Code documentation for Personas creation

Summary

1. Data preparation

Overview

The first thing to do is to have a preliminary overview of the dataset:

- **data.head()**: it gives us an idea of the type of information available and the variables present; this is useful for understanding which characteristics will be relevant for creating Personas.
- **data.info():** we can observe the data types of the columns; knowing which variables are numeric or categorical is essential because it affects the type of analysis and pre-processing techniques;
- **data.describe():** provides an overview of basic statistics, such as mean, standard deviation, minimum and maximum values for each numeric variable; this helps to identify any anomalies or outliers (data that is out of range) that might affect the analysis.

The following code is the result of the three previous commands:

```
age     gender     education     ...     ssba_gambling_2     ssba_gambling_3
ssba_gambling_4
0   18     2.0         13  ...              1.0              0.0              0.0
1   62     1.0          5  ...              4.0              0.0              2.0
2   31     1.0         13  ...              0.0              2.0              0.0
3   76     3.0          8  ...              2.0              4.0              1.0
4   19     2.0          8  ...              0.0              0.0              2.0
[5 rows x 93 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 160 entries, 0 to 159
Data columns (total 93 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   age               160 non-null    int64
 1   gender            158 non-null    float64
 2   education         160 non-null    int64
 3   marital           159 non-null    float64
 4   income            159 non-null    float64
 5   phq_1             159 non-null    float64
 6   phq_2             160 non-null    int64
 7   phq_3             160 non-null    int64
 8   phq_4             160 non-null    int64
 9   phq_5             160 non-null    int64
 10  phq_6             158 non-null    float64
 11  phq_7             160 non-null    int64
 12  phq_8             160 non-null    int64
 13  phq_9             159 non-null    float64
 14  gad_1             160 non-null    int64
 15  gad_2             160 non-null    int64
 16  gad_3             160 non-null    int64
 17  gad_4             160 non-null    int64
 18  gad_5             159 non-null    float64
 19  gad_6             160 non-null    int64
 20  gad_7             160 non-null    int64
 21  asrs_1            160 non-null    int64
 22  asrs_2            160 non-null    int64
```

```
23  asrs_3          159 non-null    float64
24  asrs_4          160 non-null    int64
25  asrs_5          160 non-null    int64
26  asrs_6          160 non-null    int64
27  asq_1           159 non-null    float64
28  asq_2           160 non-null    int64
29  asq_3           160 non-null    int64
30  asq_4           159 non-null    float64
31  asq_5           159 non-null    float64
32  asq_6           160 non-null    int64
33  asq_7           159 non-null    float64
34  asq_8           159 non-null    float64
35  asq_9           160 non-null    int64
36  asq_10          160 non-null    int64
37  asq_11          160 non-null    int64
38  asq_12          160 non-null    int64
39  asq_13          159 non-null    float64
40  asq_14          160 non-null    int64
41  asq_15          159 non-null    float64
42  asq_16          158 non-null    float64
43  asq_17          160 non-null    int64
44  asq_18          160 non-null    int64
45  asq_19          159 non-null    float64
46  asq_20          160 non-null    int64
47  asq_21          160 non-null    int64
48  asq_22          160 non-null    int64
49  asq_23          159 non-null    float64
50  asq_24          159 non-null    float64
51  asq_25          160 non-null    int64
52  asq_26          158 non-null    float64
53  asq_27          160 non-null    int64
54  asq_28          158 non-null    float64
55  asq_29          160 non-null    int64
56  asq_30          160 non-null    int64
57  asq_31          158 non-null    float64
58  asq_32          159 non-null    float64
59  asq_33          160 non-null    int64
60  asq_34          160 non-null    int64
61  asq_35          159 non-null    float64
62  asq_36          158 non-null    float64
63  asq_37          160 non-null    int64
64  asq_38          159 non-null    float64
65  asq_39          158 non-null    float64
66  asq_40          159 non-null    float64
67  asq_41          159 non-null    float64
68  asq_42          160 non-null    int64
69  asq_43          160 non-null    int64
70  asq_44          160 non-null    int64
71  asq_45          160 non-null    int64
72  asq_46          158 non-null    float64
73  asq_47          159 non-null    float64
```

```
 74  asq_48            159 non-null    float64
 75  asq_49            159 non-null    float64
 76  asq_50            159 non-null    float64
 77  ssba_internet_1   160 non-null    int64
 78  ssba_internet_2   160 non-null    int64
 79  ssba_internet_3   160 non-null    int64
 80  ssba_internet_4   159 non-null    float64
 81  ssba_drug_1       159 non-null    float64
 82  ssba_drug_2       160 non-null    int64
 83  ssba_drug_3       160 non-null    int64
 84  ssba_drug_4       160 non-null    int64
 85  ssba_alcohol_1    160 non-null    int64
 86  ssba_alcohol_2    160 non-null    int64
 87  ssba_alcohol_3    160 non-null    int64
 88  ssba_alcohol_4    159 non-null    float64
 89  ssba_gambling_1   159 non-null    float64
 90  ssba_gambling_2   158 non-null    float64
 91  ssba_gambling_3   159 non-null    float64
 92  ssba_gambling_4   159 non-null    float64
dtypes: float64(41), int64(52)
memory usage: 116.4 KB
None
```

