# **Title** **Health, medical, epidemiological**   **data mining, statistical modeling**

Kepzesi\_szint: MSc

Temavezeto: Molnár Bálint

Email: molnarba@inf.elte.hu

Temakor: Data mining, data analysis, statistics

There is a medical, biological, epidemiological collection of more than ten years

a set of data to examine which models are suitable

draw conclusions for pre-notification

There is a data set, empirical (medical, etc.) data.

We should find a statistical, data mining model,

with which to support proven results.

Beginning with descriptive statistics (correlation, normalization of input data

etc )

Find predictive models for validation purposes.   That is, a regression model

compare actual data and forecasts.

Using competitive alternative models (eg neuron nets), they are

forecasting ability, fact and predicted data

in order to support the model's predictive capability.

Data Mining: Concepts and Techniques: Concepts and Techniques

  by Jiawei Han, Micheline Kamber, Jian Pei

[https://books.google.hu/books?id=pQws07tdpjoC&pg=PA279&dq=data+mining+concepts+and+techniques&hl=hu&sa=X&redir\_esc=y#v=onepage&q=data%20mining%20concepts%20and%20techniques&f=false](https://translate.googleusercontent.com/translate_c?depth=1&hl=hu&ie=UTF8&prev=_t&rurl=translate.google.com&sl=hu&sp=nmt4&tl=en&u=https://books.google.hu/books%3Fid%3DpQws07tdpjoC%26pg%3DPA279%26dq%3Ddata%2Bmining%2Bconcepts%2Band%2Btechniques%26hl%3Dhu%26sa%3DX%26redir_esc%3Dy&usg=ALkJrhj--g2bPDT9uFoUZjZkO6du9miosg" \l "v=onepage&q=data mining concepts and techniques&f=false)

Data Mining for Business Intelligence: Concepts, Techniques, and ...

  author: Galit Shmueli, Nitin R. Patel, Peter C. Bruce

[https://books.google.hu/books?id=6WYV\_ACVFuQC&printsec=frontcover&dq=data+mining+concepts+and+techniques&hl=hu&sa=X&redir\_esc=y#v=onepage&q=data%20mining%20concepts%20and%20techniques&f=false](https://translate.googleusercontent.com/translate_c?depth=1&hl=hu&ie=UTF8&prev=_t&rurl=translate.google.com&sl=hu&sp=nmt4&tl=en&u=https://books.google.hu/books%3Fid%3D6WYV_ACVFuQC%26printsec%3Dfrontcover%26dq%3Ddata%2Bmining%2Bconcepts%2Band%2Btechniques%26hl%3Dhu%26sa%3DX%26redir_esc%3Dy&usg=ALkJrhjcqcxJEdH0upFelSY1HaNEbrNVDw" \l "v=onepage&q=data mining concepts and techniques&f=false)

Handbook of Statistical Analysis and Data Mining Applications

  by Gary Miner, Robert Nisbet, John Elder IV

[https://books.google.hu/books?id=U5np34a5fmQC&pg=PA4&dq=statistics+data+mining+concepts+and+techniques&hl=hu&sa=X&redir\_esc=y#v=onepage&q=statistics%20data%20mining%20concepts%20and% 20techniques & f = false](https://translate.googleusercontent.com/translate_c?depth=1&hl=hu&ie=UTF8&prev=_t&rurl=translate.google.com&sl=hu&sp=nmt4&tl=en&u=https://books.google.hu/books%3Fid%3DU5np34a5fmQC%26pg%3DPA4%26dq%3Dstatistics%2Bdata%2Bmining%2Bconcepts%2Band%2Btechniques%26hl%3Dhu%26sa%3DX%26redir_esc%3Dy&usg=ALkJrhjtJf3FChDzkWsvfLMpajcFOmp3uA" \l "v=onepage&q=statistics data mining concepts and techniques&f=false)

 The proposed subject of the dissertation is about the application of various data mining algorithm on medical/health/patient data. You may analyse the various algorithm, on some data set. You may find on the net some data sets that you may use.

1. ***One approaches could be***: Go to site Medline ([https://www.nlm.nih.gov/bsd/](https://www.nlm.nih.gov/bsd/pmresources.html) . Collect abstract, full text of articles that deal with e.g. colorectal cancer . Select a data mining tool with text mining capabilities (e.g. RapidMiner Community Edition) and install it. Exploiting text analytics technologies, perform on the text entity and relation extraction. Follow an established text analytics process. Summarize the results in project report. ([https://marketplace.](https://marketplace.rapidminer.com/UpdateServer/faces/index.xhtml) ).
2. ***Another example*** : Look up the site ([http://biogps.org/dataset/](http://biogps.org/dataset/tag/colorectal%20cancer/) ). Look at the data set. and then plan and design a data analytics programme based on the selected data mining process (either CRISP-DM or SEMMA). Choose and formulate the objective of the data mining process as evaluation of diagnostic methods, therapy, forecasting diseases based on results of timely diagnostic. Produce a project plan, then carry out to create a professional study about the experiments using data mining methods.
3. Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine, 34 (2), 113–127.
4. Metrics that could be used for analysing and comparing the disperse algoritms:
5. The common accuracy metrics for data mining algorithms



1. 
2. F-Measure

A measure that combines precision and recall is the [harmonic mean](https://en.wikipedia.org/wiki/Harmonic_mean) of precision and recall, the traditional F-measure or balanced F-score:

F1=(2⋅recall⋅precision)/(recall+precision)

1. <https://en.wikipedia.org/wiki/Precision_and_recall>

fallout

Fall­out The proportion of non­relevant documents that are retrieved, out of all non­relevant documents available: In binary classification, fall­out is closely related to specificity and is equal to

Some assistance to get acquainted with the topic:

Textbook és sample data are can be found:

Data Mining For The Masses With Data Sets : ​

[**Data Mining Data Science**](https://drive.google.com/drive/folders/1RGgtqzW6vu1BuRt6XFUEkQ2mmXx1PXOP)

​

 any open source data mining software can be used:

Rapid Miner: is recommended

<https://rapidminer.com/>

***In order to practice the basics, you may choose any data set from***

<https://www.kaggle.com/dataset>

or

<https://archive.ics.uci.edu/ml>

Make any arbitrary more complex analysis on the data to find interesting patterns, clusters, or predicate something using any tools, Rapid Miner or python package  [http://scikit-learn.or](http://scikit-learn.org/stable/)

