International Journal of Combinatorial Optimization Problems and Informatics, 13(2), May-Aug 2022, 14–32. ISSN: 2007-1558

www.editada.org

A Pool of Free Software Tools to Assist Business Intelligence and Analytics

*Juan Cruz Pantano [1,2], María Romagnano[1,2]**

[1] Institute of Informatics, Faculty of Exact Physical and Natural Sciences, National University of San Juan

[2] Computer Science Department, Faculty of Exact Physical and Natural Sciences, National University of San Juan

*Correspondence: maritaroma@gmail.com

Abstract. At present, enterprises face new economic models and dedicate a lot of time and resources to obtain, process, apply, and project information. If they do not collect the appropriate data, the information generated will not be accurate, the results will likely be wrong, and any decision made will not be the most appropriate. Business Intelligence and Business Analytics, used properly, can present competitive advantages, allowing organizations to know their current status and forecast future market behaviour, carrying out proactive actions based on predictive and prescriptive analysis. In this work, it is proposed to assist small and medium enterprises by integrating BI and BA into their information systems. The case of a local transport small and medium enterprise is presented where the benefits of applying free software tools, such as PowerBI Desktop, Orange, KNime, and Knowage, were analysed and evidenced.

Keywords: Data Science, Integration, Systems, Small and Medium-Sized Enterprises, Transport, Visualization.

Article Info

Received: August 8, 2021
Accepted: October 10, 2021

1 Introduction

Small and medium-sized enterprises have to deal with large amounts of data that are often generated in their daily operations. These transactional data are usually stored in spreadsheets, small databases, or even in text format. These data must be processed and transformed into information, which will be used to make decisions about strategies to follow and investments to make, among other possible actions. If the correct or most relevant data is not collected, the information produced will not be precise, results will most likely be wrong and, as a consequence, any decision taken will not be the best or the most adequate to the situation at hand. This means that a problem is presented when the necessary information to make the best decision in the best moment is not present. However, in large corporations or the so-called enterprises, this problem is usually already contemplated and resolved, given that there already exist specific areas that specialize in handling data and there is a strong investment in information science and data analytics.

Small and medium enterprises (SMEs) managers could count on proper tools for data exploitation and analysis, that allows them to obtain the needed and adequate knowledge that supports strategic decision-making processes. The investment could be focused on human training or hiring because all these proposed tools are free. This would notably reduce software costs, allowing simultaneously, to endow the enterprise with ideal personnel.

Leslie Bell-Friedel mentions: “the technology and the data are there; the problem lies in that organizations don’t know how to use them in the best possible way or they ignore the potential benefit in the application of these concepts” [1].

The objective of this work is to present, through a local SME study case, how free software visualization and data mining tools support both business analytics and intelligence when applied to information systems from small or medium-sized enterprises.

As a hypothesis, it is proposed that the application of these concepts and the use of this kind of tools might be of high benefit to information systems, whichever the size of the enterprise or the data handled may be.

2 Background

The global economy is suffering an unprecedented crisis due to the COVID pandemic, according to Damián Di Pace, economic analyst for the journal Diario Ámbito Financiero in his article "Knowledge Economy: the great winner in pandemic and key to the post-pandemic", the informatics and software sector will be the most needed now and in the next future. Citing the article "Knowledge is the future": if we revise the ongoing capacity of our entrepreneurs, the knowledge is the present. Primary economy, such as agriculture and cattle raising, and industrial economy, where jobs are usually under-qualified in our country, will need more and more knowledge economy, such as informatics, software, robotics, biotechnology, among others. In this economy, specialized workforce handles data, develops algorithms and simulated models and innovates in processes and systems that enhance productivity and competitivity of primary and secondary Argentinian economy sectors" [2].

It is complicated to trace a path post-pandemic, but there is no doubt that enterprises will have to invest in different informatics, communication alternatives and, of course, learn to optimize their businesses processes. In this current global context, it has been shown the importance of intelligence software, its growing adaptation as an analysis tool and an investment opportunity to solutions development organizations. Even for SMEs, this topic is relevant because, like all enterprises, they count on data to be analysed to make better business decisions [3].

The actual scenario is formed by globalization, high competence, dizzying changes, trying to keep customer loyalty, take maximum advantage of time, variability, and the amount of information available, among others [4]. Currently, SMEs plays a very wide economic and social role, and, because of this, have become an economic development source. The necessity to improve SMEs competitiveness to a global level is crucial. This type of enterprise is typically vulnerable and not robust enough to confront the worldwide and economic competitive advances. To survive, they should be able to monitor their businesses and utilize all their resources in the most efficient way, especially information [5],[6]. Managing an SME successfully is not an easy task. Besides maximizing incomes and operational efficiency, SMEs face strong competence, and their survival depends a lot on decisions taken [7].

To keep themselves currents in this scenario, organizations count on a large variety of both commercial and free software products, that allow them to do Business Intelligence (BI), Business Analytics (BA), visualization, data analytics and/or data mining tasks. SMEs represent the sector that can appreciate in the most tangible way the benefits that BI can generate [8]. Cloud Computing advances in the last few years are accelerating IT adoption in SMEs, including the opening to a possibility to implement BI [9].

Inside the field of Data Science, Business Intelligence and Business Analytics are two trends that are currently considered as very beneficial to an organization. This is because used properly, they can present competitive advantages to the organization, allowing it to know, with high precision, their current status (BI) and, based on that information, be able to forecast future market behaviour and take proactive actions based on predictive and prescriptive analysis (BA). These days, these concepts are gaining strong popularity and recognition, even though they are not new or from the recent apparition, particularly BI.

BI is the set of techniques, methods, strategies and tools that allow the use of data and the information produced and, from this, determine the current organization status with respect to its customers, competitors, sellers and the market itself, and be able to make a decision. The possibility to turn data and information into knowledge to finally carry out the decision-making process in an informed and accurate way can be considered a great competitive advantage, especially against other organizations that don't apply BI and, because of this, make decisions based only on personal opinions or ideas. López Benítez states that BI references the optimized handling of an organization's stored, collected, and analysed data, being able to turn them into strategic decisions that allow the design of actions oriented to achieve enterprise success [10]. The elements on which the business intelligence conceptualization is based are information systems, innovation mechanisms and decision-making processes. In each one, strategies that can help the organization to acquire knowledge and improve the way to increment products and services value are implemented [11],[12]. The implementation process of a BI system in an organization starts by selecting relevant information for decision-making, and it requires to be able to count on operative, tactic and strategic personnel participation. Once relevant information is identified, implementation continues with the consolidation process, where the ETL (Extract, Transform and Load) process is done, and it consists in the collection of data from different sources with the objective of normalizing, debugging, structuring and then storing this data. In the next stage, exploitation, the existent tools begin to be applied to leave the database's data ready in the hands of the users, that start taking advantage of them and to use the information already debugged and filtered that's on the data warehouse.

In this work, there are only briefly mentioned and described the ones that will be used in experimentation:

- Orange Data Mining is a tool that allows visualization and analysis, machine learning and data mining. This open-source software was developed by the Ljubljana University Informatics and Information Science Faculty's Bioinformatics Lab [13]. On one hand, it offers an attractive data visualization system to work with, and on the other, it reaches this visualization fast and easy, making it accessible to both beginners and experts.
- KNIME (Konstanz Information Miner) is a visualization, analysis, and data mining tool. Written in Java and prepared with Eclipse, it is considered as a highly popular tool among the international programmer's community and, compared with other data mining programs, it stands out thanks to a lot of functions: with over 1000 modules and app packages prepared, this tool allows to discover hidden data structures or integrated data analysis. In this ambit, KNIME is one of the most advanced programs because it allows the integration of numerous amounts of machine learning and data mining procedures. It also presents a notable efficiency in previous data treatment, as in their ETL process [14]. KNIME is used in the pharmaceutical investigation, the financial sector, and most notably in BI.
- PowerBI Desktop is an enterprise analysis service from Microsoft. Its objective is to provide interactive visualization and enterprise intelligence capacity with a sufficiently simple interface, so the final users create their own reports and dashboards. It provides BI services based on the cloud, called "PowerBI Services", along with a desktop interface called "PowerBI Desktop". It offers data storing capacities, including data preparation, discovery, and interactive panels. It can be concluded that PowerBI is a set of tools that unifies, sorts and analyses business information and it presents it as dashboards and reports that are easy to create [15].
- Knowage is a software tool that allows data visualization through one or more tabs called cockpits. In these cockpits, each graph, table or another data visualization method, is presented as a widget. It is easy to use and accepts a wide variety of databases and data types, making it a very versatile and convenient tool for small and medium-sized organizations, particularly SMEs. Not only allow BI through this visualization, but it also has widgets that allow the use of embedded languages (such as R or Python) that can be used to make predictions and BA. Once these codes are written and loaded, Knowage allows their visualization and customization so they can be appreciated as their best for the enterprise. Knowage has two different versions, one that is free (Community Edition) and another that needs a pay subscription (Enterprise Edition). However, for the objectives of an SME, the difference between these two is minimal, and it doesn't make it necessary to use Enterprise Edition, Community should be enough. Loading data sources, relevant data selection, use of cockpits and widgets, and other processes, are detailed in Knowage documentation and manuals [16].

Most SMEs consider these types of informatics solutions are only destined for large-sized enterprises. It is true that these last kinds of enterprises have a lot to win when implementing these tools, while they also count on the budget needed to do so. But this doesn't mean that SMEs can't benefit in an equal manner with BI when at the same time don't have to spend money and time on excess to be able to do so. SMEs count with much more focused objectives and implementation costs are usually much lower. Without a technological BI infrastructure, SMEs tend to exceed budgets, deadlines, improve performance in an area at the expense of the entire business, and reward employees on actions that do not necessarily mean an improvement in enterprise performance [17].

