



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

KDD Techniques and Open Source Tools: A Study

¹Dr. K.Venkataramana, ²Dr G.V.Ramesh Babu

¹.Assistant Professor, ²Assistant Professor

¹ Dept of Computer applications, KMM Inst. Of P.G Studies, Tirupati, ²Dept of Computer science, S.V.University, Tirupati

Abstract: Post Pandemic world is moving entirely on analysing and mining of data from various fields for returning to pre-Pandemic situation. Daily Terabytes or petabytes of data is generating online and it is to be processed with various techniques of data mining. So in this paper some of those techniques of data mining are given. Business recovering from Covid situation seeking IT solutions from Companies at low cost hence the companies adapt Open-Source ETL Tools as these Open-Source ETL Tools help businesses keep their production costs low but at same time will provide all functionalities as other ETL Tools in the market. At the same time these tools provides a simple and accurate UI (User Interface) front ends for users to set up the ETL process within a few minutes. So basics of data mining and important open source tools are discussed in this paper.

Keywords: KDD, Data cleaning, Data mining, Data Integration, open source

I. INTRODUCTION

Modern society is built on data generation and information usage to move forward for a better society. Daily Terabytes or petabytes of data are transferred into our computer networks, the World Wide Web (WWW), and various data storage devices every day from various activities like business, society, science and engineering, medicine, etc... Simply to say from every aspect of daily life. Computerization of our society along with development of powerful data collection and storage tools has led to exponential growth of data. Businesses across the globe generates enormous data sets, including sales transactions, stock trading records, product descriptions, sales promotions, company profiles and performance, and customer feedback. For example, Wal-Mart, twitter, instagram, Facebook generate high orders of petabytes of data in a continuous manner, along with other areas like remote sensing, process measuring, scientific experiments, system performance, engineering observations, and environment surveillance. The growth statistics of internet usage has increased leaps and bounds shows that 4.66 billion people are using the internet as of January 2021. That's close to 60% of the world population which is huge spike due to digitization by global economies has been very huge due to Covid-19. Major change is that most of the IT Sector are working without reporting to physical work locations. Modern Information and communication technology is capable of collecting and generating large amounts of data that need to be analyzed to become more useful or profitable [1]. Data mining concept leads to development of Potent and versatile software tools for which automatically recovers valuable information from the gigantic sets of data to organize knowledge by transformation.

II. KDD PROCESS

Data mining is part of KDD process involves in extracting the pattern, rules, regularity and constraints from the enormous amount of data which allows in analysing the data from the source to answer questions of authorized department of various organization to make decisions. Data mining uses not only analytical, subject oriented and structured approach to answer various questions but also mathematical approach such as statistics, probability, algebraic functions, mensuration, curve fitting, set theory and logical reasoning to analyse the data. These mathematical methodologies can be modified, filtered, compiled and combined to make the algorithms more efficient[2].

Data preprocessing/ preparation and mining data from data sets are the basic operations in data mining. Among the first four processes, such as data cleaning, data integration, data selection and data transformation, are considered as data preparation processes, where as the last three processes including data mining, pattern evaluation and knowledge representation are integrated into one process called data mining. The knowledge data discovery or data mining involves processing of large data from data warehouses or data centres is shown in Figure-1 which is an iterative sequence of the following steps[3]:

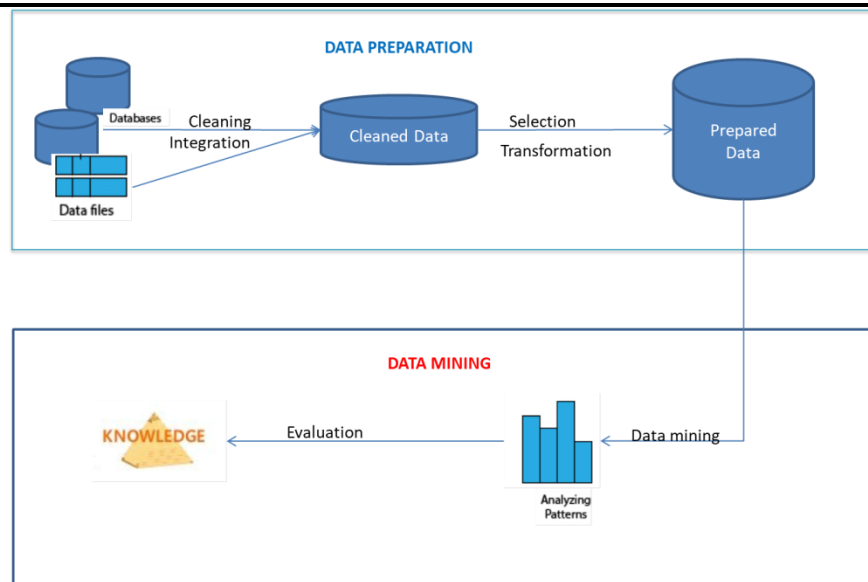


Figure.1. Steps in Knowledge Discovery Process

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined) Most of the companies perform data cleaning and data integration as a preprocessing step, and the resulting data are stored in a data warehouse.
3. Data selection (where data/ or attributes relevant to the analysis task are retrieved from the database)
4. Data transformation essential for mining is done by computing summary or aggregation operations) and sometimes in mining the steps of data transformation and consolidation are performed before the data selection process, particularly in the case of data warehousing. Some of the data attributes in can be reduced to obtain a smaller representation of the original data without losing integrity.
5. Data mining (an essential process where intelligent methods are applied to extract data patterns) essential for decision making.
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on measures important at specific situation.
7. Knowledge presentation where visualization and knowledge representation techniques are used to present mined knowledge to users in various views.

It is known that the main aim of data mining process is to extract information from large amount of data and to translate raw data into meaning full information. Data Mining helps to identify patterns among data. Data mining techniques are widely useful in different fields like games, business, medical diagnosis, science & engineering and many more [4].

Data mining can be conducted on database data, data warehouse data, transactional data, and advanced data types or any type of data as long as the data is meaningful for a target application. Complex or advanced data types for modern applications include spatial and spatiotemporal data, time-related or sequential data, data streams, text and multimedia related data, graph and networked data, and Web data for predictions in various ways. [5]. Data mining moreover popularly referred to as Knowledge Discovery in Databases (KDD), usually referred because the nontrivial extraction of implicit, antecedently unknown and doubtless helpful information from understanding in databases. Data mining is a step worried in Knowledge discovery manner [2].

III. DATA CLEANING

In the level of Data Cleaning, cleansing noisy statistics and irrelevant records are removed from the large set of data in order to increase efficiency and accuracy in outcome of analysis with results. Data that is incomplete or inaccurate is known as “dirty” data. In order to ensure high data quality, data warehouses must validate and cleanse incoming data records from external sources. The various types of anomalies classified under several types occurring in data had to be eliminated. Evaluation and comparison of existing approaches for data cleansing are done with respect to the types of anomalies handled and eliminated by them.

Border Detection algorithm proposed by Arturas Mazeika and Michael H.B ohlen in 2006 [6] which identifies a group of strings containing both occurrences of a correctly spelled string and adjacent misspelled strings. Grouping of strings is done based on the most frequent string of this group by searching in proper noun databases, including names and addresses, which are not handled by dictionaries. Center of group is computed to determine the border of the group and the experimental evaluation shows by taking proper nouns. The center calculation and border detection algorithms are robust with good results even with a very small sample sizes. It works in two steps, at first, the string data is clustered by identifying center and border of hyper-spherical clusters, secondly, the cluster strings at borders are cleansed with the most frequent string of the cluster. Strings within the overlap threshold from the center of the cluster are assigned to one cluster and new cluster is created.

Efficient cleansing technique based on fuzzy match algorithm is developed in 2003 by Surajit Chaudhuri, Kris Ganjam, Venkatesh Ganti Rajeev Motwani [7][8]. In many scenarios, clean tuples must match acceptable tuples in reference tables. In this proposed algorithm, main challenge of effectively cleaning an incoming tuple if it fails to match exactly with any tuple in the

reference relation is implemented by an accurate fuzzy match operation. A similarity function which overcomes limitations of commonly used similarity functions is given. To develop an accurate fuzzy match similarity function for matching erroneous input tuples with clean tuple, an Edit distance similarity is generalized by incorporating the notion of tokens from a reference relation. The error tolerant index and an efficient algorithm are developed for identifying with high probability with closest fuzzy matching reference tuples.

