Bachelor Degree Project

Measurement of interactive manual effectiveness

- How do we know if our manuals are effective?

Author: Henrik Ståhlberg
Supervisor: Johan Hagelbäck
Semester. VT 2018
Subject: Computer Science

Abstract

Multimedia learning is today a part of everyday life. Learning from digital sources on the internet is probably more common than printed material. The goal of this project is to determine if measuring user interaction in a interactive manual can be of use to evaluate the effectiveness of the manual. Since feedback of multimedia learning materials is costly to achieve in face-to-face interaction, automatic feedback data might be useful for evaluating and improving the quality of multimedia learning materials.

In this project an interactive manual was developed for a real-world report generating application. The manual was then tested on 21 test users. Using the k-nearest neighbour machine learning algorithm the results shows that time taken on each step and the number of views on each step did not provide for good evaluation of the manual. Number of faults done by the user was good at predicting if the user would abort the manual and in combination with the number of acceptable interactions the usability data did provide for a better classification then ZeroR classification. The conclusions can be questioned by the small dataset used in this project.

Keywords: multimedia learning, user behaviour, effectiveness measurement

Preface

With a background as a upper secondary teacher one major strategy of improving the quality of the lessons is to continuously evaluate what the students learn. This is in research one part of what is called formative assessment. My idea was that these principles may be applied in multimedia learning but in a more automatic way where continuous measurement can be done automatically. The results show that there are some indications that this might be the case. I want to thank Lars Wendelstam and Johan Wendelstam at Meridix System AB for providing me with a real-world application on which I could build my manual and for helping me in completing this project. I would also like to thank Johan Hagelbäck at Linnaeus University for introducing me to the world of machine learning and for support during the project.

Contents

1 Introduction

As the internet have expanded the capability to share information, more and more digital learning material have been created for different purposes, varying from written descriptions, how-to and video tutorials. This paper tries to provide a method for evaluating the effectiveness of specific material. It is done by developing an interactive manual for a report generator and then evaluating the manual by gathering data about how the users interact with the manual. Analysis of this data should provide answers of which parts of the interactive manual that could be improved to facilitate better user experience and an understanding of the application.

1.1 Background

In this paper a multimedia learning artefact is any digital material used for learning presented through a computer or television. It includes; video tutorials, interactive learning software (for example code academy[1]) and instructional animations. The interactive manual used in this paper is one type of multimedia learning artefact that possesses some certain characteristics. It is a step-by-step instruction where the user has to actively progress through the manual by either executing some action in the application or follow the instruction in the manual to get to the next step. It is what in multimedia learning literature is known as "student paced" even if the users are not students and an adaptation of the segmentation principle [1, Sec. 9]. The manual part means that its content is concentrated on "how to" use a certain product, in this case a web application. The focus is on "how" rather than "why", which might be the focus of other multimedia learning artefacts.

Some well-known principles from multimedia learning have been used when implementing the manual in this project. It is mostly based on the works of Richard E. Mayer. Three assumptions about multimedia learning are used as a default position to construct an efficient manual. They are; dual-channel, limited-capacity and active-processing [1, Sec. 3].

Dual-channel states that there are two channels that humans receive and process information; audio and visual. But there is also a limitation in cognitive processing of for instance words. This means that it is possible to look at a picture and listen to someone speak at the same time. However, it is

[1] Code academy, www.codecademy.com

not possible to read one text and listen to another at the same time since they are both processed by the same cognitive processing unit. The limited-capacity assumption states that there is a limit to how much information a human can receive and process at the same time. It is related to the cognitive-overload situation where too much information is presented for the learner to process and understand. Active-processing states that is not sufficient to just be in front of pictures or having spoken text around you to learn. The learner must actively process the information presented to actually achieve learning. Several principles are presented and tested by Mayer that should be included in multimedia learning material [1, Sec. 3].

The idea behind this project arose from formative assessment principles in pedagogy. It has a wide meaning but in the context of this paper it means tools that teachers use to assess the effectiveness of their teaching. The general idea is that the teacher should often assess what students have learned with the intention to improve their teaching materials and practices. Focusing on continuously improving the effectiveness of their work by getting proper feedback about how well the content was understood. For a teacher this includes practices such as exit notes, continuous testing and live short survey questions. In this project that principle have been tried on a multimedia interactive manual.

The general idea is that an interactive manual should have the ability to gather feedback from the users about the use of the manual. This can be done automatically, and the data can later be used to improve the quality. General surveys are often used at the end of a manual or learning material. The goal here is to be more precise and identifying weaker parts of the manual, not just gathering a general evaluation. For a teacher it means to not do one big test at the end of a course to evaluate what the students learned but to try to measure the outcome of each lesson. How can this be implemented in a multimedia learning environment?

Every manual has a goal, often a content that the user is supposed to have learned after reading the manual. One manual is more effective than another if it manages to fulfil a larger part of that goal to more people. Meaning the average goal fulfilment is higher. Another metric would be efficiency meaning how much time and effort is needed from the user to fulfil the goal. In this paper a method for evaluation effectiveness is in focus. Efficiency evaluation might be a by-product of this method but will not be measured.

If an interactive manual consists of two parts, A and B, users will spend a different amount of time and act differently when processing the both parts. The difference in use of the manual are in this paper referred to as “usability data” and means data about how the users interact with the manual. For a complete list of the data tested in this paper see the method chapter 2.2.3.

Machine learning is the method of applying computer algorithms in order to identify patterns in data. One common use is to classify data into categories based on attributes. If the algorithms can classify data correctly based on the attributes it means that the attributes may determine the category.

The k-nearest neighbour algorithm uses a certain number of neighbours to classify the instance. If the neighbours are of a certain category the instance is classified to also be in that category. It uses a distance function to identify the neighbours and calculates the distance for each attribute. It assumes that instances of the same category have attribute values close to one another and as such the attribute is a determiner of the category.

1.2 Related work

It is a well-established result from previous research that prior knowledge is one of the most important factors for learning [2] [3, pp. 41-42]. The cognitive load theory states that there is a limit to the amount of information humans can process and retain during learning [1, Sec. 3]. Several multimedia learning principles stated by Richard E. Mayer is built upon this theory and they show empirical validity [1, Sec. 14]. For example, the segmentation principle states that the learning material should be segmented into smaller parts and that the student should set the pace giving each student enough time to process the information independent of previous knowledge [1, Sec. 9]. For a manual given out by a company the previous knowledge of the users may differ and the manual needs to facilitate all users’ needs to effectively accomplish its goal.

The segmentation principle is also supported by an experiment with eye tracking by Nakayaman and Shimizu [4]. Their experiment shows that previous knowledge affects searching time since searching becomes slower for inexperienced learners. The mental workload seems to be higher for searching then viewing and to facilitate learning considerations should be taken to reduce searching to increase effectiveness [4]. The experiment with eye tracking is an attempt to evaluate learning material. They state that “The issue of system usability is often considered regarding various other processes, but learning materials should also be evaluated” showing that there is a lack of research about how to improve usability in multimedia learning environments. [4]

Further support of the segmentation principle can be found in research related to video game training. The results show that part-task training where participants train specific game parts separately before playing the game to a large degree reduces the post ability difference between the participants. Low-ability users with part-task training performed at the same level as high-ability users without the part-task training. [8] These findings support the theory that segmentation reduces the effect that previous knowledge has on learning.

A lot of research have been done on finding general principles for multimedia learning that of course should be used if applicable. For instance, findings suggest that animations might not provide any learning advantage compared to static information in understanding of complex computer concepts [2]. The risk of only using results from multimedia learning is that context is not taken into consideration. Maybe the principle does not provide cost justified results in this situation and therefore a less expensive method should be used [2].

Not much research has been done on how to systematically evaluate a specific systems effectiveness and identifying weak parts of the learning material. As MOOC courses became popular the problem of high dropout rates was investigated [5]. As a model to predict dropout was developed focusing on student backgrounds such as gender, age and education background. Although the online environment enables evaluation of students’ performance in objective and quantitative ways the focus is on the student not the course. The result shows that participating in the course forum as well as having friends that pass the course is indicators for high performance. [5]

In one paper with the goal of identifying MOOC creation patterns one of the advantages with MOOC is stated as using “Big data to improve teaching” but other than that leaving the questions unsatisfyingly unanswered about what data and how to use it. [6]

The interests in MOOC courses have provided some research where data is used to track student activities. “Understanding how students interact with MOOCs is a crucial issue because it affects how we evaluate their efficacy and how we design future online courses.” [7]. The focus is on the big picture and about how to design courses in relation to assignments and video and not about how to identify poor video lectures or evaluating quality in certain parts of the course material. [7]

Comparing different learning material have been done as a whole. For instance, in the US army game tips and computer-based tutorials where tested for how efficiently the participants learned a computer game used for army training. They find that a combination was most beneficial and that the two techniques trained different skills. It did not however provide any answers on how to evaluate one tutorial over another. [9]

Very little research has been done about models or frameworks for designing multimedia learning material that is measurable and evaluable from an effectiveness perspective. Since there is a lot of research about effective principles for increasing learning in a multimedia environment it is used in the creation of multimedia learning material. The danger of using principles without custom evaluation is apparent since there might be other contextual factors not included in the analysis and if no evaluation is made these factors might never be identified and addressed.

1.3 Problem formulation

We have good understanding and scientific validity for principles to use in designing multimedia learning materials. We also know how to evaluate a manual or material as a whole. What we lack are good methods for identifying differences in effectiveness of different parts of a manual. This paper tests the method of gathering data from the use of an interactive manual in order to identify weaker parts in a manual. It is also unknown what type of usability data that can be used to evaluate effectiveness of different parts and this paper aims to provide insights about this method. The hypothesis is that gathering of usability data provides useful information for evaluating a manuals effectiveness in its parts and as a whole.

1.4 Motivation

Although research in multimedia learning has been done for more than four decades some of the results are rarely introduced into online learning materials. A lot of material have been created but the evaluation of these artefacts is not fully examined. Any company presenting their customer with a manual or instruction about how to use their product should be interested to evaluate if the manual or instruction fulfil its goal. With paperback manuals there is no feedback information other than customer support. With interactive multimedia manuals there is a possibility to measure how users interact with the manual. Maybe data gathered from such user interaction could be used to identify parts of the manual that is confusing or in other ways does not enhance understanding for the users? If this is the case, construction of interactive manuals should be done in such a way as that data is gathered and later analysed. With an increased number of applications that people use in their life as well as work and with frequent updates including changes of these applications, effective ways to learn these applications are needed. Implementing a new application within an organisation causes costs in the form of time for learning how to use the application, mental effort in having to learn a new application and the risk of problems arising from wrong use of the application. All these costs can be mitigated by effective manuals and instructions.

1.5 Objectives

O1	Develop an interactive manual for one part of Meridix Systems AB report generating application
O2	Gather user data from their use of the manual
O3	Analyse and evaluate the user data
O4	Present a conclusion about the manual effectiveness

The result would hopefully show that it is possible to evaluate the manual effectiveness. Identifying weaker parts of the manual would be an expected outcome. It might identify different measurement data that could be gathered in the future or identify types of data that is not useful in determining effectiveness. Gathering ineffective data is also an interesting result as figuring out what not to do might be just as important as figuring out what to do.

1.6 Scope/Limitation

Since only one manual will be created for a specific type of program, different users and applications might generate varied results. Although a lot of applications are similar to the Meridix Systems AB web application the results will not be generalized to other types of learning material. The evaluated artefact is a step-by-step interactive manual guiding the user through the application and cannot provide answers how to measure other types of learning artefacts such as video tutorials. Also, the manual is quite specific in the sense that it is teaching the user how to generate a report. More complicated learning content where models and concerns about why, might require other types of feedback data in order to evaluate the material.

1.7 Target group

Any company interested in increasing value of their product by developing more effective ways to instruct their customers about how to effectively use their product. Especially companies that sell applications should be interested in the results and implementation of this project.

1.8 Outline

This report is outlined in the following manner. In chapter 2 the method is described where selection, data measurements and variables are discussed. The chapter also describes some limitations to the method and the validity of the results. In chapter 3 the interactive manual software produced in this project, as well as the technical tools used and developed for the project are portrayed. Chapter 4 presents the user data gathered when the interactive manual is used. In chapter 5 the results have been analysed and what kind of conclusions that is possible to gather from the results are presented. The discussion about the findings of this paper are discussed in chapter 6 and in chapter 7 the conclusions are presented as well as suggestions for future studies.

2 Method

2.1 Method

This paper will use data gathering and analysis from a custom-made manual designed for this purpose. It uses a survey as well as automatic gathering through user interaction as ways to gather the data. The dependent variable will be the general assessment of the manual by the users at the end of the manual. This is assumed to be affected by the data results gathered during user interaction in the manual.

These data results are the independent variables and are gathered as usability data during the users use of the manual. If a certain type of usability data provides a prediction of the outcome of the general assessment it means that it is a good tool for identifying issues in the manual. That data can then identify which parts of the manual that caused poor user understanding in this context meaning has low effectiveness.

2.1.1 The manual

The manual is implemented as an overlay on Meridix report system and consists of 35 steps. Each step contains some information about a specific part of the manual and may contain instructions to the user. The user is only allowed to interact with the parts of the manual acceptable in the given step. The first and last step are general greetings and the second last step contains the dependent variable general assessment question.

2.1.2 Selection

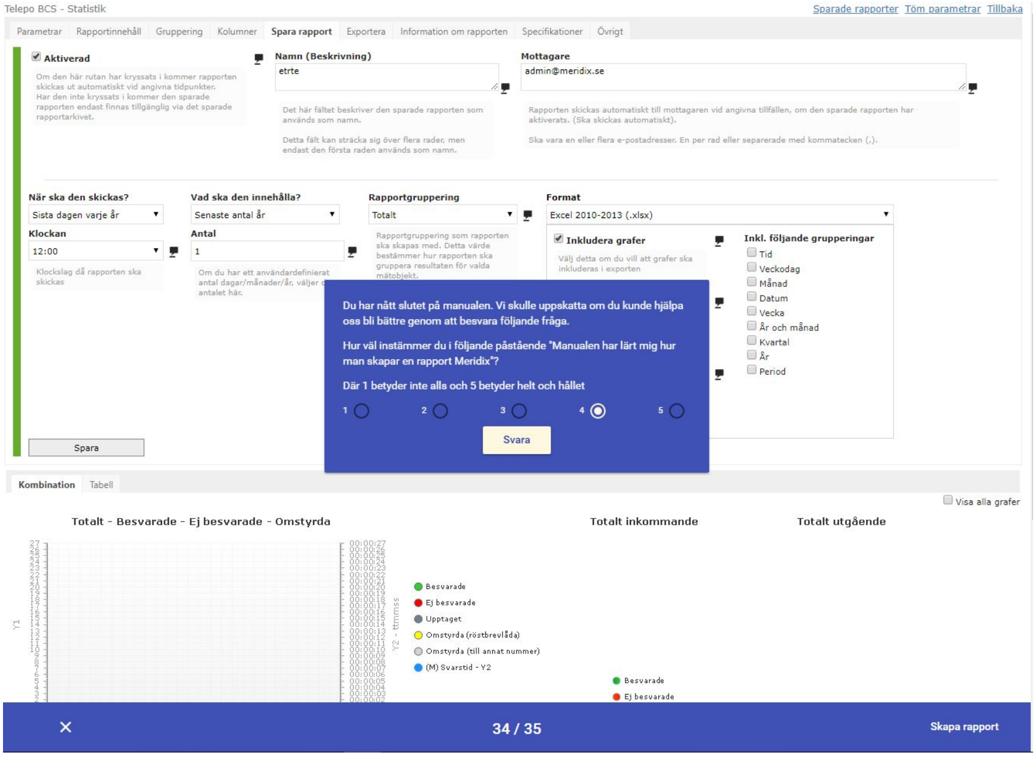
The users in this study will be test users from a mix of individuals unfamiliar to Meridix System. There are 45 test users in total that goes through the manual set up with mock statistics in order to gather usability data. Although it would have been desirable to use real novice users for testing the system the time frame did not allow it since Meridix do not have that many new customers during the weeks of testing to provide reliable results.

