MEGASKILLS

"Developing Soft Skills While Gaming"

D4.1 MODELLING OF SOFT SKILLS METHODOLOGY

Funded by the European Union

Deliverable No.	D4.1	Work Package No.	WP4	Task/s No.	Task 4.3
Project	MEGASKILLS				
Work Package Title	WP4 – Intelligent soft skills assessment				
Linked Task/s Title	T4.3 Development of the intelligent platform prototype				
Status	Final	(Draft/Draft Final/Final)			
Dissemination level	PU	(PU-Public, PP, RE-Restricted, CO-Confidential)			
Due date deliverable	30/04/2024	Submission date			
Deliverable version	1.0				

Document Contributors

Deliverable responsible			
Contributors	Organization	Reviewers	Organization
Juan Bartolome	TECNALIA	Iñaki Aliende	CEGOS
Idoya del Rio	TECNALIA	Jaroslava Kubátová	UP
Ibai Laña	TECNALIA		
Carmen Alonso	TECNALIA		
Leire Bastida	TECNALIA		

Document History

Version	Date	Comment
0.1	05/04/2024	Initial TOC and platform description
0.2	16/04/2024	AI modelling completed
0.3	22/04/2024	Ready for internal review
0.4	26/04/2024	Ready for final quality review
1.0	29/04/2024	Final version

Disclaimer: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Table of contents

Executive Summary **7**

1. Introduction **8**

2. Modelling data **9**

2.1. Data generated in video games 9

2.1.1 Integration with STEAM in MEGASKILLS 10

2.1.2 Data model 10

2.2. Standardized tests data 14

2.2.1 Sociodemographic data 16

2.2.2 PRE-TEST and POST-TEST 16

2.2.3 Data model 20

3. Artificial intelligence approaches for soft skills analysis **27**

3.1. AI and soft skills development 27

3.2. AI and soft skills development using video games 29

4. AI design approach for MEGASKILLS **30**

4.1. Data analysis 30

4.1.1 Pre-test and Post-test soft skills 31

4.1.2 Gaming data 33

4.1.3 Previous gaming experience 34

4.1.4 Socio-demographic data 35

4.2. AI experiment design & methodology 37

5. Conclusions **39**

6. References **40**

List of Figures

Figure 1 – Methodology followed in the modelling of soft skills in gaming........ 8
Figure 2 – Entity Relationship Diagram: Steam Data Model 14
Figure 3 – Data collection process for standard tests 15
Figure 4 – Sociodemographic information questionnaire........ 16
Figure 5 – Test of Critical Thinking 17
Figure 6 – Critical Thinking Assessment – Scale Short Form 17
Figure 7 – Reflective Thinking Skill Scale for Problem Solving 18
Figure 8 – I-ADAPT-M Short Version 18
Figure 9 – Time Management Questionnaire........ 19
Figure 10 – CAMBIOS mini-game........ 19
Figure 11 – User history........ 20
Figure 12 – Entity Relationship Diagram: Standard Test Data Model 26
Figure 13 – Answers for pre- and post-tests 31
Figure 14 – Distribution of answers before and after for Critical Thinking Test Part 1........ 32
Figure 15 – Distribution of answers before and after for Critical Thinking Test Part 2........ 32
Figure 16 – Distribution of answers before and after for Flexibility /Adaptability Test 32
Figure 17 – Distribution of answers before and after for Time Management Test 32
Figure 18 – Distribution of answers before and after for Problem Solving Test........ 33
Figure 19 – Time played during the last 2 weeks of the experiment by users who played these two weeks 33
Figure 20 – Number of achievements obtained by each user during test phase (orange) and in total (blue)34
Figure 21 – Distribution of achievements obtained by users during test phase (orange) and in total (blue) 35
Figure 22 – Time played during the last 2 weeks and in total for the experiment games by users who played these two weeks........ 35
Figure 23 – Type of game most played by of users 36
Figure 24 – Average time playing games of users 36
Figure 25 - Steps carried out in MEGASKILLS for the AI development (Generic overview) 37
Figure 26 - Example of possible players' behaviour from the dataset using a clustering technique........ 38

List of Tables

Table 1 – usersprofile table details........ 10
Table 2 – achievements table details 11
Table 3 – achievementspercentagespergame table details........ 12
Table 4 – catachievements table details 12
Table 5 – catstats table details 12
Table 6 – friendlist table details 12
Table 7 – ownedgames table details 13
Table 8 – stats table details 13
Table 9 – timeplayed table details 13
Table 10 – users table details 20
Table 11 – testresultssd table details 21

Table 12 – testresultsct table details 22
Table 13 – testresultscta table details 22
Table 14 – testresultssp table details 23
Table 15 – testresultsiadapt table details 24
Table 16 – testresultstm table details 24
Table 17 – cambiostoken table details 25
Table 18 – cambiosanswers table details 25

List of Abbreviations

Abbreviation	Description
API	Application Programming Interfaces
EAB	External Advisoy Board
EC	European Commission
EU	European Union
GA	Grant Agreement
GDPR	General Data Protection Regulation
IP	Intellectual Property
IPR	Intellectual Property Rights
R&I	Research and Innovation
WP	Work Package

List of partners

N.	Partner	Country	Acronym
1	FUNDACION TECNALIA RESEARCH & INNOVATION	ES	TECNALIA
2	ALL DIGITAL AISBL	BE	ALL DIGITAL
3	PALACKY UNIVERSITY OLOMOUC	CZ	UP
4	FUNDACION IBEROAMERICANA DEL CONOCIMIENTO	ES	GECON
5	CONFEDERATION EUROPEENNE DES ASSOCIATIONS DE PETITES ET MOYENNES ENTREPRISES	BE	CEA-PME
6	CEGOS ESPAÑA LEARNING & DEVELOPMENT	ES	CEGOS
7	UNIVERSITY OF SOUTHAMPTON	UK	UoS

EXECUTIVE SUMMARY

This executive summary encapsulates the key components and findings detailed in our methodology for modelling soft skills, i.e., establishing causal relationships between user behaviour and assessments of soft skills, as outlined in the accompanying document.

First, we carried out a comprehensive analysis of data obtained from experiments conducted under task T3.2, which included game achievements and standardized test results gathered through the MEGASKILLS platform. We grouped the data into the following categories:

- Pre-test and Post-test soft skills.
- Gaming data.
- Previous gaming experience.
- Socio-demographic data.

Secondly, in order to establish causal relationships between user behaviour and assessments of soft skills, we examined and evaluated various artificial intelligence (AI) methodologies designed specifically for analysing soft skills. We identified two main approaches developed, depending on the way soft skills are used, leveraging AI techniques to facilitate the assessment, identification, and prediction of soft skills:

- leveraging soft skills indicators to group and cluster individuals based on their proficiency in specific soft skills.
- inferring or predicting causal relationships between a source of information and soft skills.

With this aim we performed a data analysis in order to be used as inputs for the AI models and described the AI experiment design and methodology followed presenting the different steps carried out in MEGASKILLS for the AI development. We detail how through information gathered from the different sources, we will try to predict or infer the level of advancement of a specific soft skill in an individual. We aim to use AI models to replace test phases with player behaviour analysis, clustering players based on game preferences and demographic data to infer soft skill sets efficiently. Techniques like K-means, DBSCAN, and Spectral Clustering are considered for behaviour clustering, while regression methods like Logistic Regression (LR), Support Vector Regression (SVR), Random Forest (RF) or Gradient Boosting Regression (GBR) are used to predict skill improvements based on behaviour. Furthermore, we could analyse game achievements using Natural Language Processing (NLP) to correlate them with player skill development, aiding in recommending games for specific skill enhancement tailored to individual players.

1. INTRODUCTION

The work in this deliverable started based on the selection of the soft skills in WP2 and the rubrics designed in WP3. Once the soft skills were selected, 4 video games and 6 standard tests were selected in order to be integrated on the MEGASKILLS platform.

In this context, the objective of this deliverable is to describe the methodology followed to carry out the modelling of the soft skills (see Figure 1). First, we have analysed the data generated during the experiments organised in T3.2 that consists of the game achievements got by participants when playing the assigned video game and the standardised tests data participants completed at the beginning and the end of the experiment through the MEGASKILLS platform[1]. Secondly, due to the volume, frequency, and type of data (a combination of categorical magnitudes for standardised test and unstructured format for the game data), data transformation/feature engineering techniques have been used to translate received that collected data into numerical entities to be digested by machine learning algorithms. This process is an essential step for inferring the casual relationship between user behaviour and soft skills assessments. Finally, the different artificial intelligence (AI) approaches for soft skills analysis are reviewed and discussed, in order to introduce the AI design approach for MEGASKILLS.

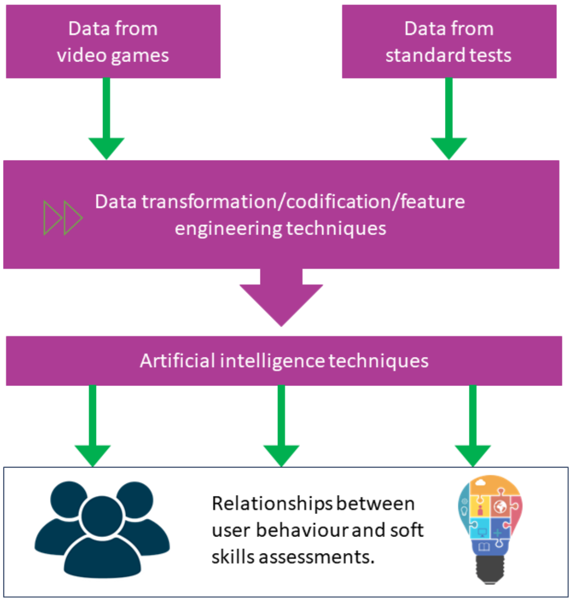

Figure 1 – Methodology followed in the modelling of soft skills in gaming

The structure of this deliverable is as follows: First, we start describing the modelling of data gathered during games on the STEAM platform[2] and through the pre- and post-tests completed by users available on the MEGASKILLS platform. Secondly, we examine in the scientific literature the AI approaches for soft skills development and paying particular attention to those who employed video games. And finally, we describe the AI design approach proposed for MEGASKILLS, starting with the data analysis, and continuing presenting the AI experiment design and methodology. We would like to point out that all the issues related to data handling, privacy, and ethics, especially considering the extensive use of personal data from gaming platforms, were dealt with according to the ethical framework included in D.1.4 *MEGASKILLS ethics and legal management report* and reviewed by EAB ethical expert.

[1] https://pilot.megaskills.eu/MegaSkills/index
[2] https://store.steampowered.com/

2. MODELLING DATA

This section details the modelling of data collected during games and through the pre- and post- tests completed by users.

2.1. Data generated in video games

Data generated in video games can be used for soft skills assessment and development, as they correlate with external measures, predict job and academic performance, enhance competencies in students, and can be analysed using various techniques to assess skills like communication, teamwork, and adaptability.

Stafford (2021, 2022) emphasized the potential of game data for understanding skill acquisition, particularly in the context of expertise development. This was supported by Chatziantoniou (2017), which focused on the design and development of an educational game for leadership assessment and soft skill optimization. These studies collectively highlighted the value of game data in assessing and enhancing soft skills, particularly in the context of leadership. Data generated in video games can be used for soft skills assessment, as the problem-solving estimates derived from the game significantly correlated with external measures, e.g., Shute et al. (2016). Nikolaou et al. showed how a game-based assessment measuring soft skills can predict self-reported job and academic performance (2019). Barr (2018) showed how commercial video games may be used to develop useful skills and competencies in undergraduate students, including communication, resourcefulness, and adaptability. Viviers et al. (2016) used a video game to effectively develop soft skills in accounting students, particularly in teamwork, communication, and time management.

Steam gaming network is the major gaming network with more than 132 million user accounts and over five million games in existence[3]. There are other online game providers networks such as Epic Games[4] or itch.io[5] with a wide catalogue of available games and registered users, but Steam provides a unique opportunity for analysis in several ways. First, because Steam is a platform for game distribution, covering a broad spectrum of gamers. Second, because Steam offers an open API for their platform, enabling the collection of data of all users and games.

The comprehensive nature of this data enables us to accurately characterize the distributions of a variety of gamer behaviours. Furthermore, apart from the game achievement Steam provides data on friendships, game ownership, and playtime data. The profile of a user shows an overview of his or her social and game statistics. Additionally, the profile also includes a "Steam level" that is displayed complemented by "badges" that denote various achievements on Steam. In fact, Rizani et al. (2023) found that digital badging (i.e., achievements) increases players' engagement with the publishing platform with good auxiliary data (such as types, rating, releases, and prices).

Previously, several researchers have harnessed the potential of the data provided by Steam. For example, O'Neill et al. (2016) highlighted that gamer behaviour is highly diverse and characterized by heavy-tailed distributions, with most players exhibiting modest behaviours in terms of playtime and money spent on games. Additionally, the research explored how playtime varies significantly over the course of a week for individual players and examined achievements in correlation with game genres. Claesdotter-Knutsson et al. (2022) discussed changes in gaming behaviour during the pandemic, showing an increase in screen time and digital entertainment consumption, particularly online gaming. The research identified positive associations with engaged gaming as well as addicted/problematic gaming, highlighting that increased time spent gaming

[3] https://store.steampowered.com/charts/
[4] https://store.epicgames.com/en-US/
[5] https://itch.io/

is a risk factor for developing gaming problems/addiction. Baumann et al. (2018) utilized a large dataset of over 100 million Steam platform users to present a comprehensive analysis of hardcore gamer profiles. Loria et al. (2021) found how the Steam user-defined tags better characterized the clusters identified to compare their network properties and their specific characteristics (e.g., genre, game elements, and mechanics) than the game genre, suggesting that how players perceive and use the game also reflects how they connect in the community.

