The discovery process of knowledge in databases

Shahnawaz Alam*[1], Mahesh Kumar Gupta[2], P. S. Sai Shashank[3], Piyush Mishra[4]
[1,2,3,4]Department of Mechanical Engineering, SRMIST, Modinagar Ghaziabad, India.

Abstract

In this article, the stages of the KDD process are described and emphasis is placed on the data mining stage and on the most commonly used techniques, such as classification, association, grouping and sequential patrons. It also details one of the reference methodologies most used in the development of data mining projects in academic and industrial environments, such as CriSP-Dm, which is made up of six phases: problem analysis, data analysis, preparation of data, modelling, evaluation and exploitation.

Keywords: CriSP-Dm, databases, data mining, KDD process.

1. Introduction

The process of extracting knowledge from large volumes of data has been recognized by many researchers as a key research topic in database systems, and by many industrial companies as an important area and an opportunity to obtain greater revenues [1]. Authors such as Fayyad, Piatetsky-Shapiro and Smith (1996, p. 89) define it as "The non-trivial process of identifying valid, novel, potentially useful and fundamentally understandable patrons by the user based on the data".

Knowledge Discovery in Databases (KDD) is basically an automatic process in which discovery and analysis are combined. The process consists of extracting patrons in the form of rules or functions, from the data, so that the user can analyze them. This task generally involves preprocessing the data, mining data (data mining) and presenting results [2-5]. KDD can be applied in different domains, for example, to determine profiles of fraudulent customers (tax evasion), to discover implicit relationships existing between symptoms and illnesses, between technical characteristics and diagnosis of the condition of equipment and machines, to determine profiles of "academically successful" students in terms of their socioeconomic characteristics and to determine purchase patterns of customers in their market baskets.

1.1. Steps of the KDD process

The KDD process that is shown in figure 1 is interactive and iterative, involving numerous steps with the intervention of the user in making many decisions. It boils down to the following steps:

- Selection.
- Pre-processing/cleaning.
- Transformation/reduction.
- Data mining.
- Interpretation/evaluation.

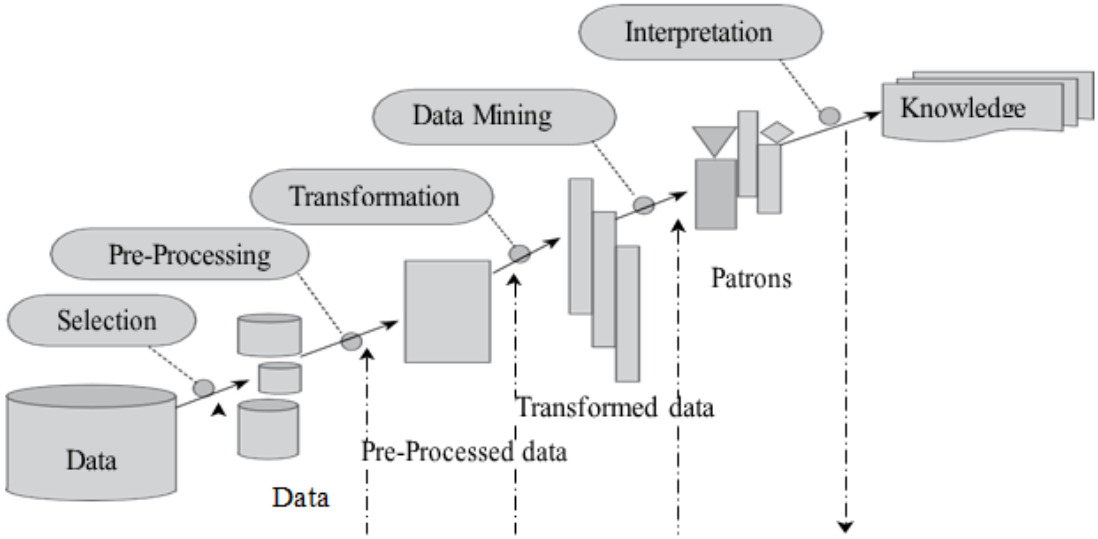

Figure 1: Steps of the KDD process.

1.2. Selection stage

In the selection stage, once the relevant and priority knowledge is identified and the goals of the KDD process are defined, from the point of view of the end user, an objective data set is created, selecting the entire data set or a representative sample of this one, on which the discovery process takes place. The selection of data varies according to business objectives.

1.3. Pre-processing/cleaning step

In the pre-processing/cleaning stage (data cleaning), the quality of the data is analysed, basic operations are applied such as the removal of noisy data, strategies are selected for the management of unknown data (missing and empty), null data, duplicate data and statistical techniques for its replacement. At this stage, the interaction with the user or analyst is of paramount importance.

Noisy data are values that are significantly outside the range of expected values; is mainly due to human errors, changes in the system, information not available at the time and sources heterogeneous data. The empty unknown data are those that do not correspond to a value in the real world and those missing are those that have a value that has not been captured. Null data are unknown data that are allowed by relational database management systems (sGBdR). In the cleaning process all these values are ignored, they are replaced by an omission value, or by the closest value, that is to say, statistics-type metrics such as average, mode, minimum and maximum are used to replace them.

1.4. Transformation/reduction step

In the data transformation/reduction stage, useful characteristics are sought to represent the data depending on the process goal. Dimension reduction or transformation methods are used to reduce the effective number of variables under consideration or to find invariant representations of the data [6].

Dimensions reduction methods can simplify a database table horizontally or vertically. The horizontal reduction implies the elimination of identical tuples as a product of the substitution of the value of an attribute by another of high level, in a defined hierarchy of categorical values or by the discretization of continuous values (for example, age by a range of ages). The vertical reduction implies the elimination of attributes that are insignificant or redundant with respect to the problem, such as the elimination of keys, the elimination of columns that are functionally dependent (for example, age and date of birth). Reduction techniques are used such as aggregations, data compression, histograms, segmentation, discretization based on entropy, sampling, among [5].

