ANALYSIS OF THE BEHAVIOUR OF THE INDIVIDUAL IN RELATION TO THEIR IMPACT ON THE ENVIRONMENT

University of Palermo, CdLM Data Algorithms and Machine Intelligence,
course Big Data Management

Francesca Maria Palazzotto, Giuseppe Termerissa, Pierfrancesco Martinello

INTRODUCTION

Nowadays, the theme of sustainability is one of the most discussed and of significant importance to which all public and private administrations continue to improve and to be part of the daily life of every citizen. The EU Commission has defined the Green Public Procurement (GPP) in which the environmental criteria in all the phases of the purchase process have been integrated, encouraging the diffusion of environmental technologies and the development of products valid from the ecological point of view. The United Nations, within the 2030 Agenda, has set the Sustainable Development Goals (SDGs), the 17 sustainable development goals, on the basis of which we have built our research.

The UI Green Metric is the world's first ranking to address sustainability concerns. Specifically, it was launched to assess universities on sustainability policies and environmental impact. Based on these indexes, we have structured and adapted useful information to put participants through sustainable questions about their everyday lives. In particular, we exploited the following categories:

- Setting and Infrastructure
- Energy and Climate Change
- Waste
- Water
- Transportation
- Education and Research

In addition to these categories, we have included a focus on product purchasing habits, in the categories of Shopping, Clothing, and Home, where information is collected about citizens' daily habits and how their purchases have a sustainable impact.

QUESTIONNAIRE AND INDEXES

Phase one is data collection. A questionnaire was developed consisting of five main sections:

- general information
- shopping
- mobility
- energy
- education and research

Each section has been carefully constructed from the indices described above with the aim of collecting as much helpful information as possible to obtain an estimate of the sustainable behavior of the average citizen and the possible causes that can lead the individual to a non-sustainable life.

Each question is labeled in order to create the indices used during the analysis and it is possible to see each specific question in Appendix 1. The indices used were created in order to cover the four main sections we were interested in: Purchases, Mobility, Energy and, Education and Research. We have made a total of 25 indices, so the entire vector used during the data analysis phase has 25 dimensions. Each index has its own meaning in order to understand a certain behavior of people involved and was created as follows.

PURCHASE INDEXES

Index	***Calculation***	***Meaning***
i_S	$\frac{\sum_{i=1}^{9} S_i}{1450}$	Total score of the grocery shopping
i_V	$\frac{\sum_{i=1}^{9} V_i}{550}$	Total score of clothing shopping
i_C	$\frac{C_{13}+C_{14}}{350}$	Total score of house shopping
i_tot_S	$(\sum_{i=3}^{9} S_i) \cdot \frac{S_2}{S_1}$	Total score of grocery shopping times the ratio between the number of times the user does the grocery and the number of supermarkets visited
i_tot_V	$V_{10} \cdot \frac{V_{11}+V_{12}}{400}$	Score about clothing shopping times the number of times the user does shopping
i_tot_C	$\frac{C_{13}}{C_{14}}$	Ratio between the behavior of house shopping and the interest of owning a green house

EDUCATION AND RESEARCH INDEXES

Index	***Calculation***	***Meaning***
i_ER	$\frac{ER_1+ER_2+ER_3}{500}$	Total score of education and research events (work, university)

MOBILITY INDEXES

Index	***Calculation***	***Meaning***
i_M1	m_9/m_8	Ratio of open space area to total area (personal)
i_M2	m_{18}/m_{17}	Ratio of open space area to total area (work)
i_M3	$m_8 \,/\, family\ size$	Total home area divided by family size
i_M4	$m_{12} \,/\, (m_{11} + m_{12} + m_{13})$	Ratio of sustainable energy consumption over general energy consumption
i_M5	m_{20}	Scoring the number of initiatives in workplace
i_M6	m_{19}	Home-work path quality
i_M7	$m_5 \,/\, m_4$	Ratio of vehicle-less trips over total weekly trips

ENERGY INDEXES

Index	***Calculation***	***Meaning***
i_e1	$eh2/5$	Efficient control systems owned at home over the total
i_e2	$ew2/5$	Efficient control systems owned by the workplace/university over the total
i_e3	$eh3 \cdot eh4$	Renewable sources owned at home times percentage of expenses avoided using renewable sources
i_e4	$ew3 \cdot ew4$	Renewable sources owned by the workplace/university times percentage of expenses avoided using renewable sources
i_e5	$eh1/eh5$	Ratio between the percentage of efficient products owned at home and the kWhs supplied by the systems
i_e6	$eh6/6$	Number of Green Building Implementation properties observed at home divided by the total
i_e7	$ew5/6$	Number of Green Building Implementation properties observed at the workplace/university divided by the total

WATER INDEXES

Index	***Calculation***	***Meaning***
i_e8	$(wh1 + wh2 + wh3 + wh4)/4$	Mean of all the percentages regarding the water waste for the house
i_e9	$(ww1 + ww2 + ww3 + ww4)/4$	Mean of all the percentages regarding the water waste for the workplace/university

WASTE INDEXES

Index	***Calculation***	***Meaning***
i_e10	$(waste1 + wasteh2 + wasteh3 + wasteh4)/4$	Mean of all the percentages regarding the waste habits for the house
i_e11	$(wastew1 + wastew2 + wastew3 + wastew4)/4$	Mean of all the percentages regarding the waste habits for the workplace/university

NORMALIZATION

Once created the dataframe containing the values of each index, we obtained different ranges for each column. For this reason, we made a normalization of the overall using the technique called *decimal scaling normalization,* in which we used the maximum value that can be obtained for each index and dividing for it the value of the corresponding index. In some cases, we used the maximum value of the column as there was no theoretical and efficient maximum value. In this way, we arranged the dataframe with values between 0 and 1.

COLLABORATIVE FILTERING

At this point, we were interested in applying the collaborative filtering technique, dividing the dataframe into a training set (80%) and a test set (20%). The *ALS - Alternating Least Squares* algorithm was used in order to estimate the ratings. It was applied to training data to build the recommendation model using as parameters 5 maximum number of iterations to run, 0.01 as the regularization parameter, and "`drop`" as the cold start strategy, which drops any rows in the dataframe of predictions that contains `NaN` values and to ensure we do not get `NaN` evaluation metrics.

