Data mining 2017 autumn project topics

Formal requirements

- Teams of 2-3 students, at least 25 hours of work per student
  - Team can have 1 student if working at least 50 hours and if this is agreed beforehand with the course instructors
- Every team must insert the project's title, description and team members into the List of Projects document (the link will be provided in Piazza by Tuesday, 5th December)
- Every team must present the project as a poster in the poster session on Jan 8, 2018
- Every team must provide access for the grading instructor to the project code repository hosted either at Github or BitBucket page
- The project will be graded by the instructors and will get maximally 20 points
- If project gets X points then each team member gets X points
- Getting at least 10 points for the project is a prerequisite for passing the course

Evaluation of projects

- Projects will be evaluated during and after the poster session on Jan 8, 2018
- The grade consists in the following two factors
- Technical quality (10 points)
  - Your project can get 10 points for technical quality, if you have:
    - stated clear objectives
    - applied relevant data mining methods on relevant data, and
    - the achievements are sufficient, considering the required working hours
- Presentational quality (10 points)
  - Your project can get 10 points for presentational quality, if the poster is:
    - explaining well the motivation and objectives of the project
    - describing the used data mining methods so that others could in principle replicate your work
    - presenting the main results of your work in a visually appealing and understandable way
  - and if your project code repository (GitHub or BitBucket) contains:
    - all the code as well as readme-files describing what the code does and how to run it

Instructions

- There are following options for choosing the topic of your project:
  - Choose a topic proposed by companies
  - Choose a topic from Social Impact Data Hack 2017 event
  - Choose a topic from Kaggle
  - Choose your own topic
- For each of those options the formal requirements differ slightly, so please look at the dedicated slide below
- There can be multiple projects on the same topic, but be aware of the following:
  - All code/analysis/results that are very similar or shared between multiple projects must be declared in a private message to the instructors in Piazza before the poster session
  - When assessing whether you have worked enough we will split the size of the task (hours) in very similar or shared parts between the teams in equal proportions
- Therefore, if multiple projects are working on the same topic then these projects should communicate between each other to be able to declare the overlap and avoid duplication

Choosing a topic proposed by companies

- To make your projects as interesting and relevant as possible, we have sent a call for project ideas to various companies and organisations
- We have obtained data and topics from the following companies:
  - Starship (https://www.starship.xyz/)
  - Rate-chain (http://rate-chain.com/)
  - Fitlap (https://www.fitlap.com/)
  - Elektrilevi (https://www.elektrilevi.ee/en)
  - DuNord (http://www.dunord.ee/)
- All these companies have given a contact person whom you can approach with potential questions about the data while working on their project
- The datasets have been given exclusively for our data mining projects and are not publicly available, hence we believe that these are very interesting to work with.
- In order to avoid everyone choosing these projects there will be an application process for these topics, see the next slide about this.

Applying for a topic proposed by companies

- If you want to do a project proposed by a company then please send a private message to instructors in Piazza with the tag ‘project’ and with the title ‘Application for project <project-name>’ with deadline 23:59 on Tuesday, 5th December. This message must include the list of team members and a brief description of your motivation and relevant skills. If you already have ideas or plans then please briefly describe these as well.
- If you are interested but you don't have a team yet, then you can still apply. Most probably the preference would be given to existing teams but individuals might also get a chance.
- Please try to follow Piazza on Wednesday, 6th December, as we might have additional questions to you.
- We will either approve or decline your application by 23:59 on Wednesday, 6th December.

Choosing a topic from Social Impact Data Hack

- For the event Social Impact Data Hack 2017 (SIDH'17) many topics were proposed and you can choose a topic from page http://sidh2017.ut.ee/ideas/.
- You cannot choose topics 1, 4, 8, 11, 12, 13, 16, 17, 23, because these topics were already chosen at SIDH'17 and it would be hard for us to grade your project due to overlap with existing work
- Additionally, you can choose your own topic based on the datasets listed at http://sidh2017.ut.ee/datasets/, as long as it is not significantly overlapping with topics 1, 4, 8, 11, 12, 13, 16, 17, 23 as explained above.