1.5. Data mining stage

The objective of the data mining stage is the search and discovery of unsuspected patrons and interests, applying discovery tasks such as classification [10-11], clustering [9-10], sequential patrons [11] and associations [1,12], among others.

Data mining techniques create models that are predictive or descriptive. The predictive models intend to estimate future or unknown values of variables of interest, which are called objective variables, dependent or class, using other variables called independent or predictive, such as for example predicting for new customers whether they are good or bad based on their marital status, age, gender and profession, or determining for New students desert or not depending on their area of origin, faculty, stratum, gender, age and average of grades. Among the predictive tasks are classification and regression. The descriptive models identify patrons who explain or summarize the data; serve to explore the properties of the examined data, not to predict new data, such as identifying groups of people with similar tastes or identifying patrons of customers in a certain area of the city. Among the descriptive tasks, association rules, sequential patrons, clustering and correlations are considered.

Therefore, the choice of a data mining algorithm includes the selection of the methods to be applied in the search for patrons in the data, as well as the decision on the models and the most appropriate parameters, depending on the type of data (categorical , numeric) unused.

1.6. Data interpretation/evaluation stage

In the interpretation/evaluation stage, the uncovered patrons are interpreted and possibly return to the previous stages for further iterations. This step may include the visualization of the extracted patrons, the removal of the redundant or irrelevant patrons and the translation of the useful patrons into terms that are understandable to the user. On the other hand, the discovered knowledge is consolidated to incorporate it into another system for further actions or, simply, to document and report it to the interested parties; also to verify and resolve potential conflicts with previously discovered knowledge.

Data mining tasks

Within data mining different types of tasks are found, which can be considered as a type of problem to be solved by a data mining algorithm [13]. Among the most important data mining tasks are classification, segmentation or clustering, association and sequential patrons.

1.7. Classification

The classification of data allows obtaining results from a supervised learning process. It is, moreover, the process by which common properties are found among a set of objects in a database and are cataloged in different classes, according to the classification model [14].

This process is carried out in two steps: the first one is built a model, in which each tuple of a set of tuples in the database has a known class (label), determined by one of the attributes of the database called attribute class. The set of tuples used to build the model is called the training set and is selected randomly from the total number of tuples in the database. Each tuple of this set is called a training example [5]. In the second step, the model is used to classify. Initially, the accuracy of the model is estimated using another set of tuples from the database, whose class is known, called the test set. This set is chosen randomly and is independent from the training set. Each tuple of this set is called a test example [5].

The accuracy of the model, on the test set, is the percentage of test examples that are correctly classified by the model. If the accuracy of the model is considered acceptable, it can be used to classify future data or tuples for which the class to which it belongs is not known. Several classification methods are proposed: rough sets, decision trees, neural networks, Bayes, genetic algorithms, among others.

The classification model based on decision trees is probably the most used and popular one due to its simplicity and ease of understanding [5, 15]. This model has its origin in machine learning studios. This is a supervised learning method that builds decision trees from a set of cases or examples called training set extracted from the database. A set of tests is also chosen, whose characteristics are known, with the aim of evaluating the tree.

The quality of the tree depends on the accuracy of the classification and the size of the tree [3]. The first method chooses a subset of the training set and forms a decision tree. If the tree does not give the correct answer for all the objects in the set, try a selection exceptions are added to the training set and the process continues until the set of correct decisions is found. The eventual result is a tree in which each sheet carries a class name and each interior node specifies an attribute with a branch corresponding to each possible attribute value.

Among the ranking algorithms for decision trees are Id-3 [7], Sprint (Shafer, Agrawal and Metha, 1996), sLIQ [16] and j48 [17]. The basic idea of these algorithms is to build decision trees in which:

- Each node in the terminal is labeled with an attribute.
- Each branch that leaves a node is tagged with a value for that attribute.
- Each terminal node is labeled with a set of cases, which satisfy all the attribute values that label the path from that node to the initial node.

The application of an attribute A as a selection criterion classifies the cases into different sets (as well as discrete values of the attribute). It's about building the simplest decision tree that is consistent with the training set T. For that, you have to order the relevant attributes, from the root to the terminal nodes, from highest to lowest ranking power. The classification power of an attribute A is its ability to generate partitions of the training set that adjust in a given degree to the different possible classes; in this way, an orden is introduced in this set. The order or the disorder (noise) of a measurable data set. The classification

power of an attribute is measured according to its ability to reduce uncertainty or entropy (degree of disorder in a system). This metric is called information gain. The attribute with the highest information gain is chosen as the attribute that forms a node in the tree [7, 14].

The decision tree was built in the following way:

- Calculate the entropy that can reduce each attribute.
- Rank the attributes from greatest to least entropy reduction capability.
- Build the decision tree by following the sorted list of attributes.

The gain of information obtained by the partitioning of the set T, according to the attribute A is defined as:

$$Gain(T, A) = I(T) - E(A) \quad (1)$$

Hence, I(T) is the entropy of the set T, composed of s emples and m different classes $C_i (i = 1, m)$ and is calculated:

$$I(T) = -\sum p_i \log_2(p_i) \quad (2)$$

Hence, p_i= s_i/s is the probability that an example which belongs to a class C_i and if is the number of examples of T from the class C_i.

E (A) is the entropy of the set T if it is partitioned by the n different values

del attribute A in n subsets, $\{S_1, S_2, \ldots S_n\}$, where S_j contains those examples of T that have the value of aj in A ands_{ij} the number of examples of the class C_i in the subset S_j.

