UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

DESIGN AND IMPLEMENTATION OF AN AGENT-BASED SOCIAL SIMULATION MODEL FOR ANALYSING THE INFLUENCE OF PERSONALITY ON STRESS MANAGEMENT

ROBERTO MARTÍN LUENGO
JUNIO 2022

TRABAJO DE FIN DE GRADO

Título:	Diseño de un modelo de simulación basado en agentes para el análisis de la influencia de la personalidad en la gestión del estrés
Título (inglés):	Design and implementation of an agent-based social simulation model for analysing the influence of personality on stress management
Autor:	Roberto Martín Luengo
Tutor:	Sergio Muñoz López
Departamento:	Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	—
Vocal:	—
Secretario:	—
Suplente:	—

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO FIN DE GRADO

DESIGN AND IMPLEMENTATION OF AN AGENT-BASED SOCIAL SIMULATION MODEL FOR ANALYSING THE INFLUENCE OF PERSONALITY ON STRESS MANAGEMENT

Roberto Martín Luengo

Junio 2022

Resumen

En la sociedad actual el estrés forma parte de nuestro día a día. Dado su enorme impacto en la salud, se han realizado numerosos estudios con el fin de analizar diferentes factores que influyen en mayor o menor medida en el estrés de las personas. La respuesta que el individuo da a las distintas situaciones que le generan estrés cada día, viene condicionada precisamente por uno de estos factores, su personalidad. Así pues, en el mismo entorno y con idénticas condiciones de estrés, nos encontramos con formas de ser que están más predispuestas a estresarse que otras. Sin embargo, debido a la complejidad de los factores que intervienen, modelar la relación entre el estrés y las respuestas de las personas es extremadamente dificil.

Este proyecto pretende resolver este desafio mediante el diseño e implementación de un modelo de simulación basado en agentes que permita estudiar cómo influye la personalidad de cada individuo en la gestión del estrés. Para ello, se parte de un dataset público con los registros diarios de estrés de 45 estudiantes a lo largo de 3 meses, teniendo cada uno de ellos una personalidad asociada mediante el test del Big Five.

Se realiza un análisis detallado de los datos desde distintos puntos de vista, obteniendo como resultado modelos diferentes dependiendo del enfoque. Estos modelos permiten estimar el estrés diario de una persona en función de su personalidad y de la carga de trabajo. Esta última se mide a partir de la cantidad de fechas de entrega y el número de horas de trabajo.

A partir de estas expresiones, se proponen varios modelos de simulación que son validados con los datos del dataset para ver cuál se ajusta más a la realidad.

El sistema de simulación está programado con Python. En concreto, se utiliza un framework de modelado basado en agentes.

Palabras clave: Diseño, Modelo de Simulación basado en Agentes, Estrés, Personalidad, Dataset

Abstract

Nowadays stress is part of our daily lives. Given its significant impact on health, numerous studies have been conducted in order to analyse different factors that influence to people's stress to a greater or lesser extent. The response that the individual gives to the different situations that generate stress every day is shaped precisely by one of these factors, his or her personality. Thus, in the same environment and with identical stress conditions, we find lifestyles that are more predisposed to stress than others. However, due to the complexity of stress factors involved, modeling the relationship between stress and people's responses is extremely challenging.

This project aims to solve this challenge by the design and implementation of an agent-based simulation model that allows us to study how the personality of each individual influences stress management. To do so, we start from a public dataset with the daily stress records of 45 students over 3 months, each of them having an associated personality through the Big Five test.

Then, a detailed analysis of the data is carried out from different points of view, obtaining as a result different models depending on the approach. These models allow us to find a person's daily stress level based on their personality and workload. The latter is measured from the number of deadlines and the number of working hours.

Based on these expressions, several simulation models are proposed that are validated with the data from the dataset to see which one best fits reality.

The simulation system is programmed with Python using an agent-based modeling framework.

Keywords: Design, Agent-based Simulation Model, Stress, Personality, Dataset

Agradecimientos

Son muchas las personas a las que debo agradecer su apoyo.

A mis padres, Isidro y Pilar, por motivarme a conseguir mis objetivos y a no rendirme ante las dificultades. A mi hermano, Javi, por demostrarme que la constancia es el camino para llegar a la meta. A mi tía, Cristi, por su gran paciencia y sus acertados consejos.

A mis amigos, por todos los momentos compartidos durante estos años, y en especial a los de siempre, que han estado conmigo en todo momento.

A mi tutor, Sergio, hacia quien sólo puedo expresar mi mas sincero agradecimiento por su guía y dedicación en el desarrollo de este proyecto.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Context

In the social, personal, and economic situations that we experience on a daily basis, multiple events that might be deemed stressful are involved. Consequently, stress is currently a popular issue in psychology and labor study.

Stress is a natural feeling of not being able to cope with specific situations [1]. It's a fight-or-flight response system that informs an individual when and how to react in an overwhelming event. However, if a person does not take the appropriate measures to manage stress, it can have a negative impact on their health. According to a survey, 7 out of 10 Spaniards acknowledge having experienced stress in the last month in some way [2]. Only 30.4% of the sample said they had not felt any form of stress, while 16.5% said they had felt it more than half of the days or practically every day.

As the type of work has significantly changed in recent decades, the volume of stress-related disorders is on the rise. Tasks that used to demand physical power now frequently involve mental effort [3]. Work is a common factor among younger and more stressed age groups. It is typically beneficial to us since it provides structure to our lives and most people find it satisfying. Workplace pressure is usually a positive thing since it motivates

you to perform better and prepares you for new challenges. However, if the pressure and demands escalate, they may result in occupational stress [4].

Occupational stress is the response to a work-related disturbing external agent. This agent is the stressor, the stimulus that triggers the stress response. It would be impossible to compile an entire list of stressors given the large number of potential factors. Nonetheless, some external factors are considerably more likely to act as stressors than others, such as workload.

Students do not have occupational stress as their main stress, but rather academic stress [5]. Academic stress is the one suffered by students mainly in secondary and higher education, and whose exclusive source is stressors related to the activities to be carried out in the school environment. It can be considered that the academic demands most frequently perceived as stressors by university students are the academic overload and the exams. These two factors, as one can see, are nearly identical to those defined in the occupational stress.

Individual differences that influence reaction or adaptation to the environment are referred to as personality [6]. Although these answers may be derived from people's regular behavior patterns over time, they do not imply that personality is the same as conduct, but rather that it is a higher-level concept. Personality can influence the impact of stress by altering one's subjective assessment of life's situations as more threatening or as manageable and potentially challenging.

Numerous studies have shown that there is an influence of certain personal characteristics in the development of stress [7]. It does not imply that owning a group of specific character traits triggers stress by itself, but rather that it raises the sensitivity of these individuals, making them more vulnerable to stress when particular events or demands emerge. The reactions of different people to the many stressors that might impact them are influenced by their personality profiles in certain ways, so not everyone will respond in the same way.

Due to the complexity of this internal factor, modeling the relationship between stress and people's responses is extremely challenging. The purpose of this research is to solve this challenge by developing an agent-based social simulation system for investigating the impact of personality on coping with stress. Besides personality, this project also considers the external factors previously mentioned, such as workload, to perform the analysis. However, the idea is not only to develop the system but also to offer the resources needed to show and evaluate the outcome.

1.2 Project goals

The goal of this project is to develop a realistic simulation system useful for analysing the influence of personality and other factors on stress management. Since not all personalities handle stress in the same manner, the system will allow users to separate by personality in order to determine the best deadline distribution that will minimize their stress levels.

To achieve this, we can identify several tasks among the project's key aims, such as:

- Study of the State of Art, research on the relationship between personality and stress.
- Investigation of simulation software and enabling technologies.
- Analysis of the Student Life Dataset.
- Design of several Agent-based Social Simulation Models.
- Validation of the simulation models with the dataset.
- Implementation of a case study with the selected simulation model.

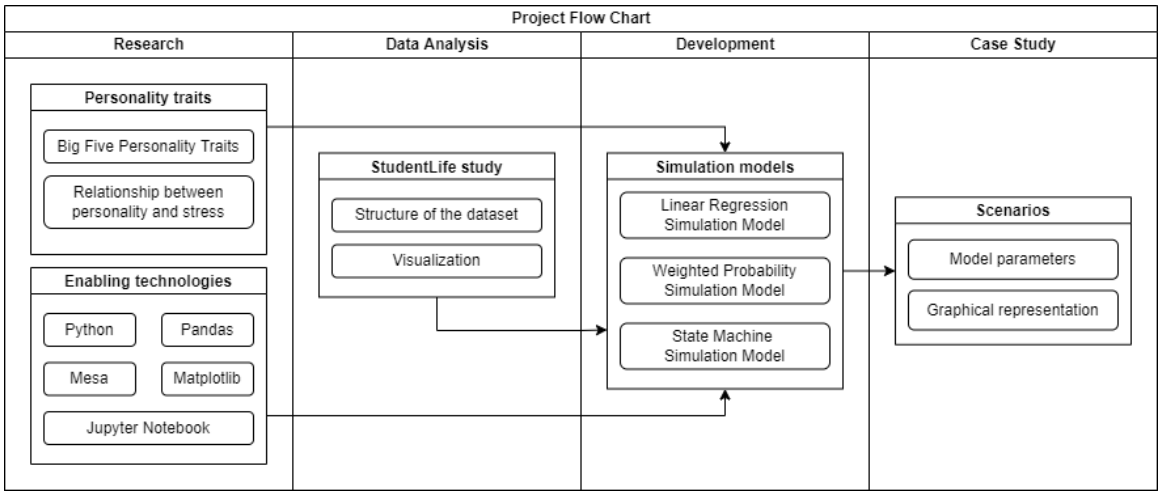

Figure 1.1: Project flow chart

1.3 Structure of this document

In this section we provide a brief overview of the chapters included in this document. The structure is as follows:

Chapter 1. Introduction. It presents an introduction of the project. Besides, it is explained the main goals of the project and the context in which this project is developed.

Chapter 2. State of the Art. It provides an overview of the recent studies about the relationship between the personality and the stress, the main technologies on which this project relies and some general background.

Chapter 3. Dataset. It presents a detailed analysis of the data from different points of view

Chapter 4. Simulation Models. It covers how the 3 Simulation Models have been designed and the analysis carried out for each one. Afterwards, it discovers which one of them best fits reality.

Chapter 5. Case study. It describes a detailed case study for the Simulation Model that most closely matches actual fact and provides a thorough analysis of the results obtained.

Chapter 6. Conclusions and future work. It joins the conclusions drawn from this project, problems faced and a brief future perspective.

Appendix A. Impact of the project. It describes the social, economic, environmental and ethical implications of the project.

Appendix B. Economic budget. It goes into detail about the economic budget required to implement the project.

CHAPTER 2

State of the Art

2.1 Introduction

Prior to doing the data analysis and designing the simulation model, this chapter provides an overview of the recent studies about the relationship between the personality and the stress, the main technologies that have made possible this project and some general background. This chapter begins with a brief review of the taxonomy for personality traits.

2.2 Personality traits

A personality trait is a consistent and stable characteristic of a person that can be defined as the explanation for a person's actions [8]. It has an impact on how we feel, think, and behave. It is based on biological factors, such as genetic influences, nutrition, gender, and social factors, like family relationships or religion. But these are not the only factors; morality, ethics, and personal convictions also play a significant role.

Gordon Allport, a psychologist, observed in 1936 that a dictionary included over 4,000 terms that might be used to define various personality traits. Since there were too many

to rank, Raymond Cattell decreased the number of main personality qualities on the list, removing unusual ones and merging common characteristics, developing the "16 Personality Factor Questionnaire."[9]. At the same time, psychologist Hans Eysenck created a personality model based on only three universal personality traits.

However, these personality models contained either too many or too few traits, which made classification problematic. Consequently, a new theory known as the "Big Five Personality Traits" [10] appeared, becoming the most popular model among psychologists for the study of personality traits.

The Big Five Personality Trait is a personality model that enables the classification of the way of being of an individual in one of the five personality traits. It is one of the most often utilized theories for defining and assessing a person's personality. The Big Five personality traits are often given the following names:

- **Openness**: People who appreciate learning and experiencing new things have a high level of openness. Personality attributes such as intuition and imagination, as well as creative and curiousity, are all examples of openness.

- **Conscientiousness**: Individuals with high levels of conscientiousness are trustworthy and punctual. Rather of being basic and messy, they are efficient, organized, methodical, and understanding.

- **Extraversion**: They are sociable, enjoy meeting friends, show a great enthusiasm for life, and love new emotions. The personality qualities that defined them are full of energy, talkative, and assertive.

- **Agreeableness**: People with high levels of agreeableness are friendly, helpful, and caring. Some of their personality attributes are considerate, sympathetic and cooperative.

