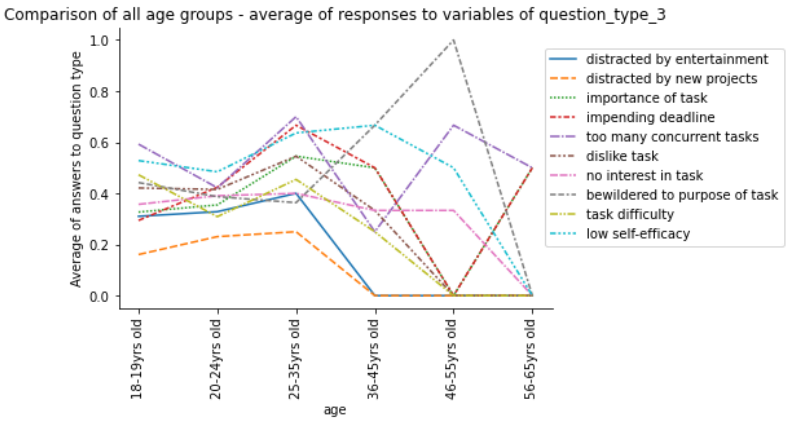


Fig. 7

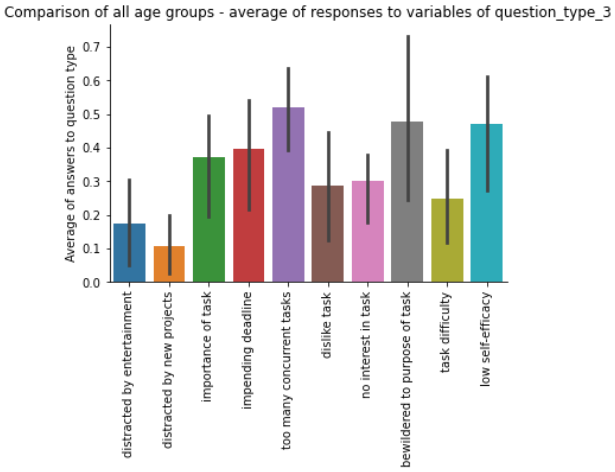


Fig. 8

From the line-plot, we can see that some of variables averaged in the medium, for age groups 18-45, for example, low self-efficacy, dislike task, no interest in task, importance of task, bewildered by task, too many concurrent tasks
We can gather from this that, on average, participants affected by these variables did not particularly fall to a specific response, but however the responses to these variables were split relatively evenly, meaning it was common as a response for participants to either 'want to do other things(not the task)' or 'give up and walk away'
It is apparent that the younger half for age groups ranging from '18-35', would fall into this group.

With regard to the older half age groups, ranging from '36-65', this is where the graph now shows some interesting contrasts
Some variables have shown to have averaged along 0, this means for these particular causable variables, the age groups ranging from 36-65 when answering 'distracted by new projects' & 'distracted by entertainment', would respond with 'It makes me want to do other things(not this task) – initially we can hypothesise that these

variables are most significant for this age group in how they are affected by the effects procrastination.

On the other hand, one particular variable averaged to the other side, and produced a mean score of 1 - this variable was 'bewildered by purpose of task' , this was particularly significant for the age group of 46-55, who responded to the question with 'It makes me give up and walk away'.
If they did not understand why they were doing the task, it would lead to them giving up and finding an escape from the task, a severe effect of procrastination.

6.2 – Stress dataset (1)

For the first stress dataset, we look at a study of the stress and wellbeing of final year medial students.

Available at: https://zenodo.org/record/4264764#.YxfF4ezMJqu

As before, this dataset too underwent data pre-processing and cleaning.

```
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   gender                                  159 non-null    int64
 1   upset due to unexpected circumstance    159 non-null    int64
 2   unable to control important aspects of life  159 non-null    int64
 3   nervous feelings                        159 non-null    int64
 4   confident about abilities to handle issues  159 non-null    int64
 5   felt things were going your way         159 non-null    int64
 6   could not cope with all the current tasks  159 non-null    int64
 7   able to control irritations             159 non-null    int64
 8   felt on top of things                   159 non-null    int64
 9   angered by incidents out of your control  159 non-null    int64
 10  felt difficulties accumulating          159 non-null    int64
 11  stress level                            155 non-null    float
64
```

As a result, we obtain a data frame of the following form.
Opposed to the procrastination dataset, we have a clear target variable, this being the ‘stress level’ feature, which we will do the majority of our focusing towards.

Responses are valued with how *often* the participant felt or thought in a particular way - with 0 being the weakest and 4 being the strongest.
The responses were ranked:
0 == Never, 1 == Almost Never, 2 == Sometimes, 3 == Fairly Often, 4 == Very Often

For the gender of the participants to convert this to numerical data, they have assigned each one a number. this genders were assigned as:
1 == Male, 2 == Female

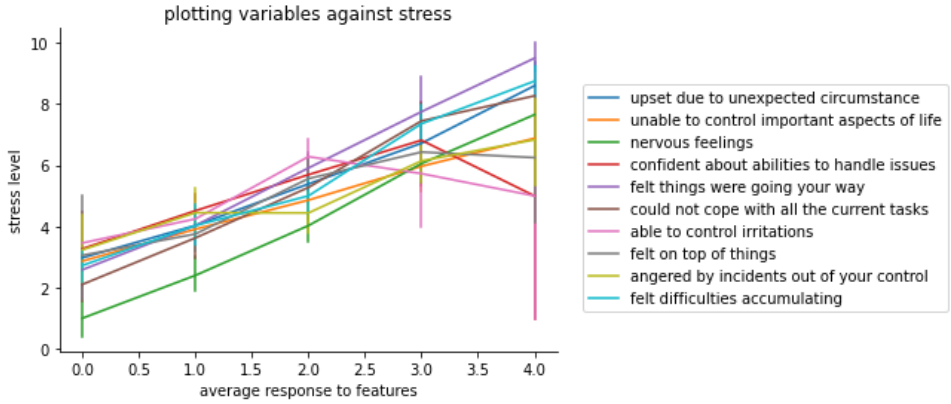


Fig. 9

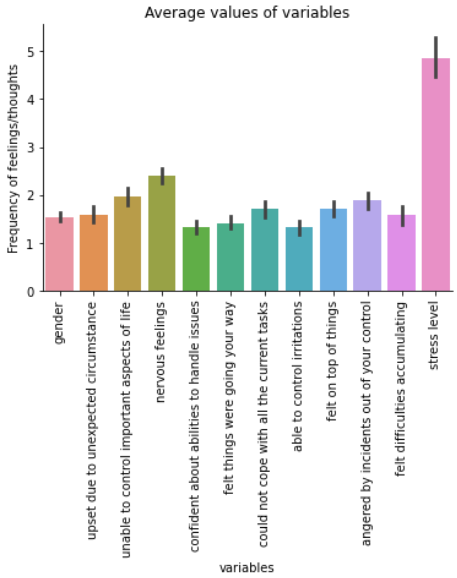
To begin, as done previously, we would take the mean values of responses from each participant and plot them accordingly.

Using a line-plot, we have plotted each of the other features against the ‘stress level’ feature, in hopes starting our initial investigation and seeing how each individual characteristic compares.

It is very apparent that all the features and variables show a strong positive correlation with stress.

There are a few variables that don't completely positively correlate in relation to stress, and this makes sense, as these variables are more 'positive driven features’, about how well participants are dealing with the variable, they higher they respond, the better they feel about dealing with the variable. These were: 'confident about abilities to handle issues', 'felt on top of things' and 'able to control irritations' - it makes sense that as these 'positive driven features ' of variables increase, that stress would then decrease and vice versa.

Fig. 10



Now, plotting the averages of each feature within a bar-plot, we can more easily see and compare the average responses from all participants. Noticeably, we can see that the variables 'unable to control important aspects in life' ‘could not cope with all the current tasks’ and 'nervous feelings' average scores were the highest amongst all the variables (excluding the target variable of ‘stress level’), average scores 2-3 which meant that they would experience these feelings 'sometimes' to 'fairly often'.

It’s also interesting to see that the lowest scoring features were ‘confident about abilities to handle issues’ & ‘felt on top of things’, showing participants rarely managed to feel control over tasks and feelings, suggesting this makes them more susceptible to stress.
This links studies about how control especially shows importance towards encouraging procrastination. Burka & Yuen and Lieberman are a few, who concluded procrastinators held such lack of control – and the first step to an aid is managing control over our emotions [5], ([7]), [2].

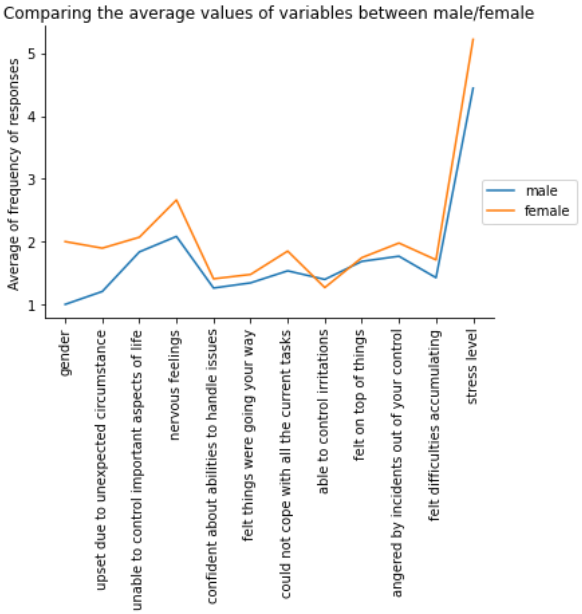
If we look back the line-plot above, we can see that these two features were strongly positively correlated along with stress levels increasing, meaning there is good insight they tend to cause stress levels to rise, making impressions of eventually leading to higher levels of procrastination.

Looking at the variables that scored lower, were : 'confident about abilities to handle issues', 'felt on top of things' and 'able to control irritations' - these being our more 'positive driven features ' variables. Since these mostly scored nearing to 1, it corresponded these participants felt that they 'almost never' felt these feelings.
As these were students at university level, this finding remains unsurprisingly with these feelings being rare, it shows participants were more commonly faced with stress, and were little experiencing control over tasks and feelings. This is good support for our finding with the earlier dataset of procrastination and aligns with the past works from researchers such as Zarick and Stonebraker (2009).

Similarly, to the procrastination dataset, we use the same sort of approach as with the different age groups.

However, for this stress dataset, we look at comparing the different gender types – and exploring how they may vary or correlate with stress. We aim to establish whether gender plays any significant role in affecting the stress level of a person and consequently their tendencies to procrastinate.

