Differentiating user groups within an educational dashboard using log data

by

Robert Brouwer

A Master's Thesis

Business and IT

University of Twente

The Implementation Group

August 2023

Abstract

Gaining complete understanding of an educational ecosystem is complex. To help with this dashboards exist that help teachers, board members and staff to gain overview of the school. Better understanding of the whole educational ecosystem is gained via these dashboards. Improving these dashboards is a complex task. Given that developers typically only have access to a small subset of user's understanding, differences in usage of the dashboard is difficult to identify.

In this research we looked at if all the users can be assigned to user groups based on measured behaviour. By measuring different features based on log data of the dashboard, different users can be grouped together into unique user groups. Grouping is performed by employing K Means and Hierarchical clustering. We found that K Means delivers better results and that there are multiple independent clustering's into which users can be grouped. The different groups give insight into distinct usage of the dashboard, but further research is necessary to better understand if the user groups differ in interests, goals and concerns.

Table of contents

Table of tables

Table of figures

1. Introduction

In education, a teacher helps students gain new insights and knowledge. Every student has their own needs which need to be considered. Different students have in general different needs and tracking the needs of many different students is complex. In secondary school, when students have a variety of teachers each giving their own subject makes it a difficult and complex task to gain a holistic overview of the performance of each individual student. Such a complete picture is necessary to identify what might cause each student's performance.

Examining each student's performance requires not only information on the student's progress in term of educational advances and individual behaviour. A student learns within an educational ecosystem, including at least teachers, staff and co-students, and also content, technology, culture and strategy. So for understanding the progress of the student, the individual performance needs to be viewed in the context of this ecosystem. Understanding the student's performance requires a holistic view on the student within such educational ecosystem, and a holistic view on the educational ecosystem as such.

Registration of grades and behavioural data enables first of all tracking the progression of the individual student. Next to administrating results, registration of progression including behavioural data helps with better students' development. Examination results and other student information can be registered in student administration systems or LAS ("Leerling Administratie Systeem").

This registration of student progress per student, and integrating this into a holistic view also helps schools to see their progress with their "school plan". School plans contain the planning of how the quality of education is assured in schools. Every school in the Netherlands is mandatory create a school plan every four years (par. 9 art. 2.88 Wet voortgezet onderwijs 2020).

To track the performance of a school, i.e., if a school is still able to deliver a certain level of education, school inspection also use different indicators for assessing the performance of the school. These different indicators must be registered by the school (Ministerie van Algemene Zaken, 2023; Ministerie van Onderwijs). Registration of this information, but also broader items such as students' development and results, are also saved within the LAS. In the Netherlands, the biggest supplier of such systems is Magister, which was the market leader with a share of around 70% in 2020 (Magister, 2023; Smit, 2020).

Examining the performance of both students and the educational ecosystems is done through dashboards. The dashboard that is the focus of this thesis aims at providing insight into the performance of the educational ecosystem of secondary education institutes. To the knowledge of the researcher, literature concerning dashboards focussed on monitoring the educational ecosystem within secondary education does not exist. Therefore this research utilizes relevant resources from literature related to tertiary education, which is education that follows secondary education (e.g., higher education or vocational education).

Registration of student performance information is not only due to outside pressure from government, but also based on an internal need. First of all, it is necessary for teachers when they are reflecting on their own practices, especially since they have to track a high number of students (Isaias & Backx Noronha Viana, 2020). Good visualization helps teachers identify possible strugglers or potential drop-outs and helps to intervene in their learning progress (Isaias & Backx Noronha Viana, 2020). Next, good insights are not only necessary for teachers but also for governing boards. Board meetings are packed already with information exchange so that how they can contribute to improving student performance is not discussed enough. (AGB, 2014; Muntean et al., 2010).

Delivering the correct information to the boards is critical for board assessment of the institutes' performance, i.e. assessing the educational ecosystem. Correct evidence from student learning, such as grades and performance, is required, while also indirect evidence such as surveys or school reviews are necessary. Dashboards are a clear way to deliver this evidence (AGB, 2014; Muntean et al., 2010).

Different dashboards have been designed for educational organisations. Such dashboard are designed with a large variety of goals in mind. Dashboards can cover a broad spectrum of information, which is not limited only to students performance but may also include financial and staff performance (Denwattana & Saengsai, 2016; Muntean et al., 2010).

Different techniques are used to research how a dashboard fits within the educational context. For good design, appropriate techniques are necessary, such as user sessions, interviews or from dashboard design literature (Chalvatza et al., 2019; Dickman et al., 2011; Iriberri & Stengel, 2021; Isaias & Backx Noronha Viana, 2020; Manwaring et al., 2017; Polikoff et al., 2018; Schellekens et al., 2022; Schwendimann et al., 2017). All these techniques try to find the user group and their requirements, aiming to make a design which is as insightful and as easy in use as possible. Evaluation after deployment is however limited and how the implementation is used in practice has not been researched fully.

Quantitative research after deployment would create a deeper understanding of how the dashboards are used. Measuring usage of the dashboard by averaging over all users is possible, but creates the illusion of a homogeneous userbase. A better understanding would be gained if the users could be split into groups, where the users within each group are rather similar, while users of different groups are distinctly different. User groups are thus those who use the dashboard in a similar way.

Identifying if there are different user groups within the users of the dashboard helps to understand how the dashboard is used. This can later on be used in further improvement. Finding out if different user groups exist and what differentiates them is necessary for the improvement of the dashboard. It allows the developers of the dashboard to optimize the dashboard for particular types of usage, and avoids to pitfall into the trap of optimizing for the non-existent average user. Finding these user groups and understanding their differences is the main focus of this research.

TIG is a company located in the Netherlands specialized in providing business intelligence solutions for education. Users of their BI applications are located in the Netherlands and TIG delivers the application as a SaaS-solution (Software as a Service). Their applications are used both in secondary education and in tertiary education. The different applications are focused on different aspects of the information needs of educational institutions.

This thesis is structured as follows: Chapter 2 discusses the Research Design. Based on the main objectives of the thesis, it formulates the research questions and the methodology that is followed to answer these questions. Chapters 3 through 7 cover each a step of the methodology. Chapter 3 investigates the different TIG needs. Chapter 4 researches both the source of the data and the data quality. Chapter 5 explains how the data from the original source is transformed to function as an input for modelling. Chapter 6 includes both the creation and the validation of the models. Chapter 7 reflects to what extend the model can fulfil the main objective. Chapter 8 covers the conclusions of this research. Chapter 9 covers the possible limitations, implications and relevance of the thesis are and lastly Chapter 10 formulates recommendations for future research.

2. Research design

2.1. Problem statement

TIG has a wide variety of dashboards for monitoring and examining data in education. For TIG it is important to deliver dashboards that are of high quality. Regular improvements are needed and provided to achieve this. Design decisions of dashboards within TIG are based on insights of TIG's developers. Market research is a regular effort for better insight in their users. The market research includes surveys, focus group sessions and interviews.

Interviews are key for understanding the users. The understanding however is limited to the insights of the developer and the interviewee. TIG expresses the concern that the interviewee is only a certain type of user, the so called super user. A super user is someone with more knowledge and uses the system more than most other users. The behaviour of super users in the software might be different from that of regular users and outcomes of interviews are too much influenced by a subgroup of the users. So the primary goal of TIG for the research reflected in this thesis is to understand if there are typical users, which are not by definition super users.

Within research, there is also a lack of knowledge measuring the usage of educational performance dashboards focussing on the performance of the educational ecosystem, in particular considering the existence of different types of users. So can user types be identified, i.e. extracted from metadata that is available concerning users in the dashboard?

Features could be defined using this metadata to measure the users' characteristics. A feature is a measurable aspect of a user that potentially distinguishes them from other users. Still it is necessary to find out if the feature defines the user. For this, an analysis of different educational performance dashboard features should be performed.

2.2. Research objectives

The goal of this research is to find out if there are different identifiable user groups that make use of educational performance dashboards. User groups are those who use the dashboard in a similar way. Those who belong to a user group don't need to have the same role, organisation or permissions.

Identification of different user groups gives a better understanding of the way different people make use of the same dashboard. It is possible to understand the complete user population from general statistics. However, splitting in user groups allows to approach the users not as one uniform mass, but see differences between users.

If user groups are found than it also helps TIG in understanding who they might need to approach in future research related to their dashboards. If multiple user groups exist than every group should be contacted separately.

In order to create sensible user groups, it is important for the research to identify what should be used for splitting. It is considered that the user actions are stored in quite some detail as log-data, therefore log-data is expected to be potentially a useful source for determining features that may be used to distinguish between user groups. Defining which features are important according to TIG is necessary for identifying their business needs for the research.

For the creation of the user groups some model needs to process these features. The user groups are unlabelled, because a priori it is unknown to which user groups someone belongs. Creating user groups out of unlabelled data requires unsupervised learning algorithms. In particular, creating groups requires a clustering algorithm.