- **Neuroticism**: These individuals are more likely to experience negative effects, including anger, irritability, or stress. This group is closely related to emotional instability. The personality qualities that defined them are coward, depressive, and low self confidence.

2.3 Relationship between personality and stress

Individual differences in stress reactions have been the subject of several studies [11]. In fact, stress, coping, and health have all been known to be influenced by personality factors for a long time. Previous studies of the Big Five's relationship to stress-related activities have centered on how these characteristics connect to the implementation of various coping techniques. Coping, in general, refers to cognitive and behavioral strategies for preventing, controlling, or reducing stress.

In comparison to other personality trails, people with a high level of neuroticism are the most vulnerable to stress [12]. People with a high level of neuroticism are far more vulnerable to negative affect, especially when a looming stressor is perceived as threatening. As the stressor last, there is a pattern toward higher negative affectivity.

Stress responses are amplified by neuroticism and are regulated by mental and emotional processes that assess the likelihood and severity of a threat [13]. While threat evaluations have substantial consequences for stress responses, people with a high level of neuroticism may be less stressed if threat perceptions can be reduced. In conclusion, higher levels of neuroticism are associated with more acute subjective stress reactions, which means a reduction in the beneficial effect of stress and a decreased sense of control in stressful situations.

Extraversion is associated with feelings of enjoyment, pleasure, and satisfaction, and are more likely to have positive affect [14]. Since there is no apparent connection between extraversion and stress overall, emotionally relevant stressors could be important to disclose any consequences.

Despite displaying more physiological stress, those with a higher openness score frequently have outstanding intelligence and a better sense of control. As a result, people with a high level of openness tend to avoid being in a vulnerable emotional state characterized by stress, difficulties, and a lack of control.

Previous studies have suggested that a positive or no relationship between agreeableness or conscientiousness and stress reactions [15]. Despite the fact that few research have shown a connection between these two personality traits and stress responses, conscientiousness, a personality attribute linked to good health and resilience to psychological disorders, was linked to a better stress response [16]. The same is valid for agreeableness, which is defined by being kind and understanding and is likewise a positive stress response.

2.4 Enabling technologies

The software and technologies used to implement this project are covered in this section. The programming language used, Python, will be introduced first. Secondly, Mesa, the agent modeling simulation software, will be described. Finally, the tools used for data analysis and visualization will be explained.

2.4.1 Python

Historically, mathematical modeling and analysis have been accomplished using a range of computer languages and frameworks. However, most modern data science libraries are now Python-based, due to the Python language's enormous development in popularity within the scientific computing community over the previous decade [17].

According to the TIOBE index for April 2022, Python is still the most popular language for analytical computing and data research, as it allows for the usage of low-level libraries, which improves efficiency and productivity [18].

Apr 2022	Apr 2021	Change	Programming Language	Ratings	Change
1	3	▲	Python	13.92%	+2.88%
2	1	▼	C	12.71%	-1.61%
3	2	▼	Java	10.82%	-0.41%
4	4		C++	8.28%	+1.14%
5	5		C#	6.82%	+1.91%
6	6		Visual Basic	5.40%	+0.85%
7	7		JavaScript	2.41%	-0.03%
8	8		Assembly language	2.35%	+0.03%
9	10	▲	SQL	2.28%	+0.45%
10	9	▼	PHP	1.64%	-0.19%

Figure 2.1: The Tiobe Index for April 2022

Python is a high-level interpreted programming language with a focus on readability that is well-known for being simple to learn whilst being able to use the capabilities of systems-level programming languages when needed.

Python is particularly appealing for applications in data science and scientific computing because of the community that surrounds the available tools and libraries [19]. It is used to implement the Agent-based Social Simulation system in this project.

2.4.2 Mesa

Over the last 20 years, agent-based modeling (ABM) has experienced tremendous growth, resulting in a plurality of ABM frameworks. There was, nevertheless, a gap. There was no easy way to create a model in Python, and there was no way to provide a model over HTTP, which would take use of current browser-based technology. Mesa, an open-source framework for developing agent-based models in Python, was created in response to this [20]. Modeling, analysis, and visualization are the three main categories of Mesa modules.

- The Model is the core of Mesa. It contains all the essential components of an ABM: the Model class in which the user defines the initial state of the model and the response of the system at each step, the Agent class for defining the agents, and the Scheduler which handles agent activation.

- The Analysis includes the DataCollector class which stores the data from the model and agents, and the BatchRunner class which performs parameter sweeps to acquire a more representative view of the model's probable outcomes.

- The Visualization component provides a browser-based visualization system. The charting modules, line charts, bar charts, and pie charts, allow us to represent the data from the DataCollector class. When the model is running, all of these modules are updated.

2.4.3 Jupyter Notebook

Jupyter Notebook is an open-source online tool that allows users to create and share interactive programs that include live code, equations, interactive visualizations, and graphics [21]. It began as an evolution of the Ipython Notebook interface, offering essentially the same features but with the addition of the ability to execute code in other languages.

A Jupyter Notebook file is defined by its ".ipynb" extension. It is in JSON format and contains all the data from the notebook. It also includes all the cell contents, plots, and document information.

For this project, Jupyter Notebook has been used for data analysis and scientific computing.

2.4.4 Matplotlib

Matplotlib is a Python library for creating two-dimensional plots [22]. It was written by John Hunter in 2003 and since then matplotlib comes along with a large community of users and developers.

This library supports common 2D plot types and interactive graphics, including xy plots, pie charts, bar charts and images, and it operates on all major operating systems.

Although matplotlib is largely developed in Python, it heavily relies on NumPy and other extension code to achieve high performance.

Matplotlib has been used for representing the results of the Simulation Model of this project.

2.4.5 Pandas

Pandas is a Python library specialized in the representation and analysis of data structures [23]. It has been in development since 2008 and it was originally created by Wes McKinney.

In terms of data analysis tools, this library tries to bridge the gap between Python and a variety of statistical computing platforms and database languages. Pandas creates new data structures based on the NumPy library's arrays, but with additional functionality. It enables users to read and write files in a variety of formats, as well as reorganize, split, and combine data sets.

Some of the main features of Pandas have been used in this project, such as label-based data access, data alignment, handling missing data, hierarchical indexing or grouping and aggregating data.

CHAPTER 3

Dataset

3.1 Introduction

Since project's overall purpose is to develop a simulation model that is as close to reality as possible, a reliable data source with sufficient records is required to solve the challenge. To accomplish this, it has been decided to use the dataset from the StudentLife Study [24], which is publicly accessible and has a wealth of information on the topic.

3.2 StudentLife Study

The study was conducted in a class of 48 Dartmouth students over a 10-week period. It consisted of evaluating the mental health of the students, such as the level of stress, depression, or loneliness, as well as their relationship with the workload. This relationship can be done because it also registers when they have mid-term exams, final exams, number of tasks, hours spent studying each day or even their results and grades for each of the tasks and exams. For psychological evaluation, a variety of well-known pre-post mental health surveys are employed, including personality tests.

All this process of data registration by students is done through a mobile phone application that contains all the questions. Considering the large number of students and the even higher number of questions, the data must be organized and structured in such a way that it can be analysed later. To do so, there is a repository called StudentLife Dataset that contains all the questions and answers in a neat organized format.

3.3 Structure of the dataset

The StudentLife dataset is divided into smaller datasets that enable clear differentiation between the different types. The four types that are available are: sensor data, Ecological Momentary Assessments (EMA) data, pre- and post-survey responses, and educational data. However, we will only use the last three in this project because the sensor data includes records of physical activity, conversations, and student location, which are out of the focus of this project.

3.3.1 EMA Data

This set has a large amount of information with the records that students enter in the application. Only the stress-related subdirectory, Stress, is of interest for our work out of a total of 26 subdirectories with different questions. These data sets have a similar structure, with two parts: the question definition and the participant's response, both in JSON array format. For example, the Stress EMA question is defined as shown in figure 3.1.

```
{
    'name': 'Stress',
    'questions': [
        {
            'options': '[1]A little stressed, [2]Definitely stressed, [3]
                Stressed out, [4]Feeling good, [5]Feeling great',
            'question_id': 'level',
            'question_text': 'Right now, I am...'
        }
    ]
}
```

Figure 3.1: Stress EMA data question

The question's name is defined in the name field. The questions field specifies the questions that participants must answer in order to complete this EMA. The question text, question id, and choices fields are all present in each item in the questions array. For example, if a participant answered, "Definitely stressed" to the first and only Stress EMA question "Right now, I am...", his corresponding response record will show "level": "2".

Regarding the structure of the responses, there is a JSON file for each of the students and the number of items in each file is the number of responses recorded. The value of each item is the answer, among the possible options exposed in the definition of the question.

As there is not enough data for the whole 10-week study, it's required to filter on the days with the most replies. The study begins on 03/25 with the last recorded answer dated on 06/25, but at this point the average response is between one and five students, so it is necessary to look back a few weeks to look for greater participation.

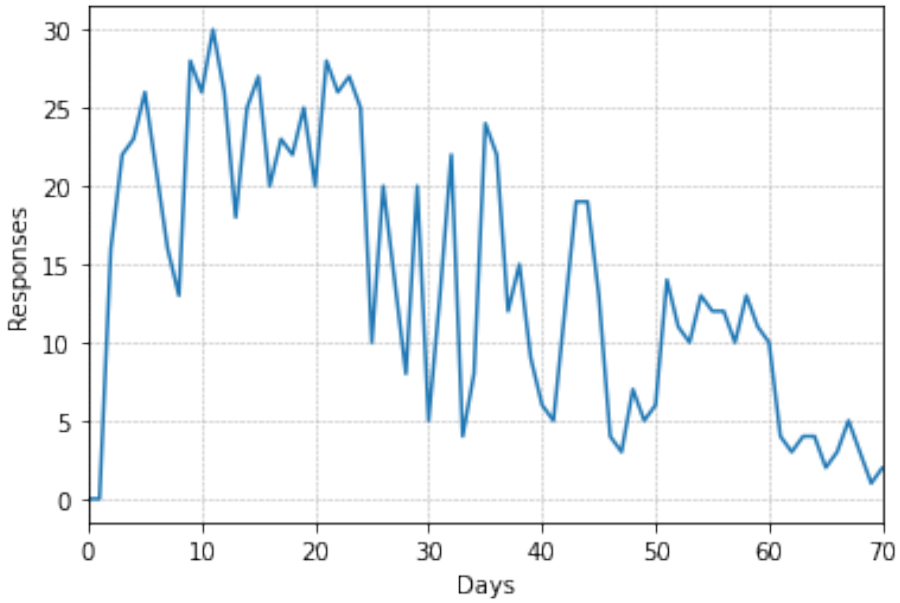

Figure 3.2: Participation in the StudentLife Study

The largest number of responses is found approximately in the first 20 days, between 03/27 and 04/20, with an average of 22 responses per day. After that, the level of participation decreases considerably in the following 20 days to an average of 12 responses per day. After 60 days, a dramatic reduction in the number of answers is seen, with only 5 records remaining, making the data worthless. As a result, the study period for this project is from 03/27 to 05/27.

Despite selecting the time period with the highest participation, some students have only recorded a few replies during the research, therefore they have been excluded since they add too much noise to the data. Furthermore, the user u00 was removed since it was the teacher, who is unaffected by deadlines.

3.3.2 Survey Data

This subset is used to determine the other essential parameter in the research, each student's personality. The directory is organized by survey names and all the files are in CSV format. From all the different surveys in the directory, the one we are interested in is the Big Five one.

Participants' responses to both pre and post mental health indicators are included in the survey responses file. The first column specifies which survey participants responded to, and the second column indicates whether the response is before starting to record the data (pre) or after (post). Each of the remaining columns correspond to a survey question. The beginning of the questions is the same in all cases, which is "I see myself as someone who..."

Uid	Type	Is talkative	Tends to find fault with others	Does a thorough job	Is depressed	Is original
u01	pre	Neither agree nor disagree	Agree a little	Agree strongly	Agree strongly	Agree a little
u01	post	Agree a little	Agree a little	Agree strongly	Agree strongly	Agree a little

Table 3.1: First 5 questions of the Big Five Survey

Within the different modalities in the measurement of personality with the Big Five, this study uses the 44-item inventory [25]. This model has a total of 44 questions with 5 possible answers, which range from "Disagree strongly" to "Agree Strongly", depending on how much the student comply with the statement.

To determine the personality, the answers provided in the file are converted to a numerical scale of 1 to 5, corresponding to "Disagree strongly" and "Agree Strongly", respectively. Then, using the answers to the 44 questions, a mathematical computation is performed for each personality, yielding five different values, one for each personality. The student's personality will be the one with the highest value, therefore the dominant one.

