INTERNATIONAL
HELLENIC
UNIVERSITY

**School of Science and Technology**

MSc in Data Science
Academic Year: 2022-2023

Data Mining Coursework

# Module: Data Mining

Zisis Lelekis
Filippos Iakovidis

# A literature survey of the state of the art in Social Media Analysis.

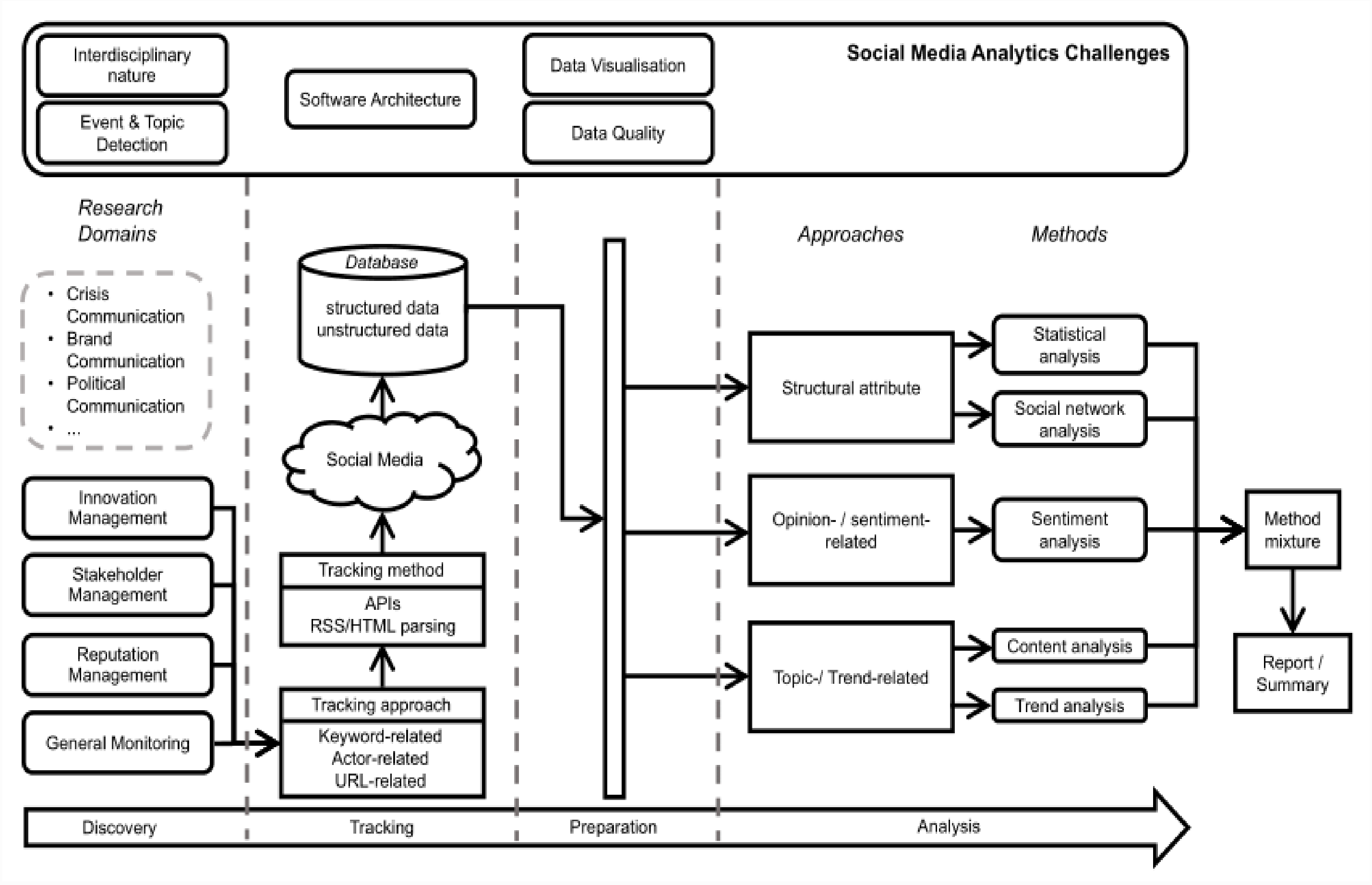
## *Introduction*

Social media is one of the fields that have been thriving intensively over the last decade. It has been an important driver for acquiring and spreading information in different domains, such as entertainment, science, business, crisis management, politics and many more. A tremendously big part of the population is making use of social media in their day-to-day lives and also social media analytics is a research area undergoing rapid change and evolution due to commercial pressures and the potential for using social media data for computational (social science) research.

The field of social media analytics is constantly evolving, driven by advancements in computing power, data availability, and analytical techniques. Researchers and practitioners have been exploring novel methods to harness the wealth of information present in social media platforms, enabling businesses to make data-driven decisions, policymakers to monitor public sentiment, and researchers to study social phenomena on an unprecedented scale. Social media consists of the tools, techniques, and technologies that use the internet to facilitate communication in an open environment. Types of social media include social networking sites(Facebook), microblogs (Twitter), blogs (Blogspot), chat (AIM), open source mapping (Wikimapia), and photo and video sharing (Flikr, Picasa, YouTube). Current research in social media analytics tends to focus around three main domains: content analysis (popular topics, and sentiment/mood), group/network analysis (define users within a group, characterize how group members interact, and identify influential users), and prediction of real-world events or characteristics (outcomes of political elections, movie revenues) . Analyzing social media, in particular Twitter feeds for sentiment analysis, has become a major research and business activity due to the availability of web-based application programming interfaces (APIs) provided by Twitter, Facebook and News services.

## Theoretical framework

The research field of social media analytics deals with methods of analyzing social media data. Researchers have divided the analytics process into several steps, whilst most of them are defining them to be: *Discovery, Collection, Preparation and Analysis*. Stieglitz and Dang-Xuan, 2013, Stieglitz et al., 2014 developed a three-stage SMA framework in the context of political communication and extended this framework in 2018 (Stieglitz et al., 2018) with the aim of creating a comparable basis for conducting social media analyses.



Discovery was added because research has increasingly highlighted the challenge of event and topic detection.
In tracking, there are different approaches, such as keyword or topic-based actor, profile or URL-based approaches. Depending on the tracking method, the database will consist of structured or unstructured social media data. For example, in social networks, the textual content belongs to the unstructured data category, while the friend/follower relationship belongs to the structured data category.

In the next step, the collected data must be prepared. Some typical challenges include data visualization and data quality.

In the final step, analysis, suitable approaches such as structural attributes, opinion- and sentiment-related or topic- and trend- related approaches must be considered. In this step, different methods, such as statistical, social network, sentiment, content or trend analysis, are included as well, based on the chosen approach. A combination of methods is advised if many research questions are posed or if the content from social media data stems from different contexts, such as politics, business, sports, or entertainment.

Next, we will provide some terminology / content analysis, plus an introduction to some of the key techniques related to analyzing unstructured textual data:

**Natural language processing**. (NLP) is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, it is the process of a computer extracting meaningful information from natural language input and/or producing natural language output.

**Sentiment Analysis.** Sentiment analysis employs dictionaries and word combinations in order to determine sentiment (positive, neutral or negative) of a conversation, a social media post, a “tweet” or a chat room. In this case, researchers have faced several problems, including “complex” human emotions such as irony, humor and sarcasm within the social media data.

Therefore, new research is looking at ways to make the analysis way more accurate.

**News analytics**. The measurement of the various qualitative and quantitative attributes of textual (unstructured data) news stories. Some of these attributes are: *sentiment*, *relevance* and *novelty*.

**Opinion mining**. Opinion mining (sentiment mining, opinion/sentiment extraction) is the area of research that attempts to make automatic systems to determine human opinion from text written in natural language

**Text analytics**. This involves information retrieval (IR), lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization and predictive analytics.

(At this point, it's best to make a short introduction to the four basic text data formats: HTML, XML, JSON and CSV.

HTML—HyperText Markup Language (HTML) as well-known is the markup language for web pages and other information that can be viewed in a web browser. HTML consists of HTML elements, which include tags enclosed in angle brackets (e.g., <div>), within the content of the web page.

XML—Extensible Markup Language (XML)—the markup language for structuring textual data using <tag>…<\tag> to define elements.

JSON—JavaScript Object Notation (JSON) is a text-based open standard designed for human-readable data interchange and is derived from JavaScript.

CSV—a comma-separated values (CSV) file contains the values in a table as a series of ASCII text lines organized such that each column value is separated by a comma from the next column's value and each row starts a new line.)

**Scraping.** Scraping is the process of collecting online data from social media and other web sites in the form of unstructured text. Also known as site scraping, web harvesting and web data extraction.

**Topic Identification.** Topic identification is a capability that is applied at the larger, real-time scale of social media data. This helps us categorize popular topics of conversation on social media, or related to specific brands, names or keywords. In addition to research on ingesting, analyzing and visualizing topics or themes in social media data, researchers are finding new ways to structure large social media datasets to organize, summarize, and interact with a body of documents, both at the topic/theme level, and the individual document level, and the relationships between those topics and documents. This includes connecting to publicly available data from Twitter, blogs, social networking sites; isolating data by geographic region; analyzing trends in topics and keywords and following and visualizing trends in real-time.

**Social Multimedia Analysis.** Although the unstructured text social media data is the most used method for social media analysis, new research is expanding this traditional approach, from sources such as Twitter and Facebook to include “social multimedia”, which is photos, videos, maps and others. Other research is designed to understand online communities surrounding social multimedia and their interactions, such as subscribers to a user’s YouTube updates, or those who routinely post and comment on Flickr. Tools and capabilities facilitate understanding large collections of multimedia data, bringing together text, images, audio, and visual information in a meaningful way.

**Group and Network Analysis:** Group and network analysis uses the interaction patterns of individuals in a network to identify groups of similar individuals. Characteristics of these groups influence the behavior of the individual group members. Group analysis consists of the following general steps:

1. Determine relationships strengths that reflect social proximity.
2. Partition the network into groups based on the relationship strength while obeying size restrictions.
3. Profile both the groups and individuals, including identifying group leaders.

Next, we will see some important factors of *Group and Network Analysis*.

**Identification of groups.** One of the many researches in social media analytics, includes the identification of new or emerging groups, looking deeper into the details in order to understand the intent of the groups, and determine similarities and differences between them. Identification of groups can include explicit relationships such as those defined by a Twitter or Instagram “follower,” or a Facebook “friend,” but also relationships via common activity, such as commenting  or posting on common online resources. Inside this analysis, things like communication style, different language usage, hash tags, emotive language or use of punctuation, etc. in an online community helps the researchers to identify actual groups. One task in this analysis is identifying a group of academics working together in a special field. Recent research in social media groups has focused on identifying key influencers within a group. Influence research has included analysis of user credibility (domain knowledge) and bandwidth (reach over social media networks) and message diffusion via follower/audience forwarding such as “retweets” on Twitter.

**Relationship Characterization.** Researchers can extract intel regarding the relationships between users of the social media, based on their communication events, such as which users have expertise, or those in a superior position to others. The language and word choice, and what / how the things are being said online may be an indication of a variety of things including the sender’s relationship to the receiver. This analysis has resulted in the ability to determine the relationships between sets of individuals based on their communication.

**User Characteristics.** Via social media, determining the characteristics of an online community, or an individual is a considerably easy task. Looking into the publicly available information from Facebook / Twitter / Instagram profiles, such as self-description, status updates, photos, interests, profession etc.

Now an equally important asset of social media analytics are *predictions*.

**Prediction of Real-World Events.** Social media data can become really useful when it comes to predictions, utilizing a combination of the scale of social media coverage, sentiment analysis and influence analysis. Predictive analytics using social media have been recently used, for example, to predict movie revenues,outcomes of political elections, financial market activity, popularity of a song on the Billboard weekly chart and the effects of product marketing.

**Determining Geo-Locations.** Because only a portion of social media data is tagged with geolocation, researchers are developing capabilities to estimate geographic regions from unstructured, non-geo-referenced text based on natural language processing, geo-statistics, and data-driven bottom-up semantics, though current capabilities are still high-level, such as differentiating users in a large city versus those at a national park.

## *Social media research and applications*

Overall, social media data is one of the largest, richest and most dynamic evidence base of human behavior, bringing new opportunities to understand individuals, groups and society.Three illustrative areas are: business, bioscience and social science.
Retail companies use social media to harness their brand awareness, product/customer service improvement, advertising/marketing strategies, network structure analysis, news propagation and even fraud detection. In finance, social media is used for measuring market sentiment and news data is used for trading. As an illustration, Bollen et al. (2011) measured sentiment of a random sample of Twitter data, finding that Dow Jones Industrial Average (DJIA) prices are correlated with the Twitter sentiment 2–3 days earlier with 87.6 percent accuracy. Wolfram (2010) used Twitter data to train a Support Vector Regression (SVR) model to predict prices of individual NASDAQ stocks, finding 'significant advantage' for forecasting prices 15 min in the future.

In the biosciences, social media is being used to collect data on large cohorts for behavioral change initiatives and impact monitoring, such as tackling smoking and obesity or monitoring diseases. An example is Penn State University biologists (Salathé et al. 2012) who have developed innovative systems and techniques to track the spread of infectious diseases, with the help of news, web sites, blogs and social media.

When it comes to computational social science applications, we have monitoring public responses to announcements, speeches and events, like political comments and initiatives; insights into community behavior; social media polling groups; early detection of emerging events. For example, Lerman et al. (2008) use computational linguistics to automatically predict the impact of news on the public perception of political candidates. Yessenov and Misailovic (2009) use movie review comments to study the effect of various approaches in extracting text features on the accuracy of four machine learning methods—Naive Bayes, Decision Trees, Maximum Entropy and K-Means clustering. Lastly, Karabulut (2013) found that Facebook's Gross

National Happiness (GNH) exhibits peaks and troughs in-line with major public events in the USA.

## *Collecting Data*

Data collection is a fundamental aspect of social media analytics, as it involves gathering relevant information from various social media platforms. Researchers and analysts employ a range of techniques to collect data for analysis, enabling them to gain insights into user behavior, trends, and sentiments.

### *Social media data providers:*

**Data access via APIs—**social media data repositories providing programmable HTTP-based access to the data via APIs (e.g., Twitter, Facebook and Wikipedia).

**Freely available databases—**repositories that can be freely downloaded, e.g., Wikipedia (http://dumps.wikimedia.org) and the Enron e-mail data set available via http://www.cs.cmu.edu/~enron/.