Choosing a topic from Kaggle

- Kaggle (https://www.kaggle.com/) hosts many interesting machine learning competitions and datasets
- You can choose to compete in one of the challenges (https://www.kaggle.com/competitions) or work on one of the datasets (https://www.kaggle.com/datasets)
- In each of the Kaggle challenges and competitions there are available ‘kernels’ where people have explained a way to analyze the data and have provided their code
- You must declare all kernels that you use in your project
- Furthermore, you must submit your project as a new kernel in Kaggle
    - The only exception to this is when you choose an ongoing competition and are able to achieve top 20% with your team

Choosing your own topic

- You can freely choose your own topic, as long as it requires you to demonstrate mastering some topics from the data mining course
- In your readme-files of your code repository page you must specify the origin of the data and provide a short description of the data
- The data do not need to be public but the grading instructor must be able to see the data, at least from your computer
- Here are some more ideas to help you:
  - https://opendata.riik.ee/et/dataset/avaliku-korra-vastased-ja-avalikus-kohas-toime-pandud-syyteod (criminal offences in Estonia against the public order and against property committed in a public place)
  - https://opendata.riik.ee/en/dataset/liiklusjarelevalve-alased-syyteod (criminal offences discovered during traffic supervision in Estonia)
  - More datasets from http://opendata.riik.ee
  - https://www.wikidata.org/
  - http://openclimatedata.net/
- Finally, have a look at the data mining project posters from this spring
  https://courses.cs.ut.ee/2017/dm/spring/Main/Projects

Topics from companies

RATECHAIN-1: Car rental demand detection based on price requests from online channels

RateChain provides price management solution in online channels for car rental companies.

Our customers are car rental companies from Iceland, India, Panama, Poland, Jamaica, Cyprus, Romania, Portugal, Turkey etc. Car rental companies have defined price rules in our system and RateChain calculates price quote for each request from online distributor. For each request we know request time, pickup date, return date, pickup location, customer source country and driver age. Today we are returning ca 2 million rate quotes per day on behalf of our customers.

By analysing and comparing price quote data and reservations history we look answers for following questions:

- To which time period customers are looking for rental cars now/yesterday/last week/last month?
- How many days in advance customers are checking prices? How it changes through seasons and years?
- How many days/weeks/months in advance bookings are made? Does it vary based on season and years?
- Are there unexpected changes (peaks, drops) in bookings or price requests?
- How many price requests could be cached (have same pickup and return location, source country and driver age) for 1 hour, 3 hours, 12 hours?
- What is actual demand after eliminating duplicated requests?
- What is “rate quotes to reservation” ratio by resellers? Does it vary by seasons or source countries?
- What is “rate quotes to reservation” ratio by car class and location?

Can we use rate quotes to reservation ratio to derive historical demand based on reservations history?

RATECHAIN-2: Market price analysis to compare car rentals offers

RateChain Solutions has solution for car rental companies to manage their rates in online distribution channels. RateChain collects market price data from different websites to monitor car rental prices in our customers' locations. Today the data set contains ca 20 million data points and every day ca 90 000 data points are added.

Car rental companies are providing prices for a car class not for a specific car model. Usually ACRISS classification codes (http://acriss.org/car-codes.asp) are used to describe the car class but there are no specific instructions in which class a car model should belong. Different car rental companies may provide different ACRISS code for the same car model. Therefore, it is not enough to use only ACRISS code if we want to compare offers. Also, we should consider that certain car models are equal alternatives for customer even if assigned ACRISS codes are different.

Based on this dataset we would like to find answers to following questions.