E (A) is calculated:

$$E(A) = \sum s_{ij} / s^* I(S_{ij}) \quad (3)$$

Where, $s_{ij}$the number of examples of the class$C_i$in the subset S_j

$$I(S_{ij}) = -\sum p_{ij} \log_2(p_{ij}) \quad (4)$$

Where $p_{ij} = s_{ij}/s_j$ this is the probability that an example of S_j belongs to the classC_i.

In other words, Gain (T, A) is the expected reduction in entropy caused by the partitioning of T according to attribute A.

Finally, the classification rules are obtained by running each branch of the tree from the root to the terminal node. The antecedent of the rule is the conjunction of the pairs collected in each node and the consequent is the terminal node.

1.8 Segmentation or clustering

The process of grouping physical or abstract objects into classes of similar objects is called segmentation or clustering or unsupervised classification [3]. Basically, clustering groups a set of data (without a predefined class attribute) based on the principle of maximizing intraclass similarity and minimizing interclass similarity. Clustering analysis helps build meaningful partitions of a large set of objects based on

divide-and-conquer methodology, which breaks down a large-scale system into small components to simplify design and implementation.

The goal of clustering in a database is the partition of this into segments or clusters of similar records that share a number of properties and are considered homogeneous. Records in different clusters are different and these last ones have high internal homogeneity (within the cluster) and high external heterogeneity (between clusters). Homogeneity means that the records in a cluster are close to each other; there the proximity is expressed by means of a measure, depending on the distance of the records to the center of the segment. By heterogeneity it is understood that the records in different segments are not similar according to a measure of similarity [18].

The segmentation, typically, allows discovering homogeneous subpopulations: for example, it is applied to a customer database, to improve the accuracy of the profiles, identifying subgroups of customers who have a similar behavior when buying.

The clustering algorithm segments a database without any indication on the part of the user about the type of clusters that are going to be found in the database, and wants any sesgo or intuition on the part of the user; thus potentiates the true discovery of knowledge. For this reason, the method of segmentation or clustering is called supervised learning. Some of the algorithms used for clustering are: K-Means [5] Clarans (Clustering Large Applications based upon Randomized Search) [9], and Birch (Balanced Iterative Reducing and Clustering using Hierarchies) [10].

Clustering is used, for example, in the analysis of cash flow for a group of customers who pay in a particular period of the month, to perform market segmentation and to discover groups of affinities. It is also used to discover homogeneous subpopulations of consumers in marketing databases.

1.9 Association

The association task discovers patrons in the form of rules, which show the items that frequently occur together in a given data set. The problem was formulated by [14] and is sometimes referred to as the market-basket problem. This problem gives a set of items and a collection of transactions that are subsets (baskets) of these items. The task is to find relationships between the items of these baskets to discover association rules that fulfill the minimum specifications given by the user, expressed in the form of support and confidence. The quantity of items purchased in a transaction is not taken into account, which means that each item is a binary variable that represents whether an item is present or not in a transaction.

Formally, if I={i1, i2, ...,im} a set of literals, called items; is a set of transactions, where each transaction T is a set of items such that T⊆I. Each transaction is associated with an identifier called TID. Sea X a set of items. It is said that a transaction T contains X itself and only X ⊆T.

A rule of association is an implication of the form X ⇨Y, where X and Y are sets of items that X⊂ I, Y⊂ I and X ∩Y = ϕ

The intuitive meaning of such rule is that the transactions of the data base that contain X tend to contain Y. The rule X ⇨Y is fulfilled in the set of transactions D with a confidence c if the c% of the transactions in

D that contain X also contains Y. The rule X⇨Y has a support without the set of transactions D si el s% de las transacciones en D contains X ⇨Y.

The confidence denotes the strength of the implication and the support indicates the frequency of occurrence of patrons in the rule. The rules with high confidence and strong support are referred to as strong rules [14]. The problem of finding membership rules breaks down into the following steps:

- Discover frequent itemsets, i.e., the set of items that support transactions above a predetermined minimum support.
- Use frequent itemsets to generate association rules for the database.

Once the frequent itemsets are identified, the corresponding membership rules can be derived directly. An example of an association rule is "the 30% of the transactions that contain beer also contain pañales; 2% of all transactions contain both items" [2]. Here the 30% is the confidence of the rule and the 2%, the support of the rule.

According to [5], there are several criteria to classify the association rules, one of these is one of the dimensions that these encompass. According to this criterion, the association rules can be unidimensional and multidimensional. An association rule is one-dimensional, if the items or attributes of the rule make reference to a single predicate or dimension. For example, if you have the following membership rule:

Beer ^ fried porridge => breadsticks, which can be rewritten as:

Buy (beer) ^ buy (fries) => buy (pañales), make reference to a sole dimensión: buy.

An association rule is multidimensional, if the items or attributes of the rule make reference to two or more criteria or dimensions. For example, there is the following association rule:

Age (30...39) ^ occupation (ingeniero) =>purchase (laptop), contains three predicates: age, occupation and purchase.

A classic use of associations is the analysis of the market basket, in which the association is a list of product affinities. For example, looking at individual customer requests for workshop supplies can generate a rule: 70% of customers who order pens and pencils also order booklets.

Other applications of association rules are the analysis of medical demands to determine medical procedures that are performed at the same time or over a period of time, for a particular diagnosis. They are also applied for text analysis, catalog design, customer segmentation based on purchase patrons, in the market, among others.

1.10 Sequential patrons

The sequential patrons seek chronological occurrences. The problem of discovering sequential patrons is addressed in [2]. It is mainly applied in the analysis of the market basket and its objective is to discover in the customers certain buying behaviors in the time. The input data is a set of sequences called data-sequence. Each of the latter is a list of transactions, in which each transaction is a set of (literal) items. Typically, there is time associated with each transaction.

