# ***CHAPTER 1: STATE OF THE ART***

## ***1.1 BUSINESS INTELLIGENCE***

The term Business Intelligence (BI) refers to a series of business processes that revolve around the data, with the collection, processing and the analysis, whose purpose is to produce information to the strategic and tactical management service, which finds support analytical, historical to forecast a efficient Data Driven Strategy. The BI was also placed in the operating subset, as it is playing an increasingly important role in the normal daily activities of the companies.

In modern business, whose main purpose is to become market leaders, companies are daily compare with realities different from their own. This is accomplished through the analysis of the behavior of competitors belonging to the same sector and studying the market in which they are located.

The BI adoption by firms allows a deeper understanding not only of themselves but also of the market.

In the current period, the "change" is on the agenda, thus being able to read in advance trends of markets can give some huge competitive factors with respect to competitors.

Given the high volume of data generated every day, becomes necessary to find a method that:

* It collects and processes data at high speed (mores often it comes to real-time processes);
* It provides a cleaning service of the data, eliminating dirty data, redundant or incorrect via the ETL processes "Extraction, Transformation & Loading" which pick up the data from the input systems (ERP, Excel spreadsheets, etc.) and carry them in a data warehouse through data quality processes.

This process will be explained in a specific manner in section 2 but in summary, It defines a consolidated and stable system storage for certificates data (data warehouse) and it turn information into a source of knowledge through business analysis on the data, determining new KPI.

The Big Data originate from different sources, both internal and external, they are often include different formats and reside in multiple positions of a legacy system or in other applications. The data can be structured (data stored in relational database, arranged in patterns and rigid tables), unstructured (stored data without any schema as free-form text such as articles and e-mail parts, audio without tags, images and video ) or semi-structured (data that contain features of both of those structured that those unstructured; an example is represented by the compiled files with XML syntax for which there are there are no structural limits the insertion of data, but the information is organized according to logical structured).

After that, the data must merged.

The next step is choose the platform and the technology to be used for big data analytics applications including queries, reports, OLAP system, data mining and visualization, including in all these applications [22].

A central role is played by big data analytics, and by the business intelligence based technologies such as:

* *CRM & Customer Analytics*: Solutions and technologies that collect, organize and synthesize customer data to help organizations solve business problems related through tools, dashboards, portals and other methods in the areas of Marketing, Sales and Customer Service; Consumers are segmented into groups based on the adopted behaviors, actions to implement customized marketing and general trends;
* *Predictive Analytics*: Advanced Analytics that implement techniques such as regression analysis, predictive models and statistics to analyze data and contents, and answer questions like "What will happen" or "What will most likely happen?";
* *Social Analytics*: Tools that automatically extract, analyze and summarize the content generated by online users;
* *Text Analytics*: Process of extracting information from texts, used for including the summary, finding key content in a large set of data, sentiment analysis or to determine what drove a particular comment of a person, so, for an explanatory purpose;
* *Web Analytics*: analytical applications used to understand and to improve the online consumer experience, the acquisition of users and the optimization of digital marketing and advertising campaigns. These offer reporting, segmentation, advertising management and integration with other data sources and processes;
* *Prescriptive analytics*: Use optimization technology to solve complex decisions with a very large number of decision variables, constraints and compromises for providing optimal actions to achieve the business objectives.

## ***1.2 DATAWAREHOUSE***

The Data Warehouse (DWH) is the main business intelligence support tool. They allow you to collect integrated, consistent and certificates data related to all business processes of a company from the operational sources. These data are suitably processed through ETL procedures and controlled through the data quality system.

Data quality is a critical requirement for the entire information system, because, if the data are dirty, can not only cause a worsening of business performance, but can also lead to take inappropriate decisions, resulting in additional costs and lost opportunities.

The goal of a data warehouse is therefore to support the "Knowledge Worker" (officer, director, manager, analyst) helping him to analyze the data aimed at implementing decision-making and improvement of information assets, to provide a single point of access to all company's data made in a consistent and reliable way through the ETL processes. The data warehouse also ensures a complete historical depth of the data, thus allowing temporal analysis.

A DWH must be carefully designed to have an efficient and effective manage of the Big Data characteristics.

Immagine che contiene screenshot

Descrizione generata con affidabilità molto elevata

Figure 2: Data warehouse

The Datawarehouse are made as the main tool for the Decision Support System (DSS), that is a system capable of providing clear information to users so they can analyze in detail a situation and take the appropriate decisions on actions to be taken in a easily and quickly way [12].

The DSS relies on data from one or more databases, often organized in different structures with non-homogeneous data.

A system of this type must support the analysis and control of management routines, the research of the causes of a problem (focused search) and complex managerial activities (decision making), besides an easy using to a user with a reduction on time and an improvement towards new technologies (especially in cases in which cannot perceive the benefits in a short time).

Let us describe in detail the features:

* *Subject oriented*: The data warehouse is organized for relevant subjects such as, for example, products, customers, suppliers and the time period, in order to provide all information pertaining to a specific area;
* *Integrated*: The data warehouse must be able to integrate seamlessly with the multitude of standards used in different applications. The data must be re-encoded, in order to be homogeneous from the semantic point of view, and must use the same units of measurement;
* *Variable time*: Unlike the operational data, those of a data warehouse have a very broad time horizon (even 5-10 years), making them reusable in different time instants;
* *Non-volatile*: The operational data is updated continuously; in the data warehouse data are loaded initially with integral processes and subsequently updated with partial loads; data, once loaded, are not modified and retain their integrity over time.