- How to classify car rental offers more accurately for price comparison by using collected data?
- Which is "correct" ACRISS code for a car model based on collected data? Does it vary by countries?
- Which car models can be considered as alternatives to each other from pricing perspective?
- How manual and automatic gearbox affects rental price by car class?
- What is price difference between car classes?
- How much price per day varies based on rental duration?

Which car rental companies are using active pricing (changing prices more often than others)?

ELEKTRILEVI-1: Quality of electricity

The dataset consists of 10 minute mean RMS value measurements from 43 primary substations with the measurement period of one year (52560 timestamped points per substation). There are different voltage and current parameters, I added a description file with brief descriptions what these values mean. I also added power quality events data for each substation that the measuring device has recorded.

Antud mõõtesüsteem on paigaldatud kahel eesmärgil – lühiajaliste rikete ja sündmuste kaardistamine ja pikaajaliste trendide jälgimine. Praegusel hetkel on projekti jaoks välja jagatud andmeid pikaajalises vaates, sellest tulenevalt ka 10 minuti rms väärtused. Kuigi mõõteseadmed mõõdavad suure sagedusega (512 samplit/20ms), siis hetkel veel puudub võimekus mõõteandmeid sellise täpsusega pika aja perioodi peale välja võtta. Hetkel Elektrilevi uurib, et kuidas neid andmeid kätte saada, siis saaksid nad juba täpsemalt võrgus toimuvaid sündmusi suure plaaniliselt analüüsida ja jälgida.

Ise oleme pingeandmeid analüüsinud, et leida vastavalt standardis (EN50160) toodud väärtustele mõõtepunktide nädalate pingekvaliteediindeksit, mis siis annab ettekujutuse kui hea/halb meie võrgu pingekvaliteet mingis mõõtepunktis on. Ühe sarnase use case'ina võiks uurida, kas nad oskavad välja pakkuda mingit võrguseisundi indeksit, mida saaks perioodiliselt arvutada ja mõõta, kas asjad lähevad paremaks või halvemaks. Lisaks oleks huvitav näha, kas andmeanalüütika abiga on võimalik leida mingeid anomaaliaid või trende mis viitaks, et võrgus või alajaamas midagi ei toimi päris nii nagu peaks ja tuleks üle vaadata ehk mingit sorti ennustav analüüs. ELV vaates oleks hea kui tänu neile mõõtmistele suudame ennetada võrgu või alajaama seadmete rikkeid ja läbi selle hoiame kokku investeeringutelt. Kuid tudengitele võib ka vabad käed anda.

Events data: 43 substations, altogether 19017 rows, 11 columns, head of data:

```
"";"ID";"Name";"Phase";"Severity";"Start.Time";"Duration";"End.Time";"Value";"Units";"Deviation"
"1";3131769;"Voltage Dip              ";"12,23,31";255;"04/07/2017 21:40:27.779";"00:00:07.8309001";"04/07/2017 21:40:35.610";1,417114;"Volt ";"99.95117%"
"2";3304106;"Voltage Dip              ";"12,23,31";255;"08/08/2017 09:53:17.062";"00:00:21.2275390";"08/08/2017 09:53:38.290";0,1637192;"Volt ";"99.95117%"
"3";3334502;"Voltage Dip              ";"12,23,31";255;"14/08/2017 11:49:00.342";"00:00:21.2627220";"14/08/2017 11:49:21.605";1,195923;"Volt ";"99.95117%"
"4";3339426;"Voltage Dip              ";"12,23,31";209;"15/08/2017 11:21:18.443";"00:00:00.7998046";"15/08/2017 11:21:19.243";1369;"Volt ";"87.54883%"
"5";3049897;"Voltage Dip              ";"12,23,31";209;"18/06/2017 12:19:17.325";"00:00:00.5797144";"18/06/2017 12:19:17.904";1284,438;"Volt ";"88.28125%"
"6";3049901;"Voltage Dip              ";"12,23,31";209;"18/06/2017 12:45:13.091";"00:00:00.5699493";"18/06/2017 12:45:13.661";1284,063;"Volt ";"88.28125%"
"7";3339428;"Voltage Dip              ";"12,23,31";207;"15/08/2017 11:43:18.632";"00:00:00.5893980";"15/08/2017 11:43:19.221";1394,313;"Volt ";"87.30469%"
"8";2340101;"Frequency Out of Range   ";"       ";189;"28/01/2017 18:58:54.662";"00:00:10       ";"28/01/2017 18:59:04.662";55,75195;"Hz   ";"11.47461%"
"9";2340106;"Frequency Out of Range   ";"       ";179;"28/01/2017 18:58:33.747";"00:00:10       ";"28/01/2017 18:58:43.747";51,32888;"Hz   ";"2.636719%"
```

