

D6.1: A demo of the tool, demonstrating the tool functions based on data samples.

WP6 – Design of web-based decision tool

Report No. D6.1 / Date: 28/06/2023



AUTHOR(S) NAME

Kalyan Ram Ayyalasomayajula

Smart Innovation, Norway



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957115.



ENCHANT Report

D6.1: A demo of the tool, demonstrating the tool functions based on data samples

VERSION: 01 / DATE: 28.06.2023

AUTHOR(S)

Kalyan Ram Ayyalasomayajula (SIN)

Quality assurance: Christian A. Klöckner

PROJECT NO.: 957115 (H2020) / PAGES/APPENDICES: 16/0

ABSTRACT

The main objective of the demo is to provide an overview of the web-tool development as part of WP6 and some additional analysis features under development. It also provides an overview of the workflow of the tasks and software life-cycle in the development of the code and machine learning algorithms.

REPORT NO.: D6.1

ISBN: NA

CLASSIFICATION: Public

CLASSIFICATION THIS PAGE: Public



DOCUMENT HISTORY:

VERSION	DATE	VERSION DESCRIPTION
1	28.06.2023	First version for quality check sent to partners
2	30.06.2023	Final version sent to the European Commission



Table of contents

1.	Introduction and Overview	5
1.1	Purpose and scope of this document.....	5
1.2	Summary of the tool capabilities and development cycles	5
2.	Training recommender system.....	7
2.1.	Data parsing and cleaning for training.....	7
2.2.	Training a recommendation system.....	8
3.	Tool demo.....	9
3.1	Online tool capabilities.....	9
3.2	Offline capabilities	11
3.3	Intervention group behaviour analysis	12
3.4	Identifying the number of groups per intervention.....	14
3.5	Future tasks.....	16



1. Introduction and Overview

1.1 Purpose and scope of this document

WP6 deals with designing a machine learning (ML) algorithm that can be used for answering a couple of socio-behavioural questions related to green energy transition within the society. As these method starts from existing trends within the data, it is interesting to explore its relevance when applied to survey data collected under selected ENCHANT interventions. Based on the data from various intervention surveys conducted in the project and contacting people through various communication channels the objective of WP6 is to:

- O1: Determine the best intervention strategy for a given group of people characterized by a combination of sociodemographic and psychological features.
- O2: Identify the most suitable channel for reaching out to these end-users.

In this report, we summarize the progress in WP6 to meet the goals mentioned above. The report provides an overview of:

- A summary of the progress in tool development in WP6.
- A description of the recommender system's role and the data used in training.
- The usage of web-tool and some of its features still under development.
- The process of identifying the group structure (homogenous groups) within the data.

This document is organized as follows. In Chapter 1 – Section 1.2, a summary of the progress made in tool development and its contributions towards O1 and O2 will be given.

Chapter 2 breaks down the idea behind the recommendation system (RS), its role in ENCHANT and challenges encountered when training such a system on ENCHANT intervention data.

Chapter 3 – Section 3.1 provides a summary of the demo of the deployed web-tool that is up and running. Section 3.2 provides an overview of the exploratory capabilities of the tool under development and not yet deployed online.

Chapter 4 presents guidelines for identifying the underlying target groups in the population based on the survey data.

1.2 Summary of the tool capabilities and development cycles

As the current report describes the inner working of the tool usage that is currently under development, we summarize its capabilities here:



Table 1 Summary of the tool capabilities

Deployment cycles	Capabilities	Utility
Online	RS trained with the ability to predict from incomplete data	Proof of concept of RS and its ability to learn tangible numerical representation to data for downstream tasks like intervention/channel predictions
Offline	In addition to incomplete data prediction this RS can also determine group patterns paving way to realise O1 and O2	Determine exact number of possible behavioural groups per intervention

The software development cycles followed along two paths:

- Offline code repository: This is the research branch of the ML algorithms developed in the project. Data collected is parsed and cleaned followed by some sanity checks. The data is then used to train the ML models to run predictions and inferences. As this setup is experimental once the features are more stable and are working as intended, they are transferred to the online web tool.
- Online web tool: Here the trained ML models are deployed for open access for the user. The features exposed in this tool are in line with the project outcomes from ENCHANT. However, some of the more experimental features developed might not be available. These functions will be made open once their usage is finalized through experiments run on the offline repository.