It is possible that a data warehouse is divided into different data marts, each specific to a single business process among those inside the company (orders, sales, customers, marketing, etc.). In Chapter 2 we will see a data mart related to the sales of a fashion retailer.

### ***1.2.1 Data Warehouse Architecture***

In the design phase it is essential to determine which types of architecture adopted. Clearly, a DWH must be constructed in accordance with modern principles [10] and the patterns described in this paragraph are still of the bases:

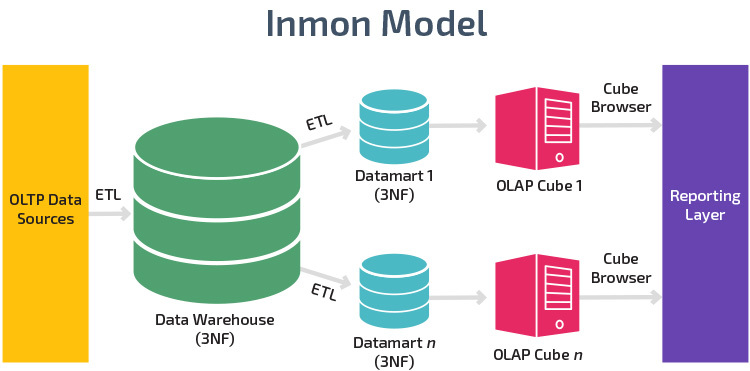
***Model Inmon - Corporate Information Factory****:* The Datawarehouse are constructed in their entirety from the beginning as a single monolithic block; you can not see how the composition of the DM. It adopted a top-down view.

Figure 3: Model Inmon

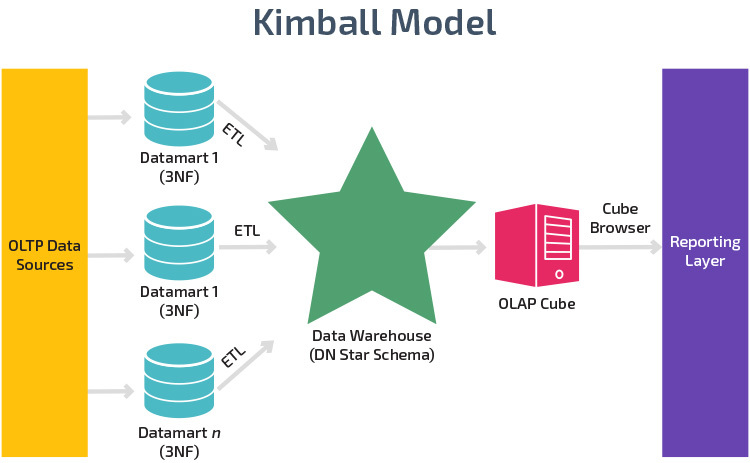
***Model Kimball - Dimensional Model****:* Adopts a bottom-up approach in which the data warehouse born from the union of the various data marts that each refer to a specific business area.

Figure 4: Model of Kimball

It has been shown that the Inmon and Kimball approaches work to successfully deliver data warehouses. But there are excellent organizations where there is a combination of both, in a hybrid model: the data warehouse is created using the Inmon model and the business data mart processes are created using the star schema has been implemented for report creation. We can not generalize and say that one approach is better than the other; both have their advantages and disadvantages, and both work well in different scenarios. For any approach to be successful, it must be carefully studied and discussed in detail and designed to meet the reporting needs of the organization's BI and should also integrate with the organization's culture.

So, The architect has to select an approach for data warehouse based on several factors;

### ***1.2.2 Extraction, Transformation and Loading (ETL)***

The role of ETL tools is to feed a single, detailed and comprehensive high quality data source that it can in its power once the data warehouse. The operations to be carried out they are often referred to the term reconciliation that, during the data warehouse feeding process is carried out on two occasions: when the DW is populated for the first time and when it is periodically updated. The reconciliation consists of four distinct said processes respectively:

* Extraction or Capture;
* Cleaning or Scrubbing;
* Transformation;
* Loading.

In general, the boundary between cleaning and transformation is quite hazy therefore, for simplicity, we assume that the cleaning operation is essentially aimed at the correction of the data values, while the transformation is concerned more properly of their format.

#### **1.2.2.1 Extraction**

The Data Integration consists of extraction and cleanup.

During the first step, the relevant data are extracted from different sources and this operation can be of type:

* *Static:* It is carried out when the DW must be populated for the first time and consists conceptually in a copy of operational data;
* *Incremental:* It is used for the periodic updating of the DW and captures only the changes that have occurred in the data from the last extraction. The basic idea is to use the changes registered at the data level to update the data warehouse. The benefits derivable are the very small volume of data involved from time to time in the operation than extracting static, and that most of the data in the data warehouse remains unchanged and only the data that have changed are analyzed. The Technique that are used is the CDC (Change Data Capture) that allow you to monitor data sources with the goal of identifying the changes at the data level. This technique is particularly important for the maintenance of the data warehouse thanks to the propagation of the changes detected at the level of the source.

The Cleanup, however, is the phase that is concerned to improve the quality of data, going to eliminate "dirty" data due to duplication, inconsistencies, missing data, incorrect values etc.