**Data access via tools—**sources that provide controlled access to their social media data via dedicated tools, both to facilitate easy interrogation and also to stop users 'sucking' all the data from the repository. An example is Google's Trends. These further subdivided into:

***Free sources***—repositories that are freely accessible, but the tools protect or may limit access to the 'raw' data in the repository, such as the range of tools provided by Google.

***Commercial sources***—data resellers that charge for access to their social media data. Gnip and DataSift provide commercial access to Twitter data through a partnership, and Thomson Reuters to news data.

Data access via APIs—social media data repositories providing programmable HTTP-based access to the data via APIs (e.g., Twitter, Facebook and Wikipedia).

## *Open-source databases:*

Wikipedia is the most major open source of social media, with free copies of all available content. These databases' use is for mirroring, database queries and social media analytics. Besides Wikipedia, researchers can find freely available data in the World Bank data, i.e.(http://databank.worldbank.org/data/databases.aspx ) which provides over 40 databases such as Gender Statistics, Health Nutrition and Population Statistics, Global Economic Prospects, World Development Indicators and other. These databases are usually filtered by country/region, series/topics or time.

Data access via tools is separated into free sources and commercial sources.

**Free sources:** A good example of this category is tools like Trends and InSights. Google is the largest scraper in the world. Google's strategy is to provide a wide range of packages, such as Google Analytics, rather than from a researchers' viewpoint the more useful programmable HTTP-based APIs.

**Commercial sources:** Big companies, such as Twitter are restricting free access to their data and licensing their data to commercial data resellers, such as Gnip and DataSift. They are using a payment-for-access policy in order for anyone to get their data and process with the analysis of it.

**Data feed access via APIs:** For researchers, arguably the most useful sources of social media data are those that provide programmable access via APIs, typically using HTTP-based protocols. Wikipedia is yet again a great resource for API's. They provide academics and researchers with large open-source repositories of user-generated content. These Wiki APIs also allow programmable access and searching that returns data in a variety of formats including XML. ((http://www.mediawiki.org/). This works by accepting requests containing one or more input arguments and returning strings, that can be parsed and used by the requesting client. Other formats supported include JSON, WDDX, YAML or PHP serialized. Details can be found at: http://en.wikipedia.org/w/api.php?action=query&list=allcategories&acprop=size&acprefix=hollywood&format=xml.

The HTTP request must contain: a) the requested 'action,' such as query, edit or delete operation; b) an authentication request; and c) any other supported actions. For example, the above request returns an XML string listing the first 10 Wikipedia categories with the prefix 'hollywood.' Vaswani (2011) provides a detailed description of how to scrape Wikipedia using an Apache/PHP development environment and an HTTP client capable of transmitting GET and PUT requests and handling responses.

Twitter, on the other hand, has the option of it's users to publish anything, whether that's public or private. However, less than 10% of all Twitter accounts are private. Tweets from public accounts are available in JSON format through Twitter's *Search API* for batch requests of past data and *Streaming API* for near real-time data.

**Search APIs:** Querying for recent Tweets containing specific keywords. Requires and authorized application before retrieving any results from the API.

**Streaming APIs:** A real-time stream of Tweets, filtered by user ID, keyword, geographic location or random sampling.

To summarize, in social media analytics, data collection techniques encompass both manual and automated approaches. Manual methods involve manually searching and collecting data from social media platforms, while automated methods rely on specialized tools and APIs (Application Programming Interfaces) provided by the platforms. These APIs allow researchers to access and retrieve specific data, such as user profiles, posts, comments, and engagement metrics.

## *Challenges*

At this point, researchers in social media analysis have faced many challenges. It can be argued that social media data shares many characteristics of “big” data, a term that encompasses data obtained from vastly different sources and in very different disciplines. The notion that today's “big”data poses new challenges is widely acknowledged in various fields. The key factors by which this new phenomenon differs from traditional analytics can be summarized as follows:

**volume**: the storage space required.

**velocity**: the speed of data creation coupled with the advantage gained from analyzing the data in real time.

**variety:** the fact that data takes many different forms. It is often unstructured or its structure is specific to the data source, and

**veracity**: uncertainty especially with regard to data quality

The first four V’s correspond to immediate technical challenges. For example, when the data takes up so much physical space that it does not fit into memory, many algorithms run considerably slower.

A big challenge for the social media analysis researchers is spam and missing data, which both compromise the veracity of the data, and they are not likely to benefit from a technique that is designed to cope with its velocity. Spam is data that is not related to the topic and represent e.g. advertisement. This increases the amount of data and makes the analyses more difficult.

However, social media data can also have negative side effects. This has been recently labeled as “the dark side of social media” (Jalonen & Jussila, 2016; Kalhour & Ng, 2016;Payton & Conley, 2014). Rumors and false information could have a negative influence on the behavior of other social media users. Therefore it becomes necessary to identify misinformation, rumors and fake news, and the overall credibility of a user. Therefore, Mechanisms are needed for detecting these categories of content.

Another important challenge that the recent social media analysis researchers are facing, is the Ethical and Privacy Considerations.

Ethical and privacy considerations play a crucial role in social media analytics. As researchers and practitioners harness the power of social media data, it is imperative to address potential ethical challenges and privacy concerns to ensure responsible and respectful data usage.

Ethical considerations in social media analytics involve issues such as informed consent, data ownership, and data anonymization. Obtaining informed consent from social media users before using their data for research purposes is essential to uphold ethical standards. Transparency in data usage and providing clear explanations of the research objectives and methodologies are vital in establishing trust with users.

Another ethical concern is the ownership of social media data. While the data is publicly available, it is essential to respect the rights of individuals and not exploit or misuse their personal information. Data anonymization techniques, such as removing personally identifiable information, can help protect user privacy and prevent potential harm.

Privacy considerations in social media analytics revolve around safeguarding user data and maintaining confidentiality. Social media platforms have their own privacy policies and terms of service that need to be adhered to when collecting and analyzing data. It is crucial to ensure compliance with relevant regulations, such as the General Data Protection Regulation (GDPR) in the European Union, to protect user privacy rights.

Additionally, mitigating biases and ensuring fairness in social media analytics is an important ethical consideration. Biases in data collection, algorithmic decision-making, or analysis techniques can lead to unfair outcomes or reinforce existing societal biases. Researchers must be mindful of potential biases and strive for unbiased and fair analysis practices.

By addressing these ethical and privacy considerations, researchers and practitioners can uphold the principles of responsible data usage in social media analytics. Implementing appropriate consent mechanisms, anonymization techniques, and complying with privacy regulations help protect user privacy rights. Moreover, promoting fairness and mitigating biases contribute to more ethical and socially responsible analyses.

As the field of social media analytics continues to evolve, ongoing discussions and collaborations between researchers, industry professionals, and policymakers are crucial in establishing ethical guidelines and privacy frameworks. These guidelines ensure that social media analytics research is conducted ethically, respects user privacy, and ultimately benefits society as a whole.

## *Conclusion*

Social media analytics is a relatively new research area, but it is of great interest to the Information Systems community and many researchers. It is a field, which with the right approach, can provide researchers with a great amount of important information. If used properly, it can be used to give feedback for massive community events, make predictions about upcoming events such as political elections, revenues of new music albums / movies. Nevertheless, social media analytics is facing many challenges. It has raised numerous important questions, such as: Which volume of data do we expect? How do we discover the parts which are relevant to our research? Do we have adequate infrastructure to cope with that volume when collecting and preparing the data? Which format will the data be in? If the data is unstructured, how can we extract the relevant structured information from it?

The biggest concern is that companies are increasingly restricting access to their data to monetize their content. It is important that researchers have access to computational environments and especially 'big' social media data for experimentation. Otherwise, computational social science could become the exclusive domain of major companies, government agencies and a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Arguably what is required are public-domain computational environments and data facilities for quantitative social science, which can be accessed by researchers via a cloud-based facility. If we overcome all these challenges, and ensure that we proceed researching this field with respect to ethics and privacy, the entire society can be crucially benefited by social media analysis as a whole.

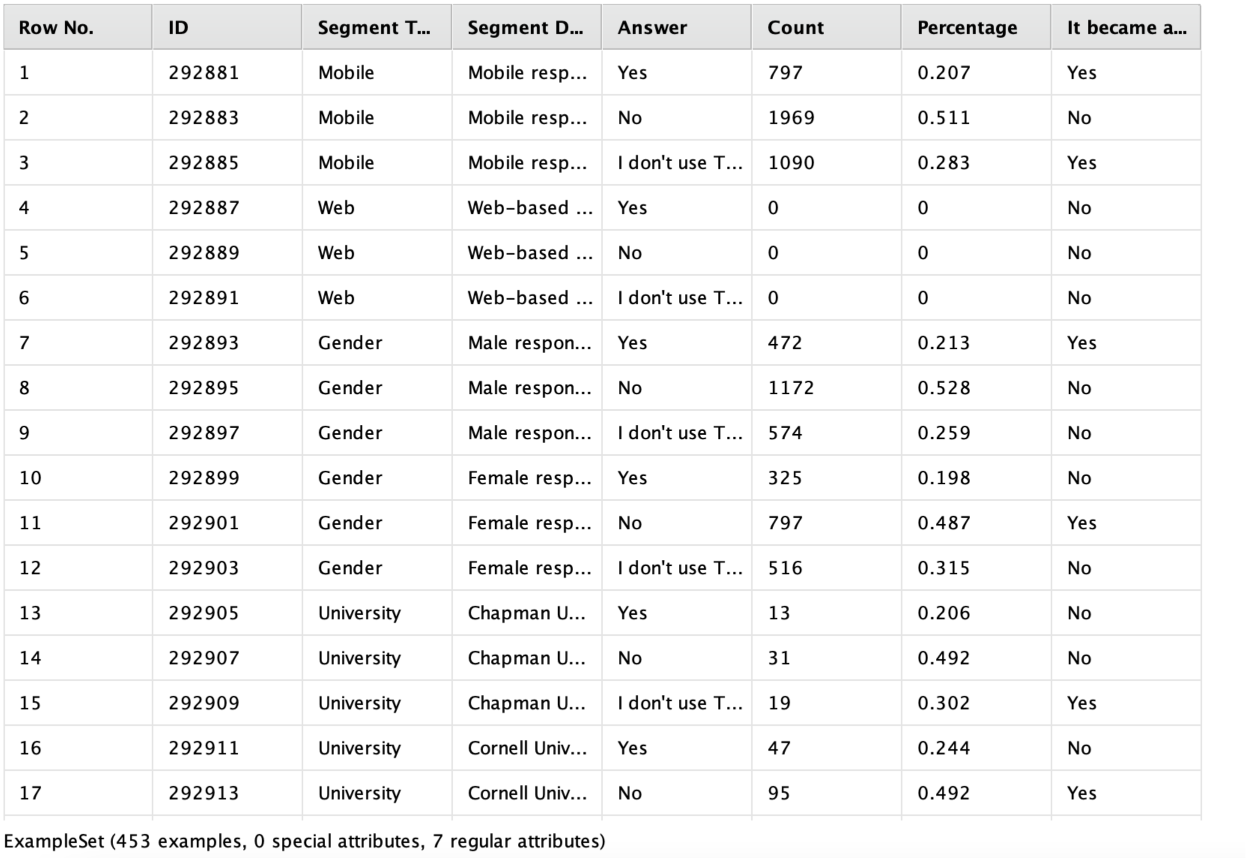
**2. Select an application domain preferably related to your topic above. Pick a data set, possibly available publicly (e.g. diabetes from the UC Irvine Machine Learning Repository) use data mining (e.g. classification, clustering, association rules etc.) to extract knowledge from it. You can use any tool you like (e.g. Weka, RapidMiner). Then go back to preprocess your data set, change algorithm, parameters etc. attempting to extract knowledge out of your data. Describe the process and discuss the results. Which method was more appropriate and why? Detail the algorithmic parameters you used, justify your choices, and discuss the challenges you met and how you overcame these.**

For the second part of the coursework and related to the social media analytic topic, a dataset from Kaggle is used. The dataset's name is Tinder Millennial Match Rate, raising the question of whether and how Many Millennials Find Someone on Tinder ? The dataset explains the match rate of individuals from different universities and whether the app has helped the person to find relationship. We did that using Rapidminder.

The dataset consists of the following attributes:

- ID -User id
- Segment type : Medium of Usage
- Segment Description: Name of Universities
- Answer: Do you use tinder ?
- Count: Number of Matches
- Percentage: % of matches
- It became a relationship: Success of relationship (Target)

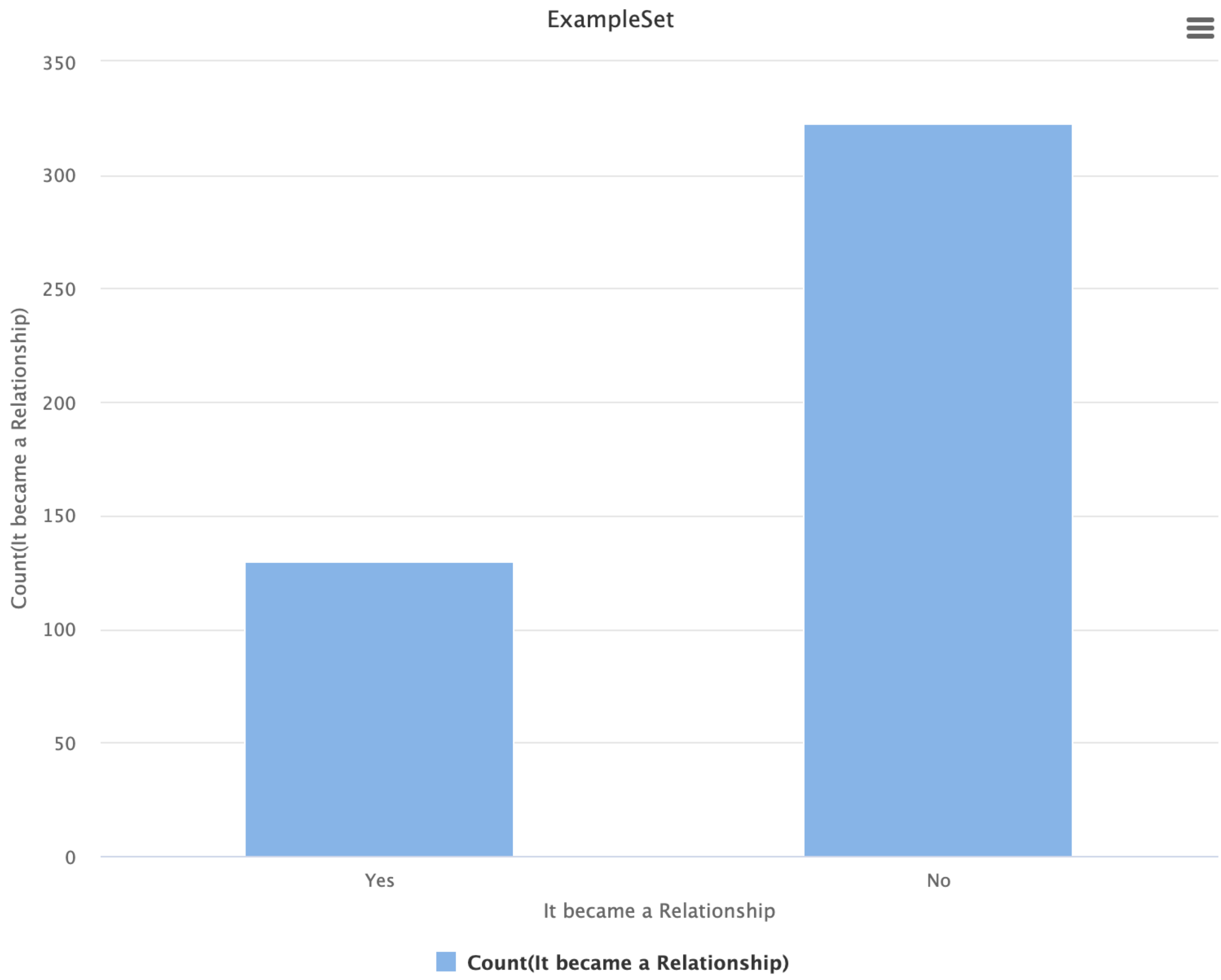
Our goal with this dataset, is to as most accurately as possible predict whether the Tinder helped the person find a relationship, by predicting the correct Yes or No class in the it became a relationship attribute (label).