Fig. 11



In this line-plot, we have plotted the average responses to the features, grouping them by their respective gender. We can see that females tend to answer ever so slightly greater of frequency to the features, more significantly when it came to 'stress level', 'upset due to unexpected circumstance', 'could not cope with all the current tasks' and 'nervous feelings'.
We could depict out of the genders, females show to be slightly more vulnerable and responsive to the features, thus it makes sense that their stress levels also appear greater.
Having said this however, the shapes of the plots are extremely similar and there is not a lot to tell apart from each line-plot, suggesting that gender as a variable does not have a great significant impact to stress and further procrastination
Zarick & Stonebrakter (2009), also sought out to test this theory. Through their study, they had found some people would say that males tend to procrastinate greater than females. However, as with our findings, Zarick & Stonebraker's research found limited evidence to support this [5]. In fact, they too showed that in the general case there is little significance between male and females that play a role in relation to procrastination [5].

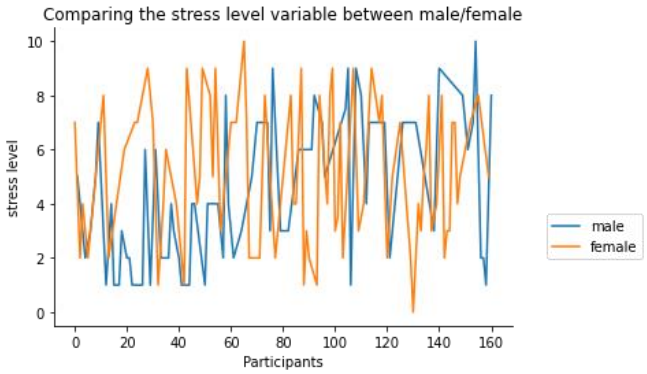


Fig. 12

Looking at one final comparison for the two gender types, we look at the general responses of the participants towards the 'stress level' target variable.
Unfortunately, we do not get much specific detail from these plots, however we can see the responses from both genders are very widespread, possibly suggesting participants were feeling different ranges of stress and perhaps that features causing this may affect everyone slightly differently.
This however is not enough to suggest that 'stress level' is a significant factor that would differ between genders.

6.3 – Stress dataset (2)

For the second stress dataset, we look at a student stress survey study on university students in the UK.

Available at: https://doi.org/10.6084/m9.figshare.11559528.v1

#	Column	Non-Null Count	Dtype
0	stress	218 non-null	int64
1	low energy	218 non-null	int64
2	headaches	218 non-null	int64
3	digestion problems	218 non-null	int64
4	anxiety	218 non-null	int64
5	sleep problems	218 non-null	int64
6	increased heart rate	218 non-null	int64
7	irritability	218 non-null	int64
8	concentration problems	218 non-null	int64
9	sadness	218 non-null	int64
10	illness	218 non-null	int64
11	aches/pains	218 non-null	int64
12	loneliness	218 non-null	int64
13	coping mechanisms help?	218 non-null	int64
14	uni work overloaded	218 non-null	int64
15	competition between peers	218 non-null	int64
16	difficulties with tutor	218 non-null	int64
17	unpleasant working environment	218 non-null	int64
18	criticism about work	218 non-null	int64
19	lack of time for relaxation	218 non-null	int64
20	financial issues	218 non-null	int64
21	lack of confidence - academic performance	218 non-null	int64

However, with good perseverance, we were able to from this new data frame, adjusting the column names to better represent what each feature is showing. We can see that similarly there is a particular target feature for us to focus on, the ‘stress’ column.

Alike the previous datasets, responses to the features and characteristics were based off frequencies and how often participants would relate or act upon them.

- For the ‘stress’ column, the possible responses were: 'not at all'==0, 'to a small extent'==1, 'somewhat'==2, 'to a large extent'==3, 'completely'==4.
- For the ‘coping mechanisms help?’, the possible response was: 'no'==0, 'not sure'==1, 'yes'==2.
- Features for the rest of the columns had possible responses: 'never'==0, 'almost never'==1, 'sometimes'==2, 'fairly often'==3, 'very often'==4.

The aim as before is to transform the data in numerical features, but retain the severity level, as each response increases.

In this dataset participants where asked whether their personal coping mechanisms helped at all to managing their levels of stress or even relieve them from it.

With three distinct responses, ‘yes’, ‘not sure’ and ‘no’, we believed it would be interesting to group participants by these responses to seeing how much variation and difference there is between participants who believe their coping mechanisms aid in their relief to stress and to those who do not believe so. It would also be interesting to look at the ‘stress level’ for those who do believe that their coping mechanisms work, and whether this really plays a significant role, or if this is just something the respective participant ‘believes’.

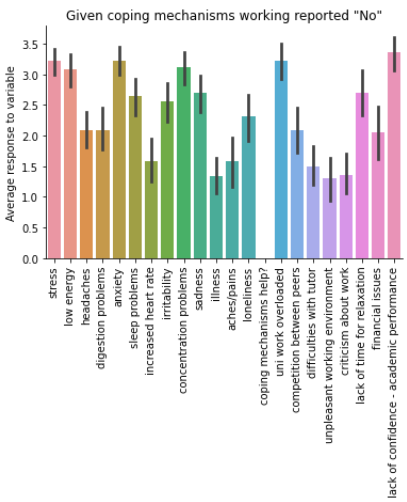


Fig. 13 (a)

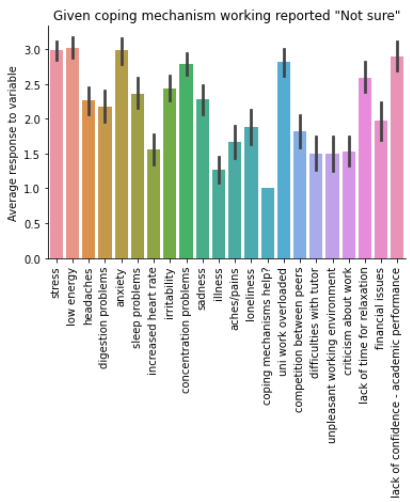


Fig. 13 (b)

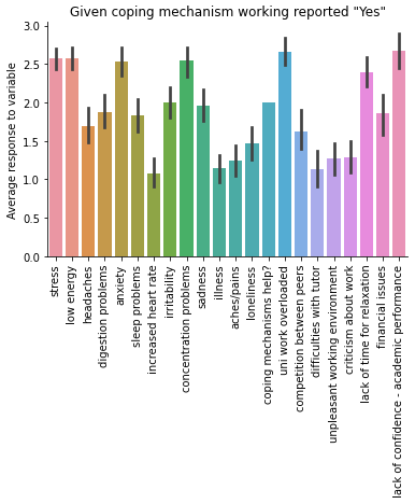


Fig. 13 (c)

Plotting a bar-plot for each grouped response, we can see the average response to features and characteristics for the respective participants.
As a general trend for all three groups, we can see that particular features shine out considerably more and consistently across than the others. These features are: ‘stress’, ‘low energy’, ‘anxiety’, ‘concentration problems’, ‘feeling overloaded with university work’, ‘lack of time for relaxation’, ‘lack of confidence with academic performance’.
As the highest scoring features, participants would score frequency with these from ‘fairly often’ to ‘very often’. As these responses are frequently the most related to by the participants and top rated features, we can initially take them to be the most significant variables affecting participants when it comes to stress.

If we take notice of the stress feature more specifically, unsurprisingly, participants who reported ‘no’ to whether their coping mechanisms worked or not, score the highest on average to not only stress, but also all the other most significant characteristics we just discussed, and this becomes immediate noticeable, when you look at the y-axis in particular for each plot.

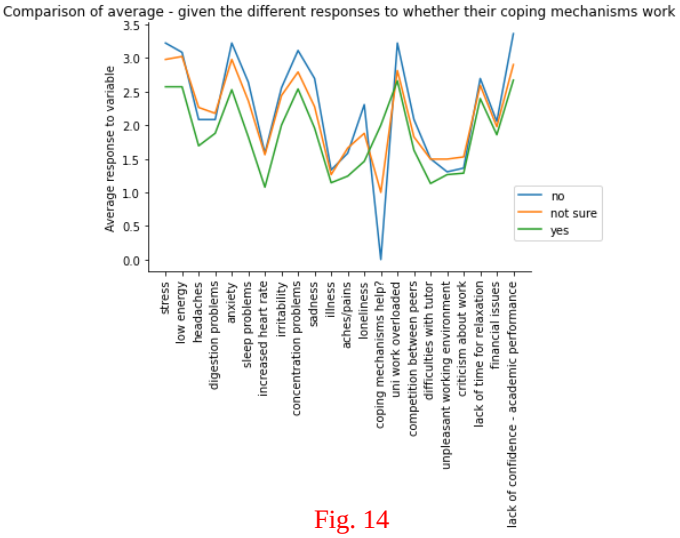


Fig. 14

In this plot, we aim to essentially combine the three plots as above in order to see what potential patterns or trends.

Just form a glance, it is immediately apparent that the shapes of each line-plot render almost identical. Moreover, looking at the distinction between the gaps of each plot, being somewhat small – we can make assumptions that regardless of whether a participants’ coping mechanisms work for them to relieving the effects of stress, there is little evidence to suggest that personal coping mechanisms play a significant role to helping with stress.

We can hypothesise, that perhaps, when it comes to stress and procrastination consequently, personal coping mechanisms are not the most optimal route in aiding in avoidance. Of course, there will be cases where they do work for a particular individual, however in terms of helping the majority, perhaps this requires external help, much more than us alone.

From the bar-plots we established a few main features, along with stress: low energy’, ‘anxiety’, ‘concentration problems’, ‘feeling overloaded with university work’, ‘lack of time for relaxation’, ‘lack of confidence with academic performance’.
Using this knowledge, we wanted to similarly to before, plot them all against the ‘stress’ feature to see how these initially ‘most causable’ variables play against stress.

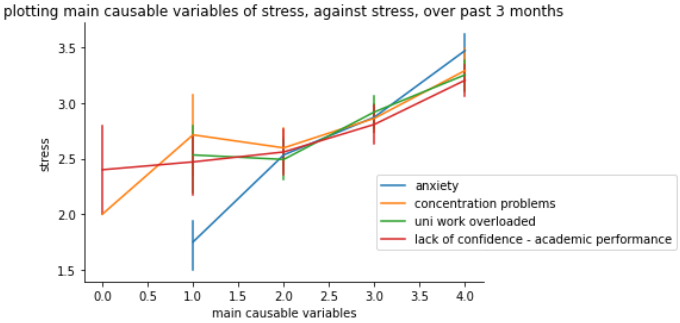


Fig. 15

Plotting each variable against ‘stress’, and using a line-plot as the visualiser, these are the results we find.
Immediately we notice that for all the other features, they express a strong correlation with the ‘stress’ feature.

We can learn from this that there are signs for strong implications when it comes to ‘stress’ and these other features.