A sequential pattern is also composed of a list of sets of items. The problem is to find all the sequential patrons that comply with a minimum support specified by the user, in which support is the percentage of data-sequences that the patron contains. For example, in a database of a library, each sequence-date can correspond to all the book selections of a customer and each transaction, to the books selected by the customer in one order.

A sequential patron can be "The 5% of customers who buy 'Foundation', after 'Foundation and Empire' and later 'Second Foundation' [2]. The data-sequence corresponding to the customer, who bought other books jointly or after these books, keep this sequential pattern. The data-sequence can also have other books in the same transaction, as well as one of the books of the patron. Elements of a sequential pattern can be sets of items; for example, "'Foundation' and 'Ringworld', followed by 'Foundation and Empire' and 'Ringworld Engineers'". However, all items in an element of a sequential pattern must be present in a simple transaction for the data-sequence to support the pattern [2].

The sequential patrons, in the field of medicine, can be used, for example, to help identify symptoms and illnesses that precede other illnesses.

1.11 Areas related to the KDD process

KDD has been developed and continues to be developed based on the investigations carried out in the fields of machine learning, patron recognition, databases, statistics, artificial intelligence, expert systems, visualization of data and high-performance computing. The common goal is the extraction of knowledge from data in the context of large databases.

KDD relates to machine learning and pattern recognition in the studio of data mining theories and algorithms for data modeling and pattern extraction. Likewise, it focuses on the extension of these theories and algorithms on the problem of finding understandable patterns that can be interpreted as useful or interesting knowledge, and makes a strong emphasis on working with large sets of real-world data.

KDD has to do with statistics, particularly with exploratory data analysis. The inference of knowledge from the data has a fundamental statistical component [19]. The statistics provide a language and a structure to quantify the degree of certainty of the results when it comes to inferring general patterns on a particular sample of a whole population.

Data warehousing is another area with which KDD relates and refers to the current tendency of businesses to collect and clean transactional data in order to find them available for online analysis and decision support. A popular method for the analysis of data warehouses (data warehouse) is OLAp (On-line Analytical Processing) [20]. The tools of OLAp are focused on providing multidimensional data analysis and are intended to simplify and support interactive data analysis, while the objective of the tools of dCBd is to automate the process as much as possible.

2. CRISP-DM methodology

2.1 Generalities

In 1993, industry leaders such as Daimler Benz, spss from England, OhRA from Holland, NCR from Denmark, consortium of European companies, and AG from Germany built the acronym CRIsp-dM (Cross-Industry Standard Process for Data Mining), whose purpose is to provide new ideas to those who decide to work with data mining. This methodology has the advantage that it has not been built theoretically or academically, but is based on real experiences of how people carry out projects (Moro, Laureano and Cortez, 2011) [21-22].

This model is one of the most used as a reference guide in the development of data mining projects. The CRIsp-dM methodology consists of a set of tasks that are organized into four levels of abstraction: phases, general tasks, specialized tasks and process instances (see figure 2). Two levels are established respecting hierarchy in tasks; starting at the most general level until finally getting to the most specific cases [23].

2.2 Lifecycle of the crisp-dm methodology

The model provides a complete representation of the life cycle of a data mining project. The process is dynamic and iterative, so the execution of the processes is not strict and with frequency it is possible to pass from one process to another, from behind to front and vice versa. These depend on the result of each phase or the planning of the next task to be carried out.

These phases help organizations to understand the process and provide a "map of the path" that must be followed, thus: knowledge of the business, knowledge of the data, preparation of the data, modelling, evaluation, deployment (see figure 3).

Each phase is structured in several general tasks, the general tasks are projected into specific tasks, in which the actions that must be developed for defined situations are finally described [24].

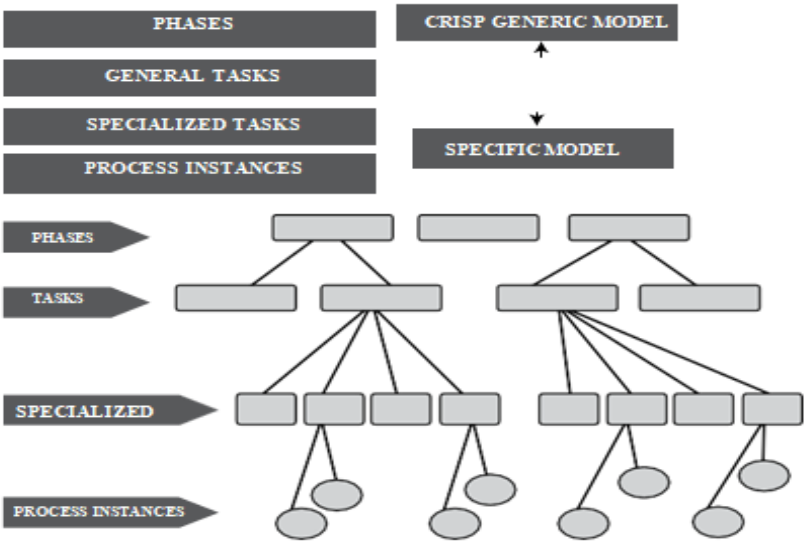

Figure 2: Scheme of the four levels of CRISP-DM.

2.3 Phases of the methodology

2.3.1 Phase 1. Understanding the business or Problem Understand the requirements and objectives of the project from a business or institutional perspective to convert them into technical objectives and a project plan, for which it is necessary to fully understand the problem to be solved (see figure 4).

- **Determine the objectives**: It determines which problem is to be solved and why data mining is used for that purpose; The success criteria must also be fixed. As for the latter, they can be qualitative or quantitative; for example, if the problem is detecting fraud in the use of credit cards, the criterion for quantitative success would be the number of fraud detections.
- **Evaluate the current situation**: In this task, the antecedents and requirements of the problem are evaluated, both in terms of the business and in terms of data mining. Some of the aspects to keep in mind could be the previous knowledge about the subject, the amount of data required to solve the problem, advantages of applying data mining to the problem, among others.