In paper by Lukasz Ciszak, 2008 IEEE[4] has proposed Data cleaning techniques such as Context-independent attribute correction implemented using clustering techniques and context-dependent attribute correction using associations based on Clustering and Association Methods. Solutions based on Attribute correction require reference data in order to provide satisfying results. In case of Context-independent attribute correction all the record attributes are to be examined and cleaned in isolation without regard to values of other attributes of a given record.

3.1 A Token Based Data-Cleaning Technique

In data cleaning, most of the techniques identifies record duplicates by computing match scores compared against a given match score value, long string comparisons are done for each record involve a number of passes and time consuming. In above techniques determining optimal match score values is hard resulting in straight long string comparisons with many passes which is inefficient. In paper [9] by Timothy E., Ohanekwu and C.I. Ezeife proposed a smart tokens a token based technique which eliminates the dependency on match value(threshold) for identifying duplicates. It also eliminates the need to use the long string records with multiple passes, for duplicate identification for cleansing. The results from the experiments show that the proposed token-based algorithm outperforms the other existing algorithms using token keys extracted from records by sorting and/or clustering.

3.2 Tools for Data Cleaning

Now-a-days various tools are available for cleaning data such as [12]

1. OpenRefine also known as Google Refine, is a popular open-source data tool free to use and customize.
2. Trifacta-Wrangler is a connected desktop application, allows to transform data, carry out analyses, and produce visualizations. It uses machine learning to find out inconsistencies and make recommendations in data cleaning process.
3. Winpure Clean & Match which is similar to above wrangler tool allows you to clean, de-dupe, and cross-match data, all via its in-built user interface. No issues in data security as it is offline application except in case of uploading your dataset to the cloud.
4. TIBCO Clarity is cloud based SaaS tools allows cleaning raw data from various data sources on cloud.
6. IBM Infosphere Quality Stage is data management tools from IBM on data quality and governance deals with the usual suspects (data matching, de-duping, etc.) it is specifically designed to clean big data for business intelligence purposes.

IV. DATA INTEGRATION

The data pre-processing technique Data Integration involves combining data from multiple heterogeneous data sources into a coherent data store and provides a uniform view of the data for further processing or mining of data. These sources may include multiple data cubes, databases, or flat files or web databases etc. The main challenge is semantic heterogeneity and structure of data to be integrated from various resources. One of the issue related to data integration is entity identification, which is difficult during processes of data integration. Matching of Schema integration and object matching equivalent real-world entities from multiple data sources is tricky mechanism. [5]

a) Redundancy and Correlation Analysis

Redundancy in attributes is an important issue in data integration, if it can be “derived” from another attribute or set of attributes in dataset or due to naming of attributes. This issue can be solved to certain extent by using correlation analysis. From the existing two attributes, we can find how strongly one attribute implies the other, based on the available data. [5]

b) Tuple Duplication

Tuple level duplication may also lead to redundancy in addition to redundancies between attributes (e.g., existence of two or more identical tuples for a given unique data entry case). The use of tables which are not normalized is another source of data redundancy.

c) Data Value Conflict Detection and Resolution

Values in attributes of dataset for same real world entity differ in representation or encoding scheme used, such as metrics, currencies etc.. So these attribute is to be Detected and should be resolve data value conflicts. For example weight or currency will be different at difference countries. When exchanging information between schools, for example, each school may have its own curriculum and grading scheme.

4.1 Open Source Tools Used In Data Integration

Data Integration process involved in moving the data from source to the destination. Various open source tools available are [13]

- 1) Airbyte is an open-source data integration solution with pre-built and custom connectors can be used on laptops and even servers to replicate data without coding via the vendor’s web application.
- 2) Apache Kafka is a distributed streaming platform that enables us for building real-time streaming data pipelines and applications and is run as a cluster on one or more servers that can span more than one datacenter.
- 3) Apache NiFi is another open source tool to process and distribute data, provides directed graphs of data routing, transformation, and system mediation logic.
- 4) Apatar is a free and open-source does not requires any programming or design to accomplish even complex integration with joins across several data sources. This tools is designed to help business users and developers move data in and out of a variety of data sources and formats.
- 5) CloverETL (now CloverDX) was java based first open source ETL tools. It’s data integration framework was built to transform, map, and manipulate data in various formats.