All participants receive the same instruction to simplify their participation. They receive a small instruction on what to do and what the manual is about. The instruction was deliberate vague and unclear to give a closer test to reality. A real user will not use the system from a clear instruction but rather try to figure out if and how it’s possible to achieve what they want. See appendix A for the instructions.

2.2 Data

2.2.1 Dependent variable - general assessment data

To answer the question if the manual was effective a survey will be used presented at the end of the manual. The survey will consist of one question “To what degree do you agree with the following statement ‘*The manual have taught me how to create a report in Meridix*’? Where 1 means not at all and 5 means completely.” The answer will be on a scale from 1 to 5 providing a baseline for the evaluation of the manual. These results are here called general assessment data. Below is a screenshot showing the question in the manual.



Telepo BCS - Statistik

☑ Aktiverad

Om den här rutan har kryssats i kommer rapporten skickas ut automatiskt vid angivna tidpunkter. Har den inte kryssats i kommer den sparade rapporten endast finnas tillgänglig via det sparade rapportarkivet.

Namn (Beskrivning): etrte

Det här fältet beskriver den sparade rapporten som används som namn.

Detta fält kan sträcka sig över flera rader, men endast den första raden används som namn.

Mottagare: admin@meridix.se

Rapporten skickas automatiskt till mottagaren vid angivna tillfällen, om den sparade rapporten har aktiverats. (Ska skickas automatiskt).

Ska vara en eller flera e-postadresser. En per rad eller separerade med kommatecken (,).

När ska den skickas?: Sista dagen varje år

Klockan: 12:00

Klockslag då rapporten ska skickas

Vad ska den innehålla?: Senaste antal år

Antal: 1

Om du har ett användardefinierat antal dagar/månader/år, väljer d antalet här.

Rapportgruppering: Totalt

Rapportgruppering som rapporten ska skapas med. Detta värde bestämmer hur rapporten ska gruppera resultaten för valda mätobjekt.

Format: Excel 2010-2013 (.xlsx)

☑ Inkludera grafer

Välj detta om du vill att grafer ska inkluderas i exporten

Inkl. följande grupperingar

☐ Tid
☐ Veckodag
☐ Månad
☐ Datum
☐ Vecka
☐ År och månad
☐ Kvartal
☐ År
☐ Period

Du har nått slutet på manualen. Vi skulle uppskatta om du kunde hjälpa oss bli bättre genom att besvara följande fråga.

Hur väl instämmer du i följande påstående "Manualen har lärt mig hur man skapar en rapport Meridix"?

Där 1 betyder inte alls och 5 betyder helt och hållet

1 ○ 2 ○ 3 ○ 4 ◉ 5 ○

☐ Visa alla grafer

34 / 35

Figure 2.1 The general assessment question in the manual

Another secondary dependent variable is users that do not finish the manual and do not provide an answer to the general assessment. They can be identified as well as the step where the manual was terminated.

2.2.2 Independent variable - usability data

The usability data will consist of four types of data; time taken of first view of the step, number of times that the step have been viewed by the user, number of faults made by the user and the number of acceptable clicks within the step. A fault is generated if the user does any unacceptable clicks of interaction outside of the manual or if they do not fulfil the task needed to progress through the manual.

2.2.3 Data comparison and analysis

Machine learning have been used to evaluate the data where each test user is one instance. The first step was omitted since it is just a starting step and that leaves 34 steps in total. Each step has four usability data attributes giving a total of 136 attributes for each user. The last attribute is the group of the user that is the dependent variable of the survey question. User not finishing the manual are placed in group 0.

The K-nearest neighbour algorithm was used with tests of neighbours one to four to evaluate the best result. It was done on all the attributes to predict the group (category). The reason for KNN was that the dataset was rather small and that there were numerical values of the usability data that was used to predict and classify a nominal category. The numerical values provide for good option to measure distance and still provide the ability to categorise nominal groups. The KNN used 10-cross fold validation. This classifies one tenth of the instances and using the rest as training data. This is repeated until each instance is classified.

Two types of select attributes evaluators was used to identify important attributes. The two select attributes methods were CfsSubset evaluation with best first method and Info gain attribute evaluation with ranker. They were used because of its ability to identify important attributes.

2.3 Reliability and Validity

One major concern for this study is the construct validity concerning the term “effective”. In this method “effective” will mean that the user values the manual high on whether they learned from the manual or not. In reality it is further use of the manual that is interesting. Maybe users find the manual very good but still misunderstand one important detail causing problems when creating reports in the future. Also, retention is not measured meaning that users might perceive that they have learned but will forget it more quickly than is desirable.

Since the test users have no relation to Meridix and will not use it in their daily life their interest in understanding the system might be limited. There is a difference in real users that will use the manual to learn the system and test users that will only test the manual. Caution should be taken when providing general assessments about manual quality based on this selection problem. It can still provide insight of the usefulness of certain kind of usability data.

Since the goal of the study is to find interesting usability data to evaluate specific parts of a manual reliability should be provided. It would at least be possible to create other manuals or interactive learning material and incorporate measurement of this data. It might be that some data is only useful under certain conditions and that other types of data not incorporated here might be even more useful. Further studies will be needed to draw general conclusions about the importance of the usability data tested and if they apply in other materials

2.4 Ethical Considerations

The reports generated, and the users are confidential information within the company and will not be gathered. The only type of user data gathered is to distinguish two users from each other's. No identifiable data about the users will be gathered and stored.

The users will be informed that they are participating in an evaluation of a new feature as part of a degree project in computer science.

3 Implementation

The manual developed in this project consist of two parts; one frontend and one backend. The implementation runs together with the Meridix System ASP.NET application that is set up in demo mode on an azure VM and is a slightly modified version of the production application of Meridix. A simple overview showing the key interaction of the system is shown in figure 3.1.

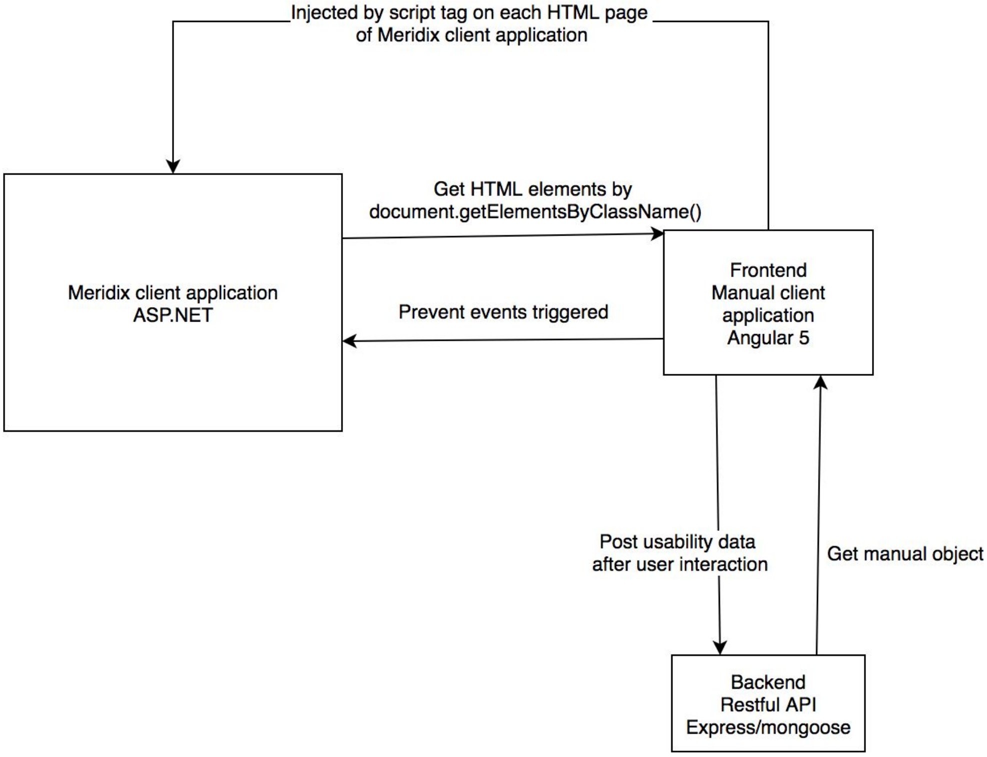

Injected by script tag on each HTML page of Meridix client application
Meridix client application ASP.NET
Get HTML elements by document.getElementsByClassName()
Prevent events triggered
Frontend Manual client application Angular 5
Post usability data after user interaction
Get manual object
Backend Restful API Express/mongoose

Figure 3.1 - Overview diagram of the system

3.1 Frontend

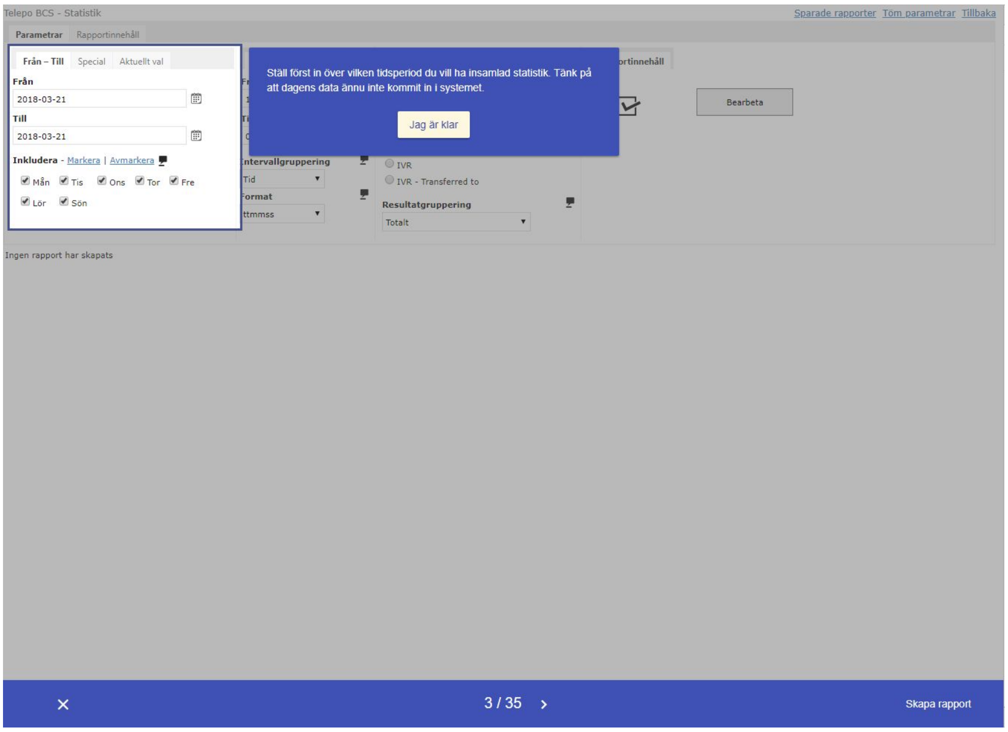
The frontend developed for this project is an angular application plugin that is run by adding a script-tag into every html page of the Meridix ASP.net client application.

3.1.1 Responsibilities

The frontend is responsible for providing the manual user interface (UI). It issues a get-request to the backend to retrieve the manuals as JSON objects.

All text in the UI is retrieved from the current manual object[2] as well as positioning and decisions on progress through the manual. For instance, each step can allow the user to go back to the previous step if the manual object specifies that it is allowed. The UI software is responsible for providing the ability for the manual object to govern the progress through the manual and all needed functionality for the manual.

Since the manual is created on top of an ASP.NET application, the manual retrieves important elements by finding manual specific class names of the HTML elements of the ASP.NET application. The manual finds these elements and highlights them or places information in relation to them. For example, in figure 3.2 the manual highlights the date selection div element of the report generator.



Ställ först in över vilken tidsperiod du vill ha insamlad statistik. Tänk på att dagens data ännu inte kommit in i systemet.

Från: 2018-03-21

Till: 2018-03-21

Inkludera: ☑ Mån ☑ Tis ☑ Ons ☑ Tor ☑ Fre ☑ Lör ☑ Sön

Intervallgruppering: Tid

Format: ttmmss

○ IVR
○ IVR - Transferred to

Resultatgruppering: Totalt

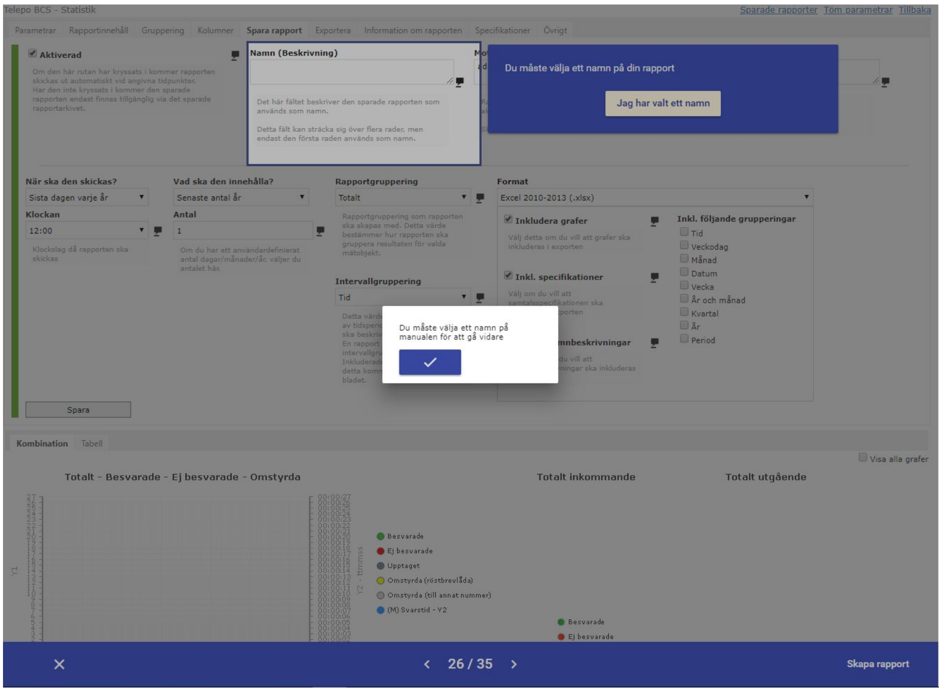
Ingen rapport har skapats

3 / 35

Figure 3.2 Step highlighting area where user should interact

The UI also handles progress control and only allows accepted user interaction. Any element can be disabled or accepted for interaction by the manual object. An example is shown in figure 3.3.