The main functions of data cleansing found in ETL tools are the correction and homogenization thank to the using of appropriate dictionaries to correct specific errors, recognize synonyms, and cleaning based on regular expressions, which apply their domain rules to establish the correct matches between values.

#### **1.2.2.2 Transformation**

It is the central phase of the reconciliation process and has the aim to convert the data from the operational source format to the data warehouse format. Among the features of this level for feeding the reconciled data level you have:

* *Converting and Normalizing*: Operating both in storage size level and at the level of measurement units in order to standardize the data;
* *Matching*: Establishing correspondences between equivalent fields in different sources;
* *Selection*: Reducing, if necessary, the number of fields and records with respect to the sources.

In the phase of feeding data warehouse we have two substantial differences: The normalization is replaced by the denormalization and the aggregation are introduced, that achieves the appropriate summary of the data.

#### **1.2.2.3 Loading**

In this phase the data are uploading to the data warehouse through two alternative methods:

* *Refresh:* the data is rewritten integrally replacing the previous ones. In general, this technique is used only during the initial phase of populating;
* *Update:* They are added to the data warehouse only the changes on the data without overwriting the entire boundary. This technique is used in combination to the incremental extraction for regular updates.

One way to reduce the load time is to parallelize the ETL process. This can occur in two ways: more steps performed in parallel or a single passage running in parallel.

* *Multiple Load Steps*. The ETL workflow is divided into several independent papers loading together. The main objective is the creation of independent jobs to create a process much safer to handle errors;
* *Pipeline*. The database can identify certain tasks that can execute in parallel. For example, the creation of an index may be generally parallel through all available process on the machine.

#### **1.2.2.4 Possible ETL Process Problems**

When the ETL system operating, some faults can occur for many reasons.

The common causes of ETL failures include:

* Network errors;
* Database errors;
* Disk errors;
* Memory errors;
* Errors in data quality;
* System updates without notice.

To protect yourself from these failures, you need a solid backup and a restart and restore system. You must plan for fatal errors when loading it happen. The system should anticipate this and provide the recovery of the capabilities, stopping and restarting the crash.

For example, for a loading process should engage relatively small set of records every time to keep track of what has been committed. The size of the set should be adjusted based on the size of the transactions and performance implications of different DBMS.

The recovery and restart system is used, of course, to take a job that is entered in error, stopped or reported it to recover it through the entire backup or asimple restart. This system is significantly dependent on the backup system capacity. When an error occurs, the instinctive initial reaction is groped to save whatever has been processed and restart the process from that point. This requires a solid and reliable ETL tool checkpoint functionality to restart the job in exactly the right point. In many cases, it might be best to get out of all the rows that have been uploaded as part of the process and restart from the beginning.

For this reason, it is recommended to design fact tables with surrogate key columns relative to the dimension tables linked. This surrogate key is a simple integer that is assigned in sequence as the rows created in the dimension tables. With the linkage of the surrogate keys with the fact table, you can easily resume a load that is stopped.

The more an ETL process is long, the more you have to be aware of the vulnerability because of an error. The design of a modular system consisting of ETL processes efficient and resistant to abnormal and unforeseen interruptions arrests, may reduce the risk of a failures resulting in remarkable recovery. Careful consideration must be given to entering physical values ​​by writing the data to the disk, when to insert the accurately prepared recovery points and when to choose the specific date / time of loading of the sequential tables with an appropriate restart logic, to not congest the system of customer.

## ***1.3 OLTP vs OLAP***

**On-Line Transaction Processing (OLTP)**

At the database level, the online transaction processing queries are based on quick and effective multi-access. The main operations performed are INSERT, DELETE, and UPDATE as they modify the data directly. The data are constantly updated and, therefore, require an efficient support for rewrites. A key feature of these systems is the standardization, which provides a quick and effective way to carry out writing into the database.

**On-Line Analytical Processing (OLAP)**

The On-Line Analytical Processing is a set of technical tools for accelerated analysis and large amounts of data, with the ability to study the problem in different points of view. These systems are very useful for the production of synthetic information, which will support and improve business decision-making. Examples of OLAP tools are the data warehouses and the multidimensional cubes.

The major differences between the two systems are shown in this table [10]:

Table 1: VS OLTP OLAP

|  |  |  |
| --- | --- | --- |
|  | **OLTP** | **OLAP** |
| **Purposes** | Support in operations | Support in decision-making |
| **How to Use** | Prompted by processes | Ad hoc query |
| **The amount of data per elementary operation** | Low: hundreds of records for each query | High: millions of records for each query |
| **Quality** | In terms of integrity | In terms of consistency |
| **Orientation** | To process / application | By Subject |
| **Refresh rate** | Continue, through actions | Sporadic, through explicit functions |
| **Time coverage** | Current data | Historical |
| **Optimization** | To read and write accesses on a data portion | For read-only access to the entire database |

According to the data storage, you will have several OLAP architectures, each with their own pros and cons [10]:

* *Relational OLAP (ROLAP):*the data is stored in a relational database as a support to the OLAP engine. The analysis is translated into queries, returning results in a multidimensional form;
* *Multidimensional OLAP (MOLAP):*it has the database and the multidimensional engine. For the Drill-Down operations it is not the ideal system, as it can generate errors;
* *Hybryd OLAP (HOLAP):*It combines the advantages of the two previous systems. In particular, pre-aggregates data in multidimensional systems for efficient and fast analysis, and are useful in a relational database in case of Drill-Down;
* *Desktop OLAP (Dolap):*the data are loaded into a client system and are calculated by the local engine.