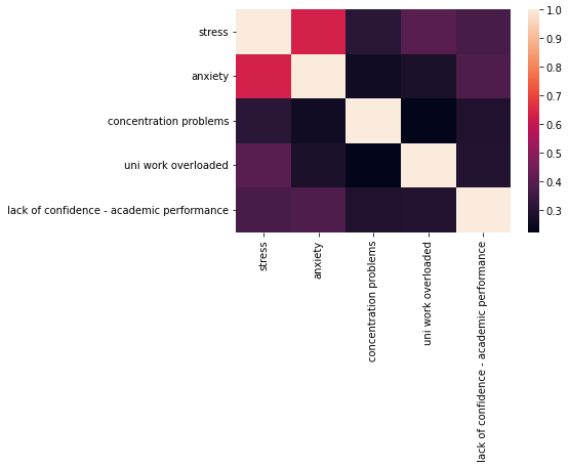


Fig. 16

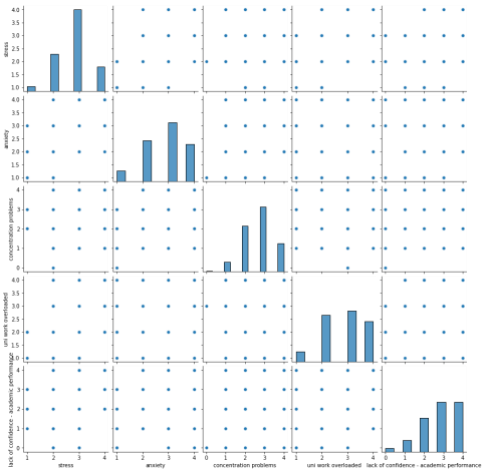


Fig. 17

Here are two more visualisations, where we have plotted the most seemingly influential features against the 'stress' variable.
A heatmap and a pair-plot can be particularly clear visuals.
Heatmaps are plots that make use of turning numerical data into colours and visualise matrix values – naturally they are very simple to understand.
Pair plots are allowing us to make use of multiple pairwise comparisons – we can grasp an understanding of the distributions between of one feature against many and the relationships hidden within.

6.4 – Motivation dataset

This dataset used academic motivations scale and intrinsic motivation questionnaire & data.
The goal from the research of this dataset, was to find what motivates university students to study, and whether their motivational levels had gone down over time, particularly over the events of COVID-19.

Available at: www.zenodo.org/record/3866179#.YxiNqOzMJqt

In this dataset, there are indeed many columns and features, so we will now do our best to break them down, using label encoding as before, to change all the entries into numerical data.

The first couple of columns that are asked to the participants are the 'level of degree' and 'year of study'.

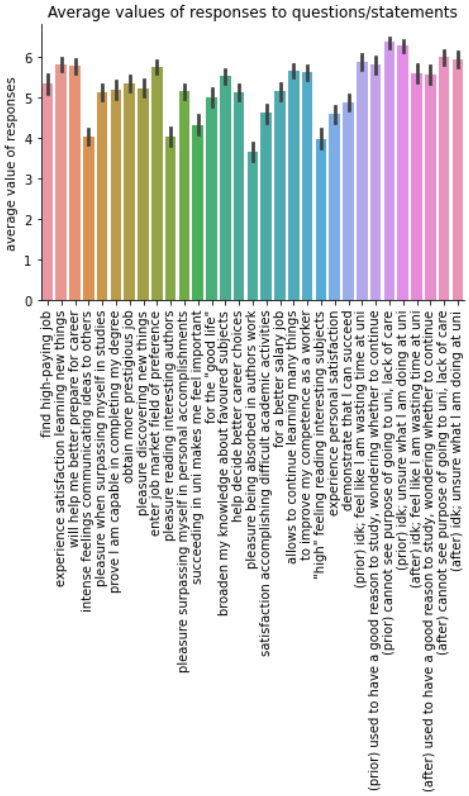
Using the survey provided by the researchers I know that for 'level of degree' these values correspond as such: 1 == Bachelors, 2 == Masters, 3 == PhD, 4 == Other.
Naturally for 'year of study', I would assign: 1 == 'first year', 2 == 'second year', 3 == 'third year', 4 == 'fourth year' and 5 == 'fifth year'.

As for the rest of the dataset, these are the questions and statements. As with the procrastination dataset, the questions and statements can be grouped together by different types.
The first type of group of questions/statements is based off the 'academic motivation' and aims to answer - "Why do you study for your university courses?". The authors have specified that 'The statement "I study..." is presumed before each of them.

The second and third type of questions are about amotivation (~lacking motivation/purpose) – these questions/statements are actually the same, however would ask the participants about motivations respectively prior/after to COVID-19 social distancing measures being put in place.

This final question asked is to sum up the types of questions before, asking about motivation levels since COVID-19 social distancing measures – this is a good sense for the gauge of motivational levels over a certain time.

Fig. 18



To begin we would plot the averages of responses from all participants to the individual features and aligned them in a bar-plot visualisation.

For questions/statements, participants responded using a Likert scale, in numerical form, which is typically used to measure either attitudes, knowledge or value changes.
Reponses are valued with how *often* the participant felt or thought in a particular way, with 1 being the weakest and 7 being the strongest.

The responses were ranked: 1 == Strongly disagree, 2 == Disagree, 3 == Somewhat disagree, 4 == Neither agree not disagree, 5 == Somewhat agree, 6 == Agree, 7 == Strongly Agree.

The higher the average score is, the stronger all participants on average, would agree with the statement given, vice versa.

Analysing the plot of the graph we can see that features that scored the highest amongst all the columns, implying that for each statement participants generally felt strongly towards each one:

- (question type 1) - 'experience satisfaction learning new things', 'will help me better prepare for my career', 'enter job market of preference', broaden my knowledge about favoured subjects', ‘for better job salary’ and ‘to improve my competence as a worker’
- (question type 2) - '(prior to covid-19 distancing measures) cannot see purpose of going to university, lack of care', '(prior to covid-19 distancing measures) idk; unsure what I am doing at university'
- (question type 3) - '(after to covid-19 distancing measures) cannot see purpose of going to university, lack of care', '(after to covid-19 distancing measures) idk; unsure what I am doing at university'

In contrast, the lowest scoring columns were: 'intense feeing’s communicating ideas to others', 'pleasure reading interesting authors', 'pleasure being absorbed by authors work', '"high" feeling reading interesting subjects'.

From question type 1, initial thoughts we can gather is that participants seem to be greatly motivated studying new knowledges and that studying is greatly based on how it will affect their future in terms of jobs - intuitively this makes sense, as participants will be greater motivated by the tasks and subjects they are most interested in, to study to work towards a good future with a job they prefer at a good salary rate.
As lowest scores were generally about reading tasks about works by authors, we can gather that this does not inspire much motivation among the participants, participants did not strongly agree with these statements (we can hypothesise that as a consequence, tasks relating to these features could tend to influence procrastination greater).

Now, by using the groups created before, for ‘level of degree’ and ‘year of study’, we can filter out results from the different question types and compare the motivations independently.

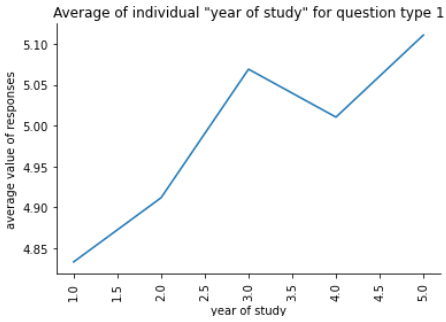


Fig. 19

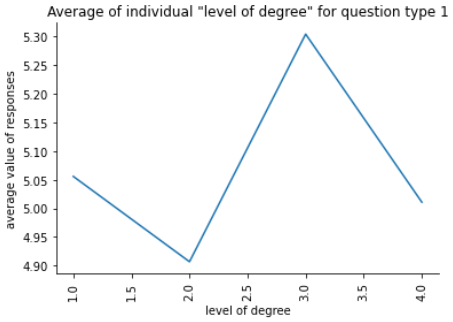


Fig. 20

On the left graph we have a line-plot for the average value of responses to the first question type, where the average of results are grouped by the participants 'year of study'.
Off quick intuition, we can see a general correlation in the data, it seems the higher the 'year of study' (i.e. generally the older the participants), the greater they would feel agreeing to the statements.
There is a clear significant difference in response from the general first year students to fifth year (who would likely be in their last year of education).
If we think about this logically, this does indeed make sense - the older we are the more we experience we have, especially at our times during university, where we learn, grow and develop mentally the most, the more we go through studying at university, the clearer we would be about why we choose to study, as well as the motivations behind it.

On the right side, we have plotted the average responses of each 'level of degree' individually towards question_type1.
As opposed to the plot for average responses of each 'year of study' individually towards question_type1 – the data shows less linear relationship. With a much more 'zig-zag' like figure, it is apparent that there is not a strong correlation between the level of degree with how the predicants reposed towards the statements by groups of 'level of degree'.
Interestingly, with level of degree == 3 == PhD scoring the highest, by quite a significant margin, we can gather that PhD students greater related to the statements, implying their signs for being the most motivated 'level of degree'.
We could hypothesise implications that PhD students are a lot less susceptible and prone to procrastination, suggesting to having much clearer views on their motivations to study.
In comparison with perhaps that of master's students, whose average score towards the statements were significantly lower - the less clear you are about your goals and studies, the lower your motivational levels will be, and this is likely to encourage high levels of procrastination.

With both the highest 'year of study' and 'level of degree' being the top groups with the greatest levels for motivations, we can initially conclude, that in fact the 'older' students are significantly less susceptible to procrastination and stress in general.
This links with the theory above that was discussed about age groups in the first stress dataset, where we had grouped that study by age ranges.
There we found that younger age groups were more likely to succumb to stress and anxiety, and as a consequence more probable to suffering the effects of procrastination. On the contrary, this would mean that older age groups would be the opposite – which is essentially what we have shown here with their levels of motivations being the greatest.

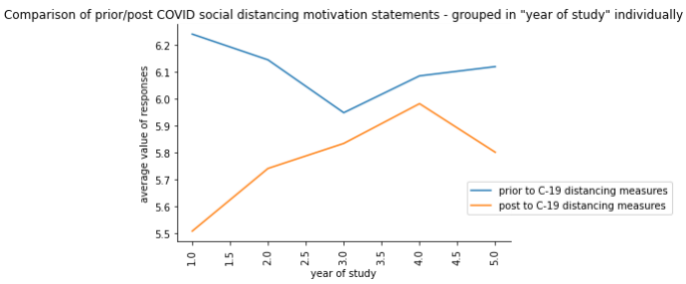


Fig. 21

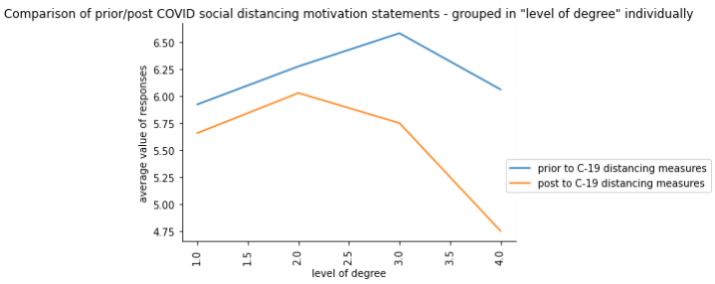


Fig. 22

The two plots above are comparisons for the questions/statements to do with motivational feelings prior/post to COVID-19 distancing measures put in place.
These statements were:
- 'Honestly, I don't know; I really feel that I am wasting my time in university'
- 'I once had a good reason for studying for university, I wonder whether I should continue'
- 'I can't see why I go to university and frankly, I couldn't care less'
- 'I don't know; I can't understand what I am doing in university'

We take the average of responses once again, about how strongly the participants associated to the statements. On the left plot, we group them by their 'year of study'. The stronger that participants scored, the stronger the participants were unsure about the motivations to study.
Hypothesising, prior to social distancing measure this means that participants are a lot more exposed to potential distractions, e.g. socialising with friends, this greater encourages people to divert and procrastinate from tasks, with less focus on why they choose to study, it is not surprising that motivational levels also drop as a result. Whereas, with social distancing measures in place, everyone would indeed have to avoid contact with people as much as possible for safety reasons - it could be that because of this, people were less likely and less prone to be distracted from their tasks - and because of this they were less susceptible to procrastination, and clearer with more time to understand their motivations behind why they study.

