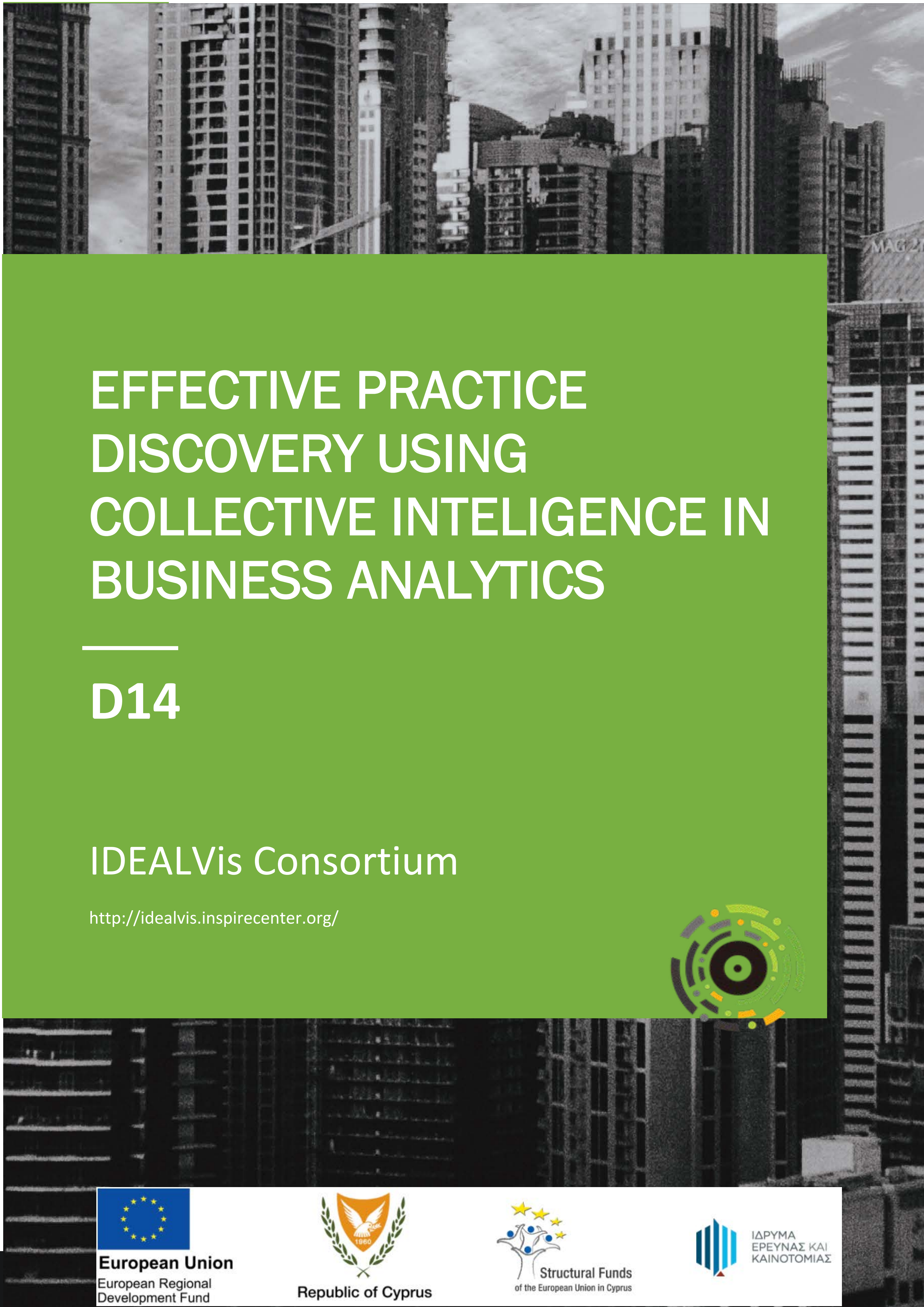

EFFECTIVE PRACTICE DISCOVERY USING COLLECTIVE INTELIGENCE IN BUSINESS ANALYTICS

D14

IDEALVis Consortium

http://idealvis.inspirecenter.org/

European Union
European Regional
Development Fund

Republic of Cyprus

Structural Funds
of the European Union in Cyprus

ΙΔΡΥΜΑ
ΕΡΕΥΝΑΣ ΚΑΙ
ΚΑΙΝΟΤΟΜΙΑΣ

Executive Summary

The overall objective of WP5 is to design and develop an intelligent data analytics component that enhances the efficiency and effectiveness of the data exploration process by quantifying the user's experience through recorded interactions, and by further identifying effective analysis practices through the utilization of machine learning techniques.