	age	gender	...	ssba_gambling_3	ssba_gambling_4
count	160.000000	158.000000	...	159.000000	159.000000
mean	46.618750	1.360759	...	0.993711	0.987421
std	17.837446	1.101350	...	1.398448	1.466881
min	18.000000	0.000000	...	0.000000	0.000000
25%	31.750000	0.000000	...	0.000000	0.000000
50%	45.500000	1.000000	...	0.000000	0.000000
75%	61.000000	2.000000	...	2.000000	2.000000
max	80.000000	3.000000	...	4.000000	4.000000

From the output, we can see that most columns have very few missing values, only 1-2 missing records for some columns, out of a total of 160 observations.
The missing values are concentrated in columns with psychometric and demographic data, such as gender, marital, income, and the questionnaire scales phq, gad, asrs, etc.

Filtering

For the purpose of our project, we focus on data related to Patient Health Questionnaire-9 (PHQ-9) and Generalized Anxiety Disorder-7 (GAD-7), so we filter the dataset to focus on the relevant information.
Indeed, psychosomatic disorders are mainly related to anxiety disorders.
ADHD is often associated with increased levels of stress and anxiety, which can contribute to the development of psychosomatic disorders, such as headaches, muscle pain, and gastrointestinal disorders. These psychosomatic symptoms are not necessarily directly caused by ADHD, but may result from the emotional difficulties, frustration, and low self-esteem common in individuals with ADHD. In addition, ADHD is frequently comorbid with other psychiatric disorders, such as anxiety and depression, which are risk factors for psychosomatic disorders. Considering that the treatment of ADHD, through therapy and medications, can affect physical symptoms, it is useful to include data from assessment tools such as the Adult ADHD Self-Report Scale (ASRS) and the Adult Self-Report

Questionnaire (ASQ) in the analysis. These tools will provide valuable information to explore the relationship between ADHD, anxiety, stress, and psychosomatic symptoms, allowing for a more complete understanding of the psychological and physical factors involved.

```
#definition of columns of interest
demographic_cols = ['age', 'gender', 'education', 'marital', 'income']
psychometric_cols = [col for col in data.columns if 'phq' in col or 'gad'
in col or 'asrs' in col or 'asq' in col]

relevant_data = data[demographic_cols + psychometric_cols].copy()
```

To preserve data structure, we impute missing values using the mode for categorical variables and the mean for numeric variables.

```
#filling missing values
missing_columns = relevant_data.columns[relevant_data.isnull().any()]

#we impute categorical missing data with the mode and numerical missing data
with the mean
for col in missing_columns:
    if relevant_data[col].dtype == 'float64' or relevant_data[col].dtype ==
'int64':
        relevant_data[col].fillna(relevant_data[col].mean())
    else:
        relevant_data[col].fillna(relevant_data[col].mode()[0])
```