Despite all the advantages that BI presents, there still are some organizations that haven't implemented it. Some reasons for this could be:

- They are afraid to leave their comfort zone.
- There is prejudice against new technologies.
- It requires additional economic resources for its implementation.
- There is ignorance about the available techniques and technologies to carry a BI enterprise project.
- They think it is only aimed at large-sized enterprises.
- They think they don't need it.

On other occasions, and fundamentally when the organization does not present a digital culture when BI implementation starts, chaos is produced. This could happen because enterprises present:

- A lot of crossed information.
- A huge amount of complex data.
- Sources and data duplicity that slows processes.
- Large volumes of "worthless" information.
- A lot of personnel are involved in Big Data handling.

In these cases, it is necessary to manage master data to leave bewilderment and disorganization. These factors are also reflected in Clarysabel Tovar's work [18] where a series of polls and interviews to different SMEs employees and employers about how much they know about BI and the reasons why they don't use or wouldn't use these tools were made.
The possibility to count with either commercial or open-source tools to carry data analysis, data mining and visualization processes, allow organizations to have different alternatives according to their necessities and economic possibilities. This scenario, then, it's highly favourable for the implementation of these tools in SMEs.

3 Local SME experimentation

Data exploration allows statistical analysis and visualization. The proposal analysis and study were carried out in a transport enterprise from the province of San Juan, Argentina, classified as medium-sized, whose origins date back to 1944 as a family business. Currently, it tracks their daily travel movements through an Excel format exportable files system. Those in charge of making decisions have an arduous task when they have to consult these extensive files and they only have their experience years to do so. Although they have an established human resources department and a developed systems area, when the research group did the interview, the enterprise revealed that they didn't know the benefits of implementing BI and data mining. So, when the topic was explained, they argued that they didn't count on the financial and technological capacities to implement it, reinforcing what was proposed in [19],[20]. As an information source from their system, an Excel file that contains records from November 2019 to February 2020 was exported. Five dimensions were identified: travels, mobilities, personnel, users and categories. For the travels dimension, five variables were identified: service exits, voucher number, cycle, cycle's costs, associated costs, responsible, season, service type, route, platforms, origin occupation, afterwards occupation, total occupation, status, cash flow, benefit, and time and date of the benefit. For the mobilities dimension, the variables identified were: bus number, capacity and description. For the personnel dimension, the variables identified were: employee id, first name, last name and category. For the categories dimension, the variables identified were: category id, description, antiquity and basic salary. Lastly, for the user dimension, the variables identified were: id user, first name, last name and category. Through this research work, the associated benefits of implementing analysis and data mining, and predictive tools in SMEs were evidenced.

3.1 PowerBI Desktop

To begin the study and to use interactive panels to visualize, it was decided to use PowerBI Desktop.

1. Data source pre-processing. The source file, originally having only one tab showing all information, was divided into four tabs considering the dimensions established. Inconsistencies and redundancies were also solved.
2. Data collecting. The tables obtained from these tabs were loaded in PowerBI Desktop environment (Figure 1 and Figure 2).

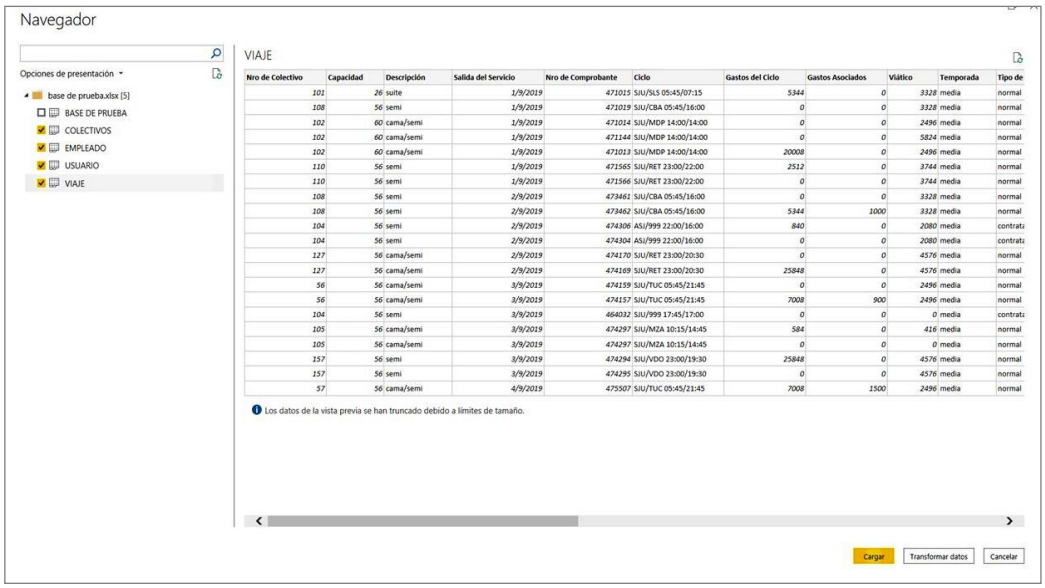


Fig. 1. Data loading in PowerBI Desktop environment

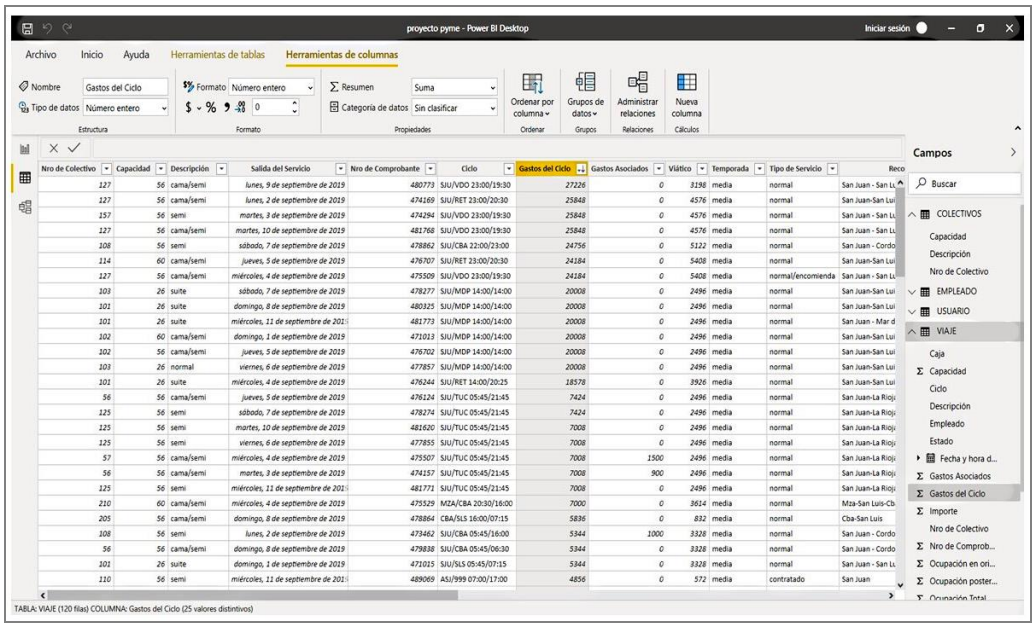


Fig. 2. Travel dimension data loading

3. Data preparation. Minimal modifications were done, such as some columns formats modifications to have better visualization. For example, service exits, add currency to the cycle's costs column, etc. (Figure 3).