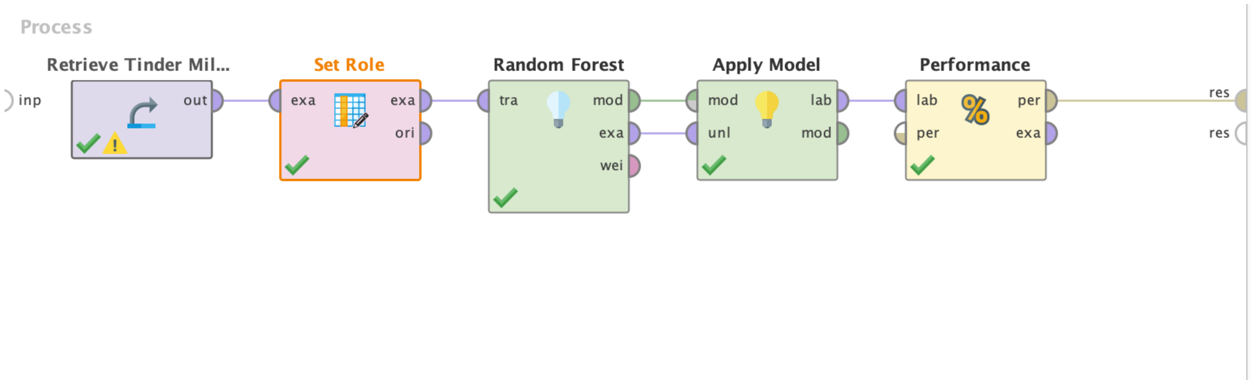
| Row No. | ID | Segment T... | Segment D... | Answer | Count | Percentage | It became a... |
|---|---|---|---|---|---|---|---|
| 1 | 292881 | Mobile | Mobile resp... | Yes | 797 | 0.207 | Yes |
| 2 | 292883 | Mobile | Mobile resp... | No | 1969 | 0.511 | No |
| 3 | 292885 | Mobile | Mobile resp... | I don't use T... | 1090 | 0.283 | Yes |
| 4 | 292887 | Web | Web-based ... | Yes | 0 | 0 | No |
| 5 | 292889 | Web | Web-based ... | No | 0 | 0 | No |
| 6 | 292891 | Web | Web-based ... | I don't use T... | 0 | 0 | No |
| 7 | 292893 | Gender | Male respon... | Yes | 472 | 0.213 | Yes |
| 8 | 292895 | Gender | Male respon... | No | 1172 | 0.528 | No |
| 9 | 292897 | Gender | Male respon... | I don't use T... | 574 | 0.259 | No |
| 10 | 292899 | Gender | Female resp... | Yes | 325 | 0.198 | No |
| 11 | 292901 | Gender | Female resp... | No | 797 | 0.487 | Yes |
| 12 | 292903 | Gender | Female resp... | I don't use T... | 516 | 0.315 | No |
| 13 | 292905 | University | Chapman U... | Yes | 13 | 0.206 | No |
| 14 | 292907 | University | Chapman U... | No | 31 | 0.492 | No |
| 15 | 292909 | University | Chapman U... | I don't use T... | 19 | 0.302 | Yes |
| 16 | 292911 | University | Cornell Univ... | Yes | 47 | 0.244 | No |
| 17 | 292913 | University | Cornell Univ... | No | 95 | 0.492 | Yes |

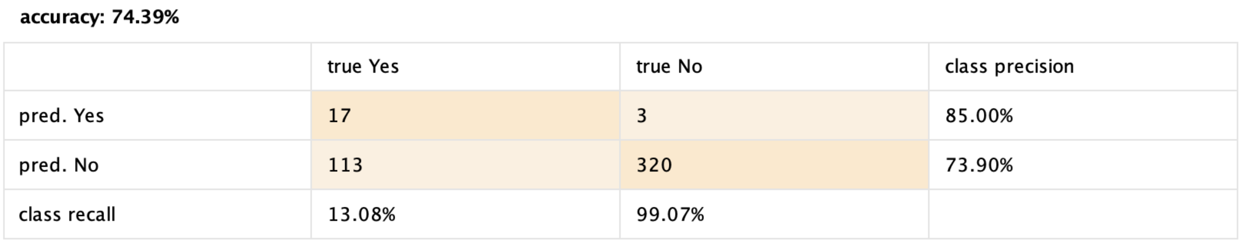
ExampleSet (453 examples, 0 special attributes, 7 regular attributes)

Also, our dataset is not unbalanced as it consists of 323 No classes and 130 Yes classes while we don't have missing values.



Firstly, we tried to extract knowledge from the dataset, applying the random forest algorithm, and without preprocessing our data, apart from assigning the label role to the it became a relationship attribute.

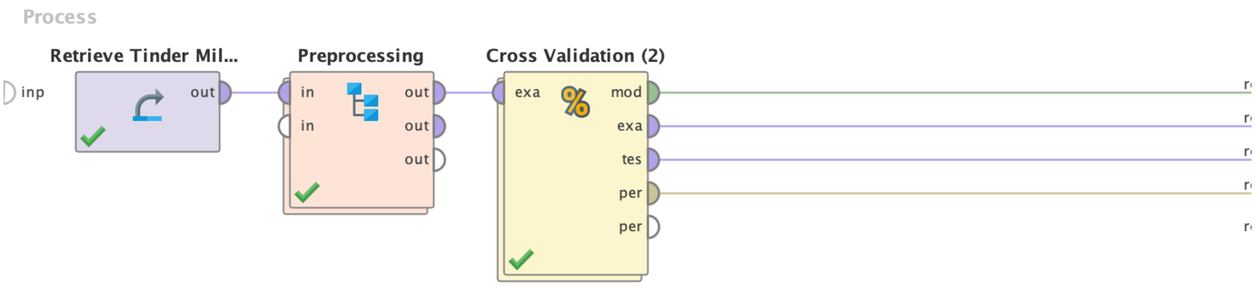


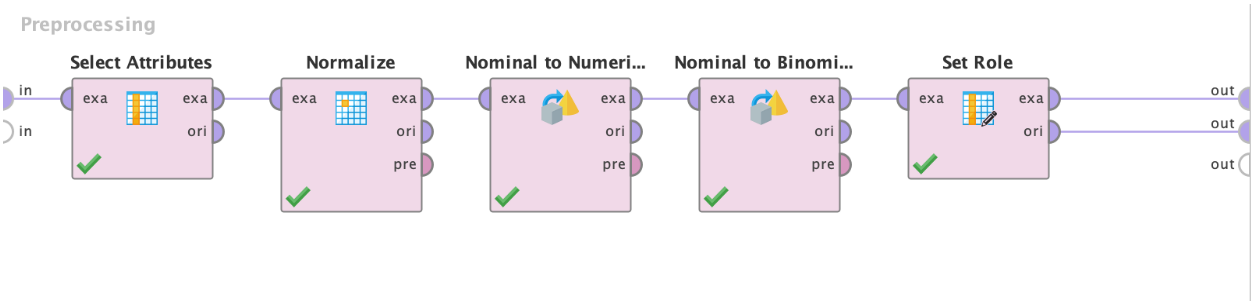

**accuracy: 74.39%**

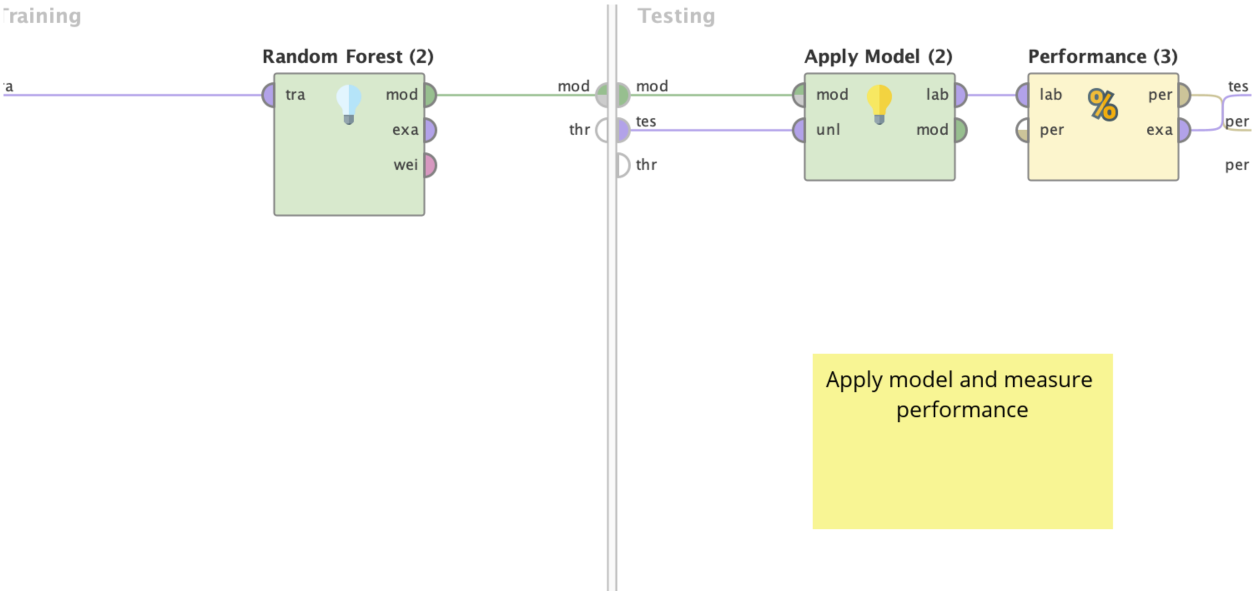
| | true Yes | true No | class precision |
|---|---|---|---|
| pred. Yes | 17 | 3 | 85.00% |
| pred. No | 113 | 320 | 73.90% |
| class recall | 13.08% | 99.07% | |

The accuracy data are not correct though, as this is the training error, and our mail goal here is the prediction error.

To proceed, we continued with the preprocess of the data set, we tested the ROC curves to find the appropriate algorithms that would maximize the accuracy and recall and finally changed the algorithm.
While preprocessing our data, and in the Preprocessing subprocess, we excluded the attribute segment segment description. There are 151 segment description values, which indicate the college which the response came from. Because there are <500 rows in this small dataset, it would not be feasible to conduct robust analysis on individual university samples.
We continued with normalizing the number of matches (count attribute) and then converted both nominal attributes answer and segment type to numerical. Furthermore, we changed the nominal attribute It became a relationship, to binominal. Finally, we have set It became a relationship as the label attribute in order to predict the target and assigned ID to ID type.

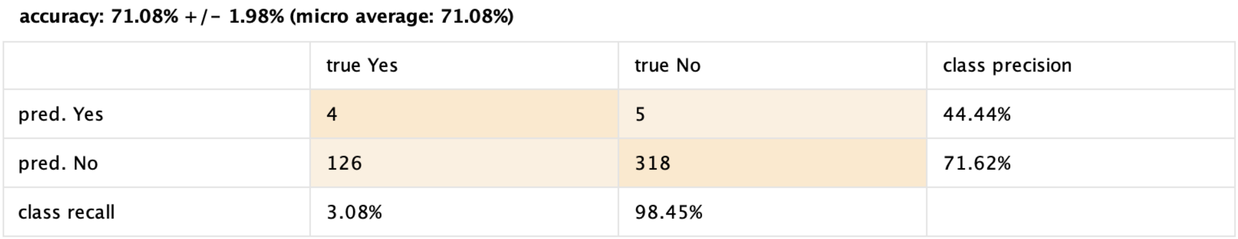






Apply model and measure performance

As a next step, we used cross validation, to get train and test data, and measure the effectiveness and accuracy of our model.