On the left plot, we group them by their 'level of degree'. This plot seems to follow the same trend as for 'year of study'. Once again, participants generally scored significantly higher and consequently felt a lot stronger with the statements for prior to COVID-19 social distancing measures were put in place.

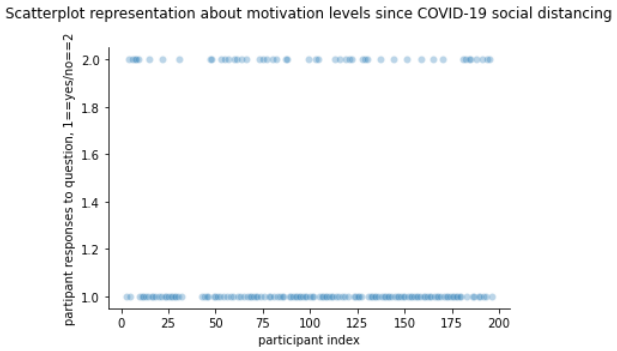


Fig. 23 (a)

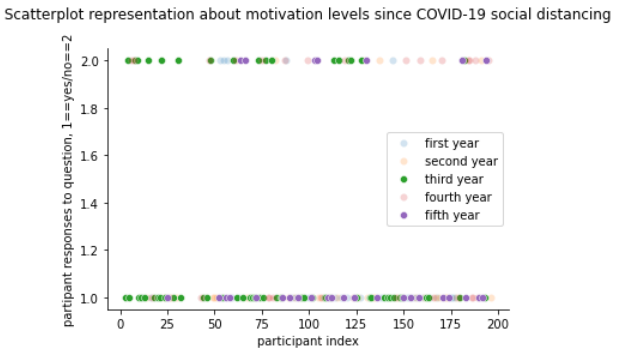


Fig. 23 (b)

Lastly, for the final question/statement of the dataset, which asked participants about where their motivation levels had gone down, post to the situation with COVID-19 social distancing measures being put in place.

Creating a couple scatter representation for all participants, they were asked to answer yes or no. For the sake of ensuring that we remained working with numerical data, we have assigned the answer for 'yes'=1, and 'no'=2. (For both plots, we have adjusted a parameter alpha, which helps us visualise the density through all the data points)

Quite evidently, most participants answer the question with 'yes', meaning the majority of students felt less motivated after the distancing measures were put in place.

This does indeed contrast slightly to what we found in the previous couple of plots.
Hypothesising a reasoning for this is because, to stayed motivated during times of social distancing can be very difficult - along with working hard for tasks, it is important to also take time off, to help keep the mind clear, whether that be socialising or exercising, however with social-distancing measures these were kept very limited to everyone

Clearly this has affected motivational levels - if participants feel that their motivational skills have gone down, it's very likely procrastination is a big factor playing into why this has happened.

Lastly, for the right-side plot we thought it would be interesting splitting the data into their respective year groups.

By adjusting the alpha levels of each year, altering the opacity and density of the points, it is interesting to see that many of the participants who felt that their motivational levels had gone down were those in 3rd and 5th year - at university level, these are usually the final years of study.

This is also a contrasting finding to what we had found early. Perhaps this also played in as a feature, as towards the end of the degree, it can be very difficult to stay motivated and refute procrastinating. It is apparent that motivational levels are much more complex than we initially believed, making us wonder whether the age of a participant really has any significance role for motivational levels.

7. Machine Learning (Supervised Learning Models)

In this section, we will go over the machine learning algorithms that we chose to use for each of the datasets including the rationale behind the choices. We will also discuss through the accuracies and metric we used to evaluate the performances of the models, as well as present what findings we were able to obtain for the most significant variables.

7.1 – Procrastination Dataset

Available at: https://doi.org/10.17028/rd.lboro.19160666.v1

Through exploratory data analysis and visualisations, we were able to make assumptions on what could be the most encouraging and influential features to procrastination.

These features were: ‘how often do you put off reading’, ‘I watch TV or films’, ‘I eat or drink’, ‘I talk with friends’, ‘I do other less important tasks’, ‘I talk about what I should do’ and ‘I plan what I should do’.

Using learning models, we can check whether these characteristics truly play the biggest roles, or whether we end up finding something different, perhaps other features that are surprisingly high in importance that we didn’t initially expect.

As briefly mentioned before, the choice of model that best fits to a dataset really depends on the nature and initial findings of the dataset we find through exploration, and what ‘kind’ of data is present. For most cases the initial decision is usually the best one, however we can very much conduct trial and testing, where sometimes we can find that other models may perform better.

For the procrastination dataset, there wasn’t a particular ‘procrastination feature’ – as the nature of procrastination is a complex culmination of many variables, such as stress, anxiety and many more.
With this being said, one particular feature did pique our interest, and this was the feature for measuring ‘low self-esteem/efficacy’. This was due to its strong relations and links to procrastination that was prevalent within the literature.
For example, Zarick & Stonebraker (2009), Berka & Yuen (1983), Lieberman (2019), Soloman & Ruthblum (1984) and many more, have all deeply expressed the significant correlations that effects of procrastination shares with low self-esteem/efficacy.

From exploratory data analysis, we can see that the feature column for ‘low self-efficacy’ and its entries only takes two values, implicating it’s of binary nature based on its categories; these are 0 or 1.
And because of this, it makes reasonable sense to use a ‘logistic regression model’ as the learning model of choice for this dataset.

In this procrastination dataset, we want to first build a model, using all the features and seeing how well they can be used to predict and preserve out target feature of ‘low self-efficacy’.

So, by assigning out features and labels: here we are assigning all the feature values apart from 'low self-efficacy' to X, and for the target feature 'y', only the column for 'low self-efficacy'.

using all features to predict outcome

dropping only irrevalant columns

```
In [3]: # features and labels
        X = df.drop(['low self-efficacy'], axis=1).values

        y = df['low self-efficacy'].values

        y = y.reshape(-1,1)

In [4]: # creating algorithm - logistic regression model
        logis_reg = LogisticRegression()

In [5]: # separating the data into training/testing
        # 0.2 = 20% of the data will be used for testing
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,

In [6]: # training the model
        logis_model = logis_reg.fit(X_train, y_train)
```

By importing libraries from scikit-learn, we can call the libraries for Logistic Regression.

Creating the algorithm, we make a new function to set up the model.

The next step for machine learning is to split the dataset intro training and testing. We do this by assigning X_train, X_test, Y_train, Y_test by calling a new function, imported by scikit-learn model sections, train_test_split. Here we have chosen a test_size of 0.2, meaning we have asked the function to split the dataset for a 20:80 ratio to testing and training respectively.

Using the fit() function, we are able to train out model, by passing through our new X_train and y_train variable splits.

We now have our model set up and ready for the procrastination dataset.

```
In [7]: logis_predictions = logis_model.predict(X_test)
        print("predictions: ", logis_predictions)

        predictions:  [1 1 0 1 0 0 1 1 1 0 1 0 1 1 1 1 0 1 1 0 0]
```

By using the predict() function along with the logistic regression model we created, since the model is now trained, these are the predicted results for which the algorithm creates for the target features.

accuracy of logis reg

```
In [8]: # we use 'score' to obtain the accuracy of the model
        logis_score = logis_model.score(X_test, y_test)

In [9]: print("logis_score (accuracy): ", logis_score)

        logis_score (accuracy):  0.7142857142857143

In [10]: # mean squared error value - score
         print("MSE: ", mean_squared_error(y_test, logis_predictions))

         MSE:  0.2857142857142857
```

Using an evaluation metric of score(), and passing through our model with the testing features splits, we are able to gauge the accuracy of the model, looking at how much roughly the model we made was able to predict and preserve the data. As shown, the model (using all the features) yields approximately at best ~0.71 or ~ 71%.

Another metric we used was by importing a library from sci-kit learn, the mean squared error (MSE). For this, we pass through the target test split, as well as the model predictions. The mean squared error metric corresponds to the error loss of the model, and we achieve a score (using all the features) for ~ 0.29.

From there, we now want to see which features seem to be the most significant. We can do this by plotting their coefficients.

```
In [11]: # coefficient
         print("coefficient: ", logis_reg.coef_)

coefficient:  [[ 5.23576043e-04 -2.55700676e-01 -1.56384731e-02  1.5602
7085e-01
  -1.26343030e-01  3.50341251e-01 -1.51400709e-01  9.07476244e-02
   2.75627064e-01  1.14832386e-01 -3.37594852e-01  1.50437526e-01
  -1.24907517e-01 -1.83483329e-01 -1.75931856e-01  2.49989031e-01
   9.18028386e-02 -1.64697721e-01  4.36376229e-01  4.23536282e-01
   2.66937316e-01 -4.85047018e-02  4.28944100e-01 -6.48973526e-01
   3.32768385e-01  1.54819822e+00]]
```

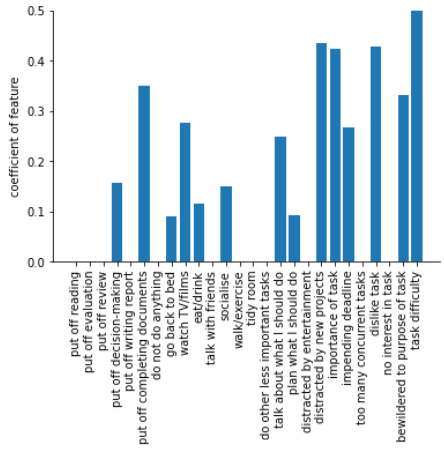


Fig. 24

From this, we can see which features in particular played the biggest roles in predicting and preserving the data. These features would be the ones with the most importance.
Our results point to features of:

- 'put off completing documents'
- 'distracted by new projects'
- 'importance of task'
- 'dislike task'
- ‘bewildered by purpose of task'
- 'task difficulty'

are the most significant features and that these play the greatest roles for preserving the data.

So quite how significant are these features? We can attempt to show this, by now using just those features alone, and seeing how well they can by themselves predict and preserve the data.
Being able to compare this with the original performance should give us the answers we need.