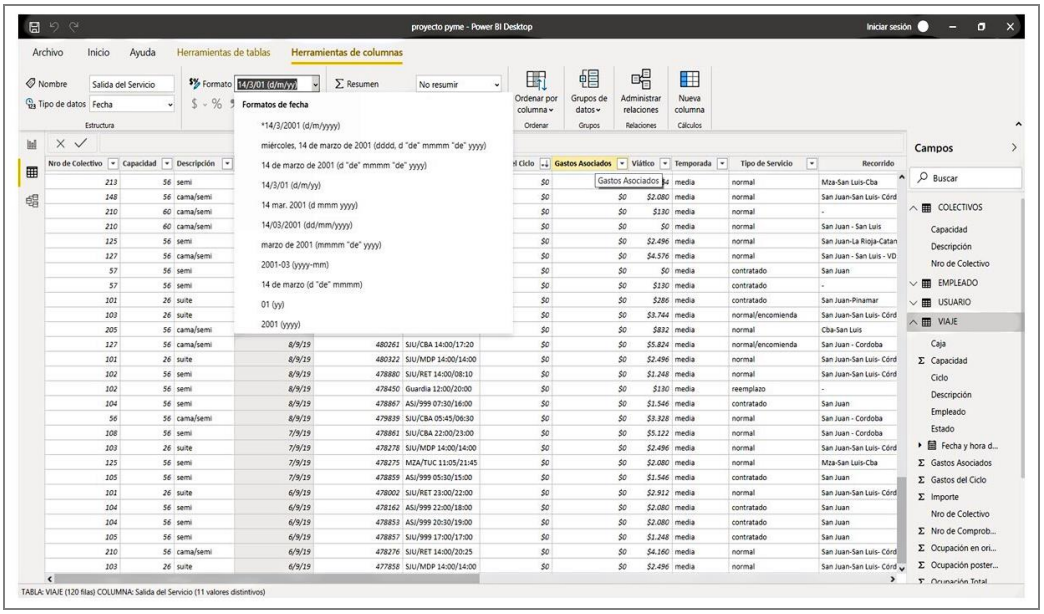


Fig. 3. Data preparation in travels table

4. Data modelling. A model was defined, establishing relationships between tables (Figure 4).

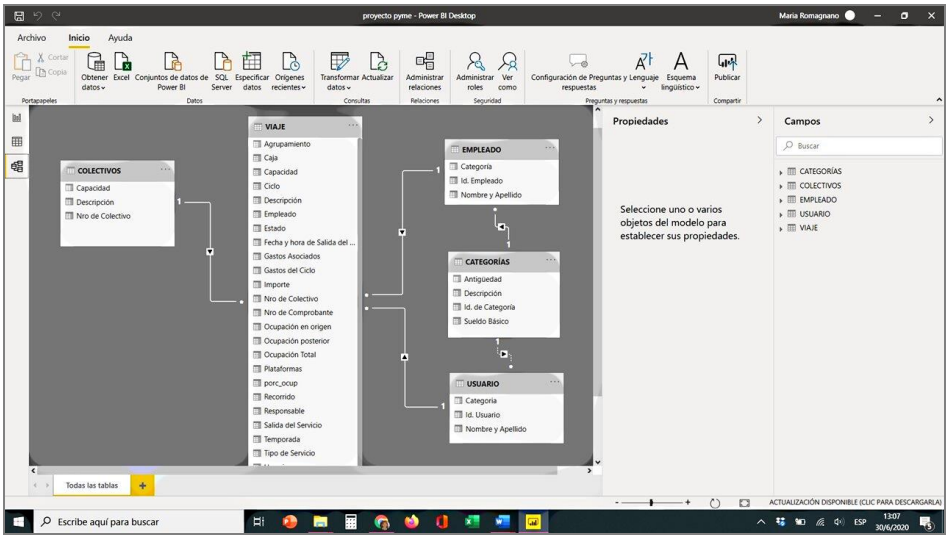


Fig. 4. Data modelling

5. Data visualization. According to the enterprise's requests, different visualizations were generated. Besides the interviews made with administrative personnel and current system users, some unidentified information needs were detected. Likewise, some inconsistencies and errors were observed, both in data loadings and analysis by eye.

   In Figure 5 it can be observed that the table presents information about mobility capacity, total occupation, and occupation percentage with respect to its capacity. These last two calculations weren't contemplated in the enterprise's records and, according to what they manifested, it was important to count on this knowledge because it allows them to analyse the convenience of assigning particular mobility with certain characteristics to a particular cycle (or route). It even gives them the possibility to observe if it's convenient to cover this route, having travels costs in mind.

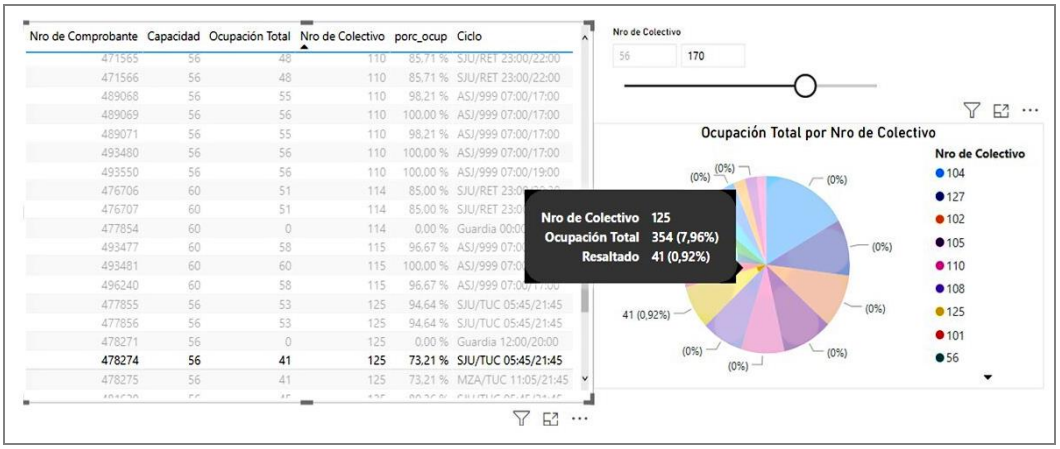


Fig. 5. Occupational capacity and mobility occupation percentage visualization, according to the data table

6. Data reports. With the integration of some visualizations presented and selected by the enterprise, a report was generated (Figure 6).

3.2 Orange and KNIME

In this section, two similar open-source platforms are analysed. Orange Data Mining and KNIME, both GPL licensed ("Table 1"). It is noted that both platforms can be executed in Windows, Linux and Mac operating systems. They also have similar features, such as:

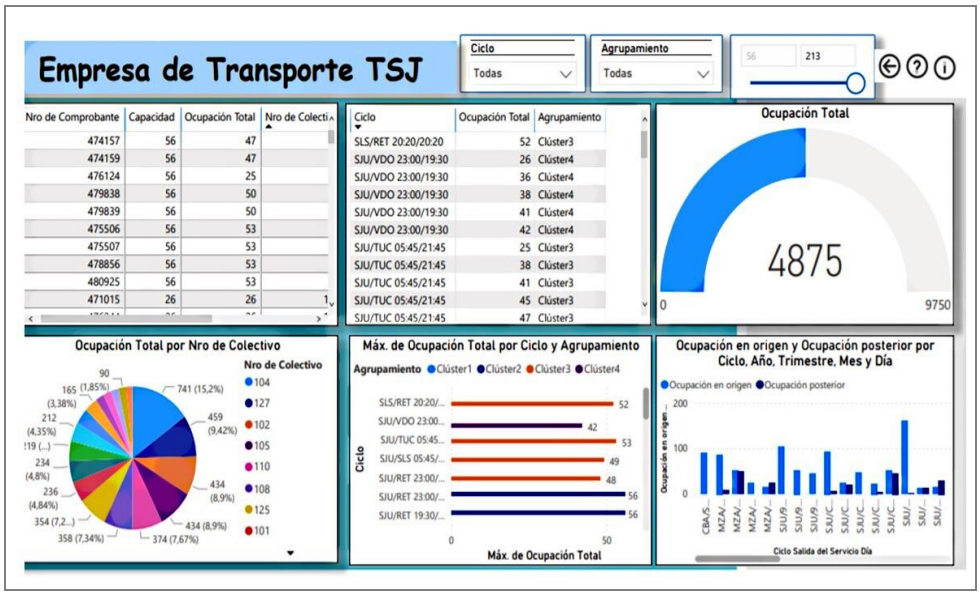


Fig. 6. Full data reports

- Workflow: where data analysts, business analysts and data scientists work directly with the data, using the component palette's drag and drop workflow (Figure 7 and Figure 8).
- Widgets: components that can relate to each other, which allows and simplifies work networks creation.
- Online documentation: both platforms offer documentation, tutorials, videos, and large support from community users.

Table 1. Orange DM and KNIME comparative

Name	Features	Development Bases
Orange Data Mining	An easy-to-learn and intuitive analysis platform. Contributes to multiple mining and visualization algorithms.	Doesn't require additional software, it can be connected to some database engines and files such as Excel or CSV. Its development base is Python.
KNIME	Complete analysis platform. Higher difficulty and complexity than Orange.	Doesn't require additional software, it can be connected to some database engines and files such as Excel or CSV and also to BI tools such as Microsoft's PowerBI. Its development base is Java.

The algorithms set, that's part of Artificial Intelligence (AI) and is contributed by both tools, is divided into two large groups: supervised and unsupervised learning algorithms. The first one is the most common and it's called like that because the developer acts as a guide to teach the algorithm the expected conclusions, namely, the expected exit is already known. It is similar to the way a kid could learn from a teacher, given that they are trained to answer to different situations. The idea behind unsupervised learning is that a computer can learn to identify complex processes and patterns without a human present to provide orientation along the way [21]. It isn't the objective of this work to delve into all technique concepts, just to present the summary in "Table 2" [22].

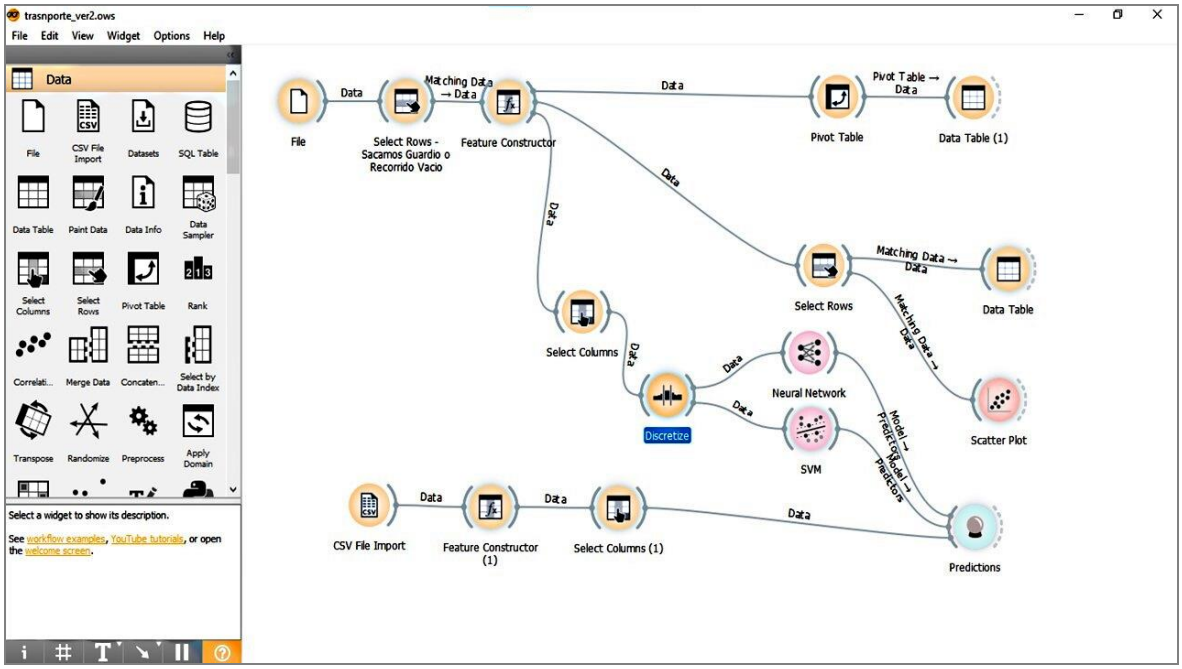


Fig. 7. Orange Data Mining Version 3.26 workflow, with an interconnected widget example