2. Training recommender system

2.1. Data parsing and cleaning for training

The data from an intervention survey can be seen as matrix with each user response as a row and each question along as a column. We denote all users as $\{u_1, u_2, \dots, u_n\}$ and all questions as items $\{i_1, i_2, \dots, i_m\}$ this can be seen as a user-item interaction matrix M of size $n \times m$ denoted as $M^{n \times m}$. An example of a typical response to a survey can be visualizes as shown below with strong responses to a question (>3) in red, weak response (<3) in green, responses with value 3 in blue and empty responses as well.

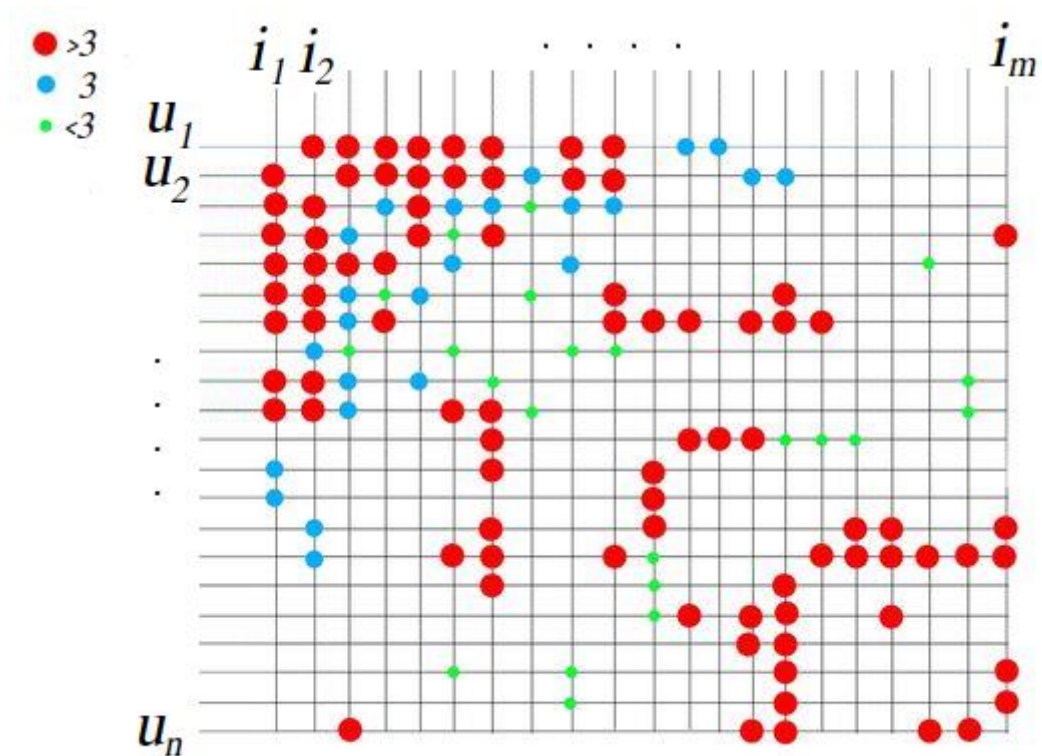


Figure 1: Visualizing user-item interaction matrix with color coded values

Typical challenges encountered with raw intervention data are:

- Handling null responses: Users in the survey have opted not to respond leaving the question unanswered or with an invalid option (out of 1-5 range)
- Pruning data: Data such as demographic, energy consumption etc. though important for answering ENCHANT objectives. Including them is beyond the scope of behavioral analysis of users. Hence, these responses were pruned in current analysis. However, these features will be employed into further analysis when answering O1 and O2.

- Sanity checks: Response rates were also gauged to eliminate entries with very few responses to the questions (typically the threshold being < 15%)

The dataset from ENCHANT interventions after the sanity checks was used to train a ML recommendation algorithm.