[2] See appendix C for example of the first step in the manual object



Du måste välja ett namn på din rapport

Du måste välja ett namn på manualen för att gå vidare

Om den här rutan har kryssats i kommer rapporten skickas ut automatiskt vid angivna tidpunkter. Har den inte kryssats i kommer den sparade rapporten endast finnas tillgänglig via det sparade rapportarkivet.

Det här fältet beskriver den sparade rapporten som används som namn.

Detta fält kan sträcka sig över flera rader, men endast den första raden används som namn.

Klockslag då rapporten ska skickas

Om du har ett användardefinierat antal dagar/månader/år, väljer du antalet här.

Rapportgruppering som rapporten ska skapas med. Detta värde bestämmer hur rapporten ska gruppera resultaten för valda mätobjekt.

Välj detta om du vill att grafer ska inkluderas i exporten

Figure 3.3 Step that requires the user to act correctly to proceed

The frontend is also responsible for constantly sending usability data to the backend for persistent storing.

3.1.2 Technology, language and frameworks

The frontend was developed with Angular 5[3] and uses a redux pattern (ngrx[4]) for state management. Redux is a state management pattern that provides a single source of truth for the application and storing all state in one place. To design the frontend application, a library called Angular Material which provided pre-styled components, and additionally custom theming was applied onto these components. Markdown (ngx-md) has been used to provide better styling options for manual texts.

3.1.3 Important design decisions

Creating a manual that could work with the ASP.NET application without interfering with it was the main goal of the application and proved to be difficult. Suboptimal solutions have been implemented for this to work. For

[3] https://angular.io/

[4] https://github.com/ngrx/platform

instance, using unique class names instead of ids to find elements as well as checking element status instead of the application status. For example, to know if a user has filled in a mandatory textbox the textbox element had to be found in the DOM and its status then checked by the manual.

3.2 Backend

The backend provides two API interfaces used by the frontend.

3.2.1 Responsibilities

The backend provides restful API interfaces for retrieving manuals in JSON and posting usability data. It provides a get/manuals interface as well as a post/usability-data interface. It persistently stores the usability data and manuals in a database. A command line application provides the ability to clear the database as well as adding manuals based on JSON files. Another responsibility is to extract the usability data, and a function within the command line application generates an Excel file from all usability data.

3.2.2 Technology, language and frameworks

The backend is created by the generator-api[5] and uses Node.js with Express and Mongoose. It is a RESTful API providing the common server requests. To create the Excel file the package json2xls[6] was used.

3.2.3 Important design decisions

Simplicity was the main reason for the choice of technology in this project. The backend only needed to support two endpoints with not very complicated data structures. It is mostly needed to support the persistent storage of the usability data through the post/usability-data endpoint.

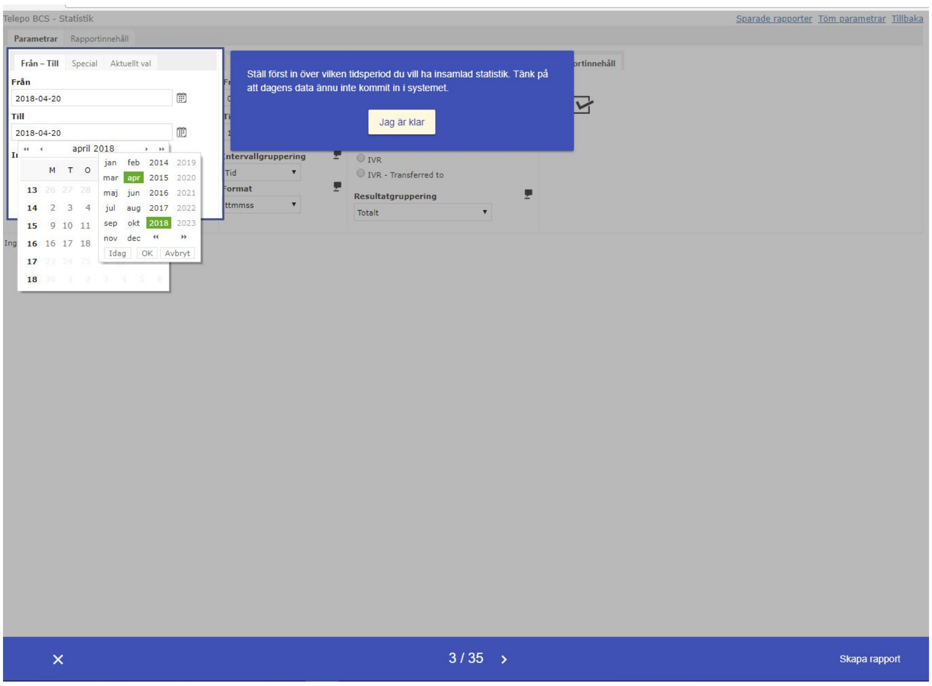
The manuals' schema was designed in a way to reduce the amount of information needed in the manual.json file that defines the manual. A lot of defaults are defined so only relevant information for the defined step are needed to be included in the json file.

[5] https://github.com/ndelvalle/generator-api

[6] https://www.npmjs.com/package/json2xls

3.3 Challenges

The use of Angular was new to the developer of this project and had to be learned from scratch. The backend consisted of technology which had previously been used by the developer, and therefore it was easier to implement. The biggest challenge was to implement the UI on top of the ASP.NET application. First, the developer of this project had no control over the underlying ASP.NET application and could only request to have elements be given a specific class name. Not all elements could be given unique class names since they were dynamically created. Also, most of the time the manual application was an overlay showed in front but sometimes there where overlays in the ASP.NET application that needed to be over the manual components.



Ställ först in över vilken tidsperiod du vill ha insamlad statistik. Tänk på att dagens data ännu inte kommit in i systemet.

Från 2018-04-20

Till 2018-04-20

Intervallgruppering Tid

Format ttmmss

Resultatgruppering Totalt

Figure 3.3 Overlay example. Blue border and box overlay over Meridix and date- and year pickers overlay over manual elements.

The ASP.NET application requests a new page from the server for every page in the application and in the current method of loading, the manual it forces the Angular application to restart. This requires the manual to store its state in the local storage to remember where the users are in the manual. It also limits some functionality, for instance users cannot move backwards in the manual after different steps where new pages have been rendered. It also cannot start the manual at the same position as it was closed since the Angular application has no way of routing in the ASP.NET application.

3.4 Machine learning tool

The Weka[7] application was used to perform the machine learning algorithms. The excel file of the usability data was exported as an csv-file and then opened in the Weka application. The group column had a number 0-5 and was transformed from numerical to nominal for the classification.

[7] https://www.cs.waikato.ac.nz/ml/weka/

4 Results

For the full report documentation see appendix B.

4.1 Statistical loss and participation

A total of 45 individuals were given login information to perform the test. Data was gathered from 21 participants with one participant aborting on the first step. The participants were grouped into six categories depending on their answer on the evaluation question, “To what degree do you agree with the following statement ‘*The manual have taught me how to create a report in Meridix*’?, and one category for participants not finishing the manual. No one entered 1 on the question and 0 was chosen to denote the aborter group. The table below show the number in each category with aborter being in majority.

Category	Number of instances
0 (aborters)	11
1 (Do not agree at all)	0
2	1
3	3
4	4
5 (Agree completely)	2

Table 4.1

4.2 ZeroR

ZeroR is a simple algorithm used as a baseline. It classifies all instances into the majority category. Out of 21 instances 11 was category 0 and gives a correct classification of 52.381% using ZeroR.

4.3 KNN - K-nearest neighbour

When using KNN different number of neighbours was tested from 1-4 where provided the best result with 57.14% accuracy

Nr of neighbours	Correctly classified categories
1	28.57%
2	47.62%
3	57.14%
4	42.86%

Table 4.2

4.4 Decision tree algorithms

Two different decision tree algorithms was tested; Random forest and J48 that is a extension of ID3. They both gave poor prediction below ZeroR with 47.6% accuracy for random forest and 38% accuracy for J48.

4.5 Attribute selection

Two different attributes selection methods were used to identify important attributes; Info gain attribute evaluation with a ranker and CfsSubset evaluation with a best first approach. These where run on the whole set of 136 attributes and the CfsSubset provided the top three attributes of the info gain evaluation. Only seven attributes where identified with a rank larger than zero on a scale from zero to one. When a KNN with 3 neighbours was performed on the subset of these seven attributes a prediction level of 33.33% was achieved.

4.6 Single usability type

The KNN with 3 neighbours was also used after singling out each usability data type in order to try to predict what type of data that was most important in trying to categorize the dataset. This was done by only keeping one of the types and still using all the steps. The results are shown in the table below.

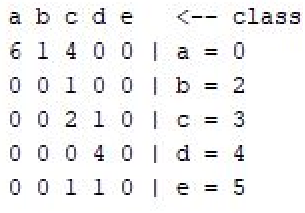
Usability data type	Correctly classified categories
Time taken	19.05%
Number of views	33.33%
Number of acceptable clicks	52.38%
Number of faults	57.14%
Number of acceptable clicks and number of faults	57.14%

Table 4.3

Even though the percentage of correctly classified categories are exactly the same for the full dataset and for only the number of faults the confusion matrices are not the same. The KNN 3 was then run on the usability data with the highest accuracy. The number of acceptable clicks and the number of faults giving the same percentage but a different confusion matrix.

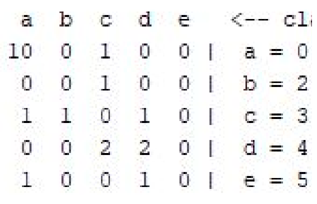
The confusion matrix shows the distribution of the classified categories. Each column represent the number classified in a certain category and the row represent the real category of that instance. As an example when looking at the confusion matrix of the full dataset the 4 in the first row represent that 4 aborters (category a = 0) was classified as belonging to category c.

Full dataset



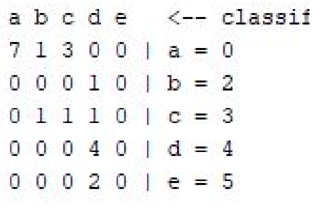
```
a b c d e   <-- class
6 1 4 0 0 | a = 0
0 0 1 0 0 | b = 2
0 0 2 1 0 | c = 3
0 0 0 4 0 | d = 4
0 0 1 1 0 | e = 5
```

Number of faults only



```
 a   b   c   d   e   <-- cl
10   0   1   0   0 |  a = 0
 0   0   1   0   0 |  b = 2
 1   1   0   1   0 |  c = 3
 0   0   2   2   0 |  d = 4
 1   0   0   1   0 |  e = 5
```

Acceptable clicks and number of faults



```
a b c d e   <-- classif
7 1 3 0 0 | a = 0
0 0 0 1 0 | b = 2
0 1 1 1 0 | c = 3
0 0 0 4 0 | d = 4
0 0 0 2 0 | e = 5
```

Table 4.4

When using the full dataset the confusion matrix shows that 6 out of 7 (85.7%) of instances in category 3 and 4 was correctly classified. Category 2 and 5 could not be correctly classified since there are too few instances in each category. No non-aborter (category 2-5) was classified as an aborter and when only considering aborters and non-aborters 16 out of 21 (76.19%) was classified correctly.

The confusion matrix for only the number of faults shows that 10 out of 11 (90.9%) aborters was correctly classified. For both the groups of aborters and non aborters the correct classification is 18 of 21 (85.7%)

4.7 Step with high faults and the discovery of a bug

A high number faults was identified at two specific steps; step 2 with an average of 3.1 faults per user and step 10 with an average of 2.55 faults per user. In comparison no other step had more than 1 as an average and many had zero. When examining step 2 a bug was identified that made the user progress through the manual but not through the site unable to comply with further instructions. It is possible to click next step until step 10 where the user then gets stuck and are unable to progress any further. Two users aborted after step 11 indicating that they may have encountered the bug. One user aborted before encountering the bug and all other users were able to continue the manual beyond step 11 and could not have encountered the bug.

One user pointed out that the instructions in the description was not possible to carry out. It instructed the user to make a saved report to be sent every three months. It is possible to generate a report that contains the last three months, but it is not possible to have it sent in other intervals then every week, every month or every year. Four users aborted the manual after the step that exposed this problem. Whether they aborted because of this or not is possible to determine from the data.

5 Analysis

5.1 Statistical loss and dataset validity

The statistical loss was significant and caused problems for the result. First of,more than 50% of test user did not participate in the test. Secondly of the 21 participants 11 ended the manual prematurely before answering the question used in the evaluation of effectiveness. This made the abort group the largest category in the classification. The statistical loss can partly be explained by the effort of taking the test not only devoting 15 minutes of time but also the effort of learning the system the manual is teaching. The large number of people aborting the test may partly be explained by the bug identified in step 2 and the unclear instructions.

5.2 KNN classification

The best classification result was 57.14% that is just marginally better compared to the ZeroR 52.38%. In comparison a totally randomized distribution would generate an accuracy of 16,7% (or 1 in 6) since there are six categories. One problem is the small dataset since the best result was made with three neighbours it is not possible for instance to classify the single person in category 2 correct and not the two users in category 5 either. It makes a maximum prediction ability of 85.7%. The quality of the classification is difficult to evaluate since the dataset is so small. Six categories on 21 instances is not enough to provide a valid result.

Still each instance has a high number of attributes (136) and the classification is severely better than the randomized distribution. When investigating the confusion matrix for the full dataset the prediction of different categories differ. In category 4 all four instances where categorized correctly and for category 3 two out of three was classified correctly. Also, no non-aborters were classified as aborters. With a larger dataset it would have been interesting to further evaluate how significant the improvement of KNN over ZeroR is.

5.3 Decision tree algorithms

The decision tree algorithms did not provide a better classification than ZeroR. Their poor performance may be explained by both the small dataset, the high number of categories and the high number a attributes each providing small importance in the determination of category. For instance in J48 important attributes are identified and classification is based on only a small subsection of the attributes. These attributes do to some extent correlate between the attributes identified in attribute selection and was not able to make good prediction on their own.