Considering that, in the vast majority of cases, the personality before and after the study was the same, only the values corresponding to "pre" have been used in this project. Students for whom no personality has been recorded in the file are excluded from this project. Therefore, we are left with 30 participants out of the 48 available in the research after excluding students with extremely few replies from EMA data and those with no personality.

When looking at the distribution of personality types among the 30 students based on the results of the Big Five survey, it is clear that they are not evenly distributed. The personality trait with the greatest percentage in the survey is openness, which got 40%, followed by conscientiousness and agreeableness, which got 26.7% each, and lastly neuroticism, which got just 6.7%. It's worth noting that none of the 30 participants had the extravert personality type.

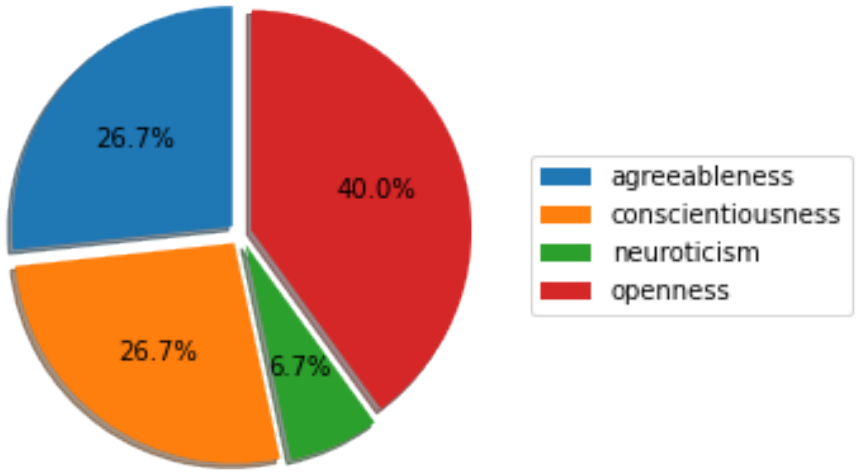

Figure 3.3: Distribution of Personality Types in Percentages

3.3.3 Educational Data

Everything related to the student's academic subject is included in this subset. At this point, having the stress records and the personality associated with the person, all that remains is to add the stressful events to the analysis. The deadlines, which consist of the quantity of tasks that the student has every day, define these stressful situations.

All of the necessary information can be found in the deadlines file, which is in CSV format. This file has a logical and organized structure, with information for each student shown separately. The number of deadlines that the individual has each day from 03/27 through the completion of the study is listed for each of them with the greatest number of deadlines per day being three and the minimum being zero. Furthermore, there are certain days with a higher concentration of deadlines than others, which might be due to the submitting of a task or in-class test.

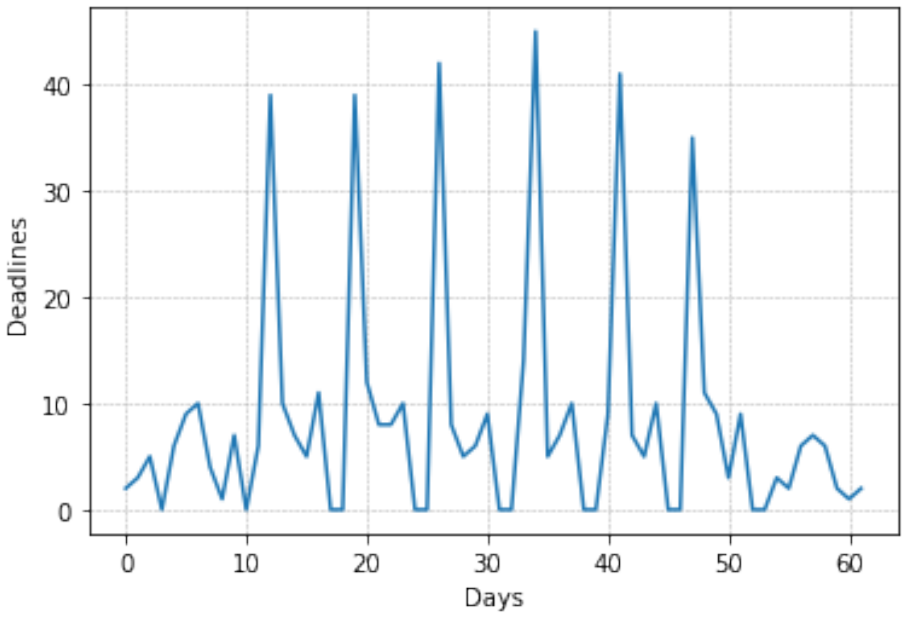

Figure 3.4: Distribution of Deadlines

The concentration of deadlines remains largely consistent throughout time and quantity, with the exception of the beginning and end of the study. As a result, the deadline peaks indicate a minimum of one assignment per student, and in some cases two or even three.

3.4 Visualization

The representation of the data is carried out after it has been filtered and thoroughly analysed, both for the stress records and the personality surveys. As previously mentioned, 30 students out of a total of 48 were selected from this analysis, since they are the ones with sufficient records for the three fundamental aspects of the project: personality, stress and deadlines.

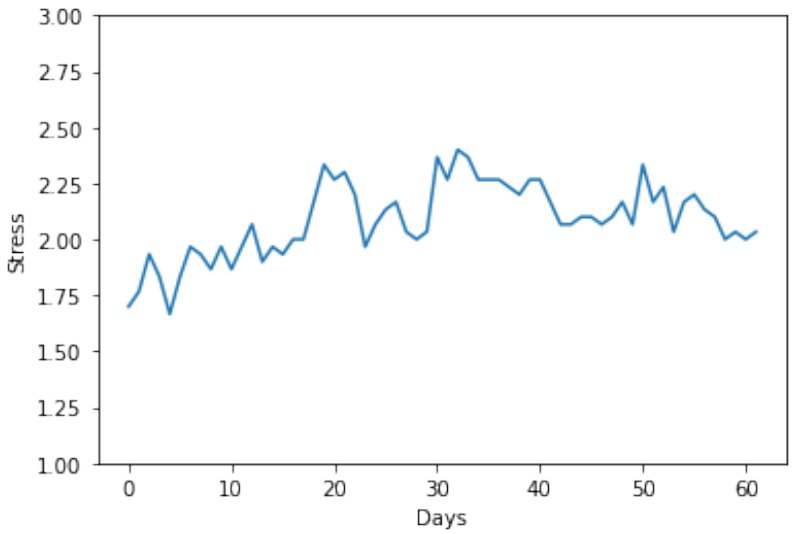

Figure 3.5: StudentLife Data Visualization

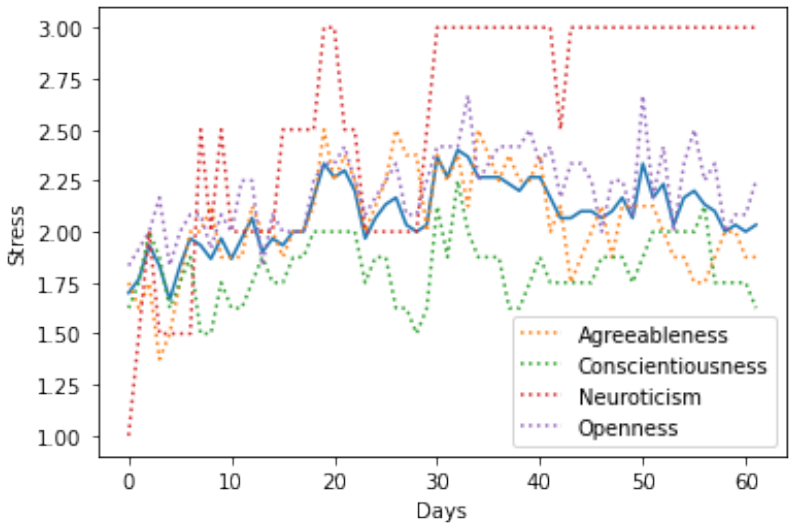

Figure 3.6: StudentLife Stress based on personality

The students' average stress levels rise throughout the course, as seen in Figure 3.5. The mean value of stress at the start of the course is 1.75, whereas the mean value at the end of the course is 2.0. Peaks of stress can also be seen in the data associated with the dates of deadline peaks. All students face at least one assignment on these dates, and in the worst-case scenario, three on the same day.

Figure 3.6 shows the average stress level of students based on their personality trait. Throughout the study, neuroticism is the personality trait that has the most stress. The rest of the personalities reach a maximum value of 2.50 during the first significant peak of deadlines, whereas neuroticism reaches a value of 3.0.

Except for some minimum stress peaks, the average level of stress of the other personalities: openness, conscientiousness and agreeableness, is generally much more similar to each other.

CHAPTER 4

Simulation Models

4.1 Introduction

This section describes the different simulation models that will be implemented in this project, as well as their corresponding components. Each of the systems is determined by a different line of research in the modeling and implementation of agents, allowing a particular approach for each of the models and a more complete global analysis. The final goal of this chapter is to validate the proposed models and to determine the simulation system with the greatest similarity to reality, in order to carry out the case study in the following chapter.

4.2 Architecture

A simulation model is divided into components or modules that determine its behavior, being the two most important ones the model and the agent. In the proposed simulation system there is only one single model, in charge of managing the whole environment. In this project, a single type of agent has been defined, the student. Each of the agents perform according to a programmed behavior and predefined personal characteristics.

In addition to the model and the agents, there are other modules that participate in the analysis and search process that is carried out with this simulation system. Therefore, the simulation system has the following components: (1) the **Model**, which is the core of the system; (2) the **Agent**, which represents the students; (3) the **Agent Behavior**, which manages the student's conduct; (4) the **Data Handler**, which records the model and agent's data; (5) the **Batch Runner**, which performs a parameter sweeps with all possible values of the model; (6) the **Analysis**, which examines the data to find the closest collection to the data set; and, finally, (7) the **Visualization**, which represents all the results in graphs and diagrams.

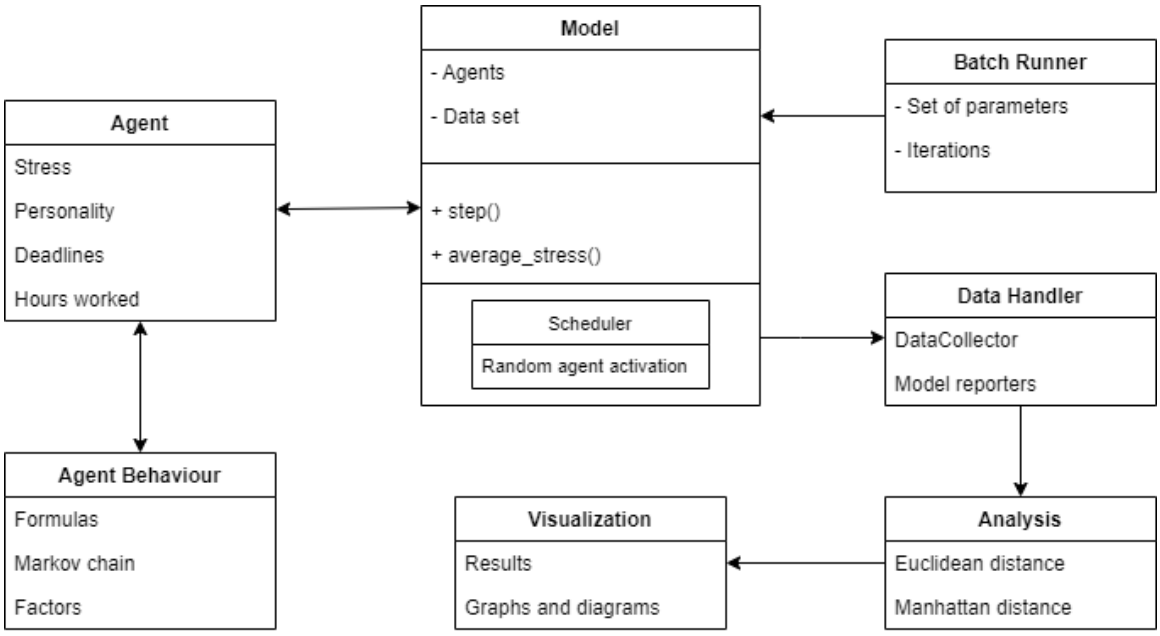

Figure 4.1: Main components of the 3 simulation models

The model has a total of 4 parameters. The seed is the first of them, and it sets the system's randomness. The same value is always input since it offers a reliable and predictable source of random numbers, resulting in the same values in every system run.

The model's second parameter is the number of agents it will contain, which must be an integer value. The third one corresponds to the simulation's start date, which should be given in the following format: 'YYYY-MM-DD hh:mm:ss'.

The last parameter is a data set that must contain the course's deadlines and hours worked. Furthermore, the dataset has to include the personalities of each of the model's agents. This data has to be imported from a CSV file.