2.2. Training a recommendation system

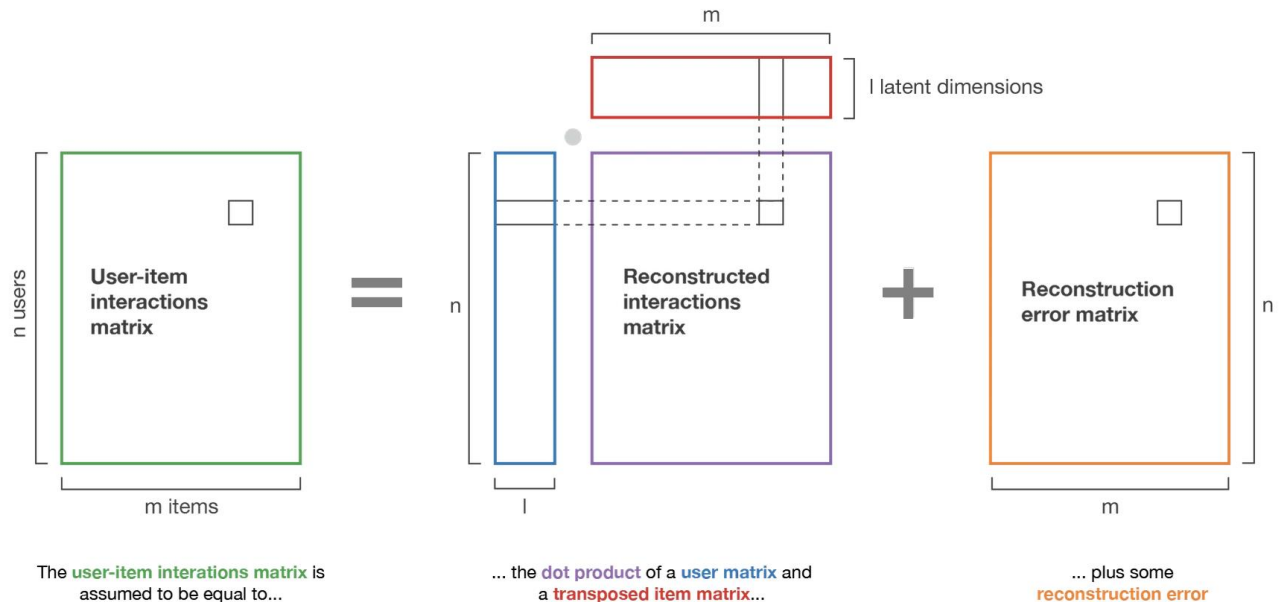


Figure 2: Matrix factorization algorithm

The algorithm used in training the RS is a *matrix factorization* algorithm that can be understood visually as shown in the figure above. The algorithm tries to decompose the user-item interaction matrix $M^{n \times m}$ into the product of the user latent matrix $U^{n \times l}$ and the item latent matrix $I^{m \times l}$ such that as per matrix multiplication we have:

$$M^{n \times m} = U^{n \times l} \cdot (I^{m \times l})^T$$

The algorithm tries to find the values for $U^{n \times l}$ and $I^{m \times l}$ such that the reconstruction error is minimized.

3. Tool demo

3.1 Online tool capabilities

This section provides a walkthrough of the web-tool deployed online.