2.1.1 Integration with STEAM in MEGASKILLS

Steam is a video game digital distribution service and storefront developed by Valve Corporation. Valve's PC gaming client offers a store, cloud saves, remote downloads, video streaming, and many other gamer-friendly features for your desktop, laptop, or Steam Deck. Steam provides several Application Programming Interfaces (APIs) so website developers can use data from Steam and allow developers to query Steam for information that they can present on their own sites. In the context of APIs, the word Application refers to any software with a distinct function. Interface can be thought of as a contract of service between two applications. This contract defines how the two communicate with each other using requests and responses. One of the APIs provides information related to the users and their stats, such as, the summaries of the players, their friend list, their achievements in the games, their stats for a game, their owned games, or their recently played games. Considering that we found this information of interest for the aims of our project, we decided to carry out the integration of the MEGASKILLS platform with Steam based on the API provided[6].

So, we developed a module on the platform integrating a series of queries that access user data through the API to collect and save it in the database of the MEGASKILLS platform. These queries are launched periodically (every day every 3 hours) through a cron (i.e., a scheduled task) designed and implemented on MEGASKILLS platform. An important requirement when making queries was that users had to make their Steam profile public. Otherwise, it was not possible to access their information via API. The information from the queries is stored incrementally, so that we can keep track of the evolution of users.

2.1.2 Data model

The following details the information gathered from Steam as well as the data model that supports it in the MEGASKILLS database:

- **User profile**: The main information about the user profile is stored in the table *usersprofile*. The fields included in the table are:

Table 1 – usersprofile table details, storing data related to the user profile

Field	Description
email	Email of the user
steamid	SteamID of the user
personaname	Name of the user
profileurl	url of the user profile
avatar	Avatar of the user

[6] https://steamcommunity.com/dev

avatarpremium	Avatar premium of the user
avatarfull	Avatar full of the user
personastate	0 - Offline, 1 - Online, 2 - Busy, 3 - Away, 4 - Snooze, 5 - looking to trade, 6 - looking to play. If the players profile is private, this will always be 0
communityvisibilitystate	1 - the profile is not visible to you (Private, Friends Only, etc), 3 - the profile is Public, and the data is visible
profilestate	If set, indicates the user has a community profile configured (will be set to 1)
lastlogoff	The last time the user was online, in unix time. Re-used with timecreated
commentpermission	If set, indicates the profile allows public comments. Re-used with loccountrycode
datestamp	Timestamp of the creation of the register
modified	Timestamp of the modification of the register

- **Achievements**: The information about the achievements made by the user in the games is stored in the table achievements. The fields included in the table are:

Table 2 – achievements table details, storing data related to the achievements made by the users in the games

Field	Description
idachievement	ID of the achievement
appid	ID of the game
steamid	SteamID of the user
nombrelogro	Name of the achievement
valorlogro	Value of the achievement
unlocktime	Date when the achievement was unlocked
datestamp	Timestap of the register

- **Achievementspercentagespergame**: Information about the percentage of achievements in a game is stored in the table *achievementspercentagespergame*. In this way we will be able to assess in a certain way the degree of difficulty related to the achievement of goals. The fields included in the table are:

Table 3 – achievementspercentagespergame table details, storing data related to percentage of achievements in a game

Field	Description
idpercentage	ID of the register
appid	ID of the game
nombrelogro	Name of the achievement
value	Value of the achievement
timestamp	Timestap of the register

- **Catachievements**: The different achievements that can be obtained in a game are stored in the table *catachievements*. The fields included in the table are:

Table 4 – catachievements table details, storing data related to the different achievements in a game

Field	Description
appid	ID of the game
nombrelogro	Name of the achievement
description	Criteria of the achievement
image	Image associated to the achievement

- **Catstats**: The different stats that can be obtained in a game are stored in the table *catstats*. The fields included in the table are:

Table 5 – catstats table details, storing data related to the different stats in a game

Field	Description
appid	ID of the game
mombrestat	Name of the stat
description	Criteria of the stat
image	Image associated to the stat

- **Friendlist**: The friends of a user are stored in the table *friendslist*. The fields included in the table are:

Table 6 – friendlist table details, storing data related to the friends of a user

Field	Description
idfriendslist	ID of the register
usersteamid	SteamID of the user (friend)

usersteamidowner	SteamID of the user (owner)
relationship	Relationship filter. Possibles values: all, friend.
friend_since	Unix timestamp of the time when the relationship was created.
datestamp	Timestamp of the register

- **Ownedgames**: The different games owned by a user are stored in the table *ownedgames*. The fields included in the table are:

Table 7 – ownedgames table details, storing data related to the games owned by a user

Field	Description
idownedgames	ID of the register
usersteamid	SteamID of the user
appid	ID of the game
playtimeforever	The total number of minutes played "on record", since Steam began tracking total playtime in early 2009.
datestamp	Timestamp of the register

- **Stats**: The different stats obtained in a game by a user are stored in the table *stats*. The fields included in the table are:

Table 8 – stats table details, storing data related to the stats obtained by a user in a game

Field	Description
idstat	ID of the register
appid	ID of the game
steamid	SteamID of the user
nombrestat	Name of the stat
valorstat	Value of the stat
datestamp	Timestamp of the register

- **Timeplayed**: The time played by a user in a game is stored in the table *timeplayed*. The fields included in the table are:

Table 9 – timeplayed table details, storing data related to the time played by a user in a game

Field	Description
idtimeplayed	ID of the register

usersteamid	SteamID of the user
appid	ID of the game
name	The name of the game
playtimeforever	The total number of minutes played "on record", since Steam began tracking total playtime in early 2009.
playtime2weeks	The total number of minutes played in the last 2 weeks
datestamp	Timestamp of the register

Figure 2 shows the entity relationship diagram of the data model that supports the information gathered from Steam:

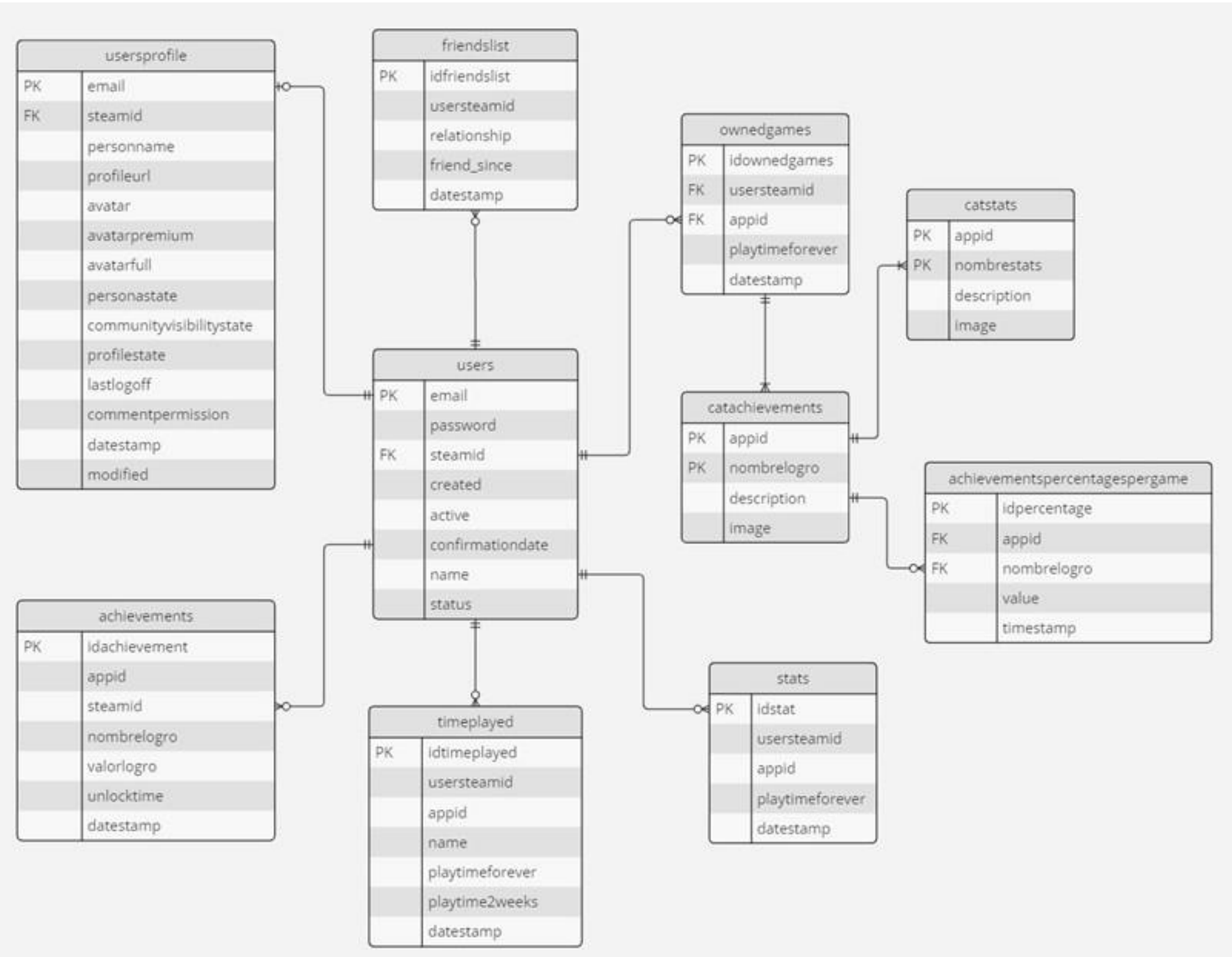

Figure 2 – Entity Relationship Diagram: Steam Data Model

2.2. Standardized tests data

In addition to data from the video games, data from a socio-demographic questionnaire and the following standardized tests have been used for the assessment of soft skills:

- Test of Critical Thinking (Bracken et al., 2003)
- Critical Thinking Assessment-Scale Short Form (Payan-Carreira, Sacau-Fontenla, Rebelo, Sebastião & Pnevmatikos, 2022)
- Reflective Thinking Skill Scale for Problem-Solving (RTSSPS) (Kizilkaya & Askar, 2009)
- I-ADAPT-M (Burke et al., 2006)
- Time Management Questionnaire (Britton & Tesser, 1991)
- CAMBIOS (Seisdedos, 2004)

Both the socio-demographic questionnaire and the standardized tests have been digitalized to gather this data and are part of the functionalities offered by the MEGASKILLS platform. The platform is accessible via the URL "https://pilot.megaskills.eu/MegaSkills/index".

In addition to these functionalities, the MEGASKILLS platform also provides user registration as well as the visualisation of the stored data. The sequence in which the data is stored is detailed below:

1. The first step is the registration of the user on the platform. During this process, the user will be asked for an email account and a password. These two fields are necessary for the identification of the user on the platform. Furthermore, the email account will be the data that will allow us to know, for each user, all the data obtained both from the video games and from the questionnaire or tests associated with them.
2. The next step is to complete the socio-demographic questionnaire (Note: data from this questionnaire were used for the assignment of the games in the study).
3. Once the questionnaire has been completed, the user will fill the standardized test. At this point, the data obtained from the tests will constitute the pre-test data set.
4. After that, the user must play the assigned game to collect the data generated by the video games.
5. Finally, when the playing time is over, the user will have to fill the standardized tests all over again. This data will constitute the post-test data set.

Figure 3 shows how this data was gathered on the platform:

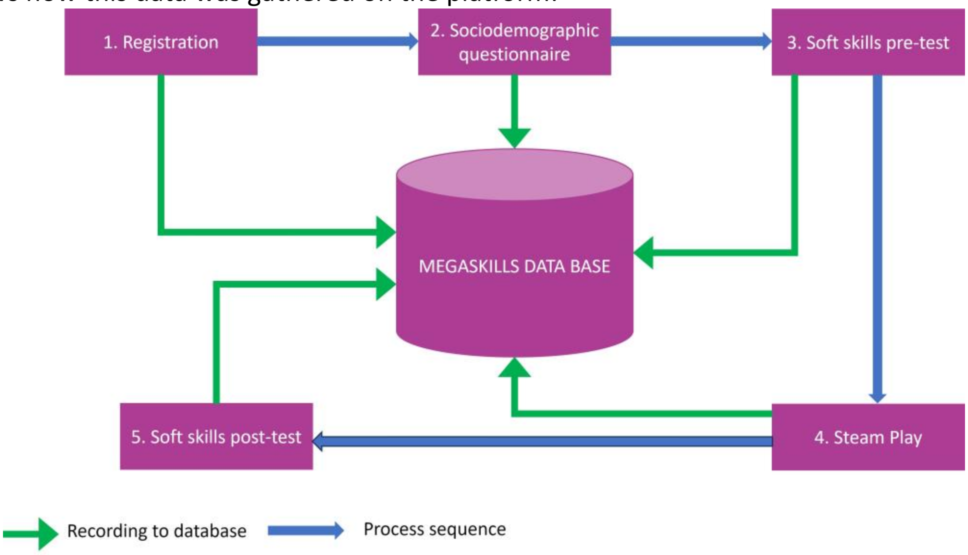

Figure 3 – Data collection process for standard tests

2.2.1 Sociodemographic data

The sociodemographic information questionnaire gathers information about the user's age, the gender, level of education and profile, the type of video game most frequently played, and the time spent playing. Age is defined as ranging from 18 to 65 and over. For the remaining data, only one value can be selected from the values offered.