## ***1.4 BIG DATA***

The term Big Data refering to a collection of extensive data in terms of volume, velocity and variety that requires specific technologies and analytical methods for the extraction of value and knowledge. The term is used to refer to the ability to analyze, extrapolate or put in relation an enormous amount of heterogeneous, structured and unstructured data, in order to discover the links between different phenomena and to predict future ones.

The size varies in different sectors, from dozens of terabytes to hundreds of petabytes (1000 terabytes), also according to the different available software tools. This size, definitely, will increase over time due to continuous advancements of technologies.

In this definition, emerge the so-called 5V that characterize the Big Data, which is the volume, speed, variety, authenticity and value [2].

The volume makes reference to the huge mass of data generated through numerous channels.

The rate refers to the rate at which data is acquired and used thanks to faster and more frequent transactions: companies not only collect data faster, but they try to exploit them as soon as possible, often in real-time.

The veracity regards the data quality and their level of security, where security is a very important challenge. To take advantage of Big Data you must know how to act in order to extract the value and increase the productivity and competitiveness of company for creating an economic surplus for consumers.

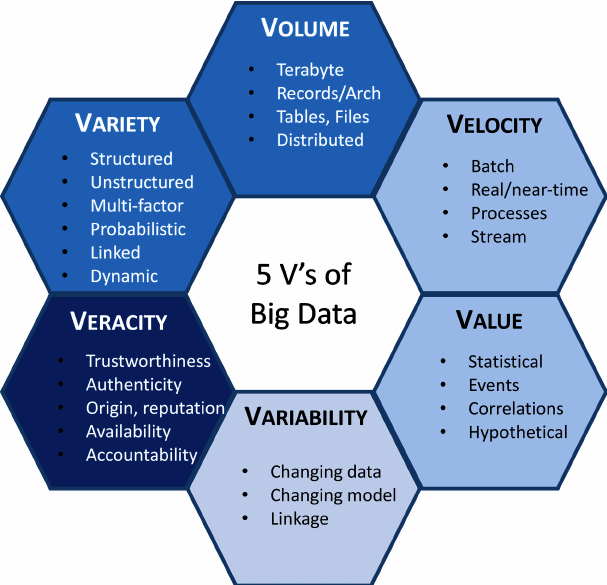


Figure 5: The 5V Big Data

The variety is related to the different types of data available from a growing number of data sources, both structured and unstructured; in particular it is possible to identify four categories of information that constitutes Big Data:

* Data generated from smartphones and other mobile device, including RFID data (radio-frequency identification) devices that track the product, and data from monitoring devices such as counters for water monitoring or gas;
* Sales and pricing data, generated by loyalty cards and promotional events;
* Computer data log, such as click streams from websites;
* Information from social media such as Twitter and Facebook or from YouTube and similar sites.

The ability to store, aggregate the data and use the results to carry out deep business analysis, is improving thanks to the availability of software tools and the increasingly of sophisticated techniques combined with an increasing computing power. We are seeing a huge change in the ability to create, communicate, share and access to the data due to the increase in the number of people and tools connected to digital networks.

### ***1.4.1 Barriers On the use and extremely beneficial Of Big Data***

The Big Data represents a great opportunity for companies and for national economies thanks to several significant benefits:

* *Revealing the variability of performance to improve them:* Creating and storing transactional data in digital form allows companies to have more accurate and detailed data on a variety of performances, from the daily state of the warehouse to staff sickness, all in real time or nearly so. Furthermore, they use the data to analyze and in order to understand the deeper causes of the variabilities;
* *Customize actions*: The Big Data allow you to create specific customer segments, called cluster, to customize products, services, promotions and advertising based on their needs;
* *Improve forecasting and supporting people in the decision-making process*: Using sophisticated analytics on entire dataset can automate and improve decision-making by the predictions of Key Performance Indicators (KPI), useful to minimize risk and discover valuable insights; These benefits cannot be pursued with the analysis and management of small samples of data via spreadsheets, but need a large number of integrated data that only a data warehouse can provide. Retailers, for example, can use algorithms that enable the auto-tuning and optimization of inventories and prices, starting from real-time data on sales.
* *Transparency*: An easy and timely access to big data makes an available larger amount of information and facilitates the sharing of data between the different organizational units of an enterprise;
* *Consumer Profiling*: The availability of near real-time data from smartphones provides detailed characteristics about customers and their complex decision-making process when shopping, identifying consumer behavior patterns and shed light on their intentions;
* *Create new products and services, new types of companies and innovative business models:* Companies can leverage big data to improve the development of future models and to create innovative after-sales services;
* *Increase productivity and profitability of companies:* The exploitation of Big Data can lead to greater efficiency and effectiveness of the companies, which will produce more output using less input, improving the level of quality.

This list of benefits highlights how the investments in Big Data leading to the creation of value for companies, obtaining competitive advantage in the long term.

Despite the opportunities offered by big data is huge, there is still some skepticism within companies on the real benefits due to the lack of results in practice [3].