This deliverable presents the machine learning mechanisms utilized as part of Work Package 5 (WP5) required for classifying users in distinct groups based on their user model characteristics. Assigning users in distinct groups, facilitates the analysis of recorded interaction data, allowing the possibility to discover analysis patters adopted by different types (i.e., groups) of users. In deliverable D13 (i) all IDEALVis tracking mechanisms were described; and (ii) the set of interaction records captured during pilot study were analyzed presenting the different interaction patters that emerged from all users as a whole.

In contrast to deliverable D13, this deliverable analyses the captured interaction data for each distinct user group that resulted from the user classification process. The main goal of this process is to detect interesting analysis interaction patterns across different groups/types of users. Additionally, this deliverable uses findings from deliverable D11, attempting to provide a set of effective/best adaptation practices that are a best fit for each of the different user groups.

Table of Contents

List of Figures

1 User Classification and Pattern Discovery

During the pilot study operation, the platform's tracking mechanisms were continuously recording the users' descriptive parameters and interaction (e.g., analysis construction and data exploration performance) across multiple steps of the data analysis process. This section provides: (i) an overview of the analysis performed using the obtained tracking records; (ii) and further reports analysis findings, including interesting patterns discovered across the interactions of different types of users.

Trying to understand the patterns emerging from the interactions of all users as a whole, does not yield insight on how different types of users (in terms of characteristics) interact with the system. In adaptive systems such as IDEALVis, the information on how different users approach their analysis is essential for suggesting effective practices and for further defining/providing related adaptations. The characteristics of each user in IDEALVis are maintained by the user model. Therefore, prior to being able to explore the data analysis interaction patterns of different user types, the term "User Type" needs to be defined based on the characteristics of the users. In the context of this work, User Type is defined as distinct group (or class) of users, where all users assigned to that group are similar in terms of their characteristics. The practice of categorizing data (in this case users) into distinct classes or groups is referred to as clustering and is one of the key facets of machine learning.

The classification of users in distinct user groups was performed using the k-means clustering approach. K-means is one of the most popular unsupervised machine learning algorithms that aims to model normal and behavior and group similar records of data in clusters. This algorithm is only able to process numerical variables, since it uses distance-based measures (e.g., Euclidean distance) for calculating the distance between different data points.

The 45 users who participated in the pilot study were classified in different groups using their prominent user model characteristics as the input to the k-means clustering algorithm. Characteristics used as input include (i) human factors (Speed of Processing, Field-Dependent Independent and Working Memory); and (ii) user demographics (Age, Educational Status and Gender). Educational status refers to a single data property that indicates if a user is a holder of a master's degree. Unfortunately, the Perceived Expertise factor had to be excluded as many users were missing this information from their user model. Additionally, the Control of Attention human factor had a strong positive correlation ($r(43) = 0.7$) with Speed of Processing and it was not used as input to the classification algorithm. Removing correlated input variables is a necessary pre-processing step for k-means to decrease bias. This is because correlated variables represent the same characteristic of a segment, and therefore, a single representative variable should be used.

Prior to classifying users with the above-mentioned characteristics, a few data preprocessing operations were applied on the input data. Those operations are listed below and pose important prerequisites, aiming to improve the output of the k-means algorithm (i.e., the segmentation of users in distinct clusters or groups).

1. **Used Numerical Variables as Input:** All categorical attributes should be transformed (e.g., Educational Status).
2. **Imputed Outliers:** K-means is very sensitive to outliers and noisy data; therefore, those should be handled appropriately.
3. **Performed Normalization/Standardization:** Variables should be the same scale — have the same mean and variance, usually in a range -1.0 to 1.0 (standardized data) or 0.0 to 1.0 (normalized data). For the k-means algorithm to consider all attributes as equal, they must all have the same scale.