5.4 Attribute selection

The attribute selection does not provide any interesting results. Neither method is able to identify important steps or usability data in any meaningful way. The KNN algorithm provides a lower correct classification rate when the identified attributes were identified then when the full set was used.

5.5 Usability type

When the KNN was used on a single type for every step a pattern is emerging. It seems like time was not an important factor since it provides for poor classification and nr of faults provided just as good classification as using the full set. One might think that fault was the only factor, but the two confusion matrices differ for full set and only number of faults. The faults are better at determining aborters and the full set is better at determining the other categories. When evaluation the confusion matrix of the combination of acceptable clicks and faults the same prediction is achieved with a confusion matrix much closer to the full dataset.

The conclusion is that number of faults has the highest impact on user quitting the manual. It predicts correctly between aborters and non-aborters in 18 out of 21 instances which indicates that it might be good for identifying aborters. Also, the most interesting usability data is the number of acceptable clicks and the number of faults where time and number of views don’t seem to provide much information about the quality of the manual.

6 Discussion

This paper is trying to determine if either or both of two claims are true; usability data can be used to measure the effectiveness of a manual and usability data can be used to identify which parts of manual that should be improved to increase effectiveness.

6.1 Question of validity of the results

First of, the dataset of test users was small in comparison to the number of categories. This questions the validity of the results and general conclusions are difficult to justify. As a consequence of the small dataset the results are subject both to the problem that some instances where impossible to classify correctly and that selection randomness have a big impact. Even if four out of four instances of category 4 was classified correctly this might be because of the sample of test users.

Each test user has a large set of attributes (136) that provided for a little bit better quality of the results in that regard. Each usability data was measured in 34 attributes providing a

6.2 The impact of bugs

After the data was gathered and analysed a high number of aborters were identified. After analysing the average number of faults and response from the users a bug was identified preventing users from continuing correctly in the manual. When examining the data further a total of six users may have been affected by the bugs.

This possible increased of the number of aborters may have created a uneven distribution of instances in each category. This decreases the ability of the results to evaluate if usability data can be used to measure effectiveness since it is harder to classify the other categories correctly because of the few number of instances in them. On the other hand it increased the ability to classify between aborters and non-aborters since the selection of aborters are higher. The results support this since the classification of aborters versus non-aborters is significantly higher får k-nearest neighbour then ZeroR.

6.3 Usability data ability to measure effectiveness

The results indicate that usability data might provide some measurement of effectiveness. At least using k-nearest neighbour algorithm it provided a better classification then the ZeroR classification. With a larger dataset of test users, the k-nearest neighbour might provide better classification but that is to be further studied.

The question of identification of parts that affects effectiveness the attribute selections techniques where not successful but identifying high averages of faults for each step provided insight of where in a manual to look to identify problems that causes users to abort a manual.

6.4 Importance of different usability data types

In this project four types of usability data were tested; time taken on each step, number of views of each step, number of acceptable clicks and number of faulty or unallowed clicks. When category classification was made for each of these types separately the results show that time taken, and number of views did not provide good classification showing a low causality between this data and the effectiveness as it was defined in this paper.

On the other hand, number of acceptable clicks and faults provided as good of a classification as all the data when combined and number of fault provided the same classification rate as the full dataset. This provides some indications that these types of usability data are better at measuring the effectiveness of a manual then the other types.

One reason that number of faults was just as good as the full dataset might be the large number of test users aborting the manual before finishing it. The confusion matrices indicates that number of faults might be better at detecting the aborters than the others usability data types.

6.5 Usability data versus direct user feedback

The manual developed and used in this project was a first iteration manual and consisted of several bugs and limitations. The nature of these bugs and problems were identified directly by the test user and forwarded to the developer through text or screen dumps. It was much more specific then the usability data. Although the average number of faults could be used to identify specific steps the problem at these steps were identified through manual testing.

One problem with the test used in this project is then that the data identifies bugs instead of effectiveness. Although one could argue that bugs limits effectiveness they should not be there if proper production tests would have been implemented.

The average number of faults for each step could be used to pinpoint the location of problems and is more useful in contexts where the developer has no direct contact to the user.

6.6 Comparison with previous research

No previous studies about the usefulness of gathering usability of this type in relation to interactive learning material have been found and therefore no comparisons can be made.

7 Conclusion

The small dataset really limits the ability to draw any conclusions about the difference between ZeroR and k-nearest neighbour. It neither proves or disproves the statement that the usability data can be useful in measuring effectiveness.

Number of faults and acceptable clicks was better predictors than time taken and number of views when identifying effectiveness. The number of faults had a accuracy of 85.7% when classifying between aborters and non-aborters. This indicates that it can be used to identify aborters or at least abortion caused by bugs since possible 6 out of 11 aborters may have been affected by known bugs.

Attribute selection do not provide any good information about important steps in a manual but average number of faults of each step do identify the location of important problems.

7.1 Future work

In further studies larger datasets should be studied. Also, implementation of two different manuals in the same subject could be one way of testing if the usability data in this study can be used to identify the better manual of the two. Testing the usability data types on other types of multimedia learning material should be done to verify the results of the types themselves rather than the specific manual created in this project.

References

[1] R. E. Mayer, *Multimedia learning*, 2nd ed. Cambridge University Press, 2009.

[2] R. Mohd Rias, H.B. *Zaman. Looking at the Effects of Various Multimedia Approach in Student Learning: A Case Study.* ICUIMC(IMCOM)'13, January 17–19, 2013, Kota Kinabalu, Malaysia.
Available: https://dl-acm-org.proxy.lnu.se/citation.cfm?id=2448583

[3] J. A. C. Hattie, *Visible learning.* Routledge, Taylor and Francis Group, 2009

[4] M. Nakayama, Y. Shimizu. Evaluation of a multimedia learning exercise using oculo-motors. ETRA 2006, San Diego, California, 27–29 March 2006.
Available: https://dl-acm-org.ep.bib.mdh.se/citation.cfm?id=1117331

[5] J. Qiu, J. Tank, T. X. Liu, J. Gong, C. Zhang, Q. Zhang, Y. Xue. Modeling and Predicting Learning Behavior in MOOCs. WSDM'16, February 22–25, 2016, San Francisco, CA, USA.
Available: https://dl-acm-org.proxy.lnu.se/citation.cfm?id=2835842

[6] B. Vassilidis, A. Kameas, C. Sgouropoulou. *A closer look at MOOC's adoption from a qualitative perspective*. PCI '16, November 10-12, 2016, Patras, Greece.
Available: https://dl-acm-org.proxy.lnu.se/citation.cfm?id=3003799

[7] A. Anderson, D. Huttenlocher, J. Kleinberg, J. Leskovec. *Engaging with Massive Online Courses*. WWW'14, April 7–11, 2014, Seoul, Korea.
Available: https://dl-acm-org.proxy.lnu.se/citation.cfm?id=2568042

[8] M. Fabiani, J. Buckley, G. Gratton, M. G.H. Colesm E. Donchin, R. Logie. *The training of complex task performance.* Acta Psychologica, volume 71 issues 1-3 August 1989.
Available:
https://www.sciencedirect.com/science/article/pii/0001691889900127

[9] J. Y.C. Chen. *Utility of game instructions*. ARMY RESEARCH INST FOR THE BEHAVIORAL AND SOCIAL SCIENCES ALEXANDRIA VA. April 2003. Available: http://www.dtic.mil/docs/citations/ADA414105

Appendix A - Test user instruction

The following letter was sent to all participants in the study.

“Hej

Tack för att du hjälper mig i mitt examensarbete i datavetenskap!

Du ska använda en manual som beskriver hur man gör en rapport i Meridix. Meridix är ett rapportgenereringssystem för telefontrafik hos ett företag eller en organisation. Meridix kan alltså användas för att ta reda på hur telefontrafiken ser ut under en vald period. Det bör maximalt ta 10 minuter att gå igenom manualen.

Föreställ dig att du har fått följande uppgift

“Du ska ta fram statistik över mars år 2012 för de olika avdelningarna; försäljning och support. Rapporttypen som du ska använda kallas för ‘User id’. Det du är intresserad av är information som berättar hur lätt det är att komma i kontakt med avdelningarna. Du ska sedan skapa en automatisk rapport som skickas var tredje månad till ditt användarnamn.”

Ditt användarnamn: TesterX@fake.se
Ditt lösenord: imantest1234

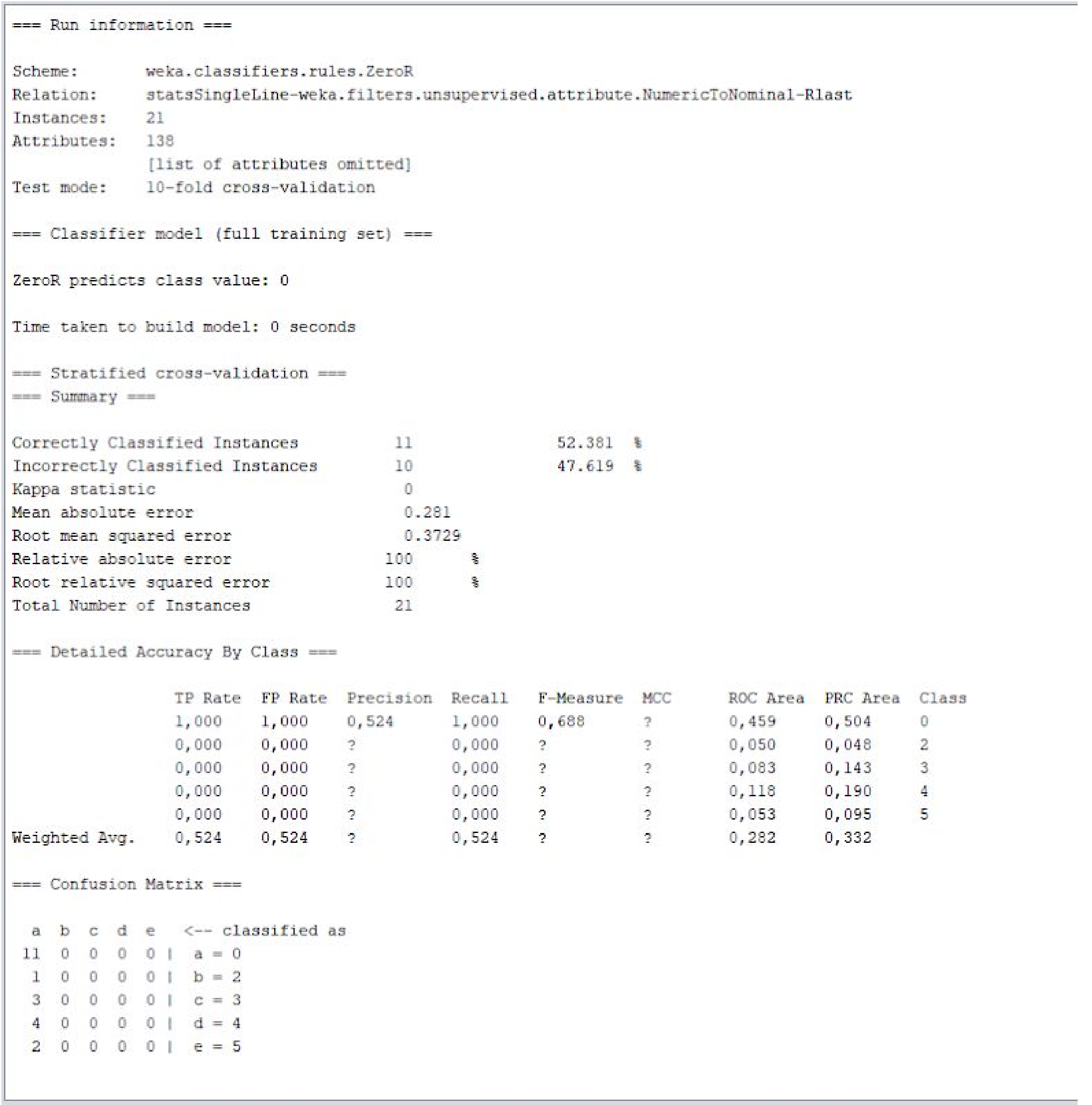
Systemet är lite långsamt första gången det körs, så ha lite tålamod och vänta några sekunder om det ser konstigt ut innan en sida laddas. Jag rekommenderar Chrome som webbläsare eftersom det än så länge endast är testat i Chrome. (Det fungerar inte i firefox)

Klicka på http://imantest.westeurope.cloudapp.azure.com för att logga in. Klicka sedan på “Starta manual” som är en knapp i mitten av startsidan.

Återigen, tack så mycket för hjälpen!”