There are therefore a number of barriers to consider, which can be classified into six categories:

* *Technical barriers*: Data integration difficulties, low-grade of business influence, poor quality of data;
* *Barriers linked to skills*: Difficulty understanding of the analytical tools and benefit quantification, shortage of talents, difficulty in choosing the suitable tool;
* *Organizational barriers / Management*: Lack of commitment of top managers who are not involved in the Big Data initiatives, towards which show little interest, resulting ineffective;
* *Cultural barriers*: Most companies are not ready and totally open to innovations that could bring big data, since their exploitation would require significant cultural and organizational changes: Inertia.
* *Economic barriers*: The Big Data initiatives require huge expenses in terms of implemented technologies and new professionals to hire or consulting;
* *Barriers related to privacy:* Consumers do not want that their personal information, such as personal location data and electronic data generated by their use of the Internet are used by companies, especially because they do not know where and how these will be exploited by the organizations, which must also consider the laws relating to statements of different countries. Tools that let track any movement of employees and their performance make the interests of organizations but not for worker who see a threat to their privacy.

### ***1.4.2 Techniques for Big Data Analysis***

So far, we have talked about the ideology and the value that Big Data can lead to a company. Here, however, they will be listed the main techniques and technologies used to aggregate, manipulate, manage and analyze the data.

* *A / B testing*: Technique in which a control group is compared with the test groups in order to determine what changes and actions will improve one target variable data, such as the response rate to a campaign of Marketing;
* *Crowdsourcing:* Technique used to collect data subjected to a large group of people or a community through, for example, the Web;
* *Data integration*: Set of techniques that integrate and analyze data from different sources in order to develop insight more efficient and accurate than those obtained by examining a single source;
* *Predictive models*: Techniques in which is created or chosen a mathematical model to predict the probability of a result;
* *Data mining*: Set of classification techniques, cluster analysis, regression and association rules, which allows to extract patterns from large datasets by combining statistical methods, machine learning and database management;
* *Machine Learning*: Part of computer science dedicated on the design and development of algorithms that allow computers to identify behaviors based on empirical data and to recognize complex patterns for predicting decisions by means of artificial intelligence;
* *Natural language processing (NLP)*: Set of linguistic science and computer techniques that use computers to analyze human language;
* *Regression*: Set of techniques that make it possible determine how the value of a dependent variable changes when one or more independent variables are changed;
* *Optimization:* Set of numerical techniques used to redesign systems and complex processes in order to improve performance with respect to one or more aspects, including cost, speed and reliability;
* *Sentiment Analysis*: Application of natural language processing and other analytical techniques to identify and extract subjective information from the texts, for example the "polarity" (positive, negative or neutral) on the characteristics of the products on which the people have expressed an evaluation;
* *Statistics*: The science of collecting, organizing and interpreting data, used to make judgments about the relationships between variables that could have occurred by chance (null hypothesis) and by causal (statistically significant);
* *Data Visualization*: Creation of images, charts, diagrams or animations that let you communicate, understand and improve the results of Big Data.

## ***1.5 BIG DATA PROJECTS IN MARKETING***

The exploitation of Big Data in Marketing is a huge potential for the companies that have a great interest in projects that provide for their use in this area. We will face on six type of projects: Direct and Digital Marketing, the Customer Micro-Segmentation, the Location-based Marketing, Price Optimization, In-store Analysis and Cross-Selling / Up- selling.

### ***1.5.1 Direct and Digital Marketing***

The Direct Marketing includes all the marketing techniques that allow companies a targeted and personalized way to communicate directly with the customer or end user. The continuous and significant growth of Internet has led to the rapid development of Digital Marketing, which takes the form of advertising, content on Facebook, video clips on You Tube, personalized email and much more. Companies to make Digital Marketing today can rely on the enormous amount of information of users who spend hours and hours on the Internet, sharing their interests, the content of their communications, the purchases they make and more [4].

The Direct Marketing uses many of Big Data techniques, as well as to identify the most profitable customers and those most likely to respond to the market, so they also to predict the behavior of those strangers. They are used both unsupervised learning techniques such as optimization models, neural networks and Bayesian decision trees and those not supervised, including clustering. For best results, the ideal is to combine the several techniques [5].

The benefits from Big Data to Direct Marketing are, in addition to the personalization of the message, the 360-degree view of the customer, the identification of content, the timing and the most appropriate channel to send the message in real time.

This results in an increase in the conversion rate, like the number of visitors who decide to click some random content or to visit a web site as a result of an action driven, and thus the maximization of Digital Return Of Investment (ROID) , the acquisition of new customers and the retention of those who already are client of the company.

### ***1.5.2 Customer Micro-Segmentation***

The variety of new types of data and the development of advanced Analytics allows for granular details and a larger number of consumer reports, generating very precise micro-segments, constituted by a small number of people [1]. Traditional segments B2C (Business to Customers) and B2B (Business to Business) based on demographics, psychographics, behavioral and on the size of the companies or the acquisition criteria are obsolete.

The most common criteria are:

* *Activity-Based Data*: Click-stream data from the web, historical purchases, the call center data, mobile data;
* *Profiles of social networks*: Historical activities and membership in groups;
* *Sentiment Data*: Associations with products and businesses (like or follows) and online comments.
* *Traditional data*: Market research and transactional data;

The goal concern into build always more narrow segments. The marketing men can therefore create offers, customized products and services tailored to each cluster, with obvious benefits on returns. This data can also be updated in real-time, thus being able to monitor customer changes and preferences.

