Eindhoven University of Technology

BACHELOR

Exploring Individual Differences and Item Characteristics in the Dutch Famous Faces Test for Alzheimer's Disease

Snoeren, Eke

Award date:
2023

Disclaimer
This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

Eindhoven University of Technology

Bachelor Thesis

Exploring Individual Differences and Item Characteristics in the Dutch Famous Faces Test for Alzheimer's Disease

Author: Eke Snoeren (1645773)

supervisor: Dr. Katrijn Van Deun

June 18, 2023

Abstract

The famous faces test, FFT, is a cognitive test which can be used to recognize Alzheimer's disease. For this study 338 participants completed the dutch famous faces test, D-FFT, which is a variation of the FFT. The test consisted of recall, where healthy older adults were asked to recall the name of a famous face (an item) presented to them, and recognition, where participants were asked to pick the right name out of four options of the item presented to them.

The influence of individual differences of participants, and possible interaction with item characteristics, were researched. Prior research on the FFT, primarily focused on comparing differences between groups of healthy and at-risk older adults, where no significant differences were found, instead of individual differences within a group.

Methods employed were the exploration of clusters in the performance of the participants and the participant-item combinations. After which the data was modeled using a decision tree to find which individual differences and interactions between individual differences and item characteristics on the performance of participants on the D-FFT. Results were two-folded, first the influence of individual differences on the total score on the recall part of the D-FFT was found to be inconclusive, as the decision tree did not have great performance. The interaction between participant differences and item characteristics had more reliable results. Features which were found to be most important mainly consisted of item characteristics and interaction between interests of participants and categories of items were confirmed.

Voor mijn opa, die me altijd zo goed begrijpt en er voor me is, zelfs als hij het moeilijk heeft.

Acknowledgements

I would like to express my heartfelt gratitude to my thesis supervisor, Dr. Katrijn van Deun, for her invaluable support and guidance throughout the entire process. Her exceptional mentorship and prompt and comprehensive responses to all of my inquiries greatly facilitated the writing of my thesis. Additionally, I am deeply grateful to her for providing me with the opportunity to work on this significant project focusing on Alzheimer's disease, a cause that holds immense personal significance for me.

I would also like to extend my sincere appreciation to Dr. Ruth Mark and Evi van den Elzen for their invaluable feedback on my thesis report and the insightful comments they shared during our research group meetings. Their input greatly enhanced the quality of my work and contributed to a more comprehensive understanding of the subject matter.

Contents

List of Figures

List of Tables

1

Introduction

The number of people with dementia has been growing for decades, from 50.000 in 1950 to 290.000 nowadays. [1] This number is only expected to grow with the aging of the general population. With Alzheimer's being the leading growing cause of death in the Netherlands [1], the diagnosis of the disease should be a priority. Currently 69% of caregivers indicate that the process of diagnosing took over a year, which is excessively long, especially when it is kept in mind that Alzheimer's disease starts years before a diagnoses is made. [1]

The preclinical stage of Alzheimer's disease is the first stage of Alzheimer's, in which changes in the brain occur, while the patient still only has small signs of the disease like subtle episodic memory loss.[2] The duration of the preclinical stage varies depending on factors as age and sex, for a 70 year old the preclininal stage of Alzheimer's Disease was estimated to take 10 years.[3] It is the stage which comes before the predromal stage, in which the symptoms become more noticeable and diagnosis becomes easier.[2]

Diagnosis of this preclinical stage is hard as the patient only shows limited signs.[4] Being able to identify the disease as early as possible can be crucial for treatments being used now, and any that will be developed in the future. Therefore being able to recognize and diagnose Alzheimer's in early cognitive decline is of great importance.

These early cognitive declines not being recognized early on is caused by current instruments used in clinic not being able to recognize the preclinical Alheimer's Disease. Due to the limited outward symptoms, cognitive test are not able to identify Alzheimer's Disease in this early stage.[3] The criteria for clinical Alzheimer's disease in the DSM-5 include: decline in memory and learning, as well as steady gradual cognitive decline. It must also be ruled out that the symptoms named above are not caused by another disease. [5]

Currently in the Netherlands, the discovery and diagnosis of Alzheimer's disease takes multiple steps. Starting with a cognitive test, one of which is the mini mental status exam (MMSE) test, at the general practitioner, together with other tests to rule out other possible causes. After which, when the General Practitioner suspects Alzheimer's disease, they will have a conversation with a person close to the patient. If this confirms their suspicion, the patient is referred to the hospital for further investigation.[6] The MMSE test is a general cognitive screening, a higher score indicates a better cognitive function and a lower score can indicate dementia.[7] In this test a personalized norm is used for age and education level as this influences the MMSE score.[8] The downsides of the MMSE test are lack of sensitivity for early signs of dementia, measurement of different memory functions is limited and repetitive use of the test may produce a practice effect (the patient gets better scores at the test, when subjected to it multiple times), especially in cases of mild Alzheimer's disease.[9]

The famous faces test (FFT) is a cognitive test which can be used to recognize Alzheimer's disease, where the performance can be assessed in two different ways: recall and recognition. During recall, the patient is shown a picture of a famous person and asked to name this person. During recognition, the patient is shown a picture of a famous person, however this time they are also presented with four possible names (these might be names of other famous people or fake names) and asked to point out the name of the famous person shown to them.

In this test semantic memory (information learned over time) is used by the participant and tested by the FFT. Decline in semantic memory is a symptom associated with Alzheimer's disease.[10] The famous faces test is unique, because it does not have a learning phase; it can't be pinpointed when the patient learned the information that is tested. As well as the famous faces test testing the semantic memory of participants over decades, which is not the case for other semantic memory tests.

In the famous faces test a time-limited temporal gradient was found, meaning that healthy older adults and older adults with declining cognitive functions were able to recognize and recall famous faces from decades ago better than recently famous faces (Orlovsky et al, 2018).[11] In this research paper, the finding also suggested that semantic measures (which are measured during the famous faces test) may be more related to cognitive decline and possible early Alzheimer's disease in healthy older adults. This proves the tests relevance.

Evidence of the famous faces test (FFT) as an instrument for early detection of Alzheimer's dementia (AD) was found in a study associating performance on the FFT with the volume of the hippocampal area of the brain. The brain integrity goes through alterations prior to manifestation of the cognitive symptoms associated with AD. The entorhinal cortex and the hippocampal areas of the brain are one of the first parts affected by Alzheimer's disease. In Seidenberg et al., 2013,[12] healthy older adults were subjected to the famous faces test twice in an 18-month period and divided in two groups depending on if their test score remained stable or declined. The stable and declining group both recognized famous faces from decades earlier in high rates, however the declining group performed worse in recognizing recent faces, here a temporal gradient was found. The declining group had smaller baseline hippocampal volume compared to the stable group, this is in line with hippocampal area shrinking in the early stages of Alzheimer's disease.

Other studies have also been done where two groups of elders, one with physical symptoms/qualities associated with early-stage Alzheimer's disease and the other consisting of healthy older adults, are compared in results for recall and recognition of famous faces. All of these find similar results to Seidenberg et al., 2013, namely that the healthy older adults are better at recalling/recognizing recent famous faces than the group with symptoms associated with early-stage Alzheimer's. The performance on recognizing famous faces from earlier decades are the same for both groups.[11] [13] [14]

From these papers it can be concluded that a patient's performance on the famous faces test can be used to identify early stages of Alzheimer's disease. With a focus on the performance on recognizing recent famous faces in particular.

For this research paper the D-FFT was used, the dutch famous faces test, which consisted of recall and recognition of famous faces (items) by participants. For recall for every correct name, points are administered, a differentiation in points might be made between full names and partial names/nicknames. The famous faces that are shown are from different decades and from different categories (ex. politicians, movie stars, athletes). A benefit of the D-FFT is that, in contrast to the MMSE test, there is no practice effect ~~was found~~ for the D_FFT and thus can be administered an unlimited number of times (with new faces) without the patient having a better score due to practice instead of cognitive capabilities. An in-depth explanation on how the test was administered, is provided in section 3.

When the FFT is administered in clinical trials, the individual differences are not taken into account. While it is expected that these differences are influential on the scores of individuals taking the test. For example, if an individual has a personal interest in sports throughout their life, they most likely have a higher probability of recognizing famous athletes. Currently it is unknown if these individual differences influence the scores, and if they do how much. Were the test to be used in a clinical setting, it is critical to recognize if individual differences influence the performance on the test, as to possibly install personal norms of expected performance on the participants performance on the D-FFT. These individual differences of the participants will also be linked to the characteristics of the faces shown to them.

2

Background & Research Questions

2.1 Background

Numerous research papers have underscored the potential of the Famous Faces Test in detecting early cognitive decline associated with Alzheimer's disease, as was highlighted in the introduction. These studies have also investigated differences in test performance with respect to individual characteristics. This topic is particularly relevant as it pertains to the exploration of personalized norms for the test. Notably, the current test employed by general practitioners in the Netherlands, the MMSE test, incorporates personalized norms considering age and education level.[8]

In the research paper Orlovsky et al. (2018)[11], the authors compared demographics such as sex, age, years of education, as well as media usage over the decades between a healthy group and a group with an accumulation of amyloid beta protein, which is associated with the development of Alzheimer's disease. The findings indicated no significant difference between the two groups.

When comparing ~~elders~~ who exhibited stable performance on the famous faces test with those who experienced a decline in performance, in Seidenberg et al. (2013).[12] Similarly, no significant differences in age, education, gender distribution, or the time interval between the baseline and follow-up tests when comparing the two groups. This finding follows through in the research paper Hays et al. (2017). Here age, gender, years of education, cognitive status, and health status did not have a significant effect on the temporal gradient.[13]

Overall, across different research papers involving the famous faces test, no significant differences have been found between groups of healthy older adults and older adults exhibiting symptoms associated with early-stage Alzheimer's disease in terms of individual differences. This includes demographic factors (age, gender, years of education), health-related statistics, and cognitive status.

This study seeks to distinguish itself from previous studies by conducting an in-depth analysis of individual differences within a group. Unlike prior research that primarily compared differences between healthy and at-risk older adults, this study focuses exclusively on a sample of healthy older adults. Significantly, this investigation will account for differences at the level of individual participants and items and explore how these variations impact overall test performance and the recognition of specific famous faces.

2.2 Research Questions

This study aims to research the potential need for personalized norms for the D-FFT. Existing literature reveals a lack of significant differences for the FFT between groups with and without dementia/Alzheimer's disease, when it comes to sex, age, education, media usage and cognitive and health status. Therefore, this project seeks to investigate between-person differences in the performance of the Famous Faces Test. These variables consist of demographics, lifespan interests, general health and a variety of indexes & scales. Consequently, the following research questions are proposed:

RQ1: How do the individual characteristics of a healthy older adult influence their overall performance on the FFT?

Hypothesis for RQ1: Individual differences will not exert an influence on the overall score of the famous faces test. Background research indicates that demographic, interests, and health factors, which are the data available for this research, seem to have no impact on test performance.
Approach for RQ1: To answer the research question, first exploring of the data will be performed. This will be done by hierarchical clustering. Here data is split based on distance, similar performances on the test will be clustered together. After which the individual differences of participants will be used to predict which cluster the participant belongs to.

After which the entire dataset will be modelled using a Decision Tree Regressor, here the total test score will be predicted by using variables on demographics, lifespan interests, general health and a variety of indexes & scales. With a decision tree the exact splitting criteria (on which variables the model splits) can be seen. Due to the interpretability, it can be observed exactly what causes the model to predict a participant to have a higher or lower test score.

RQ2: How do the descriptive features of items presented to participants in the FFT interact with their individual characteristics, and what is the impact of these interactions on their ability to accurately recall the presented items?

Hypothesis for RQ2: Individual differences and characteristics of famous faces are interrelated. Factors such as personal interests may affect the recognition of famous faces. For example, with individuals who possess a fondness for sports potentially exhibiting better recognition of athletes compared to others. However, demographics and health factors are not expected to have any influence.
Approach for RQ2: The focus of this research will be on the recall part of the D-FFT for reasons explained in section 4.2.3. Similar steps will be made for research question 1, will be applied here. First the data will be explored by clustering it on similarity. However here bi-clustering is employed as the method. Bi-clustering gives the ability to group together clusters of famous faces and individuals based on those participants ability to get the particular famous faces correct. After which the data will be modeled on using a Decision Tree Classifier. Here variables of participants and famous faces will be used to determine if a participant got a particular famous face correct on the test. This will help to understand the relationship between the individual and the famous face presented to them.

3

Data

3.1 Famous Faces (Items)

The data used for this research project incorporates a famous faces test which consisted of photos of 220 famous faces. These photos were distributed into 12 sets, two practice sets (P1 and P2) and 10 regular sets (numbered 1 to 10). P1 and P2 consisted of 10 photos, while the regular sets consisted of 20 photos. The test was divided into two sessions, in each session the participant was first presented with either P1 or P2, followed by 5 of the regular sets. To make identification easier, each famous face is associated with a unique code denoted as Cxxxx, where "x" represents a numerical value.

For each of the famous faces, attributes were recorded, including the decade in which they achieved fame, their gender, whether they are of Dutch or international origin, and their respective category. The decades span from the 1960s to the 2010s, while the categories encompass Politics, Singers & Musicians, Film & Theatre, and Sports.

It is essential that the famous faces selected for this test achieved prominence primarily within a specific decade, thereby avoiding any overlap. This aspect holds significant importance for the purpose of the test, as the recognition of a temporal gradient necessitates the precise categorization of famous faces within their respective decades. Failure to assign a famous face to a specific decade could lead to ambiguity regarding which decade the participant may recognize the famous face from. To ensure objectivity and minimize potential biases, the photographs used in the test were standardized without the inclusion of any props or hints that could guide or influence participants' responses.

3.2 Participant Data

It is worth mentioning that all participants were required to complete a set of self-report questionnaires pertaining to various aspects of their personal information. The first category of questions encompassed demographic, household, and retirement details, capturing information such as age, gender, educational level, whether they have children/grandchildren, and retirement status. These variables were encoded using specific values, for instance, gender was encoded as 1 for Male, 2 for Female, 3 for Other, and 4 for Prefer not to say.
The second category comprised a series of scaled questions pertaining to `Lifespan interests`. Participants were asked to rate their level of interest, on a scale ranging from 1 (least) to 10 (most), in eight distinct interest categories, namely Film & Theatre, Music, Politics, Sports (general), Ice skating, Tennis, Soccer, and Cycling. These ratings were collected for three different age categories: interest during young adulthood (around their 20s), interest during middle age (around their 40s), and current interests.
Additional inquiries were made regarding participants' general health, encompassing factors such as medication usage, duration of any medical conditions, body weight, smoking and alcohol consumption habits, cholesterol levels, and blood pressure. The Charlson Comorbidity Index was utilized to assess the presence of various illnesses, with a higher total score indicating a greater number of illnesses. The Lawton Instrumental Activity of Daily Living scale measured participants' ability to independently perform daily tasks, where a higher total score reflected greater independence in daily life. Participants were also questioned about their engagement in physical activities. The Pittsburgh Sleep Quality Index was used to evaluate sleep quality, where a higher total score indicated lower sleep quality, and assessed for cognitive failures, where a higher total score indicated a greater number of cognitive failures. Furthermore, participants' social involvement, feelings of loneliness, and symptoms of depression were also addressed in the questionnaires.