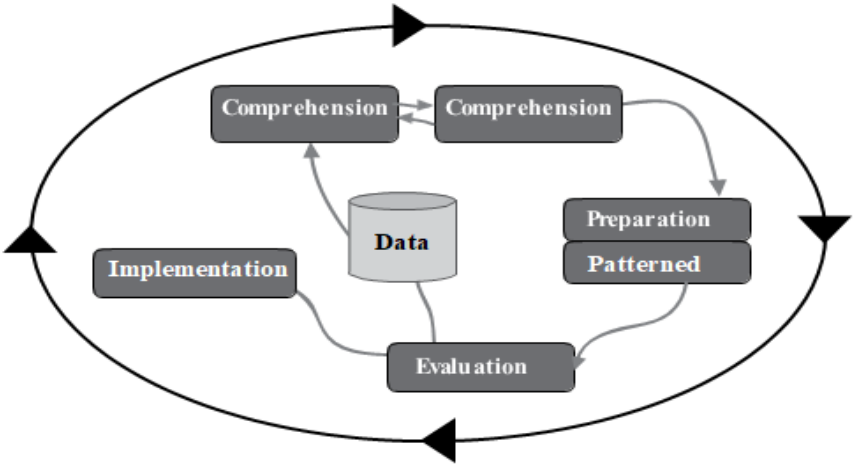

Figure 3: CRISP-DM lifecycle.

- **Determine the objectives of data mining**: The purpose of this task is to represent the business objectives in terms of the goals of the data mining project. For example, if the objective of the business is the development of an advertising campaign to increase the assignment of mortgage loans, the goal of data mining would be to determine the profile of customers with regard to their capacity for lending.
- **Produce a project plan**: The last task of this phase has the objective of developing the project plan considering the steps that must be followed and the methods to be used in each step.

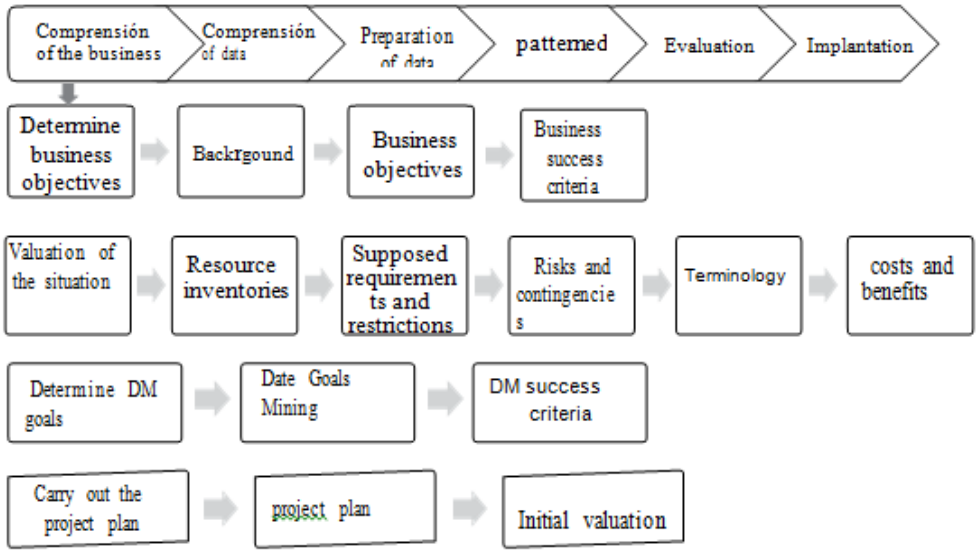

Figure 4: Business understanding phase

2.3.2 Phase 2.Understanding the data: Corresponds to the initial collection of data to establish a first contact with the problem; this phase, together with phase 3 and phase 4, demand more effort and time (see figure 5).

The main tasks that must be carried out in the data understanding phase are: collecting initial data, describing the data, exploring the data and verifying the quality of the data.

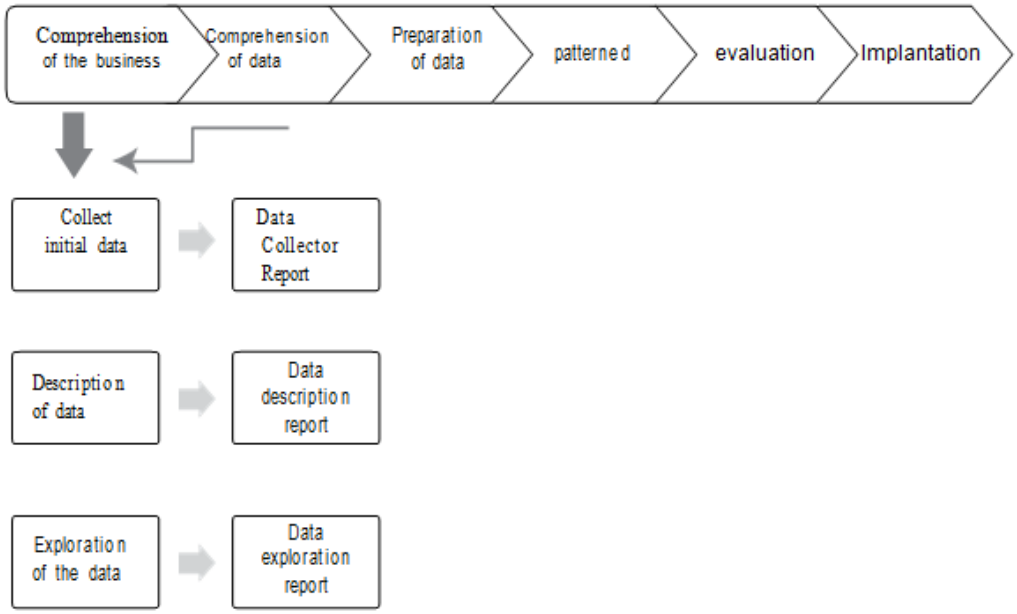

Figure 5: Data comprehension phase

- **Collect initial data**: Its main objective is the collection of initial data and its suitability for subsequent processing. Reports should be drawn up with a list of the data acquired, their location, the techniques used in their collection and the problems and solutions inherent to this process.