The three most important methods of the model are:

- ***step()***. Advances the model one cycle. A step is related with a day of the academic year in this project. The model decides when the agents move from step to step
- ***average_stress()***. The average stress of the agents is calculated using this method. To accomplish this, it extracts the set of agents from the Scheduler and records each one's associated stress. Subsequently, all the values obtained are added and then divided by the number of agents in the scheduler.
- ***step_time()***. At the same time that a step is advanced, the variable that stores the current system date is advanced by one. This allows the agents to obtain the daily value of stress, deadlines, and hours worked.

A Scheduler class instance is also included in the model, which manages the agent activation. This class has several methods for performing basic operations on the agents, such as adding and deleting them from the activation buffer. RandomActivation is the class that is used, it activates in complete random order each agent once every step, being the order reorganized after each step. An agent or student has three parameters: a unique identifier, a model class instance, and dataset data. The model, which is where the agents are initialized, provides all of these parameters. The characteristics or variables of an agent are stress, deadlines and hours worked.

Once the architecture of the simulation system has been described, we proceed to describe the simulation models: 4.3 Linear Regression Simulation Model, 4.4 Weighted Probability Simulation Model and the 4.5 State Machine Simulation Model.

4.3 Linear Regression Simulation Model

This first simulation model is developed by establishing a relationship between stress and workload in the StudentLife data. This relationship is referred to as a correlation, which is a method of detecting if and how closely two variables are related. This model aims to identify if there is a linear relation between these two variables: stress and workload.

The **Pearson correlation coefficient** [26], known as the r-value in data science, is used to calculate the correlation. This coefficient can take on any value between -1 and 1. If the number is a positive value, it can be determined that there is a positive correlation and that both variables are related in the sense that if one grows, the other increases too. On the other hand, if the value obtained is a negative number, it can be inferred that there is a negative correlation and that the variables are inversely related.

The stronger the correlation, the closer the coefficient is to the range's extremes, -1 and 1. The less correlation there is, the closer to zero is the result. Therefore, a correlation value of 0 implies that there is no linear relationship between the two variables.

The formula for calculating the Pearson's correlation coefficient is given by equation 4.1.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{4.1}$$

Once the coefficient has been estimated and the variables having a linear relation to stress have been identified, the formula can be defined using a simple statistical approach known as **linear regression** [27].

The two types of variables in this statistical approach are the input variable (deadlines or hours worked), which helps predict the value of the output variable, and the output variable (stress), which is the one to be predicted. The equation for calculating the linear regression is as shown in equation 4.2.

$$Y_e = \alpha + \beta \cdot X \tag{4.2}$$

To calculate the values of α and β, a method called **least squares** is used, by which the sum of the squared difference between Y, the real stress, and Y_e, the stress to be predicted in the simulation, is minimized. The formulas for calculating the values of α and β are given by equations 4.3 and 4.4.

$$\beta = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{4.3}$$

$$\alpha = \bar{Y} - \beta \cdot \bar{X} \tag{4.4}$$

Once the variables have been defined, a mathematical formula must be obtained that relates them in order to calculate the stress in each step. The relationships studied are stress with the other two: deadlines and hours worked.

The degree of association between both variables is obtained using the Pearson correlation coefficient. The correlation calculation results for the StudentLife data set are provided in table 4.1.

	Stress	Hours worked	Deadlines
Stress	1.000000	-0.090998	0.223849
Hours worked	-0.090998	1.000000	0.216080
Deadlines	0.223849	0.216080	1.000000

Table 4.1: Correlation results

With a coefficient value of 0.22, the data show a significant positive correlation between stress and deadlines. However, there is no remarkable correlation between stress and hours worked, which is even negative with a value of -0.09. This is possibly due to low number of responses for hours worked. Therefore, since there is a direct linear relationship between the first two variables, the deadlines is the one used as the factor in the linear regression formula.

The results of α and β for the stress-deadlines data, when the daily average is performed in both variables, without separating by user or personality, are $\alpha = 2.058319$ and $\beta = 0.098901$. However, this computation is not entirely correct, because it would be more accurate to filter by user and personality, resulting in a total of four pairs of values α and β, corresponding to the different personalities of the Big Five, with the exception of Extraversion, for which no data is available.

Personality	α	β
Agreeableness	1.989536	0.210745
Conscientiousness	1.817181	0.005898
Neuroticism	2.526483	0.137635
Openness	2.185685	0.072947

Table 4.2: Linear regression variables

In general, the relationship between stress and deadlines, grouped by personality, is quite low. Agreeableness, conscientiousness and openness vary around an average stress level of 2, increasing slightly depending the daily average of deadlines. However, neuroticism has a value of 2.52 when there is no deadline, which is relatively high. The stress of the agents is defined by equation 4.5, already particularized for this project

$$Stress = \alpha + \beta \cdot Deadlines \tag{4.5}$$

4.4 Weighted Probability Simulation Model

This simulation model is characterized by the use of weighted probabilities to determine the agent's stress level. The key change between this model and the previous one is the addition of three additional parameters: prob_level1, prob_level2, and prob_level3. These variables represent the probability of getting a stress level of 1, 2, and 3 respectively.

The model is initialized with the values of these probabilities in the Batch Runner component. Since the closest values to the dataset are undetermined, the numpy library's linspace method is used, which provides evenly spaced numbers over a defined interval.

```
params = {
    'num_agents' : 30,
    'prob_level1' : np.linspace(0.0, 1.0, 20),
    'prob_level2' : np.linspace(0.0, 1.0, 20),
    'prob_level3' : np.linspace(0.0, 1.0, 20),
    'data_dir' : ['data/mw_data.csv'],
    'seed' : 24091993
}
results = batch_run(
    StressModel,
    parameters=params,
    iterations=10,
    ...
)
```

Figure 4.2: Batch runner parameters

A range of values from 0 to 1 is defined for each variable, with a total of 20 samples, generating 8000 possible data combinations. In addition, for each combination of values,

a total of 10 iterations are determined to provide more stability and reliability to each data set and consequently to the model. Therefore, taking into account the iterations per combination, the number of simulations carried out is 80,000.

The behavior of the agent is determined by the new parameters of this system that are obtained from the model: prob_level1, prob_level2, and prob_level3. Besides the new variables, the deadlines and personality parameters are still used.

In the computation of stress, three phases of analysis are developed.

1. ***Baseline***. The first version of the agent behaviour simply uses the random library's choices method with the probability distribution provided in the model. The choices method returns an element randomly drawn from a list of possible stress levels: 1, 2 or 3. The probability that each element of the list is extracted is the same by default, but this can be modified using the weights parameter, which allows weighted percentages for each level. Therefore, the probability of each of the stress levels are the values of the parameters and x corresponds to each of the possible levels of stress: 1, 2 or 3.

$$p_x = prob\_levelx \tag{4.6}$$

2. ***Workload***. The first variable that influences the agents is introduced in the second version of the agent behaviour. The number of deadlines influences the final probabilities that are entered in the choices method by multiplying the base probability by an increase or decrease factor. The formulas proposed for each of the probabilities based on the number of deadlines are as shown in equations 4.7, 4.8 and 4.9.

$$p_1 = prob\_level1 \cdot (1 - \frac{deadlines}{4}) \tag{4.7}$$

$$p_2 = prob\_level2 \cdot (1 + \frac{deadlines - 1}{2}) \tag{4.8}$$

$$p_3 = prob\_level3 \cdot (1 + \frac{deadlines}{4}) \tag{4.9}$$

Among the elements that constitute the equations, the values of the model parameters corresponding to the stress levels are prob_level1, prob_level2 and prob_level3. On the other hand, *deadlines* is the number of deadlines on that day. The workload factor, which is in parentheses in the equations, is not applied if the deadlines are zero, hence the initial version of the agent behavior is maintained.

3. ***Personality***. Personality is introduced as the second variable that influences agent behavior in the third version of the model. It impacts behavior in the same way as deadlines do, with some increase or decrease variables. The formula for those with the neuroticism personality trait is given by equation 4.10. On the other hand, the formula for those agents with any of the other three personality traits is given by equation 4.11. The components of the equations are x, which corresponds to each of the possible levels of stress (1, 2 or 3) and $workload\_factor_x$, which corresponds to the factor of the previous version.

$$p_x = prob\_levelx \cdot workload\_factor_x \cdot \frac{1 + 2 \cdot (x - 1)}{2} \tag{4.10}$$

$$p_x = prob\_levelx \cdot workload\_factor_x \cdot \frac{3 + x}{5} \tag{4.11}$$

For example, in case of having the personality trait neuroticism with only 2 deadlines, the probability of having stress level 3 is as shown in expression 4.12.

$$p_3 = prob\_level3 \cdot (1 - \frac{2}{4}) \cdot \frac{1 + 2 \cdot (3 - 1)}{2} = prob\_level3 \cdot 1.25 \tag{4.12}$$

4.5 State Machine Simulation Model

The third simulation model is characterized by the use of a **finite-state machine** [28] to determine the agent's behavior. Several factors are added to this behavior in separate analysis sections to adjust the probability of state transitions, resulting in a more realistic simulation model.

A computing model based on a theoretical machine with one or more states is known as a finite state machine. The machine can only be in one state at a time and has a limited number of states. Each state has a set of transitions, each of which has a change probability associated with it. The transitions lead to a state, which does not necessarily have to be different, since there is a probability of staying in the same state.

The model's development is divided into three main parts:

1. ***Baseline***. In addition to the basic parameters established in the linear regression model, this initial version of the model includes two additional parameters. These two variables define the behavior of the finite state machine's early version, determining

the probability of changing state by increasing or decreasing the stress level. An important characteristic of this model is that once the parameters that provide the most realistic simulation have been found, they are kept in the following versions with those fixed values.

2. ***Workload***. Once the best values for the likelihood of increasing or decreasing state have been determined, two factors corresponding to the workload are determined, which modify the probability of state change, getting it even closer to the real data.

3. ***Personality***. Fixed values have already been specified for the four factors from earlier versions in this last phase. As a result, the new parameters are those related to personality, which will be separated into two factors of increasing and decreasing for neuroticism and two factors for the other three personalities.

A finite state machine determines the agent's behavior, with changing states based on probabilities. As behavioral programming evolves, other factors are included that modify these probabilities in order to achieve more similarity. The agent's behavior is divided into three main phases:

1. ***Baseline***. The two new parameters added in the model are prob_increase and prob_decrease, which are part of the finite state machine's state change conditions that influence the agent's behavior. Each state is a stress level and has three options: two to change states and one to remain in the same state. Each of the probabilities to change or maintain a state is shown in figure 4.3 as P_{xy}.

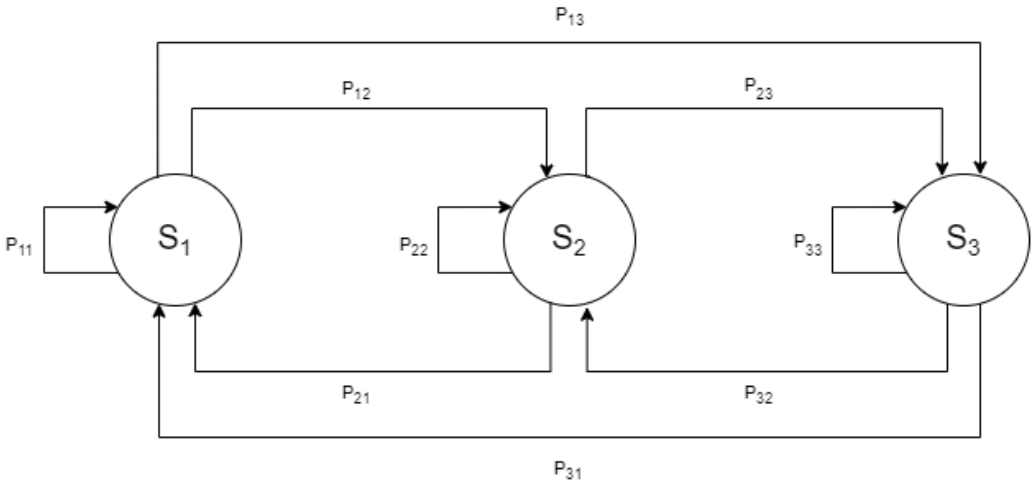

Figure 4.3: Finite state machine

In each run of the model, different values are introduced for the parameters in order to find the best combination, so the probabilities of state change vary accordingly. An additional random variable between 0 and 1 called *change* is defined to perform the transition of state.

The values of the probabilities that define the change of state are those shown in table 4.3. The initial state is determined by the rows, whereas the final state is determined by the columns. Therefore, the probability of moving from state 2 to 3 (P_{23}) is represented by the value of the row with the number 2 and the column with the number 3.

