# ***CHAPTER 3: DATA MINING ALGORITHMS FOR THE FASHION RETAIL***

In recent decades, the development of information and communications technologies have given new vitality to the company's marketing. The data to be stored and to be analyzed are increasing at a very rapid pace, probably 1000 times compared to five years ago. However, data and corporate earnings are not directly proportional.

These applications of Data Mining algorithms are referred to a Boundary Science, a variety of scientific theories based primarily on the basic disciplines of Information Technology, Marketing and study of the statistical methods, that is the basis of every possible algorithm. In addition, data mining also refers to literary and behavioral disciplines to better evaluate the characteristics of a customer, such as psychology and sociology [20].

In general, through the extraction, transformation and loading of a large amount of information, we are able to identify the interests, preferences and behaviors of specific groups or individual consumers, but above all, the forecast of consumption, orienting sales for the specific marketing.

Since automation is popular throughout the industry, the companies that manage the processes must have many operational data. The data are not collected for the purpose of analysis, but originate from commercial operations. The analysis of these data gives decision-makers the real value of the information, in order to obtain profits.

The commercial information coming from the market through various channels, for example, one of the most popular is the purchase process by credit card where we can collect consumer data, such as time, location, assets or interesting services concerned, took prices and the level of capacity receipt. In addition, companies can also purchase a variety of customer information from other consulting firms.

The marketing based on data mining can usually create specific sales promotions for the client according to his previous purchase. The most common applications are in banking, insurance, traffic system, retail and commercial matters.

As described in State of the art, technology and marketing analyzes are based on the analysis of the market, such as prediction, segmentation and classification of customer profiling and cross-selling. They can also be used for credit scoring and fraud operations.

***Immagine che contiene testo

Descrizione generata con affidabilità molto elevata***

Figure 31: Applying Data Mining in Marketing

The basic process of data mining in marketing shows as follows:

• *Prepare the primitive data:* Includes personal information (age, gender, hobbies, background, profession, address, zip code and income), the earlier shopping experience and customer relationship. The preprocessing of the early data is very important to select potential customers,

• *Establish a certain pattern*: This model can be created by using very traditional technologies mining technologies. However, the problem of these technologies can be solved to identify the best or acceptable market within source of limited information, limited time and limited expenses.

Ultimately, we are using this model to select customers and decide the marketing plan.

In our project we are going to idealize a possible prediction of the ISTAT 2018 Italian data using linear regression, starting from 2007 data until 2017 by the eponymous site [25]. In addition, I try to find a possible best geo location to open a new store. and, in a second time, I will develop a classification CART to understand future expectations of the stores are currently open.

## ***3.1 PREDICTION OF THE BEST GEO - LOCATION TO OPEN A NEW STORE***

For a complete and reliable analysis, it is very important the integrity and completeness of the data. After a search in various sites dedicated to Open Data I have chosen the data given by the National Institute of Statistics, called ISTAT [25].

The actual data have a horizon starts from 2004 and finish in 2017. For make a good analysis of 2019, the data provided can not be considered complete. Therefore, it was decided to address the problem by considering a time horizon of ten years, considering the data from 2007 to 2017 to predict also the 2018.

Research to determine the appropriate indicators to the analysis, was held following a matrix process, making a mapping divided into geographical areas and time series covered, inserting all into an explanatory table shown in Appendix A2.

The preference at regional level are expressed with a total of five indicators, one or maximum two for each sector:

• Transportation Sector: operating rail network;

• Family Sector: Average monthly household expenditure on non-food goods and services, average income;

• Macro-Economics Sector: GDP Per Capita;

• Work Sector: Unemployment Rate.

While, at the municipal level, there was evidence the resident population and tourism.

Before the implementation of the code, it was necessary to create worksheets that are unique to each indicator, favoring an efficient data prediction, given the diversity of origin of the same. The subdivision, to avoid complicated mechanisms and complicated process and any copying errors, was made with Talend Open Studio [28], ETL software already widely discussed in the previous chapter.

The job takes Complete data previously loaded into the Staging Area and through the use of queries, each table is interrogated to extract the dedicated Excel spreadsheet, that will be the regional datas used to our analysis.

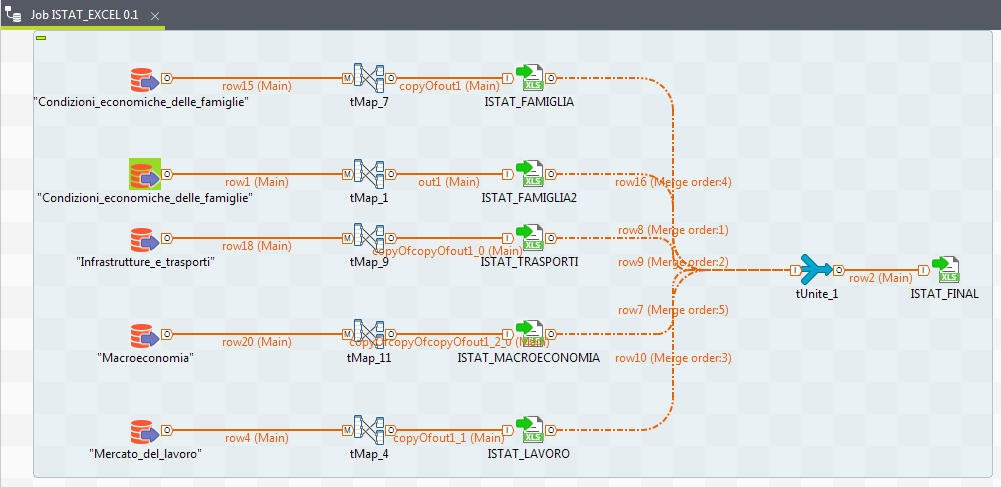


Figure 32: Job ISTAT Excel

For example, to extract only the data for the *'Spesa media mensile familiare per beni e servizi non alimentari'*, it is considered the table dedicated to the economic conditions of families data of the National Institute of Statistics highlighting the SQL IN clause. The same thing was done to exclude national data or related to membership in areas to be excluding (North Central, Northeast, ...) with a NOT IN.

Of course, the same operations were conducted for the other indicators.

