

Data mining



Kucherov Sergey
Shafeullakh Andrey
Kislinskikh Kirill

Data mining – this is a technique of revealing hidden relationships within large databases. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.



The purpose of data mining is to collect as much information as possible about any particular issue so that analysts can spot trends and predict what is likely to happen next.



Knowledge

Collected information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Data mining is useful!

Retail trade: analysis of shopping basket, creation of predictive models, exploration of temporal patterns.

Banking industry: fraud detection with credit cards, analysis of clientele.

Social insurance: fraud detection, risk analysis.



The Best Open Source Data Mining Tools

It is rightfully said that data is money in today's world. Along with the transition to an app-based world comes the exponential growth of data. However, most of the data is unstructured and hence it takes a process and method to extract useful information from the data and transform it into understandable and usable form. This is where data mining comes into picture. Plenty of tools are available for data mining tasks using artificial intelligence, machine learning and other techniques to extract data.

RapidMiner (formerly known as YALE)

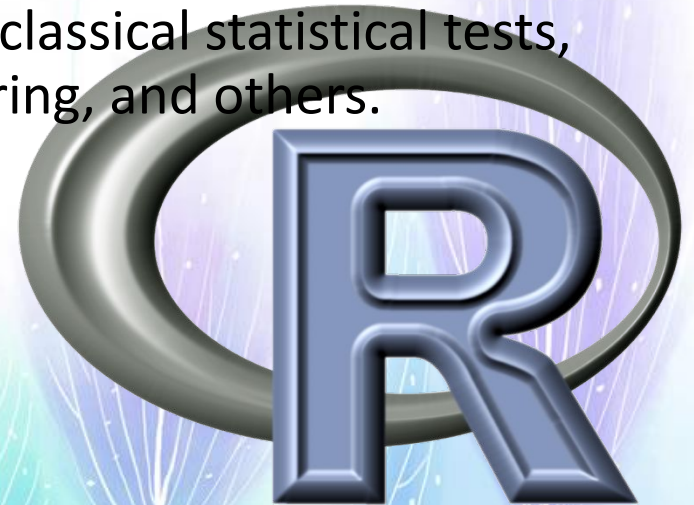
- Written in the Java Programming language, this tool offers advanced analytics through template-based frameworks. A bonus: Users hardly have to write any code. Offered as a service, rather than a piece of local software, this tool holds top position on the list of data mining tools.
- In addition to data mining, RapidMiner also provides functionality like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. What makes it even more powerful is that it provides learning schemes, models and algorithms from WEKA and R scripts.



rapidminer

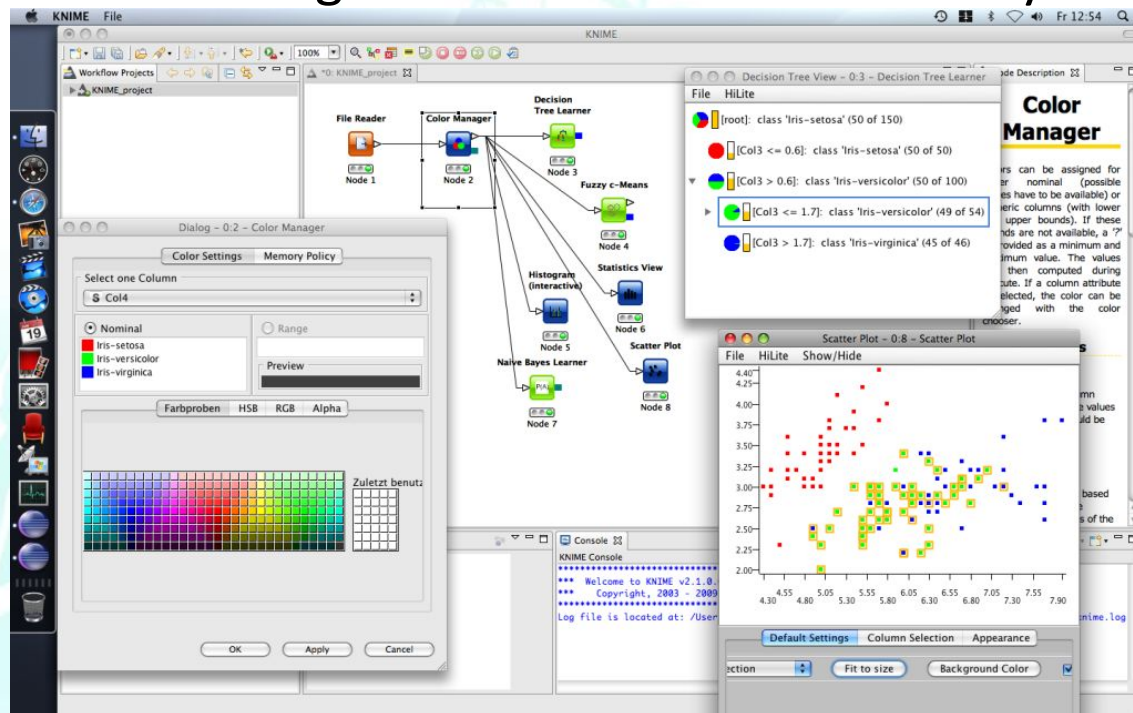
R-Programming

- What if I tell you that Project R, a GNU project, is written in R itself? It's primarily written in C and Fortran. And a lot of its modules are written in R itself. It's a free software programming language and software environment for statistical computing and graphics. The R language is widely used among data miners for developing statistical software and data analysis. Ease of use and extensibility has raised R's popularity substantially in recent years.
- Besides data mining it provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.



KNIME

Data preprocessing has three main components: extraction, transformation and loading. KNIME does all three. It gives you a graphical user interface to allow for the assembly of nodes for data processing. It is an open source data analytics, reporting and integration platform. KNIME also integrates various components for machine learning and data mining through its modular data pipelining concept and has caught the eye of business intelligence and financial data analysis.



Links

- http://www.megaputer.ru/data_mining.php
- <https://basegroup.ru/community/articles/data-mining>
- <http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/>



Thank you for your attention