To evaluate the model, it was used the *RMSE - Root-Mean-Square Error* on the test data, commonly used to measure the differences between values predicted by a model. This value calculated on our normalized dataframe, gives us a range of values between 2.50 and 5.50. Looking deep into our data, these values are consistent and accurate. This information can be seen in the recommendations given both for users and items by the model. For each feature, we have seen how the model recommends to each user with such an accurate evaluation close to 0 (see Figure 1).

	feature_id	recommendations
0	20	[(51, 0.76), (22, 0.72), (12, 0.59), (16, 0.52...
1	10	[(48, 0.24), (0, 0.24), (22, 0.23), (31, 0.18)...
2	0	[(22, 0.89), (26, 0.78), (23, 0.76), (16, 0.73...
3	1	[(32, 0.83), (47, 0.7), (36, 0.65), (34, 0.62)...
4	21	[(22, 0.92), (48, 0.89), (51, 0.61), (26, 0.59...
5	11	[(22, 0.54), (45, 0.44), (48, 0.36), (43, 0.33...
6	12	[(27, 0.96), (38, 0.89), (1, 0.89), (32, 0.88)...
7	22	[(22, 0.85), (48, 0.74), (33, 0.67), (51, 0.6)...
8	2	[(47, 0.85), (48, 0.77), (33, 0.71), (32, 0.6)...
9	13	[(32, 0.81), (47, 0.79), (7, 0.68), (45, 0.68)...
10	3	[(47, 0.39), (48, 0.36), (33, 0.27), (17, 0.24...
11	23	[(22, 1.03), (12, 0.98), (6, 0.97), (33, 0.96)...
12	4	[(47, 0.84), (33, 0.81), (26, 0.8), (48, 0.73)...
13	24	[(22, 1.05), (6, 0.98), (11, 0.98), (33, 0.98)...
14	14	[(27, 0.55), (16, 0.55), (17, 0.48), (50, 0.4)...
15	5	[(50, 0.78), (32, 0.69), (45, 0.58), (12, 0.51...
16	15	[(27, 0.8), (51, 0.66), (16, 0.66), (17, 0.64)...
17	6	[(27, 0.88), (17, 0.81), (51, 0.75), (22, 0.57...
18	16	[(48, 0.2), (22, 0.2), (32, 0.14), (34, 0.14),...
19	17	[(27, 0.45), (17, 0.27), (38, 0.25), (51, 0.23...
20	7	[(3, 0.82), (0, 0.72), (16, 0.7), (27, 0.69), ...
21	8	[(22, 1.14), (32, 0.98), (31, 0.96), (3, 0.95)...
22	18	[(3, 0.98), (1, 0.97), (49, 0.91), (45, 0.89),...
23	19	[(22, 0.6), (12, 0.55), (18, 0.47), (46, 0.44)...
24	9	[(37, 1.27), (33, 1.17), (43, 1.09), (23, 1.05...

Figure 1: collaborative filtering model recommendations

K-MEANS

During this phase, we wanted to analyze how the data can be clustered. We chose to use the *K-Means* algorithm, a prototype-based clustering technique to partition the data objects.

We couldn't use any kind of graphical technique in order to determine the value of *k*, the number of clusters to obtain, due to the high amount of features (the 25 indices). Insofar, we decided to use a value of k equal to 3, in order to divide people into three main groups: eco-friendly, eco-unfriendly, and eco-neutral. Tuning the algorithm, we obtained data divided as follows:

- 22 eco-friendly
- 7 eco-unfriendly
- 23 eco-neutral

In view of evaluating the clusters obtained, we studied the separation between them through the value of the average of each dimension for each cluster and the distance between each other. Then we plotted the values obtained, as we can see in the following Figure 2.

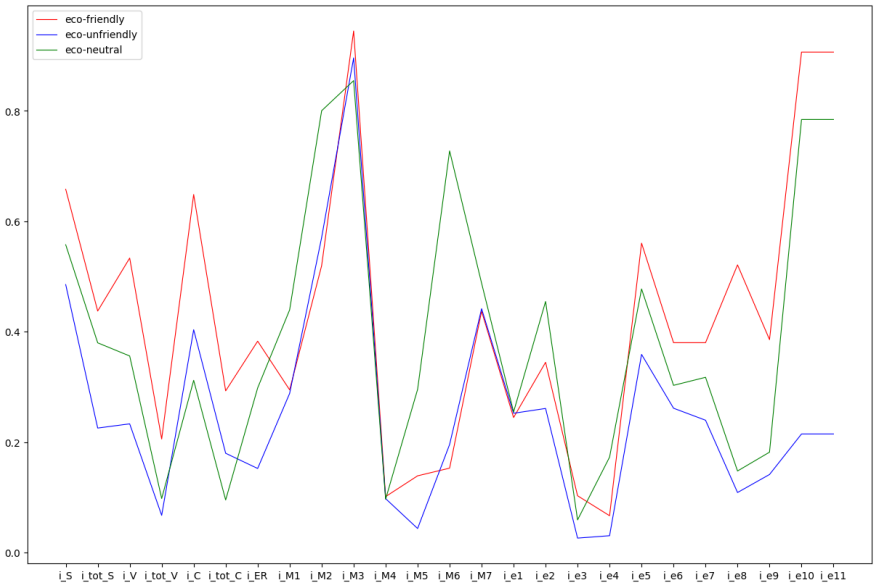

Figure 2: separation between clusters based on the average of dimensions

As we can see, we obtained well-separated clusters in which there are some overlapping around the indexes i_M2, i_M3, and i_M4 that are relative to the ratio of open space areas to total one, the total home area divided by family members and ratio of sustainable energy consumption over general energy consumption, and for i_e1 that is relative to the number of efficient systems owned at home. Both these "collisions" are expected, since the similar answers respect what can be found in a real population.

ANCILLARY VISUALIZATIONS

As the first visualization, we plotted the heatmap in order to see the correlations between each index and see how they are correlated. Looking at Figure 3, we considered and studied all the features whose correlation is higher or equal to 0.5.