The query used in the case of the average monthly expenditure is as follows:

"

***SELECT*** *\**

***FROM*** *OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie*

***WHERE*** *Indicatore* ***IN*** *(**'Spesa media mensile familiare per beni e servizi non*

*alimentari')*

***AND*** *Territorio* ***NOT******IN*** *('Nord-ovest',*

*'Bolzano/Bozen',*

*'Trento',*

*'Nord-est',*

*'Nord',*

*'Centro',*

*'Centro-Nord',*

*'Mezzogiorno',*

*'Italia') "*

By creating Excel files unique to each indicator, it is now possible to perform Linear Regression.

### ***3.1.1 Regression of ISTAT data***

To perform the prediction, is used "R", a language or a dedicated environment for statistical computing and graphics. It is a GNU project and provides a wide variety of statistical models (linear and nonlinear modeling, classical statistical tests, time series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The R language provides Open Source option that it used in our project. One of the strengths of R is the ease with which you can produce great quality well-designed plots, and mathematical symbols and formulas where necessary. The cons are that there isn't paid great attention to the defaults for the less design choices in graphics, but the user retains full control [27].

R-Studio is an integrated development environment for R, with a console and an editor syntax, that supports the direct execution of code, Monitoring tools, history, debugging and the workspace management [26].

As said before, the objective is to extract a Y predictive value refers to the year 2018, conditioned by Xi variables identified as the years 2007-2017.

The result is the shape of the multi-variable linear regression:

Y = b 0 + b + b 1X 2X2 + ⋯ b KxK + ε

A complete example of R code used for linear regression is shown in Appendix A4, where the considered indicator is the average monthly expenditure.

To get the actual results of our regression you must use the command "summary (reg)", which will create an output like the one below, which displays all of the most important values of the model to assess the real goodness:

**summary(reg)**

***Coefficients:***

***Estimate Std. Error t value Pr(>|t|)***

***(Intercept) -****64.51987 67.42781 -0.957 0.36363*

***Istat\_Famiglie$Anno\_2007*** *0.36274 0.15502 2.340 0.04402 \**

***Istat\_Famiglie$Anno\_2008*** *0.07351 0.10744 0.684 0.51106*

***Istat\_Famiglie$Anno\_2009*** *0.23755 0.12889 1.843 0.09843* ***.***

***Istat\_Famiglie$Anno\_2010*** *-0.31284 0.20341 -1.538 0.15842*

***Istat\_Famiglie$Anno\_2011*** *-0.04140 0.12748 -0.325 0.75278*

***Istat\_Famiglie$Anno\_2012*** *-0.82588 0.32808 -2.517 0.03292 \**

***Istat\_Famiglie$Anno\_2013*** *1.65055 0.38821 4.252 0.00214 \*\**

***Istat\_Famiglie$Anno\_2014 -****0.88867 0.37681 -2.358 0.04271 \**

***Istat\_Famiglie$Anno\_2015*** *0.29580 0.34225 0.864 0.40987*

***Istat\_Famiglie$Anno\_2016*** *0.51256 0.18976 2.701 0.02435* ***\****

***Residuals:***

***Min 1Q Median 3Q Max***

*-49.282 -17.568 2.277 18.774 37.037*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

***Residual standard error:*** *37.12 on 9 degrees of freedom*

***Multiple R-squared:*** *0.9957*

***Adjusted R-squared:*** *0.9908*

***F-statistic:*** *206.5 on 10 and 9 DF*

***p-value:*** *2.185e-09*

The same procedure was carried out for the other indicators, with the exception of avarage income, where a further prediction of the 2017 data was needed to provide the data of 2018, as missing both.

How reflects the command output summary, the results are very acceptable. In fact, the estimate of the results of a linear regression model could and should contain:

* An adequate number of observations;
* The estimates of the values ​​of β are acceptable parameters;
* The values of the statistic test called T-test of Student are associated with each parameter in order to assess the significance; these statistics are often accompanied by an indication of the error associated standards, as well as the p-value that is considered acceptable only if less than 0.10, 0.05 or 0.01;
* Statistics adapted to evaluate the overall goodness of the model; these may be limited depending on the case in goodness measure of the fitting which R² and R² Adjustment for degrees of freedom. R² range is between 0 and 1: 0 when the model used does not explain at all the data; 1 when the model explains the data perfectly;
* Statistics of tests such as the F-test, namely the F statistic of Fisher are associated with the null hypothesis that all the elements of β, to verify the significance of the entire model. You want to check H0: β1 = 0,. . ., Βk = 0 against the alternative that at least one of the parameters is different from zero. Under the assumption that errors are N (0, σ²), the total deviance always admits the decomposition SST = SSE + SSR.

### ***3.1.2 ETL process of ISTAT data***

Having dealt with the problem of the completeness of the data provided by ISTAT data, resolved with the prediction, you can load the data obtained with a simple ETL process, as done earlier in chapter 2.

The principle is the same. After having copied the data obtained from the regression in the spreadsheet provided by ISTAT, we will go to create the metadata in Talend [24].

This, it will be placed in a job and loaded to L0 level Staging Area, without performing any operations. The data quality will be carried out by the TMAP [25] in the Layer L1, and then be connected to each other via the Surrogate Key to level L2. In the last two processes i will work directly on the database without affecting the original Excel spreadsheet.

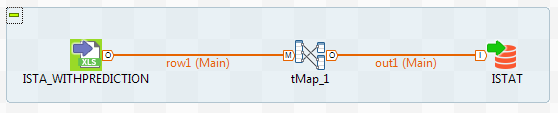


Figure 33: Job ISTAT with Prediction

Of course, the connections will be carried out through the dimension tables refers to the region, which in turn will be linked to the provinces which will be bound to the municipalities, creating a relational level hierarchy. Instead, with regard to the Town table, that already include attribute refer to the population, will be further connected to the table with the data related to tourism. Most important is to note how the table with its Istat data, will be considered as a fact table, and not as a simple dimension.