So similarly, following the same steps as before: this time however we filter the features for ‘X’ to just the significant features that we have just found.

```
# features and labels

X = df.drop(['put off reading', 'put off evaluation', 'put off review',
       'put off writing report',
       'put off completing documents', 'do not do anything', 'go back to
       'watch TV/films', 'eat/drink', 'talk with friends', 'socialise',
       'walk/exercise', 'tidy room', 'do other less important tasks',
       'talk about what I should do', 'plan what I should do',
       'distracted by entertainment',
       'impending deadline', 'too many concurrent tasks',
       'no interest in task',
       'low self-efficacy'], axis=1).values
y = df['low self-efficacy'].values

y = y.reshape(-1,1)
```

```
# we use 'score' to obtain the accuracy of the model
logis_score = logis_model.score(X_test, y_test)

print("logis_score (accuracy): ", logis_score)

logis_score (accuracy):  0.6666666666666666

# mean squared error value - score
print("MSE: ", mean_squared_error(y_test, logis_predictions))

MSE:  0.3333333333333333
```

This time, we are able to obtain an accuracy score (using just the significant features) of: ~0.67 or ~67%
And a mean square error score (using just the significant features) of: ~0.33

```
coefficient:  [[-0.00194732  0.26108361  0.72674647 -0.28675316  0.4653
4513  1.29736446]]
```

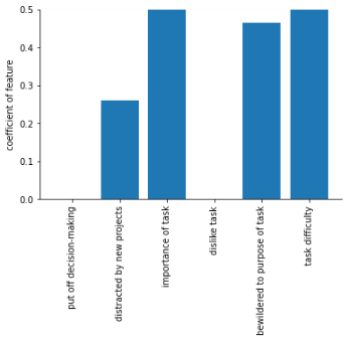


Fig. 25

Plotting the graph for the feature importance once again, we can see which features are truly the most significant when it comes to low self-efficacy and therefore procrastination in this dataset.

We can conclude these features to be:
- 'importance of task'
- 'bewildered by purpose of task'
- 'task difficulty'

7.2 – Stress Dataset (1)

Available at: https://zenodo.org/record/4264764#.YxfF4ezMJqu

For the first stress dataset, we look at a study of the stress and wellbeing of final year medial students.

In this dataset, there was a clear target variable of choice, that was the 'stress level' feature.

As discussed in the exploratory part of the report for this dataset, we talked about how there were strong linear relationships for the features along with the 'stress level' variable.

Because these relationships yielded strong positive correlations, to choose a Linear Regression model.

We import the necessary libraries needed to carry this task out, similarly to before. From sci-kit learn, we are importing the libraries for LinearRegression as well as train_test_split.

As the steps here are almost identical to that of the procrastination dataset, as well as the next couple of models, I will refrain from repeating myself over and over – and instead focus more on the accuracies and performances we find.

```
from sklearn.metrics import r2_score

# r^2 value - score
print("R^2: ", r2_score(y_test, predictions))
```

```
R^2:  0.7394457207072913
```

For this first stress dataset, by using an r2 score for the accuracy of the model when using all the features, we obtain a score for ~0.74 or 74%.
(r2 is a metric commonly used along with linear regression, it calculates the proportion of variance in the response variable that can be explain by the predictor variable)

```
from sklearn.metrics import mean_squared_error

# mean squared error value - score
print("MSE: ", mean_squared_error(y_test, predictions))
```

```
MSE:  2.1495050220942713
```

As for the mean squared error (using all the features), we obtain a score of ~2.14

```
coefficient:  [[ 0.33971036 -0.19545316  0.91381447 -0.04467548  0.6964
9197  0.60264251
   0.04871477  0.11175535 -0.32036941  0.32675502]]
```

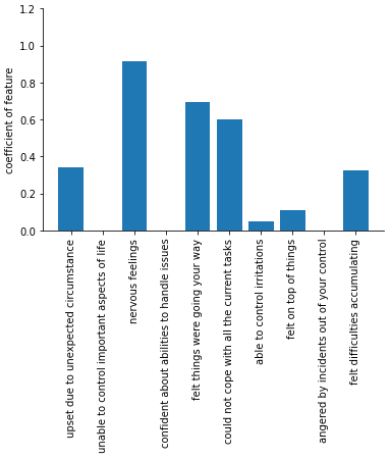


Fig. 26

Calculating the coefficients of the features, we can the plot a feature importance graph to be able to see which of the features are the most significant and play the biggest roles in predicting and preserving the data.

We achieve finds for these features:
- 'nervous feelings'
- 'could not cope with all the current tasks'
- 'upset due to unexpected circumstance'
- 'felt difficulties accumulating'.

So, to be able to truly see their significance in predicting the data, we isolate those features and model them, seeing how well just by themselves they are able to predict the data.
We can then compare their performances in the next section.

```
# r^2 value - score
print("R^2: ", r2_score(y_test, predictions))
```

```
R^2:  0.6795244882939926
```

```
# mean squared error value - score
print("MSE: ", mean_squared_error(y_test, predictions))
```

```
MSE:  2.6438396012541383
```

Using just the most significant features as found before, to predict the data, we obtain new scores for accuracy and mean squared error.

Accuracy (using just the significant features): ~0.68 or ~68%
Mean squared error (using just the significant features): ~ 2.64

```
coefficient:  [[0.2742186  0.85478911 0.65450569 0.3352973 ]]
```

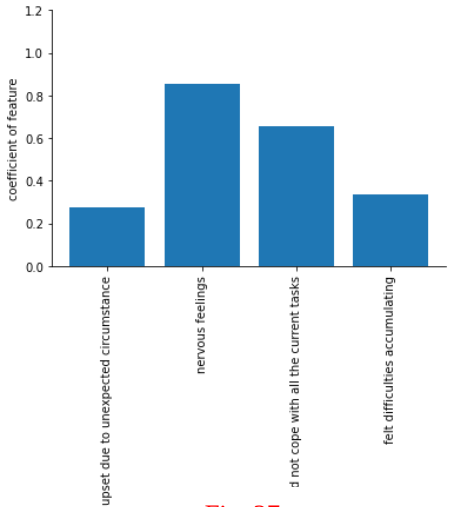


Fig. 27

Once again, by calculating the coefficients of the features and plotting.
This time are we used just the most significant features of data for the model, we obtain the true most significant features, finding:

-'nervous feelings'
-'could not cope with all the current tasks'

as the greatest influencing characteristics for this stress dataset.

7.3 – Stress Dataset (2)

For the second stress dataset, we look at a student stress survey study on university students in the UK.

Available at: https://doi.org/10.6084/m9.figshare.11559528.v1

As with the previous dataset, this stress dataset too had a very obvious choice for the target variable, the ‘stress’ feature.

Similarly, we had also found through exploration that the features would show strong positive correlations against the ‘stress’ feature.
Here we gathered that there were the strong implications for a firm linear relationship between the features and label.

Naturally again, Linear Regression was the obvious choice of model to fit for this dataset.

As previously shown, we set up the model appropriately and see how well the model can predict the target column for the ‘stress’ feature.

```
# r^2 value - score
print("r^2: ", r2_score(y_test, predictions))
```

```
r^2:  0.6282713123287191
```

```
# mean squared error value - score
print("MSE: ", mean_squared_error(y_test, predictions))
```

```
MSE:  0.23559457632885408
```

Calculating the accuracies and error loss metrics we obtain:

Accuracy (using all features): ~ 0.63 or ~63%
Mean squared error (using all features): ~ 0.24

```
coefficient:  [[ 0.0478941   0.06938793  0.03022271  0.24159546  0.0343
4032  0.08275686
  -0.02104632 -0.02138025  0.22071879 -0.09852527  0.01640881 -0.006879
76
  -0.02291103  0.10063411 -0.05316698 -0.01865526  0.02787424  0.014975
17
   0.07428174 -0.01114585  0.03166381]]
```

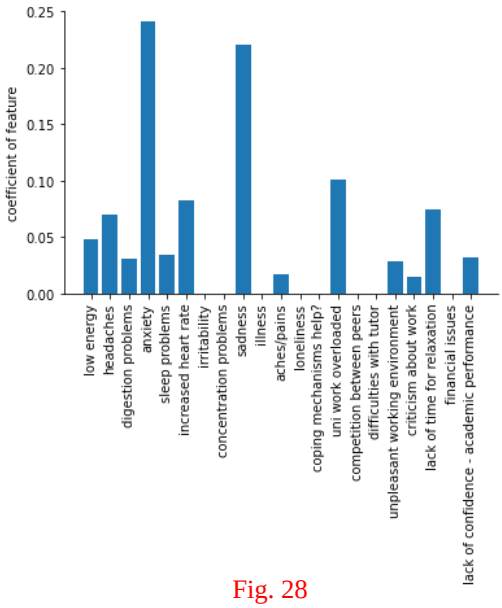


Fig. 28

Calculating the coefficients of the features, we plot a feature importance graph to be able to see which of the features are the most significant and play the biggest roles in predicting and preserving the data.

Our results point to features of:
- 'low energy'
- 'headaches'
- 'anxiety'
- 'sadness’
- 'uni work overloaded'
- 'lack of time for relaxation'.

It is quite clear the features for ‘anxiety’ and ‘sadness’ clearly reign greater for importance compared to the other features; however, it would still be interesting to see what happens in the next step when we isolate just these top features.

Once more, following the same steps as before: this time however we filter the features for ‘X’ to just the significant features that we have just found.

```
# r^2 value - score
print("r^2: ", r2_score(y_test, predictions))
```

```
r^2:  0.6187984707004818
```

```
# mean squared error value - score
print("MSE: ", mean_squared_error(y_test, predictions))
```

```
MSE:  0.24159828329055208
```

Calculating the accuracies and error loss metrics we obtain:

Accuracy (using just significant features): ~ 0.61 or ~61%
Mean squared error (using just significant features): ~ 0.24

```
coefficient:  [[0.04235284 0.08409015 0.27148504 0.20161189 0.09706202
0.08547099]]
```

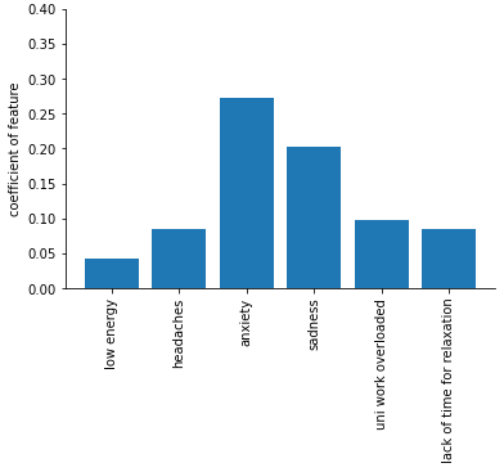


Fig. 29

Once again, by calculating the coefficients of the features and plotting.