3.3 Recall Data

Within each session, consisting of five regular sets, participants were tasked with recalling names associated with the presented famous faces. Prior to the regular sets, participants were shown either P1 or P2 as a practice set. Subsequently, they were instructed to provide their full names when able to do so. Participants were explicitly prohibited from utilizing the internet or seeking assistance from others. The study involved two sessions in total.

During the recall process, participants were initially allotted 15 seconds to indicate their familiarity with the presented individual, with response options categorized as follows: 1=Yes, 2=Yes, but it is on the tip of my tongue, 3=No, but I do recognize the face, and 4=No, and I do not recognize the face. Following this assessment, participants were granted one and a half minutes to both provide the name corresponding to the face and express their confidence level in their answer. Confidence levels were measured using the following scale: 1=I guessed, 2=I am in doubt, and 3=I am sure. If participants did not fill in an answer within the given time frame the answer was left blank.

The names provided by participants were subsequently encoded to determine whether they correctly recalled the name or not. Additionally, a separate value was assigned to indicate whether the response was incorrect, partially correct, or fully correct. The reaction time, representing the duration taken by participants to fill in their responses, was also recorded.

Upon completion of all the sets, the total number of responses, number of correct responses, and the number of faces participants claimed to recognize were counted for each participant. Furthermore, the relative number of correct recalls, expressed as the percentage of correct responses out of all the answers provided, was computed for each participant.

3.4 Recognition Data

Following the completion of the recall phase involving sets of famous faces, participants were subsequently presented with the same faces again. However, this time they were provided with four options and tasked with selecting the correct name corresponding to each famous face. Similar to the previous phase, the participants were presented with a set of 10 practice faces, after which the formal session with the regular sets started.

For each question, participants were allotted a duration of 20 seconds to provide their answer. Following their response, participants were queried regarding their level of certainty regarding the accuracy of their chosen answer. In instances where no response was provided by the participant, the corresponding answer was left blank. Additionally, variables such as the position of the correct name on the answer sheet and reaction time were measured.

Upon completion of all sets, the total number of responses given by each participant and the number of correct responses were recorded. Moreover, the relative number of correct recognitions, expressed as the percentage of accurate responses out of all the answers provided by each participant, was calculated.

4

Methodology

4.1 Data Preprocessing

4.1.1 Famous Faces (Items)

The famous faces dataset consisted of information on the famous faces and their photographs. To preprocess the famous faces dataset for this research, columns that were deemed irrelevant were initially removed. Specifically, the columns pertaining to the photograph's origin, quality, and creator, namely `Qualtrics URL`, `Image URL`,`photo_year`, `photographer`, `license`, were dropped. Additionally, `ID_new` was excluded as it merely assigned numbers from 1 to 220 without providing any substantial information. Furthermore, `no_decades` was eliminated since the column `decade` already indicated the specific decade to which the photograph belonged in a more explicit manner.

Following these exclusions, the dataset consisted of the following columns: `ID` (representing the item number), `list` (indicating the set to which the item belonged), `name` (denoting the name of the famous person in the photograph), `decade` (indicating the decade during which the famous person achieved prominence), `gender` (representing the gender of the famous person), `international` (indicating whether the famous person was of Dutch or international origin), `category_main_tekst` (indicating the category to which the famous person belonged), `lure_1`, `lure_2`, `lure_3` (representing the alternative options provided to participants in the recognition question). Importantly, none of these columns contained missing values.

The columns `decade`, `gender`, and `international` originally were the data type `integer`; however, they were converted to string values that corresponded to the respective numerical values in the columns. For instance, the values 1 and 2 in the `gender` column were replaced with 'Male' and 'Female', respectively. This conversion was carried out to ensure that these columns were treated as categorical variables during the exploration and modeling of the data, rather than as numerical variables.

Finally, in preparation for data exploration and modeling, the items from the two practice sets were removed from the dataset. These sets were intended for participants to familiarize themselves with the task and may not reflect their full capacity to answer the questions accurately. It is plausible that participants were still adjusting to the question format or not fully focused on providing correct responses for these items. The IDs of the excluded items were recorded in a separate list to ensure their removal from the famous faces test (FFT) dataset, which contains recall and recognition data.

4.1.2 Participant Data

In regard to the participant data, repetitive columns containing duplicate information were removed. For instance, `gender_session2` was dropped since the same data was already stored in the `gender` column. Prior to discarding these columns, a check was conducted to ensure the consistency of the data in both columns.

Furthermore, columns containing open-ended responses were excluded from the dataset, such as `birthcountry_4_TEXT`. These columns posed challenges for modeling purposes due to the unique and individualized nature of the answers. The values in these columns were largely distinct (with counts ranging from 1 to 3). Additionally, these columns contained a high proportion of missing values, ranging from 31.4% to 100%.

Columns that contained only a single value were also eliminated. For example, the column `medication_type_36`, where all participants responded with 'no', was dropped. Subsequently, certain columns that originally contained empty values represented by a space (' ') were converted to NaN values for numerical columns. For string columns, the missing values were retained as ' '. The data types of the columns were carefully examined, and categorical columns with the data type 'integer' were converted to strings as necessary (e.g., `gender` column in section 4.1.1).

Considering the extensive number of columns, the dataset was slimmed down by retaining only the total score columns for specific question categories. These categories included the Charlson Comorbidity Index, Lawton Instrumental Activity of Daily Living, Physician-based Assessment and Counseling for Exercise, Pittsburgh Sleep Quality Index, Cognitive Failures Questionnaire, De Jong-Gierveld Loneliness Scale, and Geriatric Depression Scale. As these scales are well-established, using only the total scores was deemed justified.

For logically connected columns, consistency checks were performed to ensure participants provided coherent answers. For example, if a participant indicated having children/grandchildren, they should not have entered '0' in the number of children/grandchildren field. Similar checks were applied to participants who reported not having children/grandchildren but provided a number greater than zero in response to queries about the quantity of children/grandchildren. No inconsistencies were detected. Subsequently, for participants who indicated not having children/grandchildren, the corresponding columns for the number of children/grandchildren were imputed with a value of 0 (if the column was not previously filled in).

The columns for weight and length were combined into one new column `BMI` and the columns for weight and length were dropped. BMI was computed as follows, where weight is in kilograms and length is in meters:

$$BMI = \frac{weight}{length^2}$$

Finally, columns were assessed for the overall extent of missing data. Given the total number of participants being relatively small (338 participants in total), if a column exhibited a missing data percentage of 30% or higher, it was dropped from the dataframe due to the substantial loss of valuable information. No columns exceeded this threshold, and therefore no additional columns needed to be dropped from the dataset.

4.1.3 FFT Dataset (Recall and Recognition)

The retained column for individual item recall in order to explore and model the data was `Cxxxx_recallCor`, which takes a value of 1 if the participant's answer was fully correct and 0 if incorrect. Conversely, other columns pertaining to the recall of individual items were dropped as they did not serve the research objectives.

To provide a comprehensive overview of participants' recall performance, the following columns were retained: `responses_recall` (indicating the total number of responses provided by each participant during recall), `recallCor` (representing the absolute number of faces for which the participant correctly filled out the first and/or last name), and `recallCor_rel` (the number of faces for which the participant correctly filled out the first and/or last name, relative to their total responses during recall).

Similarly, the preserved column for individual item recognition, intended for exploration and modeling of the data, was `Cxxxx_recog`. In this column, a value of 1 corresponded to the correct name chosen by the participant, while values 2, 3, and 4 corresponded to the lure names. As with recall, the remaining columns pertaining to the recognition of individual items were removed as they were not relevant to this research.

A new column named `Cxxxx_recogCor` was created for each item in recognition, with a value of 1 denoting a correct answer and 0 indicating an incorrect answer.

To offer a comprehensive understanding of participants' recognition performance, the following columns were retained: `Responses_recog` (representing the total number of responses provided by each participant during recognition), `recogCor` (representing the absolute number of faces for which the participant chose the correct name), and `recogCor_rel` (the number of faces for which the participant chose the correct name, relative to their total responses during recognition).

Lastly, the items included in the practice sets were removed from the FFT dataset for both the recall and recognition columns pertaining to individual items.

4.1.4 Matrix for Clustering

The FFT dataset was reformatted to generate two matrices specifically tailored for clustering purposes: one matrix for recall and another for recognition. Figure 4.1 displays the shape of these matrices. Here, the rows represent the participants' IDs, while the columns correspond to the item names (Cxxxx). The elements within the matrices are binary, with 0 indicating an incorrect response by the participant and 1 denoting a correct recall or recognition of the famous face. These scores were extracted from the `Cxxxx_recallCor` column for the recall matrix and the `Cxxxx_recallCor` column for the recognition matrix. The matrices were designated as `cluster_recall_matrix` and `cluster_recog_matrix` for recall and recognition, respectively.

It is important to note that not all participants provided responses for every item within the allotted timeframe, resulting in the presence of NaN values in the matrices. To accommodate the bi-clustering method, which does not handle NaN values, these missing values were substituted with 0s. This substitution was justified by the fact that the clustering analysis primarily focuses on the correct values (1s), while the missing values constitute a minor portion of the matrices (2.0% for recall and 1.2% for recognition).

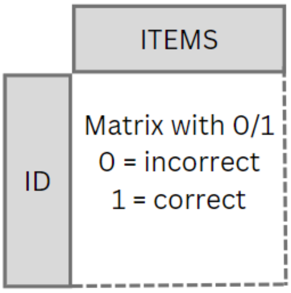


Figure 4.1: Shape of matrices `cluster_recall_matrix` and `cluster_recog_matrix`.

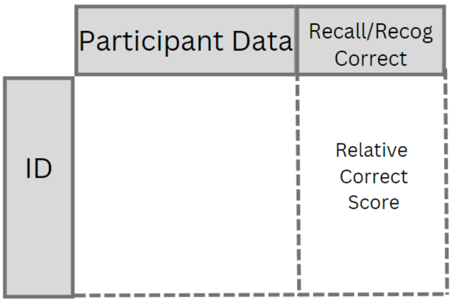


Figure 4.2: Shape of matrices `merged_ID_recall` and `merged_ID_recog`.

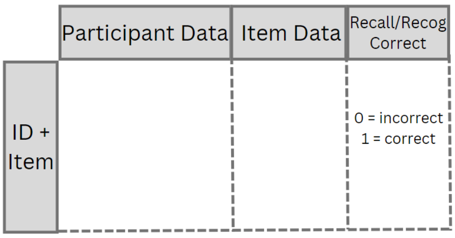


Figure 4.3: Shape of matrices `merged_ID_item_recall` and `merged_ID_item_recog`.

4.1.5 Merged Dataframes

In order to establish connections between clusters and either participant-specific or participant-item data, as well as to model the data, the dataframes needed to be reshaped. Various approaches were employed to merge the participant and item data with the recall or recognition scores, resulting in the creation of four new dataframes.

Figure 4.2 presents a visual representation of the structure of the first two dataframes, namely `merged_ID_recall` and `merged_ID_recog`. These data frames have an index consisting of participant IDs, with participant data columns forming the remaining columns. Additionally, a column indicating the relative correctness score (`recallCor_rel`/`recogCor_rel`) is included. Separate dataframes were created for recall and recognition.

Figure 4.3 illustrates the shape of the last two dataframes, denoted as `merged_ID_item_recall` and `merged_ID_item_recog`. These dataframes possess an index composed of two columns: participant ID and item. Consequently, each combination of participant and item has a distinct row in the dataframe. The columns comprise participant data associated with the participant ID and item-specific data corresponding to the item. Notably, a column indicating whether the participant answered a particular item correctly was established by using the `Cxxxx_recallCor` and `Cxxxx_recallCor` variables for recall and recognition, respectively. A value of 1 denotes a correct response, while a value of 0 signifies an incorrect response by the participant for the specific item.

It is important to acknowledge that for `merged_ID_item_recall` and `merged_ID_item_recog`, rows were omitted if the recall or recognition correct value was NaN. This decision was made, as these values could not be modelled on. Additionally, item columns related to recognition (`lure_1`, `lure_2`, and `lure_3`) were excluded from the recall dataframe, as they are irrelevant to the recall data analysis.

4.2 Data Statistics

In order to enhance the understanding of the data and facilitate the subsequent exploratory and modeling processes, a preliminary statistical analysis was conducted for each of the datasets.

4.2.1 Famous Faces (Items)

For the item data, table 4.1 was created, which presents the quantitative distribution of various columns, both for the entire dataset and for each individual subset. The analysis reveals that there are uneven distributions across several attributes, including decade, gender, international status, and category. Specifically, the decades of the 1990's and 2000's exhibit underrepresentation with only 10 and 26 items, respectively, whereas the remaining decades have a more balanced distribution with 30-50 items each.

Regarding gender, there is an overrepresentation of male famous faces. Similarly, Dutch famous faces outnumber their international counterparts. In terms of category, the majority of famous faces belong to the fields of politics and Singers & Musicians, constituting 70 and 60 items, respectively. Conversely, Sports and Film & Theatre categories have relatively fewer representations, with 30 and 40 items, respectively.

When comparing the 10 different subsets, minimal variations in the distribution of item characteristics are observed. These variations are consistent across all the analyzed columns.

Table 4.1: The count of variables for each column of the item dataset.

Column	Value	All	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
Decade	1960	48	5	5	5	4	5	4	5	5	5	5
	1970	32	3	3	3	4	3	4	3	3	3	3
	1980	40	4	4	4	4	4	4	4	4	4	4
	1990	10	1	1	1	1	1	1	1	1	1	1
	2000	26	2	3	2	3	2	3	3	3	3	2
	2010	44	5	4	5	4	5	4	4	4	4	5
Gender	Male	153	14	16	15	14	14	17	16	14	17	16
	Female	47	6	4	5	6	6	3	4	6	3	4
International	Dutch	113	11	13	13	11	11	12	11	11	12	12
	International	87	9	7	7	9	9	8	9	9	8	8
Category	Politics	70	7	7	7	7	7	7	7	7	7	7
	Singers & Musicians	60	6	6	6	6	6	6	6	6	6	6
	Film & Theatre	40	4	4	4	4	4	4	4	4	4	4
	Sports	30	3	3	3	3	3	3	3	3	3	3

4.2.2 Participant Data

The descriptive analysis of the participant data primarily focused on demographic characteristics and interest statistics.
In terms of gender distribution, the dataset predominantly consisted of female participants (203) compared to males (135). The education level of the participants was examined and represented in Figure 4.4, where the encoded values were interpreted as follows:

1. L̄ess than 6 years of primary education
2. = Finished primary education
3. = Primary education and less than 2 years of low-level secondary education
4. = Finished low-level secondary education
5. = Finished average-level secondary education
6. = Finished high level secondary education
7. = University degree

Based on this information, it can be concluded that a majority of the participants possessed at least an average-level secondary education or higher, indicating a highly educated sample.