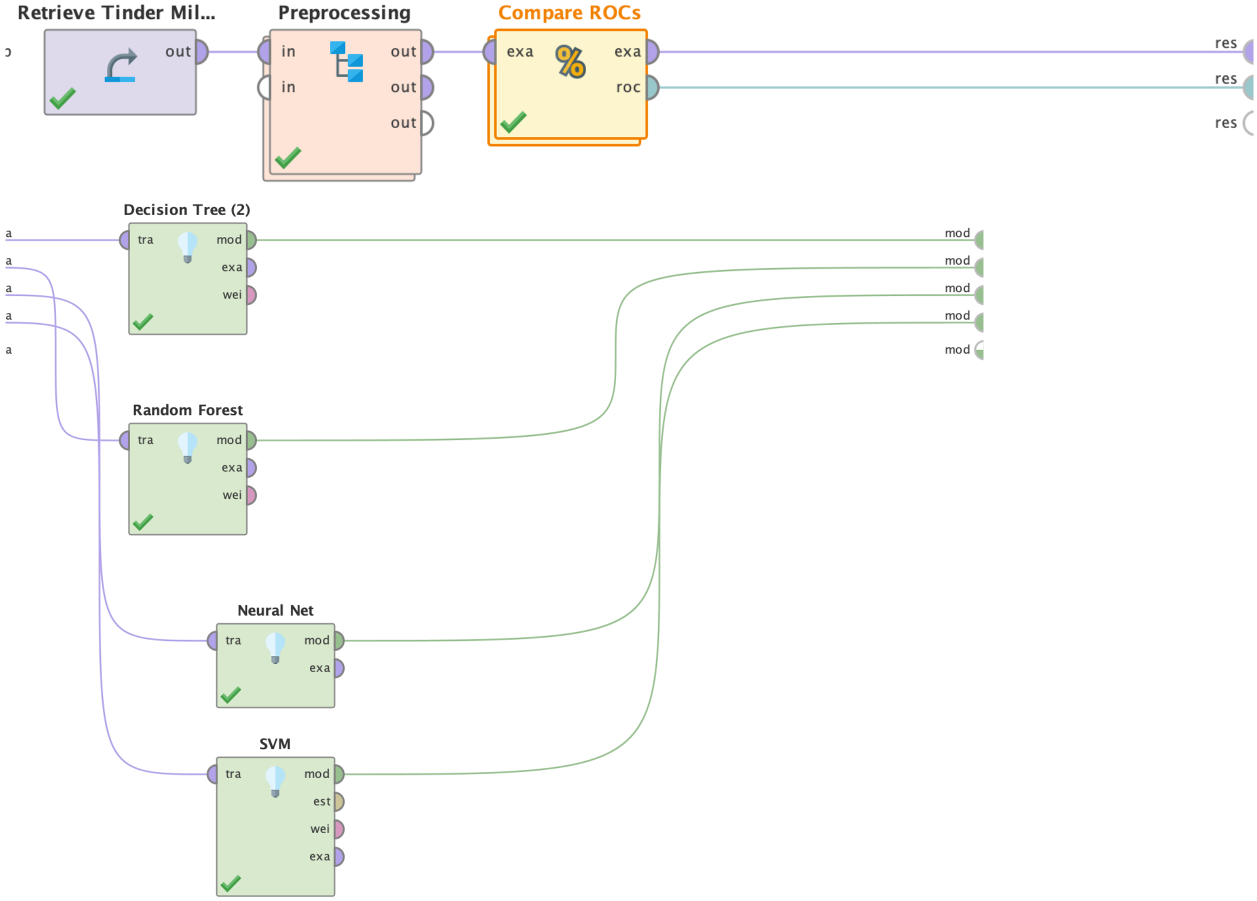


**accuracy: 71.08% +/- 1.98% (micro average: 71.08%)**

| | true Yes | true No | class precision |
|---|---|---|---|
| pred. Yes | 4 | 5 | 44.44% |
| pred. No | 126 | 318 | 71.62% |
| class recall | 3.08% | 98.45% | |

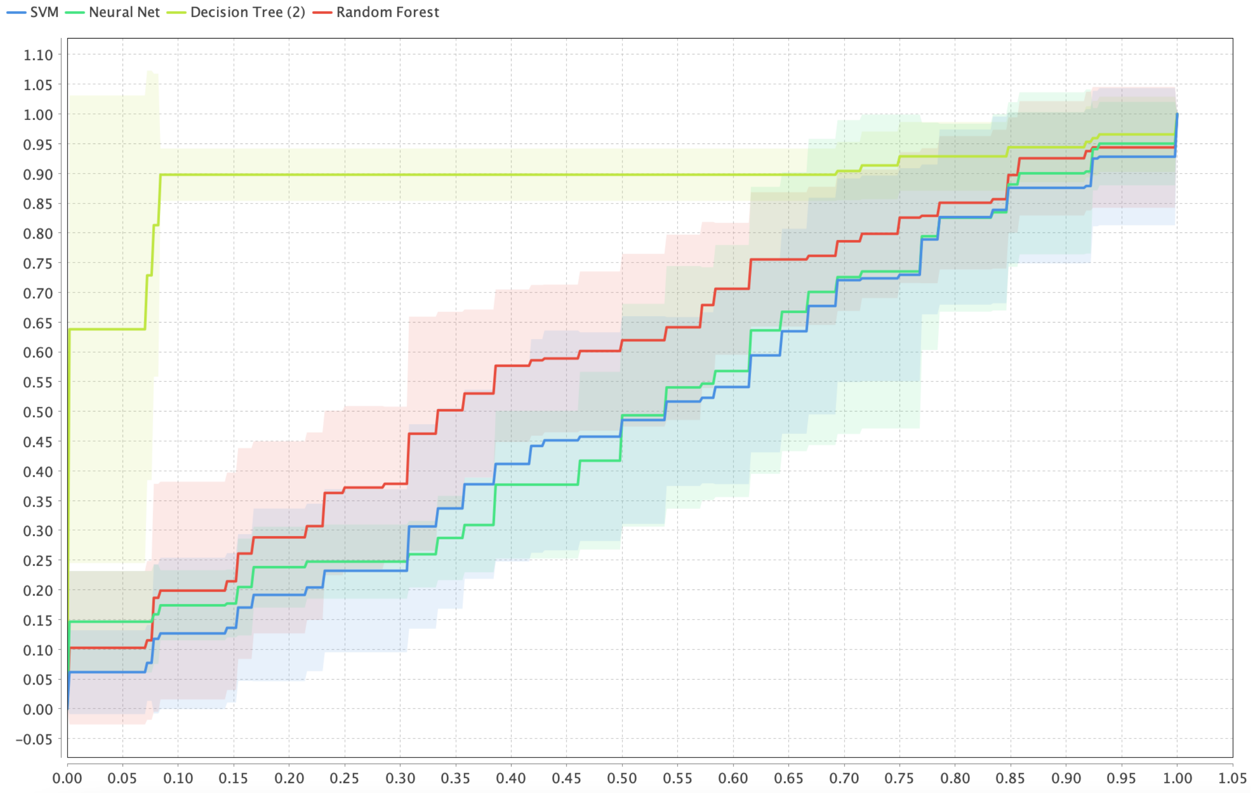
The accuracy of our dataset is 71% but our biggest concern is around class recall for class Yes, as it only predicted correct only 3% of the values.

For this reason, we proceeded with the ROC curves to find how different algorithms would compete.



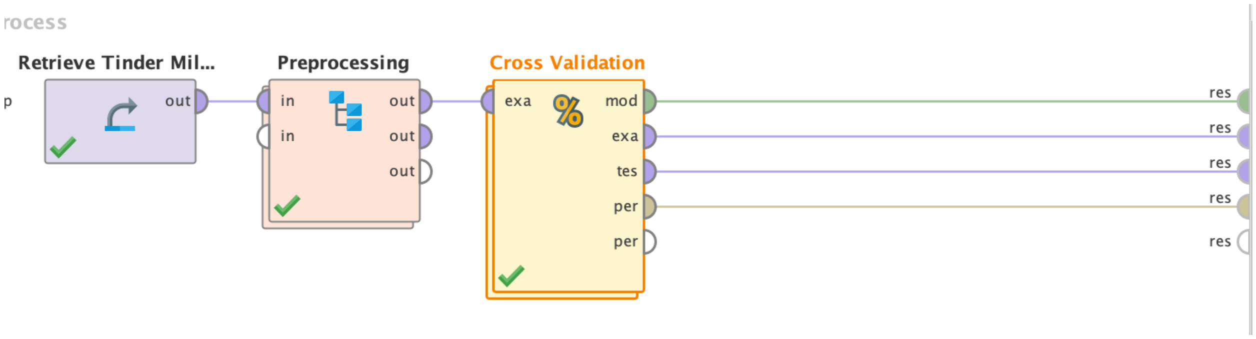
For this reason we used the compare ROCs operator in Rapidminer. Inside the ROCs operator we measured decision trees, random forests, neural nets and

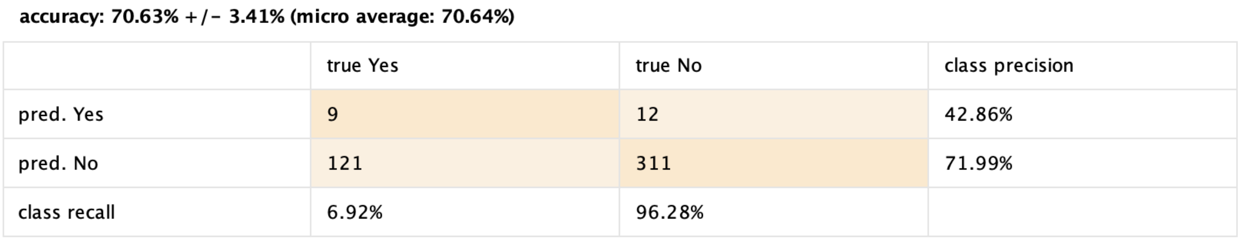
SVM.



Generating the diagram, we conclude that Decision trees outperformed the rest of the algorithms, and is probably the best algorithm for this dataset.

Based on the ROC curves, we continued with changing our algorithm from random forest to decision trees.

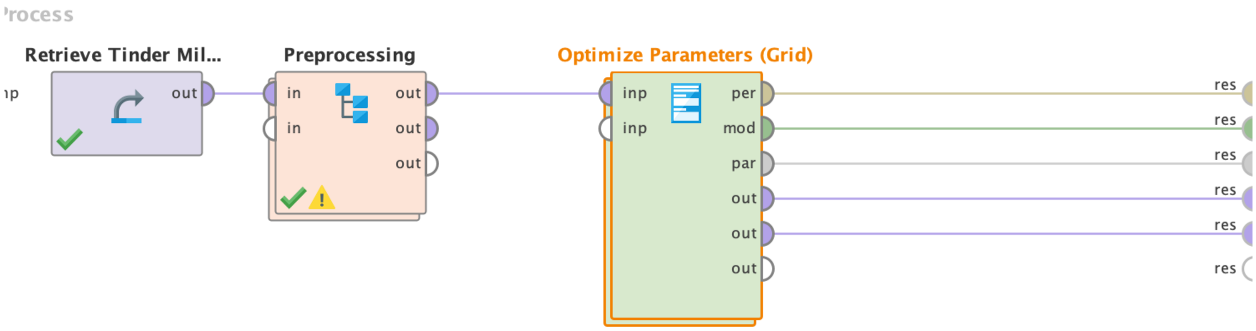


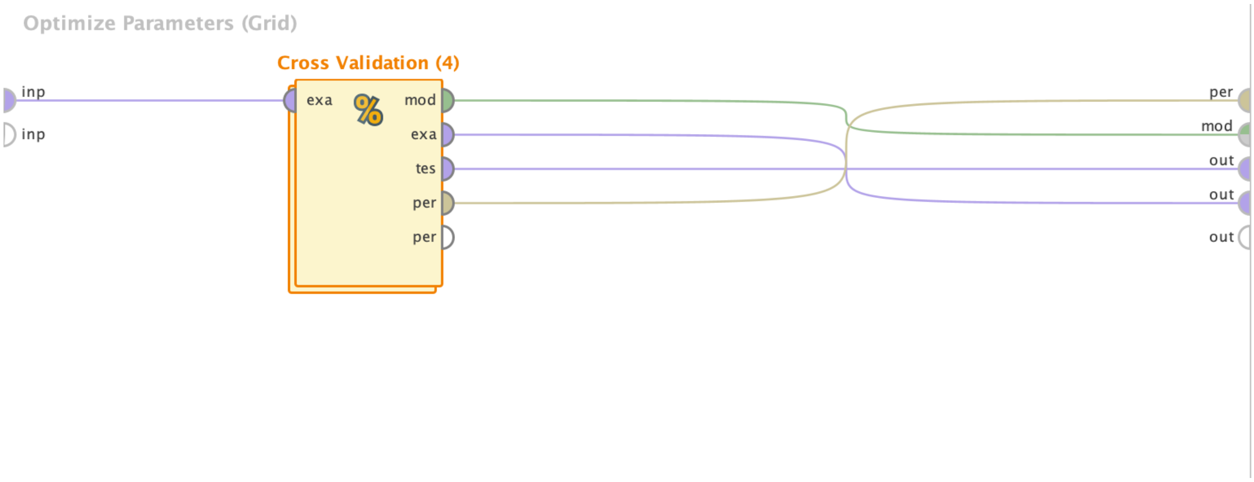


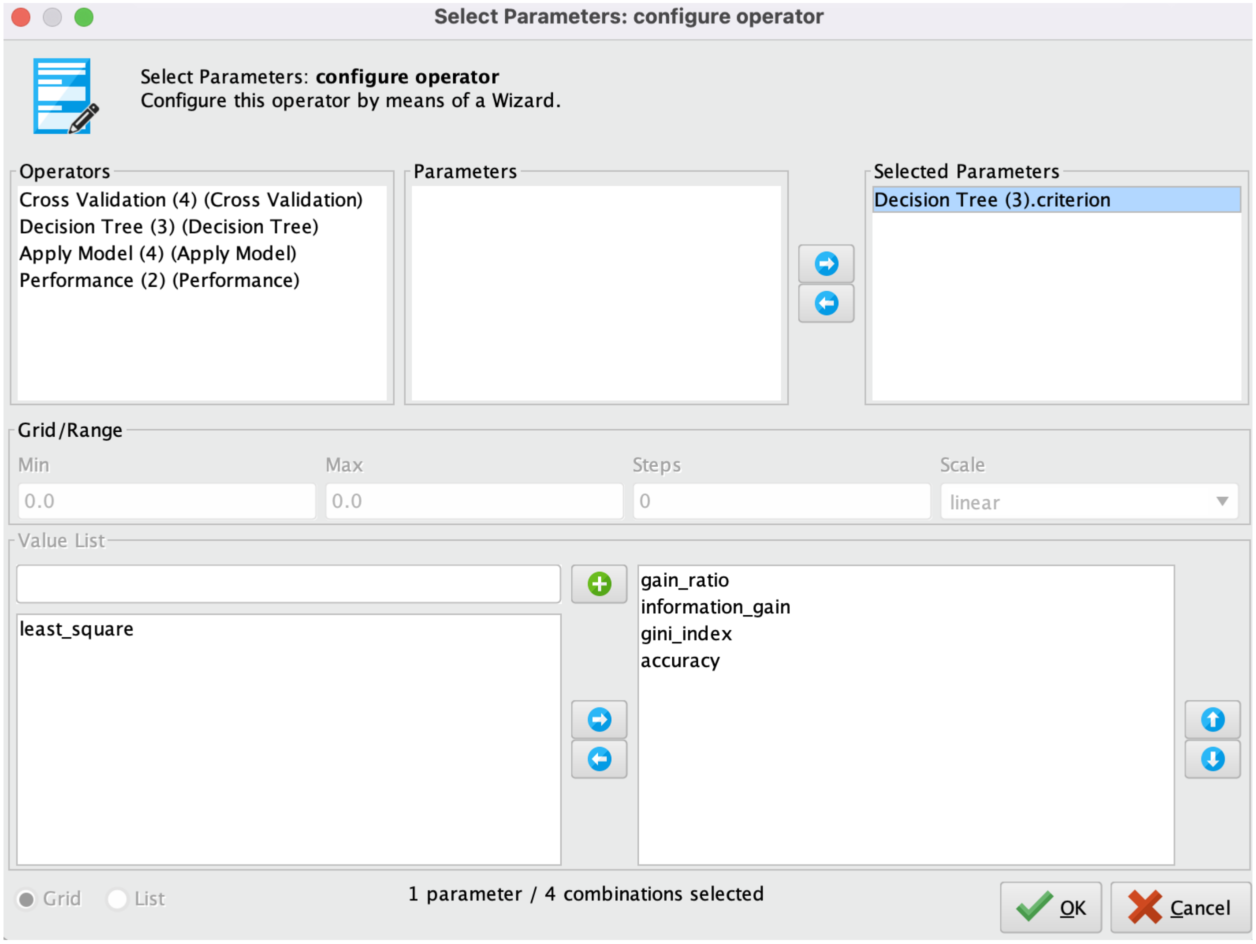
**accuracy: 70.63% +/− 3.41% (micro average: 70.64%)**

| | true Yes | true No | class precision |
|---|---|---|---|
| pred. Yes | 9 | 12 | 42.86% |
| pred. No | 121 | 311 | 71.99% |
| class recall | 6.92% | 96.28% | |

Changing our algorithm from random forest to decision tree, we notice in our confusion matrix that even though the models accuracy has slightly dropped, the class recall for yes value has increased, or even doubled to almost 7%.