1.1 User Classification Results

Applying the k-means algorithm on the user's characteristics resulted in 3 clusters of users (i.e., 3 user groups). The three clusters and the distribution of each input characteristic to each of the clusters is illustrated in Figure 1

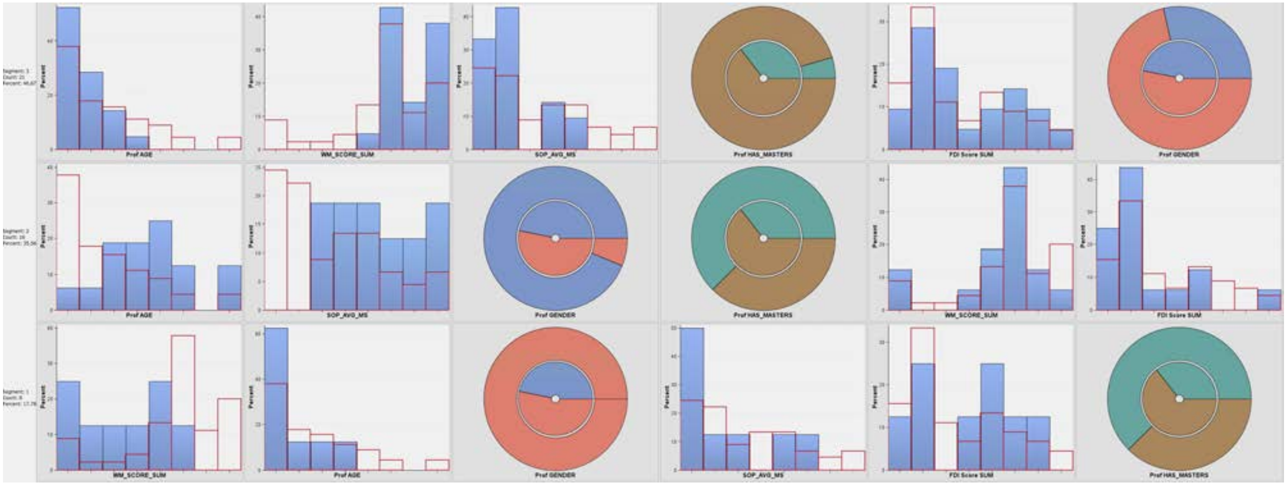

Figure 1 - User Clusters and their Characteristics (Results on Row 1, 2 and 3 belong to User Groups 1, 2 and 3 respectively)

Next, an overview of the characteristics describing each user group is provided.

1.1.1 USER GROUP 1 (SEGMENT 3)

This is the larger of the three groups, with 21 out of 45 users being assigned in this group. With regards to demographics, users in this group are described (i) as mostly being between 25 and 33 years old (80%); (ii) as mostly being male (71%); and (iii) mostly as being holders of a master's degree (95%). With regards to human factors, users in this group tend to have higher Working Memory and Speed of Processing levels compared to users of other groups. Moreover, this group seems to have both field-dependent and field-independent users, with field-independent users being higher in number compared to those of other groups. This group is generally made up of younger individuals that are most likely junior or mid-level analysts, possessing a high level of cognitive abilities.

1.1.2 USER GROUP 2 (SEGMENT 2)

This is the second group in size compared to the three groups, with 16 out of 45 users being assigned to it. With regards to demographics, users in this group are described (i) as mostly being

33 years or older (87%); and (ii) as mostly being female (93%). Also, it seems that a third of the users are holders of a master's degree (37%). With regards to human factors, users in this group tend to have (i) slightly lower Working Memory levels compared to the users of group 1, but higher levels compared to the users of group 3; and (ii) lower Speed of Processing levels compared to users of other groups. Moreover, most users in this group are field-dependent users. This group is generally made up of older individuals that are most likely senior data analysts, having a lower level of Speed of Processing compared to users of other groups.

1.1.3 USER GROUP 3 (SEGMENT 1)

This is the last and smallest group of users, with 8 out of 45 users being assigned in this group. With regards to demographics, users in this group are described (i) as mostly being between 25 and 29 years old (62%); and (ii) as being male. Also, it seems that a third of the users are holders of a master's degree (37%). With regards to human factors, users in this group tend to have (i) lower Working Memory levels compared to users of other groups; and (ii) higher Speed of Processing levels compared to users of group 2, but slightly lower levels compared to users of group 1. Moreover, users in this group are mostly field-dependent, while some tend to be intermediates in terms of the field-dependent independent scale. This group is generally made up mostly of younger individuals that might be junior or mid-level analysts, possessing a lower level of Working Memory compared to users of other groups.