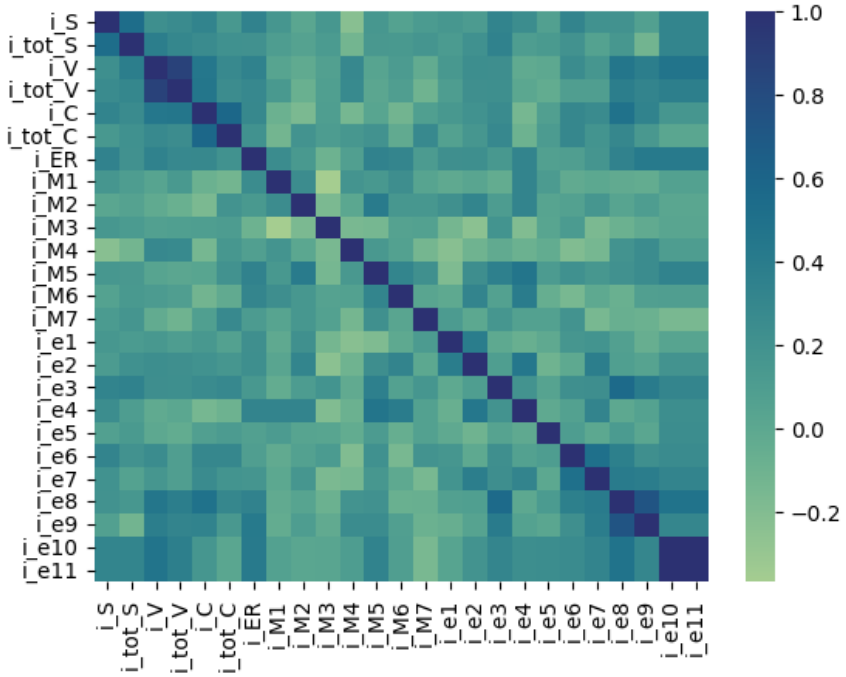

Figure 3: heatmap of dimensions/indexes

As we can see, there are three main correlations between the following features:

- i_e1, i_e3, i_e6, i_e2, i_e4, and i_e7 regarding energy consumption Figure 4: it is possible to see how these features are correlated through an asymmetric distribution (by using the kde visualization) and the correlation between couples of indexes in the non-diagonal graphics
- i_e8, i_e9, i_e10, and i_e11 regarding water and waste habits Figures 5, 6, 7: it is possible to appreciate how there is a very little correlation regarding the habits on water consumption at home and in the workplace/study place, and on the other hand there is a very high correlation regarding the habits on waste for the two considered environments. The difference between the two different habits is shown in detail in Figures 5 and 6.
- i_S, i_tot_S, i_V, i_tot_V, i_C, and i_tot_C regarding purchases Figure 8: it is possible to see how distributions (using both kde and hist) are almost symmetric, showing that the typical behavior of a user is consistent each time.