2.3. Research questions

The main research question is therefore formulated as:

Which user groups can be differentiated considering characteristics of their behaviour in an educational performance dashboard based on log-data?

This question is a summary of the goal of the research. Differentiation between user groups is necessary for a more complete picture for TIG. Characteristics of behaviour are the features that are the input for the eventual model. Behaviour is any action that someone performs with the dashboard. Educational performance dashboard is the type of dashboard that is the subject matter. The log-data is the source where all information is gathered from.

To help answer this main research question, it is broken down into the following sub questions all considering the scope of an educational performance dashboard:

1. What behavioural characteristics are relevant for differentiating between user groups?
2. What is the quality of the available data?
3. What features can be distilled from the data that is available?
4. How should features be mapped onto available data?
5. What models should be used for clustering users into user groups?
6. How stable and reliable are the clusters?
7. Which user groups are determined from the clusters?

2.4. Methodology

The methodology of this research has been based on the CRoss-Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000). This is used because it takes into account the business needs of TIG as if translates such needs into an approach of researching data for such needs. Understanding of business needs has been deemed important and thus the choice for CRISP-DM has been made (Azevedo & Santos, 2008). CRISP-DM is not a fully linear methodology but cyclical in parts. Some parts described below will be performed multiple times, however a linear description is given. The next sections describe the individual steps undertaken. Graphical overview of the method is represented in Figure 1.

2.4.1. Business understanding

Business understanding focusses on formulating the needs of a company within the scope of a project.

This starts by exploring the context of the research in order to grasp a good understanding of the domain. Knowing that the scope of this research is the identification of user groups in the population of educational performance dashboard users, the end goal of this phase is to answer the first sub-question.

The business needs from TIG's perspective are formulated as features, which are the eventual input for the model. A stakeholder focus group assisted in incorporating TIG's expertise in this formulation of features. This is a way for TIG to express what they expect from the model. Defining with TIG the input of the model is important because input determines output. The list of features will be the input for the eventual model.

2.4.2. Data understanding

Data understanding addresses the identification of the sources of the data and its quality.

Data understanding aims at answering sub questions 2 and 3. A deep understanding of where the data comes from is necessary. Checking the data quality and taking actions for improving are essential for creating trustworthy results. These steps are an integral part of the eventual creation of the features that will be the input of the model.

Within the scope of this thesis, the main source is log data available in the internal database of TIG. Within data understanding, these database are investigated. This investigation includes the assessment of the quality of the database, and selection of data that is deemed of sufficient quality.

2.4.3. Data preparation

Data preparation focuses on making sure that the data is ready to use for the eventual modelling.

Within the scope of this thesis, this means that the list of features from data understanding will be implemented in this phase. The main focus will be on describing the features and the conditions they operate under. Some transformation of the data are also necessary for the eventual model which also happens in this step. As output, a fully prepared dataset will be available for the model.

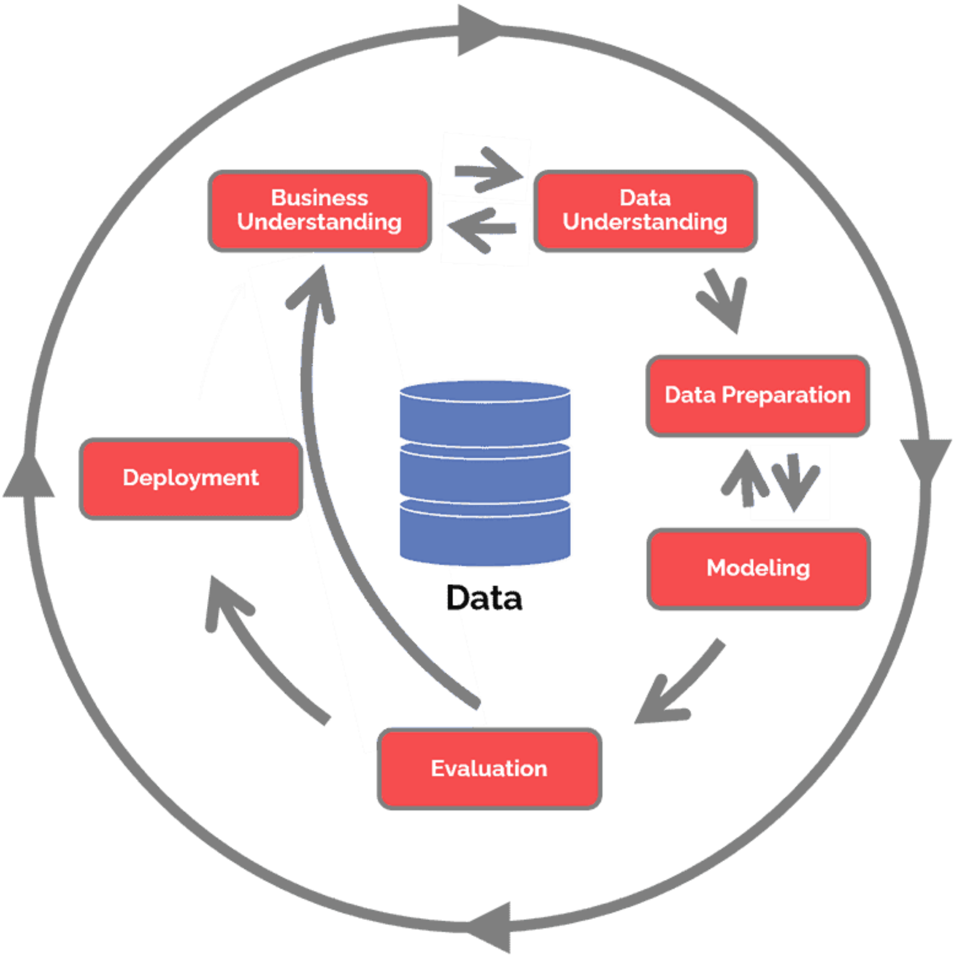

Figure 1 The different phases as defined by CRISP DM (Hotz, 2023)

2.4.4. Modelling

The modelling phase focusses on creating the model and assessing its validity.

The scope of this thesis includes potential identification of user groups from unlabelled data. So the model will be a clustering model using the features as being defined so far. The modelling considers alternative base models from literature and applies these to the available features. Analysis of the models consider the distinctness of the clusters the models create. If clusters are distinct enough then they represent possibly different user groups. The goal of this step is to end with one or multiple validated models which are able to cluster different users together in a group.

2.4.5. Evaluation

Evaluation aims at finding out if the model fulfils its objective.

This thesis models using clustering algorithms. A clustering algorithm by definition creates clusters. However, this does not yet imply that these clusters make sense. The evaluation analyses the found clusters, and addresses if these clusters describe potential user groups and how they differentiate from each other. This translates into what this means to TIG in a broader context, and what is learned.

2.4.6. Deployment

Within CRISP-DM deployment delivers the results to the business and incorporates the results in the business.

Deployment is fulfilled by presenting the results to TIG and by the elaboration of this thesis. Incorporation within the business operations is not be covered as it is not part of the research assignment. The needs for deployment will however be described in the final chapters of this thesis.

3. Business understanding

Business understanding focusses on the needs of TIG in relation to the project. The aim is to delineate the research in line with the business deeds and include the expertise of TIG in the definition of the research.

The first section of this chapter investigates the domain of research. It discusses why TIG chose a specific application as subject for research and describes the application itself as well as its use. This gives insight in the functioning of the application and how customers use it.

Characterization of users is supposed to the done based on the log data of the selected application. This raw input data is not suitable to characterize users. First useful features need to be defined that reflect users in a way that is aligned with the business needs. As this can only be successful by involving the business, TIG is involved through a stakeholder focus group. The second section of this chapter describes the stakeholder focus group and its outcome, which is a list of feature collections that TIG desires.

3.1. Domain description

Firstly, a scope for the research must be defined to have a focus for a specific domain. For this research one of the TIG applications is chosen. The choice for this application was primarily made by TIG. There are multiple motivations for choosing this application over other existing applications within the organization. Within TIG there are also other projects that investigate better understanding the application, but these focus on interviewing users directly. This makes the results directly applicable to improvements of the application. Next to this, the application contains the highest number of unique users of all available applications.

The application is used by secondary schools (approximate ages of 12-18) in the Netherlands. In the application, information is shown which is saved by schools in student administration systems or LAS ("Leerling Administratie Systeem" in Dutch). A LAS such as Magister (Magister, 2023) contains different information about the students' educational performance. Most information recorded in the system are about the grades a student receives.

Information of students which are for example grades are recorded in student information systems. The application itself has access to the data, but outside of the owner it is not accessible. Any information that comes from the LAS surrounding student grades or performance cannot be used in this research.