1.2 Group-related Analysis of Interaction Patterns

All the 45 pilot study users were assigned to one of the abovementioned user groups. The tracking records (i.e., system interactions) of all users belonging to each group were analyzed for better understanding the patterns adopted by each user group. The tracking metrics used in this analysis are listed below. For more information on how these metrics were collected by the platform please refer to deliverable D13.

- **Data Analysis Attempts**
  - The total number of times a user utilized the Analysis Wizard for producing an analysis report.
- **Select Analysis Method Duration**
  - The average time spent in the Select Analysis step of the Analysis Wizard. This is the step where the user selects the appropriate analysis method/type for their analysis.
- **Select Attributes Duration**
  - The average time spent in the Select Attributes step of the Analysis Wizard. This is the step where the user selects the appropriate data attributes for their analysis.
- **Result View Duration**
  - The average time spent by the user viewing/interpreting a specific analysis report.

Those are the tracking records captured for all user during the pilot study. The reason only 4 metrics are presented, is because during the pilot study some of the Analysis Wizard steps (Select

Dataset and Select Output) were unavailable/disabled for the users' data exploration, and thus, records for those metrics were not available for analysis. Results of this analysis regarding each of the abovementioned tracking metrics are presented below, shedding light on how distinct user groups differ in terms of their interaction.

Data Analysis Attempts: On average users of User Group 1 made less data analysis attempts for addressing all pilot study analysis tasks compared to other user groups. Their average data analysis attempts were 49.5 ± 6.4. The user with the least attempts from User Group 1 performed on average 38 attempts, while the user with the highest number of attempts performed on average 64 attempts. Moreover, users of User Group 2 made on average 51.1 ± 5.9 data analysis attempts for addressing their tasks. The user with the least attempts from User Group 2 performed on average 39 attempts, while the user with the highest number of attempts performed on average 62 attempts. Finally, User Group 3 was the group with the highest number of data analysis attempts for addressing all pilot study analysis tasks with an average of 52.4 ± 8.2 attempts. The user with the least attempts from User Group 3 performed on average 42 attempts, while the user with the highest number of attempts performed on average 66 attempts.

Select Analysis Method Duration: On average users of User Group 1 were faster compared to users of other groups when selecting the appropriate analysis method for their exploration. Their average performance in seconds was 9.1 ± 2.3. On average the fastest user for this metric in User Group 1 achieved a performance of 5 seconds, while the slowest achieved a performance of 13 seconds. Moreover, for this metric, users of User Group 2 were on average the slowest with a performance of 13.6 ± 8.8 seconds. On average the fastest user for this metric in User Group 2 achieved a performance of 5 seconds, while the slowest achieved a performance of 34 seconds. Finally, User Group 3 performed slightly faster than User Group 2 when selecting the appropriate analysis method. Their average performance in seconds was 10.6 ± 7.0. On average the fastest user for this metric in User Group 3 achieved a performance of 5 seconds, while the slowest achieved a performance of 27 seconds.

Select Attributes Duration: On average users of User Group 1 were faster compared to users of other groups when selecting the appropriate data attributes for their exploration. Their average performance in seconds was 62.5 ± 14.0. On average the fastest user for this metric in User Group 1 achieved a performance of 41 seconds, while the slowest achieved a performance of 101 seconds. Moreover, for this metric, users of User Group 2 were on average the slowest with 78.9 ± 19.6 seconds. On average the fastest user for this metric in User Group 2 achieved a performance of 52 seconds, while the slowest achieved a performance of 119 seconds. Finally, User Group 3 performed slightly faster than User Group 2 when selecting the appropriate data attributes. Their average time was 70.5 ± 34.9 seconds. On average the fastest user for this metric in User Group 3 achieved a performance of 36 seconds, while the slowest achieved a performance of 151 seconds.

Result View Duration: For the current user sample, the performance of all user groups for this metric was on average equal. Specifically, users of (i) User Group 1 achieved an average performance of 16.0 ± 5.7 seconds; (ii) User Group 2 achieved an average performance of 16.0 ± 4.7 seconds; and (iii) User Group 3 achieved an average performance of 16.0 ± 6.5 seconds. On average (i) the fastest user for this metric in User Group 1 achieved a performance of 7 seconds, while the slowest achieved a performance of 25 seconds; (ii) the fastest user for this metric in User Group 2 achieved a performance of 8 seconds, while the slowest achieved a performance of 25 seconds; and (iii) the fastest user for this metric in User Group 3 achieved a performance of 10 seconds, while the slowest achieved a performance of 27 seconds.