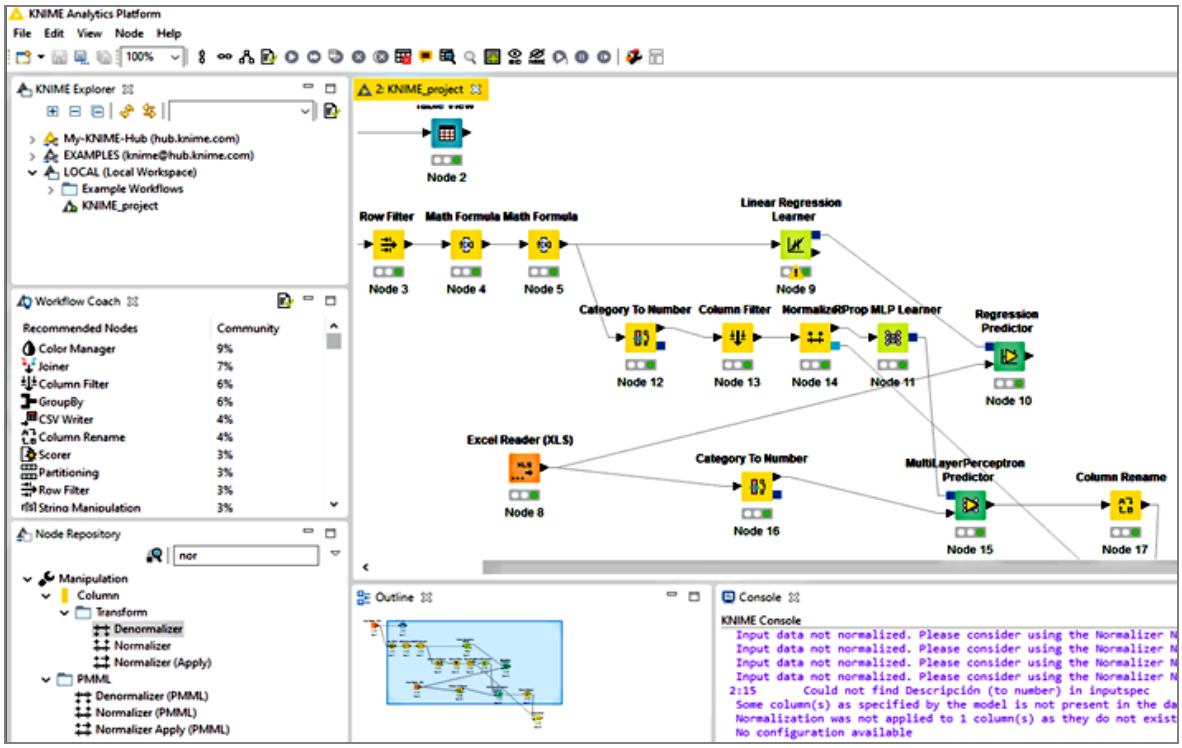


Fig. 8. KNIME Version 4.1.3, workflow, with an interconnected widget example

Data entry to the workflow was made through the widget “Data Origin”. Orange Data Mining and KNIME allow loading data directly by Excel or .csv files (“Table 3”).

Table 2. Data Mining common techniques [22]

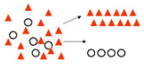
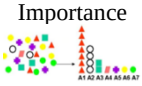
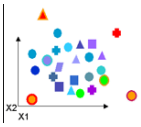
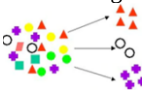
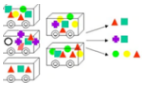
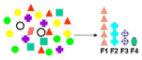
Technique	Applicability
Classification 	Commonly used technique to predict a specific result as an answer/not an answer, high/medium/low-value customer, with a probability to buy/not buy. Stand out: Generalised Linear Models (GLM) Naive Bayes Classifier Support Vector Machines (SVM) Decision Tree
Regression 	Technique to predict a continuous numerical result, such as a living place's value in time, investment return rates. Stand out: Generalised Linear Models (GLM) Support Vector Machines (SVM)
Attribute Importance 	Classifies the attributes according to the strength of their relation with the target attribute. For example, use cases include search factors more associated with customers responding to an offer, or factors more associated with healthy patients. Stand out: Minimum Description Length (MDL)
Anomaly Detection 	Identifies unusual or suspicious cases based on norm deviation. Common examples include medical attention fraud, expenses report fraud and tax non-compliance. Stand out: One-Class SVMs Covariance estimator Local Atypical Value Factor Isolation Forest
Clustering 	It is useful to explore data and find natural clustering. Members of a cluster that look more like each other than members of a different cluster. Most common examples include new customer segments search and life or medical science discovery. Stand out: K-Means Orthogonal Partitioning Clustering or Hierarchical Clustering. Expectation Maximization
Association 	Finds associated rules with items that are frequently produced, utilized for the shopping basket analysis, crossed sales, root cause analysis. It is useful for products groups, stores placement and flaws analysis. For example, if a client buys a knife and an after-shave lotion, there is an 80% chance that they also buy shaving cream. Stand out: A Priori
Feature Selection and Extraction 	Produces new attributes from a linear combination of already existing ones. Applicable for text data, latent semantic analysis, data compression, decomposition and projection, and pattern recognition. Stand out: Not Negative Matrix Factorization Principal Component Analysis (PCA) Singular Value Decomposition (SVD)
Neural Networks 	Set of nodes known as artificial neurons that are interconnected and send signals to each other. These signals transmit from entry until an exit is generated. It is used to train classification models that are more complex than a simple yes/no. Stand out: Multi-Layer Perceptron (MLP)

Table 3. Data reading in Orange DM and KNIME

Orange 	In Orange, the Excel file containing the data-sheet is directly read. It's characterized by the field type on "DateTime", "categorical", "numeric" or "text". Also, it is possible to indicate the attributes' role in "entry variable" or "target".
KNIME 	In KNIME the Excel file containing the data-sheet is directly read. It is not automatic; the user must execute the component for it to be available.

Data exploration allows statistical analysis and visualization. In Orange, through the widget "Feature Statistics" and connecting to the data origin, information was extracted immediately.

For example, in Figure 9, the field "Money exits date and time" has an outlier value, dispersion is 19 years, and the minimum value corresponds to the year 2000. This indicates that there would be an error in data entry. The analyst, then, can know if they have to clean the source table.

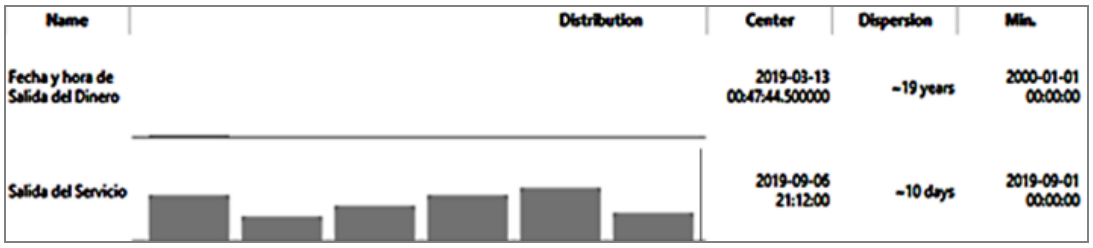


Fig. 9. Feature Statistics Visualization in Orange Data Mining

KNIME offers similar widgets to visualize data. For example, "Bar Chart" allows creating a bar chart diagram directly from the XLS widget that reads the source data. In Figure 10 it is discovered that for the "Day of Exit" variable, Wednesdays are the days with most service exits.

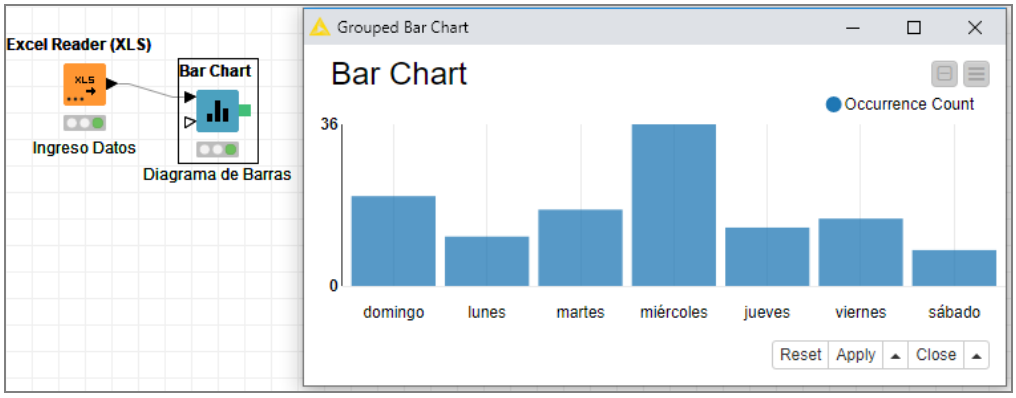


Fig. 10. Statistics visualization in KNIME

In another analysis made with Orange (Figure 11), the attribute Platform had an 8% missing data. It was also identified that, for future analysis, the "Status" attribute could be dispensed with, because it has only one value.