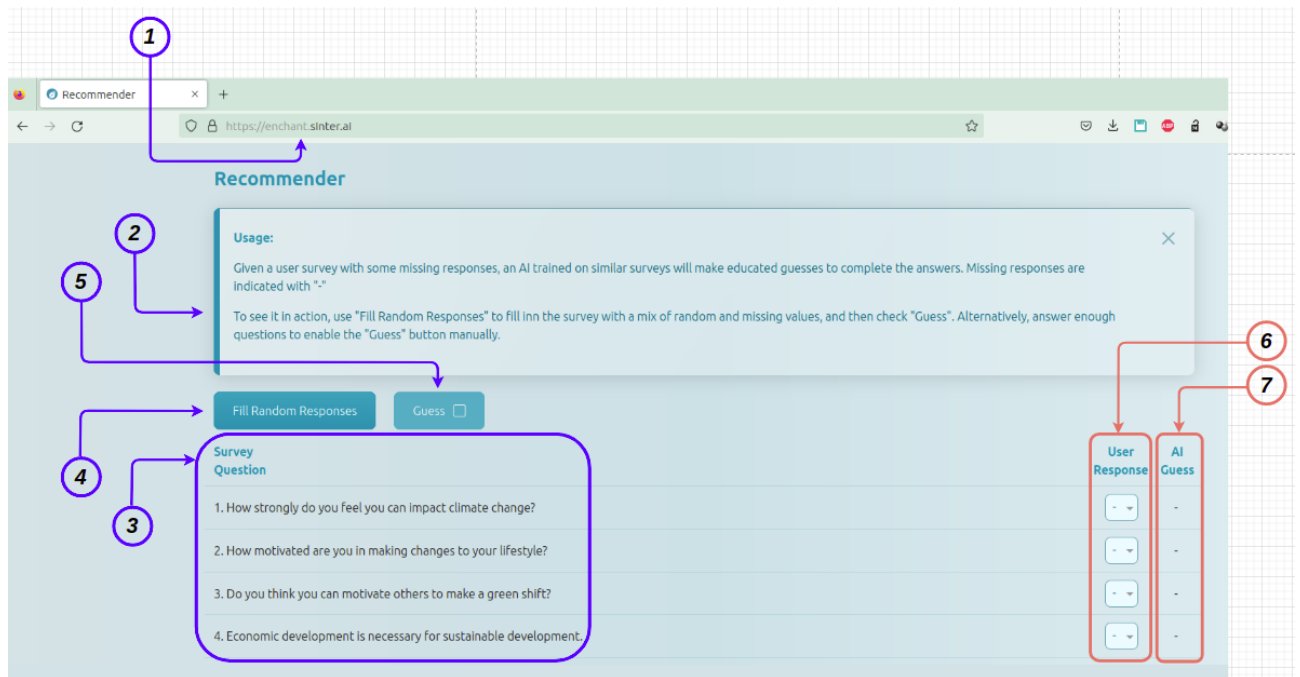


Figure 3: Web graphical user interface

The recommendation tool deployed online is as shown in the Fig. 1, with the following components:

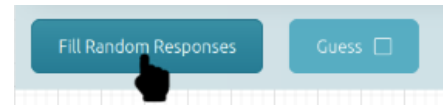
1. The URL to the tool is: <https://enchant.sinter.ai>
2. Help in using the tool is mentioned to guide new users.
3. The survey questions for which the user response is expected is as indicated.
4. Random response fill-in feature is given. This feature is only for the demo and will not be disabled in the final version.
5. The AI guess option is given to provide the recommendation for missing options as a demonstration of the capability of the tool to infer likely responses for missing information.
6. User responses for each question can be picked from the drop down from options 1-5. 1- to indicate low preference, 5- to indicate a strong preference and anything in between for a question.
7. The recommendation or guess from the recommendation system will be shown here.

Usage of the tool is as follows:

1. Open the indicated URL in 1 in a browser.
2. Read the text option in 2 to make oneself familiar with the tool.



- Fill in the response from the drop down individually. However, for the demo we proceed by clicking the “Fill Random Responses” option indicated by in the figure below.



- Responses will be updated as shown.

Survey Question	User Response
1. How strongly do you feel you can impact climate change?	1
2. How motivated are you in making changes to your lifestyle?	-
3. Do you think you can motivate others to make a green shift?	-
4. Economic development is necessary for sustainable development.	3
5. Improving people's health and opportunities contribute to sustainable development.	-
6. Reducing water consumption is necessary for sustainable development.	4
7. Preserving nature is not necessary for sustainable development.	5

Note: Some responses are not filled (as indicated by -) to replicate the real use case where the user can opt not to respond to an option.

- Clicking on the Guess option will bring up the updated output

Survey Question	User Response	AI Guess
1. How strongly do you feel you can impact climate change?	1	🔄
2. How motivated are you in making changes to your lifestyle?	-	🔄
3. Do you think you can motivate others to make a green shift?	-	🔄
4. Economic development is necessary for sustainable development.	3	🔄
5. Improving people's health and opportunities contribute to sustainable development.	-	🔄

- The values as guessed/recommended will be updated. In this step the ML algorithm can make an informed decision to fill in values that were missed out by the users previously.

Survey Question	User Response	AI Guess
1. How strongly do you feel you can impact climate change?	1	1
2. How motivated are you in making changes to your lifestyle?	-	4
3. Do you think you can motivate others to make a green shift?	-	3
4. Economic development is necessary for sustainable development.	3	3
5. Improving people's health and opportunities contribute to sustainable development.	-	3



This ability to learn from incomplete data is the basis for making recommendations to the users or user groups in general and paves way to achieve objectives O1 and O2 of WP6.

In the final version of the tool, the input will be more focused on structural variables like type of communication provider (e.g., municipality, electricity provider, NGO) or socio-demographics of a selected target group for a planned intervention. This information will then be used to infer the target group's likely psychological profile using the same technique as outlined above and then match it with the intervention strategy and corresponding communication channel most likely leading to a successful intervention.

Note: When moving from identified groups structure to the target groups we might encounter groups with very small sample size and varied psychological profiles. At this point, the generalization of the inference could be hampered. This is not a limitation on tool functionality but rather a bottle neck in sample data availability.

3.2 Offline capabilities

As the previous section has set precedence to the capabilities of RS in dealing with incomplete data collected through ENCHANT interventions. We will further discuss the usage of RS in developing and analyzing group behavior identified through ENCHANT interventions. Here is a visual representation of the group structure identified for the Norwegian (NO) and German (DE) data collected through the online interventions. These features are still in an experimental phase and will be deployed online when they are mature.