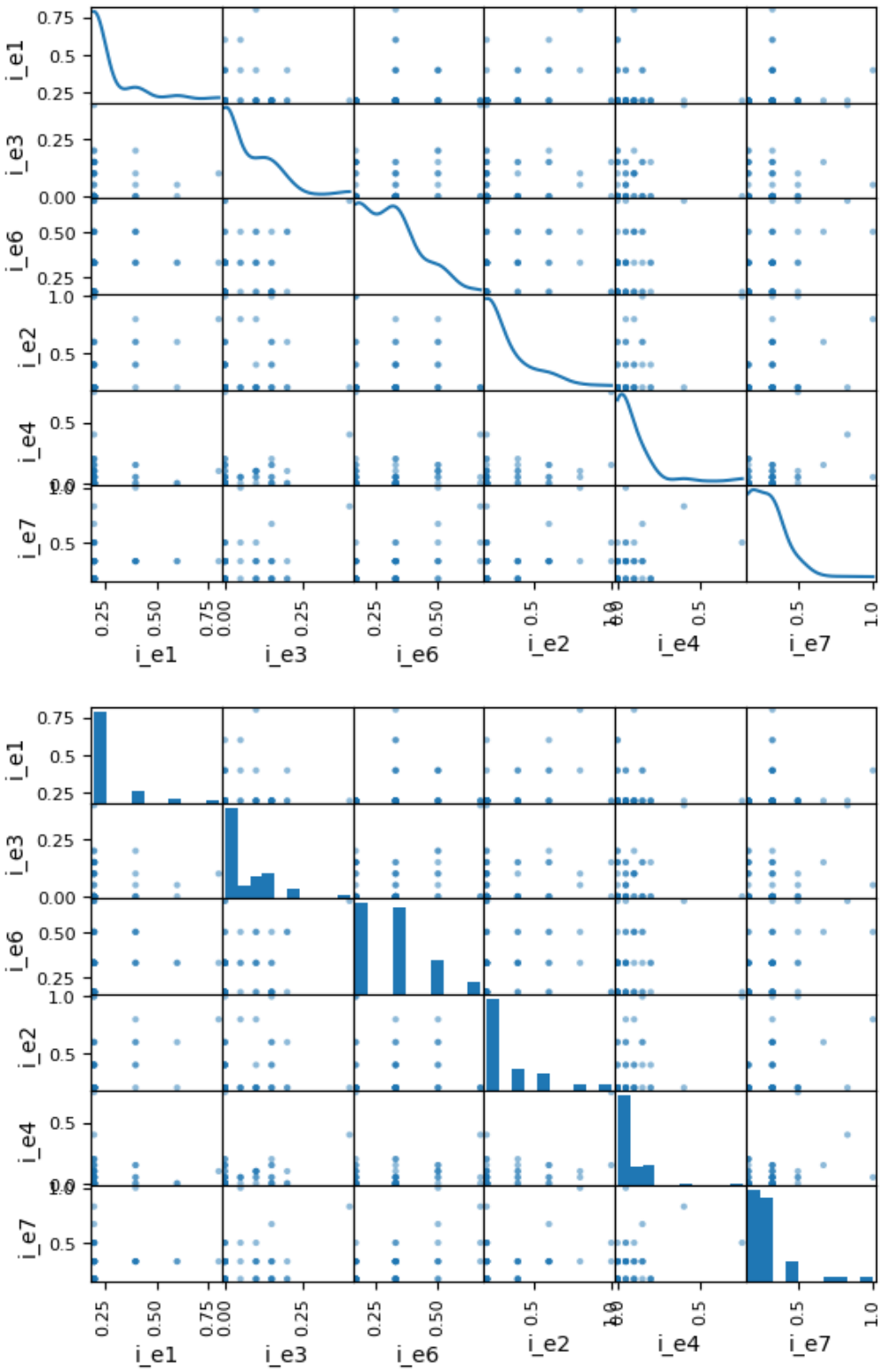

Figure 4: up - kde, bottom - hist

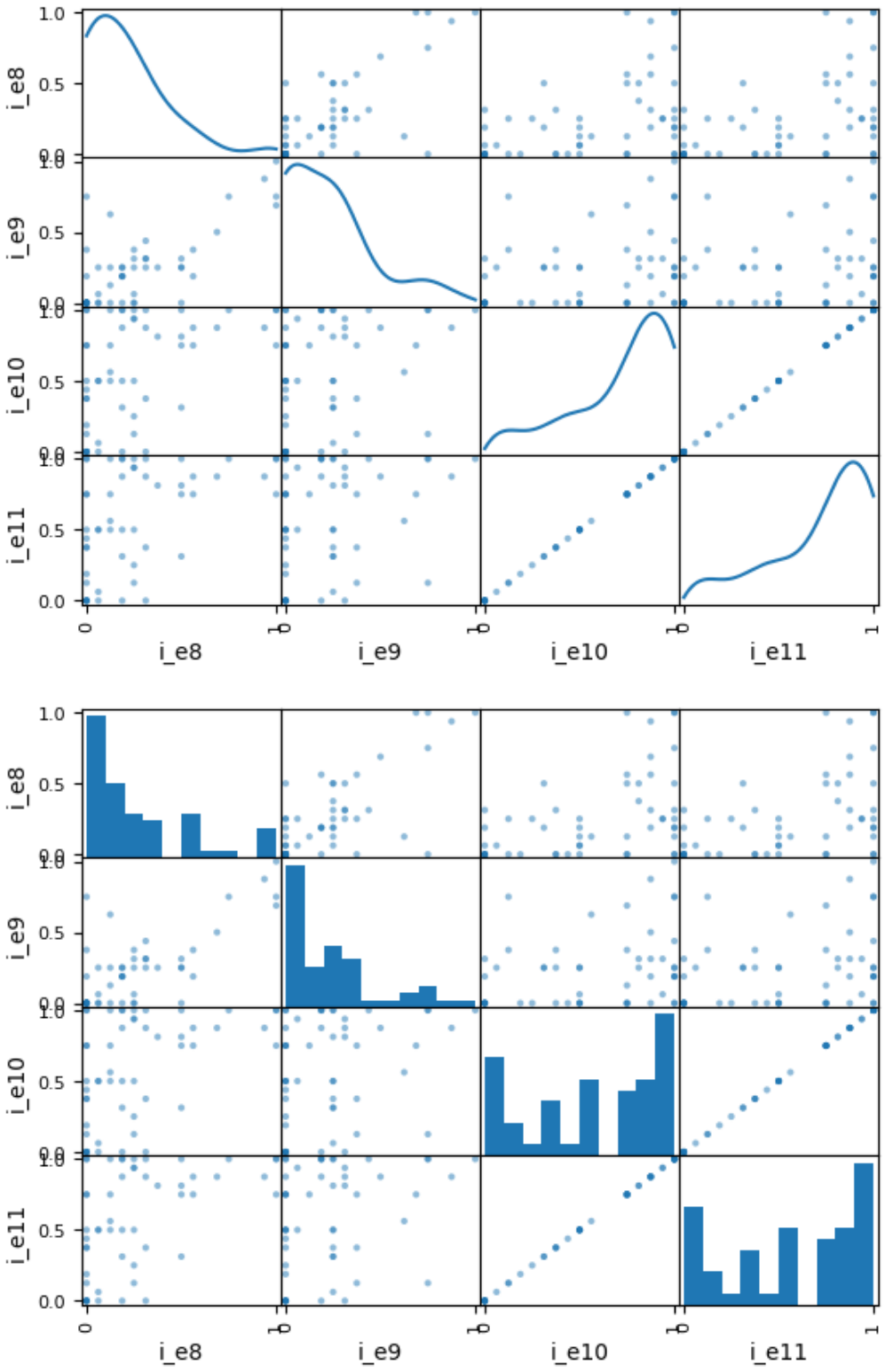

Figure 5: up - kde, bottom - hist

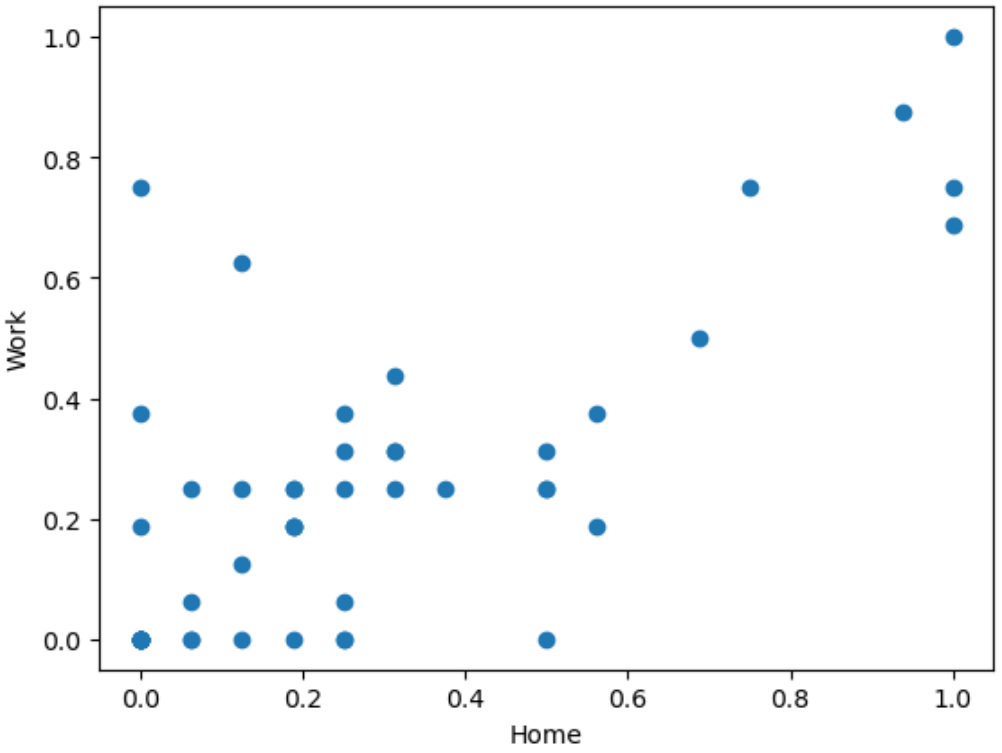

Figure 6: Showing that there is a very little correlation between the water habits in the house and at home