Select Parameters: configure operator

Select Parameters: **configure operator**
Configure this operator by means of a Wizard.

Operators
Cross Validation (4) (Cross Validation)
Decision Tree (3) (Decision Tree)
Apply Model (4) (Apply Model)
Performance (2) (Performance)

Parameters

Selected Parameters
Decision Tree (3).criterion

Grid/Range

| Min | Max | Steps | Scale |
|---|---|---|---|
| 0.0 | 0.0 | 0 | linear |

Value List

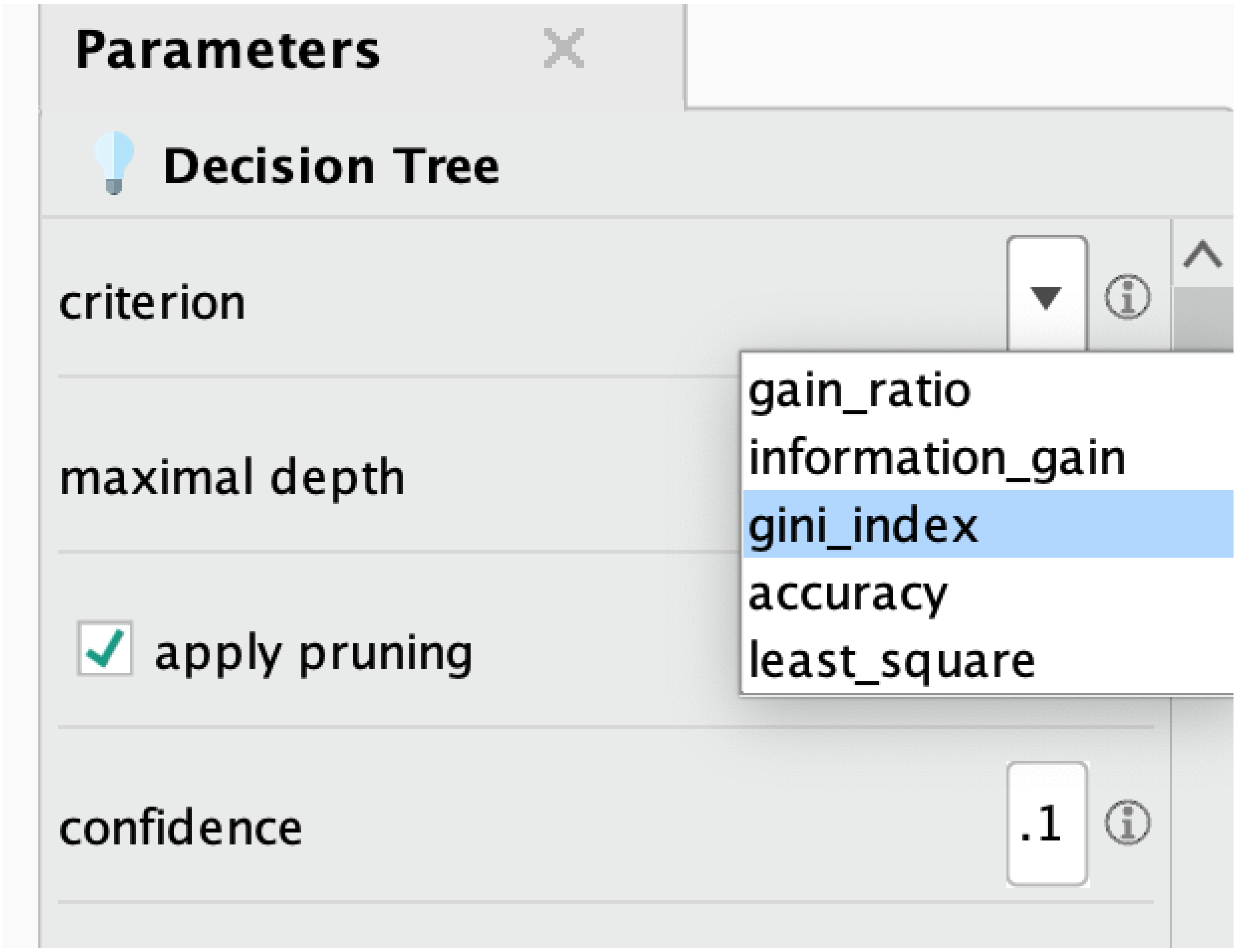
least_square

gain_ratio
information_gain
gini_index
accuracy

● Grid ○ List

1 parameter / 4 combinations selected

Therefore, changing from gain ratio to gini index  we can see an increase in Yes class' recall to almost 15%.



Parameters

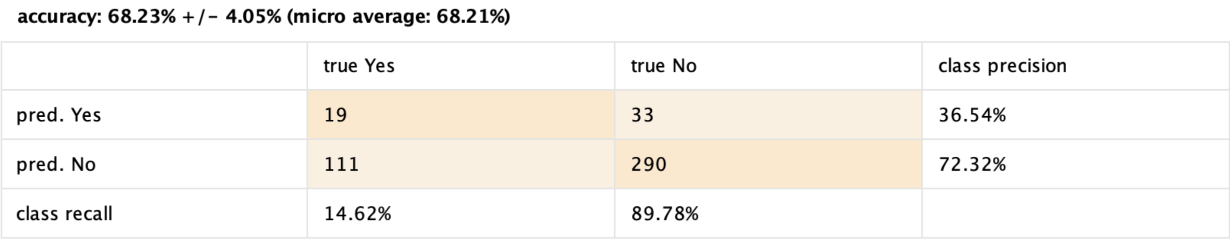
Decision Tree

criterion gini_index

maximal depth

☑ apply pruning

confidence .1



accuracy: 68.23% +/- 4.05% (micro average: 68.21%)

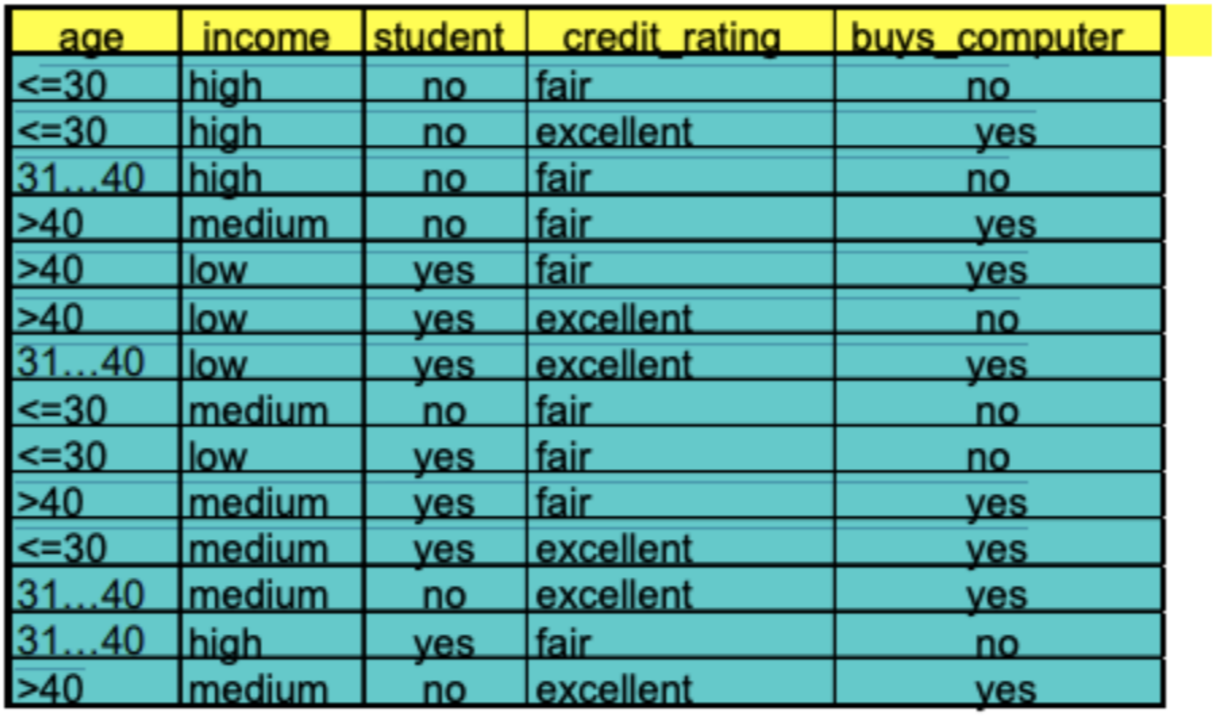
| | true Yes | true No | class precision |
|---|---|---|---|
| pred. Yes | 19 | 33 | 36.54% |
| pred. No | 111 | 290 | 72.32% |
| class recall | 14.62% | 89.78% | |

Based on our results, we can conclude that the data in the dataset is insufficient to make accurate predictions. Even using our dataset without using cross validation, our maximum accuracy was close to 74% which is rather low.

It is suggested that more data in volume and more detailed are gathered to predict whether Tinder's users will finally find their perfect match and have a relationship.

## 3. Construct a decision tree using gain ratio, information gain or gini index.

Attribute Selection using Information Gain



| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | yes |
| 31…40 | high | no | fair | no |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | no |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | yes |
| 31…40 | high | yes | fair | no |
| >40 | medium | no | excellent | yes |

| Age | income | student | credit | buys |
|---|---|---|---|---|
| below 30 | high | no | excellent | yes |
| below 30 | medium | yes | excellent | yes |
| 31 to 40 | medium | no | excellent | yes |
| above 40 | medium | no | excellent | yes |
| above 40 | low | yes | excellent | no |
| above 40 | medium | no | fair | yes |
| 31 to 40 | low | yes | excellent | yes |
| below 30 | high | no | fair | no |
| below 30 | medium | no | fair | no |
| above 40 | low | yes | fair | yes |
| below 30 | low | yes | fair | no |
| 31 to 40 | high | no | fair | no |

| | | | | |
|---|---|---|---|---|
| 31 to 40 | high | yes | fair | no |
| above 40 | medium | yes | fair | yes |

Class P: buys_computer = “yes”
Class N: buys_computer = “no”

## 1st split

### Expected information:

Info(D) = I(8,6) = -8/14*log2(8/14)-6/14*log2(6/14) = 0.98

### Information from age:

| age | pi | ni | I(pi,ni) |
|---|---|---|---|
| <=30 | 2 | 3 | 0.971 |
| 31...40 | 2 | 2 | 1 |
| >40 | 4 | 1 | 0.721 |

Info_A = 5/14*I(2,3) + 4/14*I(2,2) + 5/14*I(4,1) = 5/14*0.971 + 4/14*1 + 5/14*0.72 = 0.346 + 0.28 + 0.25 = 0.89

### Information from income:

| income | pi | ni | I(pi,ni) |
|---|---|---|---|
| low | 2 | 2 | 1 |
| medium | 5 | 1 | 0.65 |
| high | 1 | 3 | 0.81 |

Info_I = 4/14*I(2,2) + 6/14*I(5,1) + 4/14*(1,3) = 0.79

## Information from student:

| student | pi | ni | I(pi,ni) |
|---|---|---|---|
| no | 4 | 3 | 0.98 |
| yes | 4 | 3 | 0.98 |

Info_S = 7/14*I(4,3) + 7/14*I(4,3) = 0.98

## Information from credit_rating:

| credit_rating | pi | ni | I(pi,ni) |
|---|---|---|---|
| fair | 3 | 5 | 0.95 |
| excellent | 5 | 1 | 0.65 |

Info_CR = 8/14*I(3,5) + 6/14*I(5,1) = 0.82

## Gain:

Gain(age) = Info(D) - Info_A = 0.98 - 0.89 = 0.09
Gain(income) = Info(D) - Info_I = 0.98 - 0.79 = 0.19
Gain(student) = Info(D) - Info_S = 0.98 - 0.98 = 0
Gain(credit_rating) = Info(D) - Info_CR = 0.98 - 0.82 = 0.16

As a result, the first split on the decision tree will be based on the income, as it is has the largest gain ratio.
Checked

## 2nd split (high)

| **Age** | **income** | **student** | **credit** | **buys** |
|---|---|---|---|---|
| below 30 | high | no | excellent | yes |
| 31 to 40 | high | no | fair | no |

| | | | | |
|---|---|---|---|---|
| below 30 | high | no | fair | no |
| 31 to 40 | high | yes | fair | no |

## Expected information:

| income | pi | ni | I(pi,ni) |
|---|---|---|---|
| high | 1 | 3 | 0.81 |

Info(D) = I(1,3) = 0.81

## Information from age:

| Age | income | student | credit | buys |
|---|---|---|---|---|
| 31 to 40 | high | no | fair | no |
| 31 to 40 | high | yes | fair | no |
| below 30 | high | no | excellent | yes |
| below 30 | high | no | fair | no |

| age | pi | ni | I(pi,ni) |
|---|---|---|---|
| <=30 | 1 | 1 | 0.5 |
| 31...40 | 0 | 2 | 0 |

Info_A = 0.25

## Information from student:

| Age | income | student | credit | buys |
|---|---|---|---|---|
| 31 to 40 | high | no | fair | no |
| below 30 | high | no | excellent | yes |
| below 30 | high | no | fair | no |
| 31 to 40 | high | yes | fair | no |

| student | pi | ni | I(pi,ni) |
|---|---|---|---|
| yes | 0 | 1 | 0 |
| no | 1 | 2 | 0.52 |

Info_S = 0.39

### Information from credit:

| **Age** | **income** | **student** | **credit** | **buys** |
|---|---|---|---|---|
| below 30 | high | no | excellent | yes |
| 31 to 40 | high | no | fair | no |
| below 30 | high | no | fair | no |
| 31 to 40 | high | yes | fair | no |

| credit | pi | ni | I(pi,ni) |
|---|---|---|---|
| excellent | 1 | 0 | 0 |
| fair | 0 | 3 | 0 |

Info_C = 0

### Gain:

Gain(age) = Info(D) - Info_A= 0.81 - 0.25 = 0.56
Gain(student) = Info(D) - Info_S = 0.81 - 0.39 = 0.42
Gain(credit_rating) = Info(D) - Info_CR = 0.81 - 0 = 0.81

As a result, the second split on the decision tree for high income customers will be based on the credit rating.