This time are we used just the most significant features of data for the model, we obtain the true most significant features, unsurprisingly finding:

- ‘anxiety’
- ‘sadness’

as the greatest influencing characteristics for this stress dataset.
(It is worth to note the third most influential feature being ‘uni work loaded’, playing the next more significant role, which we will later discuss.)

7.4 – Motivation Dataset

Lastly we have the motivation dataset, which used academic motivations scale and intrinsic motivation questionnaire & data - finding what motivates university students to study, and whether their motivational levels had gone down over time, particularly over the events of COVID-19.

Available at: www.zenodo.org/record/3866179#.YxiNqOzMJqt

For this dataset, there was a column of which asked participants about the motivation levels, and whether it had gone down over a certain period of time.
We decided to use this column as the target column for our models to be able to predict.

As with the procrastination dataset, this target columns entries only takes two values, implicating it’s of binary nature based on its categories; these are yes or no.
And because of this, it makes reasonable sense to use a ‘logistic regression model’ as the machine learning algorithm of choice for this dataset.

As previously, we set up the model, with a logistic regression algorithm appropriately and examine how well the model can predict the target column for the ‘motivation level’ feature. Starting off by using all the features:

```
# we use 'score' to obtain the accuracy of the model
logis_score = logis_model.score(X_test, y_test)
```

```
print("logis_score (accuracy): ", logis_score)
```

```
logis_score (accuracy):  0.7777777777777778
```

```
# mean squared error value - score
print("MSE: ", mean_squared_error(y_test, logis_predictions))
```

```
MSE:  0.2222222222222222
```

Calculating the accuracies and error loss metrics we obtain:

Accuracy (using all the features): ~ 0.78 or ~78%
Mean squared error (using all the features): ~ 0.22

```
coefficients:  [[ 0.04183711  0.08192726  0.13381559  0.01673438  0.196
57163  0.10510845
   0.17046469  0.20789842 -0.23226654 -0.3997054  -0.0868913   0.154244
9
  -0.14890487  0.51337743 -0.30354352  0.36220993 -0.13788301  0.266862
18
  -0.06799211 -0.08546211 -0.1172202  -0.1218795  -0.09157634]]
```

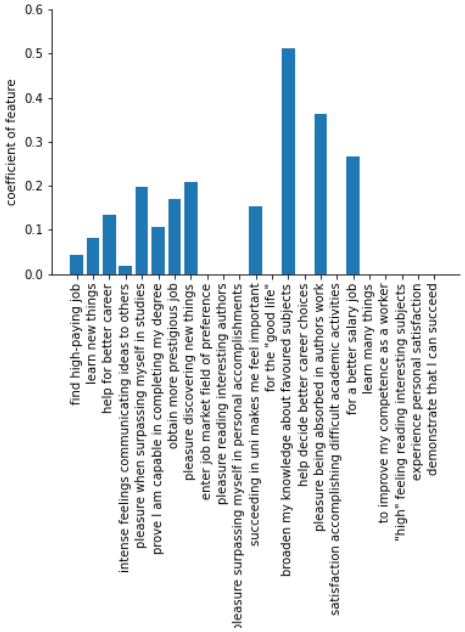


Fig. 30

Calculating the coefficients of the features, we plot a feature importance graph to be able to see which of the features are the most significant and play the biggest roles in predicting and preserving the data.
(for our model which used all the features)

Our results point to features of:
- ‘for a better career’
- 'pleasure when surpassing myself in studies'
- 'obtain more prestigious job'
- 'pleasure discovering new things'
- 'broaden my knowledge about favoured subjects'
- 'pleasure being absorbed in authors work'
- 'for better salary job'

Initially we can gather that the top two most significant features for motivation are: 'broaden my knowledge about favoured subjects', 'pleasure being absorbed in authors work' – and this is very evident in the feature importance plot.

To see their individual significances in predicting the data, we isolate those features and model them, seeing how well just by themselves they are able to predict the data. We can then compare their performances in the next section.

```
# we use 'score' to obtain the accuracy of the model
logis_score = logis_model.score(X_test, y_test)
```

```
print("logis_score (accuracy): ", logis_score)
```

```
logis_score (accuracy):  0.7222222222222222
```

```
# mean squared error value - score
print("MSE: ", mean_squared_error(y_test, logis_predictions))
```

```
MSE:  0.2777777777777778
```

Calculating the accuracies and error loss metrics we obtain:

Accuracy (using just significant features): ~ 0.72 or ~72%
Mean squared error (using just significant features): ~ 0.278

```
coefficients:  [[0.04666401 0.07757667 0.09614871 0.10201972 0.24876978
0.04522047
  0.14676685]]
```

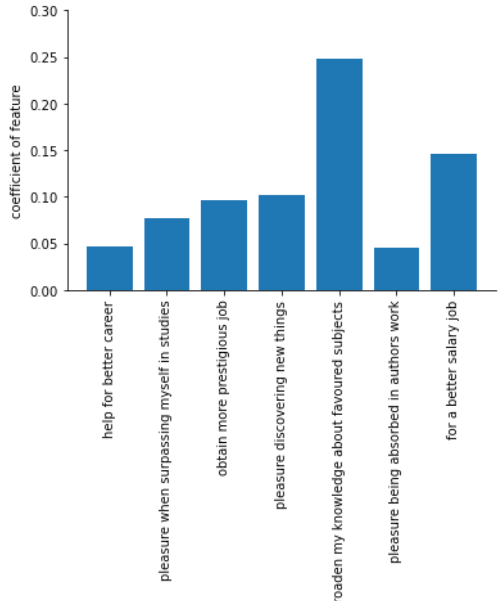


Fig. 31

Once again, by calculating the coefficients of the features and plotting.

This time are we used just the most significant features of data for the model, we obtain the true most significant features finding:

- 'broaden my knowledge about favoured subjects'
- 'for a better salary job'

as the greatest influencing characteristics for this stress dataset.
What is interesting about this is the difference in significance of features compared to the original performance using all features.

This concludes our findings from using models to see how well features can predict and preserve the data. As a results, for each dataset we have established the most significant features, the characteristics that play the most influential and encouraging roles.

Next we discuss through our findings, results, and accuracies, looking at potential trends and relationships – as well as being able to compare how each models performances compare to when we used just the topmost significant features by themselves to predict and preserve the data.

Once finally concluding our top variables that influence procrastination, stress, and motivation respectively, we may look to comparing how this relates or differs to the past literature and findings of previous works.
And finally, being in the position to provide insights about what characteristics and features are most important and essential to look into for future works, and perhaps which ones we could perhaps overlook.

8. Results & Analysis

In this section, we will be discussing through the results of our exploratory section and the results for our learning models. We will discuss the initial characteristics and features for which we first thought would be of greatest significance and in playing the biggest roles to procrastination, stress, and motivation respectively.

We will also discuss through the accuracies of the models, making comparisons between performances when using all the features and the performances for when we used just the significant features by themselves to with the aim of seeing how well on their own they were able to predict and preserve the data.
If the results are close or similar, it is a good indication that these 'significant' features are truly the most important when it comes to predicting the target variables, and we can further conclude those characteristics the most necessary for delving into when it comes to procrastination.

During the 'exploratory' portion of the project, we found that these following features were our *initial* thoughts to what may be the most significant features respectively:

Procrastination dataset: - 'how often do you put off reading' - 'I eat or drink' - 'I talk with my friends - 'I do other less important tasks' - 'distracted by entertainment' - 'distracted by new projects' - 'too many concurrent tasks' - 'bewildered to purpose of task'	Stress dataset (1): - 'nervous feelings' - 'unable to control important aspects of life' - 'could not cope with all the current tasks'
Stress dataset (2): - 'stress' - 'low energy' - 'anxiety' - 'concentration problems' - 'feeling overloaded with university work' - 'lack of time for relaxation', 'lack of confidence with academic performance'	Motivation dataset: - 'experience satisfaction learning new things' - 'will help me better prepare for my career' - 'enter job market of preference', broaden my knowledge about favoured subjects' - 'for better job salary' and 'to improve my competence as a worker'

Through the process of using supervised learning algorithms, based on these accuracies and metrics, and by calculating the coefficients of these features, we were able to gather which features in particular were of most importance.

However, we found that not all the features would remain dominant, but rather other features deemed themselves to have greater significance over some of these ones for predicting their respective target features. This led to the true results being slightly different to what we had initial hypothesised.

Dataset	ML algorithm model	Accuracy (score/r2)	Mean squared error (MSE)
Procrastination	Decision Tree	(~) -0.361	~0.333
Procrastination	Random Forest	~0.179	~0.201
Stress (1)	Decision Tree	~0.065	~6.331
Stress (1)	Random Forest	~0.573	~2.890
Stress (2)	Decision Tree	~0.074	~0.545
Stress (2)	Random Forest	~0.455	~0.321
Motivation	Decision Tree	(~) -0.833	~0.389
Motivation	Random Forest	(~) -0.099	~0.233

Note:
- Before carrying on to looking at the accuracies and performances of our models, it is worth noting that we did indeed attempt to try using other supervised learning regressor models, such as *decision trees* and *random forest's*

Table. 1

As shown in this table, the results we obtained initially were very unsuccessful. For all the models, although in general we were able to still manage retain a relatively low mean squared error, meaning most the models for this metric were able to prevent a substantial about of error loss – the accuracies and scores were still far too low, and even sometimes produced negative scores.
As r2 is the coefficient of determination, this means that the metric is used to score the performances and accuracies of regression models in particular – thus if we have a very low r2 score, this implies there is a low level of correlation with the data, and even worse for negative r2 scores, this tells us that the model fits the data very badly.
(For a couple of random forest models, we are able to show that the features used with this algorithm managed to predict and preserve a decent amount of the target variable data, however, these were still significantly lower than what we were able achieve using the other models – hence we were always led back to our initial choices.)
In hindsight, for a decision tree model, it makes especially sense that this would not work well for our datasets, not perform well. As the name suggests, this type of model revolves around decision making, solving

classification problems, even as a regressor models, (for example, a forecast prediction, deciding whether on a particular day, it will be sunny), and this does not particularly fit as a solution to our problem.

So, let's finally have a look at the summary of results of accuracies and metrics from our chosen models:

		Accuracy (score/r2)		Mean squared error (MSE)	
Dataset	ML algorithm model	All features	Only significant features	All features	Only significant features
Procrastination	Logistic Regression	~0.714 or ~71%	~0.667 or ~67%	~0.286	~0.333
Stress (1)	Linear Regression	~0.739 or ~74%	~0.679 or ~68%	~2.149	~2.644
Stress (2)	Linear Regression	~0.628 or ~63%	~0.619 or ~62%	~0.236	~0.256
Motivation	Logistic Regression	~0.778 or ~78%	~0.722 or ~72%	~0.222	~0.277

Table. 2

Here, we have constructed a clear table, which outlines our datasets, the supervised learning model of choice that was paired with the data and our accuracies & error metrics, for both modelling with all the available features, and when we used only the most significant ones.