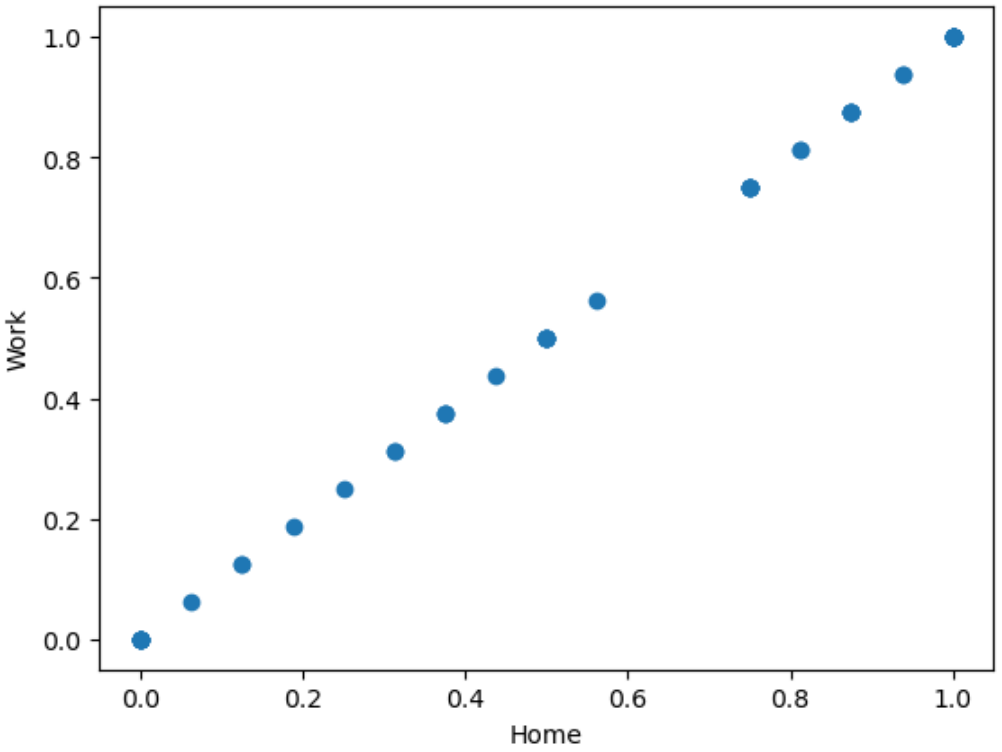

Figure 7: Shows that there is a really big correlation between the waste habits at home and work