The age distribution of the participants ranged from 60 to 90 years, with a predominant concentration between 66 and 74 years, this is represented in Figure 4.5.

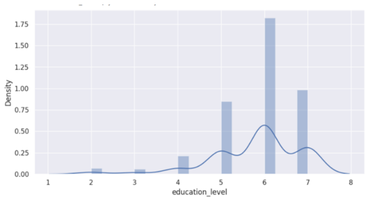


Figure 4.4: Distribution of the education level for participants.

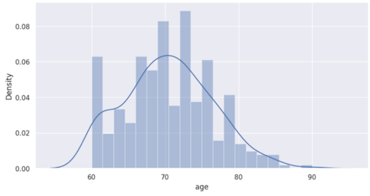


Figure 4.5: Distribution of the age for participants.

Regarding the participants' mother tongue, nearly all respondents (336 out of 338) indicated Dutch as their primary language, while two participants reported English. Similarly, Table 4.2 reveals that the majority of participants were born in the Netherlands.

Table 4.2: Count for the birth country of participants.

Birth Country	Value
Netherlands	327
Germany	0
Belgium	2
Other	9

Analyzing lifelong interests, a general trend emerged where average interest levels remained relatively stable or increased across different life stages. Table 4.3 illustrates this pattern, with Soccer, for example, exhibiting an increase in mean interest from 4.3 during young adulthood to 5.6 in the present. However, there was a notable decrease in interest observed specifically for ice skating, as the mean interest declined from 5.0 to 4.5 across life stages.

Moreover, participants tended to maintain their interests from a young age until the present day, as demonstrated in Figure 4.6. The figure specifically presents the interest levels, ranging from 1 to 10 (with 10 indicating the highest level of interest), for Film & Theatre between the young adult life stage and the current period. Similar patterns of interest were observed across different interest categories. Here the biggest count can be

Table 4.3: The mean interest per category of the participants throughout the life stages.

Category	Young Adult	Middle Age	Now
Film & Theatre	6.4	7.1	7.1
Music	7.5	7.5	7.4
Politics	5.4	6.8	7.2
Sports (general)	6.0	6.1	6.1
Ice Skating	5.0	4.9	4.5
Tennis	4.0	4.4	4.2
Soccer	4.3	4.5	5.6
Cycling	3.8	4.2	4.7

found between interest as young adult having value 7 or 8 and interest currently having value 8.

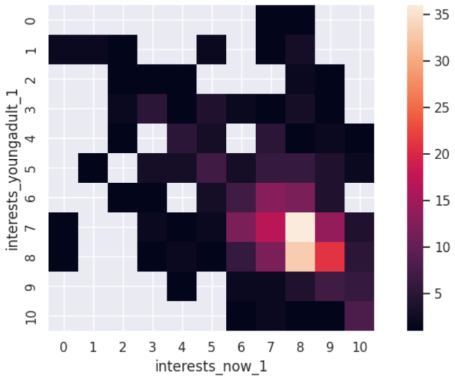


Figure 4.6: A heatmap of interest in category 1 (Film & Theatre) between life stages young adult and now.

4.2.3 FFT Dataset (Recall and Recognition

The overview of recall performance is presented in Table 4.4, which provides details on the participants' response rates, the number of correctly recalled faces, and the relative score obtained by dividing correct responses by total responses. The analysis reveals that, on average, participants completed 215.25 out of the total 220 items and accurately recalled 109 faces. Consequently, the relative correct score, with a mean of 0.51 (equivalent to 51%), indicates the proportion of correct responses. Notably, the standard deviation for both the number of correct responses and the relative correct score is high (39.64 and 0.18, respectively), indicating significant variability in the number of correctly recalled faces. This paper aims to investigate the factors contributing to these substantial score variations.

Additionally, the percentage of correct responses was calculated for each individual item. Table 4.5 presents the number of participants who provided correct answers (correct count) and the corresponding relative score (correct relative), obtained by dividing the number of correct responses by the total number of responses per item. None of the items were answered correctly by all participants. On average, 167.7 participants correctly responded to each item, yielding a relative score of 0.51 (equivalent to 51%). Similarly, the standard deviation values of 64.89 for correct count and 0.11 for correct relative reaffirm the findings from Table 4.4, emphasizing substantial disparities in participant performance and the accuracy for recall per item.

Table 4.4: Summary statistics of the overall performance of participants on recall.

	responses	correct responses	relative correct
Mean	215.25	108.87	0.51
Standard Deviation	6.73	39.64	0.18
Minimum	152	19	0.09
Maximum	220	210	0.96

Table 4.5: Summary statistics for each item on recall.

	correct count	correct relative
Mean	167.74	0.51
Standard Deviation	64.89	0.19
Minimum	36	0.11
Maximum	333	0.985

The distribution of participants who correctly recalled famous faces per item exhibits a wide range, as depicted in Figure 4.7. Although most items have a relative correct score ranging from 0.35 to 0.65, certain items were exceptionally well or poorly recalled. Table 4.6 highlights the top 5 best-recalled items, which include famous men from the 1970s and 2010s. These individuals represent a mix of international and Dutch figures and belong to the categories of Politics and Singers Musicians. The results do not reveal a clear pattern. Conversely, the top 5 worst-recalled items primarily feature male individuals, with one exception. These faces are associated with the 1960s and 2000s and once again comprise a combination of international and Dutch personalities.

Considering the insights from Table 4.1, several observations can be made in relation to Table 4.6. The overrepresentation of males in both the top 5 best and worst-recalled famous faces is not surprising, given the larger number of male faces in the item dataset compared to females. Additionally, there is a higher prevalence of famous faces from the categories of Politics and Singers & Musicians, which is reflected in the top 5 best and worst-recalled faces, with eight out of the ten belonging to these categories. However, it is unexpected to find three faces from the 2000s among the top 10 in Table 4.6, considering that the dataset has relatively fewer representations from the 2000s and 1990s, making these decades the least represented in the item dataset.

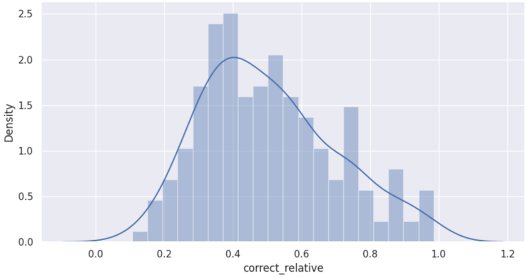


Figure 4.7: density plot of the relative correct score for recall items.

Table 4.6: top 5 items with best and worst recall correct relative score

Item	Correct Relative	Name	Gender	Decade	International	Category
C4397	0.99	Vladimir Poetin	Male	2010	Yes	Poitics
C2954	0.99	Mark Rutte	Male	2010	No	Politics
C3530	0.97	Pierre Kartner	Male	1970	No	Singers & Musicians
C1084	0.96	Elvis Presley	Male	1970	Yes	Singers & Musicians
C0184	0.96	André van Duin	Male	2010	No	Film & Theatre
C2717	0.19	Lex Goudsmit	Male	1960	No	Film & Theatre
C4289	0.18	Tonny Eyk	Male	2000	No	Singers & Musicians
C1263	0.16	François Mitterrand	Male	1990	Yes	Politics
C0053	0.16	Agnes Kant	Female	2000	No	Politics
C1799	0.11	Hosni Moebarak	Male	2000	Yes	Politics

The overall performance in recognition is presented in Table 4.7, revealing that participants, on average, responded to 218.6 out of 220 items, with an average of 207.2 correct responses. This results in a high relative correct score of 0.95 (equivalent to 95%). These scores significantly surpass the corresponding recall performance, indicating that the recognition test might be comparatively easier for participants. The standard deviation values follow the same trend, being notably lower than those observed for recall, as observed by comparing Table 4.7 with Table 4.4.

Similarly, Table 4.8 provides a parallel structure to the recall counterpart. The relative correct values range from 1.00 to 0.76 (equivalent to 100% to 76%). Remarkably, 23 out of 200 items were correctly filled in by all participants who provided an answer. On average, each item received a high accuracy rate of 0.96 (equivalent to 96%) in recognition. This exceptionally high level of accuracy suggests that meaningful insights may be challenging to extract through data exploration and modeling approaches.

Table 4.7: Summary statistics of the overall performance of participants on recognition.

	responses	correct responses	relative correct
Mean	218.60	207.23	0.95
Standard Deviation	4.46	12.90	0.05
Minimum	168	146	0.70
Maximum	220	220	1.00

Table 4.8: Summary statistics for each item on recognition.

	correct count	correct relative
Mean	319.49	0.96
Standard Deviation	19.60	0.05
Minimum	251	0.76
Maximum	338	1.00

The distribution of participants who accurately recognized famous faces per item exhibits a skewed pattern, as depicted in Figure 1. The majority of items demonstrate a high percentage of recognition, which aligns with the findings presented in Table 1. Only a small number of items exhibit relatively lower recognition performance, albeit still not reaching a significant low. The minimum recognition relative correct score recorded is 0.76 (equivalent to 76%).

Since 23 out of the 200 items achieved a perfect score of 1.00 (equivalent to 100%) in terms of correct relative responses, it was not possible to identify a top 5 list for the best recognized items. Instead, Table 2 highlights the top 5 worst-recalled items, which encompass a combination of male and female faces from different decades and categories. A notable observation is that all 5 of the least recognized items are of international figures, particularly noteworthy considering that the item dataset contains only 87 out of 200 items (43.5%)

with international faces, as indicated in Table 3.

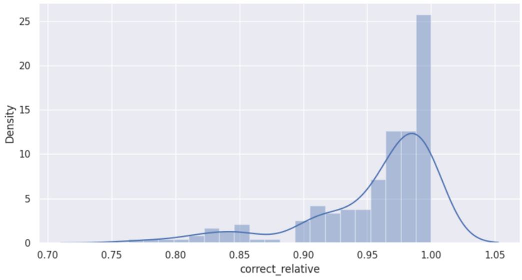


Figure 4.8: Density plot of the relative correct score for recognition items.

Table 4.9: Top 5 items with worst recognition correct relative score.

Item	Correct Relative	Name	Gender	Decade	International	Category
C2631	0.81	Lance Armstrong	Male	2000	Yes	Sports
C0036	0.81	Adele	Female	2010	Yes	Singers & Musicians
C2852	0.79	Mao Zedong	Male	1960	Yes	Politics
C3798	0.78	Roger Federer	Male	2010	Yes	Sports
C2827	0.76	Madonna	Female	2000	Yes	Singers & Musicians

Among the top 5 best recalled famous faces, all individuals achieved either a recognition score of 1.00 or 0.997 (equivalent to 100% and 99.7%, respectively).
Interestingly, the top 5 worst-performing items in terms of recall and recognition are not identical, as demonstrated in Table 4. This table highlights the recall and recognition correct relative scores for the top 5 worst-recalled and worst-recognized items. Notably, the worst-recalled faces exhibit a relatively high percentage of correct recognition scores (ranging from 0.90 to 0.99). However, these recognition relative correct scores differ significantly from those of the top 5 worst-recognized items, which range from 0.76 to 0.81. However the top 5 worst recognized items, do all have a low recall relative correct score with the exception of Mao Zedong. It can be concluded that poorly recalled items are still recognized relatively well, whereas poorly recognized items tend to exhibit lower recall scores.

In conclusion, the participants exhibited significantly higher performance in recognition compared to recall. Consequently, the decision was made to prioritize recall for data exploration and modeling, as the greater variation in this data provides a higher likelihood

Table 4.10: Top 5 items with worst recall and recognition scores compared.

Item	Name	Correct Relative Recall	Correct Relative Recognition
C2717	Lex Goudsmit	0.19	0.90
C4289	Tonny Eyk	0.18	0.90
C1263	François Mitterrand	0.16	0.97
C0053	Agnes Kant	0.16	0.99
C1799	Hosni Moebarak	0.11	0.98
C2631	Lance Armstrong	0.25	0.81
C0036	Adele	0.22	0.81
C2852	Mao Zedong	0.66	0.79
C3798	Roger Federer	0.26	0.78
C2827	Madonna	0.28	0.76

of uncovering meaningful results. Furthermore, no discernible patterns were identified regarding the characteristics that contribute to poor recall or recognition of items.

4.3 Exploring the Data

As previously emphasized, considering the substantial number of accurate responses for recognition, the exploration of the data focused exclusively on recall. This investigation was divided into three parts. The initial phase involved an exploration of the similarity among interest columns. Given the considerable volume of interest columns in the participant data, comprising 8 interest categories for 3 distinct life stages, an assessment was conducted to determine if each of these variables contributed valuable information to the research. Correlation coefficients were computed to examine the relationship between the 8 different interest categories within each life stage, as well as the correlation between the various life stages for each category. Consequently, a decision was made regarding the potential merging of highly correlated variables into a novel composite variable.

The remaining two components of the data exploration encompassed the utilization of clustering methods. Hierarchical clustering, employed to examine participant performance in relation to Research Question 1, and bi-clustering, employed to assess the interaction between participant performance and each individual item pertaining to Research Question 2.

To establish a connection between the resulting clusters and the participant (and item) data, Random Forests were employed as a means of predicting membership in the clusters generated by the hierarchical and bi-clustering analyses.

4.3.1 Correlation for Interest Levels

This section explores the feasibility of consolidating certain interest variables within the participant data, which are also represented in the matrices `merged_ID_recall` (refer to Figure 4.2) and `merged_ID_item_recall` (refer to Figure 4.3) established in Section 4.1.5. This step was taken, as there was a correlation found between interest throughout different life stages in section 4.2.2, figure 4.6

The correlation between interest variables was examined in two stages. Firstly, the correlation among the eight categories was calculated for each life stage, such as the correlation between each category within the 'young adult' life stage. Secondly, the correlation between life stages was computed for each category, such as the correlation between the three different life stages for the 'Film & Theatre' category.

The life stages considered in this study are 'young adult', 'middle age', and 'current', while the categories encompass 'Film & Theatre', 'Music', 'Politics', 'Sports (general)', 'Ice Skating', 'Tennis', 'Soccer', and 'Cycling'.