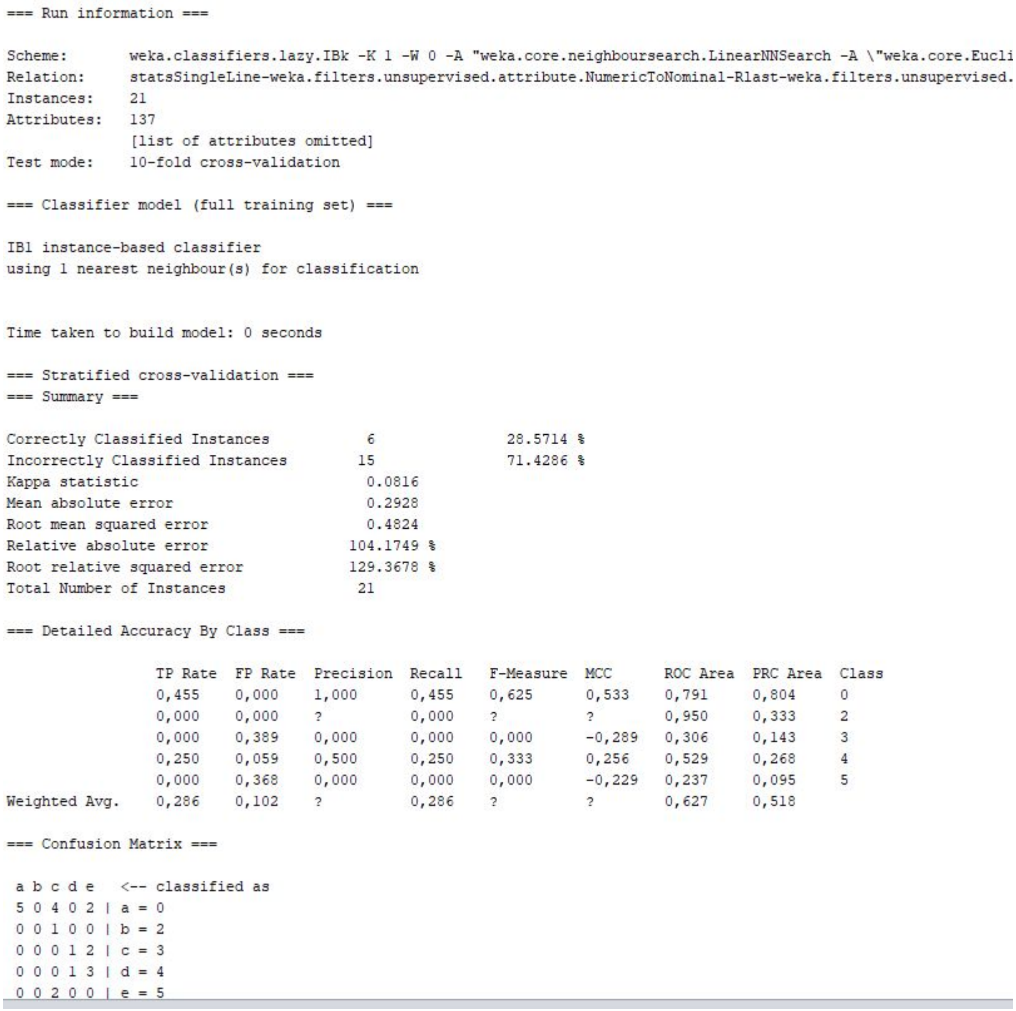
Appendix B - Result reports

ZeroR - on full set



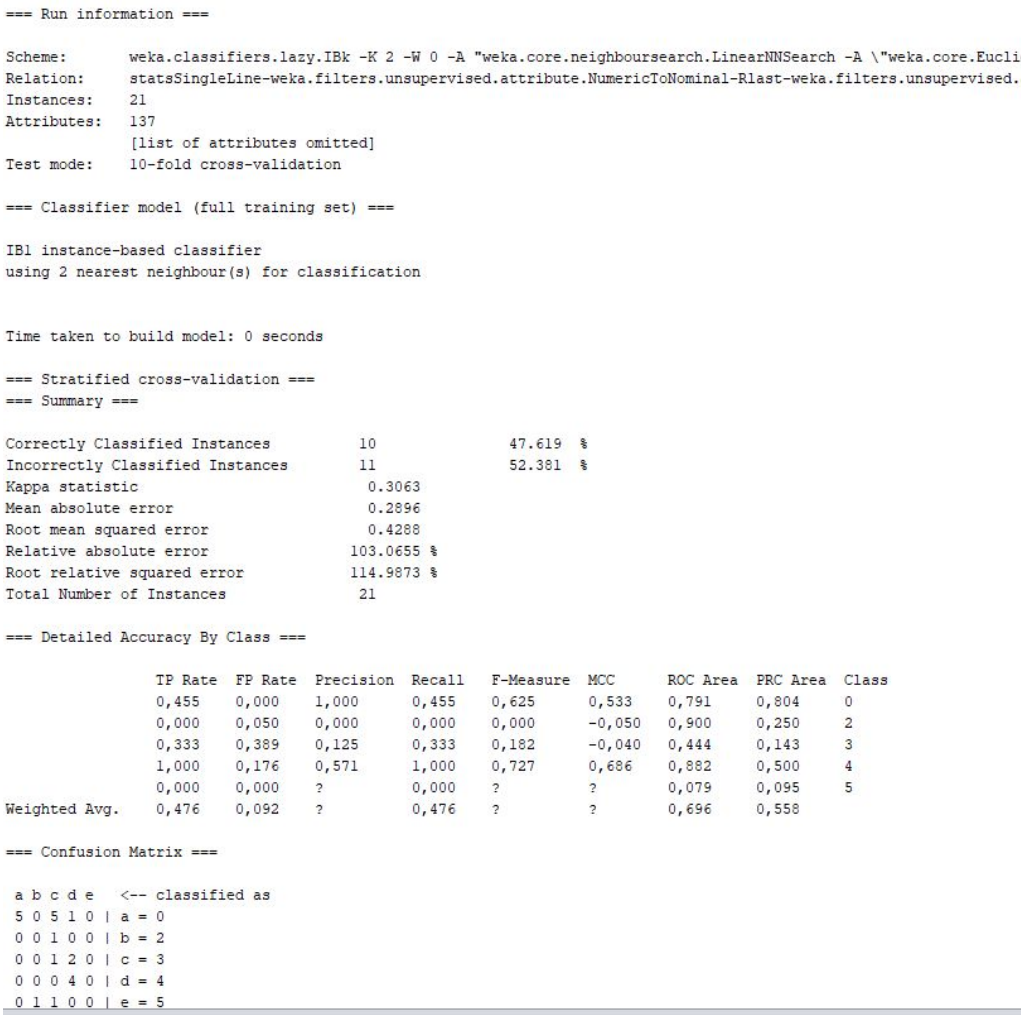
```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR
Relation:     statsSingleLine-weka.filters.unsupervised.attribute.NumericToNominal-Rlast
Instances:    21
Attributes:   138
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: 0

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          11               52.381  %
Incorrectly Classified Instances        10               47.619  %
Kappa statistic                          0
Mean absolute error                      0.281
Root mean squared error                  0.3729
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    1,000    0,524      1,000    0,688      ?        0,459     0,504     0
                 0,000    0,000    ?          0,000    ?          ?        0,050     0,048     2
                 0,000    0,000    ?          0,000    ?          ?        0,083     0,143     3
                 0,000    0,000    ?          0,000    ?          ?        0,118     0,190     4
                 0,000    0,000    ?          0,000    ?          ?        0,053     0,095     5
Weighted Avg.    0,524    0,524    ?          0,524    ?          ?        0,282     0,332

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 11  0  0  0  0 |  a = 0
  1  0  0  0  0 |  b = 2
  3  0  0  0  0 |  c = 3
  4  0  0  0  0 |  d = 4
  2  0  0  0  0 |  e = 5
```

KNN 1 - on full set



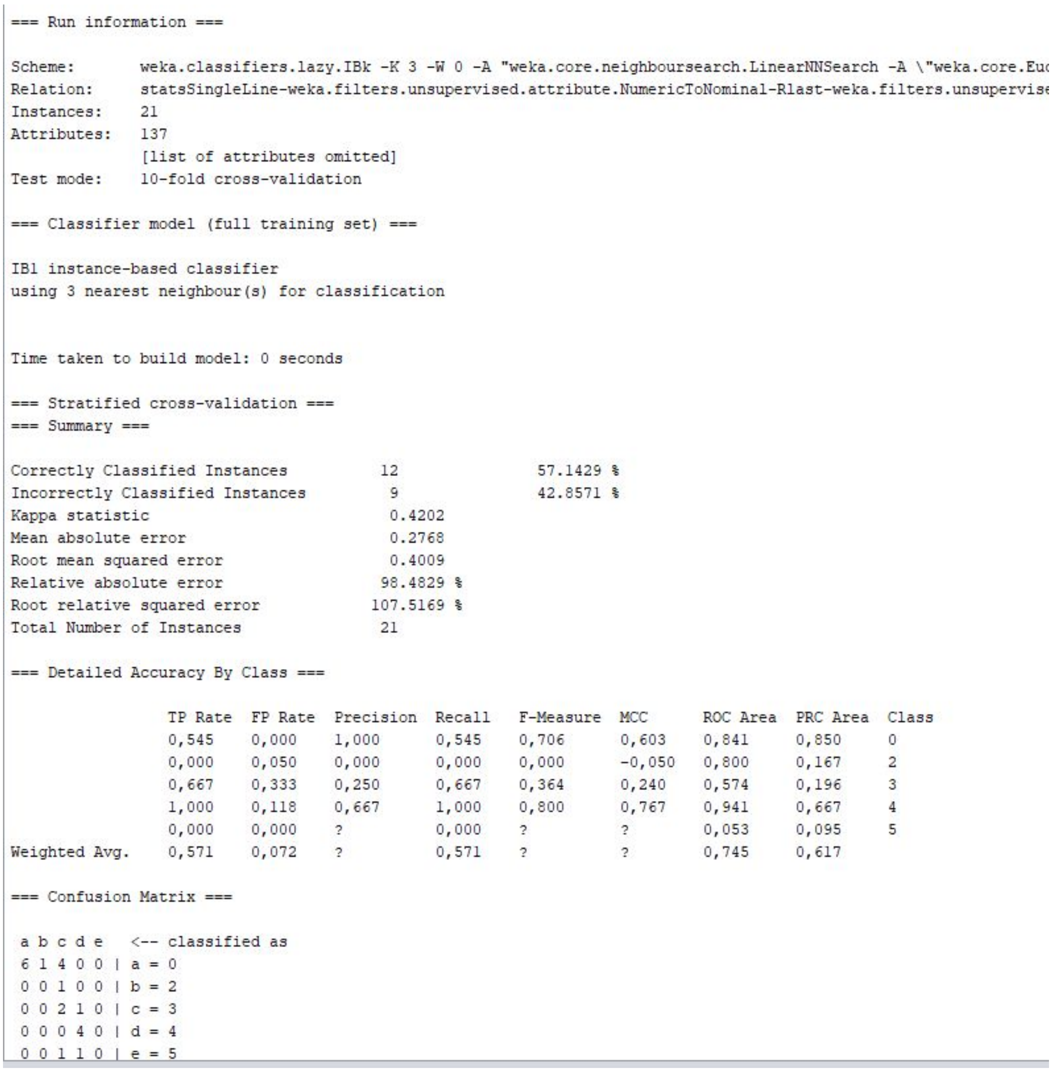
```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.Eucli
Relation:     statsSingleLine-weka.filters.unsupervised.attribute.NumericToNominal-Rlast-weka.filters.unsupervised.
Instances:    21
Attributes:   137
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           6               28.5714 %
Incorrectly Classified Instances        15               71.4286 %
Kappa statistic                          0.0816
Mean absolute error                      0.2928
Root mean squared error                  0.4824
Relative absolute error                104.1749 %
Root relative squared error            129.3678 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,455    0,000    1,000      0,455    0,625      0,533    0,791     0,804     0
                 0,000    0,000    ?          0,000    ?          ?        0,950     0,333     2
                 0,000    0,389    0,000      0,000    0,000      -0,289   0,306     0,143     3
                 0,250    0,059    0,500      0,250    0,333      0,256    0,529     0,268     4
                 0,000    0,368    0,000      0,000    0,000      -0,229   0,237     0,095     5
Weighted Avg.    0,286    0,102    ?          0,286    ?          ?        0,627     0,518

=== Confusion Matrix ===

 a b c d e   <-- classified as
 5 0 4 0 2 | a = 0
 0 0 1 0 0 | b = 2
 0 0 0 1 2 | c = 3
 0 0 0 1 3 | d = 4
 0 0 2 0 0 | e = 5
```

KNN 2 - on full set



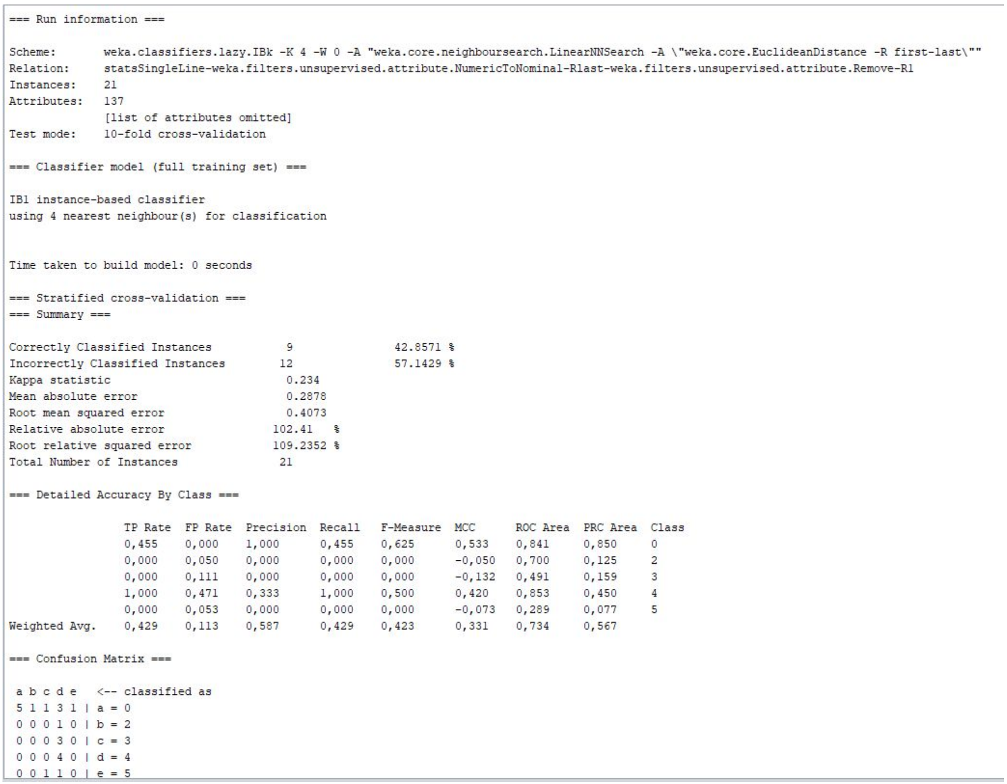
```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 2 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.Eucli
Relation:     statsSingleLine-weka.filters.unsupervised.attribute.NumericToNominal-Rlast-weka.filters.unsupervised.
Instances:    21
Attributes:   137
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 2 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          10               47.619  %
Incorrectly Classified Instances        11               52.381  %
Kappa statistic                          0.3063
Mean absolute error                      0.2896
Root mean squared error                  0.4288
Relative absolute error                103.0655 %
Root relative squared error            114.9873 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,455    0,000    1,000      0,455    0,625      0,533    0,791     0,804     0
                 0,000    0,050    0,000      0,000    0,000      -0,050   0,900     0,250     2
                 0,333    0,389    0,125      0,333    0,182      -0,040   0,444     0,143     3
                 1,000    0,176    0,571      1,000    0,727      0,686    0,882     0,500     4
                 0,000    0,000    ?          0,000    ?          ?        0,079     0,095     5
Weighted Avg.    0,476    0,092    ?          0,476    ?          ?        0,696     0,558

=== Confusion Matrix ===

 a b c d e   <-- classified as
 5 0 5 1 0 | a = 0
 0 0 1 0 0 | b = 2
 0 0 1 2 0 | c = 3
 0 0 0 4 0 | d = 4
 0 1 1 0 0 | e = 5
```