The users of the application include different parts of the educational staff. Staff members that make use of it include teachers, administrators and management. Because the information within the student information systems is of varying nature, different overviews are created which are called sheets. Every sheet is dedicated to a specific subject. For example, there is a sheet that is the first visited. This sheet shows information which describes the performance of the school with regard to certain KPI's (Key Performance Indicators). All users have access to the same sheets in the application. Other sheets might show information specific to a certain data object or certain type of data. Data from the application can be exported by the users. With this, a copy of the dashboard can be created to show to other people outside of the application. The current state of the dashboard can also be saved to facilitate retrieval for a later moment. Filtering gives the users the option to focus on a subset of the data of interest.

3.2. Stakeholder focus group session

The eventual model must fit TIG's needs. The output of any model is mainly determined by the input. Aligning the input of the model with TIG's needs makes sure that the output aligns with what TIG wants from the model. For models, the input is a selection of features. A feature is a measurable aspect of a user. The outcome will be a list of feature collections. A feature collection is a group of features that are similar, but measure something slightly different. The stakeholder focus group session is based on requirements engineering. A requirement is something that is necessary to have in the end product. A list of features does not directly represent a list of requirements, but can reflect what the stakeholders find important.

The end goal is to define a feature list in cooperation with internal stakeholders from TIG. The list of feature collections reflect the wants and needs of TIG for the model. The next section describes how the internal stakeholders together are able to create such an feature collection list.

3.2.1. Organization

The stakeholder focus group session is based on principles taken from requirements engineering. Since the features can be perceived as requirements for the eventual model this is regarded as an appropriate approach. The identified features are based on individual and communal needs.

Before starting the stakeholder focus group session, it is important to define a methodology of the session and the goal of the session. It is important to define an appropriate way to identify the features (Gottesdiener, 2002). Before the start of the goal of the session was communicated. It is best to use different ways of engaging with the requirements to increase the quality and variety of the identified features (Gottesdiener, 2002).

During a session, a trade-off needs to be made between individual stakes of all participants and more global goals (Konaté et al., 2014). It is important that individual or global goals do not overshadow each other. Although features collections are about measuring data they are meant to represent what stakeholders find important. A good balance between different goals can only be achieved by using an appropriate structure (Konaté et al., 2014).

A balance between individual and global stakes is achieved by firstly having a part where users engage individually. Herein they express what they exactly want. These are then presented to the group. To get to the global goals, the whole group then discusses the gathered features. These are ranked together and merged to obtain one total list of feature collections. By first engaging individually and then as a group a balance between individual and global needs is achieved.

It is important to use correct prompting to keep participants active. Correctly prompting the participants helps them think more broadly about different aspects of the features. The main technique used for generating information was the interrogatory technique (Browne & Rogich, 2001). The interrogatory technique tries to ask open questions when generating information is important. This helped identify a more broad set of features. The technique however does rely on the ingenuity and abilities of the questioned subject (Browne & Rogich, 2001).

The focus group session took place with a group of four participants. In the session all participants met in person except one who joined through a video call. The different roles of participant have with respect to the application are: product owner of the application, lead data scientist and two application developers. All participants are directly involved with the development of the application.

In total one hour was available for the focus group session. Writing tools in the form of whiteboards were available for all participants. The one joining online performed all activities online on their own computer, but shared the results. Later on discussions were performed in group using the same whiteboards. The outcome of the whole session was in the end dictated and the whole session was recorded for later on review.

3.2.2. Results

The outcome of the analysis from the focus group session was a list of different feature collections. During the session they were ordered and given a specific score. Scoring the relevance of different feature collections was done with involvement of the stakeholders. Giving a score allowed to order them. These are important when later on selecting the features for implementation

In Table 1 is shown the feature collections that TIG identified as important. The included features are shown as TIG defined them. Features are not precisely defined thus some interpretation needs to be employed during data preparation. Creation of the list doesn't yet take into if data is readily available or not. This list of feature collections is however too extensive for this thesis and will need to be shortened during of data understanding.

Table 1 feature collections created by the stakeholder focus group

Feature collection	Relevance	Feature	Relevance
Which combination of sheets are used	9	Diversity of sheets used in a session.	8
Which sheets are used	8	Usage of filters	8
Exports making and repeats	8	Reporting made and repeats	8
Bookmark usage	8	Mistakes made by user	8
Speed of clicks	8	G4/G40/other regions	7
Number of sessions	7	Session length	7
Function of user in their organisation	7	Selections made per sheet	7
Growth/decline of school	7	Digital proficiency	7
Colleagues within the application	6	Size of school	6
Users in organisation who are data coach	6	Support tickets	6
School type	6	Inspection indicators	5
Time of logging in	4	Consistency of logging in time	3
Number of licenses within organisation	3	Gender	3
Authorization restrictions	3	Age	3
Subject given by user	3	Years in organization	3
Usage of quality calendar	3	Salary scale someone is in	3

4. Data understanding

Data understanding focusses on how the data is collected and the quality of the data that is used for the conducted research.

The first section of the chapter describes the database containing the raw data. This forms the basis of the features on which the research is conducted. The data from the database is complemented by a small amount of additional data not available in the database. This is shortly described in the second section.

The data extracted from the database may not provide consistent quality. The third section discusses the result of the data quality analysis. It discusses the analysis, all found inconsistency in the data and how it is handled. Various actions are undertaken to improve data quality.

The business understanding resulted in a large list of potential features that might define its users. This list is too large for the analysis. The last section of this chapter discusses the selection of feature collections that are eventually used during the next steps of the research.

4.1. Database overview

A read-only database has been provided by TIG. Data contained in this database comes from different sources. Part of the data is the direct log of the application and a small part comes from other systems.

The database was read-only, however access to another database was provided for creating temporary tables. Once a day the database is updated. The earliest records in the database are from 01-01-2020. For the eventual analysis, data has been included until 31-05-2023.

In the database, data about other products are also contained. These need to be filtered out when selecting the data. All tables were necessary for the creation of the features although some fields of the tables have not been used. How features will be implemented given this information is discussed in chapter 5.

Next to information from the database itself, also some extra information was available. These is data describing the current school size, predicted school size in the future and location of the school.

4.2. Data quality

To make sure that the information provided is correct, an analysis must be performed to identify if there are structural or random errors in the data. In this section, we discuss multiple different anomalies with the data that have been found and how they are handled.

We had to make sure that all potential error’s in the data were identified is important. For this, the different aspects of data quality must be kept in mind. Quality data is accurate, complete, unique, current, valid and conform with the standard data formats (Craig Stedman, 2022). By going through the data and checking it for these aspects, errors can be found. They are appropriate for checking data quality, but quality assessment cannot guarantee all mistakes have been removed from the dataset. This is the first check on data quality on this dataset.

In this section, some of the data quality researched is provided. In some cases not all information about the user is available. The researched users are limited to those whose log data is verified to be accurate.

4.2.1. Employee

Some companies and users must be removed from the dataset. These users are employees of TIG or fake companies which only exist for TIG's testing purposes. TIG employees use their account when developing the applications. They do not represent any customers and are thus not users of interest for us.

4.2.2. Limited sessions

In the dataset, there are various different users with only a limited number of sessions. Users with a small number of sessions have more extreme data due to lack of datapoints. It is chosen that users with five or less sessions are not counted. As we assumed that users with only a limited number of sessions do not produce stable features, they have been excluded from the dataset.

4.2.3. Application changes

The application is approached as if it were a stable immutable platform. This is obviously not true and there have been a number of changes over time. To find out if any major changes have taken place, the release notes of the application have been sifted through. In them, a total of four changes were found that affect the structure.

There are multiple different types of changes. Firstly, there is the elimination of a sheet. In this case the sheet is not any part of the current dashboard anymore. In those cases it is possible that they have been merged into another sheet. If that has happened, then any visitation to that sheet is observed as visits to the sheet it was merged into. In the case the sheet is fully dropped, visits are counted towards the more general features, e.g., sheets visited per session. Dropped sheets will not be measured as a separate feature in the end.

Secondly, new sheets can be created. In those cases information about visitations to the sheet only start from its creation. If the sheet was created by splitting from an older sheet than a possibility is to fill in the non-existing data with the visits to the sheet it has been split from. We chose not to do it to prevent overcounting actions.

Besides these major structural changes, minor updates have also been made. These include updates that add new functionalities, dimensions or settings, but do not change the structure. As they do not change why someone visits a sheet, they are not acted upon.

Unfortunately, during the period in which the data was gathered, there have been changes in the application, as well as cultural changes. This means there is likely to be a difference between how people used the platform in 2020 versus 2023. Curtailing the period in which measurements are taken only entails reducing information about individual users. To keep the reliability of the features as high as possible, we considered better in this case to use the full period instead of limiting it to a shorter time frame.