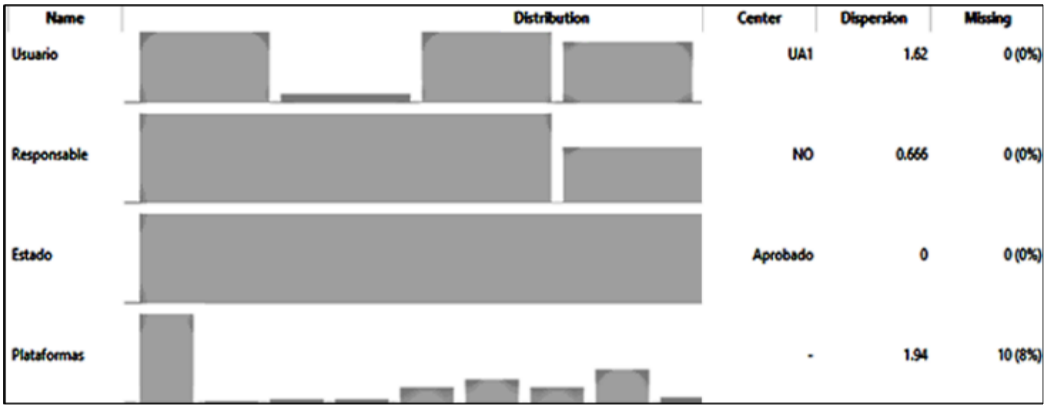


Fig. 11. Orange Data Mining's Visualization of Feature Statistics

Also, a new set of variables was created, which were calculated from the base ones that were already available. In Figure 12, the widget "Feature Constructor" from Orange was utilized to generate other variables. Figure 13 presents an example in KNIME where the "Math Formula" widget was used.

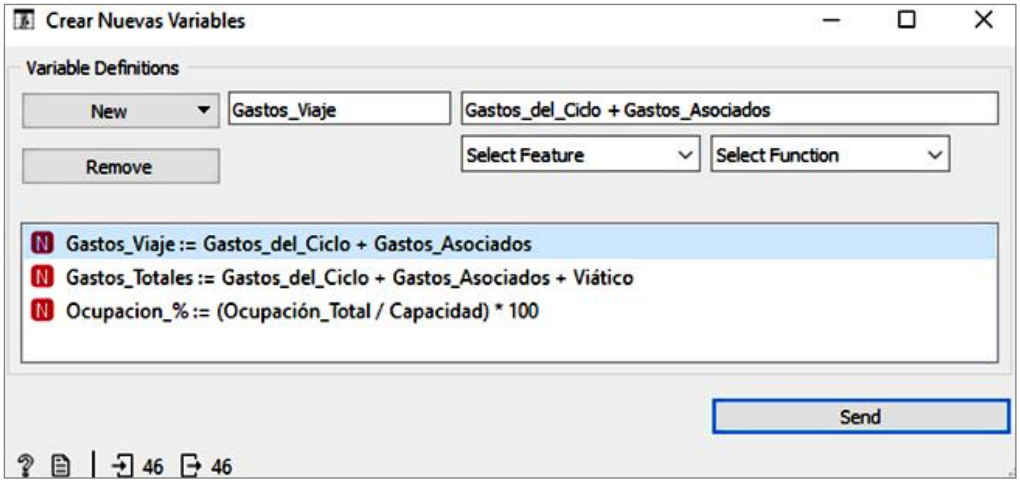


Fig. 12. Creation of 3 new calculated variables in Orange Data Mining

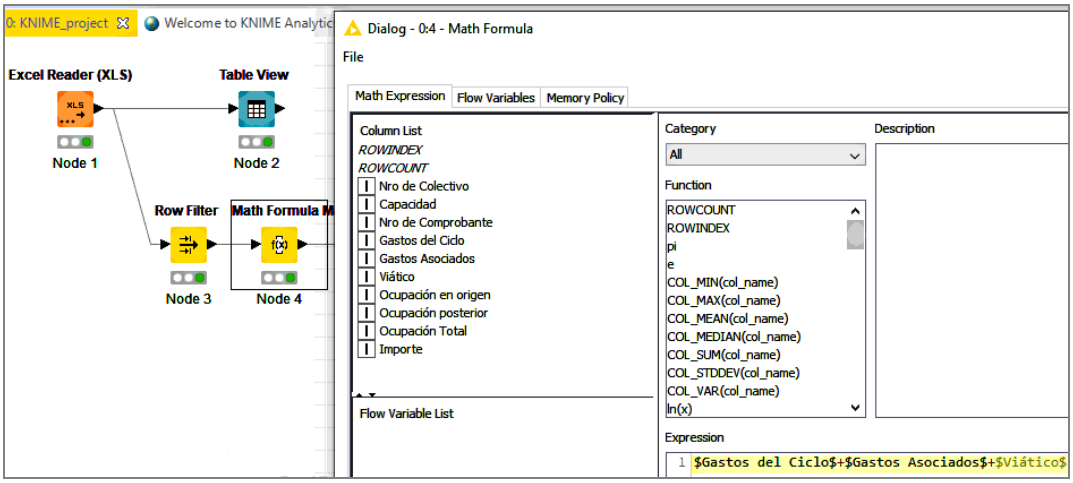


Fig. 13. Creation of a new calculated variable in KNIME

More conclusions can be extracted in the creation of this new variable. For example, the average bus occupation was around 83% (Figure 14).

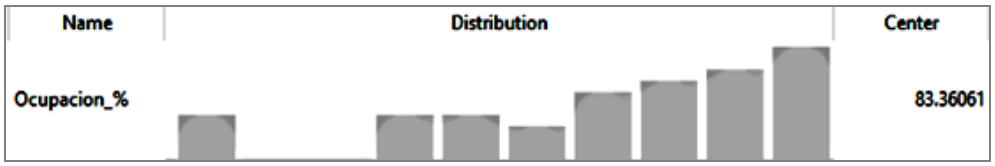


Fig. 14. New calculated variable visualization

One of the advantages of unsupervised data mining algorithms is the possibility to find not only patterns but also anomalies. These are data that "escapes" the normal set.

Orange's "Outliers" widget was also utilized, with the "Local atypical value factor" algorithm, to obtain the local density of the closest k neighbours (Figure 15).

Fig. 15. Atypical Value Factor algorithm result

According to this algorithm, three atypical values were identified. Analysing the cycle "ASJ/999 07/17" two outliers were identified. One of the outliers corresponded to a mistake in the writing of a value ("contrado" should be "contratado"). Also, the cycle SLS/RET was detected as an outlier because it was the only registry with that value in the whole data set. With this simple algorithm, the data set can be "purified" and mistakes in data entry could be easily identified.

Orange, like many data mining tools, provides a set of supervised algorithms to make forecasts and predictions. For example, the behaviour of the bus occupation can be estimated. In this case, "Select Columns" was employed to indicate the model how it must be trained, which variables are used as entry and which as target (Figure 16).

Orange's KNN widget uses the closest k neighbours' algorithm. It is a supervised classification method that is used to estimate each class predictors density function. This is, it searches k closest training examples in the features space and uses their average as a prediction. This is one of the various models that can be used in the applicative.

Figure 17 shows the result after the application of a test/prediction data set. For example, KNN established that Fridays should expect high occupation to service "semi" and Thursdays should have low occupation for the cycle SJU/TUC.

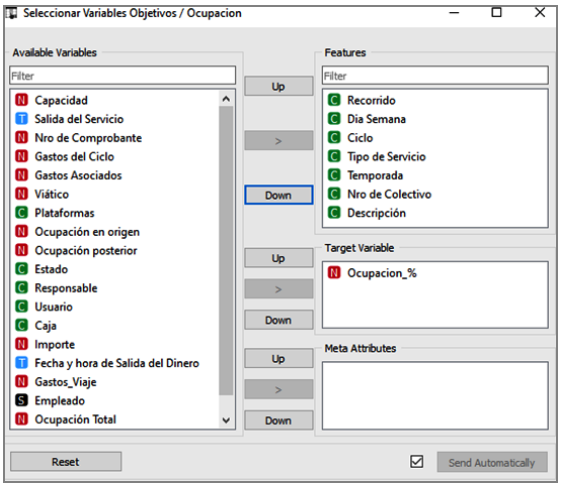


Fig. 16. Entry and target variables selection in Orange Data Mining

kNN	Num Escenario	Dia Semana	Ciclo	Descripción	Tipo de Servicio
81.786	13	viernes	SJU/TUC 05:45/21:45	semi	normal
78.450	6	viernes	SJU/TUC 05:45/21:45	cama/semi	normal
80.046	10	martes	SJU/TUC 05:45/21:45	semi	normal
74.826	3	martes	SJU/TUC 05:45/21:45	cama/semi	normal
77.439	7	sábado	SJU/TUC 05:45/21:45	cama/semi	normal
78.937	14	sábado	SJU/TUC 05:45/21:45	semi	normal
73.151	9	lunes	SJU/TUC 05:45/21:45	semi	normal
82.209	2	lunes	SJU/TUC 05:45/21:45	cama/semi	normal
73.441	8	domingo	SJU/TUC 05:45/21:45	semi	normal
80.318	11	miércoles	SJU/TUC 05:45/21:45	semi	normal
76.224	4	miércoles	SJU/TUC 05:45/21:45	cama/semi	normal
78.236	1	domingo	SJU/TUC 05:45/21:45	cama/semi	normal
72.143	12	jueves	SJU/TUC 05:45/21:45	semi	normal
66.944	5	jueves	SJU/TUC 05:45/21:45	cama/semi	normal

Fig. 17. Results of KNN algorithm in Orange Data Mining

KNIME can also apply supervised learning algorithms. A RProp MLP model can be trained. This is a Neural Network type [23]. Figure 18 shows an example where the network was trained with variables like expenses, service cost, breaking costs, among others, for it to forecast which services are the most expensive according to their route and exit.