- **Describe the data**: The initial data obtained must be described, such as the number of records and fields per record, their identification, the meaning of each field and the description of the initial format.
- **Explore the data**: Its purpose is to discover a general structure for the data. It involves the application of basic statistical tests, which reveal properties in the data, frequency tables are created and distribution graphs are constructed. Create a data exploration report.
- **Verify the quality of the data**: Verification of the data is carried out to determine the consistency of the values of the fields, the quantity and distribution of the null values, to find values out of the range that can be noisy for the process. The objective is to ensure the completeness and correction of the data.

2.3.3 Phase 3. Data preparation: It is used to adapt them to the data mining technique, through the visualization of the data and the search for relationships between the variables. This phase is for modelling, as the data requires processing in different ways; therefore, the preparation and modeling phases interact permanently (see figure 6).

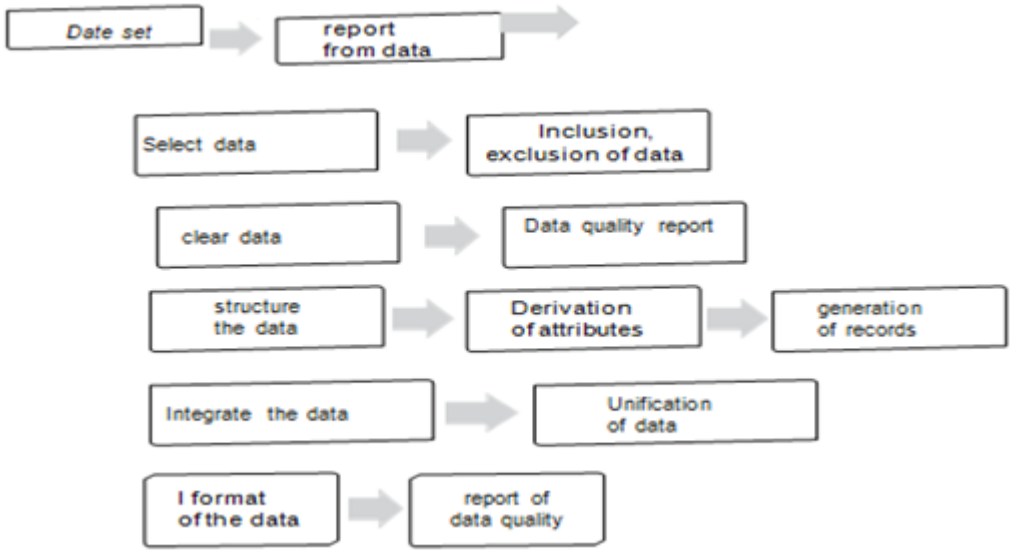

Figure 6: Data preparation phase.

The steps that are considered for the preparation of the data are: selecting, cleaning, structuring, integrating and formatting the data.

- **Select data**: A subset of data is selected considering the quality of the data, the limitation in the volume or in the types of data that are related to the mining techniques of selected data.
- **Clear data**: There is a diversity of techniques applicable to this task in order to optimize the quality of the data with a view to preparing them for the modeling phase. Some of the techniques are: normalization of data, discretization of numeric fields, treatment with empty values, reduction of data volume.

- **Structure the data**: Some of the operations to be performed in this task are the generation of new attributes from already existing attributes, integration of new records or transformation of values for existing attributes.
- **Integrate the data**: It involves the creation of new structures; for example, create new fields, new records, fusion of tables or new tables.
- **Format the data**: It mainly consists of syntactically transforming the data without modifying its meaning in order to facilitate, in particular, the use of some data mining technique; for example, remove comas, tabs, special characters, spaces, maximums and minimums for character strings, etc.

2.3.4 Phase 4: Modelled: Corresponds to the selection of a suitable and specific model; for that, techniques are used that meet the following criteria (see figure 7):

- Be appropriate for the problem.
- Availability of adequate data.
- Comply with the requirements of the problem.
- Appropriate technique to obtain a model.
- Full knowledge of the technique.

For example, if the problem is one of classification, we can choose between decision trees, k-nearest neighbor or chaos-based reasoning (CBR).

- **Generate test plan**: A plan must be generated to test the quality and validity of the built model; for example, in a task such as classification it is possible to use the error rate as a measure of quality. So, typically the data is separated into two sets, one for training and the other for testing.
- **Build the model**: The selected technique is performed on the prepared data to generate one or more models. All the modeling techniques have a set of parameters that determine the characteristics of the model for generating. The task of selecting the best parameters is iterative, based on the generated results. These must be interpreted and your income justified.
- **Evaluate the model**: The agreement models must be interpreted with the knowledge of the domain and the pre-established success criteria.