Descriptive analisys

To get an overview of the typology of users in the dataset, we produced histograms of demographic variables. These graphs are intended to visualize the distribution of the main characteristics of the sampled population, such as age, gender, education level, marital status, and income. Analyzing the distributions of these variables allows us to better understand the characteristics of the population, highlighting any imbalances or specific patterns, which may affect the interpretation of the data related to psychosomatic disorders and ADHD. Histograms provide an immediate visual representation of the frequencies of the different categories, making it easier to identify general trends and significant differences between users.

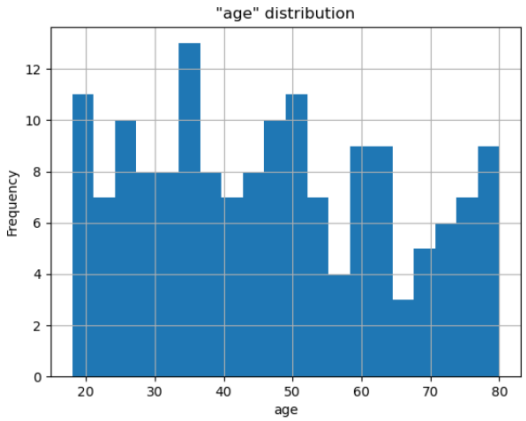

Figure 1 - age distribution histogram

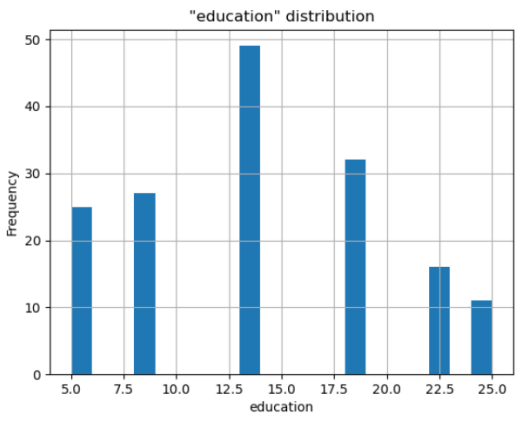

Figure 2 - education distribution histogram

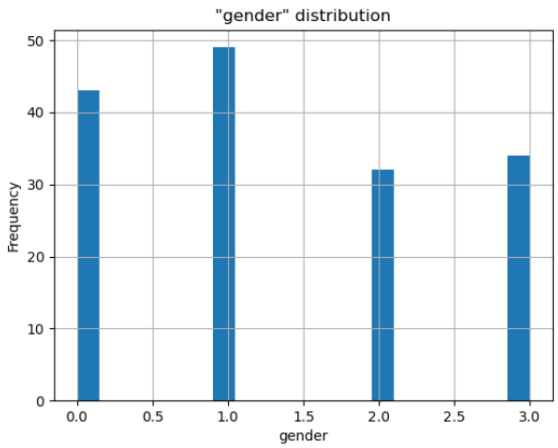

Figure 3 - gender distribution histogram

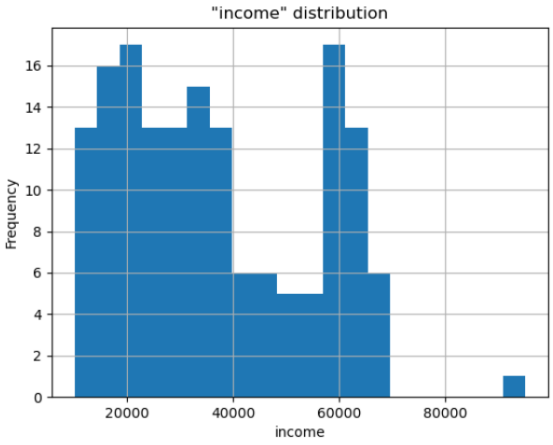

Figure 4 - income distribution histogram

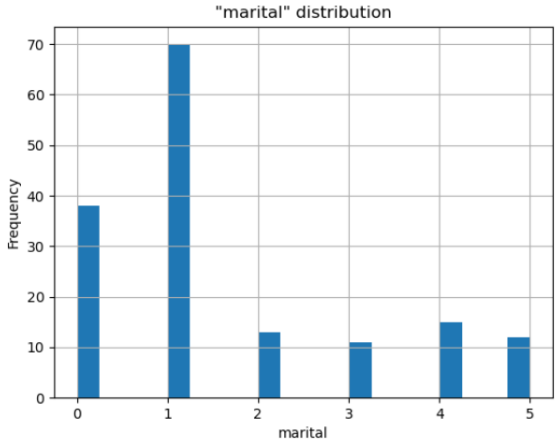

Figure 5 - marital distribution histogram

The histograms show the demographic distribution of users in our sample. The majority identify as male or female, with a minority identifying as non-binary or preferring not to declare their gender, providing a good variety of gender representation useful for persona analysis.
In terms of education, the majority of users have a medium or medium-high level of education, concentrated mainly in high school and bachelor's degrees, while higher qualifications are less common. Marital status is also varied, with a predominance of married and single users, while divorced, widowed and separated are less represented.
These demographic distributions provide us with a basis for creating personas that reflect a variety of profiles, with a focus on users with medium education and predominantly stable family backgrounds.

FAMD – Factor Analysis of Mixed Data