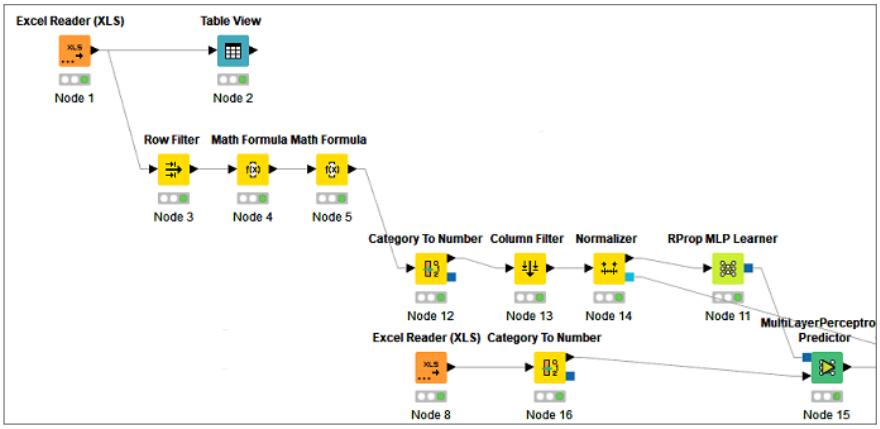


Fig. 18. Training and test MLP RProp network in KNIME

Figure 19 shows the result, where it is discovered that Monday's cost may be higher than Saturday's cost in route SJU/TUC with service "semi". This is correct, given that expenses can be higher for the drivers on weekends.

Denormalized output - 0:18 - Denormalizer

File Hilite Navigation View

Table "default" - Rows: 14 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	I Num Es...	S Dia Se...	S Ciclo	S Descrip...	D Gastos...
Row8	9	lunes	SJU/TUC 05:45/21:45	semi	25,959.877
Row6	7	sábado	SJU/TUC 05:45/21:45	cama/semi	24,185.401
Row13	14	sábado	SJU/TUC 05:45/21:45	semi	23,845.248
Row5	6	viernes	SJU/TUC 05:45/21:45	cama/semi	22,979.931
Row9	10	martes	SJU/TUC 05:45/21:45	semi	22,581.556
Row7	8	domingo	SJU/TUC 05:45/21:45	semi	21,529.724
Row4	5	jueves	SJU/TUC 05:45/21:45	cama/semi	21,035.659
Row12	13	viernes	SJU/TUC 05:45/21:45	semi	19,301.47
Row3	4	miércoles	SJU/TUC 05:45/21:45	cama/semi	18,643.544
Row2	3	martes	SJU/TUC 05:45/21:45	cama/semi	16,596.218
Row1	2	lunes	SJU/TUC 05:45/21:45	cama/semi	16,474.27
Row11	12	jueves	SJU/TUC 05:45/21:45	semi	14,229.383
Row10	11	miércoles	SJU/TUC 05:45/21:45	semi	12,945.995
Row0	1	domingo	SJU/TUC 05:45/21:45	cama/semi	12,202.149

Fig. 19. RProp MLP model result in KNIME

3.3 Knowage

Among the different software tools for BI and BA that had been used in this work, one of them it's Knowage. This tool allows data and information visualization through a wide set of tables, graphs, histograms, etc. All of these visualizations can be viewed through the so-called *widgets*, which are completely adjustable and customizable. These widgets are placed in *cockpits*, workplaces where all the widgets from the same analysis are placed together.

Knowage also includes some widgets that allow the use of different embedded programming languages, such as R or Python, that can be used to make predictions and processes like machine learning. Later, through the same widget, Knowage allows these predictions to be visualized.

Knowage allows the use of different data types, that can come from different and varied data sources, such as .xls and .xlsx files, .csv files, SQL databases, MongoDB, MariaDB, among others. All databases (except .xls, .csv and other similar files) are loaded as *data sources* and, from there, all data inside of them that is needed or wanted to be worked on and visualized are selected. These selected data are now known as *dataset*. These datasets, as their name acknowledges it, are specific sets of data selected from a particular database. These datasets can be created through queries to the different databases or, for files such as .xls, .csv or similar, being loaded (in these particular cases, all of the data is loaded in the dataset, not just a part of it).

Knowage has two versions on its site to download, a free one and a premium one that requires payment. However, to an SME, the difference between both is minimal and wouldn't bring any substantial benefits to work with the premium version instead of the free one.

For the use of Knowage, data from this enterprise was loaded as a dataset from a .xls file (Figure 20 and Figure 21). After data is loaded, it is necessary to validate the type and the way they'll be used (Figure 22). *Type* references if data will be an integer, a float number, a string, a date, among others. *The way they'll be used* means if the data will be used as an attribute where they'll represent an intrinsic characteristic that shouldn't be affected or modified by the day-a-day basic operations (as an example in the transport enterprise, the bus number, description, form number would be the attributes); or, on the other hand, another way to catalogue them would be as measures, where each data represents, as its name acknowledges, a measure, something that may be modified or altered by the daily operations or due to different particular factors (an example for the transport enterprise could be some measures, such as total occupation, benefits, costs, among others).