### 2nd split (medium income)

Regarding the medium income case is, we have to define where we are going to split first, in age, student or credit rating. We calculate again the Gain for these three.

| Age | income | student | credit | buys |
|---|---|---|---|---|
| 31 to 40 | medium | no | excellent | yes |
| above 40 | medium | no | excellent | yes |
| below 30 | medium | yes | excellent | yes |
| above 40 | medium | no | fair | yes |
| below 30 | medium | no | fair | no |
| above 40 | medium | yes | fair | yes |

## Expected information:

| income | pi | ni | I(pi,ni) |
|---|---|---|---|
| medium | 5 | 1 | 0.65 |

Info(D) = I(5,1) = 0.65

## Information from age:

| age | pi | ni | I(pi,ni) |
|---|---|---|---|
| <=30 | 1 | 1 | 0.5 |
| 31 to 40 | 1 | 0 | 0 |
| Above 40 | 3 | 0 | 0 |

Info_A = 0.16

## Information from student:

| student | pi | ni | I(pi,ni) |
|---|---|---|---|
| yes | 2 | 0 | 0 |
| no | 3 | 1 | 0.31 |

Info_S = 0.2

## Information from credit:

| credit | pi | ni | I(pi,ni) |
|---|---|---|---|
| excellent | 3 | 0 | 0 |
| fair | 2 | 1 | 0.38 |

Info_C = 0.19

## Gain:

Gain(age) = Info(D) - Info_A = 0.65 - 0.16 = 0.49
Gain(student) = Info(D) - Info_S = 0.65 - 0.2 = 0.45
Gain(credit_rating) = Info(D) - Info_CR = 0.65 - 0.19 = 0.46

As a result, the second split on the decision tree for medium income customers will be based on the age

After the second split and with regards to the third split,  as the Info Gain is the same for two of these attributes we make the split wherever we want. So we split by credit.

## 2nd split (low income)

Regarding the low income case is, we have to define where we are going to split first, in age, student or credit rating. We calculate again the Gain for these three.

| Age | income | student | credit | buys |
|---|---|---|---|---|
| 31 to 40 | low | yes | excellent | yes |
| above 40 | low | yes | excellent | no |
| above 40 | low | yes | fair | yes |
| below 30 | low | yes | fair | no |

## Expected information:

| **income** | pi | ni | I(pi,ni) |
|---|---|---|---|
| low | 2 | 2 | 1 |

Info(D) = I(2,2) = 1

## Information from age:

| age | pi | ni | I(pi,ni) |
|---|---|---|---|
| <=30 | 0 | 1 | 0 |
| 31 to 40 | 1 | 0 | 0 |
| Above 40 | 1 | 1 | 0.5 |

Info_A = 0.25

## Information from student:

| student | pi | ni | I(pi,ni) |
|---|---|---|---|
| yes | 2 | 2 | 0.5 |
| no | 0 | 0 | 0 |

Info_S = 0.5

## Information from credit:

| credit | pi | ni | I(pi,ni) |
|---|---|---|---|
| excellent | 1 | 1 | 0.5 |
| fair | 1 | 1 | 0.5 |

Info_C = 1

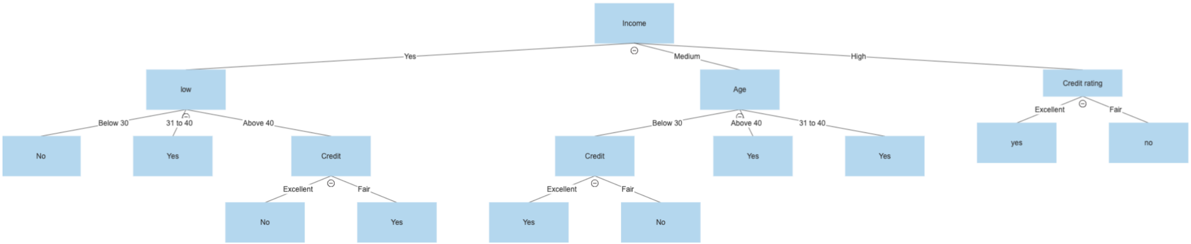
## Gain:

Gain(age) = Info(D) - Info_A = 1 - 0.25 = 0.75
Gain(student) = Info(D) - Info_S = 1 - 0.5 = 0.5
Gain(credit_rating) = Info(D) - Info_CR = 1 - 1 = 0

As a result, the second split on the decision tree for medium income customers will be based on the age as well.

After the second split and with regards to the third split,  the gain info for credit is bigger so we use credit to split.

The final tree can be see below:



## 4. Let A(-1,0), B(2,0), C(1,1), D(0,3), E(3,2) and F(3,3) be 6 entities in the 2-dimensional space.

## a) Cluster the 6 entities using agglomerative hierarchical clustering, with Euclidian distance, and single, complete, and average link methods. Discuss your observations.

With hierarchical agglomeriative clustering we start with one cluster and iteratively merge clusters until all the items belong to one cluster.

First, let's define all 3 methods that we are going to be using.

Simple link: The distance between the closest members of the 2 clusters.

Complete link: The distance between the members that are farthest apart.

Average link: Here, we are looking at the distances between all pairs and averages of these distances.

In order to perform the clustering of the 6 entities, we will have to use the Euclidian distances matrix.

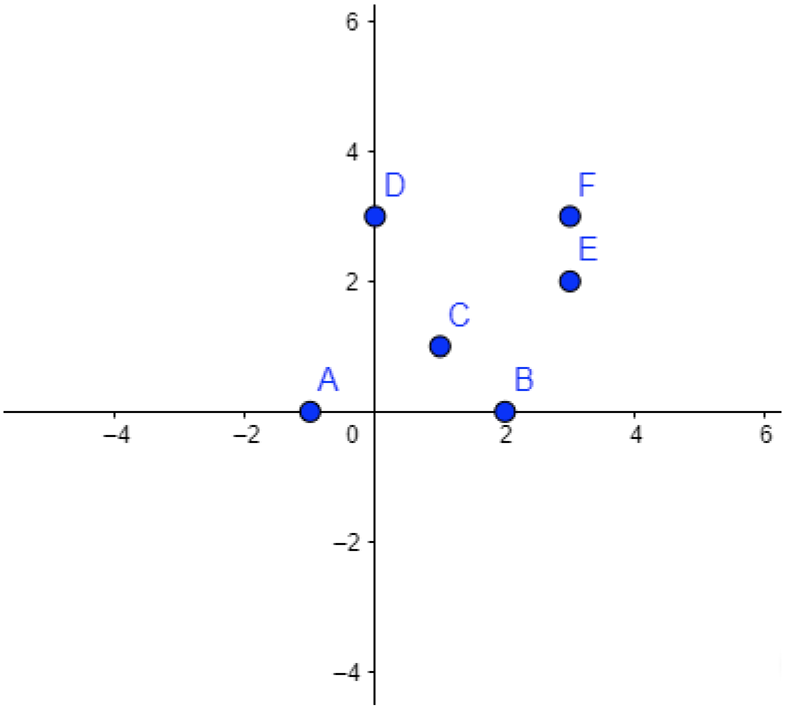
The distance between 2 points (x,y) and (a,b) in the 2-dimensional space is defined as:

Distance[(x,y), (a,b)] = $\sqrt{}(a-x)^2+(b-y)^2$

## Single Link

| Column1 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | | | | | |
| B | 3 | 0 | | | | |
| C | 2.236 | 1.414 | 0 | | | |
| D | 3.162 | 3.605 | 2.236 | 0 | | |
| E | 4.472 | 2.236 | 2.236 | 3.162 | 0 | |
| F | 5 | 3.162 | 2.828 | 3 | 1 | 0 |

We pick the lowest distance so we can form our first class. That is, between the two points F and E.



On our next step, we will cluster (group) the points E and F and calculate the distances of the remaining points from this cluster.

Since we are using the single link method, we are interested in the minimum distances between points, and clusters/points.

This actually is the minimum distance of one point from the two points that form the cluster.

So for example distance[A, (E, F)] = min[distance(A, E), distance(A, F)]

The updated distance matrix lies below:

| Column1 | A | B | C | D | (E, F) |
|---|---|---|---|---|---|
| A | 0 | | | | |
| B | 3 | 0 | | | |
| C | 2.236 | 1.414 | 0 | | |
| D | 3.162 | 3.605 | 2.236 | 0 | |
| (E, F) | 4.472 | 2.236 | 2.236 | 3 | 0 |

This time, the minimum value (distance) is 1.414. Which is the distance between points B and C. Hence, we form our second cluster: (B, C)

| Column1 | A | (B, C) | D | (E, F) |
|---|---|---|---|---|
| A | 0 | | | |
| (B, C) | 2.236 | 0 | | |
| D | 3.162 | 2.236 | 0 | |
| (E, F) | 4.472 | 2.236 | 3 | 0 |

Here, we find that the minimum distance is 2.236, but it is between many point choices. We are going to grab the one which forms the third cluster, which is ( (B, C), (E, F) ).

(We could had also clustered (B, C, A) and proceed on the new distances)

We calculate the distances once again.

| Column1 | A | B, C, E, F | D |
|---|---|---|---|
| A | 0 | | |
| B, C, E, F | 2.236 | 0 | |
| D | 3.162 | 2.236 | 0 |

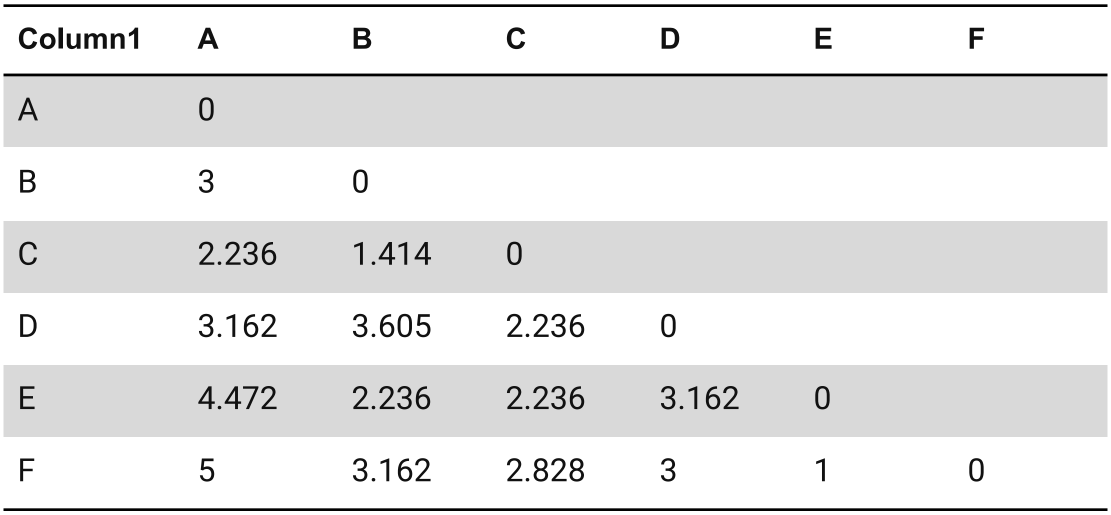
A and D has the same distance from the cluster, we pick D and we have:

| Column1 | A | B, C, E, F, D |
|---|---|---|
| A | 0 | |
| B, C, E, F, D | 2.236 | 0 |

For the single link method, we formed the first cluster to be the points E and F (E, F), the second one to be (B, C), then a bigger cluster including (B, C, E, F). Next we clustered that with point D to form the (B, C, E, F, D) cluster which we finally merged with point A.

## Complete Link

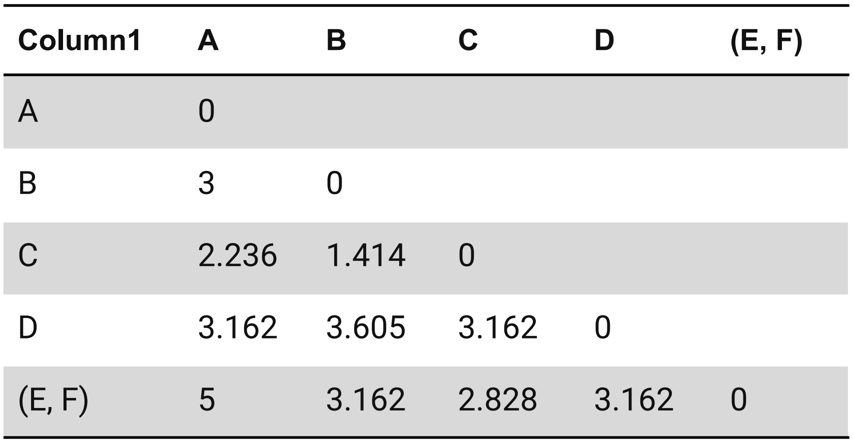
We start by doing the same thing, we calculate all the distances between every point. We have this distance table:



| Column1 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | | | | | |
| B | 3 | 0 | | | | |
| C | 2.236 | 1.414 | 0 | | | |
| D | 3.162 | 3.605 | 2.236 | 0 | | |
| E | 4.472 | 2.236 | 2.236 | 3.162 | 0 | |
| F | 5 | 3.162 | 2.828 | 3 | 1 | 0 |

Our next step is to find the minimum value, which will give us the minimum distance of the 2 points which will form our first cluster. This is the distance between E and F, with a distance of 1 unit.