2 Effective Adaptation Practices

This section provides a set of best adaptation practices which are deemed as the best fit visualization settings for each of the different user groups that emerged through user classification, using input from D11.

The motivation behind this approach, stands on the fact that each user group contains a set of users which possess similar characteristics. Therefore, by understanding the key characteristics of a user group (done in the previous section), it is also possible to assign to it, a set of adaptation practices/rules that are compatible to the group's user characteristics. Sections below present three sets of effective, collective adaptation practices/rules, each assigned to a specific group of users. If applied, those collective rules, will most likely benefit the corresponding users of each target group with regards to their ability to process data visualizations.

NOTE: The terms (i) task complexity; and (ii) data visualization elements appearing in the next sections are defined in deliverable D11.

2.1 *Effective Adaptation Practices for User Group 1*

For simple and medium complexity tasks, this group of users will most likely benefit by a column chart as the data visualization type. Additionally, according to findings of D11 for higher complexity tasks, these types of users tend to perform better specifically on pie charts and radar charts. If those charts are not an option in a specific situation, a line chart can also be used. Below a list of data visualization elements is provided and their applicability/usefulness to this group of users according to the group's characteristics.

- **Proximity:** Proximity between bars and columns will negatively affect these users when used on bar and column charts for low complexity tasks. Proximity will only be useful to this group of users on medium and high complexity tasks.
- **Element Size:** Changing the size of primary data visualization elements in medium complexity tasks for this group of users will negatively affect their performance. Changing the element size on high complexity tasks can provide a performance benefit, while it might also help such users for low complexity tasks. The latter effect is only attributed to the high levels of Speed of Processing that users of this group possess.
- **Grid Lines:** Enabling grid lines on data visualizations will benefit users of this group in terms of performance for all task complexity levels (only low and medium complexity tasks were investigated in D11 for this element). Compared to other user groups, this group is likely to benefit the most by this intervention at medium complexity tasks.
- **Data Labels:** Data labels on data visualizations will benefit users of this group in terms of their performance only when enabled with low complexity tasks.
- **Dark Theme:** Dark theme on data visualizations will benefit users of this group in terms of their performance only when enabled with low complexity tasks.

- **Sorting:** Sorting the data on a data visualization will be beneficial in terms of performance for this group of users, especially for high complexity tasks.
- **Palette 1:** Changing the colour palette setting on the data visualizations to Palette 1 can benefit user of this group achieve a higher performance across all levels of task complexity.
- **Palette 2:** Changing the colour palette setting on the data visualizations to Palette 2 is better in terms of performance for this group of users on low complexity tasks when compared to Palette 1. Palette 2 should be avoided for medium complexity tasks and instead Palette 1 should be used. Moreover, for high complexity tasks Palette 1 is superior in terms of enabling better performance and should therefore be preferred for this group of users.

2.2 Effective Adaptation Practices for User Group 2

For simple and medium complexity tasks, this group of users will most likely benefit by a column chart as the data visualization type. Additionally, according to findings of D11 for higher complexity tasks, these types of users tend to perform better specifically on pie charts and radar charts. In case those charts are not an option in a specific situation, a column chart can also be used. Below a list of data visualization elements is provided and their applicability/usefulness to this group of users according to the group's characteristics.

- **Proximity:** Proximity between bars and columns will negatively affect this group of users when used on bar and column charts for low complexity tasks. Proximity is useful to this group of users only on medium and high complexity tasks.
- **Element Size:** Changing the size of primary data visualization elements in medium complexity tasks for this group of users can be done but it will not offer any significant performance benefit. Moreover, for low complexity tasks changing the element size might slightly degrade the performance of these users. In contrast to User Group 1, for this group changing the element size is mostly suggested for high complexity tasks.
- **Grid Lines:** Enabling grid lines on data visualizations will benefit this group's users in terms of their performance for all task complexity levels.
- **Data Labels:** Data labels on data visualizations will benefit users of this group in terms of their performance only when enabled with low complexity tasks.
- **Dark Theme:** Dark theme on data visualizations will benefit users of this group in terms of their performance only when enabled with low complexity tasks. While not suggested, it is interesting to note that higher expertise users of this group might be able to benefit from the dark theme on medium complexity tasks, since expertise was found to have an interaction with this visual element on even more difficult tasks.