PQ data: 43 substations, each has 52561 rows, head of data:

```
"";"Period.Start";"Period.End";"PLT.L12";"PLT.L23";"PLT.L31";"Sys.Frequency";"K.Factor.L1";"K.Factor.L2";"K.Factor.L3";"Active.Power.L1";"Active.Power.L2";"Active.Power.L3";"Reactive.Power.L1";"Reactive.Power.L2";"Reactive.Power.L3";"RMS.Current.L1";"RMS.Current.L2";"RMS.Current.L3";"RMS.Voltage.L12";"RMS.Voltage.L23";"RMS.Voltage.L31";"Apparent.Power.L1";"Apparent.Power.L2";"Apparent.Power.L3";"Asymmetria.Voltage";"THD.Current.L1";"THD.Current.L2";"THD.Current.L3";"THD.Voltage.L12";"THD.Voltage.L23";"THD.Voltage.L31"
"1";"01.11.2016 00:00:00.000000";"01.11.2016 00:10:00.000000";0,3166066;0,3228025;0,317051;49,99787;1,0704;1,050938;1,059978;-655848,6;-738451,4;-602845,4;-95507,5;-195236,1;-212599;192,3195;192,611;184,5825;6288,814;6260,687;6250,239;662856,9;763915,9;639284,3;0,3704743;4,485761;3,685225;3,945889;0,6331791;0,5195504;0,6488776
"2";"01.11.2016 00:10:00.000000";"01.11.2016 00:20:00.000000";0,3050334;0,3124813;0,3059823;49,99787;1,0704;1,050938;1,059978;-655848,6;-738451,4;-602845,4;-95507,5;-195236,1;-212599;192,3195;192,611;184,5825;6288,814;6260,687;6250,239;662856,9;763915,9;639284,3;0,3704743;4,485761;3,685225;3,945889;0,6331791;0,5195504;0,6488776
```

ELEKTRILEVI-2: Cables, transformers, poles

There are multiple datasets below, these can be part of one or several projects:

1. The intention is to understand the cable reliability, and when in their age process they start to show more faults. However, we use number of joints that have had to be installed as a proxy for the cable condition. The more problems a cable type has, the more joints has had to be installed. Total: 47308 rows, 10 columns, head of data:

```
JID,CID,CLASSID,CABLENO,CLASSNAME,FEEDERID,AREA,SIDE,JYEAR,CYEAR
214454630,214454622,3941,810230,AXLJ-TT  3x50-16 24 kV,2052948,4,2,2011,2011
217967253,217967229,3950,710218,AXLJ-LT.3x50+16 24kV,4927808,4,2,2013,2013
215791696,199660432,3515,510092,AHXAMK-W.3x120+35Cu 20kV,2041764,3,2,2012,2012
215176183,215176119,3961,710202,AXAL-TT PRO 3x50,4881281,4,2,2012,2012
199783141,199783139,3950,810219,AXLJ-LT.3x50+16 24kV,99492197,4,2,2010,2010
```