We chose our first cluster to be points (E, F). This time we are using the complete link method, so we are interested in the **max** distances between points/clusters. We update the distance table:



| Column1 | A | B | C | D | (E, F) |
|---|---|---|---|---|---|
| A | 0 | | | | |
| B | 3 | 0 | | | |
| C | 2.236 | 1.414 | 0 | | |
| D | 3.162 | 3.605 | 3.162 | 0 | |
| (E, F) | 5 | 3.162 | 2.828 | 3.162 | 0 |

We look for the minimum value once again. That is 1.414 which is the distance between points B and C. Hence, our next cluster is formed to be the points (B, C).

Now we keep finding the max distance between points/clusters.

| Column1 | A | (B, C) | D | (E, F) |
|---|---|---|---|---|
| A | 0 | | | |
| (B, C) | 3 | 0 | | |
| D | 3.162 | 3.605 | 0 | |
| (E, F) | 5 | 3,162 | 3.162 | 0 |

The minimum value is from point A to cluster (B, C). We merge them into a cluster (B, C, A)

We do the same:

| Column1 | (B, C, A) | D | (E, F) |
|---|---|---|---|
| (B, C, A) | 0 | | |
| D | 3.605 | 0 | |
| (E, F) | 5 | 3.162 | 0 |

We chose the distance between point D and cluster (E, F) so we merge them into the cluster (E, F, D) and we update the table:

| Column1 | (B, C, A) | (E, F, D) |
|---|---|---|
| (B, C, A) | 0 | |
| (E, F, D) | 5 | 0 |

Our clusters were:

(E, F) -> (B, C) -> (B, C, A) and (E, F, D)

## Average link method

We present the distance table:

| Column1 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | | | | | |
| B | 3 | 0 | | | | |
| C | 2.236 | 1.414 | 0 | | | |
| D | 3.162 | 3.605 | 2.236 | 0 | | |
| E | 4.472 | 2.236 | 2.236 | 3.162 | 0 | |
| F | 5 | 3.162 | 2.828 | 3 | 1 | 0 |

We choose the lowest value, which is 1. This means we take the distance between points E and F. So we form our first cluster (E, F).

In order to update our distance table and because we are using the average link method, we have to find the average distance between a point and the points that form the cluster.

For example, for the point A and the cluster (E, F) we have:

distance(A, (E, F) = ½ (distance(A, E) + distance(A, F))

We apply this to update our distance table:

| Column1 | A | B | C | D | (E, F) |
|---|---|---|---|---|---|
| A | 0 | | | | |
| B | 3 | 0 | | | |
| C | 2.236 | 1.414 | 0 | | |
| D | 3.162 | 3.605 | 2.236 | 0 | |
| (E, F) | 4.736 | 2.699 | 2.532 | 3.081 | 0 |

We choose our minimum value which is 1.414, the distance between point B and C. Therefore our new cluster is the (B, C).

We update the distance table:

| Column1 | A | (B, C) | D | (E, F) |
|---|---|---|---|---|
| A | 0 | | | |
| (B, C) | 2.618 | 0 | | |
| D | 3.162 | 2.92 | 0 | |
| (E, F) | 4.736 | 2.615 | 3.081 | 0 |

Minimum value is 2.615 so our new cluster is (B, C, E, F)

| Column1 | A | (B, C, E, F) | D |
|---|---|---|---|
| A | 0 | | |
| (B, C, E, F) | 3.677 | 0 | |
| D | 3.162 | 3 | 0 |

Minimum value is 3, so we have (B, C, E, F, D)

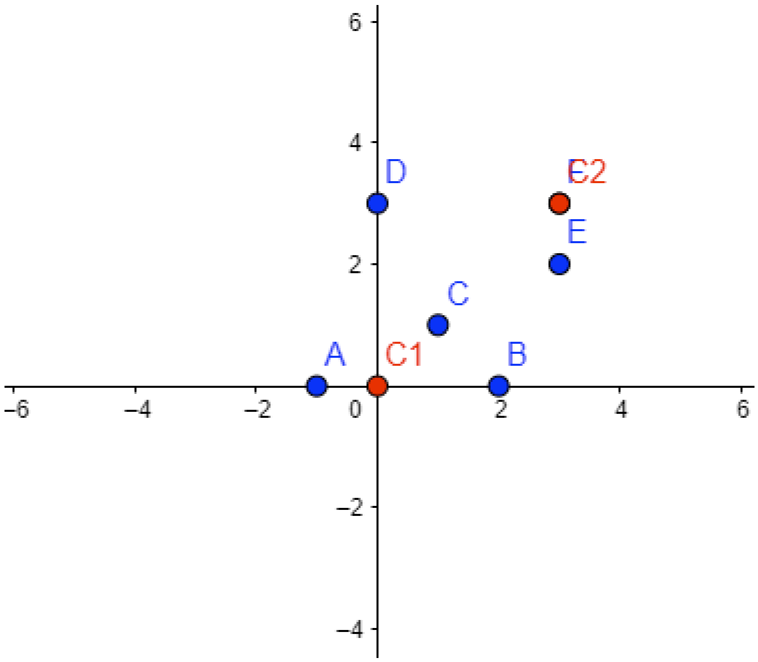
| Column1 | A | (B, C, E, F, D) |
|---|---|---|
| A | 0 | |
| (B, C, E, F, D) | 3.419 | 0 |

Comparing the 3 different methods, we can see that with single link, the distance between the final cluster and the remaining point is the least from all other 2 methods. When we used complete link, that distance was the highest (5).

b) Cluster the 6 entities using k-Means, for k=2 and initial centroids C1(0,0) and C2(3, 3) and for a maximum of 7 iterations. Compare results to these acquired at a) and discuss the selection of k and the initial centroids.

Our initial centroids are C1(0, 0) and C2(3, 3). We want to calculate the Euclidian distance between all points and both the centroids.

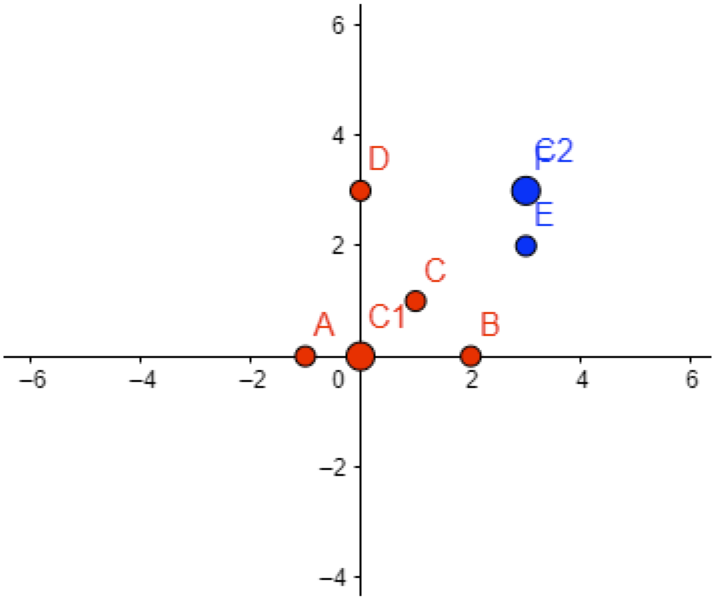
| Column1 | C1 | C2 |
|---|---|---|
| A | 1 | 5 |
| B | 2 | 3.162 |
| C | 1.414 | 2.828 |
| D | 3 | 3 |
| E | 3.605 | 1 |
| F | 4,242 | 0 |



K1 is the first cluster with centroid C1 and K2 is the second one with centroid C2.

We look into the distance table and we assign every point to the cluster which they have the less distance from.

K1 = {A, B, C, D}<br>
K2 = {E, F}



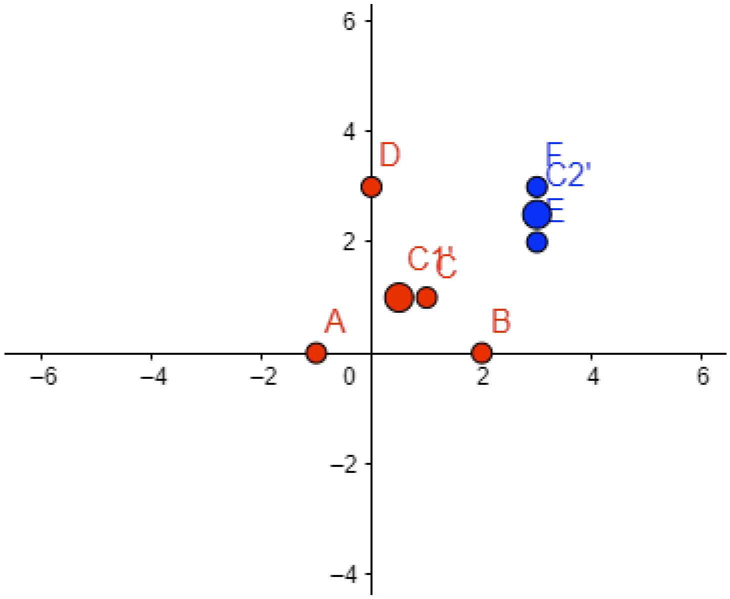
We now need to calculate our new Centroids by finding the mean from each cluster point.

So our new C1' will be
C1' = ¼ [ (-1 + 2 + 1 + 0, 0 + 0 + 1 + 3) ] = (0.5, 1)
C2' = ½ [ (3 + 3, 2 + 3) ] = (3, 2.5)

| Column1 | C1' | C2' |
|---|---|---|
| A | 1.802 | 4.716 |
| B | 1.802 | 2.692 |
| C | 0.5 | 2.5 |
| D | 2.061 | 3.041 |
| E | 2.692 | 0.5 |
| F | 3 | 0.5 |

Our clusters now become the following:

K1 = {A, B, C, D}
K2 = {E, F}



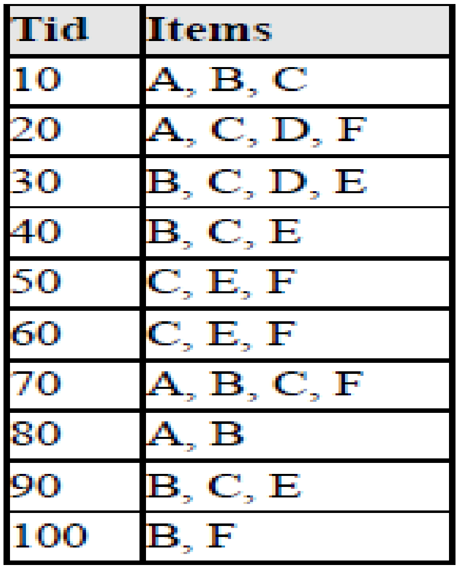
The new centroids C1” and C2” are equal to C1’ and C2’ respectively. So since there is no change in our centroids, we stop the algorithm here because it has converged.

Our final clusters are:

K1 = {A, B, C, D}
K2 = {E, F}

This may indicate that the choice of our initial Centroids might not have been optimal. If we chose different initial centroids, the distances would had been different, which would had lead in a more in-depth scan of our point, requiring more search and more changes in our clusters, with possibly more assignments of the new centroids.

**5. Given the following transactional database, produce manually association rules, using an algorithm of your choice (support=25% and confidence=75%). Also, produce manually association rules if b) support=25% and confidence=50% and c) support=35% and confidence=50%. Discuss your results.**

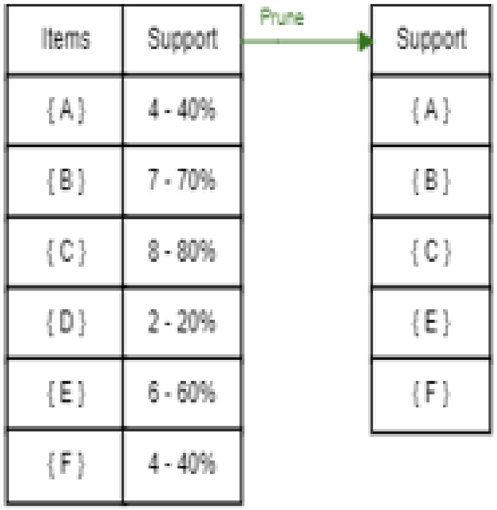


| Tid | Items |
|---|---|
| 10 | A, B, C |
| 20 | A, C, D, F |
| 30 | B, C, D, E |
| 40 | B, C, E |
| 50 | C, E, F |
| 60 | C, E, F |
| 70 | A, B, C, F |
| 80 | A, B |
| 90 | B, C, E |
| 100 | B, F |

For our transactional database, we will use the apriori algorithm to produce association rules.

*<u>1st scenario: support = 25%, confidence = 75%</u>*

We scan our transactional DB and see every item's frequency. We then proceed with removing the item with support (frequency) less than our threshold (25%)

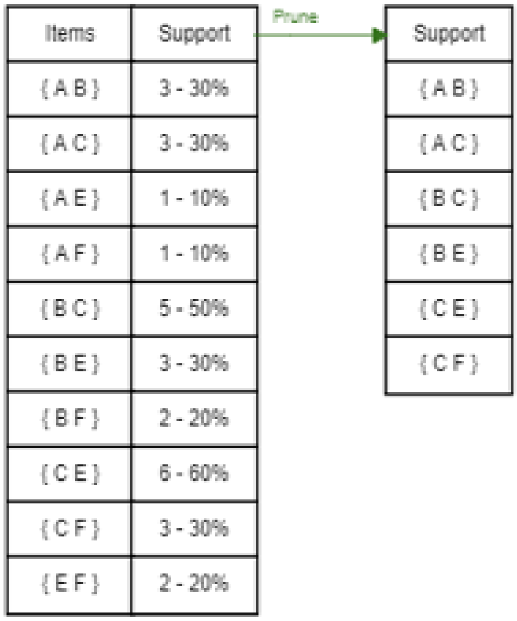


| Items | Support |
|---|---|
| {A} | 4 - 40% |
| {B} | 7 - 70% |
| {C} | 8 - 80% |
| {D} | 2 - 20% |
| {E} | 6 - 60% |
| {F} | 4 - 40% |

| Support |
|---|
| {A} |
| {B} |
| {C} |
| {E} |
| {F} |

We notice that {D} is shown in 2 out of 10 transactions (20% support), hence we remove it.