A simple visualization of the connections between the various tables is shown in the following figure:

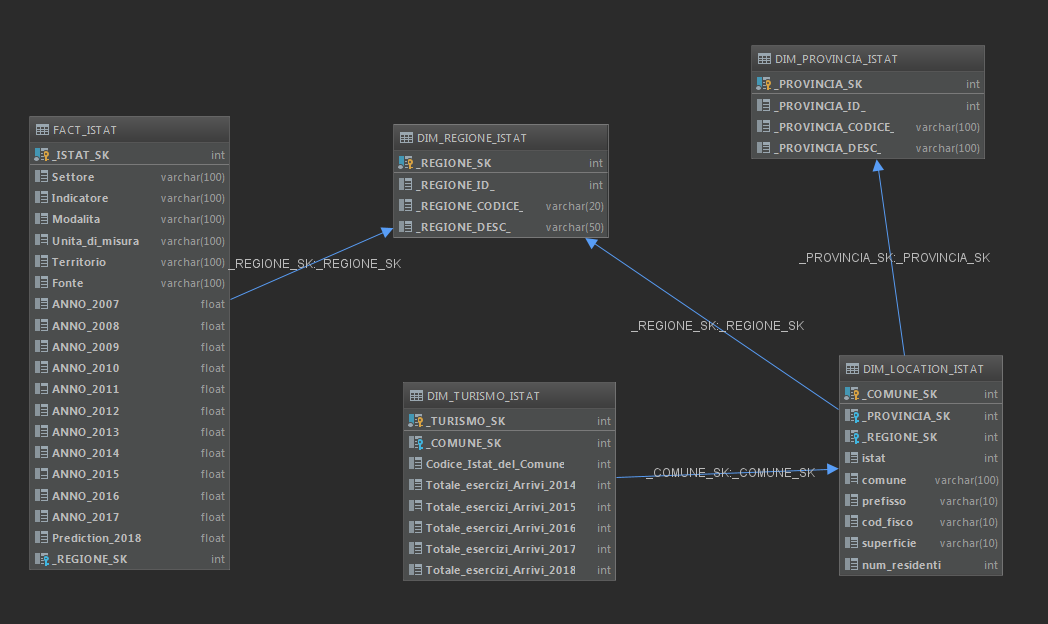


Figure 34: Star Schema ISTAT

### ***3.1.3 Best Geo-Locations***

To carry out a complete analysis and obtain a reliable result starting from ISTAT data [25], highlighted and processed via the data quality process, mainly three queries extremely connected to each other have been necessary, with the ultimate goal of finding among the thousands of cities, which are the most economically desirable to open a new store.

Initially, the first and the second queries will be on regional indicators taken from the website of ISTAT (average monthly household spending on non-food goods and services, average income, network operating railway, unemployment rate, Per Capita GDP), while the last, will be extended in depth at the municipal level, filtering for the best three regions obtained through the solution of previous database query.

The first query is a full view of the best and worst three regions for each indicator illustrating Moreover, the corresponding sector and the unit of measurement that characterizes it.

To implement the model, i will create a rank through a partition analytic function, which can be descending, if you treat of positive economic data, or increaing, whether it is data not very favorable to development as the unemployment rate.

In the second query it will instead create a summation factor of various rank grouped by region and sorted by rank descending. In this way, we will get as a result an order of development potential of Italian regions. For it, we will add up the corporate data for the historian of the number of shops closed and still open for each region, and the final value of which will be based our rankings, will take the name of Rank.

As shown in the table below, the three best regions to invest in an economic capital are Trentino Alto Adige, Valle D'Aosta and Friuli Venezia Giulia.

Table 7: Ranking of Regions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | Territory | #Close\_Shop | #Open\_Shop | Istat\_Rank | Final\_Rank |
| 1 | **Trentino Alto Adige** | 6 | 1 | 14 | 21 |
| 2 | **Valle d'Aosta** | 0 | 0 | 26 | 26 |
| 3 | **Friuli Venezia Giulia** | 4 | 1 | 37 | 42 |
| 4 | **Liguria** | 1 | 0 | 46 | 47 |
| 5 | **Piedmont** | 7 | 5 | 37 | 49 |
| 6 | **Emilia Romagna** | 19 | 7 | 23 | 49 |
| 7 | **Veneto** | 11 | 7 | 34 | 52 |
| 8 | **Umbria** | 2 | 1 | 51 | 54 |
| 9 | **Tuscany** | 12 | 5 | 38 | 55 |
| 10 | **Lombardy** | 14 | 17 | 32 | 63 |
| 11 | **Marche** | 3 | 1 | 59 | 63 |
| 12 | **Molise** | 1 | 0 | 65 | 66 |
| 13 | **Abruzzo** | 7 | 1 | 58 | 66 |
| 14 | **Basilicata** | 1 | 0 | 66 | 67 |
| 15 | **Sardinia** | 6 | 0 | 76 | 82 |
| 16 | **Lazio** | 23 | 12 | 47 | 82 |
| 17 | **Calabria** | 4 | 0 | 83 | 87 |
| 18 | **Puglia** | 7 | 4 | 80 | 91 |
| 19 | **Campania** | 11 | 3 | 87 | 101 |
| 20 | **Sicily** | 13 | 5 | 90 | 108 |

The last step to be carried out to find the ideal location to open a new store is to select the municipalities in the three regions previously chosen as the most attractive and go deep analyzing based on the number of tourists of the last year and the number of residents for each town.

The solution is the creation of a new indicator obtained through a sum of the number of residents and the tourists, to have an indicative number of potential clients.

The cities with the highest index will be the most interesting.

Table 8: Ranking of Town

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Town** | **Residents** | **Tourism 2018** | **Tot\_Possible\_Clients** |
| 1 | **Lignano Sabbiadoro** | 6616 | 691154 | 697770 |
| 2 | **Trieste** | 201148 | 414003 | 615151 |
| 3 | **Trento** | 115540 | 360388 | 475928 |
| 4 | **Riva Del Garda** | 16052 | 428198 | 444250 |
| 5 | **Bolzano** | 103891 | 337366 | 441257 |
| 6 | **Merano** | 37791 | 328265 | 366056 |
| 7 | **Grado** | 8434 | 302626 | 311060 |
| 8 | **Castelrotto** | 6540 | 301459 | 307999 |
| 9 | **Selva Di Val Gardena** | 2657 | 245928 | 248585 |
| 10 | **Badia** | 3396 | 228401 | 231797 |
| 11 | **Bressanone** | 20921 | 202598 | 223519 |
| 12 | **Pinzolo** | 3123 | 218297 | 221420 |
| 13 | **Courmayeur** | 2836 | 205460 | 208296 |
| 14 | **Nago-Torbole** | 2810 | 196793 | 199603 |
| 15 | **Canazei** | 1921 | 187328 | 189249 |

The result shows that Lignano Sabbiedoro, Trieste eTrento are the best three cities to set up a new shop in 2019.

## ***3.2 CLASSIFICATION: CART***

### The CART algorithm, as previously introduced in the State of the art, is a nonparametric procedure that builds a decision tree in order to label an attribute; In fact, the classification term refers to a process, given a collection of records called Training Set, try to build a model able to attribute a feature called Class attribute, based on the combination of other properties that characterize the specific population . Once you have the model, it can be used to predict the class of new records for instances where the class is unknown (Test Set).

