Intelligent User Assistance for Automated Data Mining Method Selection

Patrick Zschech, Richard Horn, Daniel Höschele, Christian Janiesch, Kai Heinrich

Business & Information Systems Engineering (2020)

Appendix (available online via http://link.springer.com)

Appendix A: Requirements Engineering

This appendix details the practical rationale behind our design requirements, which represent meta-requirements for the design of the TbIAS. The requirements primarily derive from exchanges with practitioners and our own experience through previous research. We have grouped them according to the three meta-categories for DSS design. They all represent functional requirements and we have categorized most of them as essential (ESS) requirements. We consider R3.1 as conditional (COND) since we assume that it would improve user acceptance by limiting the restrictiveness of the result presentation. Nevertheless, we do not intend to extend our research into the areas of explainable artificial intelligence and human-computer interaction, both of which deal with perceptions and self-understanding of users with selected aspects of machine autonomy. All requirements have been formulated according to the guidelines of the SOPHIST Group (Rupp 2014). Table A1 provides a summary and also includes the type of system activity of the requirement. Type 1 comprises autonomous system activities, type 2 comprises user interaction, and type 3 comprises interface requirements – in this case the processing of the user input.

ID	(Meta-)Requirement	Category	Type
R1	Increase decision quality by providing advice with high advice quality	ESS	n/a
R1.1	The system shall select DM methods with a higher accuracy than guessing	ESS	1
R1.2	The system shall be able to remove noise from user inputs	ESS	3
R2	Reduce human decision maker's cognitive effort by providing decision support	ESS	n/a
R2.1	The system shall provide the user with the ability to enter natural-language and domain-specific text	ESS	2
R2.2	The system shall be able to extract context and central constructs from user inputs	ESS	3
R3	Minimize system restrictiveness by allowing users to control the strategy selection	ESS	n/a
R3.1	The system should provide the user with the ability to review transparent assessment scores for DM method selection	COND	2
R3.2	The system shall be able to operate on small amounts of text	ESS	3

Table A1: Requirements for the development of TbIAS

Furthering these functional requirements, we consider the following non-functional requirements important from a user acceptance and a practical perspective. The system should be interactive, that is the processing of requests should be online rather than batch processing. The system should be platform-independent with the purpose to provide researchers and practitioners an implementation, which is not restricted to be run on a certain operating system. Finally, the system should be extensible to ensure that new methods can easily be integrated into the existing prototype. We consciously decided not to include the non-functional requirements into our meta-requirements but consider them separately as advice for the practical applicability of any instantiation.

Appendix B: Design Rationale

Table B1 summarizes the design rationale behind our design principles.

No.	Design Principle	Reason	Justification	Alternatives	Trade-offs
1	Provide the system with the functionality to process natural-language user requests automatically and in their entirety in order for the system to assist novice users in DM method selection.	Users shall be able to interact with the system in natural language.	Using natural language rather than technical terms requires less expert knowledge as a prerequisite and reduces knowledge prerequisites.	Rule-based expert systems, question answering systems.	Text analysis can be ambiguous and generating quality advice is a complex problem. Yet, requiring the user to adapt to the system opens up an even less controllable venue for errors.
2	Provide the system with the functionality to extract the embedded context from the system's learning base automatically in order for the system to recognize and discriminate user requests regardless of their locale.	Users shall not be required to translate their domain-specific request into a domain-independent request.	Translation of requests requires technical knowledge and restricts the accessibility by novices.	Domain-independent formulation of user requests.	While domain-independent user request may be easier to analyze and may even result in higher advice quality, the translation of domain-specific requests itself is a source of errors, which may lead to lower quality advice. This is true in particular if the translation is performed by DM novices that are unaware of the relevant technical terms.
3	Provide the system with the functionality to construct the learning base automatically in order for the system to be economically feasible and exhibit adequate degrees of freedom.	A sufficiently large learning base is necessary to allow for an adequate degree of freedom when providing advice.	(Labelled) learning data is not publicly available and cannot be manually modeled in an economical manner.	Ignore practical economic considerations for the sake of a scientifically optimal yet impractical solution design.	A comprehensive, user-curated learning base may provide the best results, yet due to the low availability of learning data and the cost of manual labeling, a more economical approach is more suitable.

Table B1: Design rationale

Appendix C: Exemplary Sentences for Dataset Generation

The following table contains examples for each data preparation step.

Preparation	Type	Example Sentence
Syntactic Clean (RICH)	Original	“In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set (Kriegel, 2012). There are over 100 different algorithms, which aim to find k different cluster.”
	Cleaned	“in density-based clustering clusters are defined as areas of higher density than the remainder of the data set. there are over <unknown> different algorithms which aim to find different cluster.”
Semantic Clean (DEF)	Kept	“(...) consequently one possible method for solving this same problem with completelink clustering is to perform a factor analysis eliminating variables in the usual manner and then carry out a cluster analysis on the remaining items.”
	Removed	“with this type of qualitative and exploratory research the aim was to understand the relationship between gold label clusters and whether interclustering is or could come to be a competitive advantage at the regional and national level and have an impact on the internal and external economy”
Markov Model Augmentation (AUG)	Input	“Clustering can...”
	Output	“Clustering can be employed to identify subgroups of genes according to the ith group”
Synonym Replacement Augmentation (AUG)	Original	“The main purpose of clustering is to find the structure of unlabeled data”
	Created	“The main determination of clustering is to recover the structure of unlabeled observations”

Table C1: Exemplary sentences for each data preparation step

Appendix D: Training and Evaluation of Embedding Models

In this appendix, we provide further information on the training and evaluation of different FastText and DAN configurations. To assess the quality of the embedding models, there are basically two types of evaluation. On the one hand, there is an extrinsic evaluation, which takes into account the performance of embeddings as inputs on certain downstream tasks, such as the accuracy in text classification. On the other hand, there is an intrinsic evaluation, which focuses on semantic and syntactic relations within the embeddings without applying them in subsequent downstream tasks (Schnabel et al. 2015). In this study, we took an extrinsic approach by evaluating the quality of alternative embedding models on the validation dataset. The assessment was based on a simple linear kernel SVM classifier and the resulting F1-score (F1) as the accuracy measure, which is defined by the harmonic mean of precision and recall (Sokolova and Lapalme 2009). For the training of the DAN models, we used the top five FastText models as they require word vectors as inputs. During the study, a grid search approach was applied to identify the best set of hyperparameters. Table D1 shows the top five results of both embedding models.

FastText Models	**Epochs**	**Window size**	**Minimal count**	**Hierarchical softmax**	**F1 (SVM)**
FT31	10	10	5	0	**0.700**
FT12	5	10	3	0	0.667
FT1	5	3	5	0	0.667
FT25	10	5	5	0	0.667
FT13	5	10	5	0	0.633
DAN Models	**Epochs**	**Dropout**	**Shape**	**Hidden layer**	**F1 (SVM)**
DAN436 on FT12	50	0.6	[100, 50]	5	**0.833**
DAN279 on FT31	50	0.6	[200]	4	0.800
DAN246 on FT12	10	0.4	[200]	4	0.800
DAN411 on FT12	10	0.5	[100, 50]	5	0.783
DAN47 on FT12	10	0.4	[100, 50]	2	0.783

Table D1: Training results of the embedding models

Appendix E: Training and Evaluation of Classifiers based on Keyword Extractions and Topic Models

Keyword extractions: For the training of the Topic-KNN classifier, we extracted word collections from the DEF dataset due to its definition-like content and then transformed them to vector representations via the embedding models. Based on these vectors, we used a one-nearest-neighbor approach with a cosine similarity as distance measure to map an arbitrary input sentence to the most similar topic. Additionally, we also examined the combinability of different keyword extractions to identify the best possible set of word collections. Specifically, we applied the following six algorithms: YAKE! (Campos et al. 2018), TF-IDF (Salton and Buckley 1988), TF-IGM (Chen et al. 2016), TextRank (Mihalcea and Tarau 2004), PositionRank (Florescu and Caragea 2017), and TopicRank (Bougouin et al. 2013).