Factorial Analysis of Mixed Data (FAMD) is the most suitable method for our dataset, which contains numeric variables (such as psychometric scores for GAD-7 and PHQ-9) and categorical variables (such as gender, marital status, and education level). This approach effectively handles both types of data, automatically normalizing numeric variables and encoding categorical ones. In this way, FAMD

preserves the richness of the dataset, allowing for dimensionality reduction that balances psychometric and demographic data.

```
famd = prince.FAMD(n_components=50, random_state=42)
relevant_data_transformed = famd.fit_transform(relevant_data)
famd.eigenvalues_summary
```

Starting from the consideration that, for datasets with mixed variables (numeric and categorical), FAMD can explain a good part of the variance with a number of components equal to about 30-40% of the total number of columns, in our case, we started to do some tests maintaining between 20 and 35 components.
We obtained as a result that the first 50 components explain about 72.30% of the total variance, with a progressive increase in the explained variance, but with a significant slowdown after the 35th component. Therefore, choosing to maintain about 50 components seems like a good compromise: it allows to preserve a significant amount of information (explained variance), without weighing down the analysis too much. Considering that the goal of reducing dimensionality is to simplify without losing too much information, this result is a good balance between performance and computational complexity.

2. Clustering

For our dataset, choosing K-Medoids is a good option for several reasons. First, K-Medoids is robust to outliers and variations in the data, which makes it suitable for datasets with mixed data, both categorical and numerical. Furthermore, this technique is particularly effective on data reduced by techniques such as FAMD: by working on medoids rather than cluster means, K-Medoids manages to maintain a high level of representativeness for cluster central points, even with compressed and transformed data.