The Pearson method was employed to calculate the correlation coefficient. The Pearson correlation coefficient is a statistical measure that quantifies the extent of linear correlation between two variables [15]. The Pearson correlation coefficient is calculated like:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

Where cov(X,Y) represents the covariance between X and Y, and σ_X and σ_Y denote the standard deviations of X and Y.

When the correlation coefficient $0.8 < \rho_{X,Y} < 1$ is observed, it indicates a strong positive correlation between the variables X and Y. Similarly, a correlation coefficient $0.3 < \rho_{X,Y} < 0.6$ suggests a moderate positive correlation between the variables X and Y [16].

In order for variables to be considered similar and suitable for merging, a positive correlation is required. Therefore, a threshold of $0.7 < \rho_{X,Y}$ was chosen. When the correlation coefficient meets this threshold, the variables are considered similar enough to be merged into a new variable.

The participant dataset was utilized for conducting the correlation analyses. These correlations are also applicable to the datasets `merged_ID_recall` and `merged_ID_item_recall`, as the interest variables share the same distribution, covariance, and standard deviation.

Initially, correlations were calculated between the eight categories for each life stage. The correlation results for the 'young adult' life stage are presented in Figure 4.11. This figure illustrates a symmetric matrix displaying the correlation coefficients for every possible combination of categories within the 'young adult' life stage.

Moderate positive correlations ranging from 0.3 to 0.7 are observed among the categories 'Sports (general)', 'Ice Skating', 'Tennis', 'Soccer', and 'Cycling'. This can be attributed to the fact that these categories are all related to sports. Participants interested in one sport are more likely to exhibit interest in other sports as well, and vice versa for individuals with no interest in sports. The strongest correlation is found between the categories 'Soccer' and 'Cycling', indicating a positive correlation between participants' interests in cycling and their engagement in sports during their young adulthood.

Another moderate positive correlation is observed between the categories 'Music' and 'Film & Theatre', as well as between 'Politics' and 'Film & Theatre'. However, in the 'young adult' life stage, none of the categories meet the threshold of $0.7 < \rho_{X,Y}$, implying that none of the variables exhibit sufficient similarity in values to warrant merging.

Similar patterns as in the 'young adult' life stage, characterized by moderate positive correlations between specific categories and none of the correlations meeting the threshold, are observed in the other life stages, namely 'middle age' and 'current'.

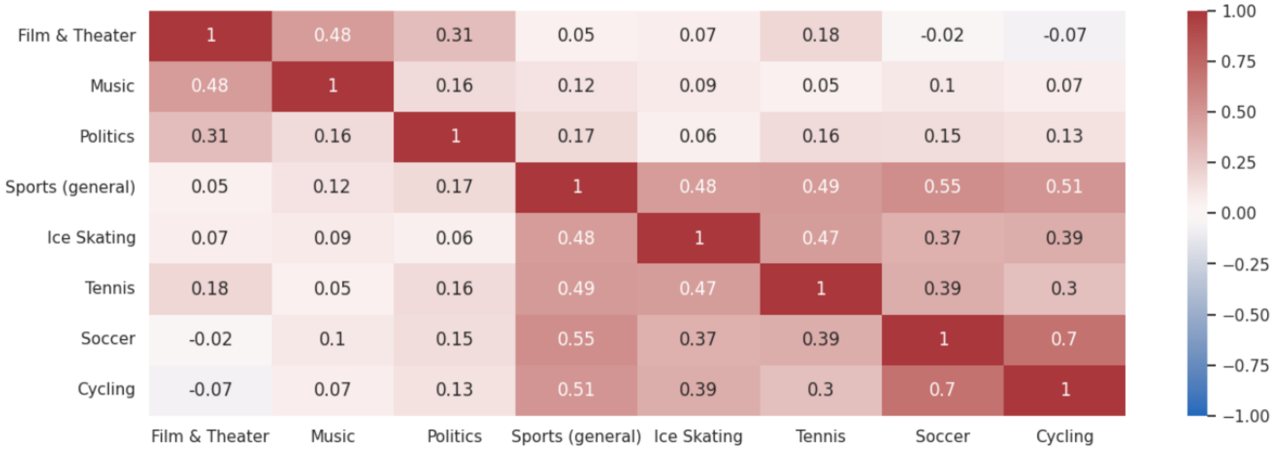


Figure 4.9: Correlation matrix for interests at life stage 'young adult'.

Subsequently, the correlation between life stages was calculated for each category. In this investigation, three categories displayed positive correlations across all life stages, surpassing the threshold of $0.7 < \rho_{X,Y}$: 'Sports (general)', 'Soccer', and 'Cycling'. Among these, the correlation for 'Cycling' is depicted in Figure 4.10. The symmetric matrix illustrates that the correlation between all life stages exceeds the threshold, indicating that the level of interest remains consistent for these three categories throughout different life stages. Considering that these three categories satisfy the threshold criteria, they were merged into a single new variable for each category: `Sports_interests`, `Soccer_interests`, and `Cycling_interests`. This merging process involved calculating the mean value across life stages for each of the three variables. Consequently, the original variables were eliminated from the dataframe.

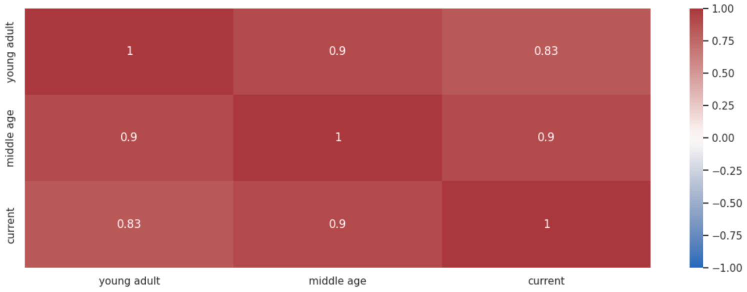
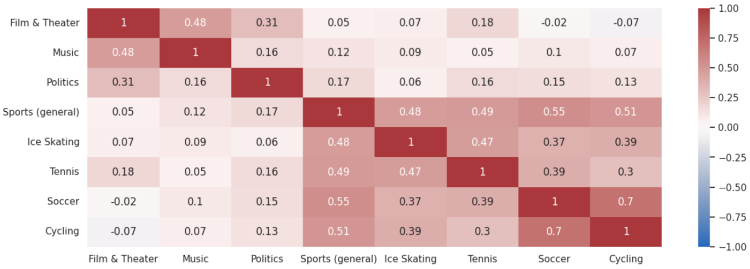


Figure 4.10: Correlation matrix for interest 'Cycling' throughout life stages.

In the case of the remaining five categories, namely 'Film & Theatre', 'Music', 'Politics', 'Ice Skating', and 'Tennis', a positive correlation above the threshold was not observed for every life stage. An example of the correlation coefficient matrix for the category 'Politics' is provided. The positive correlation coefficient only reaches the threshold between the 'middle age' and 'current' life stages, while no other correlations meet the threshold. A similar pattern is observed for the category 'Film & Theatre'. These findings indicate that, for these categories, the level of interest between the 'middle age' and 'current' life stages remains more consistent compared to the relationship between 'young adult' and either 'middle age' or 'current'. Consequently, the 'middle age' and 'current' life stages were merged into new variables (`Politics_mc_interest` and `FilmTheatre_mc_interest`), computed as the mean of these two life stages. The original columns used for these variables were then dropped.

For the categories 'Music', 'Ice Skating', and 'Tennis', two positive correlations were identified that met the threshold. Specifically, a strong positive correlation was observed between 'young adult' and 'middle age', as well as between 'middle age' and 'current'. This suggests that the interest levels for these categories do undergo some changes across life stages, albeit at a gradual pace. Nevertheless, the life stage positioned between 'middle age' still exhibited a strong correlation with both of the other life stages. Consequently, two new variables were created for each category: one merging 'young adult' and 'middle age' through the calculation of their mean (`Music_ym_interest`, `IceSkating_ym_interest`, and `Tennis_ym_interest`), and another merging 'middle age' and 'current' (`Music_mc_interest`, `IceSkating_mc_interest`, and `Tennis_mc_interest`).



	Film & Theater	Music	Politics	Sports (general)	Ice Skating	Tennis	Soccer	Cycling
Film & Theater	1	0.48	0.31	0.05	0.07	0.18	-0.02	-0.07
Music	0.48	1	0.16	0.12	0.09	0.05	0.1	0.07
Politics	0.31	0.16	1	0.17	0.06	0.16	0.15	0.13
Sports (general)	0.05	0.12	0.17	1	0.48	0.49	0.55	0.51
Ice Skating	0.07	0.09	0.06	0.48	1	0.47	0.37	0.39
Tennis	0.18	0.05	0.16	0.49	0.47	1	0.39	0.3
Soccer	-0.02	0.1	0.15	0.55	0.37	0.39	1	0.7
Cycling	-0.07	0.07	0.13	0.51	0.39	0.3	0.7	1

Figure 4.11: Correlation matrix for interest 'Politics' throughout life stages.

In conclusion, the analysis of the correlation between categories for each life stage reveals varying degrees of positive correlation, ranging from stronger to more neutral associations. No negative correlation effects were observed, with only minor negative correlations around 0 (e.g., -0.02). It is expected that a considerable portion of the correlation coefficients presented in Figure 4.11, which depicts the correlation coefficient matrix, demonstrate a neutral relationship. This is because these categories exhibit no apparent interconnectedness. For instance, individuals interested in tennis have no inherent reason to develop an affinity for music as a result of their tennis interest. Conversely, when someone is interested in a particular sport, they may be more inclined to encounter and appreciate other sports. Examining the correlation coefficients for each category across life stages, it becomes evident that a moderate to strong positive correlation exists within each category. Nonetheless, the weakest correlations consistently emerge in the relationship between the 'young adult' and 'current' life stages. This finding is not surprising, considering that interests may undergo a gradual shift throughout an individual's life. For example, as individuals age, their interest in cycling might gradually diminish.

4.3.2 Hierarchical Clustering & Random Forest

In order to examine the participant data, the technique of hierarchical clustering was employed to generate clusters based on similarity. Hierarchical clustering, an unsupervised machine learning approach, aims to group together data points that exhibit similar characteristics. The process involves providing a matrix as input, where initially all data points are considered as part of a single large cluster. Subsequently, the matrix is split into smaller clusters based on a preselected distance metric, until each cluster consists of a single element. The progression of this splitting procedure can be visualized through a dendrogram, enabling the determination of the desired number of clusters [16].

In the present study, hierarchical clustering was performed using the matrix `cluster_recall_matrix` (refer to Figure 4.2), wherein participants were grouped based on the similarity of their answer patterns regarding the accurate or inaccurate recall of famous faces. The Euclidean distance was chosen as the distance metric, which quantifies the direct distance between two points in the matrix [17]. To establish the optimal number of clusters, two methods were employed: the elbow method and the silhouette method.

The elbow method involves plotting the number of clusters on the x-axis against the total within sum of squares on the y-axis. The plot is analyzed to identify a bend or elbow point, indicating a significant decrease in the sum of squares. It is important to note that this method can introduce subjectivity, as the identification of the elbow point can vary depending on interpretation. In this study, the 'fviz_nbclust' function from the 'factoextra' package was employed to implement the elbow method.

Additionally, the silhouette method was employed to provide a more definitive determination of the appropriate number of clusters. This method computes silhouette coefficients for each data point, quantifying the similarity of a point to its own cluster in relation to other clusters. The results are visualized in a graph. As with the elbow method, the 'fviz_nbclust' function from the 'factoextra' package was utilized for this purpose.

A preliminary indication regarding the optimal number of clusters was obtained through the application of the elbow method, as depicted in Figure 4.12. The plot exhibits an elbow bend around the range of 2 to 3 clusters. Nonetheless, it should be noted that the subjective nature of this method prevents it from yielding a definitive answer.[18]

The confirmation of 2 clusters as the optimal number was obtained through the implementation of the silhouette method, as illustrated in Figure 4.13. The graph clearly demonstrates a distinct peak in the similarity of data points to their respective clusters when utilizing 2 clusters.[19]

Figure 4.14 exhibits a dendrogram depicting the hierarchical clustering of the `cluster_recall_matrix`. Each split in the tree corresponds to a division in the data, leading to the formation of clus-

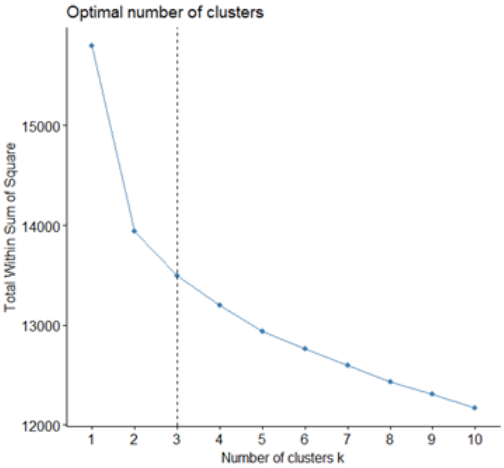


Figure 4.12: Elbow method to determine optimal number of clusters.

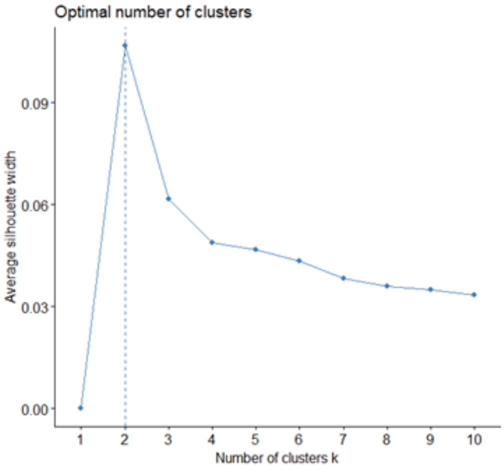


Figure 4.13: Elbow method to determine optimal number of clusters.

ters. The proximity of data points in the lower regions of the tree reflects their similarity in terms of recall responses (for example, correct and incorrect answers). In the dendrogram, the colors red and green represent the two resulting clusters, which are formed relatively early in the clustering process.

Subsequently, the two clusters were merged with the dataframe `merged_ID_recall` (Figure 4.2). A supplementary column was introduced, denoted by values '1' and '2', indicating the cluster to which each participant belongs. The distribution of participants among the clusters revealed that 60.7% belonged to cluster 1, while 39.3% belonged to cluster 2.