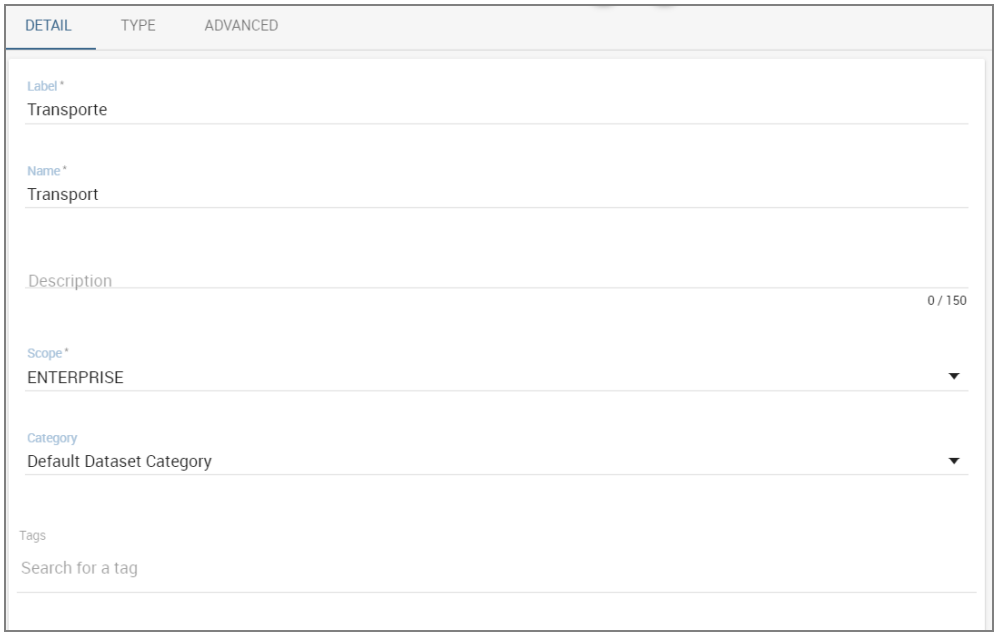


Fig. 20. Dataset preparation

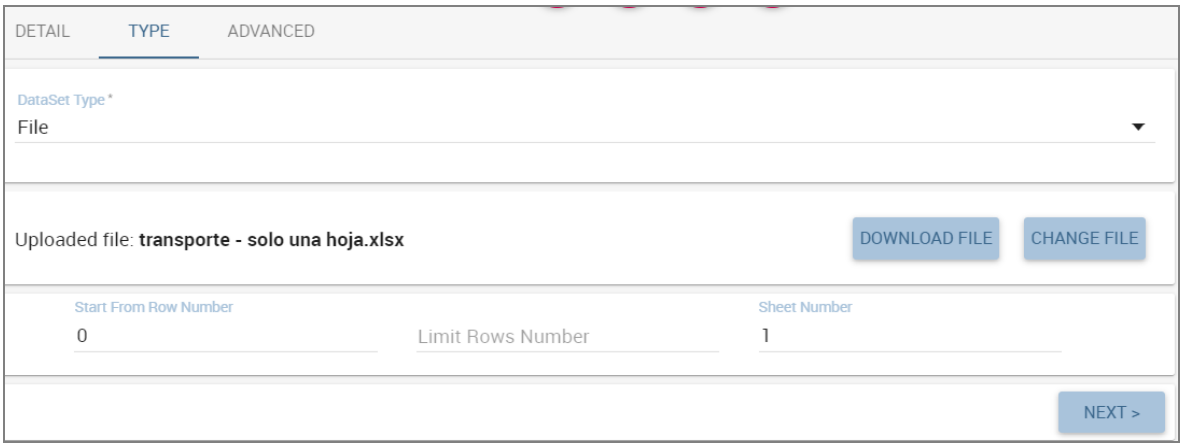


Fig. 21. Transport SME dataset load

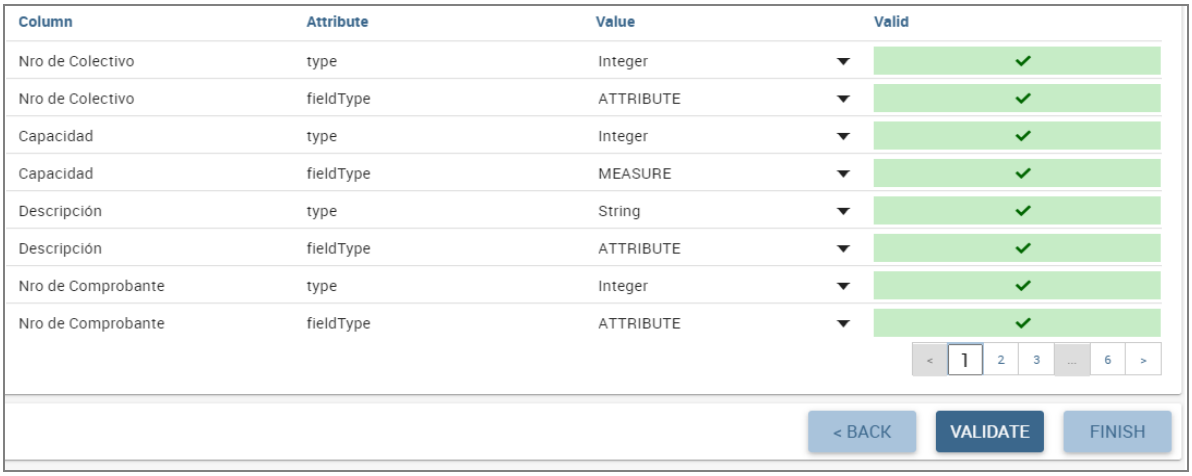


Fig. 22. Dataset validation

Before saving them, a preview of the dataset, with data, types and ways of using them, can be done, to determine if everything is correct and complete (Figure 23).

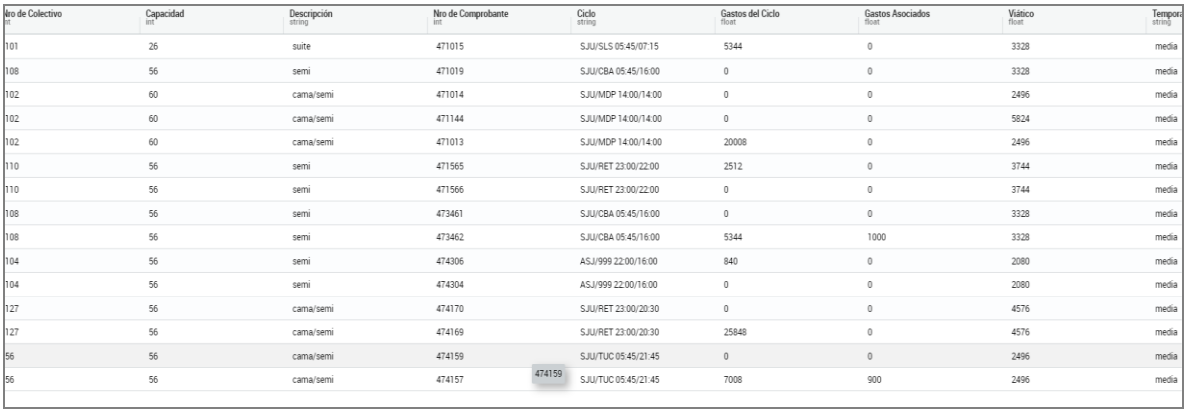


Fig. 23. Dataset preview

Once is loaded, validated and saved; data is ready to be used in many widgets across the different cockpits. In this case, data visualization was presented via pie charts, histograms, and temporal graphs. However, there are other widgets to work with, such as cross tables, comparative graphs, word clouds, among others.

The first widget used is a pie chart showing the total occupation for cycle and bus number (Figure 24). Here, the total occupation for a particular bus number can be seen and, at the same time, each cycle that the bus has participated in. This way, not only the total occupation of just a bus had can be appreciated, but also how much of that occupation occurred in each cycle, giving the visualization a higher level of detail.

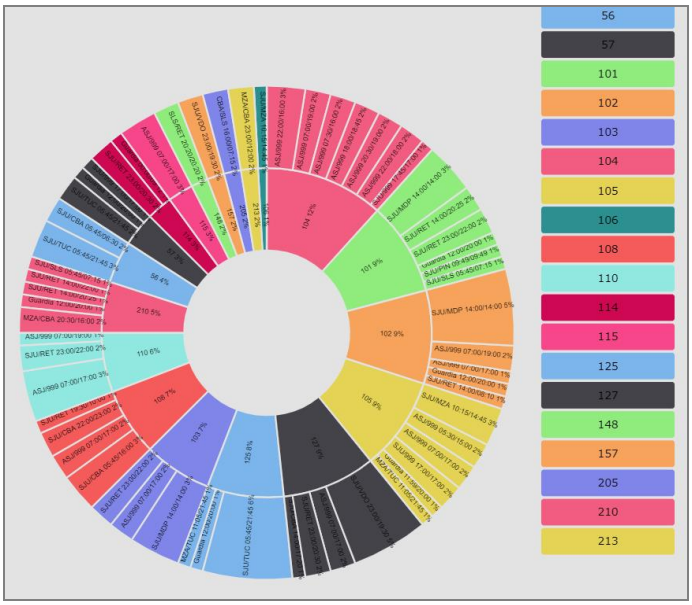


Fig. 24. Total occupation for cycle and bus number

After that, it was analysed and compared an average of benefits and an average of costs for the service type, taking advantage of the widget option that allows to use more than one measure in the same graph and to work with averages and not just total sums or counts (Figure 24).

Another analysis was made to determine and compare total benefits and costs for a service day. This allowed showing, for each individual service day, what was the overall profit or losses (Figure 25).

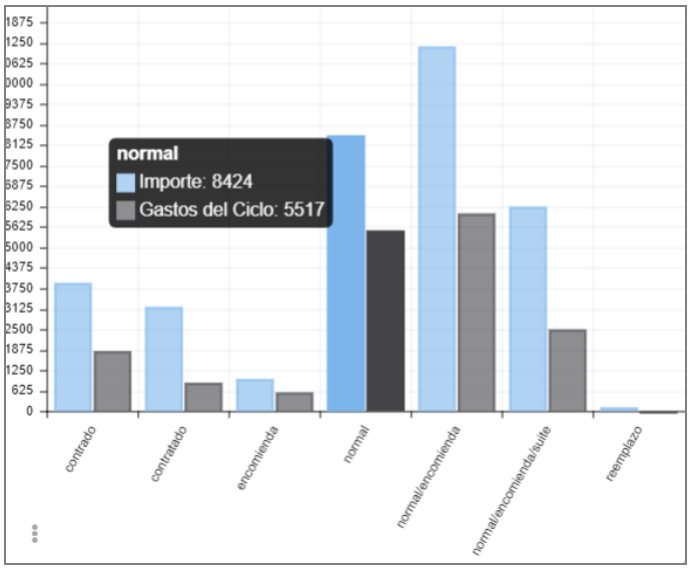


Fig. 25. Average benefits and costs for service type

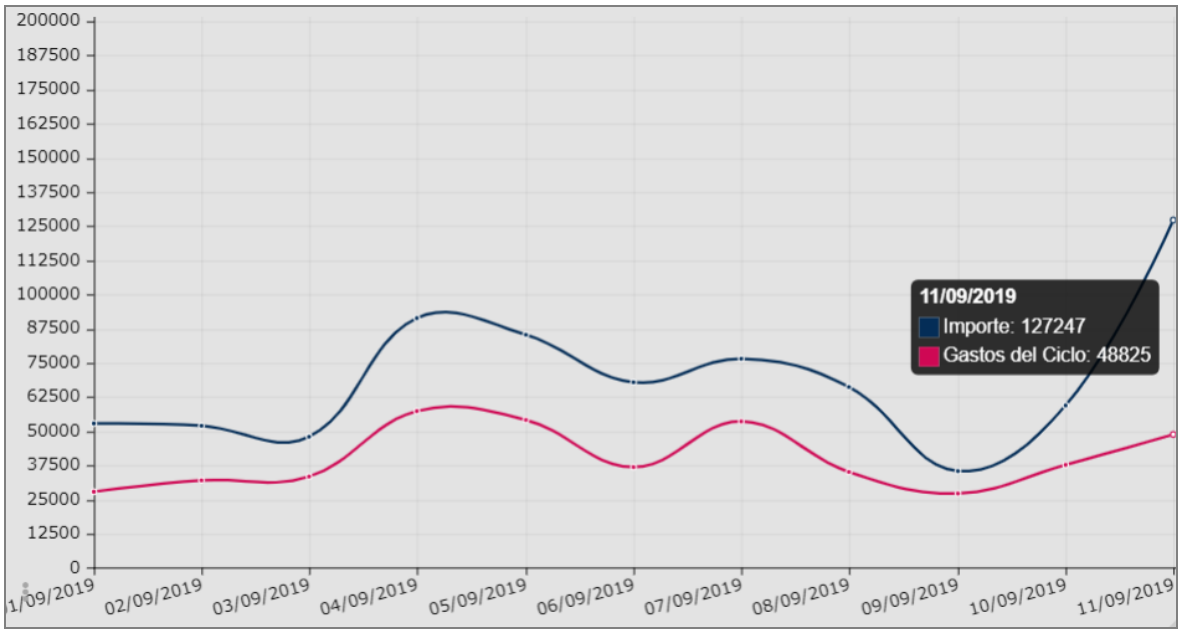


Fig. 26. Benefits and costs for service day

Finally, a cockpit with three widgets on the same page is shown (Figure 27). However, in Knowage is easy and simple to put everything in the same cockpit or, even, on different pages, like a spreadsheet. All of this is thanks to the adjustability and customizability of the different widgets.