This information is only linked to the user's platform profile and email address, without personal data. All data collected during the project will be treated anonymously. The data obtained by filling in this questionnaire are stored in the table "*testresultssd*" in the MEGASKILLS database. Its structure is described in section 2.2.3 of this document. Figure 4 shows how this data was gathered on the platform.

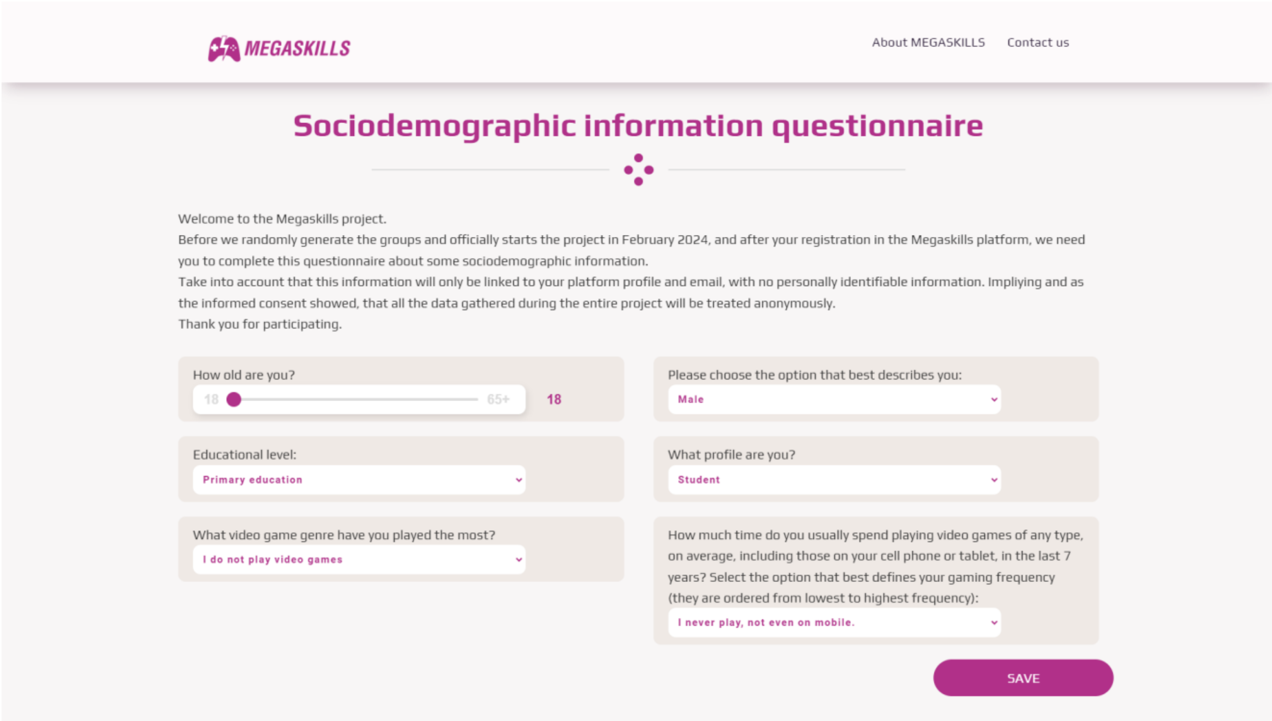

Figure 4 – Sociodemographic information questionnaire

2.2.2 Pre-Test and Post-Test

Below is a detailed description of the tests the user will take during the Pre-Test and Post-Test phases. These tests are the same in both phases, making it possible to evaluate your progress in the competition later, as the Pre-Tests are completed before the game and the Post-Tests after the game.

- **Test of Critical Thinking.** This test consists of 10 stories. Each story has several questions, making 45 questions in all. The best answer has to be chosen by the user. The data obtained from this test is stored in the "*testresultsct*" table of the MEGASKILLS database. Its structure is described in section 2.2.3 of this document. Figure 5 shows how this data was gathered on the platform.

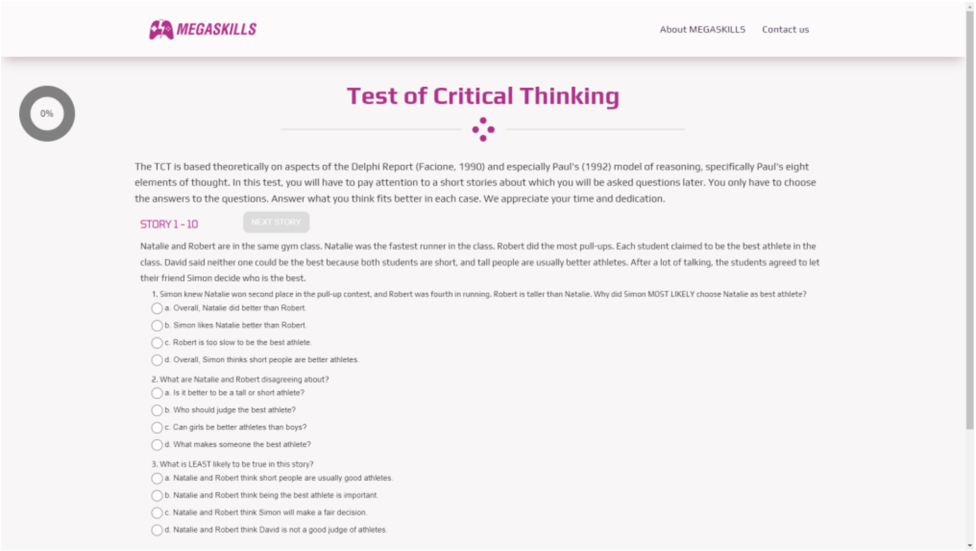

Figure 5 – Test of Critical Thinking

- **Critical Thinking Assessment-Scale Short Form**. This test consists of 60 questions that assess the user's level of critical thinking skills. The user must give an answer between 0 and 6, where 0 means "*Never*", 1 "*Rarely*", 2 "*Occasionally*", 3 "*Usually*", 4 "*Often*". 5 "*Frequently*" and 6 means "*Always*". The data obtained from this test is stored in the "*testresultscta*" table of the MEGASKILLS database. Its structure is described in section 2.2.3 of this document. Figure 6 shows how this data was gathered on the platform.

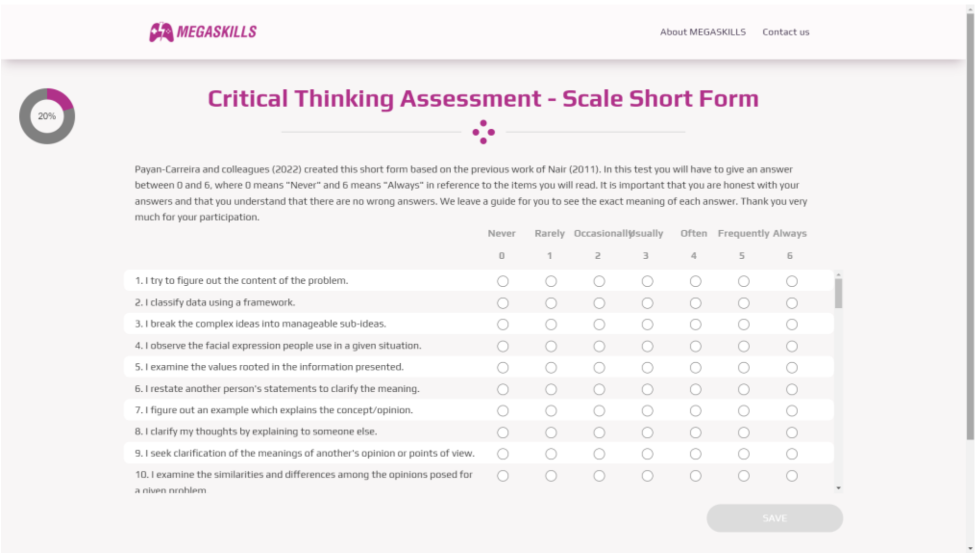

Figure 6 – Critical Thinking Assessment – Scale Short Form

- **Reflective Thinking Skill Scale for Problem-Solving (RTSSPS)**. This test consists of 14 questions that assess the user's level of complex problem-solving soft skills. The user must answer between 1 and 4, where 1 means "*Never*" and 4 means "*Always*". The data obtained from this test is stored in the "*testresultsps*" table of the MEGASKILLS database. Its structure is described in section 2.2.3 of this document. Figure 7 shows how this data was gathered on the platform.

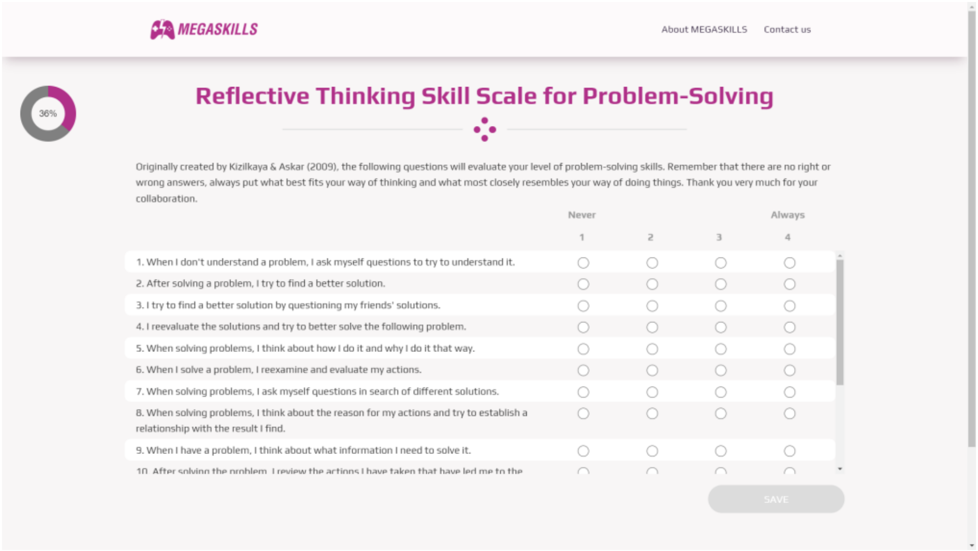

Figure 7 – Reflective Thinking Skill Scale for Problem Solving

- **I-ADAPT-M**. This test consists of 55 questions that assess the user's level of flexibility / adaptability soft skills. The user must answer between 1 and 5, where 1 means “Strongly Disagree”, 2 “Disagree”, 3 “Neither Agree nor Disagree”, 4 means "Agree", and 5 means “Strongly Agree”. The data obtained from this test is stored in the "testresultsiadapt" table of the MEGASKILLS database. Its structure is described in section 2.2.3 of this document. Figure 8 shows how this data was gathered on the platform.

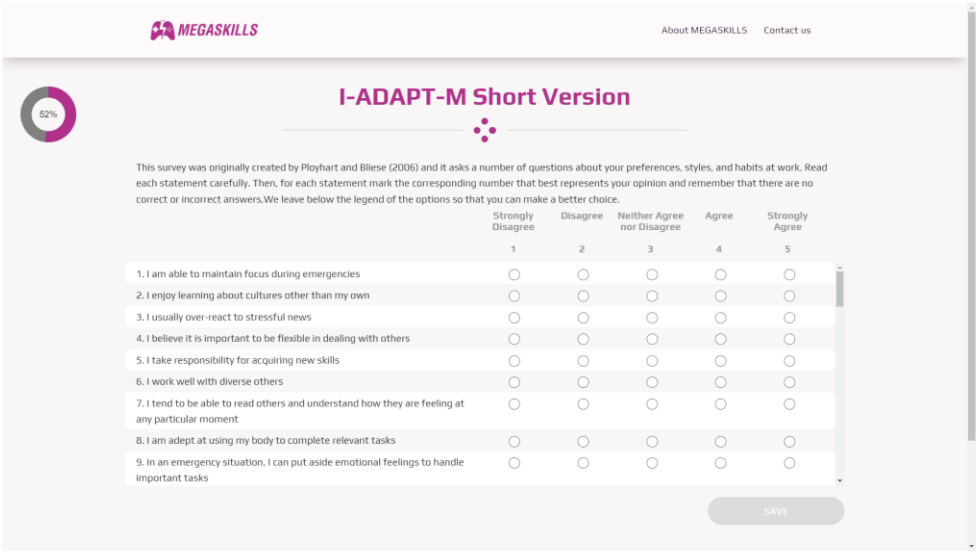

Figure 8 – I-ADAPT-M Short Version

- **Time Management Questionnaire**. This test consists of 18 questions that assess the user's level of time management soft skills. The user must answer between 1 and 5, where 1 means “*Never*”, 2 “*Infrequently*”, 3 “*Sometimes*”, 4 means "*Often*" and 5 means “*Always*”. The data obtained from this test is stored in the "*testresultstm*" table of the MEGASKILLS database. Its structure is described in section 2.2.3 of this document. Figure 9 shows how this data was gathered on the platform.