To predict the cluster membership of participants based on the data in `merged_ID_recall`, a random forest model was employed. A random forest consists of an ensemble of multiple decision trees, wherein each tree predicts a class, and the tree with the highest probability is selected. The trees are constructed using subsets of the input data during training, thereby mitigating the risk of overfitting.[20]

Prior to applying the random forest classifier, on-hot encoding was performed for object columns (non-numerical) since the classifier cannot handle categorical data as input. Here for each categorical value of a column, a new column is created with values 0 and 1, where 0 means the participant did not have this categorical value and 1 meaning the participant did have this categorical value. Furthermore, missing values were imputed with the mean values for each numerical column, as the random forest cannot process NaN values. Mean imputation was preferred over imputing null values for all missing data, as the random forest model relies on data division based on decision criteria. Imputing all missing values with 0 would result in all missing entries falling into one category during the data division

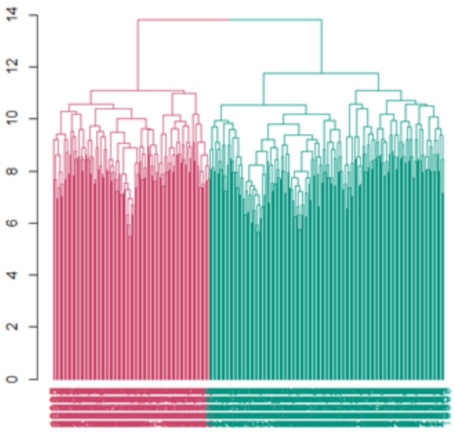


Figure 4.14: Dendrogram for hierarchical clustering.

process, potentially skewing the representation of the actual clusters. Mean imputation, on the other hand, provides more nuanced values and removes this issue. If a split is chosen for a column, for example x $>$ 3, then all imputed values of 0 will always automatically fall into the group of lower value, while if the mean is used, it gives the model more leverage to choose which group the missing values should be assigned.

The dataset was then split into training and testing data, employing stratification based on the cluster column to ensure equal representation of all clusters in both sets. The split ratio used was 80:20. Subsequently, the random forest model was fitted using the training data, aiming to predict the cluster membership of participants based on their data variables, excluding the `recall_correct` column. This exclusion was made because the clusters were formed based on the similarity of participants' recall scores. Including the `recall_correct` column as a dependent variable could potentially diminish the model's ability to identify patterns among the participant data variables specific to the different clusters.

To assess the predictive performance of the random forest, the model was used to predict the clusters for the test set, which resulted in an accuracy of 0.96. The fit of the random forest was evaluated by fitting the model on the training set, which yielded an accuracy score of 0.94. The accuracy for the training and test set being close in value, shows that the model is well-fitted without signs of under- or overfitting. This outcome aligns with expectations, as random forest models are not prone to overfitting.

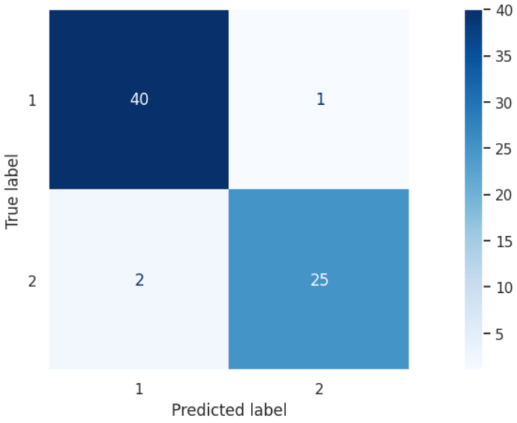


Figure 4.15: Confusion matrix for predicting of hierarchical clusters.

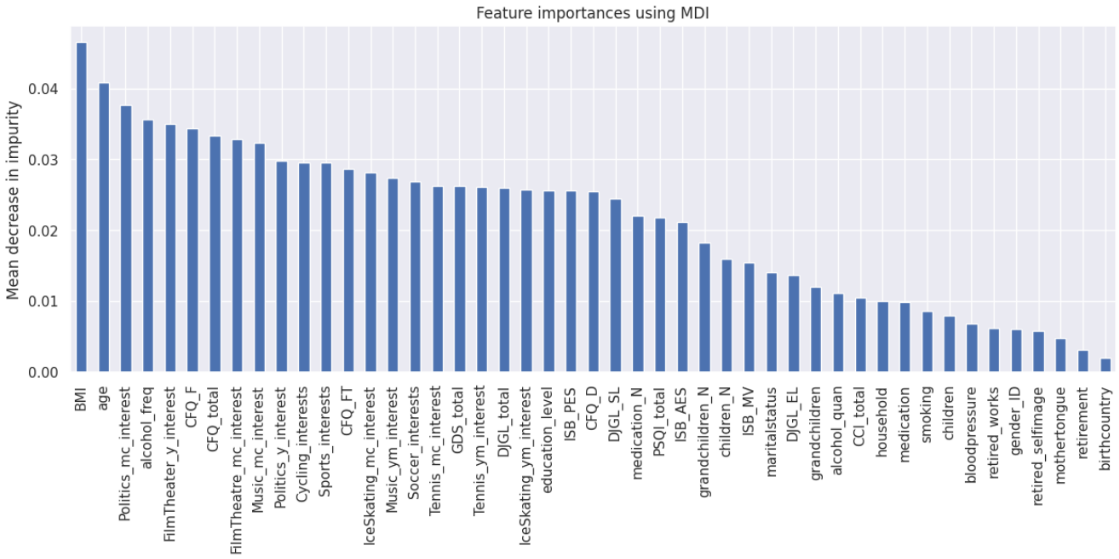


Figure 4.16: Feature importance for predicting of hierarchical clusters.

The effectiveness of the model in predicting clusters can be observed from the confusion matrix, comparing the predicted and true clusters (Figure ??). Out of the 68 participants in the test set, only 3 were inaccurately predicted, indicating a high level of accuracy. Furthermore, these prediction errors were evenly distributed across the clusters, suggesting that the model performs equally well for both classes of clusters.

Insight into the importance of individual features in determining the participant's cluster membership can be gained from both Figure 4.16 and Table 4.11. As random forest models are considered black box models, examining feature importance provides valuable insights into the decision-making process of the model. Feature importance is calculated based on the mean decrease in impurity, which is determined by the impurity function utilized during feature selection at each node of the decision trees within the random forest model. [21]

The feature importance scores have significant range, with some features having higher values and others lower values. However, there is also a substantial portion in the middle of Figure 4.16 where multiple features possess similar importance scores. In Table 4.11, the features with the highest and lowest importance values are highlighted.

Among the top five most important features for predicting the clusters, there are unexpected features such as BMI and alcohol frequency, as these do not appear to have an obvious tie with the ability to recognize famous faces. The second most important feature is age, which could potentially be attributed to the increase in forgetfulness that often accompanies aging. [22] Which could make it more difficult to recall famous faces.

Furthermore, interest in Politics during life stages 'middle age' and 'current', along with interest in Film & Theatre during the life stage 'young adult' , are also included in the top five most important features for predicting the clusters. It is noteworthy that in the item datatset the Politics category has the highest count, as illustrated in Table 4.1. This factor could contribute to the increased importance of this specific category. Additionally, while the Film & Theatre category has a lower representation compared to Politics and Singers & Musicians, it still plays a significant role in predicting the clusters.

Interest variables appear to have a relatively higher frequency of importance, as depicted in Figure 4.16. This observation could be attributed to individuals with a strong interest in a particular category being more familiar with famous faces associated with that category.

Conversely, the top five least important features primarily consist of demographic statistics that do not possess evident connections to the recognition of famous faces. The low importance of variables such as mother tongue and birth country is consistent with the findings in Section 4.2.2 concerning the statistics of participant data, where it was highlighted that the majority of participants fell into a single category for both variables (Netherlands for birth country, as shown in Table 4.2, and Dutch for mother tongue).

In summary, the application of hierarchical clustering and modeling to the clusters provided initial insights into the potential significance of various features in predicting participants' correct scores. The analysis revealed that a majority of the features play a role

Table 4.11: Top 5 dependent variables with most and least feature importance for predicting hierarchical clusters.

Feature	Importance
BMI	0.047
age	0.041
Politics_mc_interest	0.038
alcohol frequency	0.036
FilmTheatre_y_interest	0.035
gender participant	0.006
retired selfimage	0.006
mothertongue	0.005
retirement	0.003
birthcountry	0.002

in predicting the correct clusters, with particular emphasis on BMI, age, and interest variables. Which could suggest that a diverse range of features contributes to the individual differences that influence performance on the Famous Faces Test (FFT).

4.3.3 Bi-Clustering & Random Forest

Bi-clustering is a data rearrangement technique used to create clusters within a matrix. In this study. The BCQuestMet bi-clustering method was employed, this algorithm searches subgroups of values in the matrix with same or similar patterns. [23] The recall matrix `cluster_recall_matrix` was again used to perform the clustering.

Figure 4.17 illustrates the visualization of the bi-clusters extracted from the data. The dark red color represents correct recall answers (score 1), while the light yellow color represents incorrect or missing values (score 0). It can be observed that the formed clusters are relatively small in size. A substantial portion of the participant data remains unassigned to any cluster, and only a small fraction of the items as it only groups together data with a strong cluster structure, comprising 5.8% of the entire matrix, are included in the clusters. In total, eight clusters were created.

The small size of the clusters indicates that there are limited larger clusters to be formed within the data. This suggests that modeling the data to predict participants' ability to recall a famous face might pose a challenge. Since there are no distinct groups of participants and items that the model could identify as being similar in terms of correctness in recalling famous faces.

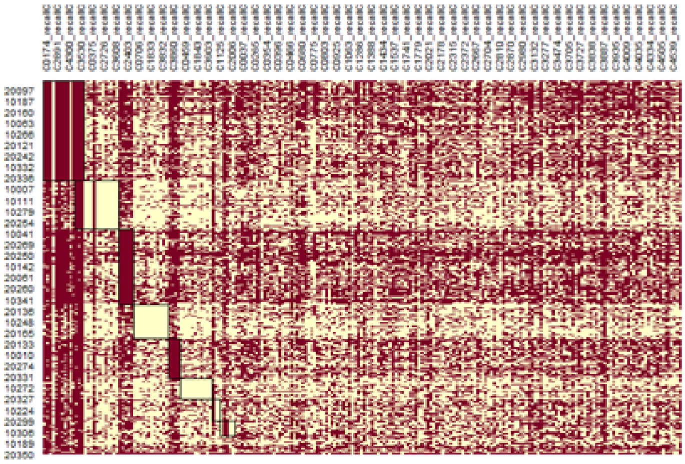


Figure 4.17: Bi-Clustering for participants and items.

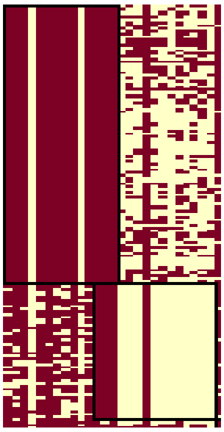


Figure 4.18: A zoomed-in picture of the first two bi-clusters

The clusters obtained from the bi-clustering process were merged with the recall dataframe `merged_ID_item_recall` 4.1.5 (refer to figure 4.3). An additional column was introduced to indicate the assigned cluster for each participant-item combination, represented by values ranging from '1' to '8'. Similar to previous steps, object columns (non-numerical) were encoded using the mean of the column, as the random forest classifier cannot handle categorical data as input. Furthermore, missing values in numerical columns were imputed with their respective column means. These imputation procedures were carried out for the same reasons stated in section 4.3.1.

To ensure the integrity of the modeling process, all participant-item combinations that were not assigned to a cluster were removed from the dataframe, resulting in a reduced dataset of 3922 rows. This measure aimed to prevent the Random Forest from predominantly classifying participant-item combinations as belonging to the 'no cluster' category, which would happen as the training set would have a big class imbalance for the 'no cluster' class.

Subsequently, the dataset was divided into training and testing sets, stratified based on the cluster column, to ensure an equal representation of all clusters in both sets. The train-test split ratio was set to 80:20. The Random Forest model was then fitted using the training data, with the exclusion of the recall correct column to avoid biasing the model with the dependent variable used in its creation.

The Random Forest model demonstrated exceptional performance in predicting the clusters for the test set, achieving a perfect accuracy score of 1.00. This remarkable accuracy is also evident in figure 4.19. Similarly, when applied to the training set, the model achieved a flawless accuracy score of 1.00. In conclusion, the model exhibits a high degree of fit to the data, accurately capturing the assigned clusters for participant-item combinations.

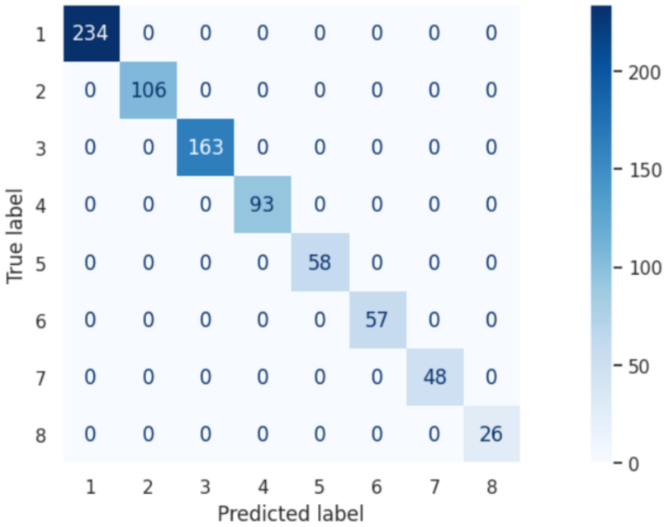


Figure 4.19: Confusion matrix for predicting of the hierarchical clusters.

Age is an outlier in terms of feature importance, as depicted in figure 4.20, showing a similar pattern to figure 4.16 in hierarchical clustering. Once again, age emerges as a

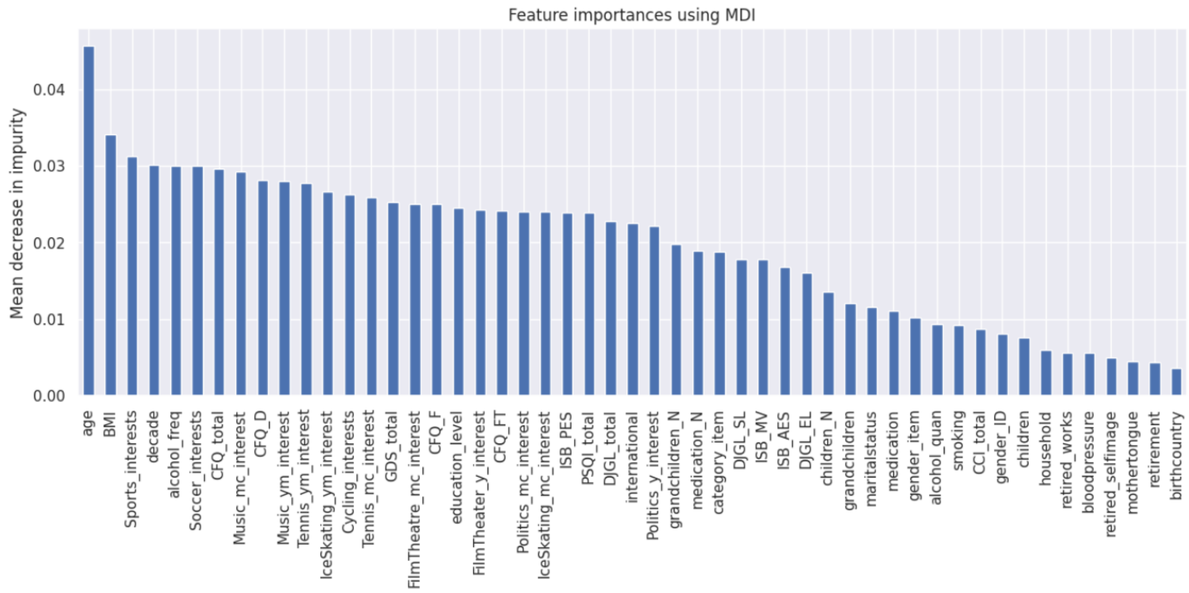


Figure 4.20: Feature importance for predicting of the hierarchical clusters.

potentially significant feature due to the well-established association between age and increased forgetfulness. This relationship suggests that age may pose challenges in the recall task.