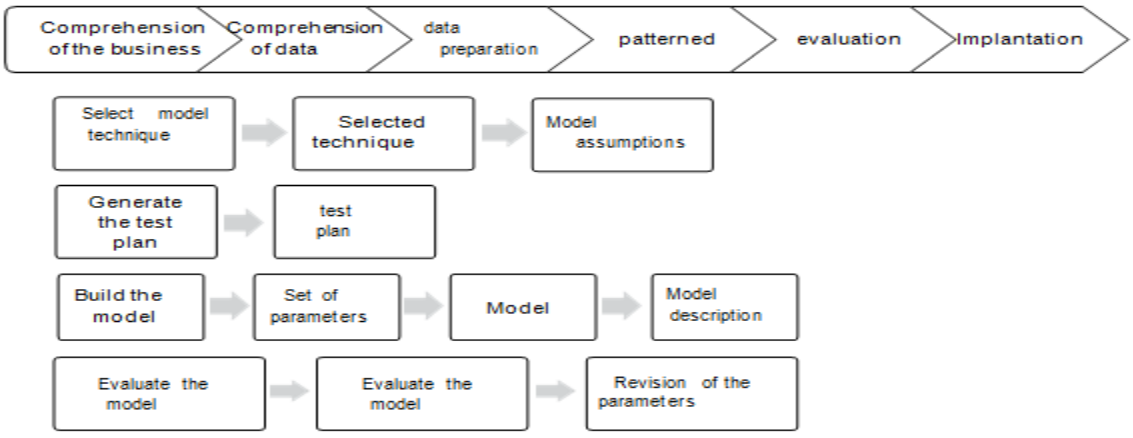

Figure 7: Modeling phase

2.3.5 Phase 5. Evaluation: Evaluate the model by considering the fulfillment of the problem's success criteria; For it, multiple tools are used to interpret the results, including the Edelstein 1999 confusion matrices, which is a table that indicates how many classifications have been made for each type. The diagonal of the table represents the correct classifications (figure 8).

If the above is valid, the model is exploited, which is the maintenance of the application and the possible dissemination of the results.

Once the model has been built and validated, the knowledge obtained is transformed into actions within the business process; The feedback generated by monitoring and maintenance can indicate whether the model is being used properly.

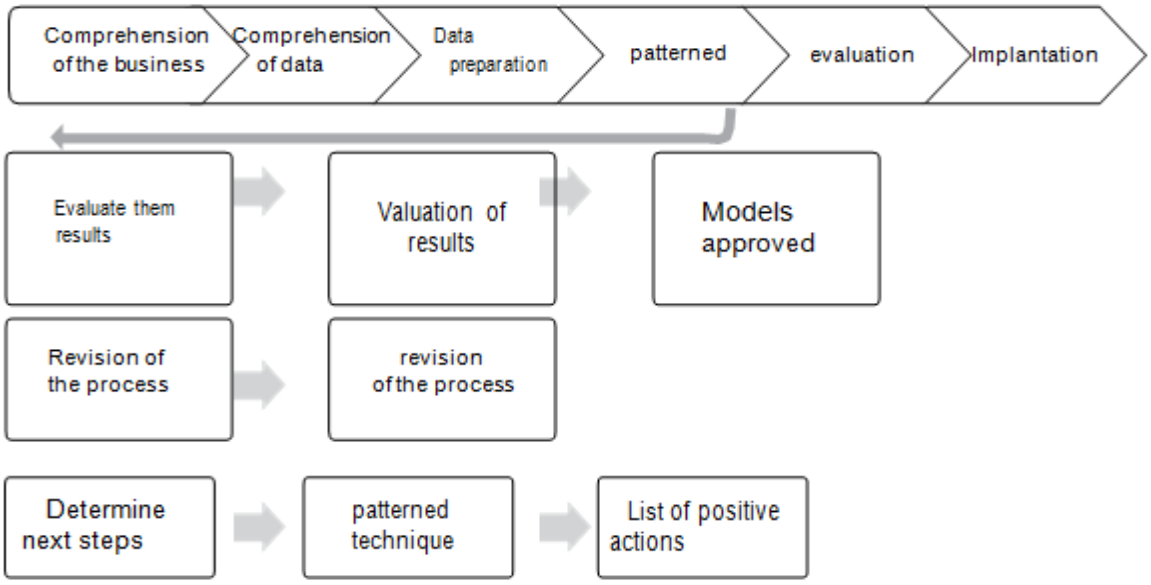

Figure 8: Evaluation phase.

2.3.6 Phase 6. Implementation: This is where the knowledge gained is transformed into actions within the business process, whether by observing the model and results, or applying it to multiple groups of data, or as part of the process. The tasks that are carried out are: planning the implementation, monitoring and maintenance, final report and review of the project.

- **Plan the implementation**: This task takes the results of the evaluation and concludes a strategy for its implementation. If a general procedure has been identified to create the model, it must be documented for its subsequent implementation (see figure 9).
- **Monitor and maintain**: Monitoring and maintenance strategies must be prepared to be applied to the models.
- **Final report**: Depending on the implementation plan, this can be a summary of the important points of the project and the experience achieved, or it can be a final presentation that includes and explains the results achieved with the project.
- **Revise the project**: If you evaluate the correct and the incorrect