	1	2	3
1	$1 - P_{12} - P_{13}$	$prob\_increase$	$prob\_increase \cdot \frac{1}{3}$
2	$prob\_decrease$	$1 - P_{21} - P_{23}$	$prob\_increase$
3	$prob\_decrease \cdot \frac{1}{3}$	$prob\_decrease$	$1 - P_{31} - P_{32}$

Table 4.3: Probabilities of the finite state machine

2. ***Workload***. This second phase is defined by the addition of two workload parameters that have an influence on the probability of the prior version. The values of prob_increase and prob_decrease are fixed at the best values found in the previous version, and the workload factors are the only variables that vary with each model run. The new increase and decrease probabilities are calculated as shown in equations 4.13 and 4.14.

$$prob\_increase_2 = prob\_increase_1 \cdot workload\_increase \tag{4.13}$$

$$prob\_decrease_2 = prob\_decrease_1 \cdot workload\_decrease \tag{4.14}$$

Variables $prob\_increase_1$ and $prob\_decrease_1$ correspond to the previous version's optimal values, whereas $workload\_increase$ and $workload\_decrease$ represents the overall workload factor for both increasing and decreasing the state, respectively.

If the number of deadlines is zero, *workload_increase* and *workload_decrease* is 1, keeping the previous version's scenario. If there are more deadlines than zero, the equations 4.15 and 4.16 are used to determine both variables.

$$workload\_increase = workload\_inc\_weighted \cdot deadline\_factor + 1 \tag{4.15}$$

$$workload\_decrease = workload\_dec\_weighted \cdot (\frac{1}{deadline\_factor}) \tag{4.16}$$

The parameters that are introduced to the model are *workload_inc_weighted* and *workload_dec_weighted*, which are the ones that are wanted to obtain the optimal values for the third version. On the other hand, $deadline\_factor$ is a factor that depends on the number of deadlines that the agent has that day and is defined as shown in expression 4.17.

$$deadline\_factor = \frac{deadlines + 1}{2} \tag{4.17}$$

3. ***Personality***. The final version of the model has four parameters, two for neuroticism and two for the other personalities, which correspond to the state increase and decrease factors in each case. The new increase and decrease probabilities are calculated as detailed in equations 4.18 and 4.19.

$$prob\_increase_3 = prob\_increase_2 \cdot personality\_factor\_increase \tag{4.18}$$

$$prob\_decrease_3 = prob\_decrease_2 \cdot personality\_factor\_decrease \tag{4.19}$$

The $prob\_increase_2$ and $prob\_decrease_2$ variables are the values obtained from the workload factor version, while the other ones represent the increase and decrease personality factors. Depending on the agent's personality, the variable associated with the increase, $personality\_factor\_increase$, can have two different values. If it is neuroticism then it takes the value $neu\_inc\_factor + 1$, while if it is the general case, for the rest of the personalities, it takes the value $aco\_inc\_factor + 1$.

The same thing happens with the other variable, $personality\_factor\_decrease$, with the difference that the plus one is not added, since it does not increase. Therefore, for neuroticism it takes the value neu_dec_factor, and for the general case, it takes the value aco_dec_factor.

4.6 Validation

4.6.1 Distance Metrics

A distance metric enables for the comparison of a real data point with a simulation point by determining their similarity or distance. Considering that the project's specified time period is two months, there will be a total of 62 real points to compare to the simulation's 62 points. The sum of the 62 point-to-point distances will equal the total distance. Several simulations are run with various parameters and values in order to determine the minimum total distance, which would be the one with the most similarity to the dataset.

In this project, two different distance metrics have been used to provide greater reliability and robustness to the validation of the models: the **Manhattan distance** [29] and the **Euclidean distance** [30].

The **Manhattan distance** between two points x and y in n-dimensional space is the sum of the distances in each dimension. The generalized formula for the Manhattan distance in an n-dimensional space is given by equation 4.20.

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{4.20}$$

In this project, the formula is applied so that the sum of the absolute value of the difference of each pair of stress records is calculated. So, the Manhattan distance to calculate the similarity between the study value and the simulation value is as shown in expression 4.21.

$$d = \sum_{i=1}^{62} |Stress_{study} - Stress_{simulation}| \tag{4.21}$$

The **Euclidean distance** is the smallest separation between two points in a two dimensional space, being called also as Pythagorean distance, by the theorem with the same name. In two-dimensional space, the relationship between the Euclidean distance and the Manhattan distance can be easily observed, the latter being the distance corresponding to the legs of the triangle, while the former is the distance in a straight line.

Figure 4.4: Euclidean and Manhattan distances in two-dimensional space

The analytical technique is comparable to the Manhattan distance in that the only difference is the method of obtaining the value. Therefore, the 62 values' similarity is examined, and the 62 distances obtained are summed, producing a total distance for that simulation. Finally, the minimum overall distance between all of the simulations' distances is determined.

The Euclidean distance between two points x and y in n-dimensional space is the square root of the sum of the distances in each dimension squared. The generalized formula for the Euclidean distance in an n-dimensional space is calculated as shown in equation 4.22.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4.22}$$

The formula is particularized in this project to calculate the square root of the sum of all the consequent subtractions between each pair of stress records, all of them squared. So, the Euclidean distance to calculate the similarity between the study value and the simulation value is given by equation 4.23.

$$d = \sqrt{\sum_{i=1}^{62} (Stress_{study} - Stress_{simulation})^2} \tag{4.23}$$

4.6.2 Linear Regression Simulation Model

The model is executed with each agent having a behavior dependent on their deadlines and their personality. The simulation's results are then gathered, and the stress is grouped by day across the two-month research period. After that, the simulation's daily stress levels are compared to the stress data from the StudentLife study.

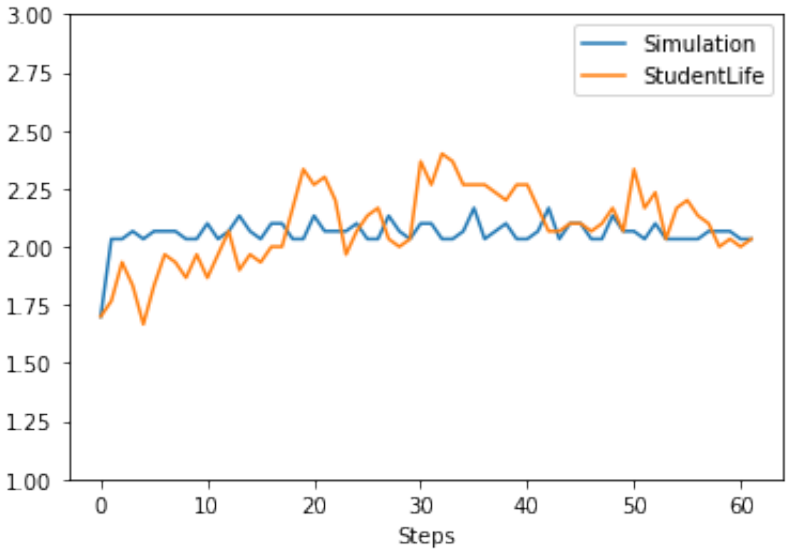

Figure 4.5: Linear regression results

The results show that the simulation remains relatively constant at stress level 2, with small peaks when there is a higher concentration of deadlines. This is mainly due to the data set's poor correlation between stress and deadlines, resulting in a beta value that is too low and does not reach the study's actual values.

4.6.3 Weighted Probability Simulation Model

The second simulation model is characterized by the use of weighted percentages to calculate the agent's stress. To discover the percentages with stress levels that are closest to the StudentLife study data, the distance metrics are employed between data sets.

The analysis begins with grouping iterations that present the same collection of data, with the intention of providing the system with stability and consistency while minimizing randomness.

The distances between each of the 8000 data combinations and the StudentLife study records are then determined. Afterward, all the distances are compared, with the lowest value representing the probability combination that is closest to the research. Since many factors influence how stress is calculated, the results change based on the agent's behavior. Therefore, the results are divided into phases in order to find the most accurate version.

1. ***Baseline***. The results of the first phase show that the three stress levels have a fairly uneven probability distribution. It's also worth noting that, since these are weighted probabilities, the sum of them does not equal the unit, or 100% if translated to percentages.

prob_level1	prob_level2	prob_level3	Manhattan	Euclidean
0.100	0.633	0.172	8.399	1.341

Table 4.4: Baseline version probabilities

This probability distribution has a Manhattan distance of 8.399, the minimum among all runs with this behavior, which represents the set of 62 stress values closest to the dataset.

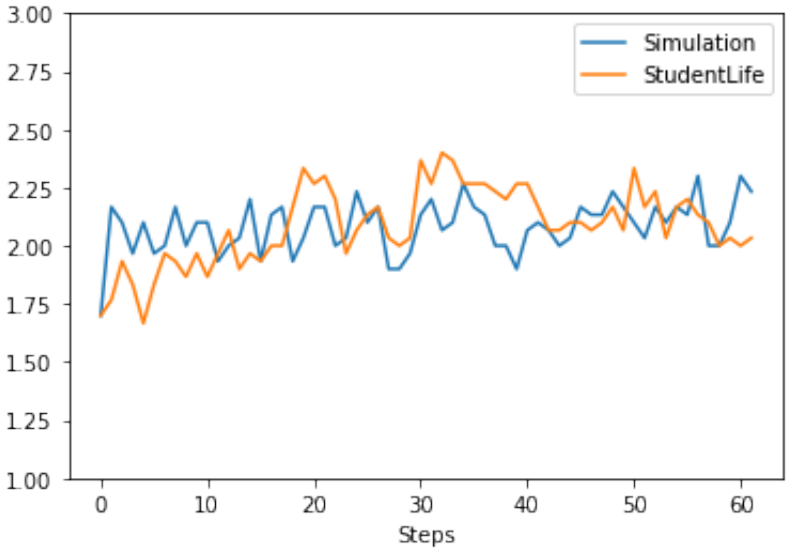

Figure 4.6: Weighted probabilities first version representation

2. ***Workload***. By including deadlines as a factor, the overall distance is reduced slightly, showing that this version improves on the baseline version. However, there is more disparity in the probability distribution, with a higher increase in the probability of stress level 2.

prob_level1	prob_level2	prob_level3	Manhattan	Euclidean
0.055	0.727	0.150	8.266	1.317

Table 4.5: Workload version probabilities

Although the values obtained for each of the distances are different, the set of simulation values closest to the dataset is the same in both cases.

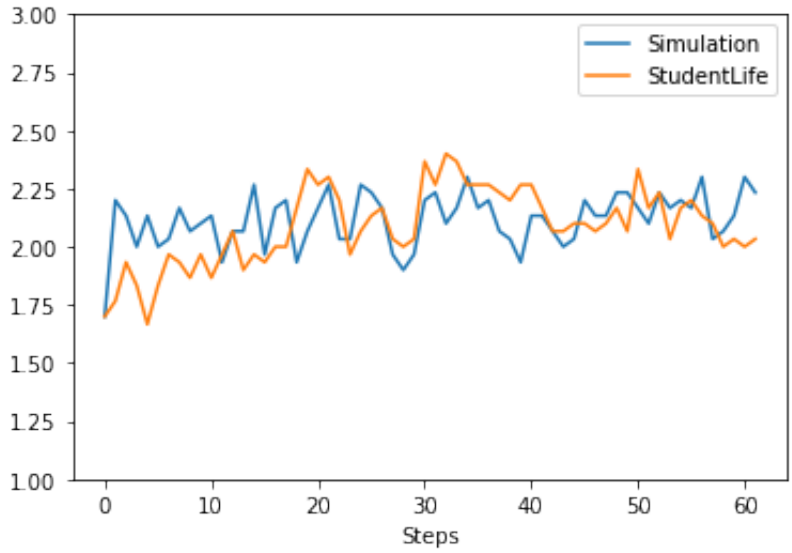

Figure 4.7: Weighted probabilities second version representation

3. ***Personality***. Stress is stabilized by adding personality as a factor in the cases of openness, conscientiousness and agreeableness; they fluctuate around a level of 2 without as many peaks as previous versions, resulting in a substantially smaller distance, even below 8.0. In the case of neuroticism, the chances of reaching stress level 1 are almost zero, with stress levels always ranging between 2 and 3 depending on the number of deadlines.