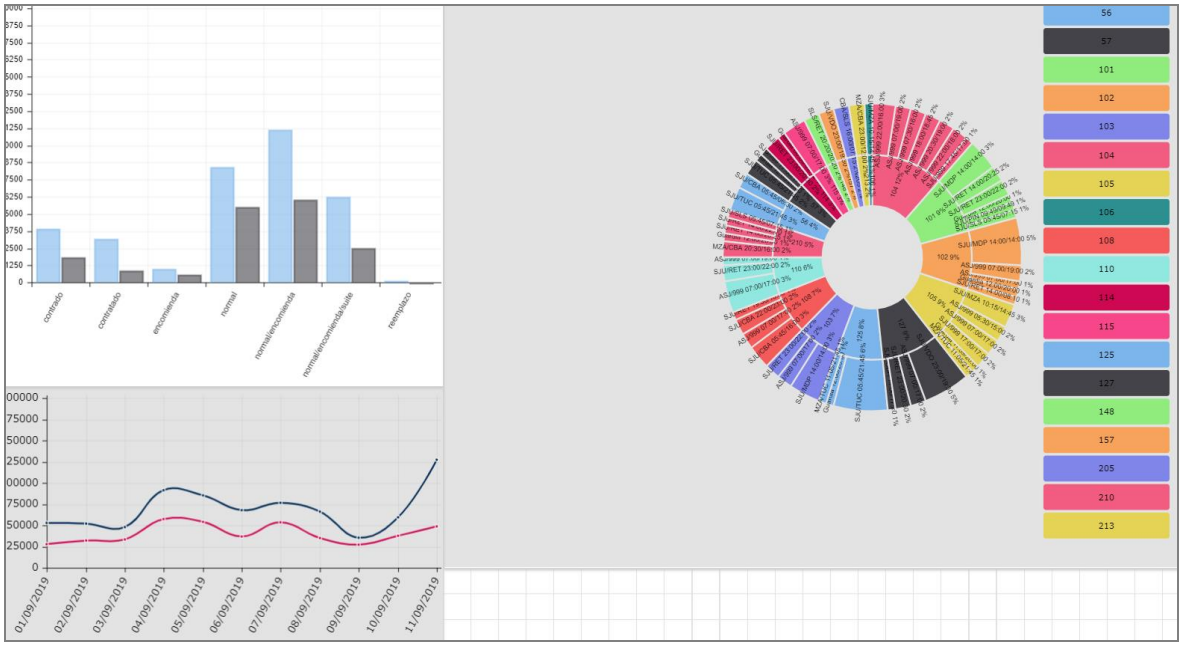


Fig. 27. Cockpit with multiple widgets at the same time

4 Conclusions

The current market in which the enterprises compete is highly complex, requiring significant competitive advantages, based on information and knowledge, to be able to both survive in it and take a leadership position against the competition. This reality manifests the necessity to enhance information systems so they can be used at the moment to make a decision, being sure that it is the best and most adequate to each opportunity. The existence of BI, BA, data mining, analytics and visualization tools,

supports this enhancement, allowing enterprises to know, with high precision, their current status and, with a base on the presented information, predict future market behaviours and make proactive actions.

On the other hand, thanks to these tests made with Orange, KNIME, PowerBI and Knowage, it could be evidenced that data mining and visualization tools are at SMEs reach. They do not require large and complex infrastructures or resources to work correctly. Also, because it is open-source software, it is not an unreachable alternative.
In the case of SME, where they do not have pre-emptively these kinds of software, training and participation from an analyst that has experience in data mining will be necessary, to be able to discover valuable information to the enterprise that can be hidden in different kinds of files. Another benefit of these tools is the feedback that can be done to the enterprise information systems, showing and bringing knowledge to make a decision based on accurate information.

It should also be noted that, if the transport enterprise doesn't follow these recommendations given, it would not know, with high precision, its current status. It would not be able to forecast future market behaviours and take proactive actions based on presented information; be able to manage its current data correctly (regarding debugging and standardization, extraction, transformation and storage); discover new data (regarding compilation and predictive analysis on new information to make projections) and make adequate reports; regarding the visualization of the processed information either.

As future works, it is expected to test the different tools presented in this work in SMEs from different natures. These applications will be analysed to determine their enhancing capability in SMEs and to compare each other.

References

1. Bell-Friedel, L. "Marine Propulsion & Auxiliary Machinery", Pp.71-72. Update date: 29/10/2017. Consultation date: 20/03/2020. https://issuu.com/rivieramaritimemedia/docs/mp-aug17_text.
2. Di Pace, Damián. "Economía del Conocimiento: el gran ganador en pandemia y la clave de la pospandemia", Artículo Diario Ámbito Financiero. Consultation date: 14/08/2020, https://www.ambito.com/opiniones/economia-del-conocimiento/el-gran-ganador-pandemia-y-la-clave-la-pospandemia-n5125013.
3. Solano, L. E. "Business Intelligence: un balance para su implementación". InnovaG, Vol. 3. Pp. 27-36. Update date: 2019. Consultation date: 20/04/2020. http://revistas.pucp.edu.pe/index.php/innovag/article/view/19742.
4. Vadell, G., Auged, A. "La economía digital en el sistema tributario argentino". Update date: 2019. Consultation date: 11/03/2020. http://www.economicas.uba.ar/wpcontent/uploads/2019/01/EconomiaDigital-en-Argentina.pdf.
5. Llave, R. "Business Intelligence and Analytics in Small and Medium-sized Enterprises: A Systematic Literature Review". ScienceDirect, *Procedia Computer Science*. Vol. 121, december 2017, Pp. 194-205. https://doi.org/10.1016/j.procs.2017.11.02
6. Pedrycz, W., Martínez, L., Espin-Andrade, R. A., Rivera, G., Gómez, J. M (Eds). "Computational Intelligence for Business Analytics", Springer, 2021. doi.org/10.1007/978-3-030-73819-8
7. Rojo, P. (s.f.). "Cómo se benefician las SMEs de una solución de Business Intelligence". Consultation date: 18/09/16. https://dataiq.com.ar/blog/se-benefician-las-pymes-una-solucion-bi/.
8. Lantares Solutions. (s.f.). Aplicaciones prácticas del análisis de datos para las SME. Consultation date: 18/09/16. http://www.lantares.com/blog/aplicaciones-practicas-del-analisis-de-datos-para-las-SME.
9. Cayón, M. (2015). Cloud Computing, propulsor en la adopción de TI en las SMEs. Consultation date: 20/09/16 http://mundocontact.com/cloudcomputing-propulsor-en-la-adopcion-de-ti-en-las-SMEs/.
10. López Benítez, Y. "Business Intelligence. ADGG102PO". IC Editorial. 1° Edición. Andalucía - España. Vol. 1, 151 pages, 2018. ISBN: 978-84-9198-467-2.
11. Ahumada Tello, E., Perusquia Velasco, J. "Inteligencia de negocios: estrategia para el desarrollo de competitividad en empresas de base tecnológica". SciencDirect, Contaduría y Administración, Vol. 61(1), january - march 2015, Pp. 127-158. dx.doi.org/10.1016/j.cya.2015.09.00
12. Pazos-Rangel, R. A., Florencia-Juarez, R., Paredes-Valverde, M. A., Rivera, G. (Eds.). "Handbook of Research on Natural Language Processing and Smart Service Systems", IGI Global, 2021. doi.org/10.4018/978-1-7998-4730-4
13. Orange. "Data Mining Fruitful and Fun". Update date: 15/08/2020. Consultation date: 15/08/2020. https://orange.biolab.si/
14. KNIME. "Open for Innovation". Update date: 15/08/2020. Consultation date: 15/08/2020. https://www.knime.com/
15. Digital Guide IONOS. "Software de data mining: realiza análisis de datos más efectivos". Update date: 2018. Consultation date: 17/08/2020. https://www.ionos.es/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas
16. KNOWAGE. Fecha de actualización: 2021. Update date: 01/05/2021. https://www.knowage-suite.com/site/
17. Flores, J. "La Importancia de la Inteligencia de Negocios Aplicada a Empresas Medianas". Update date: 30/12/2010. Consultation date: 03/06/2020. https://www.ibm.com/developerworks/ssa/local/data/dm-bi-SMEs/.
18. Tovar, C. (2017). Investigación sobre la Aplicación de Business Intelligence en la Gestión de las SMEs de Argentina. *Palermo Business Review*, 15, 79-97
19. Ministerio de Ciencia, Tecnología e Innovación, Sistema Integrado de Indicadores. "Inversión en I+D". Update date: 2018. Consultation date: 17/08/2020. https://www.argentina.gob.ar/ciencia/indicadorescti

20. Ministerio de la Producción. "SMEs registradas por región". Update date: no data. Consultation date: 28/05/2020.https://www.produccion.gob.ar/SMEsregistradas/
21. González, L. "Diferencia entre aprendizaje supervisado y no supervisado". Update date: 2018. Consultation date: 15/08/2020. https://ligdigonzalez.com/diferencia-entre-aprendizaje-supervisado-y-no-supervisado/
22. ORACLE. "Oracle Advanced Analytics' Machine Learning Algorithms SQL Functions". Update date: 2020. Consultation date: 15/08/2020. https://www.oracle.com/database/technologies/advanced-analytics/odm-techniques-algorithms.html
23. KNIME. "RProp MLP Learner". Update date: no data. Consultation date: 17/08/2020. https://hub.knime.com/knime/extensions/org.knime.features.base/latest/org.knime.base.node.mine.neural.rprop.RPropNodeFactory2