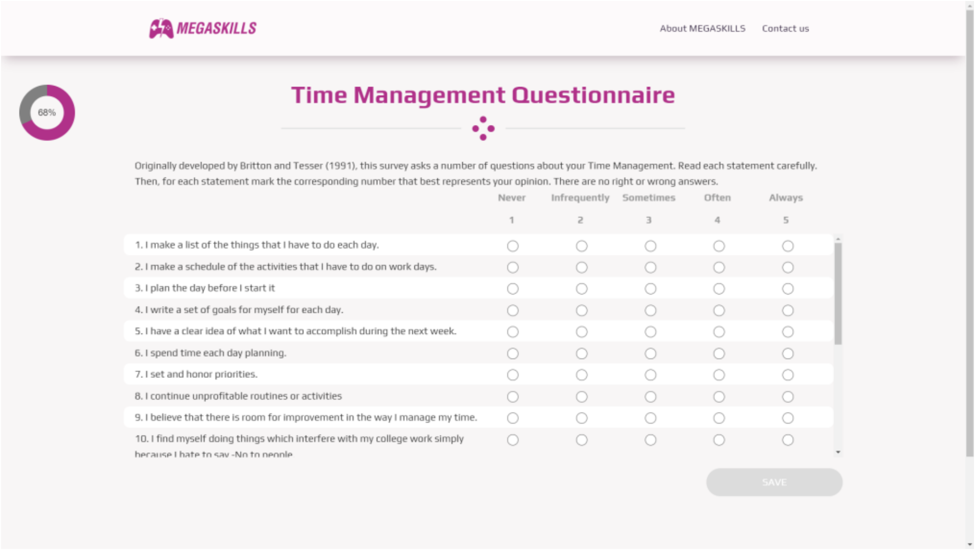

Figure 9 – Time Management Questionnaire

- **CAMBIOS**. The CAMBIOS test is implemented as a mini game where the user must evaluate the performance of a factory workers. Within a maximum of 7 minutes, the user must evaluate 27 rounds. In each round, he is shown 3 boxes prepared by the workers with their label and 2 modification instructions, labelled A and B. The user must judge whether the result of applying the modification instructions to each box produces the next box presented. So, applying change A to cell 1 produces cell 2; applying change B to cell 2 produces cell 3. The result of each round is stored in the "*cambiosanwers*" table and the result of the mini game in the "*cambiostoken*" table in the MEGASKILLS database. Figure 10 shows the first round that must be evaluated by the user.

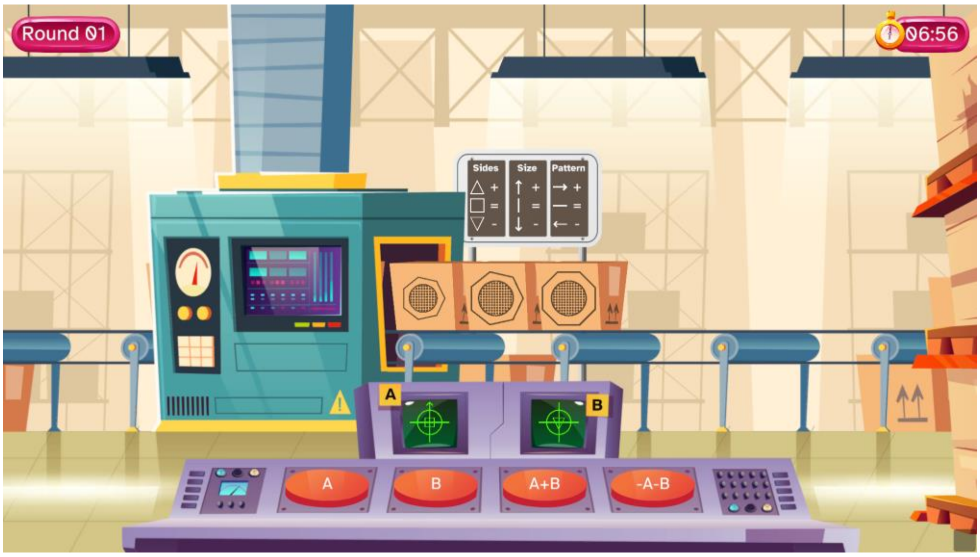

Figure 10 – CAMBIOS mini-game

Once the pre-test and post-test have been completed, users will be able to compare the scores obtained in the different tests through their profile (currently hidden) as well as consult the achievements obtained in the video games. An example of this functionality can be seen in Figure 11.

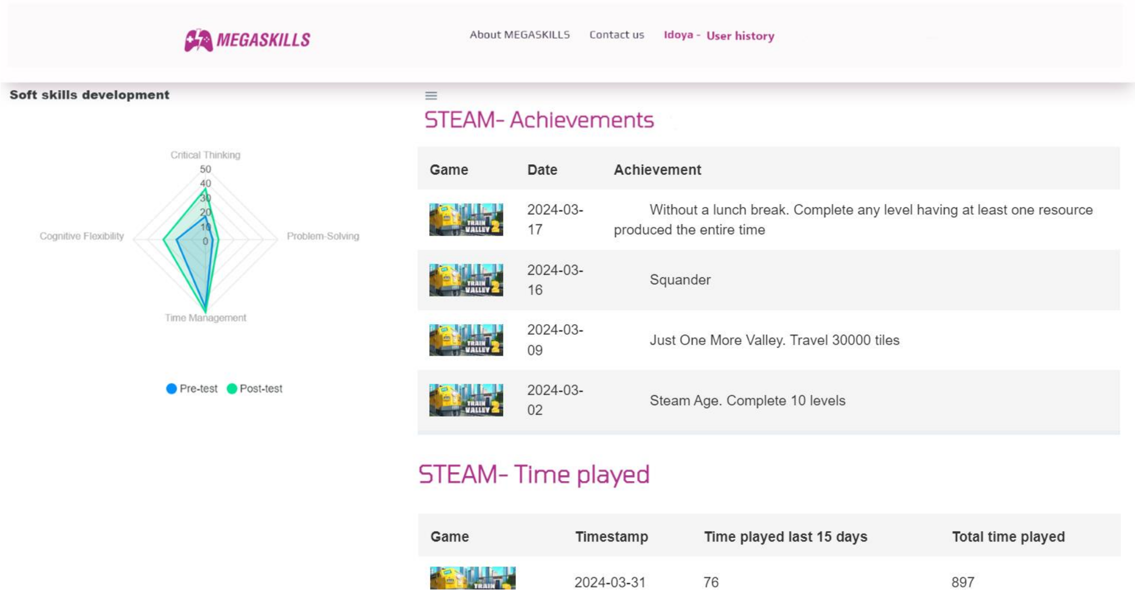

Figure 11 – User history

2.2.3 Data model

Below is the data model of the MEGASKILLS database, which stores the data from the sociodemographic information questionnaire, the various pre-tests and post-tests and user data.

- **Users**: Data from the user registration process will be stored in the table *users*. The table fields are described below.

Table 10 – users table details, storing basic data related to the user

Field	Description
email	Email of the user
password	Password of the user
steamid	SteamID of the user
created	Timestamp when the register is created
active	User active or not
confirmationdate	Timestamp when the user confirms the email
name	Name (alias) of the user
status	An indication of the step of the process the user is in

- **Testresultssd**: The data from the sociodemographic information questionnaire will be stored in the table *testresultssd*. The table fields are described below.

Table 11 – testresultssd table details, storing data related to the sociodemographic information questionnaire

Field	Description
idtestresult	Register ID
email	Email of the user
idtest	Test ID
idquestion1	ID of the question *"How old are you?"*
result1	The answer to the question *idquestion1:* user age
idquestion2	ID of the question *"Please choose the option that best describes you:"*
result2	The answer to the question *idquestion2*. Valid values: "*Male*", "*Female*", "*Other*", "*Prefer not to answer*".
idquestion3	ID of the question *"Educational level:"*
result3	The answer to the question *idquestion3*. Valid values: "*Primary education*", "*Secondary education*", "*Post-secondary non-university studies*", "*Bachelor's degree*", "*Master's degree*", "*Doctorate degree or postdoc*".
idquestion4	ID of the question *"What profile are you?"*
result4	The answer to the question *idquestion4*. Valid values: "*Student*", "*Teacher*", "*Other*".
idquestion5	ID of the question *"What video game genre have you played the most?"*
result5	The answer to the question *idquestion5*. Valid values: "*I do not play video games*", "*Action-Adventura games (such as Assassin' s Creed", Legend of Zelda or The Witcher)*", "*Fighting games (such as Tekken, Mortal Kombat or Super Smash Bros)*", "*Shooter games (such as Counter Strike, Valorant, Overwatch or Call of Duty)*", "*Platform games (such as Crash Bandicoot or Hollow Knight)*", "*Puzzle games (such as Candy Crush or Monument Valley)*", "*Racing games (such as Forza Horizon or Gran Turismo)*", "*Role playing games (such as World of Warcraft, Diablo or Final Fantasy VII)*", "*Simulation games (such as Sims, Farming simulators)*", "*Sports games (such as FIFA, NBA or Switch Sports)*", "*Strategy video games (such as Starcraft, Clash Royale or Civilization)*", "*Survival games (such as Don´t Starve, Subnautica or Minecraft)*", "*MOBA games (such as League of Legends or DOTA 2)*"
idquestion6	ID of the question *"How much time do you usually spend playing video games of any type, on average, including those on your cell phone or tablet, in the last 7 years? Select the option that best defines your gaming frequency (they are ordered from lowest to highest frequency):"*
result6	The answer to the question *idquestion6*. Valid values: "*I never play, not even on mobile*", "*About once a month*", "*A few times a month*", "*A few times a week*",

	"Every day, for less than 1 hour", "Every day, between 1 and 3 hours", "Every day, for more than 3 hours"

- **Testresultsct**: The data from the test of critical thinking will be stored in the table *testresultsct*. The table fields are described below.

Table 12 – testresultsct table details, storing data related to the test of critical thinking questionnaire

Field	Description
idtestresult	Register ID
email	Email of the user
idtest	Test ID
idquestion1	ID of the first question
result1	The selected answer to the first question
idquestion2	ID of the second question
result2	The selected answer to the second question
idquestion3	ID of the third question
result3	The selected answer to the third question
idquestion4	ID of the fourth question
result4	The selected answer to the fourth question
…	…(*Until the last question, the forty-fifth, the same question and answer structure is repeated*)
…	..
idquestion45	ID of the forty-fifth question
result45	The selected answer to the forty-fifth question

- **Testresultscta**: The data from the test of Critical Thinking Assessment will be stored in the table *testresultscta*. The table fields are described below.

Table 13 – testresultscta table details, storing data related to the test of Critical Thinking Assessment

Field	Description
idtestresult	Register ID
email	Email of the user
idtest	Test ID

idquestion1	ID of the first question
result1	The selected answer to the first question
idquestion2	ID of the second question
result2	The selected answer to the second question
idquestion3	ID of the third question
result3	The selected answer to the third question
idquestion4	ID of the fourth question
result4	The selected answer to the fourth question
…	…(*until the last question, the sixtieth, the same question and answer structure is repeated*)
…	..
idquestion60	ID of the sixtieth question
result60	The selected answer to the sixtieth question

- **testresultssp**: The data from the test of Problem-Solving will be stored in the table *testresultssp*. The table fields are described below.

Table 14 – testresultssp table details, storing data related to the test of Problem-Solving

Field	Description
idtestresult	Register ID
email	Email of the user
idtest	Test ID
idquestion1	ID of the first question
result1	The selected answer to the first question
idquestion2	ID of the second question
result2	The selected answer to the second question
idquestion3	ID of the third question
result3	The selected answer to the third question
idquestion4	ID of the fourth question
result4	The selected answer to the fourth question

…	…(*until the last question, the fourteenth, the same question and answer structure is repeated*)
…	..
idquestion14	ID of the fourteenth question
result14	The selected answer to the fourteenth question

- **Testresultsiadapt**: The data from the test of I-ADAPT-M will be stored in the table *testresultsiadapt*. The table fields are described below.

Table 15 – testresultsiadapt table details, storing data related to the test of I-ADAPT-M

Field	Description
idtestresult	Register ID
email	Email of the user
idtest	Test ID
idquestion1	ID of the first question
result1	The selected answer to the first question
idquestion2	ID of the second question
result2	The selected answer to the second question
idquestion3	ID of the third question
result3	The selected answer to the third question
idquestion4	ID of the fourth question
result4	The selected answer to the fourth question
…	…(*until the last question, the fifty-fifth, the same question and answer structure is repeated*)
…	..
idquestion55	ID of the fifty-fifth question
result55	The selected answer to the fifty-fifth question

- **Testresultstm**: The data from the test of Time Management will be stored in the table *testresultstm*. The table fields are described below.