4.2.4. Validity check

We made a sub selection of the users to keep only data that could be verified to be accurate. This means that some users are not included. A comparison was made to compare users removed with those kept within the dataset. On some of that were measured later, features differences were found such as speed of clicks and export usage. However, in most areas the removed users and those used didn't differ significantly when it comes to most to be measured features. Therefore we assumed that conclusions from the analysis can be applied to the users as a whole.

4.3. Features selection

Combining the knowledge gained from the stakeholder focus group and data understanding a selection of features can be made. The features are mostly not directly available within the data and must be created. Some features might take more time to obtain than others. Those created faster should be prioritized over those that cost more time. The amount of time something costs is assigned based on assumptions of the author.

All feature collections from the stakeholders focus group were scored by giving a 10 to those that can be quickly created and a 1 to those that would cost a long time. The two scores are afterwards averaged, which gives a score to the feature collection. The score represents the priority for obtaining. Given the limited time frame, only the feature collections at the top of the list have been collected.

Table 2 shows the feature collections with their score. It also indicates if during data preparation the feature has to be created. Most of the feature collections consist of multiple features. Which sheets are used will in the model be measured by multiple features, since a separate feature is needed for every sheet.

Table 2 overview of all features collections from the stakeholder focus group.

Feature collections	Relevance	Time	Score	Acquires creation
Diversity of sheets used in a session	8	8	8	Yes
Which sheets are used	8	8	8	Yes
Number of sessions	7	9	8	Yes
Which combination of sheets are used	9	5	8	Yes
Usage of filters	8	6	7	Yes
Reporting made and repeats	8	6	7	Yes
Exports making and repeats	8	6	7	Yes
Bookmark usage	8	6	7	Yes
Speed of clicks	8	5	6,5	Yes
Function of user in their organisation	7	6	6,5	Yes
Growth/decline of school	7	5	6	Yes
G4/G40/other regions	7	5	6	Yes
Time of logging in	4	8	6	Yes
Session length	7	4	5,5	Yes
Size of school	6	5	5,5	Yes
Selections made per sheet	7	4	5,5	No
Colleagues within the application	6	5	5,5	No
Users in organisation who are data coach	6	5	5,5	No
Mistakes made by user	8	2	5	No
Inspection indicators	5	5	5	No
Support tickets	6	3	4,5	No
School type	6	3	4,5	No
Digital proficiency	7	1	4	No
Consistency of logging in time	3	5	4	No
Gender	3	5	4	No
Age	3	5	4	No
Number of licenses within organisation	3	5	4	No
Authorization restrictions	3	4	3,5	No
Subject given by user	3	3	3	No

Years in organization	3	3	3	No
Salary scale someone is in	3	3	3	No
Usage of quality calendar	3	1	2	No

5. Data preparation

The previous chapter resulted in a list of feature collections that form the input for the model. This chapter discusses how features are created by processing the raw input data.

This chapter starts by discussing transformation techniques that apply to all features. The second section discusses how each feature is created out of the raw data.

The set of features likely includes redundant information with respect to the targeted clustering, i.e. features that express the same or nearly the same characteristic. Redundant features are removed through multicollinearity analysis, which is discussed in the last section of this chapter.

5.1. Transformation techniques

In this section three different techniques are discussed. Transformations and normalization are used on all features that are numerical. Feature encoding is used on all features that are categorical.

5.2. Transformations

Within the dataset not all features make use of the whole range of the feature, since some of the features are highly right skewed. To make the feature less skewed, log-scaling can be used (Gong, 2021), which applies a log function to the data which transforms the data. Higher values are decreased more than lower values due to how the log function works. Because the log function cannot handle 0 values, we added to all values a small non-zero value (0.001). By spreading out the data more, the differences between different datapoints become more clear. This should however only be applied to the features that are skewed.

For some features applying the log function is not able in removing the skewness. Not applying any transformation keeps the data highly right skewed. However applying log transformation makes it highly left skewed. For those, the square root method would be more appropriate (Biostats, 2017). With this a better distribution of the data is achieved.

5.2.1. Normalization

Different features have different absolute ranges for their data. The absolute distance between values is highly dependent on how the feature is measured. Measuring as a percentage gives values ranging from zero to one, while measuring occurrences can give a natural number (0,1,2,…). Two features that are measured with different units cannot be directly clustered togheter. Any type of distance measured is dominated by the feature measured that produces larger relative distances. Features with small relative differences do in those cases not influence the clustering.

In many clustering methods the absolute ranges of variables significant influence the outcome of the model. Normalization scales the values of a feature inside a predetermined range. Min-max scaling (Serafeim Loukas, 2023) can be used which uses the highest and lowest available value to set the range from 0 through 1.

$$x_{scaled} = \frac{x - \min\,(x)}{\max(x) - \min\,(x)}$$

Equation 1 min-max scaling which takes the lowest and highest values to calculate the new scaled value

Although normalization is not strictly necessary when applying clustering algorithms, it is applied whenever a big difference in variance between different features is observed. Normalization is applied to all features even after transformations are applied.

5.2.2. Feature encoding

The dataset contains a small selection of categorical features. Features that are categorical can however not directly be used in the models described. Thus, they need to be encoded before they can be used. User role is not ordinal information, and thus has been encoded in different features using dummy encoding (Pramoditha, 2023). In dummy encoding, every feature gets its binary value except one, which represents in this case the unknown value. Some features might be ordinal, however determining that the different ordinal levels are the same distance apart is not possible. These cannot be assigned a value and need to be also dummy encoded.

5.3. Features creation

This section discusses what needs to be considered when creating features from a raw dataset. They are discussed per feature collection. The order of the discussed features is the same as the priority as given in Table 2 overview of all features collections from the stakeholder focus group.2.

5.3.1. Diversity of sheets used in a session

Diversity of sheets used in a session is measured as the *unique sheets visits within one session*. The number of sheets visited is counted. Multiple visits to the same sheet are not counted. The average over all sessions is taken. The dashboard page is also counted, which is always visited during a session.

5.3.2. Which sheets are used

Which sheets are used is measured per sheet individually. There are 33 different sheets and per sheet the number of visits is counted. The value is averaged over all sessions. Multiple visits to the same sheet within a session are measured separately.

5.3.3. Number of Sessions

Number of sessions counts the number of sessions a user has in total. Number of sessions represents the experience of the user.

5.3.4. Which combination of sheets are used

Which combination of sheets are used is measured by six sets of two features totalling twelve. The most prominent combinations of sheets are used as separate features. Visualized is if someone visits a sheet (horizontal axes) then how likely is it they also visited the other sheet (vertical axes). High chance combinations are used for determining the sets.

The most prominent combinations of features are given here. These combinations are features that are used a lot together. A name thematically is given to each collection of features which represents what relates them. For every combination of sheets to features we created one feature measuring the percentage of sessions any of the sheets are visited and a second feature measuring the percentage of sessions in which all sheets are visited.

5.3.5. Usage of filters

Usage of filters is measured as the number of times a filter is used. Six different filter categories are: School, Teacher, Class, Student, Study and Subject (Course of Study). Multiple usages of a filter within a session are not counted separately. This value is averaged over all sessions of the user.

5.3.6. Reporting made and repeats

Reporting made and repeats is measured by looking at the reports made with the sheet “PDF Rapportage”. On the sheet, eight different reports can be made. For every report, two features are defined.

1. Average number of reports per session takes the total number of reports made, averaged over all the sessions of the user.

2. Unique sessions making the report type measures the percentage of sessions in which a report has been made.

All reports are also measured together by measuring the total number of reports made and the percentage of sessions in which a report is made.

5.3.7. Exports making and repeats

Exports making and repeats measures the number of exports there are made of any table by right clicking and choosing one of the three export options. Because most exports happen on a specific sheet, the use of the export function is measured for that specific sheet, all other sheets and for all sheets. For all three, this is measured in two ways:

1. Average number of exports per session takes the total number of exports made, averaged over all the sessions of the user.

2. Unique sessions making exports measures the percentage of sessions in which an export has been made.

Making reports and exporting are fairly similar actions, since both create documents that can be used outside of the application. Therefore we define a combined feature measuring the total number of exports and reports, and a feature measuring the unique sessions where in a report or export has been made.

5.3.8. Speed of clicks

Speed of clicks is measured as the average time between two logs. Only logs that are a direct result of something the user does are counted. This means that logs of the type document or action are not counted. Those are background actions and are not related to a user's input. When users use the search function every key stroke creates a separate log entries. Multiple search in a row are therefore coalesced into one. Logging off automatically is not counted either since it is not a user's action.

Some sessions are longer than others and contain more logs. We choose to take the average over all times between two user interactions. This is different from taking the average user interaction speed per session.

5.3.9. Function of user in their organization

Function of user in their organization is as the role category someone belongs to. *User role* is created as a categorical feature based on the role as defined by their respective organisation. These are split into nine main user categories who have roles which are similar. The category of teacher is further more split into three parts depending on their pay scale.