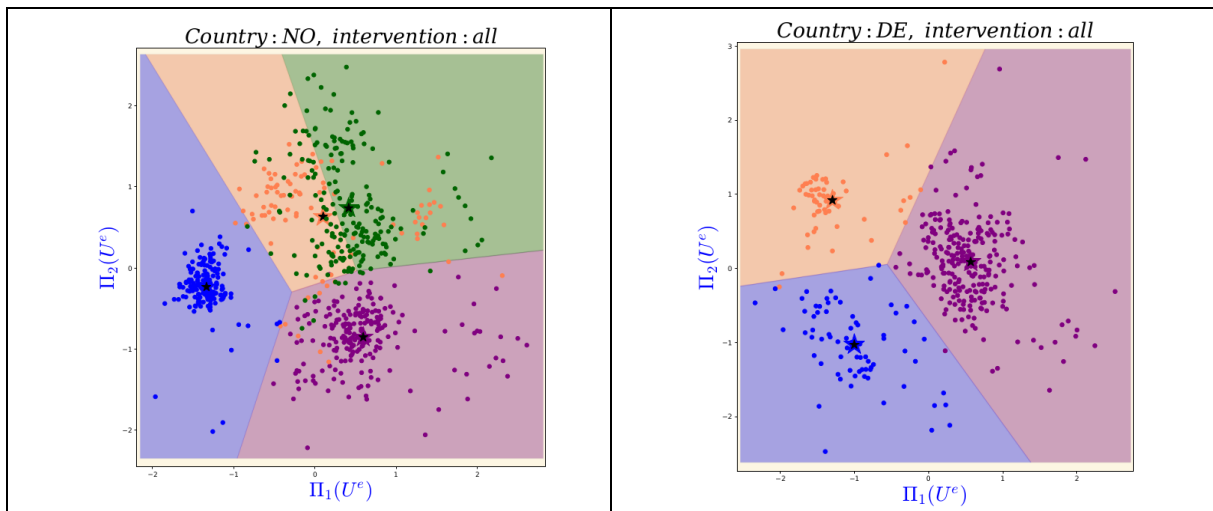


Figure 4: Comparing the behavioral groups across countries using PCA.

Here is an overview of the steps followed by the algorithm:

1. There were 4 and 3 behavioral groups identified using all the interventions data collected from Norwegian and German surveys (using *k-Mean clustering*).

2. The cluster centroids of the group clusters are marked with a star of the corresponding group color with a black star to mark its center.
3. All the data points corresponding to a group are marked with the respective group color.
4. With the group centroids as the center each group region is filled out with a faded color corresponding to the region (using *Voronoi tessellations*)
5. Marking the regions for the group will help identify the behavior changes for each group over the interventions.
6. The latent representations (*user embeddings* in some high dimensional vector space U^e) of each learned by the RS are then linearly projected (using principal component analysis [PCA]) to two dimensions ($[\Pi_1(U^e), \Pi_2(U^e)]$) and plotted along each axis as shown in the plots above.
7. A non-linear projection (using t-Distributed Stochastic Neighbor Embedding [t-SNE]) can also be used to visualize the group behavior as shown below:

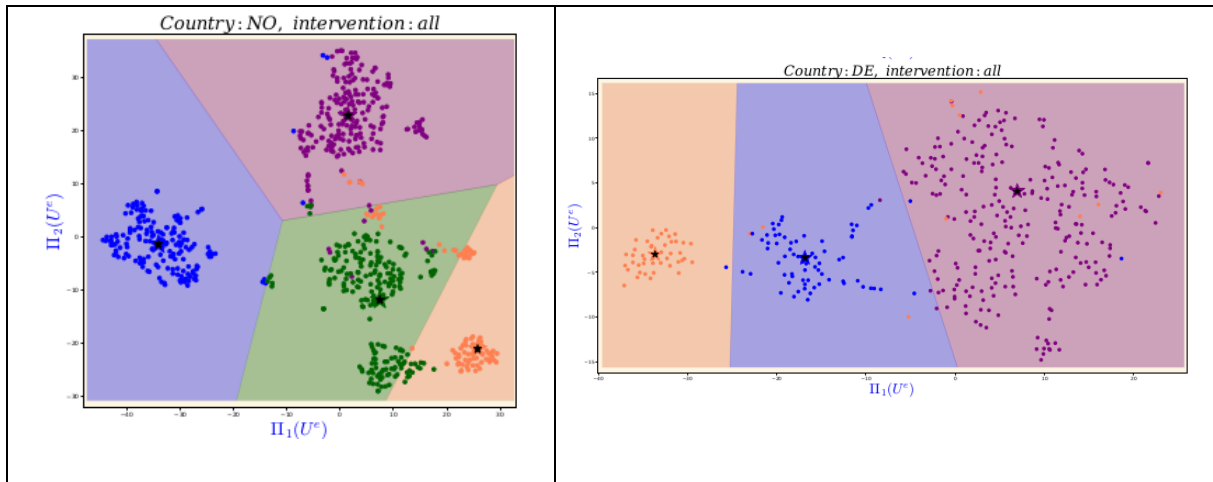


Figure 5: Comparing the behavioral groups across countries using t-SNE

Note: Although the t-SNE plots look visually more pleasing compared to PCA plots interpreting the projections used to map $U^e \rightarrow [\Pi_1, \Pi_2]$, by t-SNE can be more complicated as compared to PCA. Hence, we opt to use both methods in visualizing the results.