KNN 3 - on full set



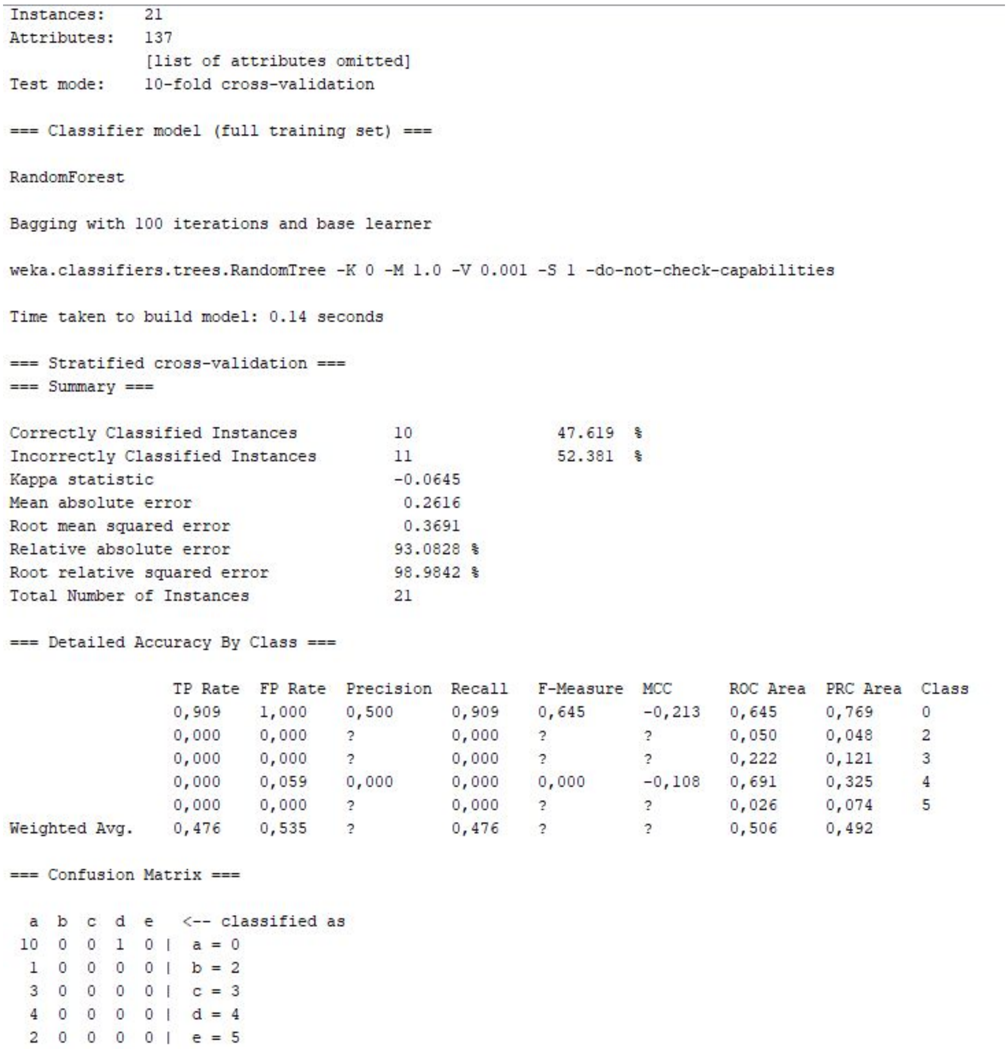
```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 3 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.Euc
Relation:     statsSingleLine-weka.filters.unsupervised.attribute.NumericToNominal-Rlast-weka.filters.unsupervise
Instances:    21
Attributes:   137
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          12               57.1429 %
Incorrectly Classified Instances         9               42.8571 %
Kappa statistic                          0.4202
Mean absolute error                      0.2768
Root mean squared error                  0.4009
Relative absolute error                 98.4829 %
Root relative squared error            107.5169 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,545    0,000    1,000      0,545    0,706      0,603    0,841     0,850     0
                 0,000    0,050    0,000      0,000    0,000      -0,050   0,800     0,167     2
                 0,667    0,333    0,250      0,667    0,364      0,240    0,574     0,196     3
                 1,000    0,118    0,667      1,000    0,800      0,767    0,941     0,667     4
                 0,000    0,000    ?          0,000    ?          ?        0,053     0,095     5
Weighted Avg.    0,571    0,072    ?          0,571    ?          ?        0,745     0,617

=== Confusion Matrix ===

 a b c d e   <-- classified as
 6 1 4 0 0 | a = 0
 0 0 1 0 0 | b = 2
 0 0 2 1 0 | c = 3
 0 0 0 4 0 | d = 4
 0 0 1 1 0 | e = 5
```

KNN 4 - on full set



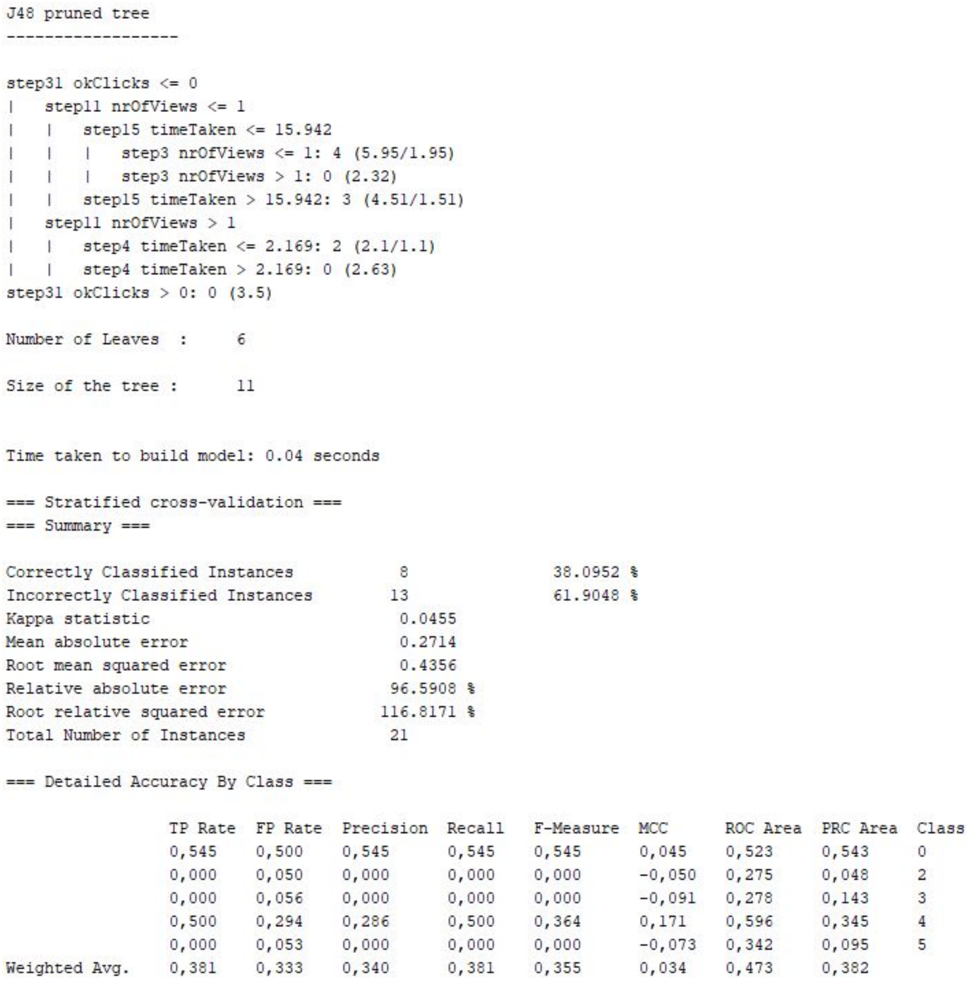
```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 4 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Relation:     statsSingleLine-weka.filters.unsupervised.attribute.NumericToNominal-Rlast-weka.filters.unsupervised.attribute.Remove-R1
Instances:    21
Attributes:   137
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 4 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           9               42.8571 %
Incorrectly Classified Instances        12               57.1429 %
Kappa statistic                          0.234
Mean absolute error                      0.2878
Root mean squared error                  0.4073
Relative absolute error                102.41   %
Root relative squared error            109.2352 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,455    0,000    1,000      0,455    0,625      0,533    0,841     0,850     0
                 0,000    0,050    0,000      0,000    0,000      -0,050   0,700     0,125     2
                 0,000    0,111    0,000      0,000    0,000      -0,132   0,491     0,159     3
                 1,000    0,471    0,333      1,000    0,500      0,420    0,853     0,450     4
                 0,000    0,053    0,000      0,000    0,000      -0,073   0,289     0,077     5
Weighted Avg.    0,429    0,113    0,587      0,429    0,423      0,331    0,734     0,567

=== Confusion Matrix ===

 a b c d e   <-- classified as
 5 1 1 3 1 | a = 0
 0 0 0 1 0 | b = 2
 0 0 0 3 0 | c = 3
 0 0 0 4 0 | d = 4
 0 0 1 1 0 | e = 5
```

Random forest algorithm



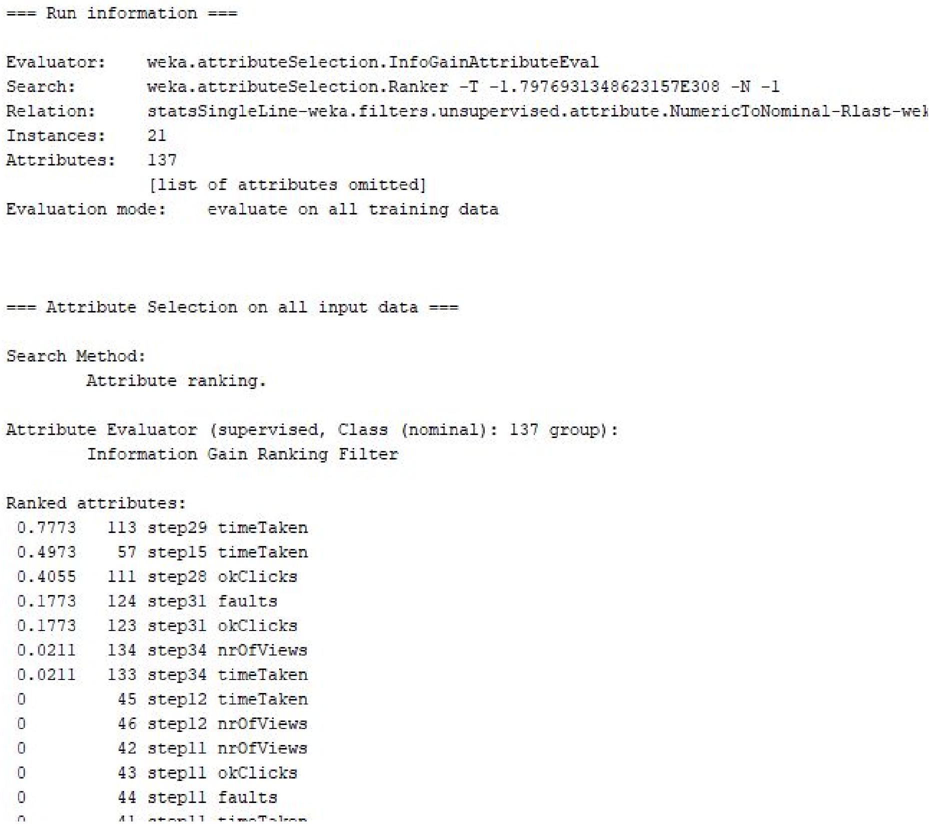
```
Instances:    21
Attributes:   137
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          10               47.619  %
Incorrectly Classified Instances        11               52.381  %
Kappa statistic                         -0.0645
Mean absolute error                      0.2616
Root mean squared error                  0.3691
Relative absolute error                 93.0828 %
Root relative squared error             98.9842 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,909    1,000    0,500      0,909    0,645      -0,213   0,645     0,769     0
                 0,000    0,000    ?          0,000    ?          ?        0,050     0,048     2
                 0,000    0,000    ?          0,000    ?          ?        0,222     0,121     3
                 0,000    0,059    0,000      0,000    0,000      -0,108   0,691     0,325     4
                 0,000    0,000    ?          0,000    ?          ?        0,026     0,074     5
Weighted Avg.    0,476    0,535    ?          0,476    ?          ?        0,506     0,492

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 10  0  0  1  0 |  a = 0
  1  0  0  0  0 |  b = 2
  3  0  0  0  0 |  c = 3
  4  0  0  0  0 |  d = 4
  2  0  0  0  0 |  e = 5
```

J48 algorithm



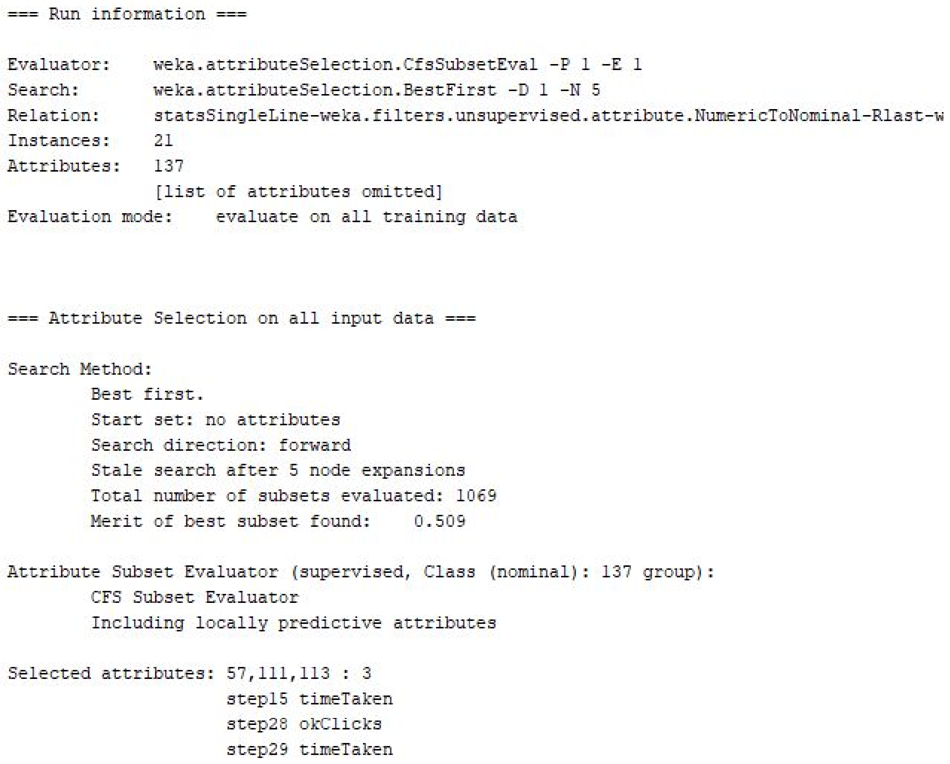
```
J48 pruned tree
------------------

step31 okClicks <= 0
|   step11 nrOfViews <= 1
|   |   step15 timeTaken <= 15.942
|   |   |   step3 nrOfViews <= 1: 4 (5.95/1.95)
|   |   |   step3 nrOfViews > 1: 0 (2.32)
|   |   step15 timeTaken > 15.942: 3 (4.51/1.51)
|   step11 nrOfViews > 1
|   |   step4 timeTaken <= 2.169: 2 (2.1/1.1)
|   |   step4 timeTaken > 2.169: 0 (2.63)
step31 okClicks > 0: 0 (3.5)

Number of Leaves  :     6

Size of the tree :      11


Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           8               38.0952 %
Incorrectly Classified Instances        13               61.9048 %
Kappa statistic                          0.0455
Mean absolute error                      0.2714
Root mean squared error                  0.4356
Relative absolute error                 96.5908 %
Root relative squared error            116.8171 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,545    0,500    0,545      0,545    0,545      0,045    0,523     0,543     0
                 0,000    0,050    0,000      0,000    0,000      -0,050   0,275     0,048     2
                 0,000    0,056    0,000      0,000    0,000      -0,091   0,278     0,143     3
                 0,500    0,294    0,286      0,500    0,364      0,171    0,596     0,345     4
                 0,000    0,053    0,000      0,000    0,000      -0,073   0,342     0,095     5
Weighted Avg.    0,381    0,333    0,340      0,381    0,355      0,034    0,473     0,382
```