### The important steps to be followed when you want a decision tree with the CART procedure are mainly two: adopt a criterion of the technical skill with which the nodes are divided from parent nodes to child nodes (split criterion) and establish a stopping rule of tree growth (stopping rule).

### To choose the split criterion is generally used a technique of Recursive Binary Splitting. For the stopping rule, you must pay attention to the type of decision tree that is considered. In fact, the trees with many nodes and split may lead to an overfitting of the data. This means that the model is difficult to interpret, because it becomes inaccurate for later forecasts and, so, needs the stopping rule. The methods to avoid this problem are to set a minimum number of training data to be used on each leaf node or set the maximum depth of the model, which refers to the length of the path longer from root node to leaf node.

### ***3.2.1 Training & Test Set***

The very first step to take when it comes to classification is to create an adequate training set for the labeling that we would expect. In the implemented project, we will use the CART process to define and predict which stores will continue to exercise and which stores will close in 2019, having as training set the shops that closed in 2018, give the data from 2017.

The table that will characterize our classification will be an aggregate table for the first three months of each year group by for shops, where you are going to analyze sales, cost of sales and operating margin, as regards the economic and financial aspect, and will also consider the number of receipts made during the period and the actual label representing the state Closed / Open of the store in the following year.

The aggregated data just listed will be our split criterion, except, of course, the store status which will be the element of our forecast.

To create the aggregate table, it had to refer to the dimension tables previously created in the ETL model (Shop and dates) and the fact table Sales, carrying out a detailed query to obtain the attributes mentioned above.

The totality of the queries is shown in Appendix A5.

### ***3.2.2 Classification and Forecast on the Causes Of Closing In Stores***

Create the reference tables, you switch to the creation of Classification and Regression Tree, more commonly known as CART. The idea, as explained above, is to take as a starting data stores from 2017 and check if they will be open throughout 2019, pulling out the table directly from the database using the commands shown in Appendix A5.

These data will be further divided into training and test set thanks to a random 80/20 partition on 100% of the analyzed elements, where you will create your models from training data set to test it on a later test data.

Once performed the test operation, will be verified its inequality distribution via the Gini index, where 0 represents perfect equality, while an index of 100 implies perfect inequality and in addition, the accuracy will be studied through the media.

**> summary(CART)**

Call:

rpart(formula = FlagOpen ~ ., data = TRAININGSet, method = "class",

control = rpart.control(minsplit = 5))

n= 78

CP nsplit rel error xerror xstd

1 0.250 0 1.00 1.0 0.1928198

2 0.100 1 0.75 0.8 0.1783112

3 0.025 3 0.55 0.9 0.1860521

4 0.010 8 0.40 1.0 0.1928198

Variable importance

GAIN COGS MARGIN nRECEIPT

28 26 25 21

**# view results**

> print(CART)

n= 78

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 78 20 OPEN (0.25641026 0.74358974)

2) MARGIN< 1235.842 5 0 CLOSE (1.00000000 0.00000000) \*

3) MARGIN>=1235.842 73 15 OPEN (0.20547945 0.79452055)

6) MARGIN< 17461.73 19 7 OPEN (0.36842105 0.63157895)

12) nRECEIPT>=113 8 2 CLOSE (0.75000000 0.25000000) \*

13) nRECEIPT< 113 11 1 OPEN (0.09090909 0.90909091) \*

7) MARGIN>=17461.73 54 8 OPEN (0.14814815 0.85185185)

14) nRECEIPT>=459 43 8 OPEN (0.18604651 0.81395349)

28) MARGIN< 39091.27 3 1 CLOSE (0.66666667 0.33333333) \*

29) MARGIN>=39091.27 40 6 OPEN (0.15000000 0.85000000)

58) COGS< 75358.5 6 2 OPEN (0.33333333 0.66666667) \*

59) COGS>=75358.5 34 4 OPEN (0.11764706 0.88235294)

118) COGS>=161920.7 18 4 OPEN (0.22222222 0.77777778)

236) nRECEIPT< 2059 2 0 CLOSE (1.00000000 0.00000000) \*

237) nRECEIPT>=2059 16 2 OPEN (0.12500000 0.87500000) \*

119) COGS< 161920.7 16 0 OPEN (0.00000000 1.00000000) \*

15) nRECEIPT< 459 11 0 OPEN (0.00000000 1.00000000) \*

**#Insert TEST set in CART**

> prevision <- predict(CART, TESTSet, type="class")

**#accuracy**

> Gini(prevision)

0.09428571

> mean (prevision == TESTSet$FlagOpen)

0.8571429

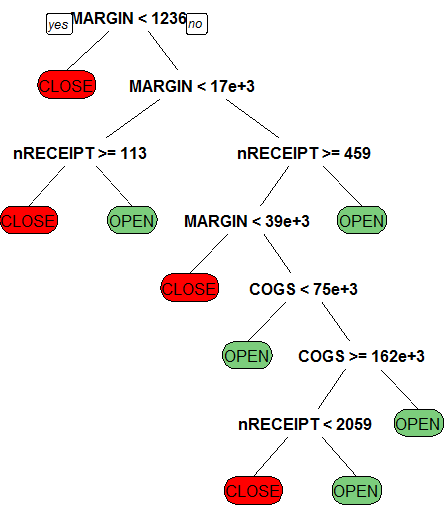


Figure 35: CART

In the obtained tree , are highlighted 4 factors: The split value in colored boxes, the final label, the probability of the input values for ​​the final label and, finally, the probability in which it is verified on the total of the initial data.

One way to avoid the problem of overfitting and bad accuracy is to set a minimum number of split for the training data to be used on each leaf node. For example, we may use a minimum of 5 to make a decision and ignore any leaf node that requires less than 5 levels.

The "rpart.control control command", do that and shows a minimum number of split to performed:

CART <- rpart (FlagOpen ~., Data = TRAININGSet,

method = "class", control = rpart.control (minsplit = 5))

Another way to set the pattern, which is take the longest path from the root node to the leaf node.

To improve the performance of the tree I can also use the Pruning technique. Simply, it removes the branches that make use of features that are of little importance. In this way, we reduce the complexity of the tree, and then we increase its predictive power (that in turn reduces the over-fitting).

The simplest of Pruning method starts from the leaves and removes any node with the most popular class in that leaf, stopping before reducing accuracy.