### ***1.5.3 Price Optimization***

Companies can take advantage from the increasing granularity of sales data and powerful analytics to optimize prices. The amount of information available to them is huge, thank to the historical demand series, the inventory data, until the current sales level. This database is constantly rising given the explosion of new online sales channels where consumers can compare prices, increasing competition between different brands on the market [6].

From these large amounts of data, through appropriate tools, pricing managers are able to extract insight to define the optimal price almost in real-time that a consumer is willing to pay for each product, based on its characteristics.

The price optimization can consider, for example, the elasticity of demand to price, with specific models that analyze historical sales data to derive insights on the pricing of each unit, which can then be used to make promotions, to reduce prices or to evaluate costs. The benefits that businesses can achieve in this way can increase revenues, margins and market share.

However, is necessary to build a confidence state with the customers, identify the most promising opportunities, for determining what exactly the consumer wants pay for a given product through customer segmentation and personalized promotions. Particular attention is focused on the correct use of adequate analytics to identify items that are often overlooked, and to determine the driving factors for each customer and product that will lead to price final choice [7].

### ***1.5.4 Location-Based Marketing***

The Location-based Marketing is based on the adoption of the growing smartphone and other mobile devices that generate personal location data, which allow you to learn about location and behavior of people in real-time using GPS or Wi-Fi technologies, encouraging the development of a marketing strategy. We need to considers also the habits and fun of worker and not only the consumer preferences. Other sources used are the signals of the triangulation of cell towers and the payments through credit and debit cards, which, through the point of sale terminal, make available personal information.

What businesses usually they make is called the Geo-Targeted Advertising, advertising or undertaking actions in real time based on the location of its customers. In fact, to get huge benefits, companies use several push notifications, for example, current offers and customized for a specific customer that, for example, walking holding smartphone inside the store. Therefore, exploitation of geo-location data may lead to an increase in sales, with an increase in profits and improved customer experience and therefore, increasing customer loyalty.

However, with respect to this project, companies are faced with two challenges: The privacy and a trade-off, that is, if users wish to receive Mobile offers when you are near to a store.

### ***1.5.5 In-Store Analysis***

The in-store analysis includes the analysis of real-time data on the behavior of consumers through the position and location of the customers inside the store are tracked through a variety of technologies: video cameras, Wi-Fi, Bluetooth, systems tools of retail outlets, payment cards, transponder carts, smartphone applications, Path Intelligence and RFID tags on purchasing cards.

By doing this insight are excerpts related to consumer behavior in the store, with a ultimate goal of improving the customer experience.

In particular, the insights obtained are related to how many customers enter the store, how they behave shoppers inside the store and to know the consumer through the attributes like sex, age, if it is the first time that enters the store, if it comes back often, where it comes from and what are his interests.

Companies use these insights to effectively improve the organization, or to optimize the layout of the store, its characteristics, shelf placement and product mix to turn one-time customers into repeat customers, to increase the frequency of their visits and their expenses by improving the store experience.

### ***1.5.6 Cross-Selling / Up-Selling***

The Big Data offer great opportunities to increase the average purchase size of a consumer, providing products or services related to the initial purchase choice improving the actions of Cross- selling and Up-selling. Data such as the demographic characteristics of customers, the real-time location, preferences, history of past purchases are used for this purpose [9].

The benefits that companies have are th increasing sales, profits, and customer loyalty.

Case example is Amazon which collects data from all users, recognizing the trend in people who make purchases through analytics tools , in order to capture potential opportunities and, according to each product or service visited in the site, suggests the client to buy other similar thing available in the website, significantly increasing sales.

## ***1.6 KNOWLEDGE DISCOVERY IN DATABASE (KDD)***

Immagine che contiene interni

Descrizione generata con affidabilità molto elevata

Figure 6: KDD Process

The KDD is an interactive and iterative process that seeks to extract implicit information from the data, a priori unknown but potentially useful.

Let's now analyze the individual steps:

* *Identify your goals*: The goal of this phase is the identification of the objectives to be pursued. It is perhaps the most difficult stage in terms of resource allocation and because it must be determined the criteria for measuring the success or the failure. Is possible make only a partial list of the many aspects that must be taken into account, like the estimated cost of the project and the choice of data mining tools to be used;
* *Selection*: The raw data are segmented and selected according to some criteria in order to create at a subset of data, which represent our target. If the original data is placed in a flat file, creating the target it is very simple. The management systems store and manipulate data transactional database, which enables information systems for making upgrades and extract information quickly. This is due to the structuring of data using relational models, whose aim are to accelerate the access to information and reduce data redundancy, through the decomposition of individual tables in the relational structures. Often, you also need to put together information extracted from multiple sources, which can make the selection process more difficult because you have to transform the data to ensure consistency in that, for example, data encryption must be equal for all records of the target data, otherwise the analysis is of little use; whose purpose is precisely to bring together data, and do not decompose in order to exploit redundancy. Often you also need to put together information extracted from multiple sources, which can make the selection process difficult because you have to transform the data to ensure consistency in that, for example, data encryption must be equal for all records of the target data, otherwise the analysis is of little use;
* *Preprocessing:* Generally, the available target date should not be analyzed entirely but just extract an appropriate sample, then performing an analysis on a sample basis. Furthermore, the data must be pre-processed, ie "clean", treating in a timely manner outliers and missing data. Should be identified incorrect values ​​of the variables; to find errors in categorical data becomes a problem when analyzing very large dataset. The data should also be simplified; these smoothing date techniques are aimed at reducing the number of values ​​for a numeric variable. Some classifiers, such as neural networks, using functions that perform the simplification during the classification process, performing so a date internal smoothing. Two simple simplification techniques are the calculation and rounding of average values;
* *Transformation*: The data to be used, often have to be transformed; this phase may take various forms and may be necessary for various reasons. It can convert data types in other or define new ones, obtained through the use of mathematical and logical operations on the variables, perform normalizations (decimal scaling, normalization min-max or with the z-score) or even eliminate the variables. In general, in fact, the DM algorithms do not work efficiently if the data contains a large amount of variables that are not able to predict the class. It is therefore useful to a search and a subsequent elimination of redundant and "unnecessary variables" for the problem in question.
* *Data mining*: Algorithms that study data to give information not trivial or obvious. These are the objectives to be achieved to give an indication of the type of technique to be applied;
* *Interpretation and evaluation*: The purpose of the DM is to determine the validity of the model obtained; in short, is not sufficient just interpret the results but you will have to understand to what extent this model or results will be useful. This can be done in various ways and through statistical analysis or experimental;
* *Data Visualization*: The ultimate goal is to use what has been learned, creating a technical report on what has been discovered, trying to figure out how to exploit what has been discovered.