For the evaluation of the keyword extraction algorithms, we applied the WordNet-based coherence measures by Leacock et al. (1998) and Wu and Palmer (1994) as well as the coherence with vector similarities based on embeddings inferred from our FastText model. By aggregating the different coherence measures per topic and algorithm, we observed that combining multiple keyword extraction algorithms yield better results than only focusing on a single one. As such, we considered all 63 possible combinations and searched for the highest scores (cf. Figure E1). Thus, the combination of TF-IGM, TextRank and TopicRank achieves a summed coherence score of > 2.5 per topic, while TF-IGM in single usage, for example, only results in a score of about 1.5 per topic. We integrated the best performing combination within the final Topic-KNN classifier.

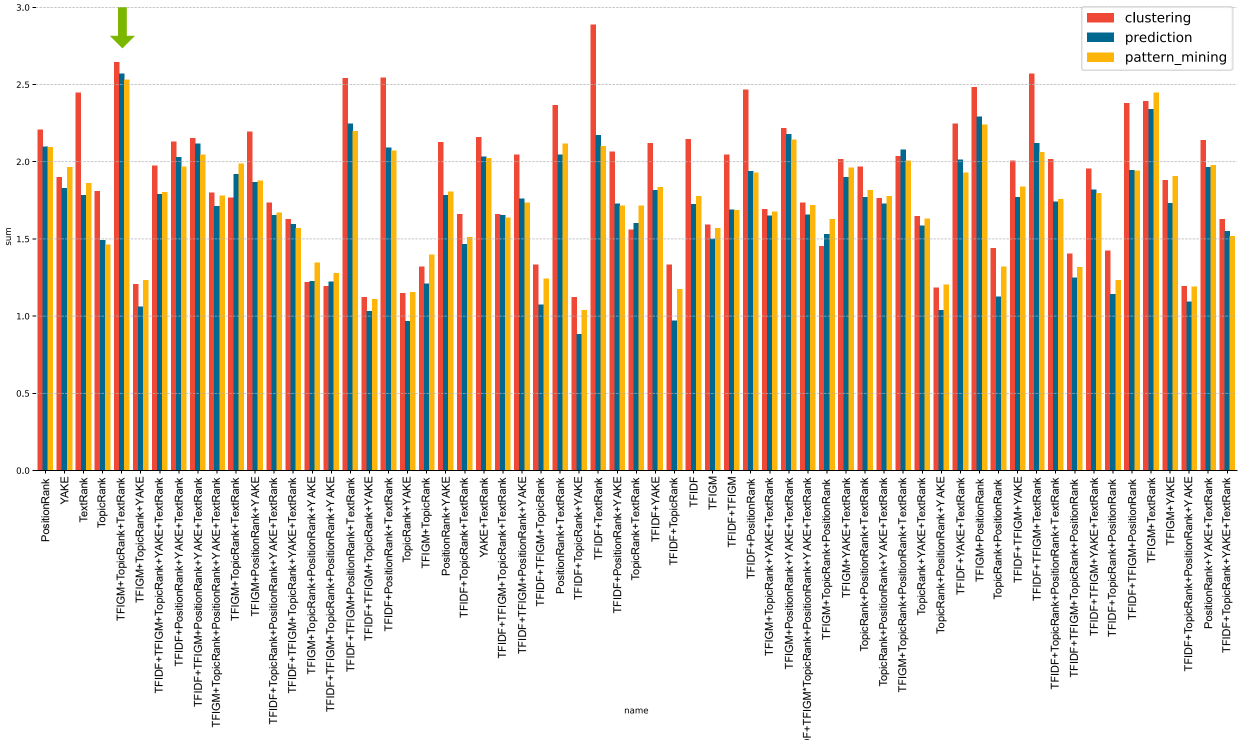
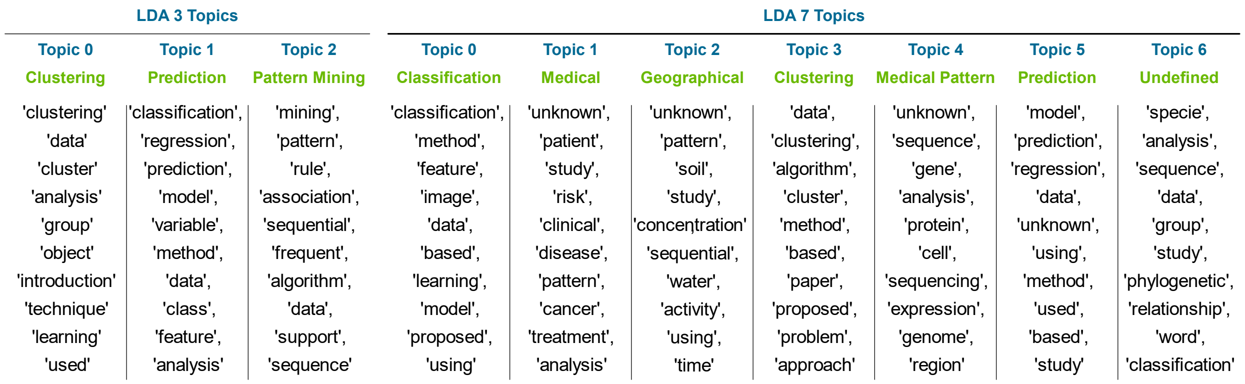


Figure E1: Evaluation results from keyword extractors

Topic models: For building the two classifiers based on LDA, we first trained different topic models and then used the derived topic probability distributions for classification purposes. The classification could be carried out in two ways: i) by directly classifying the documents with the respective topic distribution, as long as the number of topics matches the number of target classes, or ii) by using a conventional classifier as an aggregator that uses the probability scores of each given topic as inputs for predicting the target class. Therefore, we trained a first LDA model with only three topics on the DEF dataset and a second model with a variable number of topics on the RICH dataset, which was connected to an SVM classifier.

For the evaluation of the LDA models, we measured their perplexity to assess how well a probability model can predict a sample. A lower score indicates better generalizability (Blei et al. 2003). Due to various hyperparameters, we trained a total of 100 variants for the first model based on the DEF dataset with a constant number of three topics, and a total of 600 variants for the second approach based on the RICH dataset with a flexible number of topics within a range of 2-8 topics. The optimal hyperparameters were then identified using a grid search, while the best performing models reached a perplexity score of 347.48 for the first approach, and a score of 2,184.47 for the second with an optimal number of seven topics. Figure E2 illustrates the extracted topics for both models. Note that the model with three topics did not only show a lower perplexity score but also achieved good results in encapsulating the target classes, while the other model with seven topics starts to capsule the topics according to domain specific subjects.