Table 4.12 provides further insight into feature importance in bi-clustering. Similar to hierarchical clustering, age and BMI continue to be the most influential features. However, the exact reason for the high importance of BMI remains unclear.

Of particular interest in the bi-clustering feature importance analysis is the inclusion of lifetime interest in sports and soccer among the top five features. This finding is intriguing, considering that the Sports category is the smallest of all item categories (see table ??). Speculatively, a strong interest in sports and soccer may exert a more pronounced influence on the ability to recall athletes compared to the impact of other interest categories on recalling items from their corresponding categories. Other interest categories exhibit higher feature importance as well in bi-clusters, paralleling the findings in hierarchical clustering. Additionally in the top five most important features CFQ_total appears, a measure obtained from the cognitive failures questionnaire designed to assess the frequency of lapses of attention, memory, and cognition in everyday life.[?] This variable's relevance to face recall can be attributed to higher scores indicating reduced attention spans or memory failures, potentially hindering a participant from recalling a famous face correctly.

Remarkably, the top five least important features predominantly pertain to item-related variables: the item's decade, international status, category, and gender. This outcome is unexpected, particularly regarding the item category's diminished importance, as it was anticipated to relate to participant interests. If a strong relationship existed between item category and participant interests, it would be expected that the item category would have a higher feature importance.

Contrary to expectations based on the findings of Orlovsky et al. (2018)[11], which suggested that healthy older adults display better recognition and recall of famous faces from past decades compared to more recent famous faces, our analysis did not find the decade of the item to be a significant feature in predicting the cluster which the participant-item combination belonged to.

.

Table 4.12: Top 5 dependent variables with most and least feature importance for predicting the bi-clusters.

Feature	Importance
age	0.046
BMI	0.034
Sports_interests	0.031
Soccer_interests	0.030
CFQ_total	0.030
gender ID	0.008
item decade	0.003
item international	0.002
item category	0.002
item gender	0.001

Based on the results obtained from modeling the bi-clusters, it can be concluded that participant-item combinations with similar recall patterns of famous faces (visualized in `cluster_recall_matrix`), and thus assigned to the same clusters, can be effectively modelled to predict their membership to the clusters. This suggests that when modeling the data to predict the correctness of participants in recalling specific items, the resulting model may achieve high accuracy. However, it is important to note that the prediction of bi-clusters was based on only 5.8% of the available data, a percentage of the data which were clustered together on participant-item combinations with similar performance, making it difficult to accurately assess the performance of the model in predicting recall correctness for the entire dataset.

Regarding the exploration of bi-clustering, the findings indicate that age, BMI, and interest variables were the most influential features in predicting the clusters, which aligns with the observations made for the hierarchical clusters. Features related to the items demonstrated lower importance in the prediction task.

4.4 Modeling of the Data: Decision Trees

4.4.1 Research Question 1

To address the initial research question, a decision tree was constructed using the `merged_ID_recall` dataframe, which was generated in section 4.1.5 as illustrated in figure 4.2. In this analysis, the column `recallCorRelative` served as the independent variable, representing the overall score expressed as the percentage of correctly recalled faces relative to the total number of faces recalled. The dependent variables consisted of the participant data encompassing individual characteristics. To prepare the data for the decision tree the object columns (non-numerical) were encoded and NaN values in numerical columns were imputed with the mean, this was done for the same reason discussed in section 4.3.2. The first research question encompasses the following objectives:

RQ1: How do the individual characteristics of a healthy older adult influence their overall performance on the FFT?

The decision tree regressor is a modeling technique used for predicting continuous variables, making it suitable for the independent variable, as each participant's score falls within the range of 0.00 to 1.00. The primary objective is to divide the population into homogeneous subsets based on the most influential dependent variables. [24]
In this study, a decision tree regressor was employed for modeling the data, instead of using a random forest, which comprises multiple decision trees. The decision to opt for a decision tree was motivated by its transparency compared to the random forest. Although the random forest offers higher accuracy, the decision tree provides insights into the precise decision points for splits, aiding in the identification of individual differences that impact overall performance and the specific criteria for these feature-based splits [24]. Given that the independent variable is continuous and numeric, with minimal occurrences of identical relative correctness scores among participants, the test and training set splits were not stratified.
It is worth noting that decision trees are prone to overfitting on the training set. To assess

the model's fit and ensure it does not overfit the data, the accuracy of both the test and training sets was computed. Since the independent variable is continuous and numeric, accuracy was evaluated using the mean squared error (MSE).

The MSE is defined as the average squared difference between actual and predicted values, specifically in this case, the difference between the estimated total relative correct score and the actual total relative correct score (`recallCorRelative`). MSE can be expressed as:

$$MSE = \frac{1}{n}\sum(ActualValue - PredictedValue)^2$$

Decision tree regressors have several parameters. To identify the best-fitting tree with the lowest MSE for the test set, various combinations of three parameters were tested exhaustively. The parameters examined were:

1. criterion ('squared_error', 'friedman_mse', 'absolute_error')
2. max_depth (values 2 - 15)
3. min_samples_leaf (values 5, 10, 15, 20)

The criterion parameter is utilized to assess the quality of a split, offering various criteria to determine the most advantageous split. The criterion options serve as different approaches for evaluating the splits' effectiveness. Additionally, the max_depth parameter determines the maximum allowable depth of a tree, limiting its complexity. Lastly, the min_samples_leaf parameter specifies the minimum number of samples required for a node to be created and exist within the tree structure [25].

Another evaluation metric employed in this study involved generating a residual plot, which is a scatterplot to compare the true values and predicted values of the total relative correct score.This plot has the predicted values on one axis and the true values on the other. In the case of a perfect prediction, the scatterplot would exhibit a 45-degree angle from the origin, indicating that the predicted values align precisely with the true values. If there is no strong correlation between the independent and dependent variables, the scatter plot of data points will have a large spread and the MSE will be high. If there is indeed a strong linear relationship between the independent and dependent variables, the MSE will be low.

To gain insights into the decision-making process of the model and to determine the specific individual differences that influence overall recall performance, the decision tree was visualized. This visualization aids in providing definitive answers regarding the factors that contribute to the model's decisions and the subsequent impact on the overall recall performance.

4.4.2 Research Question 2

In order to address the second research question, the dataset will be employed to construct a predictive model that determines whether individual participants recalled specific items correctly. The dependent variables encompass both participant and item data. The independent variable utilized in this analysis was the `Cxxxx_recallCor` column, which takes the values of '0' (indicating an incorrect response) and '1' (indicating a correct response). The dataset employed for the data modeling process was the `merged_ID_item_recall` dataframe, which was created in section 4.1.5 (refer to Figure 4.2).
To prepare the data for the decision tree algorithm, the non-numerical object columns were one-hot encoded, and missing values in numerical columns were imputed with the mean. This procedure was conducted for the same reasons discussed in section 4.3.2. Consequently, the second research question encompassed:

RQ2: How do the descriptive features of items presented to participants in the FFT interact with their individual characteristics, and what is the impact of these interactions on their ability to accurately recall the presented items?

To establish the data modeling process, two decision tree classifiers were constructed. The first classifier served the purpose of feature selection, due to the substantial number of variables involved (a total of 54, before the one-hot encoding). The second decision tree classifier aimed to address the research question at hand. The decision to employ a decision tree classifier, instead of a random forest, aligns with the explanation provided in section 4.4.1. Resulting in a more precise identification of the specific individual differences and item characteristics that influence the participants ability to recall specific items, as well as the corresponding splitting criteria associated with each of these features. Consequently, it sheds light on the interaction between individual differences and item characteristics. Stratified sampling was employed to partition the dataset into training and testing sets, ensuring a representative distribution of the independent variable across both sets.
To evaluate the decision tree classifier for potential overfitting, the accuracy of predictions was measured on both the test and training sets. Accuracy, defined as the percentage

of correct predictions for the 'correct' or 'incorrect' classes, was an appropriate metric for assessing the model's predictive performance, as it provides insight into its ability to correctly predict the different classes. Notably, the distribution of classes in the dataset was approximately 49.4% for the 'incorrect' class and 50.6% for the 'correct' class. This distribution bears significance when evaluating accuracy. In cases where there is a severe class imbalance (e.g., 95-5), a high accuracy score could be misleading if all predictions are biased toward the majority class. However, in this scenario with a relatively balanced distribution, accuracy holds value in assessing model performance. The accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

To identify the optimal decision tree classifier, three parameters were systematically tested with various inputs, aiming to achieve the highest possible accuracy on the test set. Exhaustive testing was conducted, considering every possible combination of the following parameters:

1. criterion ('gini', 'entropy')
2. max_depth (values 2 - 15)
3. min_samples_leaf (values 5, 10, 15, 20)

These parameters have the same definition for both the classifier and the regressor, as discussed in section 4.4.1.

Following the construction of the decision tree, feature selection was performed due to the high number of variables involved. The feature importance for each column within the decision tree was computed and ranked in descending order. Features were sequentially added to a list based on their importance until the cumulative threshold of 0.90 feature importance was reached. These selected features were incorporated into the final decision tree. The underlying assumption is that numerous features possess negligible importance and can be discarded without compromising the accuracy of the results.

Subsequently, the same procedures employed for the initial tree were repeated for the final tree after the feature selection process. Exhaustive testing of different parameters was conducted, and accuracy was assessed on both the training and testing sets to evaluate potential overfitting.

To gain insight into the tree's splitting behavior, the first four layers of the tree were visualized, as depicting the entire tree was not feasible within the confines of this paper due to its extensive tree depth. The visualization of the decision tree was employed to illustrate the decision-making process, offering conclusive insights into the individual differences and item characteristics that influence participants' ability to accurately recall famous faces. To still gain insight into feature importance a bar plot was created showing the importance of every variable and the top 5 most important features were highlighted in a table. Here the one-hot encoded variables were merged together to show the feature importance for the original variables.

The evaluation of the final tree also involved the utilization of a confusion matrix, which provides information on the model's predictive performance across different classes. As well as classification report which includes the recall, precision and f1-score, and an ROC curve.

Precision shows the amount of instances which is classified correctly by the model for a class, compared to the total amount of instances which belong to the class.

$$Precision = \frac{TP}{TP + FP}$$

Recall on the other hand shows the amount of instances correctly classified by the model for a class, compared to the total amount of instances which the model classifies as that particular class. In other words, recall is the accuracy per class.

$$Recall = \frac{TP}{TP + FN}$$

The F1-score is the harmonic mean of precision and recall. It attempts to provide a metric that balances both characteristics of precision and recall into a single number. Usually, the F1-score is used as a feedback to optimize/tune machine learning models. The higher the F1-score, the better your model. The reason F1-score is used as an evaluation metric is because the classes are nearly balanced, and so we can equally consider false positives and false negatives. The F1-score is expressed as:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

An ROC curve is used to measure the ability of any binary classification method by plotting the recall (also known as the true positive rate) against the false positive rate, thus showing the trade off between sensitivity and specificity. The ROC curve is built by evaluating a model given various classification thresholds. Since ROC does not consider

class distributions, it makes it very beneficial for ailments that are rare in nature that lead to imbalanced data sets.
Furthermore, if you compute the area under the curve (AUC) of the ROC by taking the integral, it tells us the probability of a random positive case being positioned away from a random negative case. In other words, it measures the ability of a classifier to distinguish between the two classes (incorrect vs. correct). ROC-AUC is a great metric for model evaluation and selection just like the F1-score.

5

Results

5.1 Research Question 1: Decision Tree Regressor

5.1.1 Decision Tree performance

Following an exhaustive exploration of parameter combinations for the decision tree regressor, the mean squared error (MSE) scores ranged from 0.035 to 0.066. The lowest MSE score was observed for the parameters: criterion = 'friedman_mse', max_depth = 2, min_samples_leaf = 20. Notably, this configuration exhibited a particularly low maximum depth and a high minimum number of samples, indicating a straightforward tree structure.

For this specific parameter combination, the MSE of the training set was determined to be 0.027, which was lower than the corresponding test set MSE. This finding suggests that the decision tree regressor did not exhibit overfitting on the training set.

However, it should be acknowledged that this model is not an ideal fit, as evidenced by the discrepancy between the true and predicted values (see Figure 5.1). In the scatterplot, the points deviate from the ideal 45-degree line originating from the origin, although the deviation is particularly pronounced for the highest predicted value.

The independent variable `recallCorRelative` displays a mean of 0.51 and a standard deviation of 0.18, indicating significant variability in the total relative correct score (refer to Table 4.4, section 4.2.3).

The mean absolute difference between the predicted and true values is 0.16, with a standard deviation of 0.11. This implies that, on average, the model predicts the total score to be approximately 0.16 units higher or lower than the actual value. Given that the scores range from 0.09 to 0.96 (representing the minimum and maximum relative correct recall scores in Table 4.4), such a substantial deviation suggests that the model may not accurately predict the scores, especially considering the relatively high standard deviation of 0.11. This indicates significant variability in the differences between the predicted and actual values.

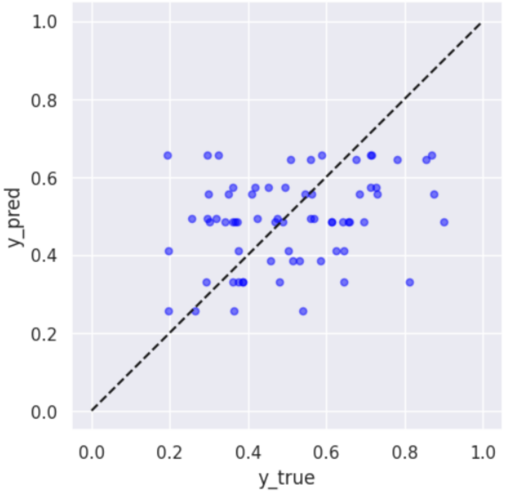


Figure 5.1: Difference between true and predicted value of the decision tree regressor.