Once the model was developed, we can extract the Test Set represented by the open shops in 2019 directly from the database just created, resulting in the forecast through dedicated command: prevision2019 <- predict (CART, DATASET18, type = "class")

Now, the model will further evaluated using the Gini index [0.1705464] and accuracy [0.8701299], and finally, compared with the initial results, in order to understand the true efficiency. The obtained result is a prediction of 5 shops that are in a state of risk of closure in a total of 29.

Immagine che contiene testo

Descrizione generata automaticamente

Figure 36: Results of Prevision

# ***CHAPTER 4: DATA VISUALIZATION***

The data vizualization is a generic term describing any attempt to help people understand the meaning of the data analyzed by placing them in a visual context. Patterns, trends and correlations that may not be detected in the text-based data can be exposed and recognized more easily by using the report data visualization software such as Microsoft Power BI.

The reporting systems are developed in complex areas which have provided for a data warehouse solution. One of the aims of a DW process is precisely to structure a hardware-software information environment capable of responding to the needs of organizational scenario.

With the growth of the data accumulated available to organizations, the advantages of centralized document processing are revealed in the execution times of individual reports: the particular hardware configuration of the workstations on which resources are physically hosted on the system, allows optimization of requests and decreases the number of activity with respect to the situation in which individual users search information on the system.

The document produced is called report and is presented as a combination of tables and graphs that show important measures for each analyzed topic, disaggregated and de-structured according to the needs of the client.

These measures constitute a common basis for subsequent analyzes. Each report once processed and generated, it is validated by the departments and is distributed (and updated periodically) to customers that will exploit the potential.

A process of implementing a reporting system is generally composed of the following phases, which can be expanded or reduced as a consequence of the particular development environments and different macro-economic contexts of the organization's activities:

* Identifying information needs and visualization;
* Identification of the information environment and sources;
* Identification of the hardware / software system configuration;
* integration of information resources;
* Preparation of the report;
* Validation of the report;
* System testing phase;
* Operating phase.

These phases are not to be construed as necessarily consecutive, because some may also take place concurrently.

## ***4.1 MICROSOFT POWER BI***

The BI of Microsoft Corporation is a complete and integrated suite that helps to reduce the complexity of interaction and organization of information and to obtain competitive advantages for the company through better decisions strategies.

Microsoft provides a number of data warehouse tools and data analysis to drive the enable users to access, understand, analyze, collaborate and act on information when they want and wherever they are. It is used to get a deeper insight for better decisions making and ultimately, to help organizations adopt agile decisions to achieve the goals.

In thesis, I will use Microsoft Power BI, a suite of business analytics tools to analyze data and share information [31].

Power BI dashboards provide a 360 degree view to business users with the most important metrics in one place, updated in real time and available on all of their devices. With one click, users can explore the data behind their dashboard using intuitive tools that facilitate the search for answers. Creating a dashboard is very simple, thanks to the hundreds connections with leading enterprise applications and pre-built templates to help you put to work immediately. Also, you can access to your data and reports anywhere using the Power BI mobile app, that update automatically after any change to the data.

To facilitate the use of the application you were created variables through the use of the DAX, Data Analysis Expressions, that indicates the formula language used in BI applications, even in the background. All the DAX formulas created for data visualization in Power BI can be found in Appendix A7.

## ***4.2 TOOLS USED AND RESULTS OBTAINED THROUGH THE DATA VISUALIZATION***

Power BI can be defined simply as a display system divided into blocks that can expand their field of definition dynamically, from a general to a detailed analysis with a simple click, with the aim of creating elaborate and complex reports tailored to the needs of customer.

The blocks that characterize the use are mainly 4:

* Views;
* Report;
* Dashboard;
* Dataset.

### ***4.2.1 Views***

When creating or editing a Power BI reports, you can use several types of visual objects. Icons for these visual objects are displayed in the Views pane.

Developers create custom visual objects through SDK. These visual objects enable business users to view data in a way that best suits their business. Users can import files of custom visual objects in reports and use them like any other visual object of Power BI, assuming a leading position. Visual objects can also be filtered, highlighted, modified, shared freely.

The custom visual objects are distributed in three ways:

* Custom file visual objects;
* Visual object organization;
* Marketplace.

In some organizations, custom visual objects are even more important, as might be necessary to communicate data and in-depth information or simply, to bring out to individuals. Therefore, these organizations need to develop custom visual objects, share them in the cloud and make sure they are handled properly.

In summary, the term views in Power BI can be defined as a visual representation of the data. This representation can be as a graph, a map or any other tool that can show your data.

The image below shows some of the views present in Power BI, providing a generic overview of the first quarter 2019:

• The left view (Multiline card), was created to relate a measurement and an attribute of the product table. As can be seen, in fact, for every manufacturing country of origin it has been associated with an equivalent number of total units sold in the first period.

• At the center, we find a quick but effective corresponding value of gain, discount, cost of goods sold and margin until 2019. Each of these values is accompanied by a weekly detailed graph that shows the comparison with the first quarter of the previous year. This chart is very useful to understand where a firm gained and where lost money, and comparing them with other company data, you can also get to the cause of mutative phenomenon.

• On the right, finally, there are two histograms, one vertical and one horizontal, which show, respectively, the total turnover filtered for the period in analysis for a store channel and for the flag weekend, obtained by means of creation of a new Power BI function with

language DAX. The last figure is the percentage of the daily flow of customers at points of sale.