2. We have the list of all the observations made on transformers, through routine inspection which happens roughly every five years for our assets. What are the connections between transformer types, their age and the types of defects we are likely to see!? This is the information related to outages as a result of failures on power lines in the years 2014-2016. Many questions can be asked about the reasons for outages, the areas, the timing etc. Total: 29983 rows, 14 columns, head of data:

```
"ID","JAOTUSALAJAAM","REPLDATE","CONYEAR","CONNECTIONDATE","REGION","OWNER","MANUFTYPE","OBSID","OBSERVATIONDATE","NAME","ENUMVALUE","DESCR","URGENCY"
5294931,"Heina:(P-Jaagupi)",13-NOV-98,1985,01-JAN-85,"P�rnu-Jaagupi","O� Elektrilevi","TOTSA 50",18229489,11-FEB-17,"Trafo �li",2,"�lin�idik mustunud, ei n�e taset",1
9626141,"Kimalase:(M�niste)",,1971,01-JAN-71,"M�niste","O� Elektrilevi","TM 400/10",18232620,10-FEB-17,"Trafo �li",0,"korras",0
21014635,"Silikaadi:(V�ru)",01-NOV-07,1989,,"V�ru","O� Elektrilevi","TNOSCTES 400",18229932,10-FEB-17,"Trafo isolaatorid",0,"korras",0
4018618,"Pol�goni:(Turba)",18-JAN-00,1989,,"Turba","O� Elektrilevi","TMG 400",18220353,09-FEB-17,"Trafo �li",0,"korras",0
4018618,"Pol�goni:(Turba)",18-JAN-00,1989,,"Turba","O� Elektrilevi","TMG 400",18220352,09-FEB-17,"Trafo isolaatorid",0,"korras",0
4018618,"Pol�goni:(Turba)",18-JAN-00,1989,,"Turba","O� Elektrilevi","TMG 400",18220351,09-FEB-17,"Trafo v�line seisukord",0,"korras",0
9334882,"Vaima IIs:(P�lva)",,1980,01-JAN-80,"P�lva","O� Elektrilevi","TM 400",18218743,09-FEB-17,"Trafo v�line seisukord",1,"tolmunud",2
```

3. We have the list of observation made on poles, with the result that the poles must be changed. Observation year and the age of pole on the year of observation is recorded. What is the lifespan of each pole type, and at what age you would retire each type? Total: 12366 rows, 5 columns, head of data (with types of poles: 602 - KP puitmast, 604 - KP raudbetoonmast, 650 - MP puitmast, 651 - MP betoonmast)

```
"ID","OBJECTID","CLASSID","AGE","OBSYEAR"
11666580,153428,604,54,2014
17203595,185378,604,48,2016
17784786,155921,650,46,2016
17784864,155923,650,46,2016
17220143,155108,650,46,2016
11823403,187180,604,34,2014
11819467,188540,602,34,2014
```

ELEKTRILEVI-3: Power outages

Information about power outages in 2014-2016. Provided as an XLSX file with spreadsheets Rikkeobjektid (6868 rows), Liinid (46639 rows), Katkestused (94934 rows) and some extra spreadsheets describing the meaning of categorical variables (Katkestustüübid, Katkestuspõhjused, Katkestusklassid, Piirkonnad, Varaklassid). Heads of spreadsheets Rikkeobjektid, Liinid, Katkestused:

ID_Katkestus	ID_Objekt	ID_Klass	Objekti vanus
203726	956367	1005	37
203742	409547	1004	43
203747	-1237742	1003	
203802	1282528	1015	44
203812	5307047	1004	41
203828	-53504689	1585	39

ID_Fiider	ID_Klass	ID_Piirkond	Liini vanus	Liini pikkus
2052849	3961	5	7	1686
2052849	1004	4	31	2970
4927636	1004	4	18	3351
2052803	1005	4	27	422
2052948	3941	4	4	451
2052803	1035	4	9	4