LDA 3 Topics			LDA 7 Topics						
Topic 0	Topic 1	Topic 2	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Clustering	Prediction	Pattern Mining	Classification	Medical	Geographical	Clustering	Medical Pattern	Prediction	Undefined
'clustering'	'classification',	'mining',	'classification',	'unknown',	'unknown',	'data',	'unknown',	'model',	'specie',
'data'	'regression',	'pattern',	'method',	'patient',	'pattern',	'clustering',	'sequence',	'prediction',	'analysis',
'cluster'	'prediction',	'rule',	'feature',	'study',	'soil',	'algorithm',	'gene',	'regression',	'sequence',
'analysis'	'model',	'association',	'image',	'risk',	'study',	'cluster',	'analysis',	'data',	'data',
'group'	'variable',	'sequential',	'data',	'clinical',	'concentration'	'method',	'protein',	'unknown',	'group',
'object'	'method',	'frequent',	'based',	'disease',	'sequential',	'based',	'cell',	'using',	'study',
'introduction'	'data',	'algorithm',	'learning',	'pattern',	'water',	'paper',	'sequencing',	'method',	'phylogenetic',
'technique'	'class',	'data',	'model',	'cancer',	'activity',	'proposed',	'expression',	'used',	'relationship',
'learning'	'feature',	'support',	'proposed',	'treatment',	'using',	'problem',	'genome',	'based',	'word',
'used'	'analysis'	'sequence'	'using'	'analysis'	'time'	'approach'	'region'	'study'	'classification'

Figure E2: Extracted topics from LDA models trained on the DEF dataset (left-hand side) and on the RICH dataset (right-hand side)

Appendix F: Technical Implementation of the Prototype

The architecture of the prototype is based on the model-view-controller design pattern (Gamma et al. 1995). Overall, the backend of the prototype is developed in *Python* in an object-oriented manner and is divided into the two packages *helper* and *core*. The first package mainly includes methods and classes which were used for training and evaluating the different methods described during the artifact's development. The second package contains all keyword extraction, embedding and classification models as pure *Python* classes. We implemented an interface as well as a facade pattern to nest the complexity in the according classes. As depicted in Figure F1, the model class defines a template for each classification model in the implemented system. With the help of a complex data type, we standardized the output of each model as prediction and combined those in the recommendation facade. The facade masks the complexity of the process and helps to adjust the final recommendation pipeline easily. For the implementation, we used *Flask* as our basic framework in combination with *Materialize*, *JavaScript* and *d3.js* to realize the frontend and build a bridge between frontend and backend. Moreover, the implementation offers the possibility to download the example analysis in a *Jupyter Notebook*[1].

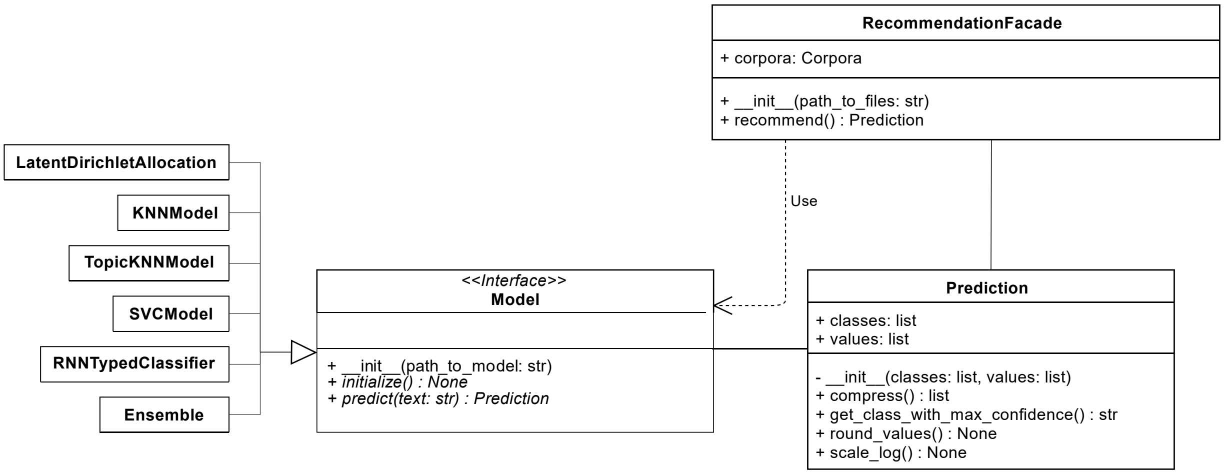


Figure F1: Excerpt of the implemented system's class structure

[1] https://jupyter.org/

Appendix G: Overview of Problem Descriptions and Evaluation Results

ID	Problem Descriptions	#W	Novice Assessment				TbIAS Baseline Configuration			TbIAS Full Configuration		
			CL	PR	FPM	NA	CL	PR	FPM	CL	PR	FPM
CL1	We want to divide our variety of coffee beans into different bundles to offer better products based on the extracted types.	21	**65**	25	5	5	17.6	9.8	**72.6**	**71.3**	26.2	2.6
CL2	I want to spot workout sessions that are similar to each other to have a better understanding of my overall performance and get new ideas on how to improve.	29	**35**	30	0	**35**	14.9	**76**	9.1	**77.0**	14.6	8.4
CL3	Our company has a variety of different customers. Based on the costumer characteristics as well of their revenues, different customer groups should be found.	24	**80**	15	0	5	46.2	1.4	**52.5**	**56.9**	22.4	20.6
CL4	The goal of the project is to clump different friendships together in similar bunches based on social media data.	19	**70**	5	15	10	**99.4**	0.2	0.4	**97.7**	1.2	1.1
CL5	We have a large number of machines to manufacture our main product and now we want to find out whether there are common groups of configuration profiles based on the many configuration parameters which are mostly set subjectively by our machine operators.	42	**50**	25	5	20	22.8	33.3	**44**	**55.3**	39.4	5.2
CL6	In this project, you are given a set of 100-dimensional data points from our machine log and you are asked to discover the hidden structure behind them based on their similarity.	31	**50**	25	5	20	11.5	0.7	**87.8**	**84.1**	5.6	10.2
CL7	We want to rearrange our staff into different divisions by bringing together colleagues with the same interests and attitudes.	19	**85**	10	5	0	4.5	1.7	**93.8**	24.7	**58.8**	16.5
CL8	I want to arrange my machines according to their energy consumption. Machines with similar energy consumption should be placed together. The number of resulting machine agglomerates does not matter.	29	**55**	30	5	10	20.9	**63.3**	15.7	**76.1**	12.4	11.5
CL9	Based on the activity data of our different employees, we want to find similar groups of activity levels, to offer new fitness opportunities to our employees.	26	**85**	5	0	10	0	0	**100**	3.2	1.8	**95.1**
CL10	The purpose of this project is to organize our supply chain partners of the automobile industry into different segments according to different partnership attributes, such as perceived quality, intensity of cooperation, etc.	32	**55**	35	10	0	24.1	**58.6**	17.2	**67.5**	17.4	15.1
CL11	Our agricultural data provides images of fruit trees that we want to organize into different classes based on their visual properties	21	**65**	30	5	0	25	**59**	16	27.0	**72.3**	0.8
CL12	We would like to group certain types of patients based on properties like length of stay, diet and medical condition.	20	**80**	10	5	5	1.2	**73.3**	25.5	**45.3**	21.4	33.3
CL13	Since the data we receive from our physical experiments are very heterogeneous, we want to build stable groups of experiments based on single experiment data that help us understand common properties and how to better prepare for certain types of experiments.	41	**50**	0	20	30	27.7	**59.6**	12.7	**41.5**	39.7	18.8
CL14	We want to organize our salesmen into different formations depending on a multitude of attributes, such as performance, willingness to travel, know-how, etc.	23	**60**	25	5	10	**70.9**	9.2	19.9	**48.9**	34.7	16.4
CL15	We would like to have an overview over common gene types based on single gene data. We would like to yield no more than four segments of gene types that are grouped based on their expressions.	36	45	30	5	20	**95.8**	1.4	2.8	**49.4**	6.5	47.7