Table 16 – testresultstm table details, storing data related to the test of Time Management

Field	Description
idtestresult	Register ID

email	Email of the user
idtest	Test ID
idquestion1	ID of the first question
result1	The selected answer to the first question
idquestion2	ID of the second question
result2	The selected answer to the second question
idquestion3	ID of the third question
result3	The selected answer to the third question
idquestion4	ID of the fourth question
result4	The selected answer to the fourth question
…	…(*until the last question, the eighteenth, the same question and answer structure is repeated*)
…	..
idquestion18	ID of the eighteenth question
result18	The selected answer to the eighteenth question

- **Cambiostoken**. The result of the mini-game Cambios will be stored in the table *cambiostoken*. The table fields are described below.

Table 17 – cambiostoken table details, storing data related to the mini-game Cambios

Field	Description
token	Email of the user
timestamp	Timestamp the user played the mini-game
time	The amount of time, in seconds, the user took to complete the game
score	Final score, sum of hits in each of the 27 rounds

- **Cambiosanswers**. The result of each of the 27 rounds of the game is stored in the table *cambiosanswers*. The table fields are described below.

Table 18 – cambiosanswers table details, storing data related to of each of the 27 rounds of the game Cambios

Field	Description
token	Email of the user
level	ID of the round

answer	Answer selected by the user
points	1 if the answer is the right one; 0 if the answer is the wrong one
time	The amount of time, in seconds, that the user took to select the answer

Figure 12 shows the entity relationship diagram of the data model that supports the information gathered from sociodemographic questionnaire and the standardized tests:

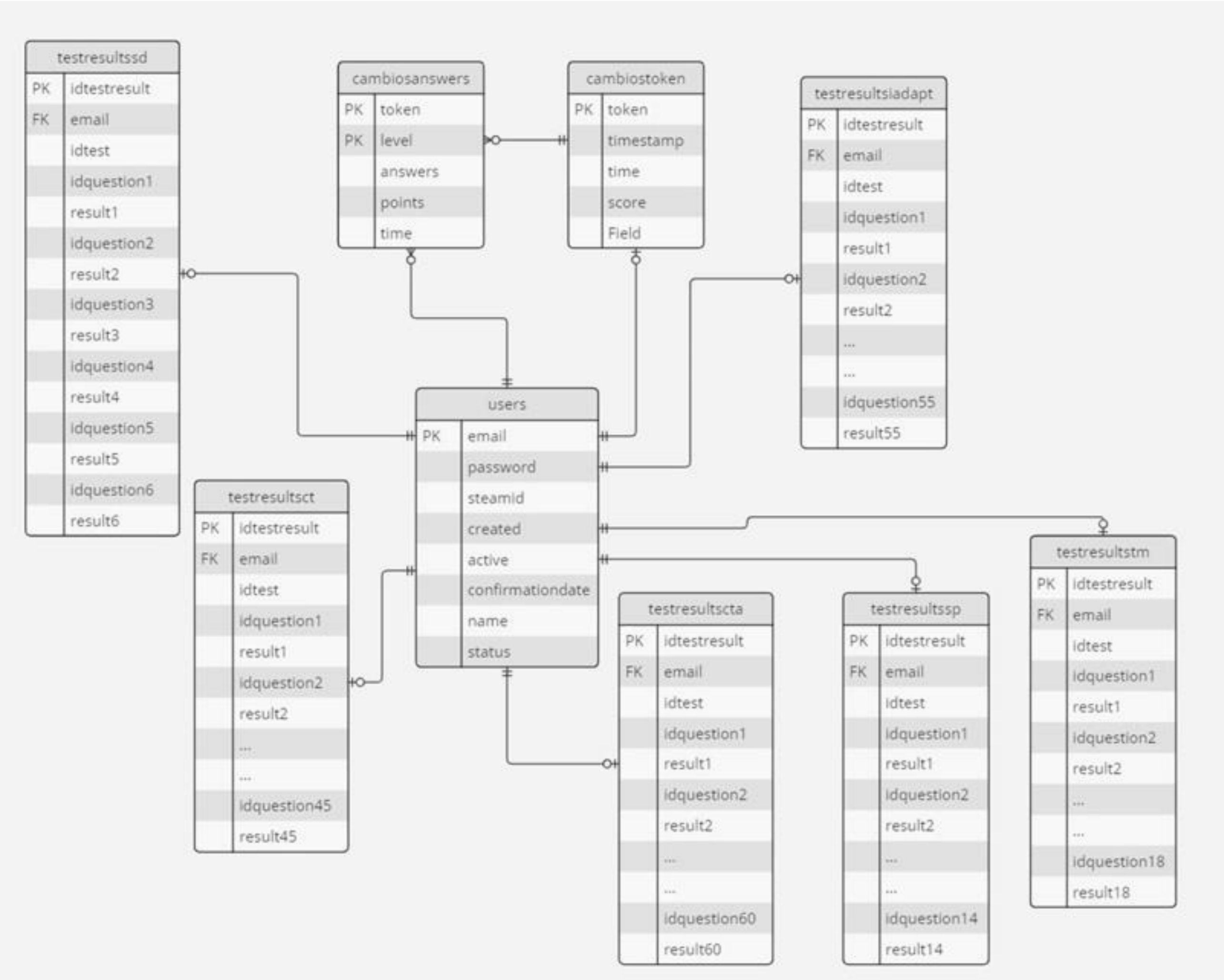

Figure 12 – Entity Relationship Diagram: Standard Test Data Model

3. ARTIFICIAL INTELLIGENCE APPROACHES FOR SOFT SKILLS ANALYSIS

While the importance of soft skills is widely acknowledged, the process of developing, refining and measure these skills can be challenging. Unlike technical skills, which can often be learned through formal education or training programs, soft skills are inherently more complex, making them harder to quantify and teach. However, despite these challenges, there has been a growing emphasis on developing innovative techniques and methodologies to improve soft skills in recent decades.

3.1. AI and soft skills development

AI techniques encompass a range of methodologies and approaches aimed at enabling machines to simulate human-like intelligence and behaviour. Here are brief explanations of some common AI techniques:

- **Machine Learning (ML):** ML involves training algorithms to learn from data and make predictions or decisions without explicit programming. This includes supervised learning (learning from labelled data), unsupervised learning (finding patterns in unlabelled data), and reinforcement learning (learning from feedback).
- **Deep Learning:** A subset of ML that uses artificial neural networks with many layers (deep neural networks) to learn hierarchical representations of data. Deep learning excels in tasks like image recognition, natural language processing (NLP), and speech recognition.
- **Natural Language Processing (NLP):** NLP focuses on enabling computers to understand, interpret, and generate human language. Techniques include sentiment analysis, machine translation, named entity recognition, and text summarization.
- **Computer Vision:** Computer vision involves teaching machines to interpret and understand visual information from images or videos. This includes tasks like object detection, image classification, facial recognition, and scene understanding.
- **Fuzzy Logic:** Fuzzy logic deals with reasoning that is approximate rather than exact. It is used in situations where precise numerical values are difficult to determine, such as in control systems, decision-making processes, and pattern recognition.

These AI techniques are often combined and adapted to solve specific problems across various domains, contributing to the advancement and application of artificial intelligence in diverse fields such as healthcare, finance, transportation, and entertainment.

In recent years, the intersection of AI and soft skills development has emerged as a promising area of research, offering new insights and methodologies to enhance interpersonal competencies in various contexts. Within this realm, two main approaches have been developed, depending on the way soft skills are used, leveraging AI techniques to facilitate the **assessment, identification**, and **prediction** of soft skills.

The **first approach** focuses on leveraging soft skills indicators to group and cluster individuals based on their proficiency in specific soft skills. By analysing data such as job roles, performance evaluations, and self-assessments, researchers can identify patterns and correlations between soft skills and job functions. In Ramos-Pulido et al., 2023, the task of job level prediction (i.e., First, Middle, and High) is carried out. For that endeavour, authors study the relationships between candidates' information, such as *current salary*, *seniority,* or *people in charge,* among others, and candidates' job level. The models used underneath varies from logistic regression to random forests. This kind of approaches enables organizations to better understand the soft skills landscape within their workforce and tailor training and development programs accordingly. Additionally, it facilitates the alignment of individuals with roles that align with their strengths in particular soft skills, thereby optimizing team dynamics and productivity. From a different perspective but same target, many of the works focused on the use natural language processing (NLP) techniques to extract information from various sources, such as resumes, job descriptions and internal forums, to infer the soft skills of candidates or employees. By analysing text data, NLP algorithms can identify keywords, phrases, and linguistic patterns associated with specific soft skills. This enables organizations to automate the process of soft skills assessment, making it more efficient and scalable. Moreover, by correlating textual data with job roles and performance metrics, organizations can gain valuable insights into the soft skills profiles of their workforce and identify areas for improvement. Some examples can be found in Aviv et al.,2021 which utilizes NLP techniques to extract information form forum data and using Latent Semantic Analysis (LSA) and clustering to assess employees' collaborative skills (Internal cohesion, Responsibility, Social Impact, Communication density, participation) by role, demonstrating the potential of AI in workplace skill evaluation.

The **second approach** involves inferring or predicting causal relationships between a source of information and soft skills. By analysing large datasets containing information about individuals' socio-economic backgrounds, educational attainment, and employment histories, researchers can identify patterns and correlations with soft skills development. Machine learning algorithms can then be used to predict the likelihood of individuals possessing certain soft skills based on their socio-economic characteristics. This approach provides valuable insights into the factors that influence soft skills development and can inform targeted interventions and policy initiatives aimed at enhancing soft skills among specific populations.

The prevalence of NLP over any other techniques becomes noticeable. Such is the case of the work carried out in Azzini et al., 2018 where authors employ different classification algorithms to identify communication soft skills (i.e., Effectiveness, dissemination, teaching, written communication, verbal communication, and internationality) from textual descriptions in CVs, offering a scalable solution for talent evaluation in research settings. In this case again a structure of preprocessing, tokenization, feature selection using evolutionary algorithms and classification (i.e., naive bayes, KNN, SVM, DTree and RF) is conducted. Harirchian et al.,2022 and Wings et al., 2021 employ an advanced natural language techniques BERT to mine soft skills from text sources like social media profiles and job offers respectively, then a classification (i.e., Logistic regression (LR), Gradient boosting classifier (GBC), Random Forest (RF), Support vector machine (SVM) and Multilayer Perceptron (MLP)) is performed. These studies highlight the potential of AI to analyse unstructured data and extract valuable insights for talent acquisition and job matching. Zhao et al. 2015 and Fareri et al.,2021 developed a Named Entity Recognition (NER) classification technique. They searched for context clues (CLUE) to help detect soft skills (problem solving, active reasoning, communication, professionalism, leadership, teamwork, and flexibility). Then, a group of experts classified the extracted phrases (assigning a class to each word), and finally, a classifier was employed to predict each class (i.e., soft skill). SVM work better for them than MLP. Finally, they studied the relationships between jobs that share soft skills or soft skills that appear in various jobs. Kannan et al., 2023 explore how AI can assist in identifying and predicting students' soft skills (professional, analytical, linguistic, communication y ethical), aiding educators in personalized learning interventions and curriculum design. In this case an analysis by means of the socio-cultural features of the students they predict their soft skills based. The Cumulative Gradient Point Average (CGPA) metric is employed to assess the extent of soft skills, alongside the utilization of psychological tests to evaluate their level.

In summary, the incorporation of AI techniques into the sphere of soft skills development has ushered in new avenues for comprehending, evaluating, and improving interpersonal competencies. Through the utilization of AI-driven methodologies, organizations can delve deeper into the soft skills landscape within their workforce, automate soft skills assessment processes, and devise strategies to nurture soft skills development among individuals. As AI progresses, the potential for innovation in this domain is vast, promising further advancements. These collective studies highlight the transformative potential of AI in soft skills development and assessment across diverse domains, including recruitment, education, and workforce training. By harnessing AI-driven approaches, organizations and educational institutions can unlock fresh opportunities for talent management and skill enhancement, thereby fostering a more adaptable and agile workforce in the digital era.

3.2. AI and soft skills development using video games

As observed, there is a considerable amount of literature focusing on analysing and extracting valuable information through the necessary AI techniques from various sources: 1) texts or human interactions, 2) sociocultural indicators, or 3) results from psychological tests based on standardized metrics in the field. As previously mentioned, video games have been gaining relevance in the acquisition and improvement of soft skills, enhancing their use in both educational and corporate settings. However, the intersection of these two areas—video games as a tool for soft skills acquisition and the use of AI techniques for analysis—has very few or almost no references that develop a methodology combining both fields. One of the references that can be found today in Alloza et al., 2017 where authors study the use of commercial videogames to enhance the development of several soft skills, concretely they use Pacman for Risk Taking, Tetris for Spatial reasoning and Flappy Bird for Persistence. In this study they monitor a small group of individuals by a telemetry system and film they face during gaming. Finally, the study is analysed by means of basic statistical rules to infer the evolution of the soft skills and they used artificial vision (VGG convolutional neural network to be fine-tuned) to infer the feeling of the user meanwhile they were playing. However, and as mentioned before, the study was performed over a reduced group (only 5 individuals per game) and they didn't use AI for the analysis of the information gathered from the study.

The development of MEGASKILLS emerges as a novel and crucial endeavour due to its integration of both video games and AI. This integration holds promise because video games have proven to be potent tools for acquiring soft skills, fostering skill enhancement, and encouraging their usage. Concurrently, AI is increasingly acknowledged as a valuable aid for domain experts, facilitating the analysis of soft skills with greater ease and yielding better results.