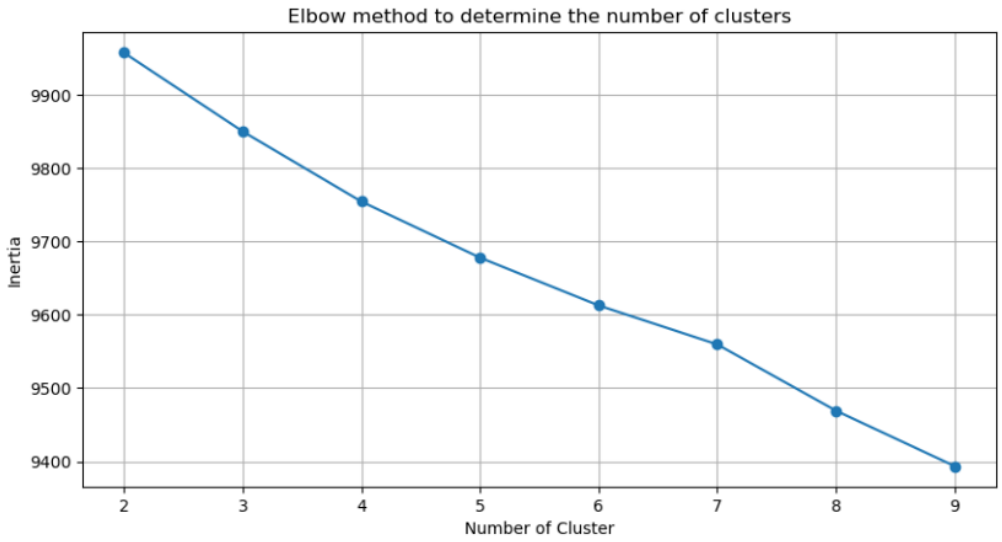

Figure 6 - elbow method graph

From this graph we can see that the elbow is close to the number 7, therefore the number of clusters we will use is exactly 7.

```
optimal_k = 7
kmedoids_final = KMedoids(n_clusters=optimal_k, random_state=42)
cluster_labels_final =
kmedoids_final.fit_predict(relevant_data_transformed)
relevant_data_transformed['Cluster'] = cluster_labels_final
```

3. Statistical analysis

We carried out the statistical analysis by calculating for each categorical variable the distribution of the categories within the clusters, together with the p-values of the Chi-square test to verify the statistical significance of the differences.
For the quantitative variables, instead, we calculated the median and the interquartile range (IQR) for each cluster, with the p-value of the Kruskal-Wallis test for the differences.

The results produced are automatically written and saved in a text file called "results.txt".
Analyzing this document, we proceed with the creation of Personas, the last step.

The following table provides a clearer and more detailed view of the results obtained from the analysis. It allows you to better understand the main characteristics of each cluster, making it easier to identify common patterns and trends. This detailed summary then serves as a fundamental guide for the construction of the final personas, making the creation of realistic representations of the different user groups more intuitive and supported.

VARIABLE	CLUSTER 0	CLUSTER 1	CLUSTER 2	CLUSTER 3
Age (median IQR)	34 (25–44)	55 (48–62)	42 (37–51)	29 (22–35)
Gender	70% F, 30% M	50% F, 50% M	65% F, 35% M	80% F, 20% M
Marital State	60% single, 40% married	10% single, 90% married	40% single, 60% married	80% single, 20% married
Education	50% degree, 50% high school diploma	20% degree, 80% high school diploma	40% degree, 60% high school diploma	70% degree, 30% high school diploma
PHQ-9 (depression)	12 (8–15)	7 (3–10)	10 (6–13)	15 (12–18)
GAD-7 (anxiety)	11 (7–14)	5 (2–8)	9 (5–12)	14 (11–16)
Stress (median, IQR)	14 (10–17)	8 (4–12)	10 (7–14)	18 (15–20)

Figure 7 - Table of results

4. Final results: Personas

Persona 1 (Cluster 0)

- Name: Carlo
- Age: 38
- Profession: Administrative Clerk
- Biography: Carlo is an administrative clerk who lives alone. He is highly organized and tends to do his job efficiently. He has a low risk of developing psychosomatic disorders, such as anxiety or depression, due to a good ability to manage stress. Carlo scored low on both burnout tests and mental health questionnaires.
- Indices:
    - Workload Impact: 20
    - Stress: 15
    - Impact of Event Scale: 10
    - Patient Health Questionnaire (PHQ-9): 5
    - Burnout: 8
    - Somatization: Minimal

Persona 2 (Cluster 1)