5.3.10. Growth/decline of school

Growth/decline of school is calculated by looking if size of the school predicted in the future is larger than the current size.

5.3.11. G4/G40/other regions

G4/G40/other regions references the different classifications of municipalities within the Netherlands. This is the classification of municipalities in the largest cities (G4), major cities (G40) and other municipalities (Centraal Bureau voor de Statistiek, 2023).

5.3.12. Session length

Session length is calculated as the time between the start of the session and logging off. Sometimes a user is however logged in for too long and is automatically logged off. When a user is automatically logged off, the last action is used to represent the end of the session. Session length as originally given in the database does not account for inactivity thus might overrepresent the users time of engagement with the system.

5.3.13. Time of logging in

Time of logging in is split in three parts: time of the day, day of the week, and year.

Time of the day is measures the percentage of sessions that took place during a certain part of the day, following the convention of dividing the day in four parts, night (0:00-6:00), morning (6:00-12:00), afternoon (12:00-18:00) and evening (18:00-24:00). A small number of time periods are chosen to not overcomplicate and create too many separate features while still creating an idea of when someone is online. Time periods are measured in "Amsterdam Time" (CEST), given that the users are from schools in the Netherlands.

Day of the week is the percentage of sessions logged on that day.

Year is measured separately following the two types of year defined in education. Calendar year going from 1 January until 31 December, while a school year goes from 1 August to 31 July.

5.3.14. Size of the school

Size of the school is measured by one feature, namely if the school is small or large. Large is defined as having more than 600 students. The value is determined by TIG.

5.4. Multicollinearity analysis

It is important to check the correlation between the different features. This is done by performing a multicollinearity analysis when preparing the data. It is possible that multiple features are highly correlated, which means that they do not truly represent different information. Correlation can be easily identified using a correlation matrix.

Complete correlation removal is however not the goal, only handling the most extreme cases. The goal is to remove redundant information; in the cases where there is still significant correlation between features, they individually contain unique information and are thus still useful.

The different features need to be checked on collinearity. They are plotted on correlation matrixes. Most correlation is within acceptable ranges, but there are some features that are highly correlated. Highly correlated features need to be addressed in the step.

5.4.1. Almost perfect correlation

There is a small selection of features that are perfectly or almost perfectly correlated. These are features which are created as variations on each other but are not capable of measuring something truly different. In case of a correlation coefficient higher than 0.95 or below -0.95 the features are considered perfectly correlated. In total, only three pairs of features were that extreme. In all three cases, one of the features has been removed from the dataset as it represents both pieces of information.

5.4.2. Highly correlated

After having dealt with the almost perfect correlated features, we made a selection by considering high correlation. Features that have a correlation coefficient of more than 0.75 or less than -0.75 are considered highly correlated.

For some features, there are two implemented variations: one variation measures the average number over all sessions and the other the number of unique sessions. In cases they are highly correlated only the average number over all sessions is kept.

Some features measured the total number of visits while others the unique amount of sessions something was visited. These are sometimes highly correlated. In these cases, we decided to drop the unique sessions variation. This is because those are more complex features to interpret.

Furthermore, the features that measure the weeks that someone is online in the two different ways (per calendar year or school year) are all highly connected. Although it could have been expected, doing the measurements twice via two different features is unnecessary. Therefore, only the features that measure weeks online per school year are kept in the dataset.

The morning sessions and afternoon sessions are highly negatively correlated. This makes sense because both features are measured as the percentage of all sessions that take place either in the morning or afternoon. Knowing the percentage of time someone is online in the morning, evening and night can be used to calculate the afternoon percentage, since this feature is already captured in the other three. Thus, to remove correlation and without knowledge loss afternoon sessions are removed as feature.

Although a next step to investigate correlated features could be possible, we decided to stop at this point. This is because features are expected to be sometimes somewhat correlated. Correlations between different features are also less easily explainable when considering lower correlation coefficients.

6. Modelling

The goal of this chapter is to present the models we created and assess the quality of the models.

The first section of the chapter explains theoretical models that can be applied to the clustering problem of this research. These two modelling methods are K-means and Hierarchical clustering. The discussion includes some additional methods that are necessary when using them.

The second section describes the creation of models by applying the modelling methods. The modelling is performed using categorical classifiers, using the full dataset, and using thematic subsets.

The last section of this chapter discusses the validation of the models. Testing stability results in a final selection, leaving only those models that create valid user groups. The outcome thus are models that can split the users in multiple distinct user groups.

6.1. Clustering models used

The goal of this thesis is to find if multiple different user groups can be identified. To identify those user groups clustering is used. For clustering a variety of different methods can be used. Two clustering models have been chosen: K-Means and Hierarchical. K-Means is chosen because it is the easiest model to interpret. K-Means uses centroids. A centroid is a point that represents the centre of a cluster. Hierarchical clustering is used because more complex clusters could potentially detected based on distances between points and not a single centroid.

There are other types of clustering techniques. Notable types are distribution based and density based models. We chose not to use distribution based model, because they assume that the data approaches a distribution (for example gaussian). Assuming a type of distribution is a hard claim that is not easily made. Density-based clustering depends on areas with high numbers of data points to detect if there is a cluster. Cluster shapes can be more flexible just as Hierarchical and should handle outliers better. Density-based models have the major drawback that they find clusters with varying density difficult to handle. Given the distribution of different features similar density between different clusters is not expected and thus Hierarchical is preferred over Density-based.

6.1.1. K-Means

K-means (MacQueen, 1967) is one of the better known methods of clustering different groups of data points. It is easily programmed and computationally economical for identifying K different sets (MacQueen, 1967). This is achieved by first creating K different points. Next all data points are assigned to the point whose mean is closest in distance calculated, then the centroid for each cluster is calculated. These centroids are finally used as the new points used to assign the data points to a cluster. This is repeated until the centroids do not change anymore. Given that K-means has random elements, running it multiple times can generate different results. The number of clusters created is not determined by the algorithm, but is set by choosing a value for K.

K-means was chosen because it is easily understandable. Clusters are only defined by their distance to a centroid. Other techniques might have clusters that are defined by an irregularly confined area and are more difficult to understand. This strength is also the main weakness of K-means. Clusters with more complex shapes are not always nicely identified by K-means.

6.1.2.Silhouette score

The number of clusters cannot be determined by K-means. An appropriate number of clusters cannot be determined via observation either, because visualization of more than 3-dimensional data is difficult. The silhouette score (Rousseeuw, 1987) however, can be used to determine how many different clusters there might be within the data available. This score represents both the cohesion within a cluster and the separation between clusters. By calculating the silhouette score for a variety of different clusters and observing which one has the highest score the number of clusters can be chosen.

The silhouette score is calculated on K-means result. The score can thus differ slightly due to randomness. If scores are close, the advised value for K can change. The silhouette score can also be used to give a measure to how well the clusters have been created, so it is a score that can be used to validate clusters.

Because the silhouette score is an abstract measure, for interpretation some reference is necessary to better understand its meaning. For this the used interpretation of different scores is based the proposal by Kaufman and Rousseeuw (Kaufman & Rousseeuw, 1990). Their proposed interpretation is shown here in Table 3 . These are rules of thumb and should not be seen as strict measures.

Table 3 Silhouette interpretation table following as by Kaufman and Rousseeuw (Kaufman & Rousseeuw, 1990)

Silhouette score	Proposed Interpretation
0.71-1.00	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial; please try additional methods on this data set
≤ 0.25	No substantial structure has been found

6.2. Hierarchical clustering

Because K-means has some limitations another clustering method has been used in addition. Hierarchical clustering has the advantage over K-means that the clusters created can have more complex shapes. This makes hierarchical clustering more difficult to interpret, but more flexible towards complex clusters.

In hierarchical clustering, every point starts in their own cluster. Stepwise clusters are merged for the two clusters for which merging increases the sum of squared error the least. Clusters are merged until only one cluster is left. Results are then generally visualized within a dendrogram. The merging point where the summed error of squares is the largest is the place used to split the results into multiple different clusters. The type of hierarchical clustering described here is called Ward's minimum variance method (Ward, 1963).

6.3. Model creation

6.3.1. Impact of categorical classifiers

As a first approach K-means is applied to the whole dataset, and the silhouette score is calculated. K=3 gives the highest Silhouette score. The Silhouette score for two K values (K=2,3) are close, which implies that selecting the highest not automatically relates to the best number of cluster when considering the dataset. Due to the random nature of K-means, different silhouette scores can be the highest depending on the starting conditions.

However, since the found silhouette score is only 0.10 it means that no or hardly any clustering can be applied to the full dataset.

In this clustering attempt, categorical features such as user role were still part of the analysis. However, when looking at the statistics behind the clusters it was found that the results have been subdivided by the categorical features, which are heavily influencing the clustering ability. Clustering purely on the binary representation of a small selection of features does not yield useful results.