4. AI DESIGN APPROACH FOR MEGASKILLS

The transition from traditional testing methods to AI-based tools represents a significant paradigm shift in the assessment practices. Traditional tests rely on predetermined questions, scoring rubrics, and fixed answer keys. The procedures to obtain soft skills assessment through these tests are costly, but their design allows to capture the degree of development of a soft skill with detail. The objective is to determine if, at some extent, when a soft skill is improved, this improvement can be measured with less costly methods.

The broad concept of what a soft skill is can lead to its improvement from an extensive amount of sources, most of which cannot be captured during gaming sessions, i.e., even if the user plays and achieves certain achievements, it is not possible to certify individually that the increase in a certain soft skill level is attributable to the game experience. The challenge lies, thus, on the degree to which the objective improvement can be attributed to gaming, and as a result, the degree objectively measured impact of gaming in this soft skill level.

AI-based tools leverage machine learning algorithms that adapt and learn from data. To train these algorithms effectively, a vast amount of diverse data is essential. Within the context of MEGASKILLS, these data will include on the one hand, in-game data, but also the results from the standard tests and the socio-demographic contextual data that are intended to provide the nuances that cannot be captured only in the gaming data. AI models are intended to learn patterns that ultimately provide the means to estimate the competence on certain skills, task for which the main assumption is that a certain skill level will be increased after a gaming period, and that the gaming experience itself will have been part of that skill increment. These relationships between gaming and soft skills are expected to be captured and represented through AI models.

It is also worth mentioning that there are some potential limitations within AI methodologies and data collection processes that will be considered and managed, such as:

- **Bias and Representational Issues**: AI methodologies are susceptible to biases present in training data, leading to biased outcomes and representations. Data collection processes can inadvertently capture and reinforce biases, resulting in skewed or incomplete datasets that do not adequately represent diverse populations or scenarios.
- **Data Quality and Quantity**: Limitations in data quality, such as missing or noisy data, can adversely affect the performance and reliability of AI models. Additionally, the quantity and diversity of data may not always be sufficient to capture complex patterns or edge cases, leading to potential inaccuracies or limited generalizability.
- **Ethical and Regulatory Challenges**: AI methodologies and data collection processes raise ethical concerns, including privacy violations, unintended consequences (e.g., discrimination), and regulatory compliance issues (e.g., GDPR, fairness standards). Ensuring ethical use of AI requires addressing these challenges through responsible design, governance, and stakeholder engagement.

4.1. Data analysis

As aforementioned, the training of AI models requires an extensive set of input data to be able to represent the underlying reality without bias. In order to model the soft skill development of the experiment participants and considering that in principle only data that can be collected automatically are available, the most detailed data possible of the gaming platform and gaming experience have been captured during the experiment phase. The data model explained in previous Section 2 shows the architecture required to capture these input data, that can be summarized as follows:

- A measurement of the assessed soft skills prior to the gaming stage (pre-test phase), measured through specific tests for each of the considered soft skills.

- A measurement of the same assessed soft skills after the gaming stage (post-test phase), measured again through the same tests.
- All gaming data that reflect the gaming experience during the gaming stage.
- All previous gaming experience, in order to characterize the gaming habits of each subject. These data allow to level all users establishing a baseline relative to the frequency a user plays.
- Socio-demographic contextual factors that could have an impact in the previous soft skill assessments.

These data have been captured during the experimentation phase and will act as inputs for the AI models. The main sources of data are analysed below.

4.1.1 Pre-test and Post-test soft skills

The analysed 4 soft skills are measured before and after the gaming experience through the tests described in Section 2. The first analysis is performed in terms of responses to each kind of test, which reveals that only a fraction of the original subjects has submitted test answers. This last user count around 150 is the one that will limit the experimental framework, as the answers to the post-test phase are crucial to observe evolution of each soft skill, as detailed in Figure 13.
This analysis used the data stored in the following tables: testresultsct, testresultscta testresultsiadapt, testresultssp, testresultstm, posttestresultsct, posttestresultscta, posttestresultsiadapt, posttestresultssp, posttestresultstm.

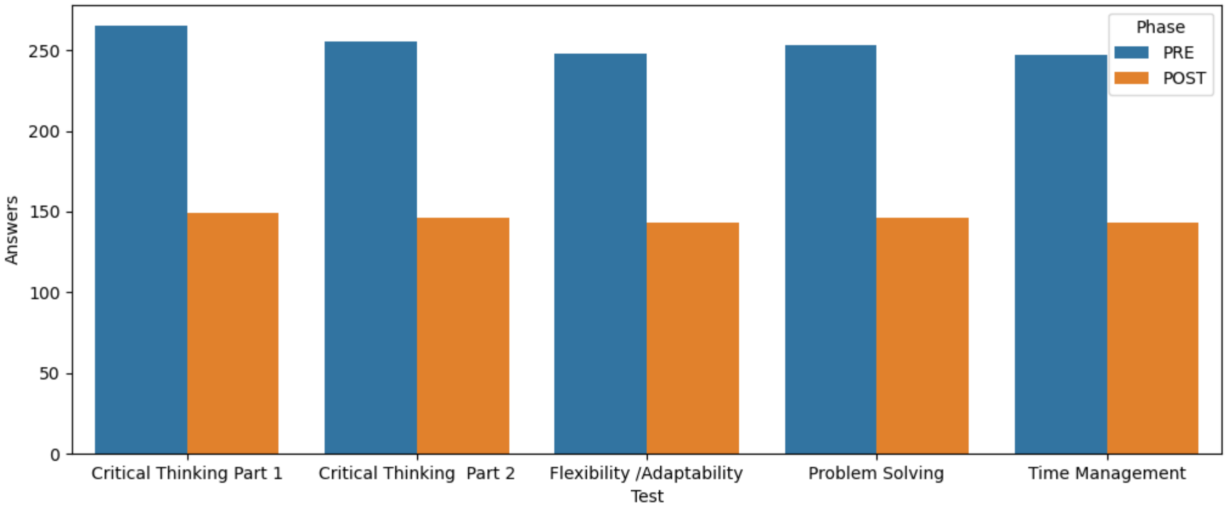

Figure 13 – Answers for pre- and post-tests

The results of the tests are aggregated into a single metric for each user that ultimately reflects his/her level in a certain soft skill. This aggregation is performed afterwards, but it is possible to perform a statistical analysis of the evolution in the answer distribution, presented in following figures: **Figure 14** and **Figure 15** for Critical Thinking (as it comprises two different tests), **Figure 16** for Flexibility / Adaptability,

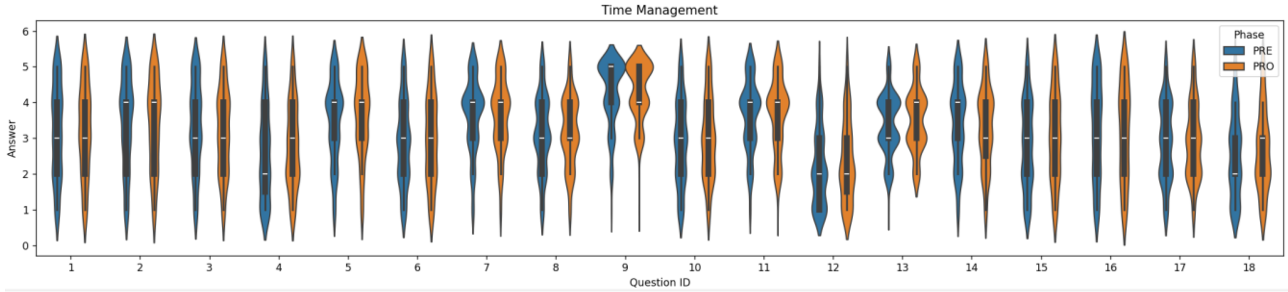

Figure 17 for Time Management and **Figure 18** for Problem Solving:

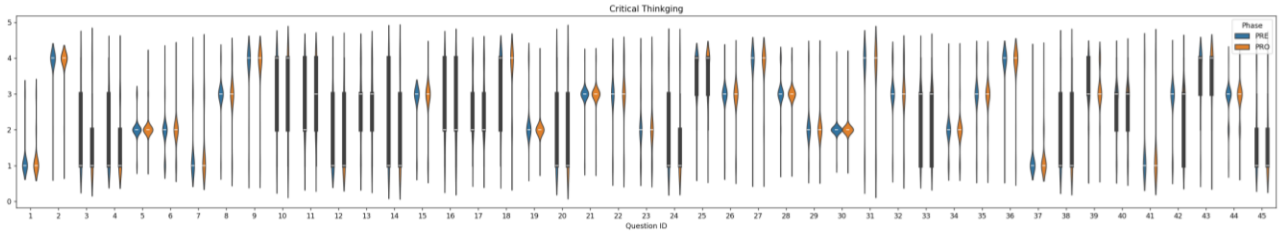

Figure 14 – Distribution of answers before and after for Critical Thinking Test Part 1

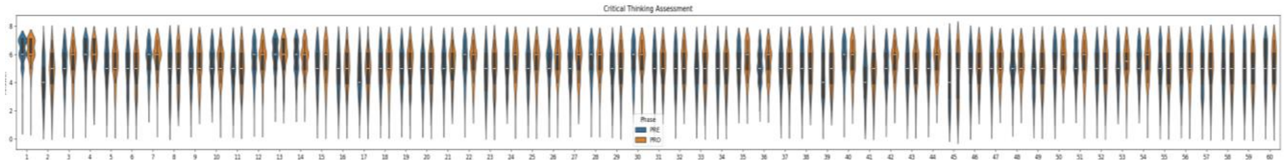

Figure 15 – Distribution of answers before and after for Critical Thinking Test Part 2

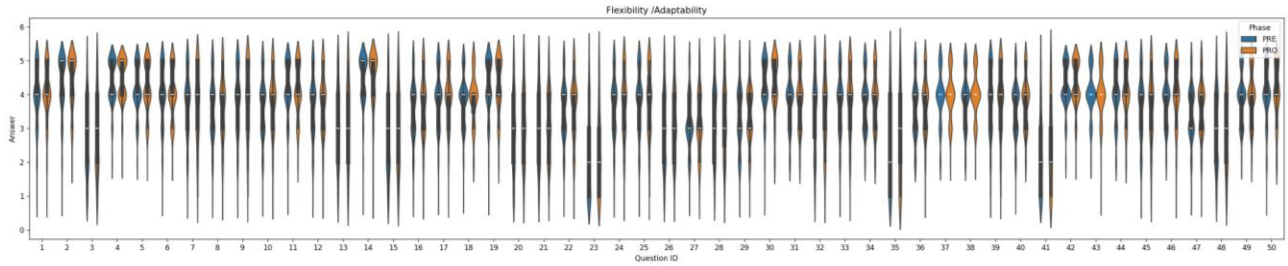

Figure 16 – Distribution of answers before and after for Flexibility /Adaptability Test

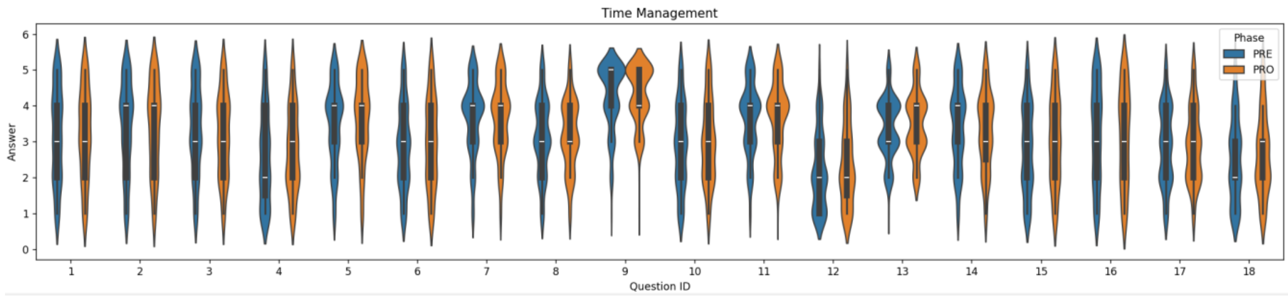

Figure 17 – Distribution of answers before and after for Time Management Test

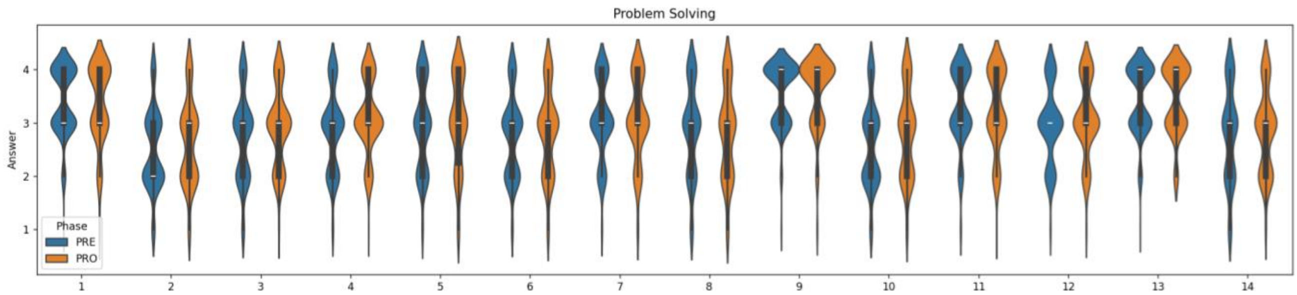

Figure 18 – Distribution of answers before and after for Problem Solving Test

As it can be observed in these violin plots that reflect the distribution of each set of answers for every question of each test, distributions are very similar in both phases, even considering that pre-test phase has up to 50% more answers (users that answered the questionnaire). This is in principle an expected behaviour, as, at a population level, the distribution of these answers should remain similar, although the aggregation of each set of questions in a single metric is expected to provide enough nuances to finely define the soft skill level. AI models will work mainly on these aggregated metrics but will also consider the user-question level evolution (i.e., tracking the changes per user at the question level), in order to detect questions that could be relevant markers for the task at hand.