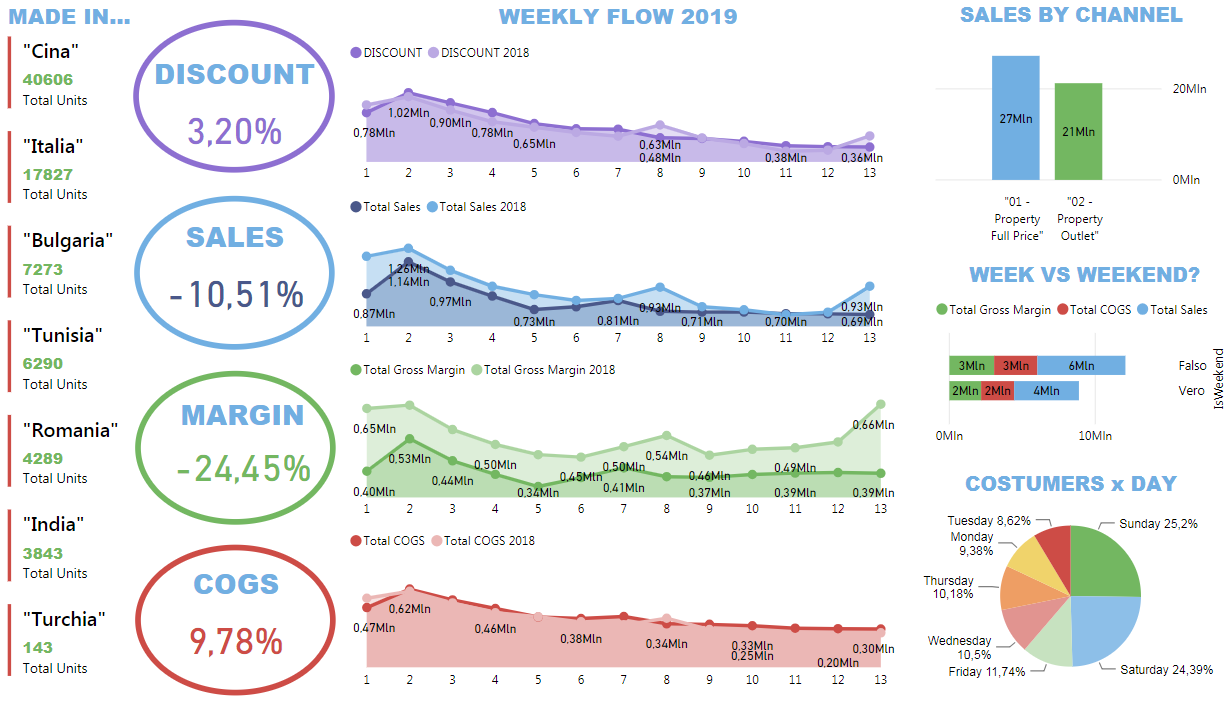


Figure 37: Visualization of 1Q 2019

From this kind of views, the results are clear and hard to confuse. We can say, for example that in this quarter the general performance of the company is below expectations, that most of the products sold are made in China, the sales channel with highest grossing is the Property Full Price and, finally, that the best flow of people occurs on the weekend.

With a small demonstration, you can understand the enormous potential that can bring data visualization, and in general the Business Analytics, for managerial decisions of a company, to improve, for example, over the years in the distribution of goods and of turnover.

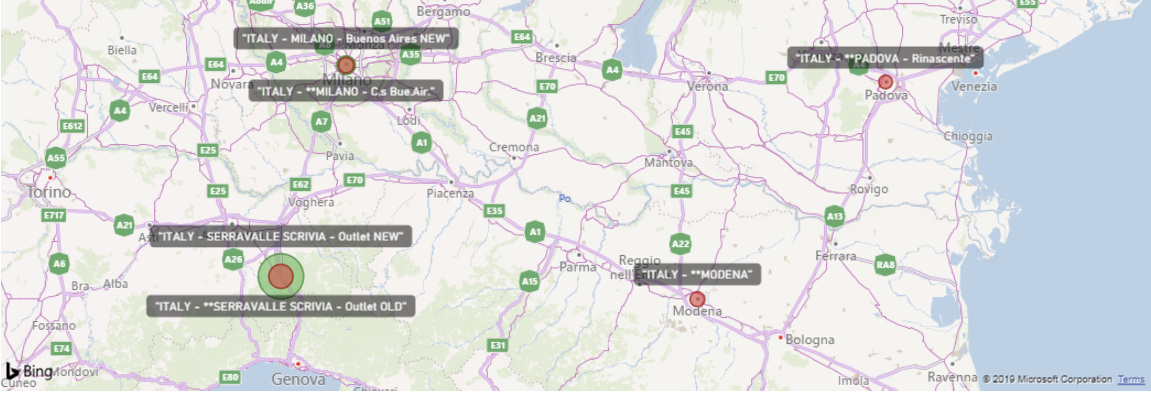
### ***4.2.4 Report***

### 

A report is a compilation of views displayed together on one or more pages. Reports help you organize your views in a way that tells the story of your data in the way you want. For example, if you want to show the sales of your company within the various sales points of your country, you can have a relationship consists of a number of charts (pie, line or bar)and maps. In the following example are shown the results obtained thanks to an analysis of the closed stores in Italy from 2017 to the present day of the client analyzed in the thesis project , identifying the cause of the closure of them, classified as a closure for a bad profit margin, or a closure for a covered market in the years or the birth of a neighbor new store. As starting data, we were used the data implemented in the ETL phase of the data warehouse, previously explained in Chapter 2, and the data obtained from CART algorithm illustrated in chapter 3.

The result is visible through the underlying map that encloses some examples for each type of closure. The shop of Serravalle show a closure caused by new opening, while the shops of Padova and Modena caused by a margin not adequate.

The size of the dots in the figure is proportionate to the total turnover for each point of sale.



*Figure 38: Map of Closing Stores*

In connection with the map, an explanatory table was created which highlights the closure type and description for each stores.

Immagine che contiene screenshot

Descrizione generata automaticamente

*Figure 39: Table Of Closing Stores*

### ***4.2.3 Dashboard***

A dashboard is a collection of views on a single page, which you can share with others. Although visually similar to a report, a dashboard must fit on a single page and can be shared with other users who will be able to interact with the data presented in it. By creating and sharing a dashboard for a sales manager, for example, he or she should be able to interact with it and see new information other than that which is clearly visible on the dashboard to start, according to the data.

The following images show an example of a dashboard in Power of BI. The underlying figure, includes a general analysis of sales channels associated with the relative gains of the products, and the made in. It will be the default view for the customer.

The first graph (BarChart) includes the view of the total gain of 2019 for each shop channel, divided into Full Price Property and Property Outlet. The second chart is the Pareto 80/20 representation, where evidence that most of the sales derived from the bags. The last graph, instead, shows the percentage of the made in of the products.