Attribute selection - Info gain attribute ranker



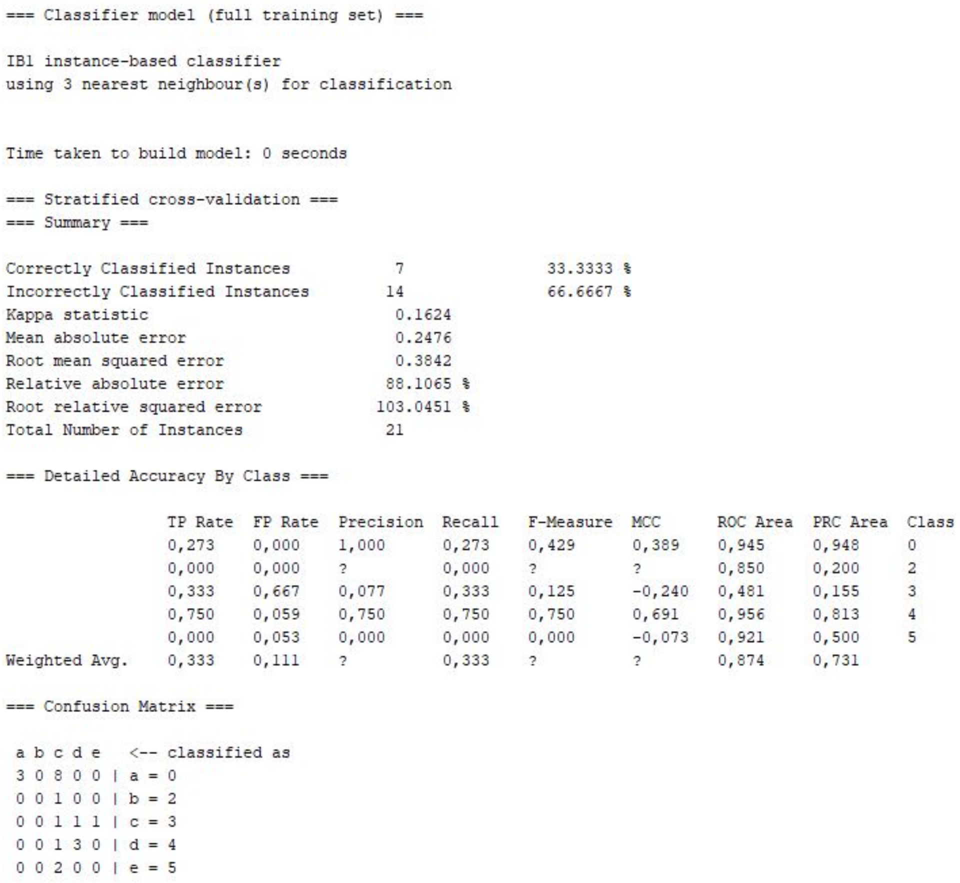
```
=== Run information ===

Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     statsSingleLine-weka.filters.unsupervised.attribute.NumericToNominal-Rlast-wek
Instances:    21
Attributes:   137
              [list of attributes omitted]
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 137 group):
        Information Gain Ranking Filter

Ranked attributes:
 0.7773   113 step29 timeTaken
 0.4973    57 step15 timeTaken
 0.4055   111 step28 okClicks
 0.1773   124 step31 faults
 0.1773   123 step31 okClicks
 0.0211   134 step34 nrOfViews
 0.0211   133 step34 timeTaken
 0         45 step12 timeTaken
 0         46 step12 nrOfViews
 0         42 step11 nrOfViews
 0         43 step11 okClicks
 0         44 step11 faults
 0         41 step11 timeTaken
```

Attribute selection - CfsSubsetEval best first



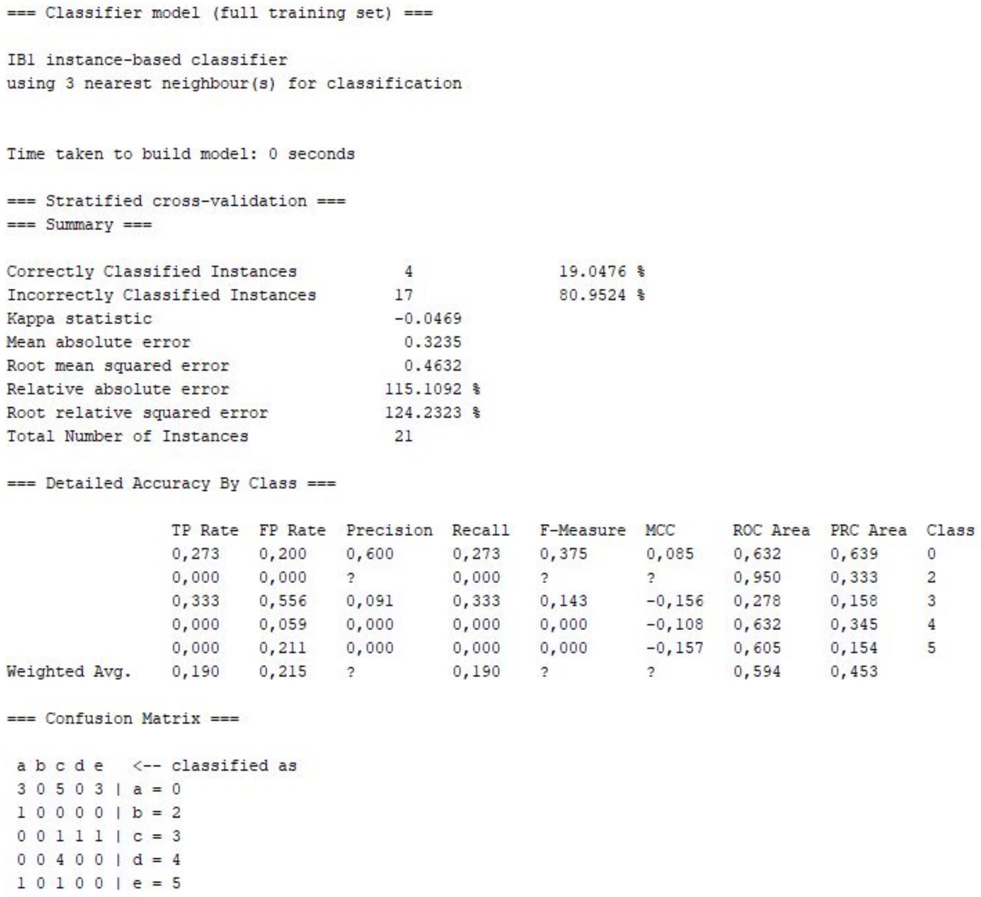
```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     statsSingleLine-weka.filters.unsupervised.attribute.NumericToNominal-Rlast-w
Instances:    21
Attributes:   137
              [list of attributes omitted]
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 1069
        Merit of best subset found:    0.509

Attribute Subset Evaluator (supervised, Class (nominal): 137 group):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 57,111,113 : 3
                     step15 timeTaken
                     step28 okClicks
                     step29 timeTaken
```

KNN 3 - on subset based on attribute selection



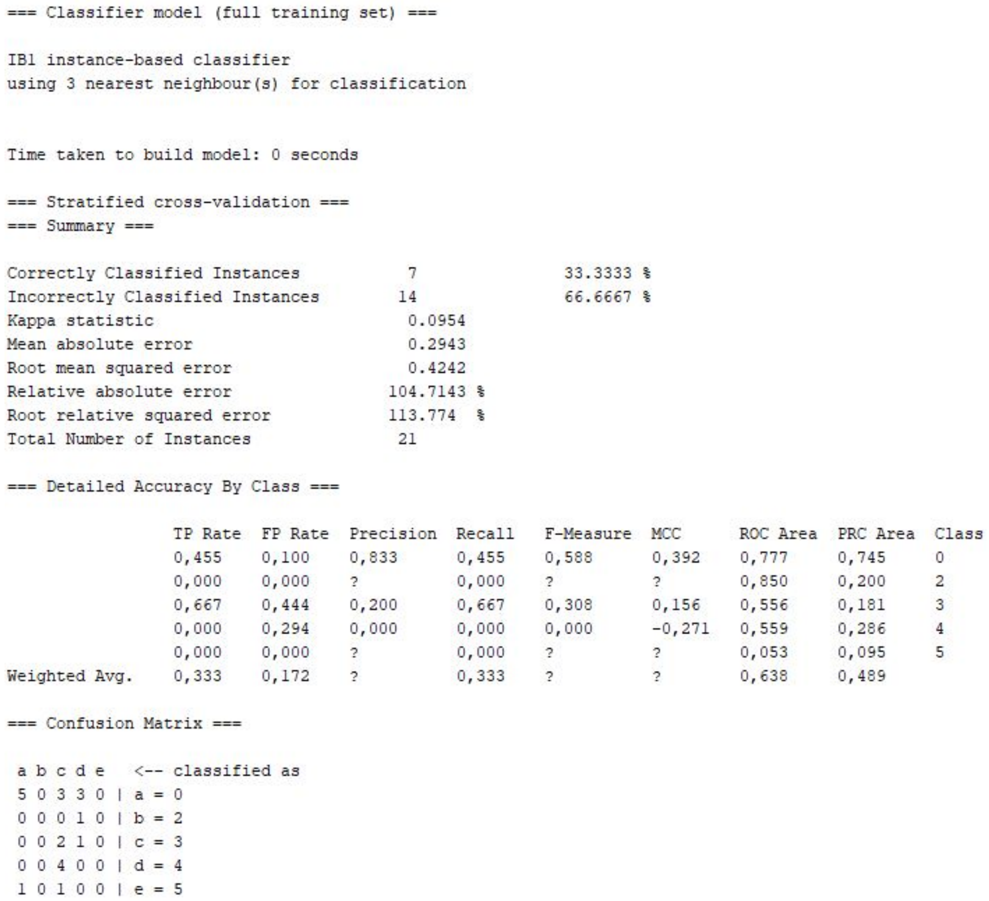
```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           7               33.3333 %
Incorrectly Classified Instances        14               66.6667 %
Kappa statistic                          0.1624
Mean absolute error                      0.2476
Root mean squared error                  0.3842
Relative absolute error                 88.1065 %
Root relative squared error            103.0451 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,273    0,000    1,000      0,273    0,429      0,389    0,945     0,948     0
                 0,000    0,000    ?          0,000    ?          ?        0,850     0,200     2
                 0,333    0,667    0,077      0,333    0,125      -0,240   0,481     0,155     3
                 0,750    0,059    0,750      0,750    0,750      0,691    0,956     0,813     4
                 0,000    0,053    0,000      0,000    0,000      -0,073   0,921     0,500     5
Weighted Avg.    0,333    0,111    ?          0,333    ?          ?        0,874     0,731

=== Confusion Matrix ===

 a b c d e   <-- classified as
 3 0 8 0 0 | a = 0
 0 0 1 0 0 | b = 2
 0 0 1 1 1 | c = 3
 0 0 1 3 0 | d = 4
 0 0 2 0 0 | e = 5
```

KNN 3 - on time taken



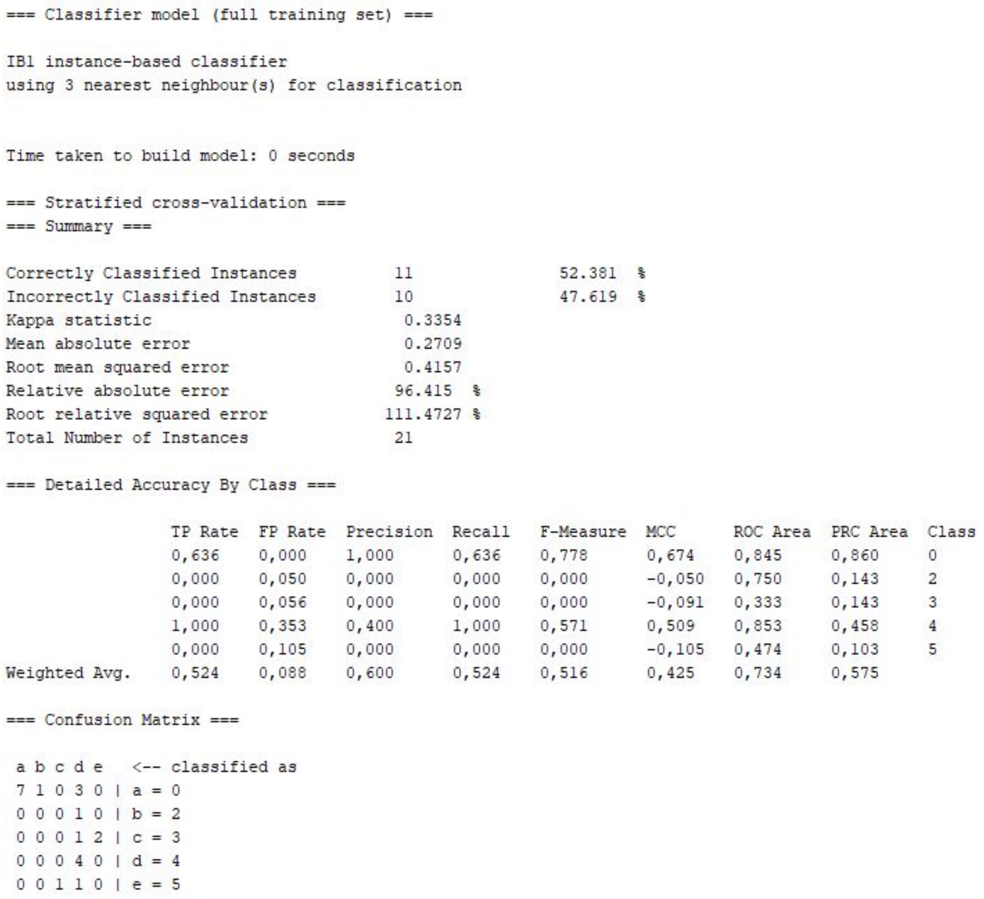
```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           4               19.0476 %
Incorrectly Classified Instances        17               80.9524 %
Kappa statistic                         -0.0469
Mean absolute error                      0.3235
Root mean squared error                  0.4632
Relative absolute error                115.1092 %
Root relative squared error            124.2323 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,273    0,200    0,600      0,273    0,375      0,085    0,632     0,639     0
                 0,000    0,000    ?          0,000    ?          ?        0,950     0,333     2
                 0,333    0,556    0,091      0,333    0,143      -0,156   0,278     0,158     3
                 0,000    0,059    0,000      0,000    0,000      -0,108   0,632     0,345     4
                 0,000    0,211    0,000      0,000    0,000      -0,157   0,605     0,154     5
Weighted Avg.    0,190    0,215    ?          0,190    ?          ?        0,594     0,453

=== Confusion Matrix ===

 a b c d e   <-- classified as
 3 0 5 0 3 | a = 0
 1 0 0 0 0 | b = 2
 0 0 1 1 1 | c = 3
 0 0 4 0 0 | d = 4
 1 0 1 0 0 | e = 5
```