The probabilities of this last version are shown in table 4.6.

prob_level1	prob_level2	prob_level3	Manhattan	Euclidean
0.050	0.900	0.100	7.666	1.204

Table 4.6: Personality version probabilities

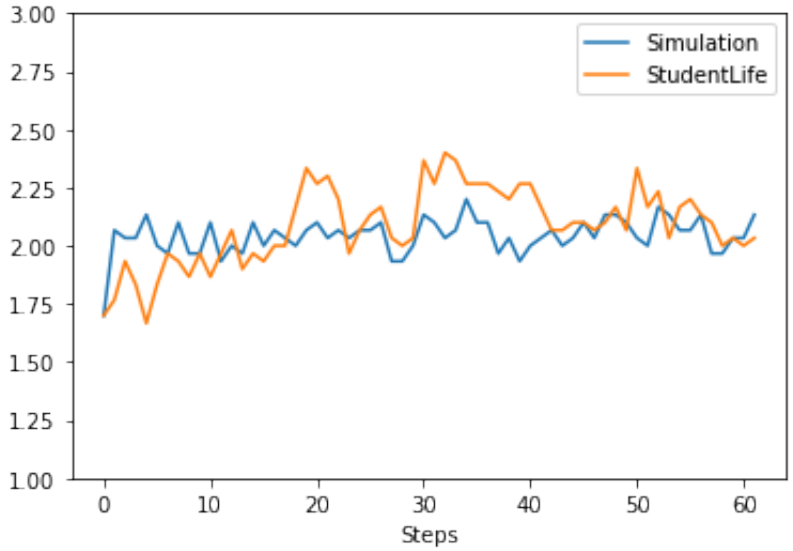

Figure 4.8: Weighted probabilities final version representation

The second version gets extremely close stress levels in deadline concentrations but loses similarity throughout the simulation, resulting in a considerably larger overall distance value that does not compensate for the proximity in the peaks.

Although the study's stress data peaks do not correspond with the simulation's values in the third phase, there is significantly less distance and hence higher overall similarity. Personality factors stabilize the simulation's stress values, affecting the two big deadline concentrations but improving the rest of the simulation.

4.6.4 State machine Simulation Model

The distance metrics have been used to analyse the results obtained in each phase of the process of developing the agent's behaviour. The version with the shortest distance will be the most similar to the study's results. Therefore, the results are shown divided by phases:

1. ***Baseline***. *prob_increase* and *prob_decrease* are the variables for which a parameter sweep is used to discover the best values for a smaller Euclidean distance. The model is run 6250 times, taking 25 values for each variable in the range of 0 to 0.7 and conducting 10 iterations for each combination of values.

   The minimum distance is determined by the pair of values shown in table 4.7.

prob_increase	prob_decrease	Manhattan	Euclidean
0.1166	0.0875	7.033	1.097

Table 4.7: Increase and decrease probabilities

   The minimum Euclidean distance among all combinations is 1.097, which represents the closest set of 62 simulation values to the dataset in this version.

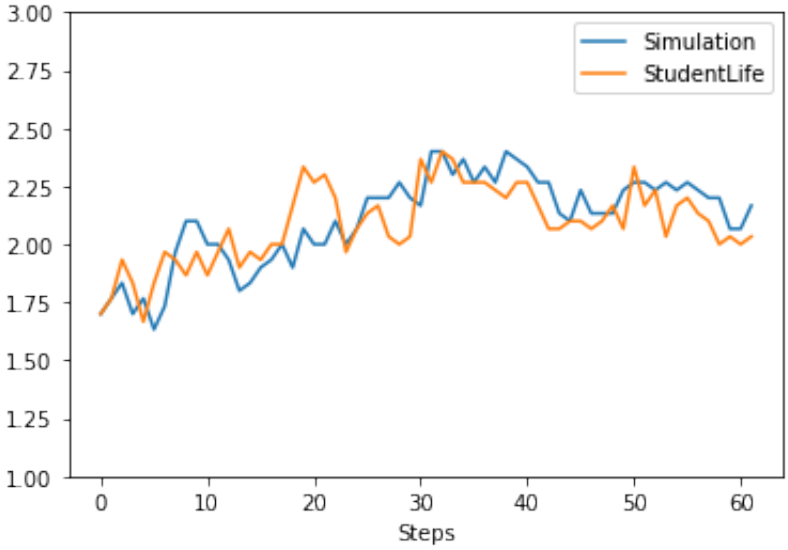

Figure 4.9: State machine first version representation

   The values obtained for the probabilities of increasing or decreasing the state remain fixed for the next version with values of 0.1166 and 0.0875, respectively.

2. ***Workload***. *workload_inc_weighted* and *workload_dec_weighted* are this second version's variables. A parameter sweep is used to find the value of the Euclidean distance that improves the prior version's situation.

   The model is run 9,000 times, with 30 values for each variable and 10 iterations for each value combination. The parameter ranges for increasing and decreasing status are 0.0 to 0.5 for increasing and 0.7 to 1.2 for decreasing. This disparity in ranges is

due to the fact that the increase factor's formula includes a plus one, which provides the probability of increase a value equal to or greater than the first version, whilst the decrease factor's formula does not.

The pair of values shown in table 4.8 define the minimum distance.

workload_inc_weighted	workload_dec_weighted	Manhattan	Euclidean
0.1421	1.0421	6.066	0.948

Table 4.8: Workload increase and decrease factors values

In this version the value of both distances are reduced compared to the previous phase. Therefore, this version considerably improves the proposed model, obtaining more similar results.

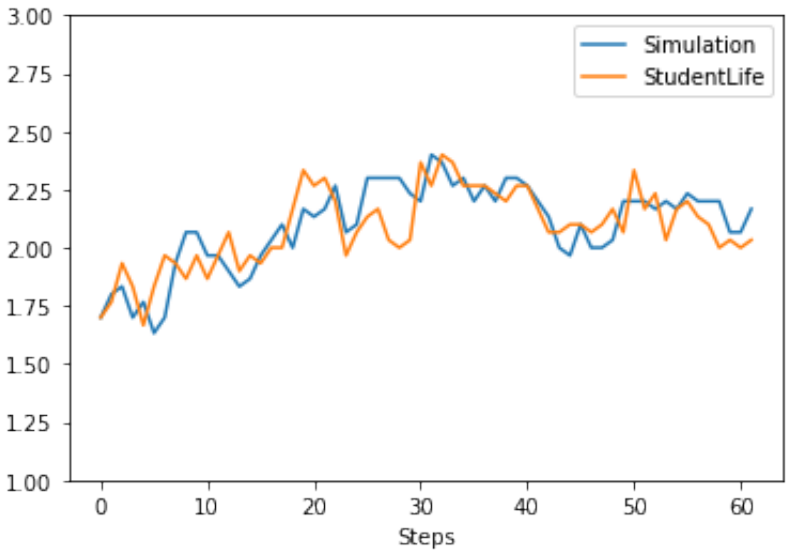

Figure 4.10: State machine second version representation

Once the best values have been found, they are fixed for the following phase, establishing the baseline.

3. ***Personality***. The final version of the model is distinguished by the inclusion of four personality factors. Since neuroticism stands out as the most stressed personality, while the rest of the them tend to remain reasonably stable, the factors have been separated into two categories. For the case of neuroticism, there are two increase and decrease factors, while for the other three personalities: openness, conscientiousness and agreeableness, correspond the other two.

Since running the parameter sweep with four variables would result in an impractical number of model runs (about 8 million), due to the number of combinations, the procedure for obtaining the best values of these factors must be divided.

First, the execution of the model is carried out having as variables the two factors of neuroticism, neu_inc_factor and neu_dec_factor. The other two factors that correspond to the other personalities remain constant, not interfering in the development of this analysis. Therefore, the model is run in the same way as it has been done in the two previous versions, performing the parameters sweep for two variables.

Using 30 values for each variable and 10 iterations for each value combination, the simulation was performed 9,000 times. The increasing and decreasing state parameter ranges are 0.0 to 0.6 for increasing and 0.7 to 1.2 for decreasing.

The set of values that define the minimum distance for this case are shown in table 4.9.

neu_inc_factor	neu_dec_factor	Manhattan	Euclidean
0.3263	0.9105	6.029	0.942

Table 4.9: Neuroticism increase and decrease factors values

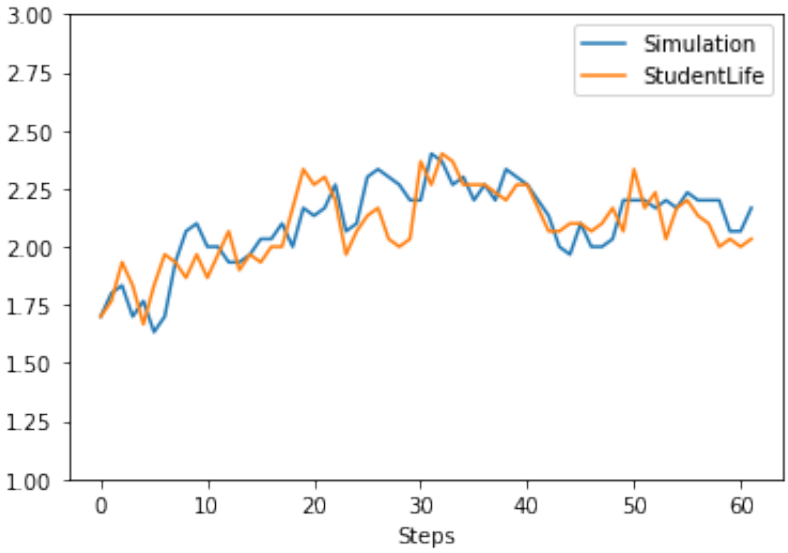

Figure 4.11: State machine neuroticism version representation

It has been possible to improve the results obtained in the workload version. It is, however, only slightly reduced from 0.948 to 0.942 for the Euclidean distance and

from 6.066 to 6.029 for the Manhattan distance, a minimal change compared to the reduction obtained from the first to the second version.

Once the optimal factors for neuroticism have been obtained, these values are fixed in the next part of the model run, performing the parameter sweep with the other two factors: aco_inc_factor and aco_dec_factor, which correspond to the other three personalities: agreeableness, conscientiousness and openness.

The simulation was run 9000 times with 30 values for each parameter and 10 iterations for each combination of values, as in the previous simulation. The ranges of the parameters are from 0.0 to 0.3 for the increase one and from 0.8 to 1.2 for the decrease one.

The minimal distance is defined by the set of values shown in table 4.10.

aco_inc_factor	aco_dec_factor	Manhattan	Euclidean
0.0157	0.9736	5.733	0.899

Table 4.10: Aco increase and decrease factors values

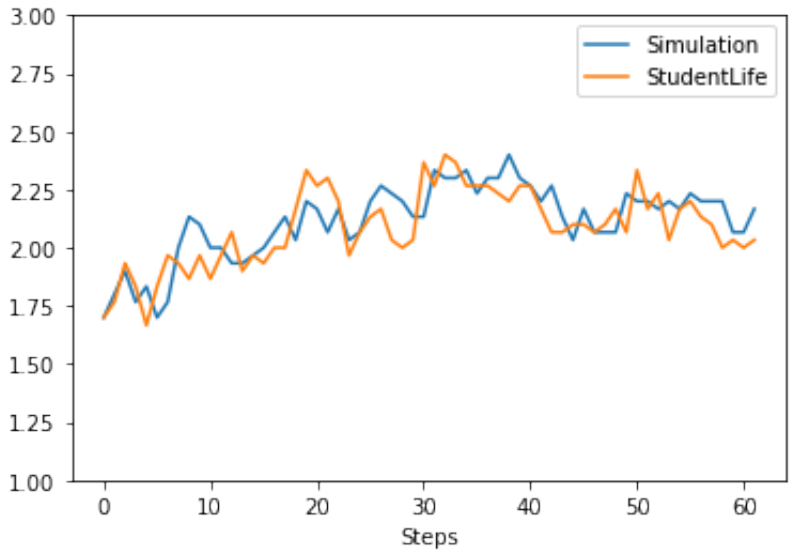

Figure 4.12: State machine final version representation

Since the increase factor is very close to zero and the decrease factor is quite close to one, the values obtained slightly modify the probability of increasing and decreasing state. However, since most students have some of the personalities contained in these factors, this latest version achieves a significantly smaller distance in both cases.

4.6.5 Comparison of the three simulation models

In the linear regression simulation model no distance metric was used since there were no different versions of the model to compare because stress is calculated in a fixed way. Therefore, both distance metrics are used in the results obtained in figure 4.5, comparing the 62 values of the simulation with the 62 values of the dataset. The results for the Euclidean distance are shown in table 4.11, while the results for the Manhattan distance are shown in table 4.12.

	Linear regression	Weighted probability	State machine
Baseline	1.225	1.341	1.097
Workload	-	1.317	0.948
Personality	-	1.204	0.899

Table 4.11: Euclidean distance results

	Linear regression	Weighted probability	State machine
Baseline	7.834	8.399	7.033
Workload	-	8.266	6.066
Personality	-	7.666	5.733

Table 4.12: Manhattan distance results

In both cases, the state machine simulation model is the simulation system with the shortest distance and that most closely resembles the Student Life research.

CHAPTER 5

Case study

5.1 Model parameter values

The State Machine simulation model is the simulation system chosen for the case study out of the three proposed since it is the most realistic. In the simulations performed in this chapter, different values are entered for the same input parameters.