CL16	In recent years, we have accumulated a lot of data about different types of wine. We would like to see whether an algorithm can come up with the same groups that experts choose for wines when considering different attributes of a wine.	42	**70**	25	5	0	**63.8**	28.4	7.8	**51.2**	31.3	17.5
CL17	I want to divide my harvested olives in different types based on similar colors and shapes. The resulting profiles should help to choose the right treatment for better oil quality.	30	**60**	25	15	0	3.1	**95.3**	1.5	**51.6**	47.2	1.3
CL18	I want to extract a set of colors from an image, where each set contains similar colors. This way I can have different color schemes.	25	**60**	5	0	35	42.5	12.6	**44.9**	**74.9**	18.8	6.3
CL19	In this competition, you will use a more realistic acoustic data set. You are provided with three different subsets. The purpose of this analysis is to group similar sound together. The golden number of groups that should be found is around 4-5 clumps.	43	**45**	25	5	25	39.8	1.2	**59**	**56.1**	26.7	17.2
CL20	The analysis should subdivide all students into several intellectual groups to determine which students need more advice.	17	**70**	30	0	0	0.1	0	**99.8**	**92.7**	4.7	2.6
PR1	In this competition, you will develop models capable of classifying mixed patterns of proteins in microscope images.	17	**35**	30	15	20	1.2	**88.8**	10	1.2	**97.9**	0.9
PR2	We are a company specialized in implementing customized e-mail clients. We want to predict or classify if an incoming email is spam or not.	24	5	**75**	5	15	10.2	**51.5**	38.3	6.8	**55.4**	37.8
PR3	In terms of a new product introduction, the price of a product should be predicted based on different characteristics like the size, equipment features or the buying power of the area.	31	20	**35**	15	30	1.7	**96.8**	1.6	5.6	**89.0**	5.4
PR4	The analysis should evaluate the relationship between the weather and the amount of sold ice cream. More specific, it should be examined how much the temperature influences the revenue of the organization.	32	0	**65**	30	5	18.2	**59.4**	22.4	7.8	**86.7**	5.5
PR5	The XY-Company is specialized on implementing different document management systems. To offer our customers a new feature, we want to forecast the type of our documents. Because of the available huge amount of data, we aim to use new machine learning, to make those kinds of prediction.	47	10	30	10	**50**	6.4	**82.7**	10.9	7.4	**86.7**	5.9
PR6	We are looking for the future returns and would like to know whether the price goes up or down the next day.	22	0	**50**	30	20	6.7	**87**	6.3	21.7	**63.6**	14.7
PR7	We would like to classify different images of plants into categories and predict the plant type solely from the image. Later on we can provide data to annotate the prediction with features like sepal length or sepal width.	38	35	**45**	5	15	0	**99.9**	0	1.0	**97.9**	1.0
PR8	We imagine a solution where we can anticipate the return on investment for a project based on the project properties.	20	10	**40**	25	25	0.8	33.1	**66.2**	13.2	**74.4**	12.5
PR9	For our project, we would like to see an estimate of the number of patients arriving at the hospital given the weekday and other circumstances like weather or special events taking place in that region (e.g. rock festivals).	38	15	**50**	25	10	12.5	**80.5**	7	16.0	**54.6**	29.4
PR10	Forecasting the number of students that will attend our class based on past experiences should be the main priority within the project.	22	0	**40**	**40**	20	1.1	1.8	**97.2**	1.6	**88.9**	9.5

PR11	The aim of the analysis is to find out whether the price of each player in the FIFA Ultimate Market will rise or fall in the light of the performance of recent matches.	33	0	**50**	35	15	21.4	**66.4**	12.2	14.8	**55.7**	29.6
PR12	We need a good estimation model for the quality of our manufactured goods by taking into account the health condition of our plant’s machinery.	24	5	**85**	5	5	5.3	**93.8**	0.9	25.0	**74.2**	0.8
PR13	In our project, we want to find out whether there is a connection between the features and health conditions of a car and the number and intensity of involved traffic accidents. Particularly, we are interested in the impact of the presence of individual electronic system components on the observed damage level.	51	5	30	**50**	15	0.8	**94.9**	4.3	6.0	**74.8**	19.2
PR14	You are given over 65,000 games worth of anonymized players, split into training and out-of-sample data and asked to determine final placement from final in-game stats and initial player ratings.	30	10	**50**	10	30	10.9	28.3	**60.9**	10.6	**85.6**	3.9
PR15	In this competition, you have to anticipate the adoptability of pets - specifically, how quickly will a pet be adopted using visual attributes?	23	10	**50**	25	15	0.4	**92.9**	6.8	3.2	**71.2**	25.6
PR16	We need a model that provides information on how potential changes in our production parameters will affect our plant’s productivity.	20	0	**70**	5	25	0.5	26.9	**72.6**	6.2	**82.5**	11.3
PR17	Given an arbitrary text of the world library, the software should determine the author who wrote the book.	18	5	**45**	15	35	40.5	17.9	**41.7**	1.6	**97.0**	1.4
PR18	The objective is to build a discriminatory model for our new logistics center which is able to proactively distinguish between our main product categories based on different shipping characteristics, such as size, weight, shape, etc.	35	40	**50**	0	10	**75.5**	18.9	5.7	3.2	**92.8**	4.1
PR19	The project describes the problem of recognizing objects in images. The main objective is to detect different car brands in one picture.	22	35	**40**	5	20	7.9	21.9	**70.2**	19.2	**75.0**	5.8
PR20	I want to know which clients are most likely to leave our company in the future.	16	15	30	**50**	5	5.6	29.1	**65.4**	23.5	**65.1**	11.5
AR1	As a shop for digital articles, we want to find out which products are bought together in which combination and in which sequence. We want to use the information to promote the most promising products better.	36	25	10	**55**	10	0.4	3.8	**95.8**	10.9	**48.1**	41.0
AR2	Based on the communication data of our customers, we want to find similar patterns in their communication behavior.	18	20	0	**75**	5	0	0	**100**	1.4	0.8	**97.8**
AR3	We want to analyze retail basket or transaction data and are intended to identify strong rules discovered in transaction data using data of our customers.	25	5	20	25	**50**	0	0	**100**	0.8	0.8	**98.4**
AR4	The last 2 years the XX-Company collected data of their machine errors. Based on those sequential data, an analysis should be performed to find different frequent error patterns.	28	10	10	**80**	0	0	0.1	**99.8**	1.4	3.9	**94.7**
AR5	As an emerging digital company, we are facing a lot of new competitors in the market. Our recent market analysis is showing some interesting data about the marketing strategies of our competitors. The ongoing project should find market strategies in terms of sequential actions, to outrival our competitors.	48	20	10	30	**40**	0	0	**100**	5.6	**48.0**	46.4
AR6	We would like to implement the amazon systems for our customers to boost cross selling – so we want to promote products for each customer individuality in the form: “customers who bought product A were also interested in product B”.	40	15	15	**70**	0	0.3	2.4	**97.3**	2.4	2.3	**95.3**