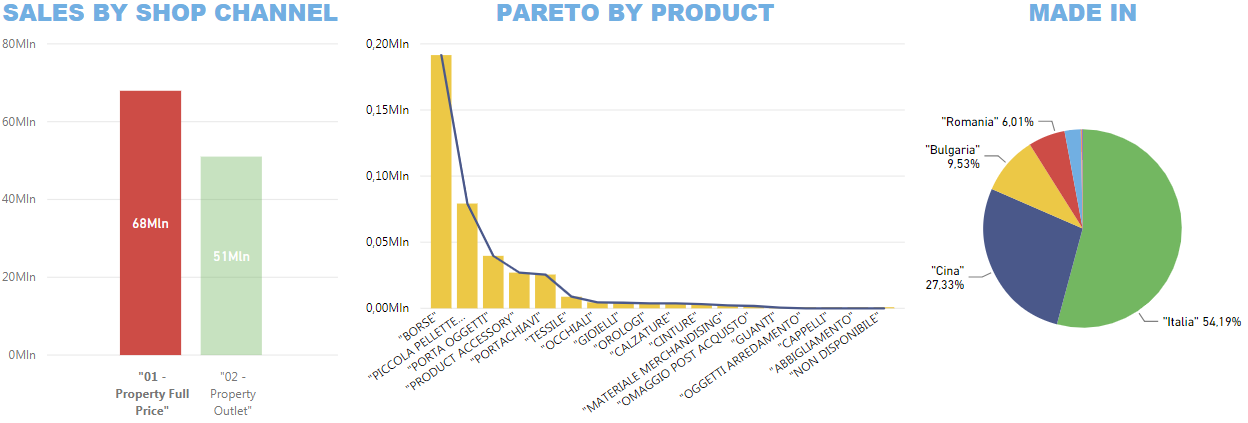
Immagine che contiene mappa, testo

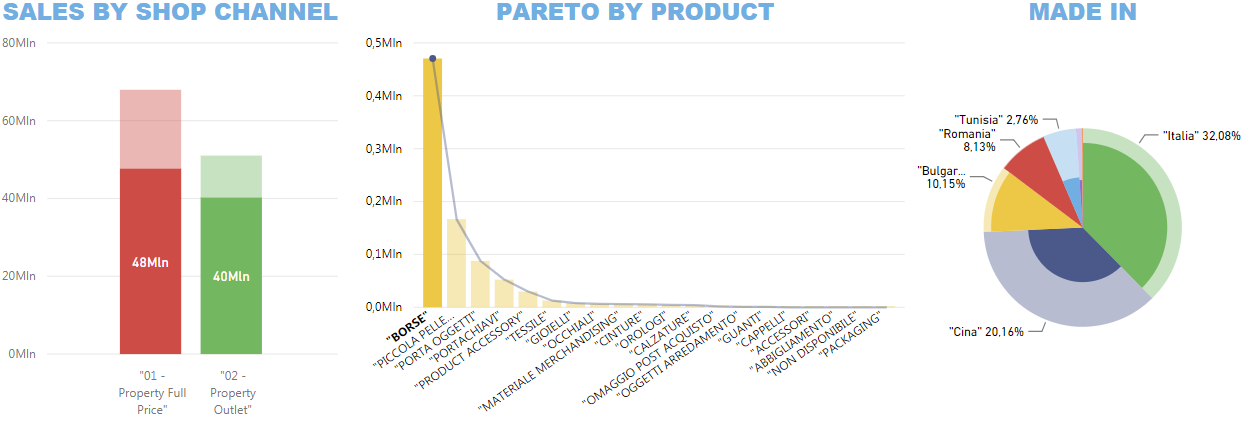
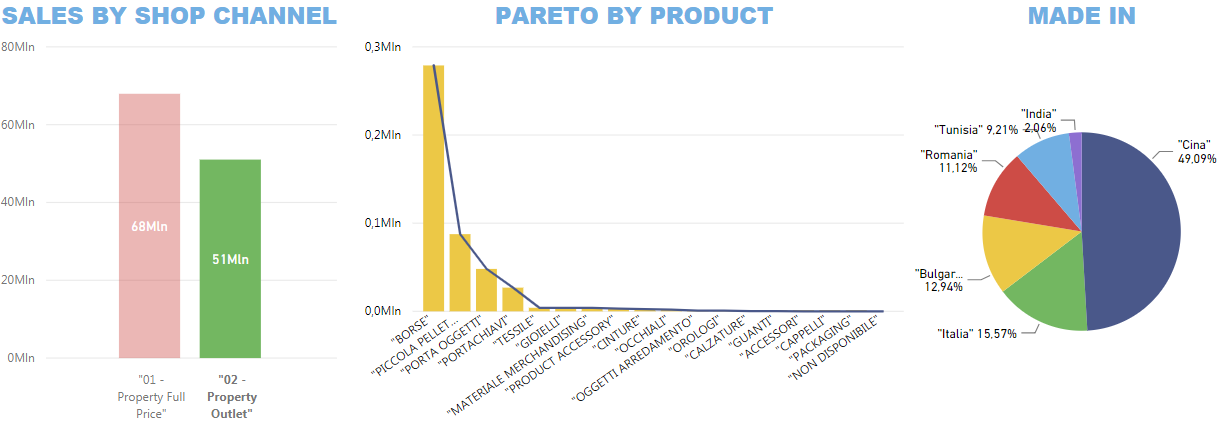
Descrizione generata automaticamente

Figure 40: Default Dashboard

Through the use of filters or simple navigation you can go from a very general analysis to a detail analysis in every single aspect. Following, will be explained briefly three different views, each resulting from the default dashboard.

The initial image and the next explain the variations of products sold and made in thanks to a selection of a specific store channel that you want to consider. The last representation uses the same idea, but the selection is made for the bags. It is very important observe how the graphics interact with each other leading to a rapid and effective analysis, chosen according to the need of the customer.





*Figure 41: Drill-Trought Of A Dashboard*

### ***4.2.4 Dataset***

A set of data, commonly called Dataset, are the data behind a chart or a map in your report. For example, if you have a chart that shows the sold in each month of the year, the data used to produce the graph are known as data records or dataset. It is important to notice, that does not necessarily datasets must be from a single source. Sometimes, it is a filtered collection of combined data from different sources with the goal of producing a unique collection that can be used in Power BI to show a feature that can be useful in the business decisions of a company. This is possible thanks to an impressive number of connectors included in Power BI.

The following image shows a set of applications which considers all economic and financial values of the customer, regarding sales, cost of sales and gross margin of 2017 and 2018 grouped for each month.

Consequently, will be display also the variation in terms of percentage value of 2018 with respect to 2017. The green and the red bar indicate, respectively, conditional formatting values of gain or loss. The last row of the table shows the totals. This Dataset can be compared to a partial income statement of the company.