Therefore, in the analysis hereafter categorical features are only used when comparing the clusters found.

6.3.2. Full dataset without categorical data

Again the same clustering is applied to the dataset, but in this case without any of the categorical features. These are user role, G4/G40/other region, school size and school growth.

Using the clustering again on the subset reveals that there are not really any identifiable clusters with this method. The maximum Silhouette score found is less than 0.08, which is far lower than the references described in Table 3 . This reduction in score compared to the score for the full dataset also shows that any structure using the K-means method on the full dataset was only due to the nature of categorical features.

Applying hierarchical clustering might be able to create better clusters in the dataset. There seems to be four different clusters. There are multiple options for different clusters within the dataset, while either 2 or 4 clusters seem most logical. The choice between two options must be done on insight. Using the slightly higher value suggests two cluster and this will be used.

6.3.3. Thematic subsets

Clusters might only be identified in specific subsets of the features. We created different subsets based on logical themes within the data. This implies that thematic subsets are created on which the analysis is done individually. We discuss the clustering in parallel.

- *Filtering features:* This subset represents how users apply filters. Filter usage is measured by six individual features which measure different types of filtering.
- *Sheet usage features:* There are two sets of features covering information specifically about visiting sheets: *sheet usage* counts how often certain sheets are visited and *sheet combinations* measures how often certain combinations of sheets are used.
- *Time features:* Lots of different features cover when and how much different users are on the platform. These include features that measure: time of logging in; day of the week; weeks per year; and logins per year.
- *Reports and exports features:* Which reports and exports a user makes and how often this is measured by a large selection of features.
- *General features:* Some general features have been identified. These are the features that measure behaviour in different ways more generally. These are *number of sheet visits, session length, number of sessions* and *time between clicks.*

K-means

Firstly, on all the different thematic subsets the K-means clustering is performed. For each one amount of clusters was chosen based on the highest silhouette score.

For the sheet features and time features subsets can we conclude that there are no clusters found by K-means, since the silhouette score is too low for both.

For the other, three clusters are found with scores between 0.25 and 0.50. These indicate that there are potential clusters within the datasets. The found clustering is not certain however. Therefor further evaluation is needed.

Hierarchical

Hierarchical clustering has been also applied. For all of the feature subsets, the same number of clusters are found as with K-means. However, for all the subsets the score has decreased. Validation needs to be performed to find out if they are truly valid clusters.

Table 4 Found clusters for hierarchical clustering for different feature subsets

Feature subset	Found clusters
Filtering features	2
Sheet features	2
Time features	5
Reports and exports features	2
General features	3

6.4. Validation

There are several ways of assessing the validity of different found clusters. When constructing the different clusters with K-means the silhouette score was used. Silhouette score is an internal validity measure. However, because it is used for the construction of the clusters, using it for assessment would not be appropriate.

To validate the clustering, a second independent metric is required. There are other metrics for internal validation using different aspects, but as the clustering algorithm tries to optimize on such metrics, these metrics are not independent and therefore not appropriate for validation (Orlov, 2017). Therefore the stability check based on the Jaccard coefficient has been chosen for validation.

6.4.1. Jaccard coefficient stability

To measure the stability of the clusters, the approach of Mucha (Mucha, 2007) is followed for calculating cluster stability using the Jaccard coefficient. This method was designed originally for hierarchical clustering. Similar methods also exist for K-means clustering (Yu et al., 2019). Therefore, the same measurement of stability is applied to both clustering approaches. To calculate the stability of the clusters, we took the following steps:

- Subsampling the total dataset. This is done by randomly selecting a certain percentage of the datapoints without replacement. Aligned with Mucha (Mucha, 2007), this percentage was set to 75%.
- Perform the same clustering as originally, and split the subsample in the same number of clusters. The number of clusters was kept equal to allow a solid stability analysis.
- Calculate the Jaccard coefficient between the original cluster and the newly created subsample cluster. The calculation of the coefficient is only done using data points within the subsample. Data points outside the subsample are ignored. The Jaccard coefficient measures the percentage of all labels that appear in both clusters.
- For each cluster in each subsample, only the maximum value of Jaccard coefficient was kept.
- Calculate different metrics to give insight into the value.

$$\tau(\varepsilon, f) = \frac{|\varepsilon \cap f|}{|\varepsilon \cup f|}$$

Values from this metric range from 0 to 1. All metrics have been calculated separately for each individual cluster. The number of different subsamples created was set to 1,000.

6.4.2. Full dataset without categorical data

Applying the Jaccard coefficient stability shows if the clustering on the full dataset without categorical data was valid. The Jaccard coefficients found were however below 0.5 which is considered too low. The low scores show that no valid clusters were identified.

6.4.3. Thematic subsets

K-means

When applying K-means to the thematic subsets of the data, three collections of features were found that might contain clusters. For all of them, the stability is calculated. Considering the Jaccard coefficient statistics, all three subsets show that the clusterings are highly stable. Even in the worst case scenario (the minimum) almost all features still have a Jaccard coefficient of at least 0.8. Therefore these found clusters are observed as being stable.

Hierarchical clustering

For the same features, the stability of the hierarchical clustering was also calculated. We compared the Filtering, Reporting and General Features, and these three had far worse results than the K-means clustering approach. For all subsets, the average Jaccard coefficient is significantly lower than for K-means. This clearly shows that this method is less capable of clustering the data. The other statistics confirm this.

For the other two thematic subsets, clustering was also performed. Scores of the hierarchical clustering are low. Therefore no valid clusters were created.

7. Evaluation

This section examines the outcome of the models. It describes the knowledge that can be extracted from the clusters. The discussion evaluates to what extent the models can show TIG if multiple user groups are identifiable.

The modelling resulted in three models. All three models are discussed individually. The discussion highlights differences between the clusters, and interprets why the differences exist.

7.1. Reports and exports

In the reports and exports features subset two different clusters where found with K-means. The main difference between the two clusters is how much reports someone has made. The second cluster contains users that make use of the reporting functionalities. The features show clearly that users in the second cluster use the report and export functions more often.

There is a big difference between the two clusters in size. The first cluster contains about five times as many users as the second cluster. High usage of the reporting functionalities appears to have happened only on a smaller group of users.

In features outside of the cluster also appear to be different. Those who create more exports have in general longer sessions. Export creation is a time costly endeavour.

For some reason, there seems to be a difference in that users who export more, use filters more on subject than other user groups. This increase in filter use is only clearly visible in the subject selection and not in any of the other groups.

In the categorical features, there seemed to be no differences between the two clusters. These user groups seem independent of role or school characteristics. This differs from the expectation that exports are advanced and belong to certain organizational roles. The main difference between these two clusters is that the second group appears to contain users who use the export functionalities more often.

7.2. Filtering

In the clustering on the filtering options, most of the clustering is due to filtering on schools. In study filtering, there is also a difference showing that the users in the second cluster use the cluster option more often. In the other filtering features, this is not visible. Interestingly there is no apparent difference between the number of sessions. This is unexpected, given that the second cluster uses more filters than the first cluster.

About two thirds of the users belong to the first cluster. Of the policy makers, about fifty percent fall in the first cluster. This shows that there is a connection between the user role of policy maker and which selection options they use. Other user roles have less extreme connection to the clusters. Given that thit is the only role with a relationship with clustering, this might be a random result.

7.3. General

The general features selection created in total three different clusters. Table 5 shows the relative differences between the three found clusters. Only is shown if the feature has an average that is low, middle or high compared to the other clusters.

The features with the most prominent differences are the number of sessions and the sheet usage. Session length does show some distinct differences between the clusters but less clearly. Differences in time between clicks are negligible.

Table 5 relative differences between the three clusters from the general clustering

	Cluster 1	Cluster 2	Cluster 3
Number of sessions	High	Low	Low
Sheet usage	Middle	Low	High
Session length	Middle	Low	High
Time between clicks	Middle	Middle	Middle

The first cluster stands out, because only the first cluster contains a high number of sessions. For the other cluster, only users with a limited number of sessions are contained. The first cluster seems to contain those which use the dashboard on a regular basis.

Both the second and third cluster contain users that do not use the dashboard a lot. The main distinction between the two clusters comes from how much they use the dashboard. The second cluster contains those who use the dashboard not much and only visit a small part. This in contrast to the third cluster that does not use the dashboard a lot but visits many different sheets.

The difference between the clusters is also visible in other metrics. Usage of different selection features is connected to the number of sheet usage. Therefore these metrics also increase, given that people who have longer sessions are probably more likely to do more filter options. The same applies to the sheet usages.

7.4. Comparing results

Clustering results can be compared by looking at agreement between how users are clustered. The different clusters created based on different feature subsets are not the same. However, clusterings with different feature subsets might not create fully independent results. It is looked at if the different clustering results are fully independent or not.