3.3 Intervention group behaviour analysis

To compare the changes in the group behavior over each intervention we retain the Voronoi tessellation pattern using all the interventions data in the background and plot the group users identified in each intervention in the foreground as shown below.



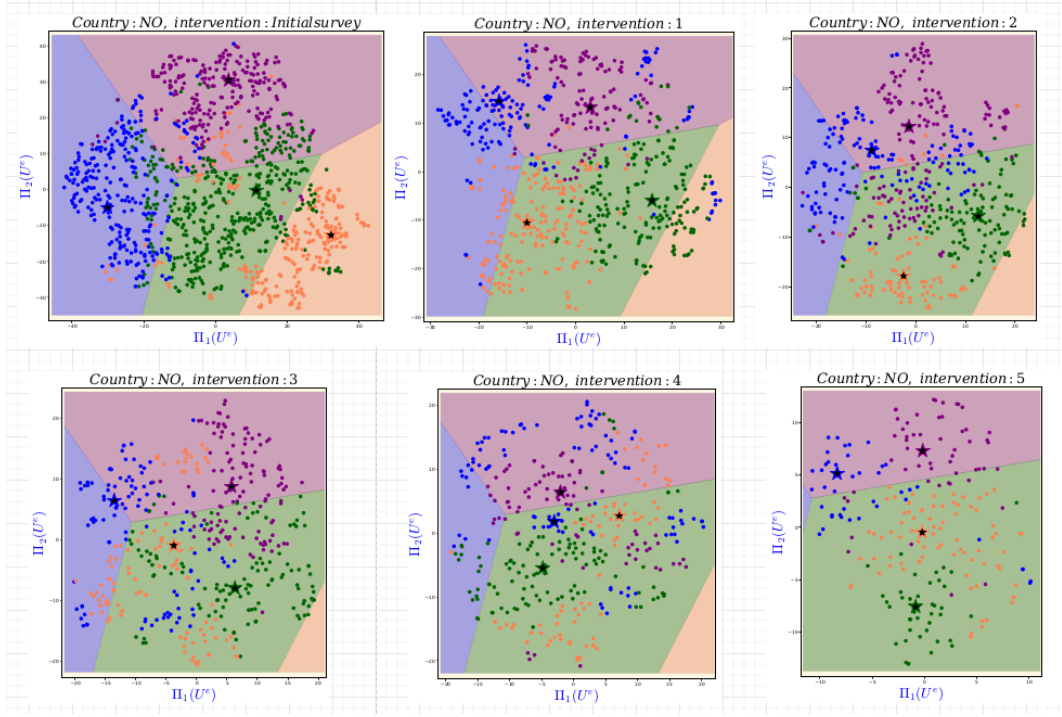


Figure 6: Tesselation for NO interventions with t-SNE embeddings

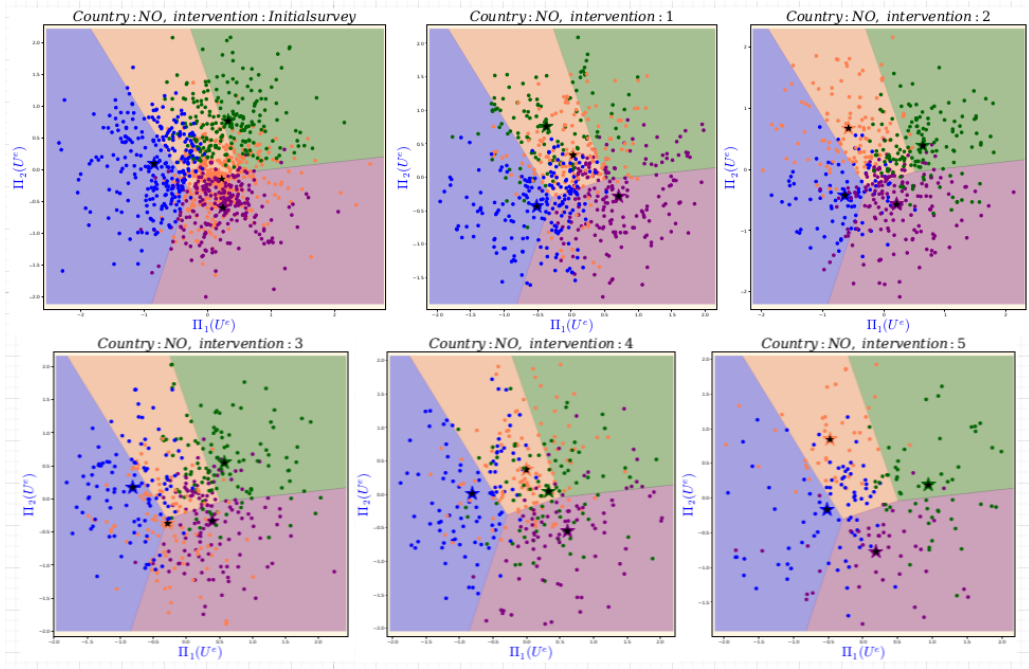


Figure 7: Tesselation for NO interventions with PCA embeddings



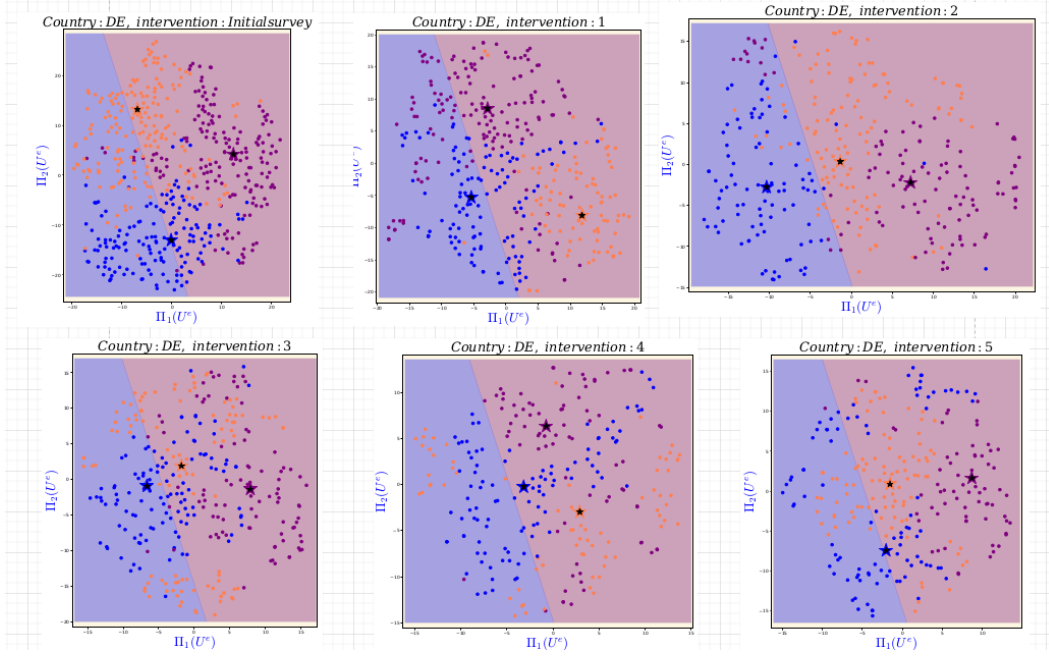


Figure 8: Tesselation for DE interventions with t-SNE embeddings

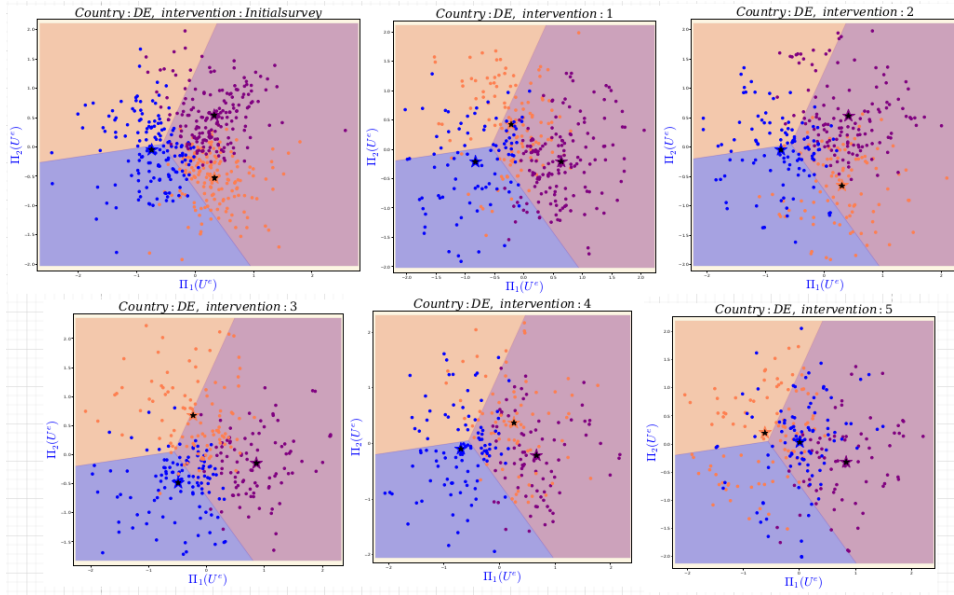


Figure 9: Tesselation for DE interventions with PCA embeddings

3.4 Identifying the number of groups per intervention

We start with the learned embedding for each user from the interventions U^e . We run the k-Means algorithm on the embeddings space to identify the clustering structure. As



we do not know the number of clusters or groups in the data to begin with, we employ two methods to determine the total number of groups:

Elbow method: We iterative run a k-Means algorithm to identify number of groups from an enumerated list starting from $[2, 3, \dots n]$, then create a plot with the number of clusters on the x-axis and the total within sum of squares error among clusters on the y-axis and then identifying where an “elbow” or bend appears in the plot as shown using all the intervention data. This approach useful in identifying the possible number of clusters.

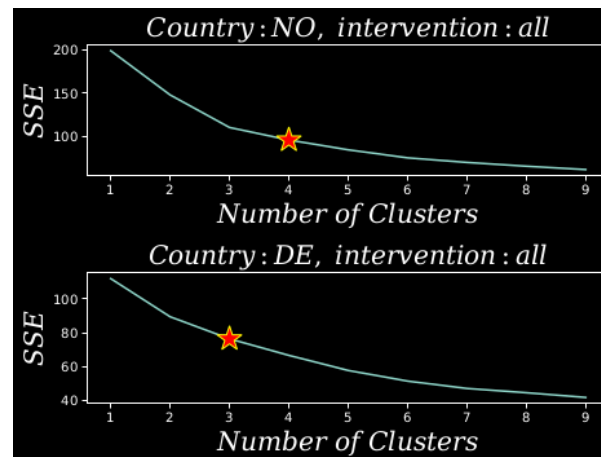


Figure 4: Elbow plots for NO and DE data

Silhouette score: In the range identified we use the silhouette coefficient or silhouette score of k-means as a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation). The Silhouette score are run for the identified range in this case $[2, 3, \dots 10]$. The number of clusters with highest coefficient is selected as the relevant number groups identified in the data.

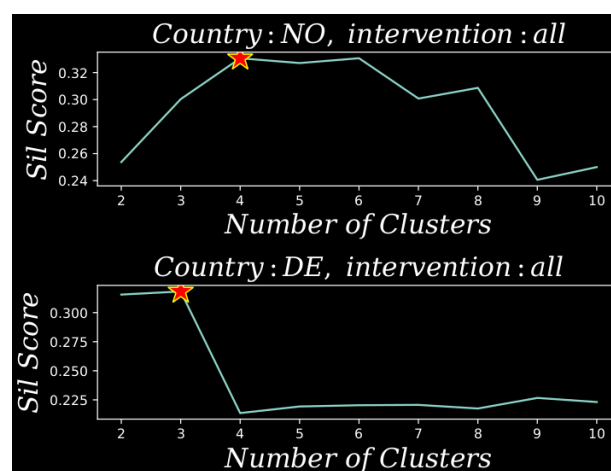


Figure 5: Silhouette score plots for NO and DE



From this we conclude that there are 4 and 3 behavioural groups in the Norwegian and German intervention data respectively.

3.5 Future tasks

The following steps are planned to enhance the recommendation tool in the next deliverables:

1. Work towards ENCHANT objectives: The learned embeddings can discern a pattern from the data as shown in the previous sections. As they are numerical representations of the users, a mapping can be learned from the user embeddings to measurable values like energy consumption values collected in the interventions. This regression approach can answer question related to best suitable interventions and channels per user group. In the other direction, the numerical representations of the users will be linked to easily identifiable profiles of structural and socio-demographic data to serve as the entry to the final tool. This converse direction is susceptible to data generalizability and needs investigation.
2. Updated features for web-tool: The analysis plots from previous section and the regression algorithm discussed in the previous step will be added to the online web-tool so that end-user may interact with it.