4.1.2 Gaming data

The gaming data during the experiment can be obtained from data captured in the following tables: achievements, achievementspercentagespergame, catachievements, catstats, stats, timeplayed.

Data have been captured during the gaming phase every 3 hours. The **Figure 19** above represents the gaming minutes spent in experiment games (orange) and other games during the last 2 weeks up to April 8th. Only 71 of the subjects of the experiment played during this time, and as the plot shows, most of them spent most of the time in games not related to the experiment. From this plot it can be elucidated which subjects have more gaming habits and experience and continue playing after the experiment, and that many users stopped playing after the experiment.

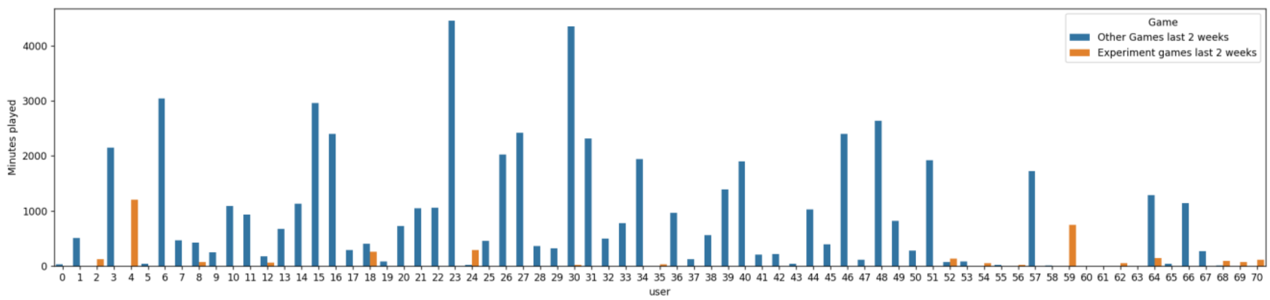

Figure 19 – Time played during the last 2 weeks of the experiment by users who played these two weeks

As these data was captured every 3 hours, deeper analyses will be performed during the AI model training phase, to conform variables that represent the gaming habits and the evolution of time spent playing both experiment games and other games. These habits are quite important to characterize the type of users and establish the impact of their gaming sessions in their measured soft skills. We can anticipate that gaming has different impact on a subject that is accustomed to gaming than on a subject that is playing for the first time, so it is important to separate and discern between subject categories. Nonetheless, the analysis can become deeper when considering also the previous gaming experience.

4.1.3 Previous gaming experience

The gaming experience can be obtained in an analogous way to the previous one, although the in-game experience previous to the experiment phase is considered, as well as other key indicators as the number of total games owned, the types of owned games, the overall time the user spends playing and the total achievements the user has.

When considering the complete gaming experience of each subject, we can better characterize the way in which the games impact their soft skills. For instance, **Figure 20** shows how the achievements obtained during the experiment can be considered relative to the ones achieved beforehand. Thus, more experienced users are going to be more prone to get more achievements, and their achievements in the experiment must be relativized when using them in AI models.

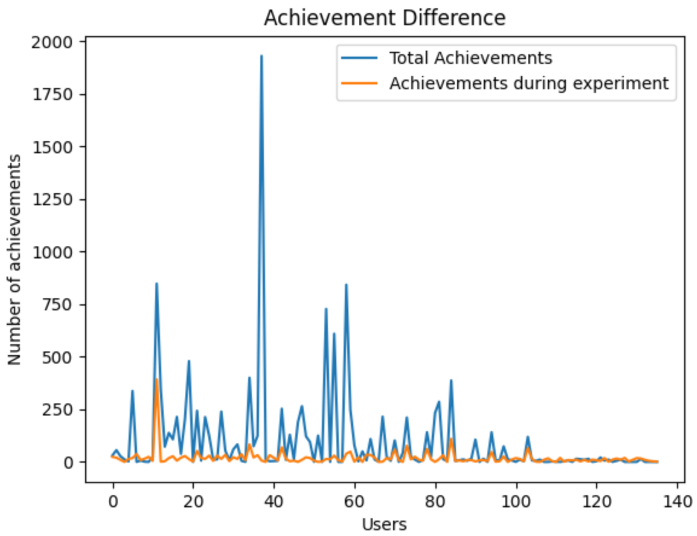

Figure 20 – Number of achievements obtained by each user during test phase (orange) and in total (blue)

When these achievements are considered in terms of volume, as detailed in **Figure 21** the distribution shows how the majority of users obtained a few achievements both during and prior to the experiment (most users obtained a single achievement during the experiment), reason for which they must be considered separately from the users that gather hundreds or thousands of achievements. The difficulty of obtaining achievements is not the same if the previous experience is regarded.

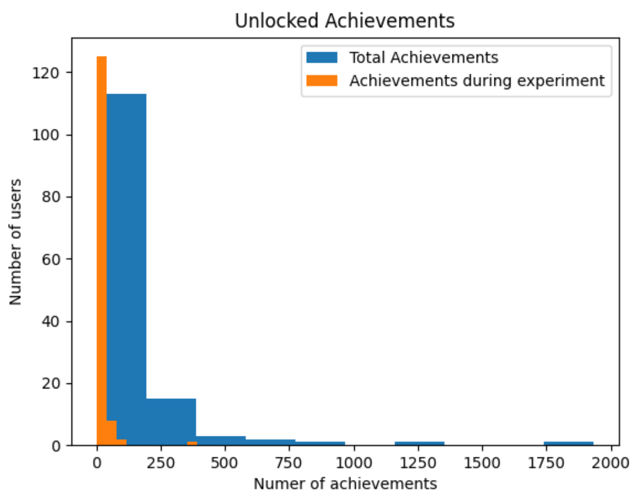

Figure 21 – Distribution of achievements obtained by users during test phase (orange) and in total (blue)

Lastly, an analysis similar to the one presented in previous section is shown in **Figure 22**, comparing the time played by different users in the experiment games but not only in the previous 2 weeks but in total (blue).

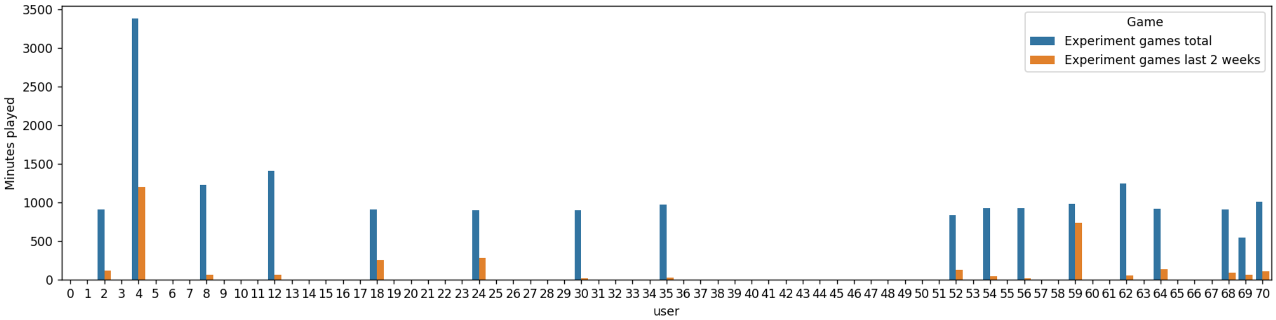

Figure 22 – Time played during the last 2 weeks and in total for the experiment games by users who played these two weeks

This plot shows that the time spent in these games is usually related to the time spent in them previously, and for subjects who did not spent time previously, it is very rare that they spend additional time during the experiment. This figure shows that subject 4 has a significant number of minutes played to one of the games in the experiment, which means that probably he/she already played before the experiment, thus the achievements of this user and the evolution of his/her soft skills should be relativized or their impact reduced. When building the AI models, all this experience related data must be carefully curated into datasets, so models do not learn biased relations between gaming experience and soft skills improvement.

4.1.4 Socio-demographic data

Socio-demographic data help establishing a common context baseline in order to cluster users and relativize their progress in soft skills to this context. More details can be checked in the deliverable *D3.2. Report on evaluation and training of soft skills with commercial video games.* This socio-demographic data were captured through a form in the beginning of the experiment and are stored in the following table: testresultssd.

The data captured with the questions of this form are summarized in the deliverable D3.2, where clear patterns can be observed. For instance, most users can be found in the "*under 30*" bracket of the population, entailing young profiles that in principle could be closer to gaming than older profiles. It is interesting that there is a proportion of users older than 60, which can also lead to analyse the relations of the soft skill improvements at this age, or the impact of age in this assessment. Regarding gender, there is a skew in the population towards male subjects, which should be considered in the AI models to avoid biases. Similarly, when the focus is set on the educational level of the subjects, most of them have a high education degree, and there is a strong skew in what regards having a PhD. The proportion of users with this kind of education does not reflect its usual proportion in the general population, reason for which this important variable should be carefully examined.

The type of game that is most played by users, as shown in **Figure 23**, reveals the type of achievement the users are more used to. For instance, shooting games will benefit eliminating enemies and aiming skills, while strategy games will promote long term strategy and planning. Not all gamers that are used to gaming develop the same skills, reason for which being able to cluster them considering the games they play more often is crucial to gauge the impact of the experiment games.

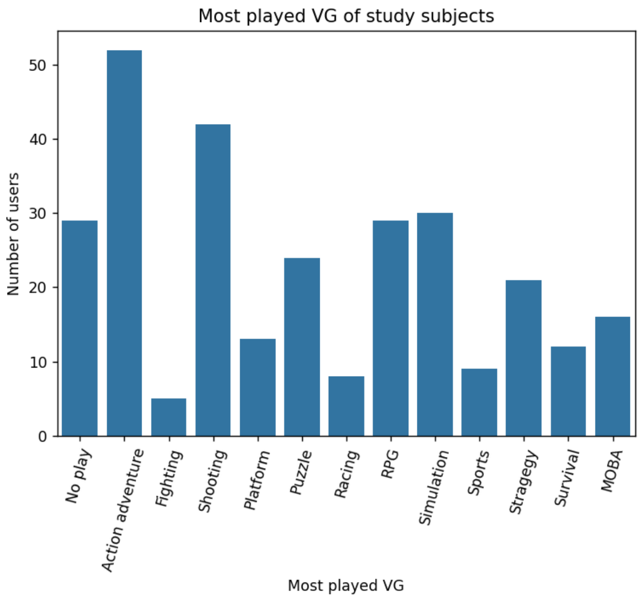

Figure 23 – Type of game most played by of users

Analogously, the subjective consideration of amount of time played by users (shown in **Figure 24**) is of paramount importance to separate users that have a long history of playing against users that play rarely or never. These data are relevant also in terms of comparison with the data obtained from Steam, as many subjects could have a longer gaming history than the one reflected by Steam, due to playing in other platforms or in consoles. The case of the previous experience out of Steam can only be captured thanks to this socio-demographic assessment.

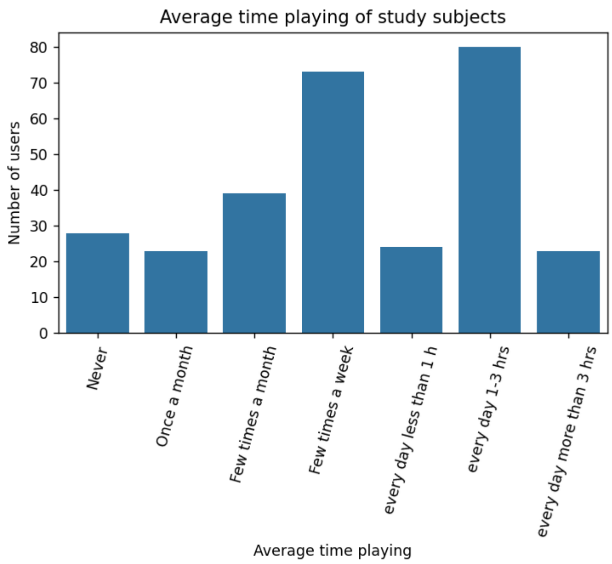

Figure 24 – Average time playing games of users

These data have been considered in a disaggregated way (each dimension separately), but as each of these 6 dimensions has a set of categories, combinations of them can render pre-configured profiles that help establishing different datasets for AI models (for instance, female subjects younger than 30 who play regularly mostly simulation or strategy games). This kind of analysis to observe the existing groups and their degree of vicinity will be performed during the AI modelling phase.