- **Sorting:** Sorting the data on a data visualization will be beneficial in terms of performance for this group of users especially for high complexity tasks. It should be noted that results indicate that users from User Groups 1 and 2 are likely to benefit more when data is sorted on high complexity tasks, compared to this group. This effect is attributed to the higher expertise possessed by this group of users.
- **Palette 1:** Changing the colour palette setting on the data visualizations to Palette 1 can enable users of this group achieve a higher performance especially for low and high complexity tasks.
- **Palette 2:** Changing the colour palette setting on the data visualizations to Palette 2 is better in terms of performance for this group of users on low complexity tasks when compared to Palette 1. Palette 2 should be avoided for medium complexity tasks and instead the default palette or Palette 1 should be used (the default palette is provided as an alternative because for the characteristics of these users on medium complexity tasks, Palette 1 did not provide a significant performance increase compared to the default palette). Moreover, for high complexity tasks Palette 1 is superior in terms of enabling better performance and should therefore be preferred for this group of users.

2.3 Effective Adaptation Practices for User Group 3

For simple and medium complexity tasks, this group of users will most likely benefit by a column chart as the data visualization type. Additionally, according to findings of D11 and similar to users of group 2, for higher complexity tasks, these types of users tend to perform better specifically on pie charts and radar charts. If for some reason those charts are not an option in a specific situation, a column chart can also be used. While these practices for selecting a data visualization type seem similar to those of User Group 2, users in this group tend to have more performance benefit when using the column chart compared to users of group 2. Below a list of data visualization elements is provided and their applicability/usefulness to this group of users according to the group's characteristics.

- **Proximity:** Similar to the rest of the user groups**,** enabling proximity between bars and columns on medium and high complexity tasks will benefit these users' performance. Enabling proximity on low complexity tasks is not suggested.
- **Element Size:** Changing the size of primary data visualization elements in low and medium complexity tasks for this group of users will negatively affect their performance. For this group changing the element size should be done only on high complexity tasks.
- **Grid Lines:** Similar to other user groups enabling grid lines on data visualizations will benefit this group's users in terms of their performance for all task complexity levels.

- **Data Labels:** Data labels on data visualizations will benefit users of this group significantly in terms of performance when enabled with low complexity tasks. Moreover, this visual element may also slightly help these users when performing medium complexity tasks. This latter effect is reported since multiple human factors of this user group (low Working Memory levels, medium Speed of Processing levels and mid-level expertise) have a positive interaction with this visual element in terms of performance. This visual element is not suggested for these users when performing high complexity tasks.
- **Dark Theme:** Dark theme on data visualizations will benefit users of this group in terms of performance when enabled with low complexity tasks.
- **Sorting:** Sorting the data on a data visualization will be beneficial in terms of performance for this group of users especially for high complexity tasks.
- **Palette 1:** Changing the colour palette setting on the data visualizations to Palette 1 can help user of this group achieve a higher performance across all levels of task complexity.
- **Palette 2:** Changing the colour palette setting on the data visualizations to Palette 2 is better in terms of performance for this group of users on low complexity tasks when compared to Palette 1. Palette 2 should be avoided for medium complexity tasks and instead Palette 1 should be used. Moreover, for high complexity tasks Palette 1 is superior in terms of enabling better performance and should therefore be preferred for this group of users.

3 Conclusions

The present deliverable’s focus was to illustrate the machine learning techniques used as part of WP5 for performing user classification, as means to further explore the users’ interaction data collected during the pilot study. The process of classification enabled the separation of users to distinct user groups (based on the similarity of their characteristics), and thus facilitated the process of analyzing the users’ interaction data; allowing the discovery of analysis patters adopted by different groups (i.e., types) of users. More specifically, the deliverable (i) presented the user classification approach used; (ii) provided the description of characteristics for each of the user groups that resulted from classification; (iii) discussed the analysis results of the captured interaction data which was analyzed for each separate group of users, for discovering adopted analysis patterns; and finally (iv) using findings from deliverable D11, three sets of effective adaptation practices were presented, each set attempting to provide the best fit visualization and visual element settings, for each of the user groups.