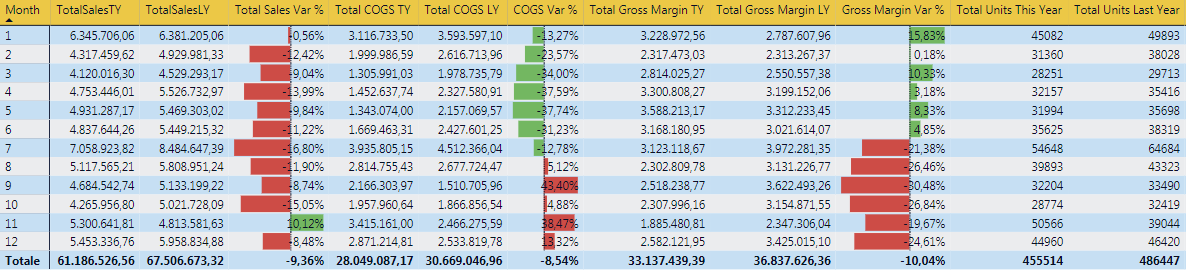


Figure 42: Income Statement Dataset

# ***CONCLUSIONS***

## ***RESULTS***

The purpose of the thesis project was to implement a data mart with clean and dedicated data concerned the sales of a fashion company, on which were carried out some helpful business analysis for future decisions strategy.

In particular:

1. Creating a ETL best practices data mart to get the best possible performance in the execution of queries and data ingestion made with an entire process of data quality to perform some relevant and efficient business analysis;
2. Data Mining and Machine Learning softwares that using Artificial Intelligence to independently implement different algorithms that can combining data warehouse and open data for studying the causes of store closures and providing the ideal location to open new ones;
3. Creating a dashboard Power BI to have a data visualization helpful to understand the meaning of the data analyzed by placing them in a visual context.

The first goal was definitely achieved because the data mart is currently operating. The performance goal has instead been achieved only in part. Certainly, with respect to the multidimensional analyzes conducted on the operational data, previously performed by the company, we faced with a substantial improvement. The process is optimal, but the software used, being open source, and the fact of working on a local server, slowed much the level of the final performance of the ETL process.

The machine learning algorithms and the prediction have reached a good level of analysis. Through the CART decision tree of paragraph 3.2, we reached a demonstration of the major causes of the client's stores closures, obtaining as optimal value Gini index 0.17, representing the inequality of the distribution of data, and an accuracy of 0.87. Both are acceptable values, because, the Gini index and the ideal accuracy are respectively equal to 0 and 1. The result obtained has identified 5 out of 29 shops risk to closure by the end of 2019.

Instead, as regards the prediction of data ISTAT 2018, as seen above in section 3.1.1, the analysis goodness values are excellent. Through the use of the query and the prediction of open data, it could be concluded that the best location to open a new store, are Lignano Sabbiedoro, Trieste and Trento.

In general, we can conclude that the thesis project achieved effective results for future business strategies, all displayed in smart and intuitive way through the use of dashboards created in Power BI.

## ***FUTURE DEVELOPMENTS: REAL-TIME BUSINESS INTELLIGENCE***

While the real-time analytics and big data are both trendy, analysis of big data in real time, which is the combination of them, is the future.

The real-time is often confused with the instantaneously. In fact, the engine in real-time processing is not always able to import the streaming data but can be designed to extract new data has just been placed in the source file. The time between these queries is highly dependent on business needs and can range from milliseconds to hours. For example, the analytical system of a bank would allow several minutes to evaluate the creditworthiness of an applicant and the dynamic price a retailer can take up to an hour to upgrade. However, all of these examples are considered in real time.

Unlike traditional models, that examine the historical data for patterns, real-time analysis focuses on understanding the information created to help make faster and better decisions [33].

The real-time business intelligence is the use of analytical and other data processing tools to enable companies to access to the data and the most recent and relevant views. To successfully provide better data it using a combination of server-less analysis (where data is transmitted directly to a dashboard or display) and a data warehouse, enabling the dashboard to show data historical and real-time in a complementary way.

For organizations that produce gigabytes or terabytes of data, many of them lose their relevance information once they are stored. The information on inventory levels, customer needs, ongoing services, and more, can be incredibly useful, but even more so if analyzed as they are generated.

The real-time analysis and BI also allow users to do custom queries, and use the available data, including the ability to perform ad-hoc analysis on existing data or create specific views for new flows, helping to better understand the trends and create more accurate predictive models.

There are several areas where the use of BI can optimize an organization:

* *Customer Relationship Management*: The relationship management suite with clients (CRM) can use real-time data to provide better service to consumers. This includes a better involvement of the services and conversations to the known the preferences of each consumers. A significant example is the Disney company has launched its innovative MyMagicPlus program, after years of testing to Disney World. Now, every guest gets their own MagicBand bracelet, which serves as the key identification, credit card and pass. Customers simply pass it on the band sensors located around the park to entry to attractions or to pay for souvenirs. In this way, Disney giving a large amount of data on where its guests, what they are doing and what they might need.
* *Location Analytics*: The geographic data and locations often hide a lot of useful information and the extraction of these can help a company to optimize their business processes and increase profits with a better resource management. For example, sensors, such as GPS tracking systems linked to the vehicles, periodically emit data on its location. The position analysis can transform these information to detect congestion, delays prediction, detect vehicles inactive, suggest alternative routes, breaking rules and transportation guidelines (speed control), identifying the routes more profitable.
* *Service transformation:* The companies being able to collect data in real time by machinery and production lines and see how they behave, improving both efficiency and productivity and solving maintenance problems before they become an emergency. A case of transformation service doing by technology is to Rolls Royce (aero and marine engines) that has always increased the use of related products, big data and analytics in a systematic and advanced in all the three areas of activity: Product Design, manufacturing and after-sales processes. For over a decade, the company has changed its business model, developing maintenance plans and innovative services that can connect directly to the performance of its products. The increasing use of these technologies has led Rolls Royce to the launch of Intelligent Insights and to create a set of data collected regarding the operation of engines of customer aircraft. The data are transferred to the cloud and analyzed automatically using the data mining algorithms, with the aim of creating automatically connections from data captured by different sources, allowing preventive and predictive actions [35].

Today, the real-time business intelligence is becoming an increasingly important aspect of the decision-making process of the organizations, implementing the right solution, understanding and collecting the correct data, creating a solid infrastructure and allowing your team to use it, creating a real competitive advantage over competitors.

In conclusion, with the development of new technologies and the passing of the years, we will find ourselves faced with an evolutionary scenario that will no longer consider the use of big data a competitive advantage but will consider them almost indispensable for any company throughout the world.