4.2. AI experiment design & methodology

As stated throughout this document, MEGASKILLS will employ AI models to leverage some aspect related to soft skills, such as assessing the impact of video game playing on specific soft skill development. For that, a huge amount of data, derived, among others, from each player's pre- and post-tests, their socio-demographic information and pre-experiment playing habits is required. The main objective is to observe how, through information gathered from various sources, one can **predict or infer the level of advancement of a specific soft skill in an individual**. While earlier sections have addressed the methodologies for data collection and storage, this section delves into the data analytics process, offering insights from the AI's perspective.

Figure 25 - Steps carried out in MEGASKILLS for the AI development (Generic overview)

In **Figure 25** the general data analytics process is depicted, serving as the foundation for all AI endeavours undertaken within MEGASKILLS. This process delineates a sequence of essential steps vital for the effective evolution of solutions. Subsequently, a more intricate, step-by-step analysis is conducted to delve deeper into each stage:

1) **Data exploration**: MEGASKILLS uses data collected from individuals, their steam stats, and tests. This initial stage is pivotal in gaining a comprehensive understanding of the collected data to formulate the most effective solution. Beyond simply acquiring knowledge of the data, it is imperative to discern its inherent characteristics and uncover any underlying patterns or trends. Identifying such insights plays a crucial role in shaping the development of impactful solutions within MEGASKILLS, facilitating informed decision-making and strategy formulation.

2) **Preprocessing**: In this crucial step, the collected data undergoes various transformations to prepare it for training AI models effectively. Within MEGASKILLS, several preprocessing techniques are employed to ensure the data's suitability for analysis. Feature selection is applied to identify the most influential variables that significantly impact the model's predictive accuracy. Simultaneously, data cleaning techniques are utilized to rectify errors and address missing values, ensuring the integrity and reliability of the dataset. Additionally, text processing methods such as tokenization, stemming, and lemmatization may be applied as needed to handle textual data appropriately. These preprocessing steps lay the groundwork for robust AI model training and contribute to the overall effectiveness of the solution developed by MEGASKILLS.

3) **AI Model**: This phase encompasses all aspects related to the development and training of AI models. It includes fine-tuning and parameter tuning to optimize the models' performance and adapt them to the data's complexities. Model training involves exposing the models to a large volume of data and iteratively adjusting their parameters to minimize errors and enhance performance. Subsequently, model validation evaluates the models' performance and generalization capabilities on unseen data, ensuring their effectiveness and identifying areas for improvement.

4) **Results/Statistics**: In this final step, the statistics generated by the AI models are carefully analysed to extract valuable insights and patterns from the data. By delving into the statistical results, trends and correlations that may not be immediately apparent can be uncovered. These insights serve as the foundation for informed decision-making, guiding the development of strategies and interventions to enhance soft skill development effectively.

One of the main objectives in MEGASKILLS is to replace some aspects of the tests conducted in the PRE and POST phases by AI models. As mentioned earlier, those tests are just one part of the total data collected in the project, information about previously played games, most played games, or socio-demographic information are of utmost importance in the definition of different player behaviours. A player behaviour could be represented by their tendency to play a certain game gender, by their interests or their wish to get most of the achievements of the games they play. Reasonably, players sharing similar behaviours are more likely to share same soft skills set. Thus, clustering players by their behaviours (see **Figure 26**)can be useful to replace some parts from the tests, by computing the distance to the cluster centroids a new player can be easily assigned a behaviour and, consequently, a set of soft skills linked to it. Methods such as **K-means**, which is computationally efficient and easy to implement, **DBSCAN**, which is robust to noise or **Spectral Clustering**, that can capture complex structures and it is not sensitive to outliers, are good candidates for MEGASKILLS, choosing among them depending on data characteristics. These techniques can analyse and detect patterns in the data, as depicted in **Figure 2** and assign new players to a known behaviour. Moreover, regression techniques, including **Logistic Regression** (LR), **Support Vector Regression** (SVR), **Random Forest** (RF) or **Gradient Boosting Regression** (GBR), can be leveraged to forecast the anticipated enhancement for a player, taking in account their behaviour and other related data.

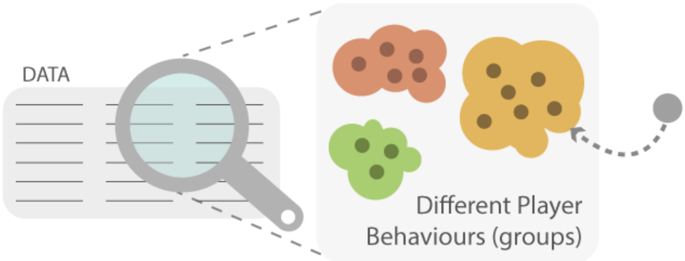

Figure 26 - Example of possible players' behaviour from the dataset using a clustering technique.

Beneath each game, a selection of achievements defines the most complex actions and challenges that players may encounter during gameplay. Some of these challenges require players to give their best to overcome them, and in some cases, a high level of a particular soft skill may be necessary to achieve success. Detecting and being able to correlate the completion of an achievement to a player that made a higher progress in the soft skill development is another interesting way that could be developed. Quite linked to that, it could also be predicted the soft skill to which an achievement could contribute the most. For that, Natural Language Processing (NLP) techniques could be employed, leveraging not only the text of an achievement but also extra game related information, such as steam tags. For that, a great effort should be put in designing a good embedding or techniques to generate that embedding, such as a fine-tuned BERT (based on transformers). Continuing in this line, a tool could be devised not only to forecast the soft skills associated with each achievement but also for entire games. Such a tool could serve as a **recommendation engine**, tailored to individual players based on their behaviour, guiding them towards the game's best suited for developing specific soft skills.

5. CONCLUSIONS

In summary, the MEGASKILLS project harnesses AI models to enhance soft skill development, utilizing a comprehensive data analytics process. This process involves data exploration to understand collected data, preprocessing to prepare it for analysis, AI model development and training, and analysis of results and statistics. The findings outlined in this deliverable could have several important implications:

- **Enhanced Soft Skills Assessment**: By leveraging AI methodologies and player behaviour analysis, organizations could improve their ability to assess soft skills in individuals without relying solely on traditional testing methods. This could lead to more comprehensive and nuanced evaluations of candidates' abilities.
- **Tailored Learning and Development**: The ability to recommend games for skill enhancement based on individual player profiles and achievements could revolutionize personalized learning and development strategies. Organizations could use this data to create tailored training programs that cater to specific skill gaps identified through gameplay.
- **Efficient Talent Management**: Clustering individuals based on soft skill proficiency using AI techniques allows for more strategic talent management. Employers can better match individuals with roles that align with their demonstrated strengths and areas for improvement.
- **Insights into Player Behaviour**: The use of NLP to analyse game achievements and correlate them with skill development provides valuable insights into how specific activities impact soft skill acquisition. This understanding could inform game design and learning content creation.
- **Advancements in AI-Based Skill Assessment**: The methodologies described could contribute to advancements in AI-driven skill assessment technologies beyond gaming contexts. This could lead to broader applications in education, recruitment, and workforce development, where soft skills are increasingly recognized as critical for success.

In summary, these findings suggest a transformative potential for AI-driven soft skills assessment, personalized learning interventions, and more effective talent management strategies, ultimately influencing how organizations approach skill development and recruitment.

Furthermore, as an AI application gains more users, its forecasting improves due to increased data volume and diversity, refining algorithms, and feedback loops. However, scalability challenges may arise, requiring efficient infrastructure handling. Continuous adaptation to evolving user behaviour ensures ongoing accuracy in predictions.

Regarding the following steps, once the data from video games and standard tests have been consolidated and transformed, machine learning algorithms will be implemented and trained. Depending on the amount and quality of historical data (training data set) different machine learning methodologies will be proposed and analysed based on the review carried out included in this deliverable. The need for one or more learning algorithms will be determined according to the soft skills selected in WP2 and the rubrics designed in WP3.

6. REFERENCES

Alloza, S., Escribano, F., Delgado, S., Corneanu, C., & Escalera, S. (2017). XBadges. Identifying and training soft skills with commercial video games. arXiv preprint arXiv:1707.00863.

Aviv, I., Barger, A., & Pyatigorsky, S. (2021, October). Novel machine learning approach for automatic employees' soft skills assessment: Group collaboration analysis case study. In 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS) (pp. 1-7). IEEE.

Azzini, A., Galimberti, A., Marrara, S., & Ratti, E. (2018). A classifier to identify soft skills in a researcher textual description. In Applications of Evolutionary Computation: 21st International Conference, EvoApplications 2018, Parma, Italy, April 4-6, 2018, Proceedings 21 (pp. 538-546). Springer International Publishing.

Barr, M. (2018). Student attitudes to games-based skills development: Learning from video games in higher education. Comput. Hum. Behav., 80, 283-294. https://doi.org/10.1016/j.chb.2017.11.030.

Baumann, F., Emmert, D., Baumgartl, H., & Buettner, R. (2018). Hardcore Gamer Profiling: Results from an unsupervised learning approach to playing behavior on the Steam platform. Procedia Computer Science, 126, 1289-1297.

Bracken, B. A., Bai, W., Fithian, E., Lamprecht, M. S., Little, C., and Quek, C. (2003). The test of critical thinking. *Center for Gifted Education, College of William y Mary*.

Britton, B. K., & Tesser, A. (1991). Effects of time-management practices on college grades. Journal of educational psychology, 83(3), 405.

Burke, C. S., Pierce, L. G. y Salas, E. (Eds.). (2006). Understanding adaptability: A prerequisite for effective performance within complex environments. JAI Press.

Chatziantoniou, N., Politopoulos, N., & Stylianidis, P. (2018). Designing and developing an educational game for leadership assessment and soft skill optimization. In Interactive Mobile Communication Technologies and Learning: Proceedings of the 11th IMCL Conference (pp. 258-265). Springer International Publishing.

Claesdotter-Knutsson, E., André, F., & Håkansson, A. (2022). Gaming activity and possible changes in gaming behavior among young people during the COVID-19 pandemic: cross-sectional online survey study. JMIR Serious Games, 10(1), e33059.

Fareri, S., Melluso, N., Chiarello, F., & Fantoni, G. (2021). SkillNER: Mining and mapping soft skills from any text. *Expert Systems with Applications*, *184*, 115544.

Harirchian, M., Amin, F., Rouhani, S., Aligholipour, A., & Lord, V. A. (2022). AI-enabled exploration of Instagram profiles predicts soft skills and personality traits to empower hiring decisions. *arXiv preprint arXiv:2212.07069*.

Kannan, R., Jet, C. C., Ramakrishnan, K., & Ramdass, S. (2023). Predicting Student's Soft Skills Based on Socio-Economical Factors: An Educational Data Mining Approach. JOIV: International Journal on Informatics Visualization, 7(3-2), 2040-2047.

Kizilkaya, G., and Askar, P. (2009). The development of a reflective thinking skill scale towards problem solving. *Egitim ve Bilim, 34*(154), 82.

Loria, E., Antelmi, A., & Pirker, J. (2021, August). Comparing the structures and characteristics of different game social networks-the steam case. In 2021 IEEE Conference on Games (CoG) (pp. 1-8). IEEE.

Nikolaou, I., Georgiou, K., & Kotsasarlidou, V. (2019). Exploring the Relationship of a Gamified Assessment with Performance. The Spanish Journal of Psychology, 22. https://doi.org/10.1017/sjp.2019.5.

O'Neill, M., Vaziripour, E., Wu, J., & Zappala, D. (2016, November). Condensing steam: Distilling the diversity of gamer behavior. In Proceedings of the 2016 internet measurement conference (pp. 81-95).

Payan-Carreira, R., Sacau-Fontenla, A., Rebelo, H., Sebastião, L., & Pnevmatikos, D. (2022). Development and Validation of a Critical Thinking Assessment-Scale Short Form. *Education Sciences, 12*(12), 938. https://doi.org/10.3390/educsci12120938

Ramos-Pulido, S., Hernández-Gress, N., & Torres-Delgado, G. (2023, February). Analysis of Soft Skills and Job Level with Data Science: A Case for Graduates of a Private University. In Informatics (Vol. 10, No. 1, p. 23). MDPI.

Rizani, M. N., Khalid, M. N. A., & Iida, H. (2023). Application of meta-gaming concept to the publishing platform: Analysis of the steam games platform. Information, 14(2), 110.

Seisdedos, N. (2004). CAMBIOS Test de Flexibilidad Cognitiva. Madrid: TEA Ediciones.

Shute, V., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. Comput. Hum. Behav., 63, 106-117. https://doi.org/10.1016/j.chb.2016.05.047

Stafford, T., & Vaci, N. (2021). Digital games as a platform for understanding skill acquisition from novice to expert. PsyArXiv [Preprint]. doi, 10.

Stafford, T., & Vaci, N. (2022). Maximizing the potential of digital games for understanding skill acquisition. Current Directions in Psychological Science, 31(1), 49-55.

Viviers, H., Fouché, J., & Reitsma, G. (2016). Developing soft skills (also known as pervasive skills): Usefulness of an educational game. Meditari Accountancy Research, 24, 368-389. https://doi.org/10.1108/MEDAR-07-2015-0045.

Wings, I., Nanda, R., & Adebayo, K. J. (2021). A context-aware approach for extracting hard and soft skills. *Procedia Computer Science*, *193*, 163-172.

Zhao, M., Javed, F., Jacob, F., & McNair, M. (2015, January). SKILL: A system for skill identification and normalization. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 2, pp. 4012-4017).