ID_Katkestus	ID_Katkestusklass	ID_Katkestuspõhjus	ID_Katkestustüüp	Katkestuse kuupäev	Katkestuse kestus
450099	2	104	210	02-Jan-14	33
450151	2	211	217	02-Jan-14	17
450169	2	220	101	02-Jan-14	34
450186	2	102	101	02-Jan-14	44
450563	1	220	208	07-Jan-14	29

We are looking at the outage data for customers in 2015. Customers experiencing more than 72 hours of continuous outage will be compensated. That means a single outage is either longer than 72 hours, or two or more consecutive outages have less than two hours between them and add up to 72 hours. Explore the outages that need to be compensated. Total: 15682 rows, 5 columns, head of data:

```
"EIC","START","END","DURATION","LAG"
"00021274-8",2015-09-03 22:52:38,2015-09-04 01:11:45,139.116666666667,
"00021274-8",2015-09-04 01:11:45,2015-09-04 01:24:19,12.5666666706403,0
"00021274-8",2015-09-04 01:24:21,2015-09-04 01:37:28,13.116666674614,0.0333333293596903
"00021274-8",2015-09-04 01:38:09,2015-09-04 13:02:46,684.61666667064,0.683333325386047
"00021274-8",2015-09-04 13:02:48,2015-09-04 15:06:38,123.833333333333,0.0333333293596903
"00021274-8",2015-09-04 15:06:38,2015-09-04 23:45:24,518.78333332936,0
```

STARSHIP: Measuring pavement quality in Tallinn

The goal of this project is to study pavement quality in Tallinn using 1 month of Starship delivery robot data. The dataset includes sensor data from the delivery robots and additional localization data. Data from the delivery robots includes robot id, timestamp, orientation, readings from the accelerometer(s), magnetometer(s) and gyroscope(s) and some additional fields, all represented in the JSON format. The provided dataset is more than 100GB because it covers many cities. However, it can be easily reduced to much smaller sizes by narrowing down to data only from Tallinn. Localization data is only from Tallinn. Putting these together it is possible to study how delivery robot's sensors act on a particular road in Tallinn, potentially visited by multiple delivery robots during this time period. An example row from the sensor data is the following:

{"meta": {"botid":"6D80","secs":1498913579,"nsecs":917155995},
"data":{"stamp":{"secs":1498913579,"nsecs":901259179},"orientation_delta":{"x":4.231929779052734e-06,"y":6.126239895820618e-06,"z":-2.557411789894104e-06,"w":1},"accel_vec":{"x":-0.004735172260552645,"y":0.007237337529659271,"z":-1.00299870967865},"magnetometer_azimuth":2.708369493484497,"magnetometer_azimuth_updated":true,"standstill_detected":true,"gyro_model_name":"mpu6050","nr_of_active_gyros":2,"estimated_gyro_stdev_per_sec":6.170670530991629e-05,"estimated_gyro_systematic_error_stdev_per_sec":2.801343180180993e-05,"estimated_gyro_sensitivity_error_stdev":0.005236493423581123,"gyro_iio_bytes":{"layout":{"dim":[],"data_offset":0},"data":[]},"accel_iio_bytes":{"layout":{"dim":[],"data_offset":0},"data":[]},"magnetometer_iio_bytes":{"layout":{"dim":[],"data_offset":0},"data":[235,255,63,0,239,255]},"invensense_gyro_iio_bytes":{"layout":{"dim":[],"data_offset":0},"data":[150,15,229,111,2,0,128,0,64,178,68,158,253,118,244,33,30,60,149,245,53,205,20,0,32,217,255,13,0,219,255,33,30,60,149,245,53,205,20,0,64,102,68,132,253,100,244,72,195,212,149,245,53,205,20,0,32,218,255,13,0,220,255,72,195,212,149,245,53,205,20,0,64,150,68,194,253,114,244,124,104,109,150,245,53,205,20,0,32,218,255,14,0,219,255,124,104,109,150,245,53,205,20,0,64,102,68,150,253,122,244,171,13,6,151,245,53,205,20,0,32,219,255,13,0,220,255,171,13,6,151,245,53,205,20,3,0,128,0,64,100,197,166,253,20,240,255,230,15,149,245,53,205,20,0,32,213,255,255,255,201,255,255,230,15,149,245,53,205,20,0,64,136,197,214,253,82,240,57,82,168,149,245,53,205,20,0,32,213,255,255,255,200,255,57,82,168,149,245,53,205,20,0,64,104,197,146,253,2,240,104,189,64,150,245,53,205,20,0,32,213,255,255,255,200,255,104,189,64,150,245,53,205,20,0,64,136,197,188,253,254,239,134,40,217,150,245,53,205,20,0,32,213,255,254,255,200,255,134,40,217,150,245,53,205,20]},"reference_gyro_iio_bytes":{"layout":{"dim":[],"data_offset":0},"data":[]}}}