5.1.2 Decision Tree features

The decision tree (see Figure 5.2) reveals the splitting criteria that contribute to the prediction of the correct relative score. Notably, the decision tree is characterized by simplicity, with only three splitting criteria, two of which pertain to the variable "age." Considering that age ranked among the top five most important features for predicting hierarchical clusters (refer to Table 4.11), it was expected to play a significant role in predicting the total relative correct score for recall. Analysis of the splitting criteria indicates that the model tends to assign lower scores to participants of higher age.

The final splitting criterion is based on the variable `FilmTheatre_y_interest`, representing interest in Film & Theatre during the life stage of 'young adult'. The model assigns a lower value of 0.33 if the interest level is equal to or below 3.5, and a higher value of 0.569 if the interest level exceeds 3.5. The choice of 'Film & Theatre' as a splitting criterion over other interest categories among participants is unexpected, considering the item data distribution presented in Table 4.1, which shows higher counts for categories such as Politics and Singers & Musicians.

An examination of the decision tree reveals that the predicted total relative correct scores exhibit a narrow range, ranging from 0.263 to 0.569, in contrast to the broader range of true values, spanning from 0.09 to 0.96.

The majority of the output (61.1%) is assigned the value of 0.569, which is the highest predicted outcome. This distribution is expected, given that the mean of the actual values is 0.51, indicating that a significant portion of participants would have a true value above 0.569. It would be assumed that all participants with scores around or above 0.569 would be assigned to this cluster of predicted outcomes.

However, the model not only assigns predicted scores to participants falling within this range but also to those with very low scores, as evident in Figure 5.1. The scatterplot illustrates that, for each of the four predicted values, the true values are dispersed throughout the entire range. This observation indicates that the model may not be effective in accurately predicting the total relative correct score for recall.

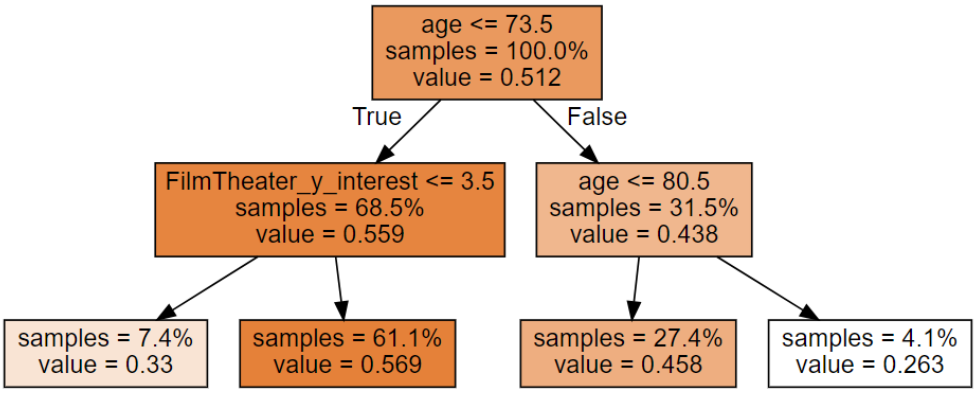


Figure 5.2: Decision Tree for prediction correct relative score (left split = True, right split = False)

In summary, the decision tree regressor utilizes age and interest in Film & Theatre during young adulthood as the variables for assigning predicted values. However, it is evident that the decision tree falls short in accurately predicting the total scores. Consequently, questions arise regarding the reliability of these results and whether they genuinely reflect the influence of the variables as described by the decision tree.

The diminished accuracy of the decision tree in predicting the scores may stem from two potential factors. Firstly, it is possible that the decision tree regressor failed to identify sufficient patterns within the data to adequately fit the model. This limitation could be attributed to various factors, such as a relatively smaller sample size, as the dataset comprises only 338 participants and corresponding data points. Alternatively, it is conceivable that the participant data holds limited sway over the total relative correct score, thereby hindering the decision tree's ability to establish an accurate model.

5.2 Research Question 2: Decision Tree Classifier

5.2.1 Feature Selection

The decision tree, employed for feature selection, underwent testing with various parameters to identify the optimal configuration that achieved the highest accuracy. Throughout the evaluation, accuracy values ranged from 0.53 to 0.66. Remarkably, the highest accuracy of 0.66 was attained when utilizing the parameters: criterion = 'gini', max_depth = 14, min_samples_leaf = 15. Notably, this decision tree exhibited an accuracy of 0.70 when applied to the training set, indicating the absence of overfitting.
Figure 5.3 shows there is no imbalance in the decision tree's ability to predict the classes. As the classes 'correct' and 'incorrect' have an almost even split and the confusion matrix has a similar amount of 'correct' and 'incorrect' values it predicts incorrectly.

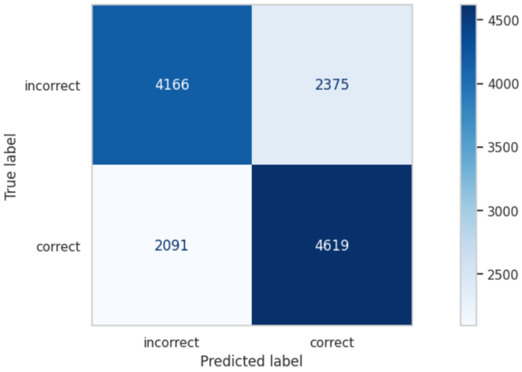


Figure 5.3: Feature selection: difference between true and predicted value for the decision tree classifier.

The current decision tree involved the computation of feature importance and cumulative feature importance. A partial representation of the feature importance list, highlighting the initial entries, can be found in Table 5.1. Analysis reveals that the top five features, out of a total of 54 variables, contribute to a cumulative feature importance of 0.421. Given that the cumulative feature importance of all variables sums up to 1.00, a threshold of 0.90 was employed to exclude features with negligible importance. Consequently, 22 variables were eliminated through the feature selection process, and the complete list of these eliminated features can be found in the appendix.

The eliminated features predominantly encompass demographic variables, including the number of children and grandchildren, marital status, and retirement variables. Additionally, variables such as '=`mother tongue` and `birth country` were discarded these variables were previously established to have a high volume of the same values, as evident in Table 4.2 and Section 4.2.2. Surprisingly, certain variables related to interests, such as `FilmTheatre_mc_interest` (interest in Film & Theatre at life stages 'middle age' and 'current') and `Politics_mc_interest` (interest in Politics at life stages 'middle age' and 'current'), were unexpectedly eliminated through feature selection. These variables were initially speculated to bear significance, as an individual's interest in poltics, for instance, could potentially influence their ability to recall politicians accurately.

Table 5.1: Feature Selection: feature importance for top 5 most highest ranking variables.

Feature	Importance	Cumulative Importance
item decade	0.116	0.116
item category	0.109	0.225
FilmTheatre_y_interest	0.085	0.310
age	0.058	0.368
BMI	0.053	0.421

5.2.2 Decision Tree

5.2.2.1 Decision Tree performance

The decision tree for classifying participant-item combinations had an accuracy of 0.67 on the test set and an accuracy of 0.70 on the training set. This was the decision tree with the highest accuracy of all the parameter combinations tested, with accuracy on the test set ranging from 0.55 to 0.67. The parameters for the decision tree were: criterion = 'gini', max_depth = 15, min_samples_leaf = 15. With the accuracy of the test and training set being close together, there is no overfitting.

From the confusion matrix in figure 5.4 it becomes clear that the model predicts both classes almost equally well, with wrong predictions for 'correct' and 'incorrect' true values being similar.

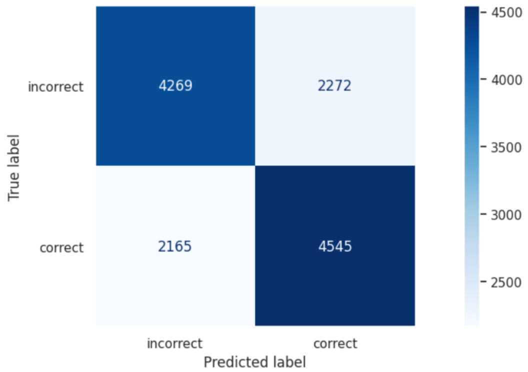


Figure 5.4: Difference between true and predicted value for the decision tree classifier.

Table 5.2: Classification report of the decision tree model.

Class	Precision	Recall	F1-score	Instances
'incorrect'	0.66	0.65	0.66	6541
'correct'	0.67	0.68	0.67	6710

Due to the consistency in scores across the metrics precision, recall and F1-score, in table 5.2, we see that the decision tree model performs relatively the same across the two classes. The reason the model performs slightly better on the class 'correct' is due to the excess instances (+169) over class 'incorrect'.
Figure 5.5 shows that the decision tree has a AUC score of 0.73. Given two data points, one belonging to the class 'incorrect' and one to the class 'correct', the AUC score shows that the model has a 73% chance of predicting both data points correctly.

Table 5.3: Predicted probability scores of 5 instances in the test dataset.

Data point	Predicted class	$P(Y = 0\|X)$	$P(Y = 1\|X)$
x_1	correct	0.34	0.66
x_2	correct	0.14	0.86
x_3	incorrect	0.55	0.45
x_4	incorrect	0.70	0.30
x_5	incorrect	0.97	0.03

For 5 instances the probability scores per class were calculated using the `predict_proba` function, the results of which can be seen in table 5.3. $P(Y = 0|X)$ and $P(Y = 1|X)$ showing the probability of a data point being classified as 'incorrect' (0) or 'correct' (1), respectively. From this it can be seen that the probability scores vary per data point.

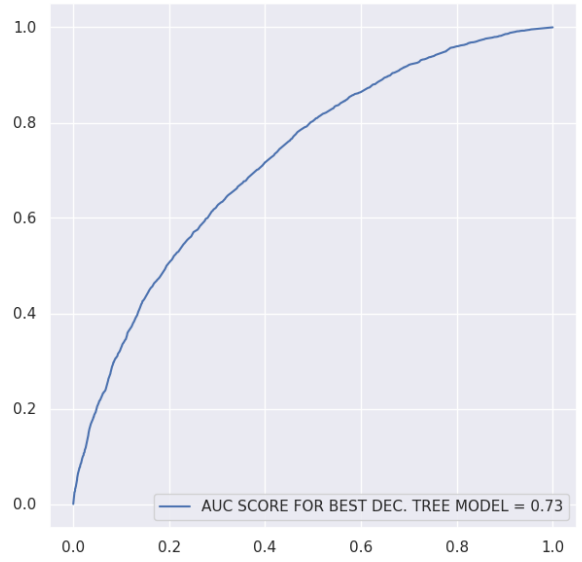


Figure 5.5: ROC curve for the best decision tree model.

For x_5, the probability of being sorted in class 'incorrect' is 0.97, a high probability score showing that the decision tree is very confident of its decision. While for x_3, the probability scores for classes 'correct' and 'incorrect' are close, showing that the model is not as sure of its classification as 'incorrect'.

5.2.2.2 Decision Tree features

The most influential features in predicting participants' recall ability for an item primarily pertain to the item variables. Specifically, the category and decade of the famous face demonstrate the highest feature importance within the decision tree, as evident in Table 5.4 and Figure 5.6. These two variables exhibit the most prominent peaks in the feature importance graph, surpassing other factors such as interest variables. Notably, certain interest variables, including `FilmTheatre_y_interest` (interest in Film & Theatre as a young adult) and 'Soccer_interests' (interest in soccer throughout a participants life), also contribute to the overall feature importance. `FilmTheatre_y_interest` is especially curious as the variable `FilmTheatre_mc_interest` was dropped from the dependent variables during the feature selection. This could possibly be due to these two variables having a high correlation, as was established in section 4.3.1.

It was initially hypothesized that the interaction between participants' interests and item categories would impact their ability to accurately recall an item. This hypothesis was validated by the decision tree, where `Soccer_interests` (indicating participants' interest in Soccer throughout a participants life), and `category_main_tekst_Sports` (representing whether an item belongs to the Sports category) are shown to interact. Here `Soccer_interests` is located at level 3, the 2nd node from the left, and `category_main_tekst_Sports` is located at level 4, the 3rd node from the left on the decision tree 5.8. A zoomed-in picture of the interaction can be seen in figure 5.7 When an item from the sports category is a soccer player, participants with high interest in soccer might be more likely to correctly recall the item.

Additionally, another feature of significant importance is age. It is plausible that this might be due to an increase in forgetfulness as a person ages[22], which might make it harder to recall an item correctly.

Table 5.4: Feature importance of top 5 most important variables for the decision tree classifier.

Feature	Importance
item category	0.126
item decade	0.106
FilmTheatre_y_interest	0.064
age	0.063
Soccer_interests	0.047

In summary, the decision tree model aimed at predicting whether a participant answered a specific item correctly revealed the significance of item variables, such as the decade and category of the famous face, as well as participants' interests. Importantly, these two types of variables were found to interact with each other. Moreover, the age of the participant emerged as a factor influencing participants' ability to accurately recall a face. Notably, a considerable number of features were deemed insignificant in determining the 'correct' and 'incorrect' classes, leading to their elimination during the feature selection process. Nevertheless, it is essential to acknowledge that the decision tree model achieved an accuracy of only 0.67, which falls short of perfection. Consequently, drawing definitive conclusions regarding the actual influence and magnitude of these variables on participants' ability to recall items accurately remains challenging.