The factors that impact the agent's behavior, such as aco_inc_factor and aco_dec_factor, are the first parameters to have their values changed. In this case, it is not necessary to enter a range of possible values to run a parameter sweep since the optimal values for these parameters were determined in Section 4.6.

The StudentLife study dataset is no longer necessary, since it was used only for two reasons: to obtain the values of the factors that influence the agent's behavior and to validate the simulation model. Therefore, the CSV file with the study data is no longer entered in the corresponding parameter. Instead, a different CSV file must be imported with the information of the users that are part of the simulation.

The value of the parameter related to the seed remains unchanged, as it was solely used to predict the system's randomness.

The CSV file must include deadline distributions as well as the personality of each model user. The number of agents input in the corresponding parameter must, of course, equal the number of users in the file.

The start date of the simulation is the last parameter to be mentioned. This parameter allows the CSV file to have a data recording start date that is different from the simulation start date. The simulation would begin at the first record in the file if the parameter was not included in the model.

In case of not having the CSV required to run the model, a configuration file has been developed that allows the user to create a CSV file from some basic inputs. The following parameters must be entered:

- ***Number of agents***. This parameter is necessary to determine the total number of users that will be in the file.

- ***Deadlines***. It requires the entry of a list of days with high concentrations of deadlines. The list must include the number of days from the start of the simulation until the workload peaks are located. If the value "[10, 20]" is entered, for example, it indicates that the deadline concentrations are established at 10 and 20 days after the simulation begins. The distribution of deadlines is random and different for each user, with daily amounts ranging from 0 to 3. Method *get_deadlines*() calculates the list of deadlines for each of the users throughout the simulation.

- ***Personality probabilities***. It requires the entry of a list with the probabilities of each of the Big Five personalities. The *get_personality*() method determines the personality of each of the users based on the probabilities.

  Since the probability corresponding to extraversion has not been analysed in this project due to lack of data, it has not been included as an option among the probabilities. However, since people with this personality actually exist, something must be done to solve the problem. The personality trait that stood out in stress in the previous chapter's analysis was neuroticism, while the other three remained grouped under the same factor. Therefore, extraversion can be included in the model within the agreeableness, conscientiousness, and openness group.

5.2 Scenarios

The scenarios consist of the execution of the model changing the content of the CSV file. Two fundamental aspects of the agents vary in each scenario of this project: deadlines and the probability of each personality. Regarding the number of agents, it is increased with respect to the previous version to a sufficiently significant number. This quantity remains constant across all scenarios at 80 agents.

Based on the distribution of deadlines, the scenarios are divided into three categories:

- ***Low workload***. This first category is distinguished by the presence of only one deadline concentration date. The simulation reaches its peak of deadlines 30 days after it begins.

- ***Medium workload***. The existence of three concentration of deadlines distinguishes the second category. At 10, 30, and 50 days, the simulation reaches its three deadline peaks.

- ***High workload***. The last group is distinguished by the presence of five concentration of deadlines. The deadlines peaks are evenly distributed across the two months of simulation. At 10, 20, 30, 40, and 50 days, the simulation reach its peaks.

Within each category of deadlines, the scenarios are divided according to the distribution of probabilities for each of the personalities. The probability distributions are as follows:

- ***Balanced personalities***. This first group is determined by the premise that all personalities have the same probability. As a result, the personality distribution is uniform, with 20 percent for each personality.

- ***Neuroticism majority***. The second group is defined by the fact that neuroticism is the most common personality. The probability of neuroticism is 80 percent, while the probability of any of the other four personalities is 5 percent.

- ***Neuroticism minority***. The fact that neuroticism is a minority personality characterizes this last category. The probability of neuroticism is 4 percent, whereas the probability of any of the other four personalities is 24 percent.

5.3 Results

The results for each scenario are shown in this section. Considering that there are three classification groups based on deadlines, each with three subcategories with different probability distributions, a total of nine possible scenarios are defined.

The results from the three model runs of each category are presented together in the same graph to make data analysis simpler. The y-axis range has also been shifted from [1.0, 3.0] to [1.8, 2.8]. This is due to the fact that there are no stress levels below 1.8 or over 2.8. This adjustment allows us to examine the differences between each scenario in more depth while also preventing data from the three cases from overlapping.

- ***Low workload***. This category shows the result obtained from model runs with only one deadline peak. Given that each category is divided according to the distribution of probabilities, three model executions are carried out in this section.

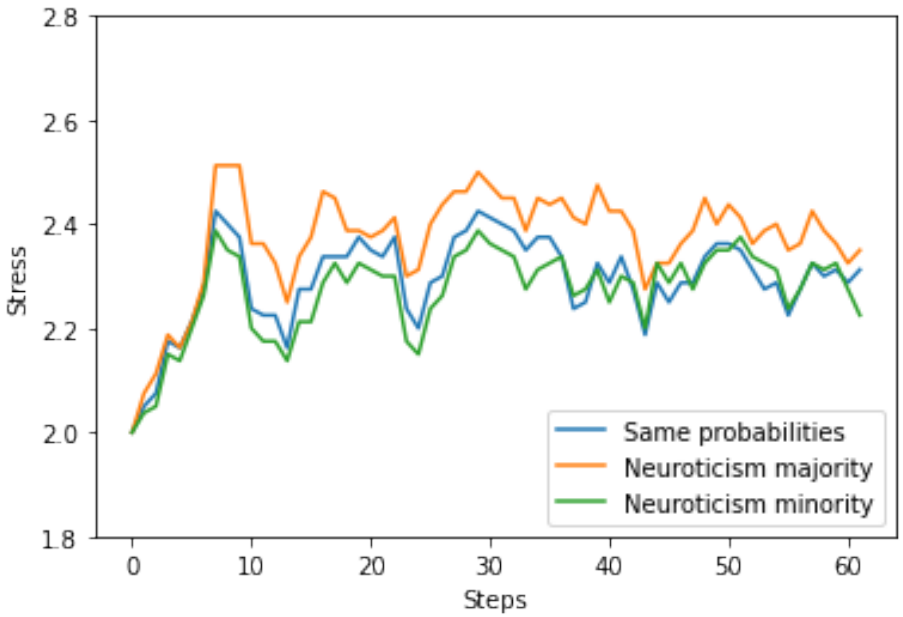

Figure 5.1: Low workload representation

Figure 5.1 shows that despite having a concentration of deadlines 30 days after the start of the simulation, there is no significant stress peak at that time. In fact, the peak is similar to the highest level of stress experienced on a regular day, and it blends in with the others.

This behavior is common of personalities agreeableness, conscientiousness, openness, and extraversion, that are not overly dependent on deadlines. As can be seen in

scenario "Neuroticism minority", the average stress of these personalities differs by approximately no more than a 6%.

The resemblance between scenarios "Neuroticism minority" and "Same probabilities" is worth noticing. The results are nearly identical in both cases, with the exception that the scenario "Same probabilities" had slightly greater stress levels. This is because neuroticism was included in this situation.

Given that there are no concentrations of deadlines for the duration of the simulation, students with neuroticism experience stress from everyday tasks as if they were a concentration of deadlines. Since there is not a great workload because it is usual to have 0 or 1 deadline, stress increases sharply when a task appears and gets down when it is completed.

- ***Medium workload***. The results of the model runs with three deadline peaks are shown in this section. The group is divided according to the probability distribution, as in the prior case, generating three model executions.

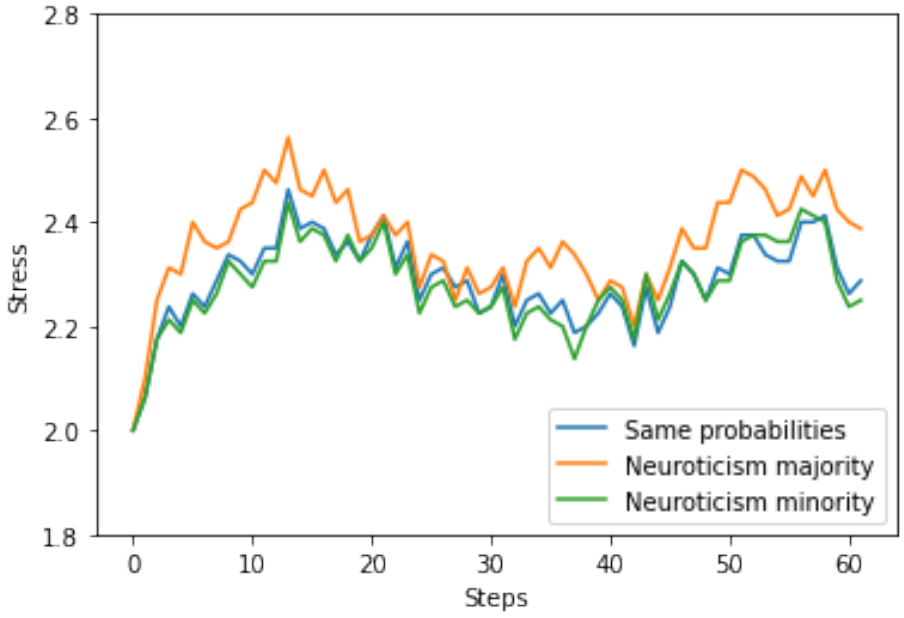

Figure 5.2: Medium workload representation

The similarity between scenarios "Neuroticism minority" and "Same probabilities" increases in this group, significantly reducing the distance seen in the prior category. The average stress stress in these two situations is 2.3, with a tenth increase when deadlines are at their highest.

The stress levels in scenario "Neuroticism majority" are significantly lower than in the

case of a single deadline peak, and they are also more similar to the other scenarios, which is noteworthy. When concentrations of deadlines occur, this scenario increases the average stress level by around a 3%. This increase is most noticeable at 10, 30, and 50 days, when the line that illustrates this scenario is more different from the other two. However, when this peak of deadlines is exceeded, the stress level stabilizes and returns to the levels seen in the other scenarios. From 20 to 30 days and 40 to 50 days are the ranges with similar values amongst scenarios.

- ***High workload***. In the last category, the results of model runs with five deadline peaks are represented. The deadline peaks are separated by ten days. The group is divided into three subcategories based on the probability distribution, resulting in three model executions.

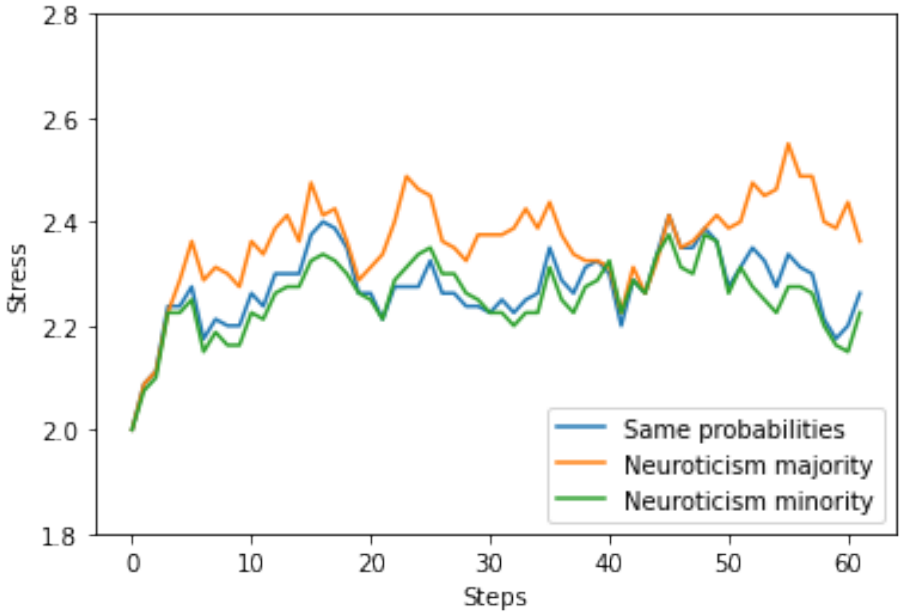

Figure 5.3: High workload representation

The resemblance between scenarios "Neuroticism minority" and "Same probabilities" remains, but to a lesser extent than in the previous case. When the probabilities are equal, peaks appear for the average level of stress due to the high values obtained for neuroticism.

Since stress peaks are more common, the average level of stress in the scenario "Neuroticism majority" is higher. Also because deadlines are so close together, there is no time for the agents to de-stress, resulting in significant differences in stress levels at the peaks.

CHAPTER 6

Conclusions and future work

This chapter describes the conclusions extracted from this project and presents the thoughts about future work and potential improvements that could be made.

6.1 Conclusions

In this project, an agent-based social simulation model has been developed to analyse the influence of personality on stress management. The system allows users to determine the best deadline distribution that minimizes stress based on their personality.