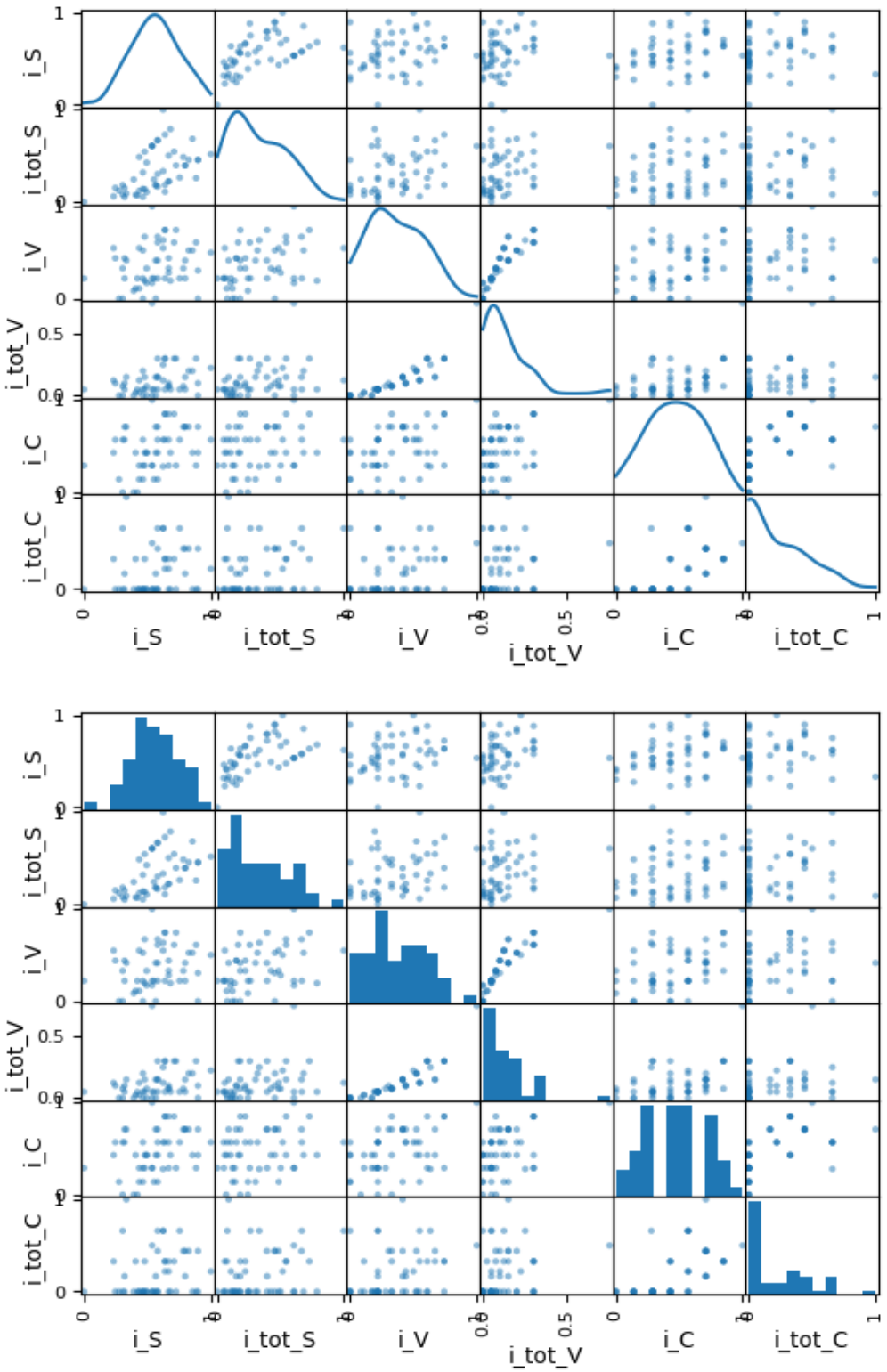

Figure 8: up - kde, bottom - hist

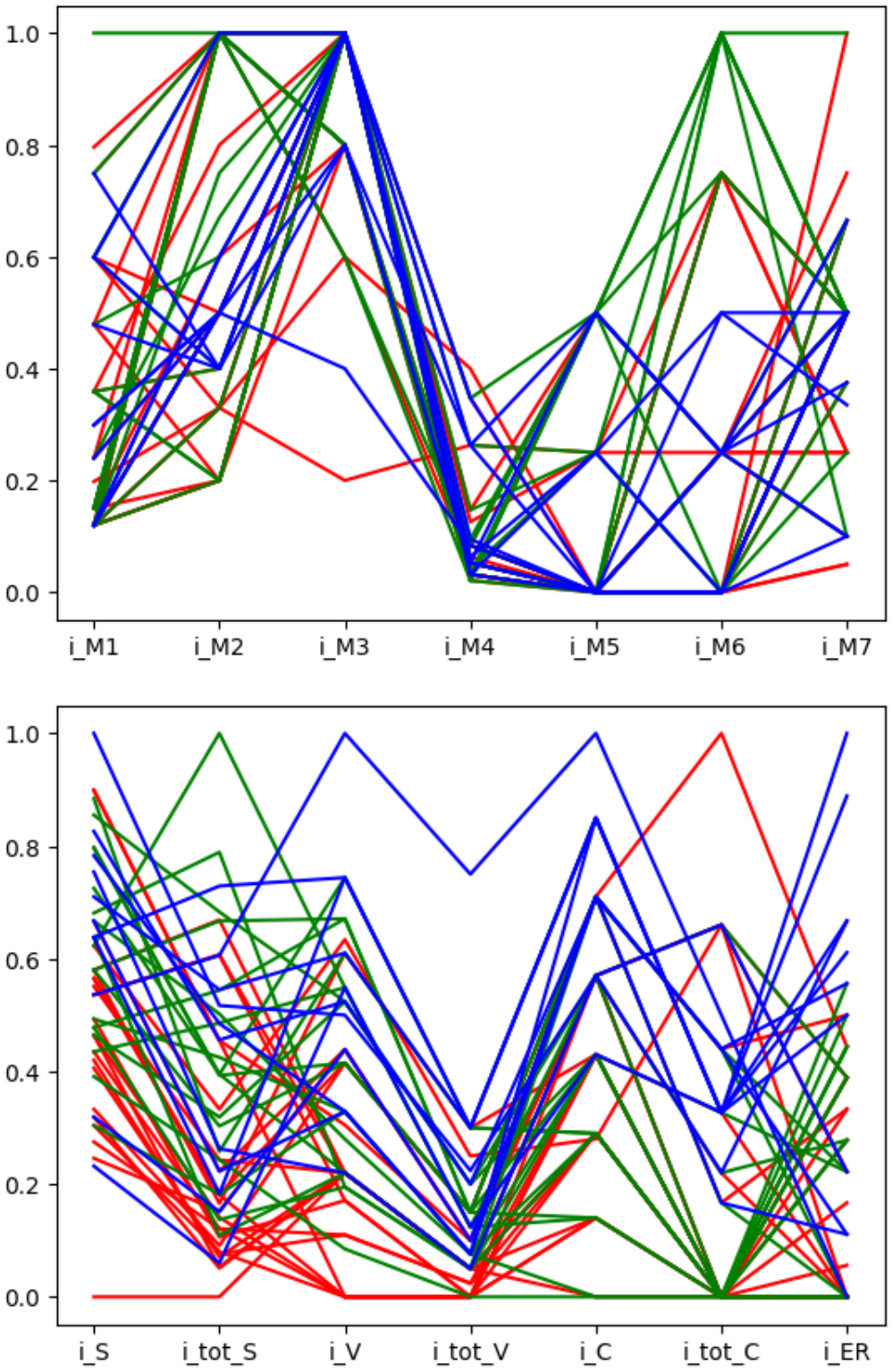
1.0
0.8
0.6
0.4
0.2
0.0
i_M1
i_M2
i_M3
i_M4
i_M5
i_M6
i_M7
1.0
0.8
0.6
0.4
0.2
0.0
i_S
i_tot_S
i_V
i_tot_V
i_C
i_tot_C
i_ER

Figure 9-10

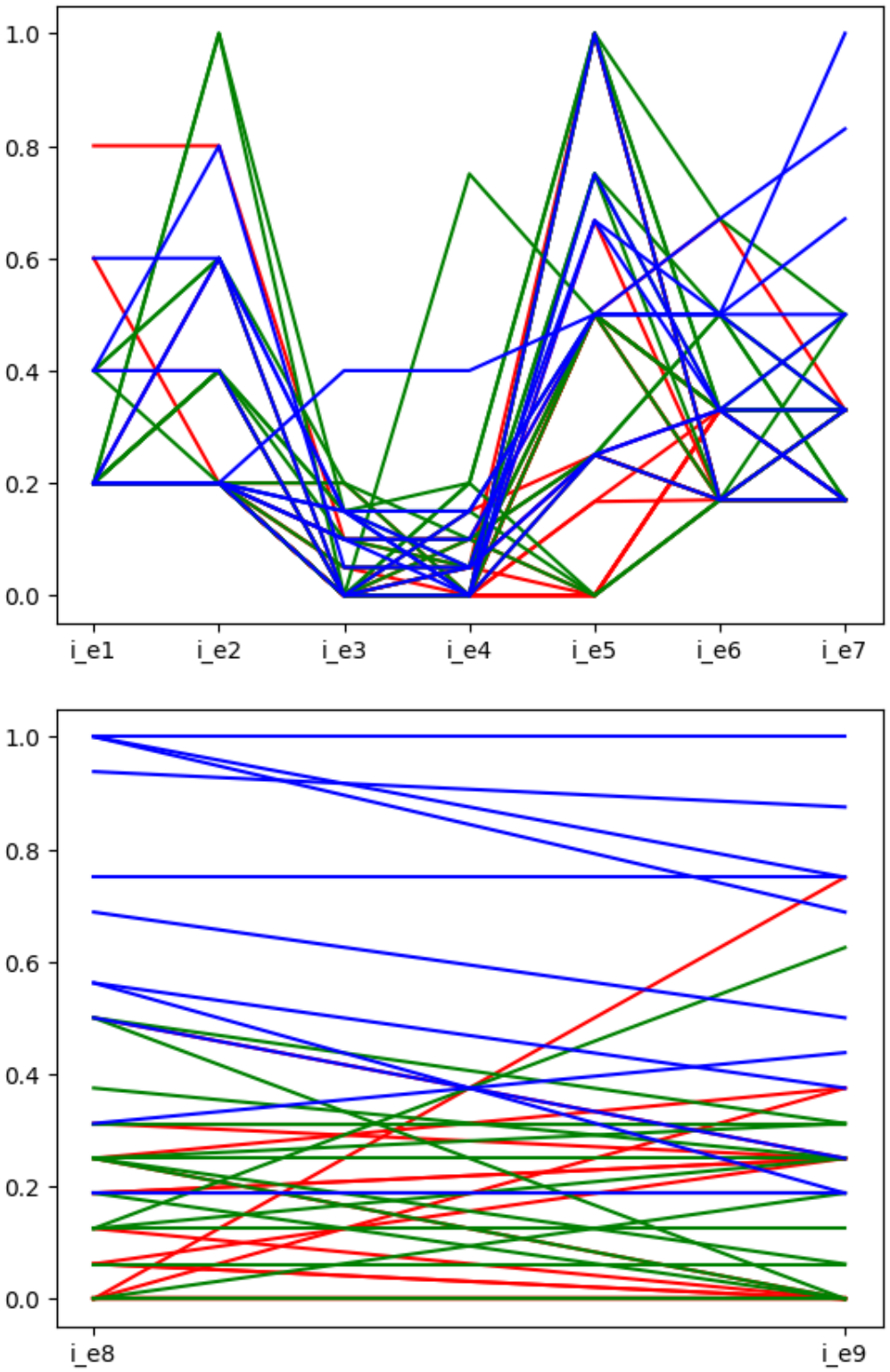
1.0
0.8
0.6
0.4
0.2
0.0
i_e1
i_e2
i_e3
i_e4
i_e5
i_e6
i_e7
1.0
0.8
0.6
0.4
0.2
0.0
i_e8
i_e9