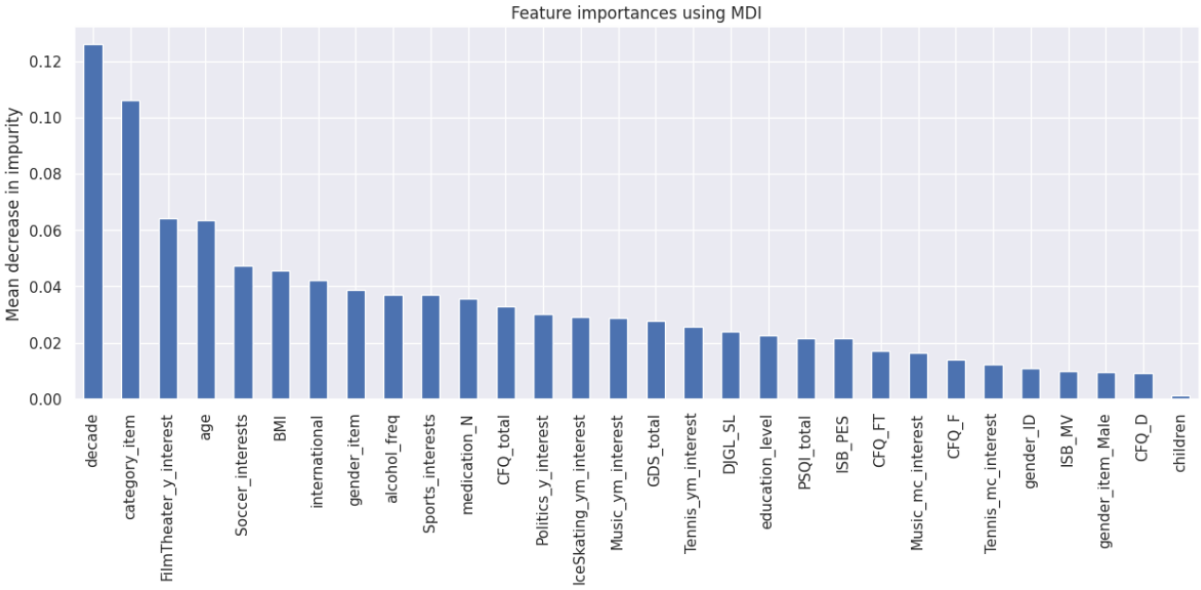


Figure 5.6: Feature importance for decision tree classifier.

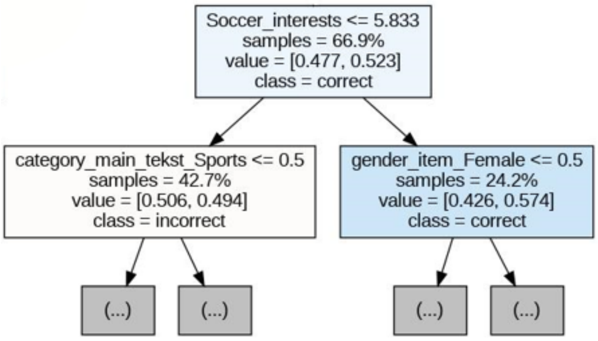


Figure 5.7: Interaction between interest in soccer and item being from the sports category, zoomed-in from the decision tree.

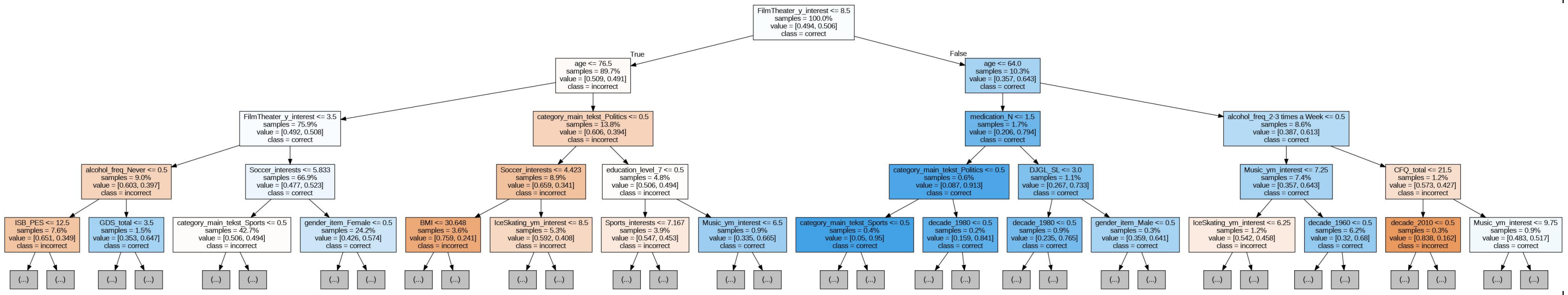


Figure 5.8: The first four levels of the decision tree classifier.

6

Conclusion

The primary objective of this research was to examine the impact of individual differences among participants on their performance in the Famous Faces Test (FFT), as well as the interaction between these individual differences and item characteristics. This investigation was conducted through statistical analyses on the datasets, exploring the data through clustering and random forests and lastly modeling the data to predict the total relative correct score and the correctness of each participant-item combination for recall.

The first research question aimed to shed light on how individual differences influence the overall performance on the FFT. When modeling the data, relationships were discovered between variables age and interest in Film & Theatre as a young adult, and the total relative correct score. However, the accuracy of the model in predicting the total relative correct score cast doubts on the validity of these findings.

In light of these inconclusive results, attention was turned to the prediction of participant clusters in Section 4.3.2. Here the participants were split into two clusters based on their similarity in performance on the FFT for recall. Through the application of a random forest algorithm on the participant data, it was successfully predicted which cluster a participant belonged to. The feature importance analysis pertaining to this prediction is presented in Figure 4.16.

To address the first research question, it appears challenging to predict the total relative correctness score solely based on the individual differences provided in the participant data. Consequently, no definitive conclusions can be drawn regarding the specific individual differences that influence FFT scores. As discussed in the results section 5.1, it is plausible that individual differences have limited influence on the overall score. However, when participants are grouped based on their performance similarities, the individual differences that distinguish between these groups can be extracted, as highlighted in Figure 4.16.

The second research question investigates the interaction between participants' individual differences and item characteristics, and how this interaction affects their ability to accurately recall famous faces. Partial insights into this interaction can be observed in Figure 5.8, as this shows the first 4 layers of the decision tree which predicts participants ability to recall an item correctly. Furthermore, Figure 5.6 provides an overview of how these variables influence participants' capacity to correctly recall items, with the category and decade of the item emerging as the most influential factors. This suggests that item characteristics outweigh participants' individual characteristics in terms of importance. The decision tree classifier created for research question 2 has interpretability on feature importance, as it shows exactly which features it splits on and the feature importance overall. Besides this, an advantage of the decision tree is its transparency on how well the model is able to predict the classes, as shown in figure 5.2. It can also show the chance of each data point being predicted in the classes 'incorrect' and 'correct', with the function `predict_proba`.

7

Discussion

The correlation between the total relative correct score on recall and individual differences among participants was found to be unreliable. This could be attributed to issues with data quality and modeling, or it could indicate a lack of a relationship between the two variables, as discussed in the results and conclusion sections. However, when participants were clsutered based on similarity in their recall performance, more reliable results were obtained regarding the prediction of the total score and its relationship with individual differences, as presented in Section 4.3.2.

The ability of participants to accurately recall an item is influenced by both their individual differences and the characteristics of the item, with the latter exerting a greater influence. These findings regarding the influence of individual differences and the interaction between participants and items on recall scores in the D-FFT can contribute to the further development of this test in clinical practice. When administering these tests, it is crucial to consider the influence of individual differences and the interaction between individual differences and item characteristics when evaluating a participant's performance. This consideration can prevent participants with lower scores, potentially resulting from their individual differences and interactions with the presented items, from being unjustly flagged as having early-stage Alzheimer's disease. Similarly, it is important to recognize cases where participants are expected to have higher scores than the norm. Failure to account for personal norms in expected recall scores may lead to the failure of identifying participants who may have early-stage Alzheimer's in these cases. Therefore, this research provides initial insights into the nature of these personal norms.

One limitation of this research is the small number of participants in the dataset. Figure 4.17 shows that only small clusters were identified in the data, suggesting that there may be limited participant-item combinations with similar patterns of correct recall. A larger dataset may reveal more patterns, leading to improved predictive models and more conclusive results regarding the variables that influence participants' performance on the D-FFT.

The individual differences used in this research were based on self-reported data, which introduces a potential source of bias. Participants may provide incorrect answers for certain variables, thus compromising the reliability of the data. Furthermore, self-reporting can introduce subjectivity, particularly with regard to scale questions that require participants to rank themselves. For instance, an "8" on an interest level scale may have different interpretations for each participant, making comparisons between participants less reliable.

To enhance the obtained results, future research can focus on achieving better predictive models to enhance the reliability of performance predictions on the FFT. This can be accomplished through two potential avenues: improving the input data and enhancing the prediction model.

Improving the input data involves increasing the volume of participant scores. The small dataset used in this research was identified as a limitation, as it may have restricted the model's ability to identify similar participant-item combinations and thus hindered accurate predictions. A larger dataset would allow for the identification of more patterns among participants.

Enhancing the prediction model is the second approach. One option is to employ a boosting model, which combines multiple weaker models to create a more robust predictive model. In this case, the weaker model would be the decision tree, which has limited predictive power. Boosting models work iteratively, with each new iteration aiming to correct the mistakes made by the previous weaker model. To rectify these mistakes, the subsequent model assigns greater importance to misclassified instances from the previous model. [26] Examples of popular boosting models include AdaBoost, XGBoost, and lightGBM.

AdaBoost is an adaptive boosting model that improves the model with each iteration by assigning higher weights to instances that were misclassified in the previous iteration. [26] Gradient boosting is another boosting model applicable to classification and regression models. In gradient boosting, each iteration fits the weaker model to the residuals, which represent the difference between the true and predicted values, of the loss function with respect to the current weaker model. Both XGBoost and lightGB are gradient boosting models. .[27] [28]

Future work could also focus on improving the visualizations provided in this research, to improve explainability. One way to do this, would be by employing SHAP, which is a visualization tool to gain insight into how predictive models arrive at their decision for regression and classification. Not only is SHAP able to visualize the feature importance for the entire dataset, but also which exact features contributed to the prediction of a single instance. SHAP can be used for boosting models, as well as decision trees.[29]

References

[1] Alzheimer Nederland. **Factsheet cijfers en feiten over dementie**. 1

[2] Douglas W. Scharre. **Preclinical, Prodromal, and Dementia Stages of Alzheimer's Disease**. 1

[3] et al. Vermunt L. **Duration of preclinical, prodromal, and dementia stages of Alzheimer's disease in relation to age, sex, and APOE genotype**. *Alzheimer Disease Neuroimaging Initiative.* 1

[4] Mayo Clinic. **Alzheimer's stages: How the disease progresses**, 29-04-2021. 1

[5] American Psychiatric Association. In *Diagnostic and statistical manual of mental disorders (5th ed.).* 1

[6] Alzheimer Nederland. **Diagnose dementie vaststellen**. 2

[7] healthcare brands Dementia.org. **Diagnoses dementia: the mini mental status exam (MSSE)**. 2

[8] Alzheimer Nederland. **MMSE test**. 2, 5

[9] M. Simard. **The mini-mental state examination: strengths and weaknesses of a clinical instrument.** *The Canadian Alzheimer Disease Review.* 2

[10] Annalena Venneri, Micaela Mitolo, and Matteo De Marco. **Paradigm shift: semantic memory decline as a biomarker of preclinical Alzheimer's disease**. *Biomarkers in Medicine*, **10**(1):5–8, 2016. PMID: 26642376. 2

[11] et al. Orlovsky I. **The relationship between recall of recently versus remotely encoded famous faces and amyloidosis in clinically normal older adults.** *Alzheimers Dement (Amst)*, **10**:121–129, 23-11-2017. 2, 3, 5, 39

[12] et al. Seidenberg M. **Recognition of famous names predicts cognitive decline in healthy elders**. *Neuropsychology.*, **27**(3):333–342, May 2013. 3, 5

[13] et al. Hays CC. **Temporal gradient during famous face naming is associated with lower cerebral blood flow and gray matter volume in aging**. *Neuropsychologia.*, **107**:76–83, Dec. 2017. 3, 5

[14] Hodges JR. Greene JD. **Identification of famous faces and famous names in early Alzheimer's disease. Relationship to anterograde episodic and general semantic memory**. *Brain.* 3

[15] Karl Pearson. **Note on regression and inheritance in the case of two parents**. *Proceedings of the Royal Society.* 26

[16] V.A. Profillidis and G.N. Botzoris. **Chapter 5 - Statistical Methods for Transport Demand Modeling**. In V.A. Profillidis and G.N. Botzoris, editors, *Modeling of Transport Demand*, pages 163–224. Elsevier, 2019. 26, 30

[17] Barrett O'Neill. **Chapter 1 - Calculus on Euclidean Space**. In Barrett O'Neill, editor, *Elementary Differential Geometry (Second Edition)*, pages 3–42. Academic Press, Boston, second edition edition, 2006. 30

[18] C. David J., Shook. **The application of cluster analysis in strategic management research: an analysis and critique**. *Strategic Management Journal.* 30

[19] Peter J. Rousseeuw. **Silhouettes: A graphical aid to the interpretation and validation of cluster analysis**. *Journal of Computational and Applied Mathematics.* 30

[20] Hannes Kisner, Yitao Ding, and Ulrike Thomas. **Chapter 4 - Capacitive material detection with machine learning for robotic grasping applications**. In Qiang Li, Shan Luo, Zhaopeng Chen, Chenguang Yang, and Jianwei Zhang, editors, *Tactile Sensing, Skill Learning, and Robotic Dexterous Manipulation*, pages 59–79. Academic Press, 2022. 31

[21] Farzin Saffarimiandoab, Riccardo Mattesini, Wanyi Fu, Ercan Engin Kuruoglu, and Xihui Zhang. **Insights on features' contribution to desalination dynamics and capacity of capacitive deionization through machine learning study**. *Desalination*, **515**:115197, 2021. 34

[22] Ballard J. **Forgetfulness and older adults: concept analysis**. *Journal of advanced nursing.* 34, 53

[23] Sebastian Kaiser. **BiCluster Algorithms (package 'biclust')**, May 2023. 35

[24] **Chapter 2 - One-Year PD**. In Tiziano Bellini, editor, *IFRS 9 and CECL Credit Risk Modelling and Validation*, pages 31–89. Academic Press, 2019. 40

[25] Martin Kryzywinski and Naomi Altman. **Classification and regression trees**. *Nature Methods.* 41

[26] Robert E. Schapire. **A Brief Introduction to Boosting**. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, page 1401–1406, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. 59

[27] Tianqi Chen and Carlos Guestrin. **XGBoost**. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, aug 2016. 59

[28] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. **LightGBM: A Highly Efficient Gradient Boosting Decision Tree**. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, **30**. Curran Associates, Inc., 2017. 59

[29] Scott Lundberg and Su-In Lee. **A Unified Approach to Interpreting Model Predictions**, 2017. 60

Appendix

Dropped features:

1. `Cycling_interests`
2. `retired_works`
3. `IceSkating_mc_interest`
4. `Politics_mc_interest`
5. `retired_selfimage`
6. `smoking`
7. `CCI_total`
8. `household`
9. `DJGL_total`
10. `DJGL_EL`
11. `children_N`
12. `alcohol_quan`
13. `maritalstatus`
14. `grandchildren_N`
15. `grandchildren`
16. `bloodpressure`
17. `FilmTheatre_mc_interest`
18. `ISB_AES`
19. `medication`
20. `retirement`
21. `birthcountry`

22. `mothertongue`