AR7	As a credit card company, we want to mine transaction records for intercontinental purchase co-occurrences of consumers to spot fraudulent use of credit cards.	24	5	5	**55**	35	0.2	1.8	**98**	0.8	0.9	**98.3**
AR8	We run a large server landscape and we want to extract rules about which malfunctions often occur together in the different systems.	22	20	30	**45**	5	0	0	**100**	1.5	17.5	**81.0**
AR9	My administrator has the feeling that whenever our processing server reaches a high number of requests, our frontend server also throws a high number of exception handlings. We want to know whether this is true.	35	0	**35**	**35**	30	13.8	39.7	**46.5**	40.2	17.2	**42.6**
AR10	We have the impression that customers who visit our photo gallery on our homepage also visit our web shop. We want to investigate this phenomenon with real data.	28	5	15	**65**	15	0.2	3.1	**96.7**	36.3	15.6	**48.1**
AR11	For our maintenance department, we search for interpretable rules in the form of “if machine A fails, then B and C fail as well”.	24	0	25	**75**	0	1.9	9.5	**88.6**	4.7	42.4	**53.0**
AR12	Based on the market basket of our customers, the current tasks aim to uncover associations between items. Those items should be offered the customers as extra product.	27	20	10	**45**	25	0	0	**100**	0.8	0.8	**98.4**
AR13	As the marketing team of a financial services provider, we are interested in intensifying the use of cross-selling. This requires deeper insights into entry-level products, follow-up purchases and chains of product deals. The goal is the automatic generation of customer-specific product recommendations.	42	15	15	**35**	**35**	16.4	40.6	**43**	4.8	43.0	**52.3**
AR14	We want to confirm that clients who are interested in our physical products are also interested in our after sales services.	21	10	30	**50**	10	0	0.4	**99.6**	10.8	9.3	**79.9**
AR15	The analysis should deal with the question of whether certain vehicle deficiencies are typically found in combination with each other, and whether this may occur more frequently with certain vehicle types.	31	5	10	**65**	20	0	0.1	**99.9**	29.7	16.2	**54.1**
AR16	The objective of the project is to find patterns and structures based on the behavior of our employees. All relevant data are stored as sequences. The goal is to use those extracted patterns to increase the overall efficiency of the corresponding department.	42	0	5	**70**	25	0	0	**100**	1.5	1.4	**97.1**
AR17	Our research goal is to find common word patterns in the abstract of scientific articles from the domain of health care information systems. We are specifically interested in co-occurring words that describe research methodology.	34	15	15	**50**	20	7.3	34.4	**58.4**	15.5	11.9	**72.6**
AR18	As emerging company in the energy sector, we want to capture weak signals of potentially threatening events to identify connections between error events and other occurring events in our energy process pipeline.	32	0	15	**70**	15	3.5	66.5	**30**	25.5	**62.0**	12.6
AR19	Our customers frequently engage in buying different types of products. We want to recommend combinations of products for customers based on their history.	23	35	15	**45**	5	0	0.1	**99.9**	0.8	0.9	**98.3**
AR20	For our restaurant's buffet, we are interested in the combinations, which dishes and drinks are often consumed together.	18	10	30	**60**	0	0	0.2	**99.8**	5.7	2.2	**92.1**

Table G1: Problem descriptions and evaluation results

Appendix H: Questionnaire

Please fill out the following fields.

Course of studies:

____

Semester:

____

Previously attended courses at the chair:

____

____

Self-assessment of Data Mining know-how:

☐ low (I have heard about the field of Data Mining, but do not know any method)

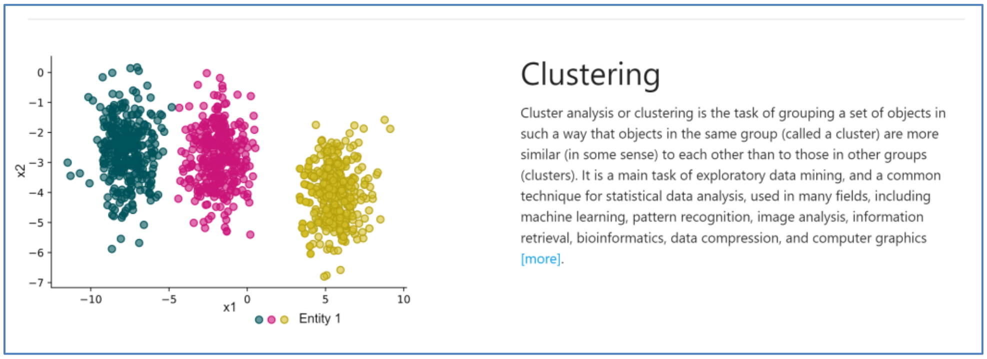
☐ medium (I have basic knowledge in Data Mining and know some methods)

☐ high (I am well aware of a wide variety of Data Mining methods)

In the following, we will provide 60 problem descriptions that can be addressed with different classes of Data Mining methods. In this context, we would like to ask you about your previous knowledge/ experience regarding the following three method classes. Please indicate your level of knowledge/ experience on the scale between 1-7, with 1 = "I have never heard of this method class" and 7 = "I have good experience with the method class and I have already applied it several times".

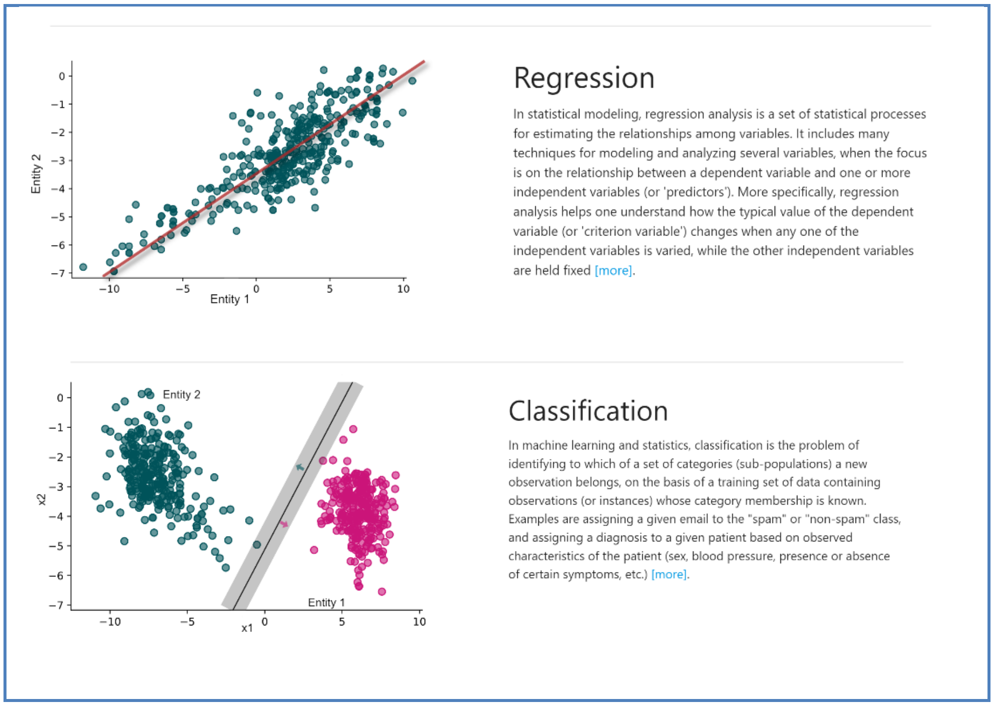
Clustering	I have never heard of this method class.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	I have good experience with the method class and have already applied it several times.
Prediction (including Regression and Classification)	I have never heard of this method class.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	I have good experience with the method class and have already applied it several times.
Frequent Pattern Mining (including Association Rules and Sequence Mining)	I have never heard of this method class.	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	I have good experience with the method class and have already applied it several times.