KNN 3 - on number of views



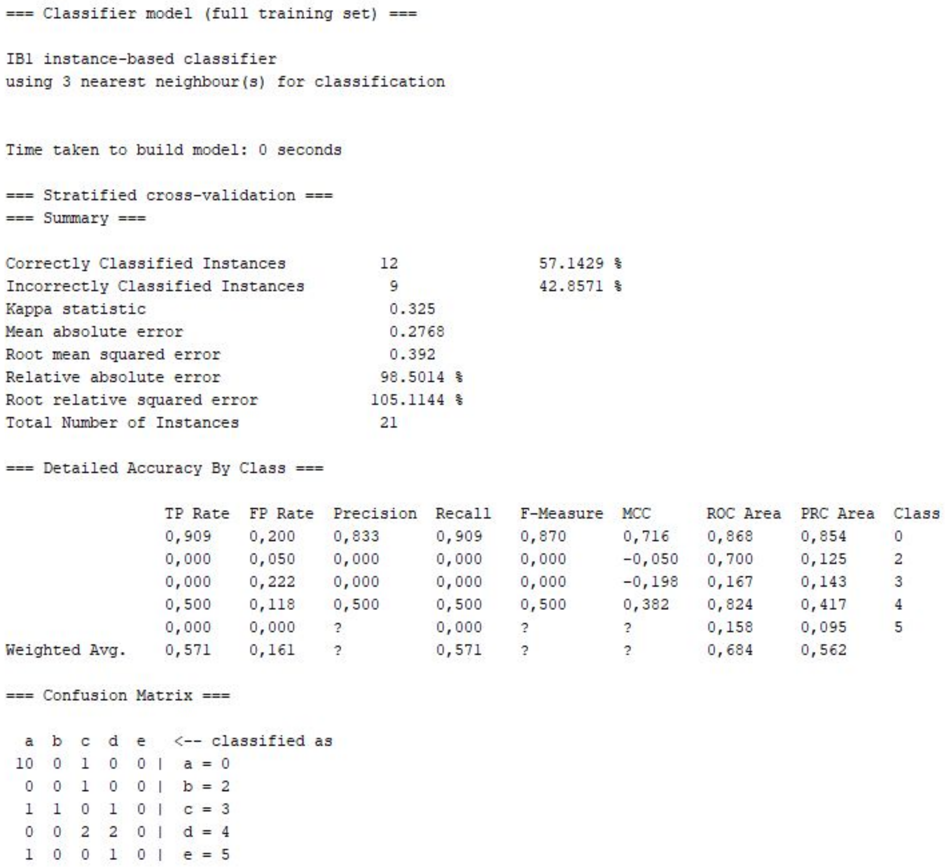
```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           7               33.3333 %
Incorrectly Classified Instances        14               66.6667 %
Kappa statistic                          0.0954
Mean absolute error                      0.2943
Root mean squared error                  0.4242
Relative absolute error                104.7143 %
Root relative squared error            113.774  %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,455    0,100    0,833      0,455    0,588      0,392    0,777     0,745     0
                 0,000    0,000    ?          0,000    ?          ?        0,850     0,200     2
                 0,667    0,444    0,200      0,667    0,308      0,156    0,556     0,181     3
                 0,000    0,294    0,000      0,000    0,000      -0,271   0,559     0,286     4
                 0,000    0,000    ?          0,000    ?          ?        0,053     0,095     5
Weighted Avg.    0,333    0,172    ?          0,333    ?          ?        0,638     0,489

=== Confusion Matrix ===

 a b c d e   <-- classified as
 5 0 3 3 0 | a = 0
 0 0 0 1 0 | b = 2
 0 0 2 1 0 | c = 3
 0 0 4 0 0 | d = 4
 1 0 1 0 0 | e = 5
```

KNN 3 - on acceptable clicks



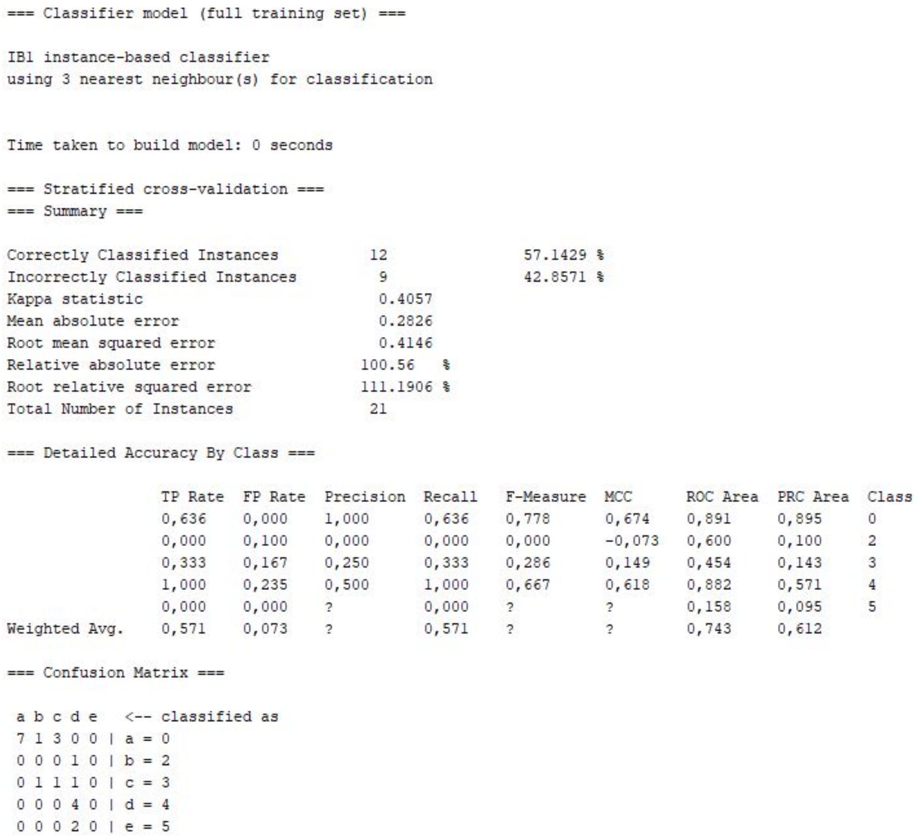
```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          11               52.381  %
Incorrectly Classified Instances        10               47.619  %
Kappa statistic                          0.3354
Mean absolute error                      0.2709
Root mean squared error                  0.4157
Relative absolute error                 96.415  %
Root relative squared error            111.4727 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,636    0,000    1,000      0,636    0,778      0,674    0,845     0,860     0
                 0,000    0,050    0,000      0,000    0,000      -0,050   0,750     0,143     2
                 0,000    0,056    0,000      0,000    0,000      -0,091   0,333     0,143     3
                 1,000    0,353    0,400      1,000    0,571      0,509    0,853     0,458     4
                 0,000    0,105    0,000      0,000    0,000      -0,105   0,474     0,103     5
Weighted Avg.    0,524    0,088    0,600      0,524    0,516      0,425    0,734     0,575

=== Confusion Matrix ===

 a b c d e   <-- classified as
 7 1 0 3 0 | a = 0
 0 0 0 1 0 | b = 2
 0 0 0 1 2 | c = 3
 0 0 0 4 0 | d = 4
 0 0 1 1 0 | e = 5
```

KNN 3 - on faults



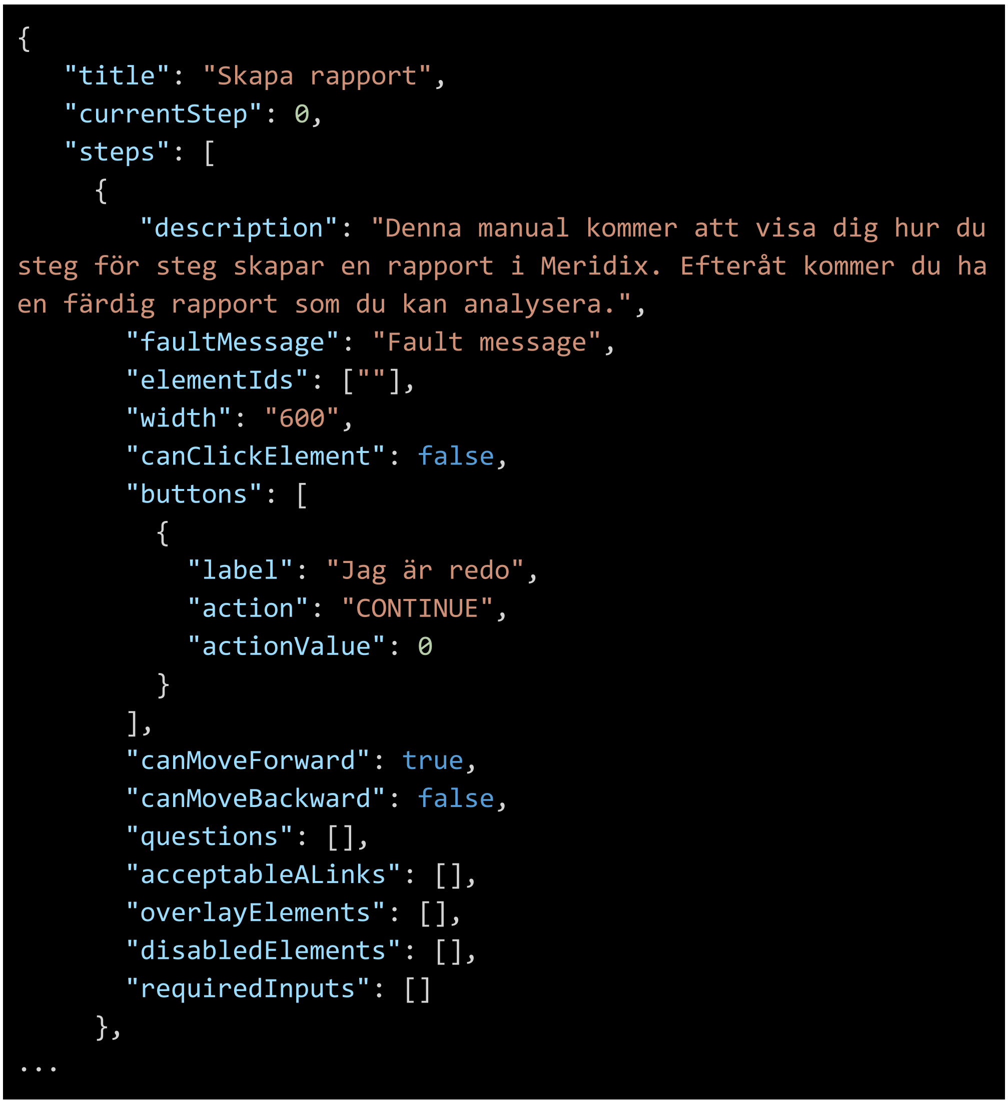
```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          12               57.1429 %
Incorrectly Classified Instances         9               42.8571 %
Kappa statistic                          0.325
Mean absolute error                      0.2768
Root mean squared error                  0.392
Relative absolute error                 98.5014 %
Root relative squared error            105.1144 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,909    0,200    0,833      0,909    0,870      0,716    0,868     0,854     0
                 0,000    0,050    0,000      0,000    0,000      -0,050   0,700     0,125     2
                 0,000    0,222    0,000      0,000    0,000      -0,198   0,167     0,143     3
                 0,500    0,118    0,500      0,500    0,500      0,382    0,824     0,417     4
                 0,000    0,000    ?          0,000    ?          ?        0,158     0,095     5
Weighted Avg.    0,571    0,161    ?          0,571    ?          ?        0,684     0,562

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 10  0  1  0  0 |  a = 0
  0  0  1  0  0 |  b = 2
  1  1  0  1  0 |  c = 3
  0  0  2  2  0 |  d = 4
  1  0  0  1  0 |  e = 5
```

KNN 3 - on faults and acceptable clicks



```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          12               57.1429 %
Incorrectly Classified Instances         9               42.8571 %
Kappa statistic                          0.4057
Mean absolute error                      0.2826
Root mean squared error                  0.4146
Relative absolute error                100.56   %
Root relative squared error            111.1906 %
Total Number of Instances               21

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,636    0,000    1,000      0,636    0,778      0,674    0,891     0,895     0
                 0,000    0,100    0,000      0,000    0,000      -0,073   0,600     0,100     2
                 0,333    0,167    0,250      0,333    0,286      0,149    0,454     0,143     3
                 1,000    0,235    0,500      1,000    0,667      0,618    0,882     0,571     4
                 0,000    0,000    ?          0,000    ?          ?        0,158     0,095     5
Weighted Avg.    0,571    0,073    ?          0,571    ?          ?        0,743     0,612

=== Confusion Matrix ===

 a b c d e   <-- classified as
 7 1 3 0 0 | a = 0
 0 0 0 1 0 | b = 2
 0 1 1 1 0 | c = 3
 0 0 0 4 0 | d = 4
 0 0 0 2 0 | e = 5
```

Appendix C - Manual object example



```
{
   "title": "Skapa rapport",
   "currentStep": 0,
   "steps": [
     {
        "description": "Denna manual kommer att visa dig hur du
steg för steg skapar en rapport i Meridix. Efteråt kommer du ha
en färdig rapport som du kan analysera.",
       "faultMessage": "Fault message",
       "elementIds": [""],
       "width": "600",
       "canClickElement": false,
       "buttons": [
         {
           "label": "Jag är redo",
           "action": "CONTINUE",
           "actionValue": 0
         }
       ],
       "canMoveForward": true,
       "canMoveBackward": false,
       "questions": [],
       "acceptableALinks": [],
       "overlayElements": [],
       "disabledElements": [],
       "requiredInputs": []
     },
...
```