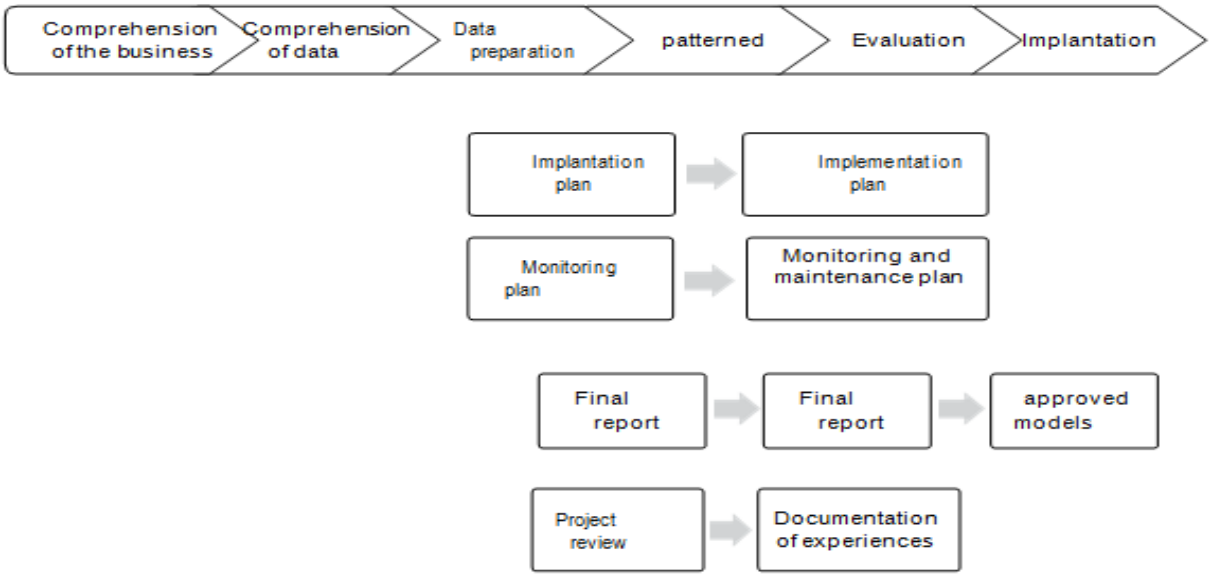

Figure 9: Implementation phase.

3. Conclusion

In this article, the fundamental concepts of the process of discovering knowledge in data bases are described, emphasizing the data mining stage where the tasks and techniques of mining the most important data are specified. The CRIsp-dM methodology that served as the basis for this investigation was also described.

References

1. Timarán, R. (2009). A look at the discovery of knowledge in databases.
2. Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules. vLdb Conference, Santiago de Chile.
3. Chen, M., Han, J., and Yu, P. (1996). Data Mining: An Overview from Database Perspective.ieee Transactions on Knowledge and Data Engineering.
4. Piatetsky-Shapiro, G., Brachman, R., and Khabaza, T. (1996). An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications. Association for the Advancement of Artificial Intelligence [AAAI], MIT Press. Retrieved from http://www.aaai. org/Papers/KDD/1996/KDD96-015.pdf
5. Han, J. and Kamber, M. (2001). Data Mining Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.
6. Fayyad, U., Piatestky-Shapiro, G., and Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the aCm, 39(11), 27-34.
7. Quinlan, J. (1986). Induction of Decision Trees. Machine Learning Journal, 1(1), 81-106.

8. Wang, M., Iyer, B., and Scott, J. (1998). Scalable Mining for Classification Rules in Relational Databases. International Database Engineering and Application Symposium - Ideas. Cardiff, Wales.
9. Ng, R. and Han, J. (1994). Efficient and Effective Clustering Method for Spatial Data Mining.vLdb Conference. Santiago de Chile, Chile.
10. Zhang, T., Ramakrishnan, R., and Livny, M. (1996). BIRCh: An Efficient Data Clustering Method for Very Large Databases. aCm siGmOd International Conference on Management of Data. Montreal, Canada.
11. Agrawal, R. and Srikant, R. (1995). Mining Sequential Patterns. The 11th International Conference on Data Engineering iCde, Taipei, Republic of China.
12. Srikant R. and Agrawal, R. (1996). Mining quantitative association rules in large relational tables, aCm siGmOd, Montreal.
13. Hernández, J., Ramírez, M. and Ferri, C. (2005). Introduction to Data Mining. Madrid: Editorial Pearson Educación sA.
14. Agrawal, R., Ghosh S., Imielinski, T., Iyer, B., and Swami, A. (1992). An Interval Classifier for Database Mining Applications. Proceedings VDLB Conference, Vancouver.
15. Sattler, K. and Dunemann, O. (2001). sql database primitives for decision tree classifiers. En Proceedings of the tenth international conference on Information and knowledge management (pp. 379-386). Atlanta: CIKM. Retrieved from http://dl.acm.org/citation.cfm?id=502650 Shafer J., Agrawal R., Metha M. (1996). spRINT: A Scalable Parallel Classifier for Data Mining.Proceedings of the vLdb Conference. Bombay, India
16. Metha M., Agrawal R., Rissanen J. (1996). SLIQ: A Fast Scalable Classifier for Data Mining. Proceedings EDBT96. Avignon, France.
17. Hall, M., Frank, E., and Witten, I. (2011). Practical Data Mining: Tutorials. University of Waikato. Available at: www.micai.org/2012/tutorials/Weka%20tutorials%20Spanish.pdf
18. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi A. (1998). Discovering Data Mining from Concept to Implementation, Prentice Hall. Retrieved from http://dl.acm.org/citation. cfm?id=270298
19. Elder, J. and Pregibon, D. (1996). A Statistical Perspective on Knowledge Discovery in Databases. En Advances in Knowledge Discovery and Data Mining, aaai Pres/ The MIT Press.
20. Gill, H. and Rao, P. (1996). Data warehousing: information integration for better decision-making. Prentice-Hall.
21. Martínez, D, and Podestá, C. (2014). Academic performance study methodology through data mining. Campus Virtuales, 3(1), 56-73.
22. Raus, N., Vegega, C., Pytel, P. and Pollo-Cattaneo, M. (2014). Proposed methodology for predicting university dropout through information exploitation (pp. 1014-1158). WICC 2014 Xvi Workshop of researchers in computer sciences. Retrieved

23. Chapman, P., Clinton, J., Randy, K., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRIsp-dM 1.0 Step-by-Step Data Mining Guide. Retrieved from http://www.crisp-dm.org/CRISPWP-0800.pdf
24. Larose, D. and Larose, Ch. (2014). Discovering Knowledge in Data: An Introduction to Data Mining (2nd ed.). New Jersey: John Wiley & Sons.