Please take a look at the following three classes of Data Mining methods to get an idea of how the methods work and which kind of problems can be solved by their application. Please note that "Regression" and "Classification" belong to the superior group of "Prediction", whereas the group of "(Frequent) Pattern Mining" consists of the two sub-groups "Association Rule Mining" and "Sequence Mining".



Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics [more].

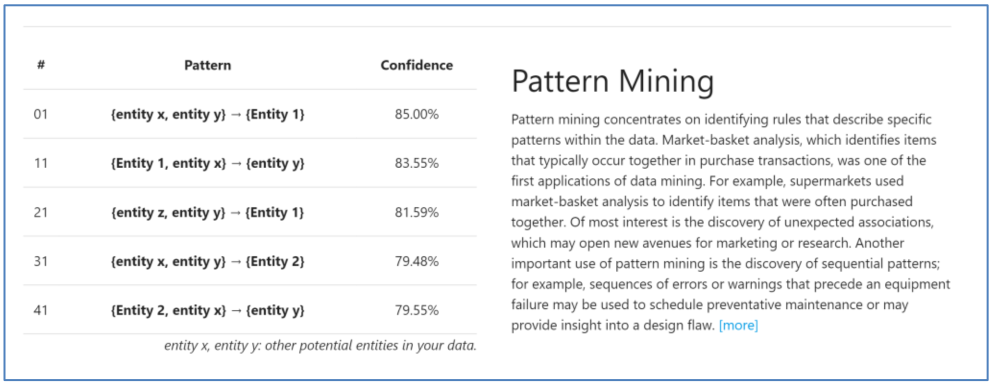


Regression

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed [more].

Classification

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Examples are assigning a given email to the "spam" or "non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.) [more].



#	Pattern	Confidence
01	**{entity x, entity y} → {Entity 1}**	85.00%
11	**{Entity 1, entity x} → {entity y}**	83.55%
21	**{entity z, entity y} → {Entity 1}**	81.59%
31	**{entity x, entity y} → {Entity 2}**	79.48%
41	**{Entity 2, entity x} → {entity y}**	79.55%

entity x, entity y: other potential entities in your data.

Pattern Mining

Pattern mining concentrates on identifying rules that describe specific patterns within the data. Market-basket analysis, which identifies items that typically occur together in purchase transactions, was one of the first applications of data mining. For example, supermarkets used market-basket analysis to identify items that were often purchased together. Of most interest is the discovery of unexpected associations, which may open new avenues for marketing or research. Another important use of pattern mining is the discovery of sequential patterns; for example, sequences of errors or warnings that precede an equipment failure may be used to schedule preventative maintenance or may provide insight into a design flaw. [more]

Now, please read carefully each problem description and select the class of Data Mining methods that you think best addresses the problem. In some cases it may seem to you that an unambiguous class assignment is not possible. So, if you cannot judge the matching well enough, please select the option "I am not sure". Please DO NOT select multiple answers. If you have too much trouble filling out the majority of the questionnaire, you can ask the staff member of the chair for the cheat sheet again with an overview about the three method classes.

ID	PROBLEM DESCRIPTION	SUITABLE DATA MINING METHOD
1	We would like to group certain types of patients based on properties like length of stay, diet and medical condition.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
2	We need a good estimation model for the quality of our manufactured goods by taking into account the health condition of our plant's machinery.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
3	As emerging company in the energy sector, we want to capture weak signals of potentially threatening events to identify connections between error events and other occurring events in our energy process pipeline.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
4	Our agricultural data provides images of fruit trees that we want to organize into different classes based on their visual properties	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
5	You are given over 65,000 games worth of anonymized players, split into training and out-of-sample data and asked to determine final placement from final in-game stats and initial player ratings.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
6	I want to know which clients are most likely to leave our company in the future.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
7	In terms of a new product introduction, the price of a product should be predicted based on different characteristics like the size, equipment features or the buying power of the area.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.

8	The XY-Company is specialized on implementing different document management systems. To offer our customers a new feature we want to forecast the type of our documents. Because of the available huge amount of data, we aim to use new machine learning, to make those kinds of prediction.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
9	In our project, we want to find out whether there is a connection between the features and health conditions of a car and the number and intensity of involved traffic accidents. Particularly, we are interested in the impact of the presence of individual electronic system components on the observed damage level.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
10	The objective of the project is to find patterns and structures based on the behavior of our employees. All relevant data are stored as sequences. The goal is to use those extracted patterns to increase the overall efficiency of the corresponding department.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
11	Our research goal is to find common word patterns in the abstract of scientific articles from the domain of health care information systems. We are specifically interested in co-occurring words that describe research methodology.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
12	For our project, we would like to see an estimate of the number of patients arriving at the hospital given the weekday and other circumstances like weather or special events taking place in that region (e.g., rock festivals).	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
13	In this project, you are given a set of 100-dimensional data points from our machine log and you are asked to discover the hidden structure behind them based on their similarity.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
14	In this competition, you have to anticipate the adoptability of pets - specifically, how quickly will a pet be adopted using visual attributes?	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.

15	In recent years, we have accumulated a lot of data about different types of wine. We would like to see whether an algorithm can come up with the same groups that experts choose for wines when considering different attributes of a wine.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
16	We would like to classify different images of plants into categories and predict the plant type solely from the image. Later on we can provide data to annotate the prediction with features like sepal length or sepal width.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
17	I want to divide my harvested olives in different types based on similar colors and shapes. The resulting profiles should help to choose the right treatment for better oil quality.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
18	We have the impression that customers who visit our photo gallery on our homepage also visit our web shop. We want to investigate this phenomenon with real data.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
19	We need a model that provides information on how potential changes in our production parameters will affect our plant's productivity.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
20	As an emerging digital company, we are facing a lot of new competitors in the market. Our recent market analysis is showing some interesting data about the marketing strategies of our competitors. The ongoing project should find market strategies in terms of sequential actions, to outrival our competitors.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
21	The aim of the analysis is to find out whether the price of each player in the FIFA Ultimate Market will rise or fall in the light of the performance of recent matches.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
22	We want to confirm that clients who are interested in our physical products are also interested in our after sales services.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
23	I want to extract a set of colors from an image, where each set contains similar colors. This way I can have different color schemes.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.

24	Our customers frequently engage in buying different types of products. We want to recommend combinations of products for customers based on their history.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
25	We have a large number of machines to manufacture our main product and now we want to find out whether there are common groups of configuration profiles based on the many configuration parameters which are mostly set subjectively by our machine operators.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
26	Based on the communication data of our customers, we want to find similar patterns in their communication behavior.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
27	As the marketing team of a financial services provider, we are interested in intensifying the use of cross-selling. This requires deeper insights into entry-level products, follow-up purchases and chains of product deals. The goal is the automatic generation of customer-specific product recommendations.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
28	We run a large server landscape and we want to extract rules about which malfunctions often occur together in the different systems.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
29	As a credit card company, we want to mine transaction records for intercontinental purchase co-occurrences of consumers to spot fraudulent use of credit cards.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
30	Based on the activity data of our different employees, we want to find similar groups of activity levels, to offer new fitness opportunities to our employees.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
31	The objective is to build a discriminatory model for our new logistics center which is able to proactively distinguish between our main product categories based on different shipping characteristics, such as size, weight, shape, etc.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.