From this analysis, it is understandable how the process of extracting knowledge is long and rather complex, therefore, the choices that are made for the treatment of anomalies or errors in the data and the clear identification of the objectives to be pursued are fundamental.

### ***1.6.1 Data Mining vs Machine Learning***

Data mining refers to the extraction of knowledge from large amounts of accurate, new and useful data. It is an iterative process of creating a predictive and descriptive model, through the discovery of previously unknown trends and patterns with large amounts of data to support decision making. It can also be defined as the subset of business analysis, similar to an experimental research. The data mining sources are the databases and statistical methods.

The Machine Learning indicates an area of research in artificial intelligence and, thanks to data-driven, involves the study of algorithms that are able to extract information automatically. Two data sources are required: training data and test data. Usually, the machine learning using data mining techniques and another learning algorithm to build models of what is happening behind some data so that it can predict future results.

But we see in the table the various differences:

Table2: Data Mining vs Machine Learning [13]

|  |  |  |
| --- | --- | --- |
|  | **Data mining** | **Machine learning** |
| **Definition** | Extract Knowledge from a large amount of data | Introduce a new algorithm to data and past experience |
| **History** | Introduced in 1930 | Introduced in 1950 |
| **Responsibility** | Data mining is used to get the rules from existing data. | Automatic learning that teaching to the computer to understand the rules given. |
| **Origin** | Traditional database with unstructured data | Existing algorithms and data. |
| **Implementation** | Develop models where we can use data mining techniques. | We can use the algorithm of the decision tree machine learning, neural networks, and in some other area of ​​artificial intelligence. |
| **Nature** | Manual | Automatic |
| **Application** | Used in the cluster analysis | Used in web search, filter spam, credit scoring, fraud detection, computer design |
| **Techniques** | Data mining is more than a search that uses methods to give not obvious information. | Auto learning and teaching done by intelligent task. |
| **Purpose** | Limited Area | Large area. |

## ***1.7 Data Mining Algorithms***

The goal of data mining is to extract new information from existing data. As we shall see, there are two approaches: supervised learning and unsupervised learning [14].

* ***Supervised learning***: Machine learning methodology in which it is passed to the machine data containing the original data and the expected result. The task of the machine is to find the rule (or function model) with which to create a relationship between the two in such a way that, at the occurrence of a previously unknown example, can obtain the correct result. The data is previously labeled or assigned to a certain category. The supervised learning is mainly used for classification problems, such as, for example, is used in marketing to classify potential customers and offer products that could be interested on the basis of the profile and history of purchases. Another example is the anti-spam email systems.;
* *Unsupervised learning:* Unlike the previous one, does not use classified data labeled before; we do not know, then, what category they belong. Therefore, to the machine is requested to extract a rule that groups the cases presented in accordance with characteristics which derives from the data themselves. For this, it is also defined characteristics of learning (learning feature). The algorithms in this case seek a relationship between the data to understand if and how they are linked together. Since it contains no preset information, the algorithm is called to create "new knowledge" (knowledge discovery). One of the main applications is the clustering or grouping data into homogeneous groups called clusters. The unsupervised learning, therefore, generally serves to extract information not yet known.

Immagine che contiene screenshot

Descrizione generata con affidabilità molto elevata

Figure 7: Data Mining Algoritms

Using some of the techniques mentioned above we can create some predictive models. Whatever their application, the predictive models use the experience to give a score, confidence levels and some interesting results for the future. To do this, you have to divide the process into two phases:

The first stage is the creation, in which the model is created using data from the past, while the second is the score of the model created.

We must never forget that is important to get good results (score) in the test data and not in the training data. The overfitting is the situation that occurs when the model explains the training data but can not generalize to test data.

The innovations that use artificial intelligence and machine learning are the major technology trends in the retail world. They are having a big impact on the industry, particularly in e-commerce companies that rely on online sales, where the use of some form of machine learning is very common today, especially in retail.

Large online retailers such as eBay, Amazon and Alibaba have successfully integrated AI technologies throughout the sales cycle, from storage logistics to customers after-sales service.