Independency is calculated using the conditional probability of belonging to a cluster given other clustering results. If the conditional probability is significantly different from the normal probability of belonging to a cluster then the results are not truly independent.

Comparing the clustering results of report and exports and filtering there are no clusterings that are dependent. Belonging to one cluster from clustering on report and exports does not relate to belonging to a cluster within filtering. This shows that these are two independent ways of clustering users.

When looking at filtering features and general features, it is clearly visible that user belong to one cluster or another. Almost all conditional probabilities are significantly different from random and thus the results of the two clustering methods should not be seen as independent from each other. Some large differences between the observed users in a cluster combination and the number that would be expected.

The same is observed also between report and exports features and the general features. The two clusterings are not fully independent. Especially knowing that someone belongs to a specific report and export cluster is a good indicator of to which general cluster someone belongs to.

Although some clusterings are not fully independent there result to predict the other results is limited. For example, in one of the clusterings about 24% are within one cluster. Knowing other clustering results the conditional probability increases to 40%. This is indeed a significant difference, but one result can not be fully predicted from another one. This example was the case with the highest absolute difference between the independent and conditional probability. Although not all results are independent all three clusterings do contain unique information.

8. Discussion

This section discusses the results of this research. It discusses the three results and what could be learned from it. Next it reflects on the design of the research and discusses the different limiting factors that may influence the quality of the results.

8.1. Analysis results

In total the research found three different ways that can be used to create user groups. All three ways of creating user groups are valid independently from each other. Different insights can be gathered from each and every one individually.

Firstly, clustering on the reports and exports functionalities created two distinct clusters. One cluster with high report and export usage and one with low usage. Those with low usage form a far larger group than those that have high usage. Given that the functionality only belongs to a small group of users creates the impression that it is mainly an expert functionality. This finding may lead to two alternatives approaches related to this functionality. The functionality may be considered for expert users only, and may be moved to a location that fits this description. The other option would be to promote it to a broad user group. Possibly the function is too much hidden at the moment and changes could be made to inform more users of the existence.

Secondly, filtering creates nicely two groups that differ mainly on the usage of the school filter. With respect to the amount of usage of the filter functionality, only for policy makers some relationship was visible. Because filtering can be seen as a zooming-in action on a specific area, more relationships to specific user roles might be expected. As no other relationship was found than the one mentioned, the user role might only have limited relationship to whether someone is interested in only a specific school or not. User roles as they are defined might also be limited in their accuracy of representing the current tasks of the user.

Thirdly, the clear split in three clusterings creates clear divides in the amount of usage and to what extend users use of the dashboard. Each of the three subdivisions are a clearly different group that should be approached in a distinct way.

The users that have high dashboard usage are very likely to includes the super users, and include current users. These are users that are probably satisfied with the dashboard in its current form. Those with whom TIG has direct contact with are also probably within this group. If TIG is considering contacting users that can act as motivators towards other users than looking throughout these users could be a good first step.

The is a group of users that have a low usage of the dashboard and only use part of the dashboard. These users may have not been fully introduced to most of the dashboard or do not see use to most of its functionality. Not looking through the different functionalities by themselves means they might be difficult to motivate. Investigating the reasons behind such usage and looking how this group can be motivate to use the dashboard more broadly might be a way to move these users towards regular users.

There are also users with low usage, but engaging with large parts of the dashboard. These users have great curiosity for the application, but are not regularly using it. They might only need it for small periods of time, or want to use it but don’t know how. This group might not need to be motivated, but needs assistance in understanding how to integrate the application within their normal tasks.

For all low usages users it is important to see if the user is still a customer or not. These users might also be customers that have gone, or who only had temporary accounts. Information about this was not available.

8.2. Research design

Different choices needed to be made on the how the research was performed that may influence the quality of the research.

The research focussed on all the users and did not focus on smaller groups of users. This means that most lessons learned look only at the whole population and not certain specific sub groups. There was no separate clustering performed on teachers or any other user role. As the interest was in the population as a whole, gaining more specific knowledge in specific groups is still difficult.

In total two types of clustering algorithms have been used. Both do complement each other nicely, but extending the research to other clustering techniques might produce other user groups. Without applying other algorithms it is probably impossible to determine if other clusters exist. Changing the transformations and normalisations might also lead to finding other user groups. Even if applied found user groups might be the same or almost the same. Some choice needs to be made how many different techniques are applied and given the two used techniques were most appropriate the research was limited to only the two.

Transformations and normalisations were necessary for creating distributions that are fit for clustering. Changes made to the data should always keep the key characteristics of the original data. The applied techniques within the research do change some of the aspects of the data, but such change is regarded not to have changed their key characteristics. Without applying transformation and normalisation, finding any user groups would not have been possible. All data would have been too much grouped together to find any differences between user groups. The techniques used do keep the most important characteristics of the original data intact.

For the selection of the users the choice was made to exclude users with less than five sessions. This creates the problem that some group of users is not analysed within the cluster analysis. It was deemed necessary for creating results, but these users must not be forgotten. This group might actually be of very much interest to focus on, as these users might need to be motivated to use the dashboard more frequent. Determining how they use the dashboard can be done on a group basis, but given the limited data not on individual level. If TIG wants to understand them they would need to extend their sources with extra information outside of the log data.

The method used in this research, CRISP-DM has a cyclical design. The results of the research can be seen as the end point of a first cycle and have created new insights. Results should also be seen as only a first cycle and should be the input for the next one.

8.3. Limiting factors

Via the stakeholder focus group session different feature collections were extracted. The stakeholders being TIG employees, means the result incorporates the knowledge of TIG. Though being important for the research to utilize this knowledge it also encompasses the risk of this research being biased to the known needs of TIG, whereas answers to unknown needs remain undiscovered. Upfront, the choice was made to incorporate this knowledge, which thus may have limited the capacity of finding new insights, but made sure that the results are useful for TIG.

The results of the dashboard are now interpreted as reflections of the dashboard and the users as if both have always been like this. This is not the case as there have been many changes to the dashboard and in society in the period of 2020-2023, which is the time period included in the data. The problem of a changing application cannot be removed as it is an active product that is updated regularly. Observing over a shorter time period as countermeasure does not work as this would limit the amount of data too much and would have a negative impact on the quality of the analysis. As some parts of the application could be viewed as being static over longer periods of time, reducing the research to such a smaller scope may counteract change and improve the learning potential for analysis on such focused area.

Throughout the research some different types of data quality problems where identified and circumvented by removing data. By definition, this decreased the quantity of the data. A potential side effect is that the removal is a selection of a source. This selection might have created a bias towards certain types of users, but it is not known what kind of users. It is assumed that it does not affect the found results, but there exists some bias due to this data exclusion.

9. Conclusions

Answered in this section is the original research question:

Which user groups can be differentiated considering characteristics of their behaviour in an educational performance dashboard based on log-data?

Different features were extracted from stakeholders that are related to certain behavioural characteristics. These were directly based on log-data from the educational performance dashboard. Multiple different user groups were identified, but the found user groups depended on the chosen selection of features. There were clear differences between user groups and the clustering developed useful insight for the further development of the platform.

9.1. Learned lessons

Many different new insights were developed during the research of which the most important ones were:

- Sub selections of features can contain clusters even if in the total set no clusters could be found.
- If a cluster could be found depends on the clustering method selected, not all are equally successful in the same situation.
- For a detailed look into specific areas of use, smaller specific groups should be selected.
- The export and report functionality is used by a small group of users and is thus possibly an expert functionality.
- Behaviour such as filtering on school is not directly related to the type of school someone belongs to.
- The users with low usage outnumber those with high usage.
- There are clear differences in behaviour even between those with limited usage.

9.2. Future work

The research was able to answer its main research question. However there are different improvements and further research that could be performed.

- Applying other clustering methods that are able to use categorical data. Categorical data had to be excluded, but might contain still hidden information.
- Outlier filtering to create clearer clusters that better represent a user group.
- Applying UX expertise to look into how the results could be transformed into UX changes.
- Looking into if the export and report functionality need to be changed to apply to a broader user base.
- Research into why a certain group of users have low usage and how their application usage can be increased.
- Comparing the user that TIG thinks are super users with the created user groups.
- Investigate a broader variety of feature sub selections and other features.
- Combining results with and comparing it with qualitative results from interviews.