32	The analysis should deal with the question of whether certain vehicle deficiencies are typically found in combination with each other, and whether this may occur more frequently with certain vehicle types.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
33	The analysis should subdivide all students into several intellectual groups to determine which students need more advice.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
34	We want to divide our variety of coffee beans into different bundles to offer better products based on the extracted types.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
35	My administrator has the feeling that whenever our processing server reaches a high number of requests, our frontend server also throws a high number of exception handlings. We want to know whether this is true.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
36	We are looking for the future returns and would like to know whether the price goes up or down the next day.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
37	For our maintenance department, we search for interpretable rules in the form of “if machine A fails, then B and C fail as well”.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
38	We would like to have an overview over common gene types based on single gene data. We would like to yield no more than four segments of gene types that are grouped based on their expressions.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
39	We want to rearrange our staff into different divisions by bringing together colleagues with the same interests and attitudes.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
40	We want to analyze retail basket or transaction data and are intended to identify strong rules discovered in transaction data using data of our customers.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.

41	The purpose of this project is to organize our supply chain partners of the automobile industry into different segments according to different partnership attributes, such as perceived quality, intensity of cooperation, etc.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
42	I want to arrange my machines according to their energy consumption. Machines with similar energy consumption should be placed together. The number of resulting machine agglomerates does not matter.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
43	We would like to implement the amazon systems for our customers to boost cross selling – so we want to promote products for each customer individuality in the form: "customers who bought product A were also interested in product B".	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
44	Forecasting the number of students that will attend our class based on past experiences should be the main priority within the project.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
45	The analysis should evaluate the relationship between the weather and the amount of sold ice cream. More specific, it should be examined how much the temperature influences the revenue of the organization.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
46	In this competition, you will develop models capable of classifying mixed patterns of proteins in microscope images.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
47	The project describes the problem of recognizing objects in images. The main objective is to detect different car brands in one picture.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
48	In this competition, you will use a more realistic acoustic data set. You are provided with three different subsets. The purpose of this analysis is to group similar sound together. The golden number of groups that should be found is around 4-5 clumps.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
49	We imagine a solution where we can anticipate the return on investment for a project based on the project properties.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.

50	The last 2 years the XX-Company collected data of their machine errors. Based on those sequential data, an analysis should be performed to find different frequent error patterns.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
51	Since the data we receive from our physical experiments are very heterogeneous, we want to build stable groups of experiments based on single experiment data that help us understand common properties and how to better prepare for certain types of experiments.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
52	For our restaurant's buffet, we are interested in the combinations, which dishes and drinks are often consumed together.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
53	Based on the market basket of our customers the current tasks aims to uncover associations between items. Those items should be offered the customers as extra product.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
54	We are a company specialized in implementing customized e-mail clients. We want to predict or classify if an incoming email is spam or not.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
55	As a shop for digital articles, we want to find out which products are bought together in which combination and in which sequence. We want to use the information to promote the most promising products better.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
56	The goal of the project is to clump different friendships together in similar bunches based on social media data.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
57	Our company has a variety of different customers. Based on the costumer characteristics as well of their revenues, different customer groups should be found.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
58	We want to organize our salesmen into different formations depending on a multitude of attributes, such as performance, willingness to travel, know-how, etc.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.

59	Given an arbitrary text of the world library, the software should determine the author who wrote the book.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.
60	I want to spot workout sessions that are similar to each other to have a better understanding of my overall performance and get new ideas on how to improve.	☐ Clustering ☐ Classification/Regression ☐ Frequent Pattern Mining ☐ I am not sure.

Thank you very much for your participation!

Appendix I: Robustness Check

ID	Problem Description	Modification	CL	PR	FPM
CL1	We want to **divide** our variety of coffee beans into different **bundles** to offer better products based on the extracted types.	Original text	**71.26**	26.18	2.56
	We want to **break down** our variety of coffee beans into different **collections** to offer better products based on the extracted types.	Modification with **weak** keyword replacement from **same** DM method class	**59.13**	36.47	4.41
	We want to **organize** our variety of coffee beans into different **clusters** to offer better products based on the extracted types.	Modification with **strong** keyword replacement from **same** DM method class	**98.21**	0.86	0.92
	We want to **determine** the **category** of our coffee beans to offer better products based on the extracted types.	Modification with keyword replacement from **another** DM method class	38.76	**55.08**	6.16
PR8	We imagine a solution where we can **anticipate** the return on investment for a project based on the project properties.	Original text	13.16	**74.37**	12.48
	We imagine a solution where we can **determine** the return on investment for a project based on the project properties.	Modification with **weak** keyword replacement from **same** DM method class	24.21	**67.95**	7.84
	We imagine a solution where we can **estimate** the return on investment for a project based on the project properties.	Modification with **strong** keyword replacement from **same** DM method class	8.29	**87.59**	4.12
	We imagine a solution where we can **group** the return on investment for a project based on **similar** project properties.	Modification with keyword replacement from **another** DM method class	**50.73**	41.79	7.48
AR8	We run a large server landscape and we want to extract **rules** about which malfunctions **often** **occur** **together** across the individual systems.	Original text	2.38	21.62	**76.00**
	We run a large server landscape and we want to extract **heuristics** about which malfunctions **occur together** across the individual systems.	Modification with **weak** keyword replacement from **same** DM method class	16.14	20.72	**63.15**
	We run a large server landscape and we want to extract **rules** about which malfunctions **frequently** **occur** **in combination** across the individual systems.	Modification with **strong** keyword replacement from **same** DM method class	1.12	5.14	**93.74**
	We run a large server landscape and we want to **predict upcoming** malfunctions across the individual systems.	Modification with keyword replacement from **another** DM method class	1.01	**58.74**	40.25
CL11	Our agricultural data provides images of fruit trees that we want to organize into different **classes** based on their visual properties.	Original text	26.95	**72.28**	0.78
	Our agricultural data provides images of fruit trees that we want to organize into different **segments** based on their visual properties.	Modification with keyword from the **CL class**	**51.11**	48.21	0.68
	Our agricultural data provides images of fruit trees that we want to organize into different **groups** based on their visual properties.	Modification with keyword from the **CL class**	**55.98**	43.33	0.68
	Our agricultural data provides images of fruit trees that we want to organize into different **clusters** based on their visual properties.	Modification with **strong** keyword from the **CL class**	**97.2**	1.95	0.86