Other tools include HPCC Systems, Jaspersoft, SourceForgeKETL, Pentaho Kettle, Scriptella

V. DATA SELECTION/REDUCTION

Depending on application and to make analysis more efficiently reduction of data size in terms of attributes is necessary. So an Attribute Subset Selection is a technique which is used for data reduction in data mining process. Large data sets used for analysis possibly contain hundreds of attributes, many of which may be irrelevant to the mining task or may be redundant to process. So in order to reduce volume of data set a specific attributes are selected, also the integrity of the original data is maintained. So techniques in mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results. Various data reduction strategies such as dimensionality reduction, numerosity reduction, and data compression are used. As application demands it is required that number of attributes or random variable are to be reduced, which is known as Dimensionality reduction. Dimensionality reduction methods include wavelet transforms and principal components analysis, which transforms or allows us to view original data in different perspective in a smaller space. Another technique of dimensionality reduction, an Attribute subset selection removes irrelevant, weakly relevant or redundant attributes or dimensions from data set.

Numerosity reduction techniques modify the original data volume into less significant forms of data representation. These techniques may be parametric or nonparametric. In parametric methods such as Regression and log-linear models, we pass parameters to model to estimate, typically only that data parameters need to be stored, instead of the actual data. Nonparametric techniques such as histograms, clustering, sampling, and data cube aggregation are used for storing reduced representations of the data in simple and efficient way.

The reduced or compressed representation of the original data obtained by Data Compression techniques on data should produce a resultant data which may be lossless or lossy. Even though there are several lossless algorithms for string compression; but when used will typically allow only limited data manipulation. As we already discussed Dimensionality reduction and numerosity reduction techniques can also be considered forms of data compression. The methods like Stepwise forward selection, Stepwise backward elimination or the combination of both forward selection and backward elimination can be used so that, after some iterations we can get a reduced data set at containing the best attributes and the worst attributes are removed which gives reduced data set.[5]

5.1 Open Source Data Selection/Reduction Tools

- 1) PyPEIT is a library in Python for semi-automated reduction of astronomical, spectroscopic data. The techniques used in PyPEIT are based on well-established and long development of previous data reduction pipelines by the developers (Bernstein et al., 2015; Bochanski et al., 2009). The reduction procedure involves a complete list of the input parameters and available functionality[15].
- 2) KNIME Analytics Platform is the open source software for creating data science projects. It provides extraction and selection features (or construct new ones) to prepare dataset for machine learning with genetic algorithms, random search or backward- and forward feature elimination. By using KNIME tool one can easily understand data and designing of data science workflows and reusable components [16].
- 3) GGobi is another data visualization open source the key tool for successful data mining. It is particularly suited for data mining and explorative data analysis. It allows user's selection in all opened visualizations provides two-dimensional visualizations and in a movie-like fashion shifts between two different projections [17].

VI. DATA TRANSFORMATION

The data collected in a data set may not be useful enough for a Data Mining algorithm as some of the raw attributes selected from original data are not enough for modeling even though they are suitable for operational system. Hence it is required to do a series of manipulation steps to transform the original attributes in data which will help in predictive modeling. For example one-hot encoding transformation method can be applied to transform categorical variables into numerical ones to facilitate the development of prediction.

For KDD process to be success the methods of data transformation such as dimension reduction (such as feature selection and extraction, and record sampling), and attribute transformation (such as discretization of numerical attributes and functional transformation) are useful. In application related to medical examinations, the quotient of attributes may often be the most important factor, and not each one by itself. If right transformation is not applied from the beginning, results obtained gives a surprising results that hints to us about the transformation needed (in the next iteration). Thus the KDD process reflects upon itself and leads to an understanding of the transformation needed (like a concise knowledge of an expert in a certain field regarding key leading indicators). The data transformation comprises following steps such as:

1. Smoothing Dataset involves in applying algorithms is used to remove noise from the dataset
2. Aggregation: Data collection or aggregation is the method of storing and presenting data in a summary format.
3. Discretization: It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes.
4. Attribute Construction: Where new attributes are created & applied to assist the mining process from the given set of attributes. This simplifies the original data & makes the mining more efficient.
5. Generalization: It converts low-level data attributes to high-level data attributes using concept hierarchy.
6. Normalization: Data normalization involves converting all data variable into a given range.