First, a research process was conducted by reading articles and publications on the subjects of this project: personality and stress. The required context was obtained in order to perform the subsequent data analysis as well as the design and interpretation of the simulation models.

The results of the data analysis for the StudentLife study matched the conclusions of the majority of research papers. Neuroticism is the personality that is most susceptible to become stressed, whereas agreeableness, conscientiousness, openness, and extraversion all respond normally to stressful events.

Based on this analysis, three different simulation models were developed: the linear regression simulation model, the weighted probability simulation model and the state machine simulation model.

The linear regression simulation model relied entirely on correlations between data from the study dataset to create formulas that defined the agent's behavior. Therefore, neuroticism's high stress level and the other personalities' usual stress response matched the conclusions of the previous sections.

On the other hand, the weighted probability simulation model calculates stress using a probability approach. The agent's stress level is determined from the probabilities associated with each of the possible stress values. These probabilities are modified depending on the number of deadlines and the personality of the agent.

The state machine simulation model is distinguished by the use of a finite state machine to determine the agent stress. In different analysis phases, different parameters are introduced to the agent's behavior to change the probability of state transitions.

Once the simulation models have been developed, the models were validated with the data from StudentLife study. The distance between the values obtained from the simulation and the real values is computed. Two different distance metrics, Euclidean and Manhattan, were used to increase the dependability of the results. The simulation model with the smallest distance was the state machine simulation model.

In the case study, the state machine simulation model is used since it has the shortest distance. To analyse the behavior of the agents, many simulations with various input values are run. Several conclusions can be made based on the findings obtained from the execution of the nine scenarios:

- ***Deadlines distribution***: The distance between deadlines should not be too big or too small. On the one hand, instability in stress levels is caused by a lack of workload. When there are almost no deadlines and a task appears, stress levels increase drastically. On the other hand, if the workload is overwhelming, it leads to a constant state of tension, which worsen in neurotics causing extremely high levels of stress. Therefore, a moderate amount of deadlines should be chosen to obtain the lowest levels.

- ***Personality probabilities***: The personalities' probability distribution should be uniform. The results indicate a strong resemblance between two probability distributions: minority of neurotics and same probabilities. As excluding neurotics does not

significantly reduce the group's mean stress, it is preferable to include them rather than putting neurotics in a separate group, as in the last probability distribution (majority of neuroticism).

- ***Neuroticism***: In order to reduce their stress, neurotics should be placed in work groups with other personalities and a balanced distribution of deadlines. In most circumstances, a group made primarily of neurotics produces inferior outcomes. A neurotic person has the best results with three deadline peaks in two months, nearly equaling the other personalities.

This model helps in the identification of the factors that cause stress and facilitates the definition and validation of stress regulation strategies. This management techniques are based on the relocation of work groups and the redistribution of tasks. A group of people that manage their stress well and have low levels of it improves both personally and professionally [31].

6.2 Future work

This section covers the potential future upgrades for this project:

- ***Increase simulation time***: The simulation model runs for two months by default and cannot be changed unless the model files are manually modified. The improvement consists of the addition of a new parameter that controls the amount of steps that the system will take.
- ***Design of a graphical interface***: A graphical user interface would allow for a more thorough and clear representation of the results. A web service could be created to allow the model to be visualized and run graphically. The process could be performed by entering parameters into a form and then clicking the execute button to see the results.
- ***Add new factors***: Only the agent's personality and the quantity of deadlines affect his behavior. However, new factors could be added, in order to obtain a much more accurate value of stress. There are external factors that are not taken into account in this model and that could be incorporated, such as social changes, unforeseen events, environment or even important changes in the person's life.

APPENDIX A

Impact of this project

This appendix considers the qualitative effects that the implementation of this project might have. It focuses on the social, economic, environmental, and ethical impacts.

A.1 Social impact

The improvement in interpersonal relationships determines the social impact. The consequences of stress affects not just the individual who is stressed, but also for those around them.

Stress can have a variety of consequences, such as a negative impact on the control of their social relationships. It can also result in deterioration of family connections or even global consequences for the family unit. Also, stress can lead to a reduction or loss of commitment to society's standards, the breakdown of friendships, or a lack of interest in social activities.

This research helps with all of these issues by identifying the ideal stress-reduction distribution, resulting in calm and rational conduct. The model is especially recommended for neurotics, as it will significantly improve their quality of life and personal relationships.

A.2 Economic impact

This project's economic impact is focused on the business environment. A positive workplace relationship is vital in a company since greater results are produced in a pleasant and calm environment. People with high levels of stress, which leads them to be irritable and restless, might disrupt this productive work atmosphere.

Another concern with stressed employees is their poor performance, which can be caused by any of the numerous stress-related symptoms, such as concentration issues, sleeplessness, fatigue, or depression.

This project helps to reduce stress levels in the workgroup, resulting in better outcomes for the organization and hence increased income.

A.3 Environmental impact

Since this is a simulation software, the project has a low environmental impact. Some elements, however, have a minimal direct or indirect impact on the environment. A higher amount of data will require more computational resources capable of supporting the model's execution. Therefore, a certain quantity of fossil and chemical fuels is required to create this hardware.

Furthermore, energy consumption is required for data processing and proper cooling of the hardware during both its creation and everyday use.

A.4 Ethical impact

The ethical impact of this project is related to the collection and protection of data. The project's ethical implications are determined by the correct and responsible use of data while respecting the users' privacy at all times.

Since personal or confidential data might be submitted into the project, its usage and exploitation must be done with consent.

APPENDIX B

Economic budget

This appendix details an adequate budget to bring about the project. The costs of physical and human resources, as well as licenses and taxes, are covered.

B.1 Physical resources

This section describes the computational resources that were used in the project's development. Since the model was run with a small number of agents over a short period of time (less than two months), the computer components used were standard PC components.

However, if the model's execution is to be carried out on a larger scale and with more data, the machine used for the execution must be improved. The necessary prerequisites for a system to be able to run the model without problems, are the following:

- ***CPU***: 10th Gen Intel Core i5 or i7
- ***RAM***: 16 GB DDR4 and 2,666 MHz frequency
- ***Hard Disk***: 512 GB SSD

A computer with this specifications can cost around **800€**.

B.2 Human resources

The project's human resource expenses are detailed in this section. These costs correspond only to an engineer's salary for research work.

Two factors are taken into consideration when calculating the costs: the hours spent analysing the data and developing the simulation system, and the average salary of an internship engineer.

Considering that the project lasted six months and that approximately 21 days of each month were dedicated to its development, the total number of days spent on the analysis and development was 126. However, since the average time spent on the project was 4 hours each day, the total number of hours spent does not equal the total number of hours for all days. Therefore, the total time spent on this project was 504 hours.

An internship engineer's monthly salary for 20 hours per week is considered to be around 500 euros. As a result, the estimated overall cost of developing the simulation model is around **3,000€**.

B.3 Software and licenses

All the enabling technologies used for the development of this project is open source software, which indicates there are no license fees to pay. Also, the dataset used for data analysis is also public, so there is no need to pay any rights.

Therefore the cost for software and licenses is **0€**.

B.4 Taxes

The simulation system can be sold to another company after the model development is completed. In that scenario, certain taxes derived from the software's sale must be taken into account.

This simulation model is classified as a computer software, and its sale is subject to a tax. The tax rate applicable to software is **21%**, according to article 8 of Law 37/1992 of the Tax Agency's regulations [32].

Bibliography

[1] Adam Felman. Stress: Why does it happen and how can we manage it? https://www.medicalnewstoday.com/articles/145855, 03 2020.

[2] Aegon and General Psychology Council of Spain. The iv health and lifestyle study. https://fr.zone-secure.net/149562/1404875/, 2021.

[3] Irene Houtman. Work-related stress. https://www.eurofound.europa.eu/publications/article/2005/work-related-stress, 02 2005.

[4] María Angeles del Hoyo Delgado. Occupational stress. https://dialnet.unirioja.es/servlet/libro?codigo=229576, 1997.

[5] Belkis A. Águila María C. Castillo Roxana M. de la Guardia Zaida N. Achon. Academic stress. https://dialnet.unirioja.es/servlet/articulo?codigo=5023824, 2015.

[6] Mercedes Jiménez Benítez. Mechanisms of relationship between personality and health-disease processes. https://dialnet.unirioja.es/servlet/articulo?codigo=5280335, 2015.

[7] Julie A. Penley Joe Tomaka. Associations among the big five, emotional responses, and coping with acute stress. https://www.sciencedirect.com/science/article/abs/pii/S0191886901000873, 05 2002.

[8] Kendra Cherry. Trait theories in psychology. https://www.verywellmind.com/trait-theory-of-personality-2795955, 02 2022.

[9] Kendra Cherry. Cattell's 16 personality factors. https://www.verywellmind.com/cattells-16-personality-factors-2795977, 11 2019.

[10] Dr. Edwin van Thiel. Big five personality traits. https://www.123test.com/big-five-personality-theory/, 04 2022.

[11] Mari Ervasti Johanna Kallio Ilmari Määttänen Jani Mäntyjärvi and Markus Jokela. Influence of personality and differences in stress processing among finnish students on interest to use a mobile stress management app. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6707572/, 05 2019.

[12] Kate A. Leger Susan T. Charles Nicholas A. Turiano and David M. Almeida. Personality and stressor-related affect. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4956603/, 01 2016.

[13] Tamera R.Schneider. The role of neuroticism on psychological and physiological stress responses. https://www.sciencedirect.com/science/article/abs/pii/S0022103104000435, 11 2004.

[14] Adam Bibbey and Douglas Carroll. Personality and physiological reactions to acute psychological stress. https://www.sciencedirect.com/science/article/pii/S0167876012006472, 10 2013.

[15] Oswald L. Zandi P. Nestadt G. Relationship between cortisol responses to stress and personality. https://www.nature.com/articles/1301012, 2006. (Accessed on 04/13/2022).

[16] Gloria García and Mateu Servera. Prosocial personality traits and adaptation to stress. https://www.sbp-journal.com/index.php/sbp/article/view/2267, 10 2011.

[17] Statistics Times. Top computer languages 2021. https://statisticstimes.com/tech/top-computer-languages.php, 12 2021.

[18] Tiobe. Tiobe index for april 2022. https://www.tiobe.com/tiobe-index/, 04 2022.

[19] Sebastian Raschka. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. https://www.mdpi.com/2078-2489/11/4/193, 2020.

[20] Andrew Crooks Nick Malleson Ed Manley A.J Heppenstall. Agent-based modelling and geographical information systems: A practical primer. https://www.researchgate.net/publication/330886643_Agent-based_Modelling_and_Geographical_Information_Systems_A_Practical_Primer, 01 2019.

[21] Brian E. Granger and Fernando Perez. Jupyter: Thinking and storytelling with code and data. https://ieeexplore.ieee.org/document/9387490, 03 2021.

[22] John D. Hunter. Matplotlib: A 2d graphics environment. https://ieeexplore.ieee.org/document/4160265, 05 2007.

[23] Wes McKinney. Pandas: a foundational python library for data analysis and statistics. https://www.semanticscholar.org/paper/pandas%3A-a-Foundational-Python-Library-for-Data-and-McKinney/1a62eb61b2663f8135347171e30cb9dc0a8931b5, 2011.

[24] Wang Rui Fanglin Chen Zhenyu Chen Tianxing Li Gabriella Harari Stefanie Tignor Xia Zhou Dror Ben-Zeev and Andrew T. Campbell. Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. https://studentlife.cs.dartmouth.edu/, 2014.

[25] Oliver P. John and Sanjay Srivastava. The big-five trait taxonomy: History, measurement, and theoretical perspectives. https://pages.uoregon.edu/sanjay/pubs/bigfive.pdf, 03 1999.

[26] Malawi Med J. A guide to appropriate use of correlation coefficient in medical research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/, 09 2012.

[27] Lorraine Li. Introduction to linear regression in python. https://towardsdatascience.com/introduction-to-linear-regression-in-python-c12a072bedf0, 10 2018.

[28] Matyas Lancelot Bors. What is a finite state machine? https://medium.com/@mlbors/what-is-a-finite-state-machine-6d8dec727e2c, 03 2018.

[29] Susan Craw. Manhattan distance. encyclopedia of machine learning. https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8_506, 2011.

[30] Srik Gorthy. Euclidean distance or pythagorean distance. https://medium.com/mlearning-ai/euclidean-distance-8fae145ef5f3, 10 2021.

[31] Ehijiele Ekienabor. Impact of job stress on employees' productivity and commitment. https://www.researchgate.net/publication/334559841, 07 2019.

[32] Law 37/1992 of the spanish tax agency's regulations. https://www.boe.es/buscar/act.php?id=BOE-A-1992-28740, 04 2022.