Figure 11-12

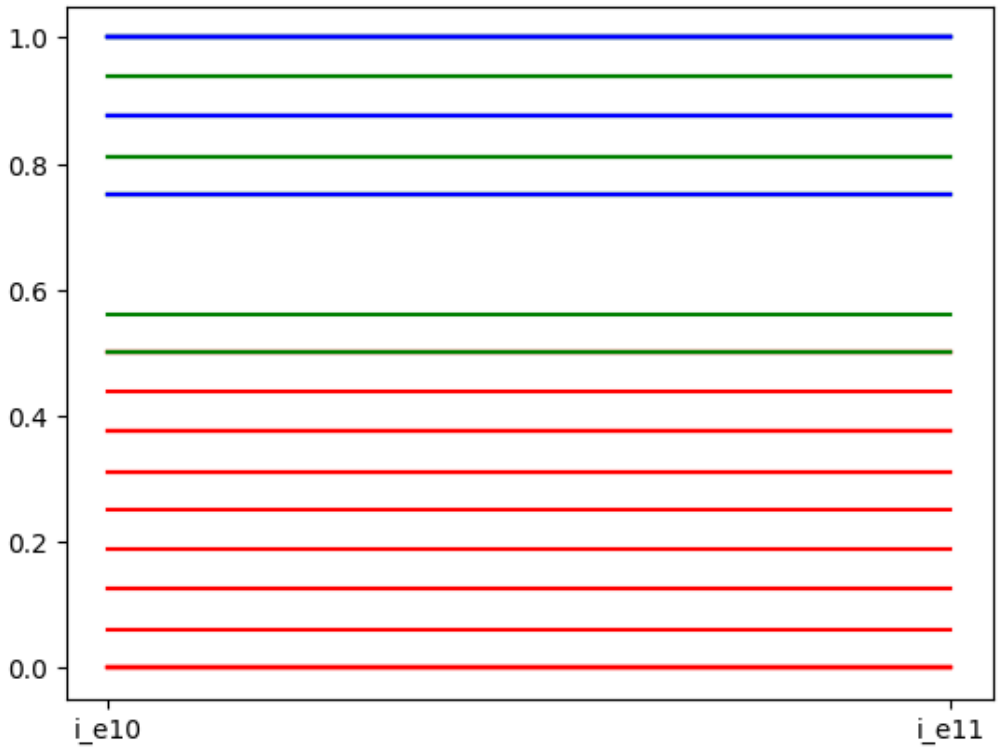

Figure 13

Figures 9 to 13 are our attempts to visualize 25 dimensions in 2 dimensions.

CONCLUSIONS

The purpose of the analysis was to examine a user's behavior from a sustainability perspective. It is possible to deduce how most people own ecological knowledge and behavior in different fields of daylife, in particular during purchase phases, waste and water consumption, and energy consumption.

We found no particular correlation between mobility behaviors, thus showing how to introduce special attention to what sustainable mobility could be built on a day-to-day basis.

APPENDIX 1

Question Text	***Question Code***
Quante volte alla settimana fai la spesa?	S_1
Quanti supermercati visiti quando fai la spesa?	S_2
Per una vita più sostenibile, una buona abitudine per gli acquisti è quello di servirsi di prodotti locali e/o di propria produzione. In quale delle seguenti condizioni ti rispecchi maggiormente?	S_3
Per incentivare la spesa sostenibile, sono stati creati dei servizi online per l'acquisto e spedizione a casa di tutti i prodotti biologici, siti come BioDiscount, Greenweez, La Boutique del Biologico. Sfrutti tali servizi?	S_4
All'interno dei supermercati, presti attenzione ai prodotti con etichetta ecologica (basso consumo CO2, allevamento a terra)?	S_5
Sacchetti della spesa	S_6
Prodotti per la casa e per indumenti	S_7
Acqua potabile	S_8
Molte associazioni coinvolgono i supermercati per la raccolta di cibo all'uscita dei supermercati. Quanto spesso dai in beneficenza cibaria?	S_9
In media, quante volte al mese acquisti vestiti?	V_10
Etichette sostenibili sono presenti anche per il vestiario. Quale delle seguenti opzioni ti rispecchia maggiormente?	V_11
La presenza di mercatini dell'usato (anche online) è una buona pratica, come anche il riutilizzo di indumenti vecchi o danneggiati, per il riciclo di indumenti. Quale delle seguenti opzioni ti rispecchia maggiormente?	V_12
Le case green sono un altro tema discusso in ambito sostenibilità. Seleziona quale delle seguenti condizioni ti rispecchia maggiormente:	C_13
Un aspetto importante che caratterizza una casa green è la presenza di dispositivi elettronici sostenibili (illuminazione a LED, pannelli solari) e di forniture ecologiche (etichette ecologiche). Quanto sei interessato all'acquisto di tali dispositivi e forniture? (Selezionare valori intermedi in base anche alla possibilità di acquisto, ad esempio, sono molto interessato all'acquisto ma sono estremamente costosi = 3)	C_14
Nella tua giornata tipo, a che ora esci di casa?	m_0
Qual è il veicolo che usi quotidianamente?	m_1
Cosa alimenta il tuo veicolo?	m_2
In quale classe ambientale si trova il veicolo che usi quotidianamente? (basati sull'anno di immatricolazione se non conosci la classe)	m_3