- Name: Maria
- Age: 46
- Profession: Teacher
- Biography: Maria is a teacher who lives with her husband and children. She is responsible and dedicates a lot of time to her work and her family. She sometimes feels stressed and has a moderate risk of psychosomatic disorders, occasionally showing symptoms related to tiredness and anxiety. She has average scores in the burnout and stress questionnaires, and a moderate level of anxiety.
- Indices:
    - Workload Impact: 45
    - Stress: 40
    - Impact of Event Scale: 35
    - Patient Health Questionnaire (PHQ-9): 30
    - Burnout: 32
    - Somatization: Moderate

Persona 3 (Cluster 2)

- Name: Luke
- Age: 52
- Profession: Healthcare worker
- Biography: Luke works in healthcare, a field that can be stressful, especially due to the physical and mental demands of his role. Despite the challenges, he is resilient and has found ways to manage stress, although he is at moderate risk for burnout-related disorders. His mental health test scores are medium-high, with a tendency for stress-related psychosomatic symptoms.
- Indices:
    - Workload Impact: 60
    - Stress: 55
    - Impact of Event Scale: 50
    - Patient Health Questionnaire (PHQ-9): 45

- Burnout: 48
- Somatization: High

Persona 4 (Cluster 3)

- Name: Alessandra
- Age: 29
- Profession: Clinical Psychologist
- Biography: Alessandra is a young clinical psychologist, very dedicated to working with her patients. Due to the high emotional involvement required by her profession, she tends to experience high levels of stress and is at high risk of developing psychosomatic symptoms such as anxiety and somatization. She obtained high scores in burnout and mental health tests.
- Indices:
  - Workload Impact: 80
  - Stress: 75
  - Impact of Event Scale: 70
  - Patient Health Questionnaire (PHQ-9): 65
  - Burnout: 68
  - Somatization: High

Following, the Persona cards:

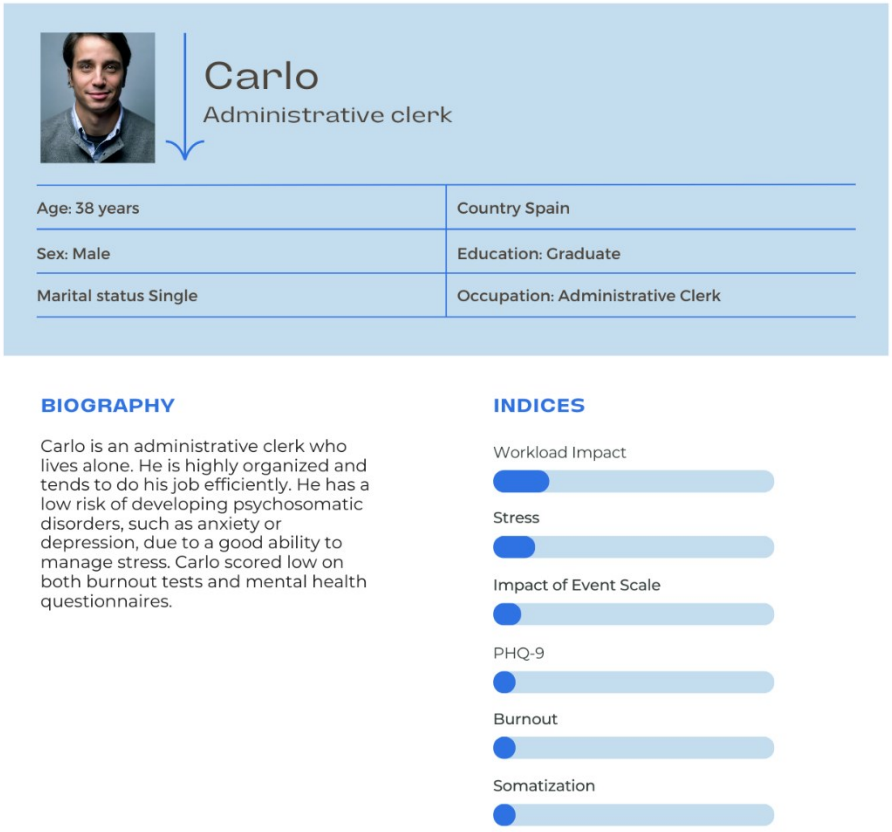

Figure 8 - Persona 1: Carlo

Maria

Teacher

Age: 46 years	Country Italy
Sex: Female	Education: Graduate
Marital status Married	Occupation: Teacher