Our next step is to merge the itemsets with combinations of length 2.

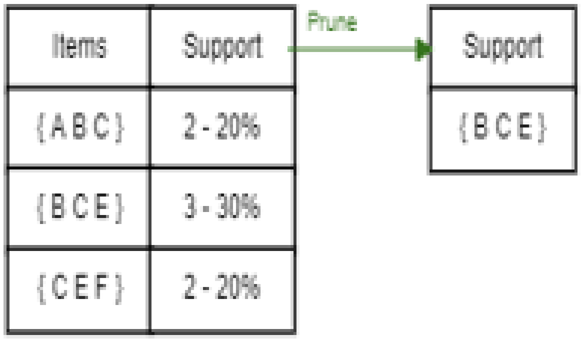


| Items | Support |
|---|---|
| { A B } | 3 - 30% |
| { A C } | 3 - 30% |
| { A E } | 1 - 10% |
| { A F } | 1 - 10% |
| { B C } | 5 - 50% |
| { B E } | 3 - 30% |
| { B F } | 2 - 20% |
| { C E } | 6 - 60% |
| { C F } | 3 - 30% |
| { E F } | 2 - 20% |

| Support |
|---|
| { A B } |
| { A C } |
| { B C } |
| { B E } |
| { C E } |
| { C F } |

After examining all the length-2 Itemsets, we scan them against our DB to find their corresponding support. Then, we remove the itemsets with support less than our threshold (20%).

We keep doing what we did before, merging the remaining itemsets.

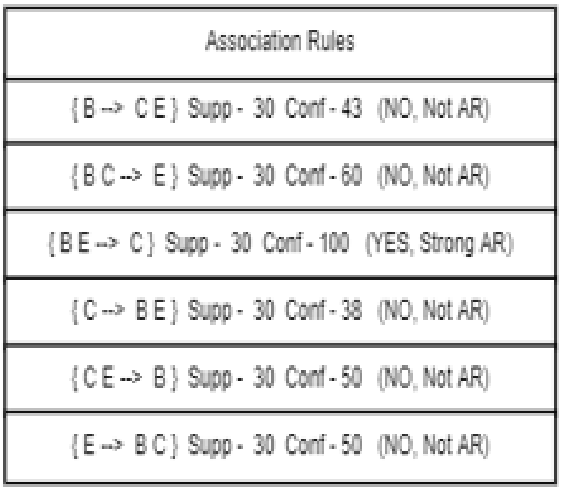


| Items | Support |
|---|---|
| { A B C } | 2 - 20% |
| { B C E } | 3 - 30% |
| { C E F } | 2 - 20% |

| Support |
|---|
| { B C E } |

As we can see, after our last scan, we end up with the itemset {B, C, E} with corresponding support of 30%. We can safely double-check our transactional database in order to confirm that this itemset indeed shows in 3 out of 10 transactions.

The Association Rules are presented in the following table.

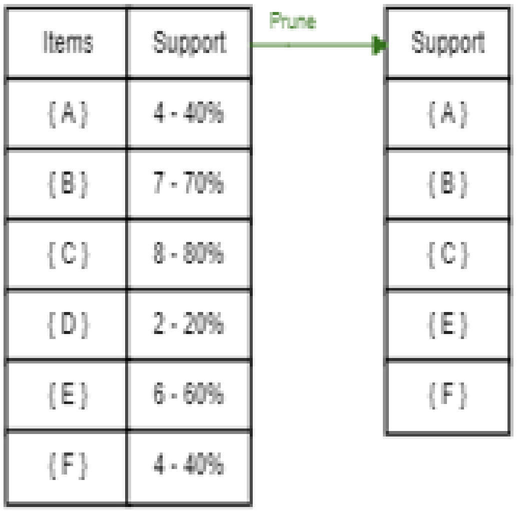


| Association Rules |
|---|
| { B --> C E } Supp - 30 Conf - 43 (NO, Not AR) |
| { B C --> E } Supp - 30 Conf - 60 (NO, Not AR) |
| { B E --> C } Supp - 30 Conf - 100 (YES, Strong AR) |
| { C --> B E } Supp - 30 Conf - 38 (NO, Not AR) |
| { C E --> B } Supp - 30 Conf - 50 (NO, Not AR) |
| { E --> B C } Supp - 30 Conf - 50 (NO, Not AR) |

***Discussion:*** In the first scenario, we can safely assume that whenever we have the items B and E in our transaction database, with 100% confidence, the item C will be present.

*2nd scenario: Support = 25%, confidence = 50%*

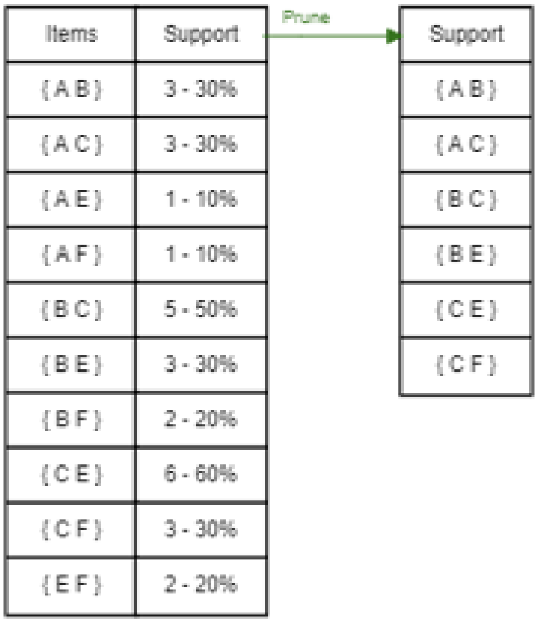
We start by doing the same thing as before. We check all our items as an itemlist of length 1 and we find their corresponding frequency.



| Items | Support |
|---|---|
| { A } | 4 - 40% |
| { B } | 7 - 70% |
| { C } | 8 - 80% |
| { D } | 2 - 20% |
| { E } | 6 - 60% |
| { F } | 4 - 40% |

| Support |
|---|
| { A } |
| { B } |
| { C } |
| { E } |
| { F } |

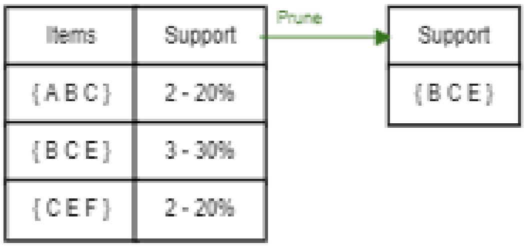
Since the minimum support threshold remains the same (25%), the frequent itemsets will remain the same.



| Items | Support |
|---|---|
| { A B } | 3 - 30% |
| { A C } | 3 - 30% |
| { A E } | 1 - 10% |
| { A F } | 1 - 10% |
| { B C } | 5 - 50% |
| { B E } | 3 - 30% |
| { B F } | 2 - 20% |
| { C E } | 6 - 60% |
| { C F } | 3 - 30% |
| { E F } | 2 - 20% |

| Support |
|---|
| { A B } |
| { A C } |
| { B C } |
| { B E } |
| { C E } |
| { C F } |

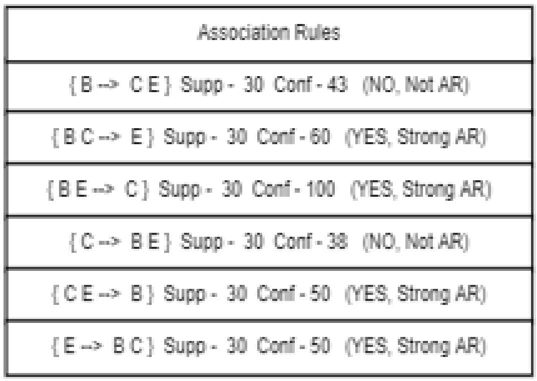
And then:



| Items | Support |
|---|---|
| { A B C } | 2 - 20% |
| { B C E } | 3 - 30% |
| { C E F } | 2 - 20% |

| Support |
|---|
| { B C E } |

Even though the support remains the same, in this scenario we have a confidence threshold of 50%, which will lead to us making different Association Rules. Let’s look at the table below:



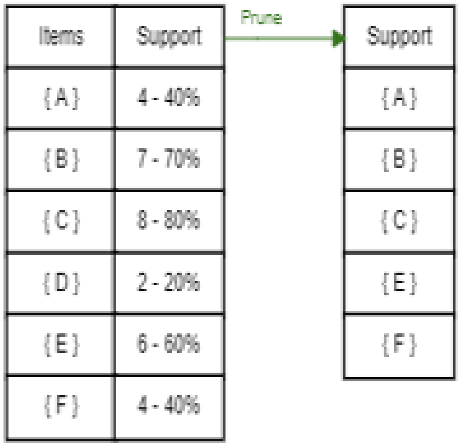
| Association Rules |
|---|
| { B --> C E } Supp - 30 Conf - 43 (NO, Not AR) |
| { B C --> E } Supp - 30 Conf - 60 (YES, Strong AR) |
| { B E --> C } Supp - 30 Conf - 100 (YES, Strong AR) |
| { C --> B E } Supp - 30 Conf - 38 (NO, Not AR) |
| { C E --> B } Supp - 30 Conf - 50 (YES, Strong AR) |
| { E --> B C } Supp - 30 Conf - 50 (YES, Strong AR) |

***Discussion:*** When we change our confidence threshold to 50% our association rules differ: 3/5 transactions which include items B, C tend to include item E as well. A stronger rule presents itself in the case of having B and E in the same item list. In that case, 3/3 transactions with items B and E also have item C in them. 3/5 transactions having items C and E, also include item B. Same thing when we have E, items B and C appear 60% of the time.

*<u>3rd scenario: Support = 35%, confidence = 50%</u>*

This time, the support threshold is higher than before, hence we would expect to see (probably) different itemsets making it throughout the scanning phase.

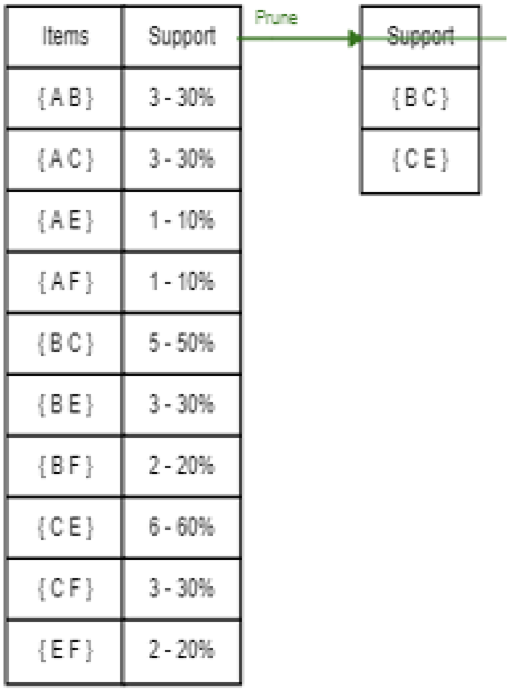
We proceed with the same first step as always:



| Items | Support |
|---|---|
| {A} | 4 - 40% |
| {B} | 7 - 70% |
| {C} | 8 - 80% |
| {D} | 2 - 20% |
| {E} | 6 - 60% |
| {F} | 4 - 40% |

| Support |
|---|
| {A} |
| {B} |
| {C} |
| {E} |
| {F} |

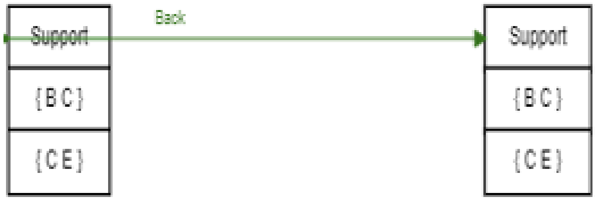
{D} is not frequent, once again, so we remove it. We continue our process:



| Items | Support |
|---|---|
| {A B} | 3 - 30% |
| {A C} | 3 - 30% |
| {A E} | 1 - 10% |
| {A F} | 1 - 10% |
| {B C} | 5 - 50% |
| {B E} | 3 - 30% |
| {B F} | 2 - 20% |
| {C E} | 6 - 60% |
| {C F} | 3 - 30% |
| {E F} | 2 - 20% |

| Support |
|---|
| {B C} |
| {C E} |

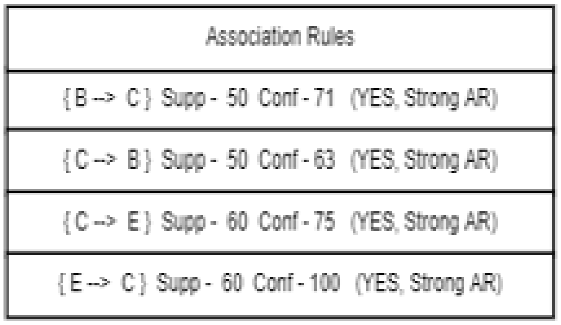
Since our minimum support has increased, we ought to remove more itemsets than we did before (support=25%). Eventually, we keep only the itemsets {B, C} and {C, E}.



| Support |
|---|
| {B C} |
| {C E} |

| Support |
|---|
| {B C} |
| {C E} |

So our Association Rules are the following:



| Association Rules |
|---|
| { B --> C } Supp - 50 Conf - 71 (YES, Strong AR) |
| { C --> B } Supp - 50 Conf - 63 (YES, Strong AR) |
| { C --> E } Supp - 60 Conf - 75 (YES, Strong AR) |
| { E --> C } Supp - 60 Conf - 100 (YES, Strong AR) |

***Discussion:*** B appears 7 times in our transactional database. 5 out of these 7 times, C belongs in the same transaction. 5/8 of the transactions having item C, item B is there as well, same applies for C->E. Lastly, when item E appears in our transactional database (5 times), all of those 5 times it goes along with item C which makes it a very strong association rule with 100% confidence.

# References

Tinder Millennial Match Rate | Kaggle

https://dumps.wikimedia.org/

http://www.cs.cmu.edu/~enron/

https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=analysis-group-overview

https://www.researchgate.net/publication/321996972_Social_media_analytics_-_Challenges_in_topic_discovery_data_collection_and_data_preparation#pfb

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-22171.pdf

https://www.sciencedirect.com/science/article/pii/S0148296322001321#b0450

https://link.springer.com/article/10.1007/s00146-014-0549-4