Table I1: Robustness check I - replacement of keywords

ID	Problem Description	Modification	CL	PR	FPM
CL3	Our **company** has a variety of different **customers**. Based on the **costumer** characteristics as well of their **revenues**, different **customer** groups should be found.	Original text	**56.93**	22.43	20.64
	Our **business** has a variety of different **clients**. Based on the **client** characteristics as well of their **revenues**, different **client** groups should be found.	Modification with context replacement to **similar** domain entities	**67.76**	21.5	10.75
	Our **infrastructure** has a variety of different **servers**. Based on the **server** characteristics as well of their **performance**, different **server** groups should be found.	Modification with context replacement to **dissimilar** domain entities	**75.05**	23.91	1.04
	Our **online community** has a variety of different **users**. Based on the **user** characteristics as well of their **behavior**, different **user** groups should be found.	Modification with context replacement to **dissimilar** domain entities	**86.85**	8.99	4.15
	Our **institute** has a variety of different **researchers**. Based on the **researcher** characteristics as well of their **reputation**, different **researcher** groups should be found.	Modification with context replacement to **dissimilar** domain entities	**79.95**	18.93	1.12
PR4	The analysis should evaluate the relationship between the **weather** and the amount of **sold ice cream**. More specific, it should be examined how much the **temperature** influences the **revenue of the organization**.	Original text	7.44	**86.72**	5.4
	The analysis should evaluate the relationship between the **climate** and the amount of **sold sweets**. More specific, it should be examined how much the **rain** influences the **earnings of the firm**.	Modification with context replacement to **similar** domain entities	1.82	**96.64**	1.54
	The analysis should evaluate the relationship between **client characteristics** and **profitability**. More specific, it should be examined how much the **income** influences the **revenue of the firm**.	Modification with context replacement to **similar** domain entities	5.68	**52.04**	42.27
	The analysis should evaluate the relationship between the **vehicle properties** and the amount of **traffic accidents**. More specific, it should be examined how much the **vehicle size** influences the number of **injured people**.	Modification with context replacement to **dissimilar** domain entities	19.97	**73.84**	6.19
	The analysis should evaluate the relationship between **machine conditions** and the **product quality**. More specific, it should be examined how much **unhealthy machine conditions** influence the number of **rejects**.	Modification with context replacement to **dissimilar** domain entities	25.18	**72.63**	2.19
AR14	We want to confirm that **clients** who are interested in our **physical products** are also interested in our **after sales services**.	Original text	10.84	9.27	**79.9**
	We want to confirm that **customers** who are interested in **cross-selling** are also interested in **up-selling**.	Modification with context replacement to **similar** domain entities	1.32	0.86	**97.82**
	We want to confirm that **users** who are interested in our **website** are also interested in our **web gallery**.	Modification with context replacement to **dissimilar** domain entities	7.26	1.21	**91.53**
	We want to confirm that **researchers** who are interested in **mathematics** literature are also interested in **geographical documentaries**.	Modification with context replacement to **dissimilar** domain entities	13.64	5.41	**80.95**

Table I2: Robustness check II - replacement of domain entities

ID	Problem Description	Modification	Keywords/ Words	CL	PR	FPM
CL	**Clump** friendships.	Reduction to important keywords	1/2	**98.27**	0.81	0.92
	Clump different friendships **together** in **similar** **bunches**.	Reduction to the central statement	4/7	**98.04**	0.81	1.15
	The goal of the project is to **clump** different friendships **together** in **similar** **bunches** based on social media data.	Original text	4/19	**97.70**	1.21	1.09
	The goal of the project is to **clump** different friendships **together** in **similar** **bunches** based on social media data. And then we add random noise to see how robust our results are.	Enrichment with additional noise	4/32	**96.47**	1.88	1.65
	The goal of the project is to **clump** different friendships **together** in **similar** **bunches** based on social media data. And then we add random noise to see how robust our results are. This project is financed by HORIZON 2020. Thanks go out to Dr. Strange, Hulk, Iron Man and Black Widow. We would also like to thank the initial project coordinator Stan Lee.	Increase of additional noise	4/63	**66.84**	10.09	23.07
PR16	How **will changes affect** productivity.	Reduction to important keywords	3/5	1.17	**86.75**	12.08
	How **will** potential **changes** in our production parameters **affect** our plant's productivity.	Reduction to the central statement	3/12	3.03	**84.07**	12.9
	We need a model that provides information on how potential **changes** in our production parameters **will** **affect** our plant's productivity.	Original text	3/20	6.19	**82.51**	11.3
	We need a model that provides information on how potential **changes** in our production parameters **will** **affect** our plant's productivity. And then we add random noise to see how robust our results are.	Enrichment with additional noise	3/33	14.61	**77.10**	8.3
	We need a model that provides information on how potential **changes** in our production parameters **will** **affect** our plant's productivity. And then we add random noise to see how robust our results are. This project is financed by HORIZON 2020. Thanks go out to Dr. Strange, Hulk, Iron Man and Black Widow. We would also like to thank the initial project coordinator Stan Lee.	Increase of additional noise	3/64	9.51	**67.18**	23.31
AR7	Purchase **co-occurrences**.	Reduction to important keywords	1/2	0.86	0.86	**98.27**
	We want to mine transaction records for intercontinental purchase **co-occurrences** of consumers.	Reduction to the central statement	2/12	0.81	0.81	**98.38**
	As a credit card company, we want to mine transaction records for intercontinental purchase **co-occurrences** of consumers to spot fraudulent use of credit cards.	Original text	2/24	0.81	0.87	**98.33**
	As a credit card company, we want to mine transaction records for intercontinental purchase **co-occurrences** of consumers to spot fraudulent use of credit cards. And then we add random noise to see how robust our results are.	Enrichment with additional noise	2/37	1.03	1.72	**97.25**
	As a credit card company, we want to mine transaction records for intercontinental purchase **co-occurrences** of consumers to spot fraudulent use of credit cards. And then we add random noise to see how robust our results are. This project is financed by HORIZON 2020. Thanks go out to Dr. Strange, Hulk, Iron Man and Black Widow. We would also like to thank the initial project coordinator Stan Lee.	Increase of additional noise	2/68	1.19	3.17	**95.64**

Table I3: Robustness check III - modification of length by increasing noise

References

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research 3:993–1022

Bougouin A, Boudin F, Daille B (2013) TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing, Nagoya, Japan, pp 543–551

Campos R, Mangaravite V, Pasquali A, et al (2018) A Text Feature Based Automatic Keyword Extraction Method for Single Documents. In: Pasi G, Piwowarski B, Azzopardi L, Hanbury A (eds) Advances in Information Retrieval. Springer International Publishing, Cham, pp 684–691

Chen K, Zhang Z, Long J, Zhang H (2016) Turning from TF-IDF to TF-IGM for Term Weighting in Text Classification. Expert Systems with Applications 66:245–260. https://doi.org/10.1016/j.eswa.2016.09.009

Florescu C, Caragea C (2017) PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Vancouver, Canada, pp 1105–1115

Gamma E, Helm R, Johnson RE, Vlissides J (1995) Design Patterns: Elements of Reusable Object-oriented Software. Addison-Wesley, Reading, Mass

Leacock C, Miller GA, Chodorow M (1998) Using Corpus Statistics and WordNet Relations for Sense Identification. Computational Linguistics 24(1):147–165

Mihalcea R, Tarau P (2004) TextRank: Bringing Order into Text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Barcelona, Spain, pp 404–411

Rupp C (2014) Requirements-Engineering und -Management: Aus der Praxis von klassisch bis agil, 6th ed. Hanser, München

Salton G, Buckley C (1988) Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management 24(5):513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Schnabel T, Labutov I, Mimno D, Joachims T (2015) Evaluation Methods for Unsupervised Word Embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, pp 298–307

Sokolova M, Lapalme G (2009) A Systematic Analysis of Performance Measures for Classification Tasks. Information Processing & Management 45(4):427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Wu Z, Palmer M (1994) Verbs Semantics and Lexical Selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Las Cruces, New Mexico, pp 133–138