As the accuracies we managed to find whilst using all features are scoring about ~70%, and this implies the models with the datasets are fitting pretty well. This being a relatively high accuracy score, means that when using all the features with their respective supervised learning algorithms, the models created were generally able to preserve as well predict the target features data well.
Looking at the mean squared errors metrics scores, we can see that they all reasonably low, with the exception to one, the other three are extremely close to zero. We know from before, that when we have an extremely low mean squared error, and one that is especially close to zero, this means the closer we get to a perfect model. It expresses the high accuracy of the models, showing that each one had very little error loss.

One reason why I believe that the accuracies aren't slightly higher, - {say towards predicting ~80,90%} - is due to the nature of the datasets, which we found to be relatively small, each one containing only at most a couple hundred entries – thus with more time and resources for the project and more data collected, we believe that the models would be able to predict with higher accuracies

If we examine the columns at where we built the models using 'only' the most significant features that were found from the performance of the original model, we can see the scores and accuracies are remarkably similar and not far off from the original performances. By isolating the specific features out and using just those, we were expecting their scores and accuracies to be slightly lower, however, to truly test the importance of the features, if would depend on how much lower these scores were.
Since for both R2 scores and mean squared errors, the values obtained are extremely close to that of their original performances when using all the features, we can conclude that the features used the second time around, are the most significant characteristics for their respective datasets, as by themselves, they are able to preserve and predict almost all the accuracy of their original counterpart scores.

From the *procrastination dataset*, we found that when using all features to predict the target outcome: 'put off completing documents', 'distracted by new projects', 'importance of task', 'dislike task', 'bewildered by purpose of task' and 'task difficulty' – were the most significant features.

Using only these significant features, and by themselves modelling the data we then narrowed this list down to: **'importance of task'**, **'bewildered by purpose of task'** and **'task difficulty'**

From the *stress dataset (1)*, we found that when using all features to predict the target outcome: 'nervous feelings', 'could not cope with all the current tasks', 'upset due to unexpected circumstance', 'felt difficulties accumulating'.

Using only these significant features, and by themselves modelling the data we then narrowed this list down to: **'nervous feelings'** and **'could not cope with all the current tasks'**

From the *stress dataset (2)*, we found that when using all features to predict the target outcome: 'low energy', 'headaches', 'anxiety', 'sadness', 'university work overloaded' and 'lack of time for relaxation'.

Using only these significant features, and by themselves modelling the data we then narrowed this list down to: **'anxiety'** and **'sadness'**

From the *motivation dataset*, we found that when using all features to predict the target outcome: 'for a better career', 'pleasure when surpassing myself in studies', 'obtain more prestigious job', 'pleasure discovering new things', 'broaden my knowledge about favoured subjects', 'pleasure being absorbed in authors work' and 'for better salary job'.

Using only these significant features, and by themselves modelling the data we then narrowed this list down to: **'broaden my knowledge about favoured subjects'** and **'for a better salary job'**

9. Discussions, Summary & Conclusions

9.1 – Discussions & Summary

As a quick reminder of the aims and objectives of our project, we sought out to determine what variables and features seemed to have the greatest roles in influencing and encouraging procrastination and stress in our day to day lives. We wanted to understand the rationale behind both the conscious and unconscious decisions we as humans naturally take when it comes to facing the effects of procrastination.

Through exploratory data analysis, and use of our distinct procrastination, stress, and motivation datasets, we were able to produce initial investigations to what we believed could have potential be the features that we were looking for.
Thereafter, we constructed our models and ran through our datasets, finally being able to establish characteristics for each respective datasets that proved the most significant, and that were able the predict the target variable data the best.

From the procrastination dataset, when we were able to narrow our findings to the few top-most significant features, we found that only one of them matched and remained from our initial thoughts and investigations via exploratory data analysis. This feature was 'bewildered by purpose of task'. As a top characteristic, it shows when we are at a lost for understanding behind what the objective of tasks we must accomplish are, this pushes us towards the likeliness of procrastination. Taking notice of the other top significant features of this dataset, we have 'importance of task', and 'task difficulty', completing a trend about any assumptions we may have about procrastination being solely revolved about the task at hand.

This is a result that is prevalent throughout research, so it's unsurprising that we had found similar results. Lieberman (2019) has expressed many a time how procrastination is strongly influenced by the tasks at hand that we avoid. She mentions how procrastination can affect all of us, even on different levels. For example, when she narrows the behaviour to two main reasonings: the first being that that task just comes to us as naturally desirable; then the second emerging from much deeper explanations that closely relate to the task [2].

O'Donoghue & Rabin add to this, for these reasons it is common for us to have the tendency to procrastinate for more important goals, rather than less important ones [5], ([11]).

From stress dataset (1), filtering the features and only leaving behind the most significant ones, we found that actually, both of them were among the list that we initially created during the exploratory section for this dataset. These features were, 'nervous feelings' and 'could not cope with all the current tasks'. Again, one of these features is task-related following strongly from the procrastination dataset, helping our case for the hypothesis of the unbreakable link of stress and procrastination and their causations being strongly based on the tasks at hand.
Nervous feelings can be ambiguous and difficult to measure, however, one thing sure about it is its unmistakeable relations to the other core features such as stress and anxiety. This feature may not be a main cause or initiator to procrastination; however, it certainly stems off commonly when it comes to procrastination and stress effects – it has settled its importance towards the two issues by displaying its value in the data.

From the stress dataset (2), we only found that one of the features remained after narrowing down the features to the topmost significant ones. Approaching naturally as one might assume, this feature was 'anxiety' – commonly associated with stress and procrastination, it was no surprise that this was amongst the most influential characteristics and not only that, by being the most significant by quite some margin compared to the rest. As with the previous studies mentioned, many others have expressed the intensely great links anxiety, stress, and procrastination that bond them together. Solomon & Rothblum (1984) and Lay et al. (1989), have shown through their research supporting evidence for the deep correlations that they share with one another [4] [48, 49]).
The other most significant characteristic for this stress dataset was 'sadness', which is a more interesting concept compared to the rest we have establish thus far. As sadness is a display of emotion, it is peculiar how much emotion has popped up and linked within the past literature.
Lieberman (2019) was the core study to which we read about establishing the links of emotion as the cause to that of the effects of procrastination saying that in the causations of procrastination is not just a coping mechanism or a productivity defect, but that the answer truly begins with gaining a control over our emotions [2]. She adds that, the effects of procrastination needs be dealt with in oneself and therefore we cannot depend on external factors [2].

Lastly, from the motivation dataset again, we were able to find that the topmost influential features were again amongst the list from the exploratory section of this dataset. These features were 'broaden my knowledge about favoured subjects' and 'for a better salary job'; suggesting that motivation for studying is largely based on the subject at hand and the aspiration for job seeking after education. This further supported by the other significant features that were brought to attention after the first round of using the supervised learning algorithms. These other features were, 'experience satisfaction learning new things', 'will help me better prepare for my career', 'enter job market of preference' and 'to improve my competence as a worker. Very evidently the rest of the significant features cohort also link towards preference of subject/task and desires towards certain jobs.
This is good support for a hypothesis we pondered earlier in this study, that: the more we study and are clear about our aspirations towards studying, the greater will and motivation we have to completing tasks to reaching these end goals. The older we are, the more we experience, the clearer we would be about why we choose to study, as well as the motivations behind it.
Zarick & Stonebraker (2009), have talked about how researchers have established task aversion to be a significant causable feature, they followed with this saying that, we delay in acting with tasks in which for the better, we should be doing. Generally, we are less prone to acting where we do not want to [5].
This links towards what they had established with 'uncertainty' being another highly causable feature. When we are uncertain of things, we tend have to be upfront with the task, planning well in advance, accompanied with the fear to act, and the fear of the possible outcomes [5].

9.2 - Conclusions

Procrastination is indeed a complex phenomenon, a perplexing thing that is both difficult to define and measure. Along with past researchers, procrastination is far too complicated, we too cannot narrow down the explanations and causations to just one feature, but the effects seem to be accretion of many characteristics.
Perhaps the effect of procrastination is something we may never be able to find one true solution for, however, as with the aims of past research, this one alike aimed to shed greater light on the matter and provide insights to establishing which features would be best to look into.

Using EDA and the supervised learning models of linear regression & logistic regression, we better understand the significance and importance of particular features; backed by the data, we believe the features established in our findings are most important for future works.
These key characteristics are: ‘importance of task’, ‘bewildered by purpose of task’, ‘task difficulty’, ‘nervous feelings’, ‘could not cope with all the current tasks’, ‘anxiety’, ‘sadness’, ‘broaden my knowledge about favoured subjects’ and ‘for a better salary job’.

As for the less important features, even if for the majority as first glance they may seem crucial for exploration and analyses, we do not say that these bear no significance, but perhaps the roles that they do play are of little relevance, and maybe it would be of greater optimality to focus efforts elsewhere. For example, in the procrastination dataset, we too would’ve thought features such as ‘distracted by entertainment’ would have great influence in encouraging procrastination, however, the results point to little significance – suggesting perhaps distractions or entertainments possibilities in general do not really need to be looked into, but something found to have a significant role such as ‘task difficulty’ would be better explored.

Although this project had carried out extensive research and distilled an understanding into the topic of procrastination, and carefully implemented explorations and modelling, there were of course some limitations. As briefly mentioned this mostly came with the datasets being of a relatively short nature. As the overall time for the project was brief, we strongly believe that with more time, we would be able to grasp more datasets, providing stronger evidence for our findings.

Another approach that we could’ve taken, had time allowed it, was to produce our own survey/questionnaire – with procrastination being such a complex figure, there are countless routes we could take; to gather our own substantial amount of data; to explore what initial insights we would find from this; and to compare this to the work of others.

As a final note, we understand that this project and the results may not be perfect – however the main goal was always to provide useful insights, insights into understanding procrastination and what features bare the greatest significance by extracting from a variety of different datasets – to aid in future work, seeing what kind of information I can relay in helping to avoid procrastination.

References

[1] Harriott, J., & Ferrari, J. (1996). “Prevalence of procrastination among samples of adults”. Psychological Reports, 78, 611–616.