References

AGB. (2014). *Overseeing Educational Quality: A How-to Guide for Boards of Universities and Colleges*. Association of Governing Boards of Universities and Colleges. http://ezproxy2.utwente.nl/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED549820&site=ehost-live http://agb.org/sites/agb.org/files/OverseeingEducationalQuality.pdf

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European Conf. Data Mining,

Biostats, A. (2017). *Transforming skewed data: How to choose te right transformation for your distribution, biostatistics and bioinformatics anatomise biostats*. Retrieved 30-05-2023 from https://anatomisebiostats.com/biostatistics-blog/transforming-skewed-data/

Browne, G. J., & Rogich, M. B. (2001). An Empirical Investigation of User Requirements Elicitation: Comparing the Effectiveness of Prompting Techniques. *Journal of Management Information Systems*, *17*(4), 223-249. https://doi.org/10.1080/07421222.2001.11045665

Centraal Bureau voor de Statistiek. (2023). *Gemeentegrootte en stedelijkheid*. Retrieved 27-06-2023 from

Chalvatza, F., Karkalas, S., & Mavrikis, M. (2019). Communicating learning analytics: Stakeholder participation and early stage requirement analysis. CSEDU 2019-Proceedings of the 11th International Conference on Computer Supported Education,

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T. P., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Corporation, M. (2023). *Download SQL Server Management Studio (SSMS)*. In https://learn.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver16

Craig Stedman, J. V. (2022). *What is data quality and why is it important?* TechTarget. Retrieved 22-06-2023 from https://www.techtarget.com/searchdatamanagement/definition/data-quality

Denwattana, N., & Saengsai, A. (2016, 14-17 Dec. 2016). A framework of Thailand higher education dashboard system. 2016 International Computer Science and Engineering Conference (ICSEC),

Dickman, A., Allen, V., & Henken, R. (2011). *Measuring up Education: Community-Driven Accountability in Milwaukee*. http://ezproxy2.utwente.nl/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED524214&site=ehost-live

Foundation, P. S. *Python Language Reference, version 3.11.* In http://www.python.org

Gong, D. (2021). *3 common techniques for data transformation Medium*. Retrieved 30-05-2023 from https://towardsdatascience.com/data-transformation-and-feature-engineering-e3c7dfbb4899

Gottesdiener, E. (2002). *Requirements by Collaboration: Workshops for Defining Needs.* Addison-Wesley Professional.

Harris, C. R. a. M., K. Jarrod and van der Walt, Stéfan J and Gommers, Ralf and Virtanen, Pauli and Cournapeau, David and Wieser, Eric and Taylor, Julian and Berg, Sebastian and Smith, Nathaniel J. and Kern, Robert and Picus, Matti and Hoyer, Stephan and van Kerkwijk, Marten H. and Brett, Matthew and Haldane, Allan and Fernández del Río, Jaime and Wiebe, Mark and Peterson, Pearu and Gérard-Marchant, Pierre and

Sheppard, Kevin and Reddy, Tyler and Weckesser, Warren and Abbasi, Hameer and Gohlke, Christoph and Oliphant, Travis E. (2020). *Array programming with NumPy*. In

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, *9*, 90-95. (IEEE)

Iriberri, A., & Stengel, D. N. (2021). Closing the Loop: Development of a Dashboard for Quality Improvement of Business Education Programs. *International Journal for Business Education*(161), 23-38. http://ezproxy2.utwente.nl/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1312339&site=ehost-live

Isaias, P., & Backx Noronha Viana, A. (2020). On the Design of a Teachers' Dashboard: Requirements and Insights. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 12205 LNCS, pp. 255-269).

K. T. Hanna. (2021). *What is a user group?* WhatIs.Com. Retrieved 19-06-2023 from https://www.techtarget.com/whatis/definition/user group

Kaufman, L., & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction To Cluster Analysis*. https://doi.org/10.2307/2532178

Kleehammer, M. *Pyodbc 4.0*. In https://github.com/mkleehammer/pyodbc/wiki

Konaté, J., Sahraoui, A. E. K., & Kolfschoten, G. L. (2014). Collaborative Requirements Elicitation: A Process-Centred Approach. *Group Decision and Negotiation*, *23*(4), 847-877. https://doi.org/10.1007/s10726-013-9350-x

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations.

Magister. (2023). *Magister*. Retrieved 24/05/2023 from https://www.magister.nl/

Manwaring, E., Carter, J. N., & Maynard, K. (2017). Redesigning educational dashboards for shifting user contexts. SIGDOC 2017- 35th ACM International Conference on the Design of Communication,

McKinney, W. a. o. (2010). *Data structures for statistical computing in python* (Vol. 445).

Ministerie van Algemene Zaken. (2023). *Verplichte eindtoets basisonderwijs*. Retrieved 24/05/2023 from https://www.rijksoverheid.nl/onderwerpen/schooladvies-en-eindtoets-basisschool/verplichte-eindtoets-basisonderwijs

Ministerie van Onderwijs, C. e. W. *Indicatoren*. Retrieved 24/05/2023 from https://www.onderwijsinspectie.nl/onderwerpen/onderwijsresultatenmodel-vo/indicatoren

Mucha, H.-J. (2007, 2007//). On Validation of Hierarchical Clustering. Advances in Data Analysis, Berlin, Heidelberg.

Muntean, M., Sabau, G., Bologa, A.-R., Surcel, T., & Florea, A. (2010). Performance Dashboards for Universities.

Onderwijsbond, A. A. *Salaris*. Retrieved 26/05/2023 from https://www.aob.nl/starters/salaris/

Orlov, K. (2017). *How to select A clustering method? how to validate a cluster solution (to warrant the method choice)? Cross Validated.* Retrieved 31-05-2023 from https://stats.stackexchange.com/questions/195456/how-to-select-a-clustering-method-how-to-validate-a-cluster-solution-to-warran

par. 9 art. 2.88 Wet voortgezet onderwijs 2020. https://wetten.overheid.nl/jci1.3:c:BWBR0044212&hoofdstuk=2&paragraaf=9&artikel=2.88&z=2022-08-01&g=2022-08-01

Pedregosa, F. a. V., Gaël and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent and others. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*, 2825-2830.

Polikoff, M. S., Korn, S., & McFall, R. (2018). *In Need of Improvement? Assessing the California Dashboard after One Year. Technical Report. Getting Down to Facts II*. http://ezproxy2.utwente.nl/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED594705&site=ehost-live

Pramoditha, R. (2023). *Encoding categorical variables: One-hot vs dummy encoding, Medium*. Retrieved 28-05-2023 from https://towardsdatascience.com/encoding-categorical-variables-one-hot-vs-dummy-encoding-6d5b9c46e2db

QlikSense. In. QlikTech International AB. https://www.qlik.com/us/products/qlik-sense

QlikView. In. QlikTech International AB. https://www.qlik.com/us/products/qlikview

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7

Schellekens, L. H., van der Schaaf, M. F., van der Vleuten, C. P. M., Prins, F. J., Wools, S., & Bok, H. G. J. (2022). Developing a digital application for quality assurance of assessment programmes in higher education [Article]. *Quality Assurance in Education*. https://doi.org/10.1108/QAE-03-2022-0066

Schwendimann, B. A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., Gillet, D., & Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research [Review]. *IEEE Transactions on Learning Technologies*, *10*(1), 30-41, Article 7542151. https://doi.org/10.1109/TLT.2016.2599522

Serafeim Loukas, P. (2023). *Everything you need to know about min-max normalization in Python, Medium*. Retrieved 28-05-2023 from https://towardsdatascience.com/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79

Smit, M. (2020). *Scholen boos om prijsverhoging onlinesysteem Magister*. Retrieved 24/05/2023 from https://www.rtlnieuws.nl/economie/bedrijven/artikel/5157796/scholen-boos-om-prijsverhoging-onlinesysteem-magister

Virtanen, P. a. G., Ralf and Oliphant, Travis E. and Haberland, Matt and Reddy, Tyler and Cournapeau, David and Burovski, Evgeni and Peterson, Pearu and Weckesser, Warren and Bright, Jonathan and van der Walt, Stéfan J. and Brett, Matthew and Wilson, Joshua and Millman, K. Jarrod and Mayorov, Nikolay and Nelson, Andrew R. J. and Jones, Eric and Kern, Robert and Larson, Eric and Carey, C J and Polat, İlhan and Feng, Yu and Moore, Eric W. and VanderPlas, Jake and Laxalde, Denis and Perktold, Josef and Cimrman, Robert and Henriksen, Ian and Quintero, E. A. and Harris, Charles R. and Archibald, Anne M. and Ribeiro, Antônio H. and Pedregosa, Fabian and van Mulbregt, Paul and SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261-272. https://doi.org/10.1038/s41592-019-0686-2

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236-244. https://doi.org/10.1080/01621459.1963.10500845

Wyne, M. F., Reeves, J., Montes, F. X., & Gurbach, T. J. (2015). Business intelligence dashboard for academic program management. ASEE Annual Conference and Exposition, Conference Proceedings,

Yu, H., Chapman, B., Di Florio, A., Eischen, E., Gotz, D., Jacob, M., & Blair, R. H. (2019). Bootstrapping estimates of stability for clusters, observations and model selection. *Computational Statistics*, *34*(1), 349-372. https://doi.org/10.1007/s00180-018-0830-y