6.1 Open Source Tools for Data Transformation

In improving business, IT companies should use a correct ETL (Extract, Transform and Load) tool as they help to combine and enrich data from numerous data sources, allowing you to carry out an insightful analysis & gain actionable insights for customers[14].

a)Hevo Data, is a simple No-code Data Pipeline helps to load data from any data source such as Databases, SaaS applications, Cloud Storage, SDK's, and Streaming Services and simplifies the ETL process.

b)Apache NiFi is a simple open-source java based ETL tool used to process and distribute data. Its high-level features for data transformation and user friendliness make the tool reliable.

c)Apache Camel is a tool developed as an integration framework to integrate different systems that consume or create data. It is optimized to work with the majority of enterprise integration patterns systems.

d)KETL is an XML-based open-source ETL and best interoperable tool. It is used in development and deployment of data from and to different platforms. KETL is fast and efficient and it helps you manage even the most complex data in minimal time.

e) PowerCenter from Informatica is an advanced open-source ETL tool for enterprise. It was developed for on-premises data integration initiatives such as app migration, data warehousing, and analytics. As it supports large enterprise level applications provides universal connectivity

Other tools such as Singer, Apatar, Scriptella, Tallend, Jaspersoft ,CloveDX, GeoKettle etc.,

VII. DATA MINING

Due to the sheer volume of information generated in organizations and institutions, it can be very difficult to grasp and analyse to strengthen the organization at different levels. So by using Data mining in businesses one can uncover data trends to obtain knowledge that is important to their business essentialities in terms of business intelligence and data science[12].

Various techniques for data mining are used by organizations to convert raw information into a usable insights. Application data mining encompasses everything from advanced artificial intelligence to fundamental data planning, which forms basis in optimizing the value of data investment. Data mining may be predictive or descriptive for extracting hidden information from huge data.

a)Classification: It is predictive data analysis process for finding a model that describes and differentiates data classes and concepts. By data extraction approach helps to classify data into different categories. Various techniques like classification by decision tree induction, Bayesian Classification, Back propagation, Support Vector Machines (SVM), Classification Based on Associations etc. are used.

b)Clustering: It is a descriptive process used in data extraction to classify related data. This method helps to consider gaps between data and similarities. Hierarchical (divisive) Methods, Partitioning Methods, Density Based Methods etc. are used in clustering.

c)Regression: It is predictive regression analysis for data extraction process in which the relationship between variables is defined and analyzed. Methods like Multivariate Linear and Non-Linear Regression, Nonlinear Regression are used in regressing process.

d)Association rules: It is a data extraction technique helps to find interrelated objects and patterns in the dataset. Association rules are derived from methods like Multilevel association rule multidimensional association rules.

e)Outlier detection: In this type of data extraction technique the observation of data elements that do not fit the predicted behavior pattern in the data collection.

f)Sequential patterns: In a specific time frame to find similar patterns or trends in transaction data this technique is used.

g)Prediction: The prediction uses various techniques for data mining, such as patterns, sequences, clustering, grouping, etc. It helps in analysing past events or circumstances in the right order to predict a future occurrence.

7.1 Open Source Data Mining Tools

a) DataMelt also known as DMelt is open-source software for numeric computation, mathematics, statistics, symbolic calculations, data analysis and data visualisation. DMelt supports combination of various scripting languages such as Python, Ruby, Groovy with several Java packages for usage and development.

b)ELKI Environment for Developing KDD-Applications Supported by Index-Structures. Java Programming language and libraries are used in developing ELKI is an open-source data mining tool. The main aim of this platform to research in algorithms, related to unsupervised methods in cluster analysis and outlier detection and emphasis is on better analysis of various mining techniques.