[2] Lieberman, C. (2019). “Why You Procrastinate (It Has Nothing To Do With Self-Control)”. Available at: https://www.nytimes.com/2019/03/25/smarter-living/why-you-procrastinate-it-has-nothing-to-do-with-self-control.html?smid=url-share

[3] McLeod, S. A. (2017). “Behaviourist Approach”. SimplyPsychology. Available at: https://www.simplypsychology.org/behaviorism.html

[4] Tice, D. M., & Baumeister, R. F. (1997). “Longitudinal Study of Procrastination, Performance, Stress, and Health: The Costs and Benefits of Dawdling”. Psychological Science, 8(6), 454–458. Available at: http://www.jstor.org/stable/40063233

[5] Zarick, L. M., & Stonebraker, R. (2009). “I’LL DO IT TOMORROW: THE LOGIC OF PROCRASTINATION”. College Teaching, 57(4), 211–215. Available at: http://www.jstor.org/stable/25763397

[6] Chu, A. H. C., & Choi J. N., (2005). “Rethinking Procrastination: Positive Effects of "Active" Procrastination Behaviour on Attitudes and Performance”. The Journal of Social Psychology. 145:3, 245-264. Available at: 10.3200/SOCP.145.3.245-264

[7] Burka, J. B., & L. M. Yuen. (1983). “Procrastination: Why you do it, what to do about it”. Addison-Wesley. Sourced via: [5]

[8] Sirois, F. & Pychyl, T. (2013). “Procrastination and the priority of short-term mood regulation: Consequences for future self”. *Social and Personality Psychology Compass, 7*(2), 115–127. Available at: https://eprints.whiterose.ac.uk/91793/1/Compass%20Paper%20revision%20FINAL.pdf

[9] Akerlof, G. A. (1991). “Procrastination and obedience”. American Economic Review 81 (2): 1-19. Available at: https://pages.ucsd.edu/~aronatas/project/academic/akerlof%20on%20procrastination.pdf

[10] Xia, R. & Sun, Y. & Zhang, F. (2020). “Do8Now: An Intelligent Mobile Platform for Time Management using Social Computing and Machine Learning”. 165-173. Available at: https://aircconline.com/csit/abstract/v10n12/csit101217.html

[11] O'Donoghue, T., & M. Rabin. (1999). “Incentives for procrastinators”. Quarterly Journal of Economics 114(3): 769-816. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1040.8680&rep=rep1&type=pdf

[12] Hershfield H. E., (2011). “Future self-continuity: how conceptions of the future self-transform intertemporal choice”. Ann N Y Acad Sci. 1235:30-43. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3764505/

[13] Ellis, A. & W. J. Knaus. (1977). Overcoming Procrastination. New York. Signet. Available via: [5]

[14] Rothblum, E. D., L. J. Solomon, and J. Murakami. (1986). “Affective, cognitive, and behavioural differences between high and low procrastinators”. Journal of Co. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.901.8939&rep=rep1&type=pdf

[15] Senecal, C, Koestner, R., & Vallerand. R. J.. (1995). “Self-regulation and academic procrastination”. Journal of Social Psychology 135 (5): 607-20. Available at: s://selfdeterminationtheory.org/SDT/documents/1995_SenecalKoestnerVallerand_JSP.pdf

[16] Tuckman, B. W. (1998). “Using tests as an incentive to motivate procrastinators to study”. Journal of Experimental Education 66 (2): 141^47. Available at: https://www.tandfonline.com/doi/abs/10.1080/00220979809601400

[17] Orpen, C. (1998). “The causes and consequences of academic procrastination”: a research note. Westminster Studies in Education 21:73-75. Available at: https://www.tandfonline.com/doi/abs/10.1080/0140672980210107

[18] Schraw, G., L. Olafson, and T. Wadkins. (2007). “Doing the things, we do: A grounded theory of academic procrastination”. Journal of Educational Psychology 99 (1): 12-25. Available at: https://vt.instructure.com/files/317681/download?download_frd=1

[19] Baumeister, R. F., Heatherton, T. F., & Tice, D. M. (1994). “Losing control: How and why people fail at self-regulation”. San Diego: Academic Press. Available at: http://courses.ucsd.edu/gkoob/psych188/188_losing_control.pdf

[20] Knaus, W. J. (2000). “Procrastination, blame, and change”. Journal of Social Behaviour and Personality, 15, 153–166. Available at: https://www.proquest.com/docview/1292314415?&imgSeq=1

[21] Ferrari, J. R., Parker, J. T., & Ware, C. B. (1992). “Academic procrastination: Personality correlates with Myers-Briggs types, self-efficacy, and academic locus of control”. Journal of Social Behavior and Personality, 7, 595–602. Available at: https://www.researchgate.net/profile/Joseph-Ferrari/publication/232553014_Academic_procrastination_Personality_correlates_with_Myers-Briggs_Types_self-efficacy_and_academic_locus_of_control/links/5556b15408ae6943a8734cce/Academic-procrastination-Personality-correlates-with-Myers-Briggs-Types-self-efficacy-and-academic-locus-of-control.pdf

[22] Lukas, C. A. & Berkling, M. (2018). “Reducing procrastination using a smartphone-based treatment program: A randomized controlled pilot study”. Internet Inventions. Volume 12, Pg 83-90. Available at: https://doi.org/10.1016/j.invent.2017.07.002

[23] Ben-Zeev, D., Schueller, S.M., Begale, M. et al. (2015) “Strategies for mHealth Research: Lessons from 3 Mobile Intervention Studies”. Adm Policy Ment Health 42, 157–167 Available at: https://doi.org/10.1007/s10488-014-0556-2

[24] Ly, K.H., Dahl, J., Carlbring, P. et al. (2012). “Development and initial evaluation of a smartphone application based on acceptance and commitment therapy”. SpringerPlus 1, 11. Available at: https://doi.org/10.1186/2193-1801-1-11

[25] Juarascio, A. S., Manasse, S. M., Goldstein, S. P., Forman, E. M., & Butryn, M. L. (2015). “Review of smartphone applications for the treatment of eating disorders”. European Eating Disorders Review, 23(1), 1-11. Available at: https://onlinelibrary.wiley.com/doi/10.1002/erv.2327

[26] Porta, M. (2007) Human–Computer input and output techniques: an analysis of current research and promising applications. Artif Intell Rev 28, 197–226. Available at: https://doi.org/10.1007/s10462-009-9098-5

[27] E. Mattila et al., (2008). "Mobile Diary for Wellness Management - Results on Usage and Usability in Two User Studies," in *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 4, pp. 501-512, Available at: https://ieeexplore.ieee.org/document/4530642

[28] Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P. & Waibel, A., (1990). “Machine learning. Annual review of computer science”, 4(1), pp.417-433. Available at: https://www.annualreviews.org/doi/abs/10.1146/annurev.cs.04.060190.002221

[29] Jordan, M.I. & Mitchell, T.M., (2015). “Machine learning: Trends, perspectives, and prospects”. Science, 349(6245), pp.255-260. Available at: https://www.science.org/doi/full/10.1126/science.aaa8415?casa_token=O2711I4mv_UAAAAA%3AVelORjUmB6TdKCnAqQ6M0Nm4f4f04bajTg9RY8Lg5OA0EOGce1-N01oGoHdJkGLMiYtc-O7eSD9Tcg

[30] Brown, S. (2021). “Machine learning explained”. MIT Sloan. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

[31] Yufeng, G. (2017). “The 7 Steps of Machine Learning”. Towards Data Science. Available at: https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e

[32] Bi, Q., Goodman. K. E., Kaminsky, J. & Lessler, J. (2019). “What is Machine Learning? A Primer for the Epidemiologist”, American Journal of Epidemiology, Volume 188, Issue, Pages 2222–2239, Available at: https://doi.org/10.1093/aje/kwz189

[33] Maulud, D. and Abdulazeez, A.M., (2020). “A review on linear regression comprehensive in machine learning”. Journal of Applied Science and Technology Trends, 1(4), pp.140-14 Available at: https://jastt.org/index.php/jasttpath/article/download/57/20

[34] Vellido, A., Martín-Guerrero, J.D., and Lisboa, P.J., (2012). “Making machine learning models interpretable”. In ESANN (Vol. 12, pp. 163-172). Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.431.5382&rep=rep1&type=pdf

[35] Singh, N. (2020). "Advantages and Disadvantages of Linear Regression". OpenGenus. Available at: https://iq.opengenus.org/advantages-and-disadvantages-of-linear- regression/

[36] LaValley, M.P., (2008). “Logistic regression”. Circulation, 117(18), pp.2395-2399. Available at: https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.106.682658

[37] Rout, A. R. (2022). “Advantages and Disadvantages of Logistic Regression”. GeeksforGeeks. Available at: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

[38] Somvanshi, M., Chavan, P., Tambade, S. and Shinde, S. V. (2016). "A review of machine learning techniques using decision tree and support vector machine," 2016 International Conference on Computing Communication Control and automation (ICCUBEA), pp. 1-7, Available at: 10.1109/ICCUBEA.2016.7860040.

[39] Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A. and Brown, S.D., (2004). “An introduction to decision tree modelling”. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6), pp.275-285. Available at: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.873?casa_token=8HgJ2CS_qQcAAAAA:7A4Pds23ZvRSvQGJHO6g3Owl-SIrffFIHQkx8fWF8DarDJQ2ywmA2L_qI3wpShcKp7xWhwqEG7gDBg

[40] Karabiber, F. “Gini Impurity”. Learndatasci. Available at: https://www.learndatasci.com/glossary/gini-impurity/

[41] Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M.J.O.G.R., (2015). “Machine learning predictive models for mineral prospectively: An evaluation of neural networks, random forest, regression trees and support vector machines”. Ore Geology Reviews, 71, pp.804-818. Available at: https://www.sciencedirect.com/science/article/pii/S0169136815000037

[42] Segal, M.R., (2003). “Machine learning benchmarks and random forest regression”. escholarship. Available at: https://escholarship.org/uc/item/35x3v9t4

[43] Kasuya, E., (2019). “On the use of r and r squared in correlation and regression”. (Vol. 34, No. 1, pp. 235-236). Hoboken, USA: John Wiley & Sons, Inc.. Available at: https://esj-journals.onlinelibrary.wiley.com/doi/full/10.1111/1440-1703.1011

[44] Rowe, W. (2018). “Mean Square Error & R2 Score Clearly Explained”. sci-kit learn guide. Bmc. Available at: https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/

[45] Hodson, T. O., Over, T. M., & Foks, S. S. (2021). “Mean squared error, deconstructed”. Journal of Advances in Modelling Earth Systems, 13, e2021MS002681. Available at: https://doi.org/10.1029/2021MS002681

[46] Murphy, A.H., (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly weather review, 116(12), pp.2417-2424. Available at: https://journals.ametsoc.org/view/journals/mwre/116/12/1520-0493_1988_116_2417_ssbotm_2_0_co_2.xml

[47] Alasadi, S.A. and Bhaya, W.S., (2017). "Review of data preprocessing techniques in data mining". Journal of Engineering and Applied Sciences, 12(16), pp.4102-4107. Available at: https://www.academia.edu/download/54509277/4102-4107.pdf

[48] Solomon, L.J., & Rothblum, E.D. (1984). "Academic procrastination: Frequency and cognitive behavioural correlates". Journal of Counselling Psychology, 31, 503-5

[49] Lay, C.H., Edwards, J.M., Parker, J.D.A., & Endler, N.S. (1989). "An assessment of appraisal, anxiety, coping, and procrastination during an examination period". European Journal of Personality, 3, 195-2. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/per.2410030305

Appendix

All the files and datasets for the project can be found on the git server of the University of Birmingham at the following URL: https://git-teaching.cs.bham.ac.uk/mod-msc-proj-2021/dzl199.git

The code was produced using Jupyter notebooks.
At the time of this project, we used Python version, 3.9.7.
Libraries/modules installed: NumPy, Matplotlib, sci-kit, Seaborn.