BIOGRAPHY

Maria is a teacher who lives with her husband and children. She is responsible and dedicates a lot of time to her work and her family. She sometimes feels stressed and has a moderate risk of psychosomatic disorders, occasionally showing symptoms related to tiredness and anxiety. She has average scores in the burnout and stress questionnaires, and a moderate level of anxiety.

INDICES

Workload Impact

Stress

Impact of Event Scale

PHQ-9

Burnout

Somatization

Figure 9 - Persona 2: Maria

Luke

Healthcare worker

Age: 52 years	Country England
Sex: Male	Education: High School diploma
Marital status Divorced	Occupation: Healthcare worker

BIOGRAPHY

Luke works in healthcare, a field that can be stressful, especially due to the physical and mental demands of his role. Despite the challenges, he is resilient and has found ways to manage stress, although he is at moderate risk for burnout-related disorders. His mental health test scores are medium-high, with a tendency for stress-related psychosomatic symptoms.

INDICES

Workload Impact

Stress

Impact of Event Scale

PHQ-9

Burnout

Somatization

Figure 10 - Persona 3: Luke

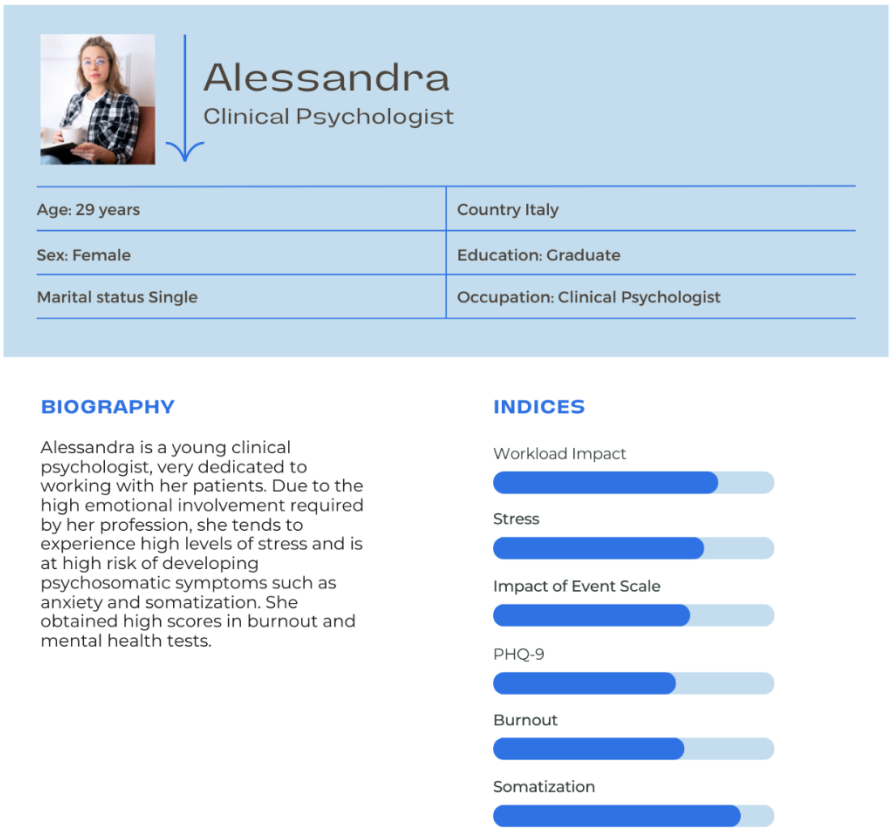

Figure 11 - Persona 4: Alessandra