c) KNIME is an open source tool is developed in Java and based on Eclipse IDE, which is an Analytics Platform for carrying tasks in data science. It is a multi-language software supports enhancements through an extensible plug-in system. The main purpose of this tool is better served in data analytics, reporting and as a integration platform .

d) Orange is an open-source, reusable component-based data mining software used in machine learning and data visualisation. It supports wide range of data visualisation, exploration, preprocessing and modeling techniques. It can be used as a module for the Python programming language and can incorporate new developments.

d) WEKA or Waikato Environment for Knowledge Analysis is a simple open-source machine learning and mining software written in Java language. It is simple and can be accessed via a graphical user interface, standard terminal applications, or as a Java API. It comprises various machine learning algorithms for solving real-world data mining problems. It is designed as a machine independent runs on almost any platform.

Other tools include Scikit-Learn,Rattle, Apache Mahout,RapidMiner, IBM Cognos etc.,

VIII. CONCLUSION

In this paper we have studied important task of Data mining which is essential task for pattern finding, prediction, discovery of knowledge, etc., in different fields of society along with open source tools used in each process. Its applications are related in terms of range of data types, from text to photos, warehouses, and different databases and data structures. Various data mining techniques are employed for extracting patterns and knowledge from various databases. So much of work has to be done in data mining with respect to Artificial intelligence, Security, Cloud computing etc. which we can treat as open areas for further research work.

REFERENCES

- [1]. <https://techjury.net/blog/how-much-data-is-created-every-day/>
- [2]. Avinash Pandey et.al, Developing Efficient Data Mining Algorithms, Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017), IEEE Xplore, ISBN:978-1-5386-1959-9
- [3]. <https://www.wideskills.com/data-mining-tutorial/data-mining-processes>
- [4]. Pooja akulwar et.al, Survey on Different Data Mining Techniques for Prediction, Proceedings of the Second International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2018) IEEE Xplore; ISBN:978-1-5386-1442-6, page no.513-519,
- [5]. Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann is an imprint of Elsevier
- [6]. Surajit Chaudhuri, Kris Ganjam, Venkatesh Ganti, and Rajeev Motwani. 2003. Robust and efficient fuzzy match for online data cleaning. In Proceedings of the 2003 ACM SIGMOD international conference on Management of data (SIGMOD '03). Association for Computing Machinery, New York, NY, USA, 313–324. DOI: <https://doi.org/10.1145/872757.872796>.
- [7]. Rohit Ananthakrishna Surajit Chaudhuri Venkatesh Ganti, Research Eliminating Fuzzy Duplicates in Data Warehouses. Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002
- [8]. Lukasz Ciszak: Application of Clustering and Association Methods in Data Cleaning .Proceedings of the International Multiconference on ISBN 978-83-60810-14-9 Computer Science and Information Technology, pp. 97 – 103.
- [9]. Timothy E. Ohanekwu, C.I. Ezeife: A Token-Based Data Cleaning Technique for Data Warehouse Systems. o algorithms
- [10]. Fan, Cheng et.al, A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data ,Frontiers in Energy Research, Vol-9, ISSN=2296-598X ,2021 models.
- [11]. Mustafa Abdalrassual Jassim, Sarah N. Abdulwahid, Data Mining preparation: Process, Techniques and Major Issues in Data Analysis, IOP Conf. Series: Materials Science and Engineering 1090 (2021) 012053. IOP Publishing, doi:10.1088/1757-899X/1090/1/012053
- [12]. <https://careerfoundry.com/en/blog/data-analytics/best-data-cleaning-tools>.
- [13]. <https://solutionsreview.com/data-integration/top-free-and-open-source-etl-tools-for-data-integration>
- [14]. <https://rigorousthemes.com/blog/best-open-source-etl-tools/>
- [15]. Prochaska et al., (2020). PyeIt: The Python Spectroscopic Data Reduction Pipeline. Journal of Open Source Software, 5(56), 2308. <https://doi.org/10.21105/joss.02308>
- [16]. <https://www.knime.com/knime-analytics>
- [17]. https://www.researchgate.net/publication/5656246_Open-Source_Tools_for_Data_Mining
- [18]. <https://analyticsindiamag.com/8-best-open-source-tools-for-data-mining/>