Quante volte a settimana esci di casa?	m_4
Di queste volte, quanto spesso usi il mezzo di trasporto che hai specificato poco sopra?	m_5
Quanto traffico incontri quando esci di casa?	m_6
Quanti minuti impieghi solitamente per trovare parcheggio per il tuo veicolo?	m_7
Indicativamente, quanto è grande in m² la tua residenza?	m_8
Indicativamente, quanti m² di spazi verdi sono presenti nella tua residenza?	m_9
Indicativamente, che percentuale della tua residenza è riservata al parcheggio?	m_10
Quanto spendi in carburante ogni mese?	m_11
Quanto spendi in trasporti ecosostenibili ogni mese? (per esempio, caricando l'auto elettrica)	m_12
Quanto spendi mensilmente in trasporti pubblici?	m_13
Seleziona tutti i mezzi di trasporto pubblici di cui fai uso tipicamente	m_14
Quanto è distante, in chilometri, il tuo posto di lavoro/studio?	m_15
Quanto spesso raggiungi il posto di lavoro/studio a piedi?	m_16
Indicativamente, quanto è grande il tuo posto di lavoro/studio? (in m²)	m_17
Indicativamente, quanti m² di spazi verdi sono presenti nel tuo posto di lavoro/studio?	m_18
Il tragitto tra casa tua e il tuo posto di lavoro/studio:	m_19
Quante iniziative esistono da parte del tuo posto di lavoro/studio alla riduzione dell'uso dei mezzi di trasporto? (ad esempio, offrendo soluzioni alternative come navette o car pooling)	m_20
Qual è la percentuale di prodotti elettronici efficienti che possiedi?	eh1
Seleziona se possiedi uno dei seguenti sistemi?	eh2
Seleziona quali sistemi di fonti rinnovabili possiedi?	eh3
Quanto, in percentuale, pensi che gli impianti rinnovabili che si possiedono producano rispetto a ciò che consumi?	eh4
Qual è il kilowattaggio del tuo impianto elettrico?	eh5
La tua casa rispecchia i seguenti elementi di Green Building Implementation?	eh6

Qual è la percentuale di prodotti elettronici efficienti che il tuo luogo di lavoro/ la tua università possiede?	ew1
Seleziona se il tuo luogo di lavoro/la tua università possiede uno dei seguenti sistemi?	ew2
Seleziona quali sistemi di fonti rinnovabili possiede il tuo luogo di lavoro/la tua università?	ew3
Quanto, in percentuale, pensi che gli impianti rinnovabili del tuo luogo di lavoro/della tua università producano rispetto a ciò che viene consumato?	ew4
Il tuo luogo di lavoro/la tua università rispecchia i seguenti elementi di Green Building Implementation?	ew5
Consumo dell'acqua - Lavoro o Università [Programma di conservazione dell'acqua]	ww1
Consumo dell'acqua - Lavoro o Università [Programma di riciclo dell'acqua]	ww2
Consumo dell'acqua - Lavoro o Università [Utilizzo di strumenti e elettrodomestici efficienti (asciugamani ad aria, scarico a doppio bottone, ecc)]	ww3
Consumo dell'acqua - Lavoro o Università [Consumo di acqua trattata comparata con tutte le altre fonti d'acqua]	ww4
Consumo dell'acqua - Casa [Programma di conservazione dell'acqua]	wh1
Consumo dell'acqua - Casa [Programma di riciclo dell'acqua]	wh2
Consumo dell'acqua - Casa [Utilizzo di strumenti e elettrodomestici efficienti (asciugamani ad aria, scarico a doppio bottone, ecc)]	wh3
Consumo dell'acqua - Casa [Consumo di acqua trattata comparata con tutte le altre fonti d'acqua]	wh4
Rifiuti - Lavoro o Università [Vengono implementate le 3R: Riduzione, Riuso e Riciclo]	wastew1
Rifiuti - Lavoro o Università [C'è un'attenzione nel ridurre l'uso della plastica]	wastew2
Rifiuti - Lavoro o Università [C'è un'attenzione alla gestione di materiali organici]	wastew3
Rifiuti - Lavoro o Università [C'è un'attenzione alla gestione di materiali inorganici]	wastew4
Rifiuti - Casa [Vengono implementate le 3R: Riduzione, Riuso e Riciclo]	wasteh1
Rifiuti - Casa [C'è un'attenzione nel ridurre l'uso della plasica]	wasteh2
Rifiuti - Casa [C'è un'attenzione alla gestione di materiali organici]	wasteh3

Rifiuti - Casa [C'è un'attenzione alla gestione di materiali inorganici]	wasteh4
Inserire il numero di eventi (conferenze, workshop, sensibilizzazione, formazione pratica, festival, etc.) legati ai problemi di sostenibilità e dell'ambiente organizzati dalla propria azienda e/o università (media per anno degli ultimi 3 anni). Selezionare una delle seguenti opzioni:	ER_1
Inserire il numero totale di attività organizzate dalle organizzazioni studentesche o da eventuali organizzazioni interne al proprio ambiente di lavoro per anno. Ad esempio, seminari, webinars, formazione, eventi sportivi, bazaar di materiali riciclati, comunità di divulgazione, etc.. Selezionare una delle seguenti opzioni:	ER_2
La presenza di strutture green all'interno del proprio campus e/o della propria azienda che siano accessibili al pubblico, per esempio durante attività culturali, indicano un impatto più ampio di tali strutture nei loro dintorni (Festival Culturali, teatro, performance musicali, esibizioni, attività virtuali, etc.). Selezionare una delle seguenti opzioni:	ER_3