Localisation data from Tallinn has about 5 million rows and 6 columns, head of the table is the following:

botid,timestamp,coordinates_long,coordinates_lat,heading,stdev
6E5,1499097601.454,24.663108,59.397973,-0.716641,0.027637
6E5,1499097602.063,24.663102,59.397976,-0.724855,0.031432
6E5,1499097602.262,24.663096,59.397979,-0.728302,0.026824
6E5,1499097602.726,24.663091,59.397982,-0.719937,0.029236
6E5,1499097603.255,24.663083,59.397987,-0.70824,0.031799

DUNORD: Demand prediction for Liivi 2 cafeteria

The goal of this project is to predict the sales quantities in the Du Nord cafeteria at J. Liivi 2 (next to our lecture room). Good predictions can help the cafeteria to order more precise amounts of food, making sure that there is enough food ordered and not too many leftovers remaining from each day. Currently the staff of the cafeteria makes predictions based on their experience and based on the lecture plan in room 111. The objective is to come up with an automatic method based on a machine learning model. To train the model we are providing full sales data for past 2 years (2016-17), accompanied with the lecture room occupancy information from all of Liivi 2 building, kindly provided by the administrators of the study information system (Kersti Roosimäe).

The sales data specifies for each hour how many items were sold in each category. Examples of categories are: Külmad joogid, Leib, sai, Lisandid praadidele, Pagaritooted, Praed linnulihast, Praed sealihast, Taimetoidud, Praed veiselihast, Šokolaad ja maiustused, Soojad joogid, Supid. For instance, on October 30 during 12:00-13:00 (after the data mining lecture) there were 11 items sold from the category Praed sealihast and 3 items from the category Praed veiselihast. The dataset from the study information system says that in addition to our data mining lecture in room 111 with 126 registered students there are many other classes during the same time 10:15-12:00, including practice session of group 2 in MTMM.00.327 in room 206 with 29 students registered.

FITLAP

Project FITLAP-1: What are the habit patterns and parameters of those who lose the most weight with Fitlap?

What kind of a person is most likely to lose weight with Fitlap? And what kind of a person is most likely not to lose weight with Fitlap? What are the reasons for both situations. Are there any conclusions to be made?

Project FITLAP-2: Why people quit Fitlap and when do they quit?

Is there some kind of conclusions to be drawn here? For example if we see that people tend to drop out the most after 2 weeks and the reason for that is that they haven´t lost any weight, we could start autocommunication on that day to help them out more regarding that issue. Or if the weight number hasn´t changed in a period of 3 consecutive weighing, people tend lose motivation and so on.

Project FITLAP-3: Behavioural patterns and habits of different personas in Fitlap.

How different people/personas use Fitlap? What are the differences? It could be age related, weight related, related to client blogging activity etc. Sky is the limit.

Project FITLAP-4: Take a look at our collected data and propose your own project topic that would help Fitlap to serve their clients better, so that they would meet their goals

Questions or comments?

Please ask in Piazza under the tag ‘project’.