Companies using the recommendation systems achieve sales increases as a result of a better customer experience. The recommendations, in general, accelerate research and make it easier to acquire and retain customers by sending e-mails with links to new offers that meet the interests of the recipients and adapt to their profiles.

When the user begins to feel understood, it is more likely to purchase additional products. Knowing what a customer wants and show immediately to him the product, it is less likely that he leaves the platform. This result give a greater chance of purchase and a decrease in threat of losing a customer that move to a competitor.

By including the offer, seasonality, external events related to your business (such as a concert, a match, a festival), demand forecast and market supply, an automatic pricing system can efficiently adjust prices.

We see in detail the most common by machine learning algorithms used.

### ***1.7.1 Clustering***

The goal of clustering is to organize the objects examined in groups that sharing similar properties. Clustering can be considered one of the most important methods of unsupervised learning and, like any method belonging to this category, it does not use certain clasificators prior to guess the possible structure of the data.

There are various forms of clustering [15]:

1. *Exclusive Clustering*: Each element can only belong to a cluster, that is, the intersections between the clusters are always empty sets; This procedure is also called Hard Clustering;
2. *Inclusive Clustering*: Each item can belong to multiple clusters simultaneously, with an index that determine the degree of membership in each cluster. This procedure that is called Soft and Fuzzy Clustering;
3. *Partitional Clustering*: It uses the concept of distance between the elements, which belong to a particular group based on their relationship with a significant point of the dataset;
4. *Hierarchical Clustering*: It builds a hierarchy of partitions both for aggregation that by division, by means of a tree representation that takes the name of dendrogram. There are other more detailed subdivisions with respect to the Clustering partition, which differ in the evaluation of the distance between the elements and the related cluster [19] creation. This technique is divided into two approaches:

* Agglomerative: The process begins by considering each point as a cluster, at each step are unified points according to a particular arbitrary function of similarity, to obtain a single cluster and its dendrogram. This approach is based on the development of a Proximity Matrix that take into consideration the function for calculating the similarity between two clusters;
* Divisive: Complementary case in which one starts from a single cluster that is divided at each iteration, until obtaining a number of clusters equal to the number of points that constitute the data base.

The complexity is in the order of O (N3), and as in K-Means the presence of outliers create some very important problem to this approach.

Next, it will be show the main clustering strategies and algorithms used in Fashion.

#### **1.7.1.1 K-Means Clustering**

The centroid clustering is based is represented by a prototype called centroid which typically is the average of the distances of the points in the cluster. One of the most popular clustering algorithms in this category is the K-Means that requires you to specify the number of K clusters to be obtained. The algorithm iteratively elects the K centroids of the clusters, and each element is associated with the closest centroid. The algorithm is as follows:

|  |
| --- |
| **K-MEANS ALGORITHM** |
| 1: Function K-Means (clusters K) |
| 2: *Election K centroids* |
| 3: **repeat** |
| 4: *Assignment of each element to the point K more*close |
| 5: *Recalculation of the K centroids* **until***The centroids do not vary* |

Initially, the centroids are chosen randomly while, in later iterations of the algorithm, they typically consist in the average between the distances of the points in the cluster. There are different methods to calculate this distance: Euclidean Distance, Cosine Similarity, Correlation. The algorithm converges to the similarity measures listed. This convergence occurs mainly in the early iterations, followed by a phase of adjustment. In it, in fact, often the stop condition is relaxed, admitting a minimum threshold of change between the centroids.

The choice of centroids is a very sensitive stage, in fact are applied the following techniques to solve, even if not completely, the problem:

* It Performing multiple executions estimating the centroids in different ways or just randomly. After, we must evaluating the quality of the results obtained by means of validation tools that will be described later;
* You use the Hierarchical Clustering procedure to perform subdivisions K and calculate the centroid of the obtained clusters. These are the starting points for the algorithm K-Means;
* It is estimated a number of centroids N> K;
* Postprocessing techniques, such as elimination of small clusters, merge some cluster with very similar between them and breakdown of clusters too large;
* It uses the bisecting algorithm. It consists of a hierarchical approach through which, starting from a single cluster, is divided by a 2-Means algorithm an arbitrary number of times. The iteration that produced the best cluster is taken, and the algorithm is applied recursively until you get the desired K cluster.

Immagine che contiene testo, mappa

Descrizione generata con affidabilità molto elevataImmagine che contiene testo

Descrizione generata con affidabilità molto elevata

Figure 8: K-Means Algorithm

The complexity of the algorithm is O (n \* K \* I \* d) where n is the number of points, K the number of clusters, I the number of iterations and d is the number of attributes based on which the function for the calculation of the distance.

In conclusion, the K-Means algorithm presents some difficulties in the management of data whose presence of outliers is too high. In fact, are often performed procedures Preprocessing to mitigate the problem. In addition, as mentioned earlier, the choice of centroids is often difficult, especially when it has to do with high-density data. However, K-Means is one of the most widely used algorithms especially with regard to the problem of Customer Segmentation.

#### **1.7.1.2 Density-Based Clustering**

The density-based clustering is based on the concept of density. The basic idea is to find clusters defined implicitly by high density regions separated by low-density regions. One of the most famous algorithms of this category is the DBSCAN that uses two parameters to identify dense areas: a ε-range, which is used to identify an area around a given point, and a minimum number of that must be present at the internal radius of ε. Each point is labeled in accordance with 3 different categories:

* *Core Point*: All the points that exceed the threshold MinPts within the radius ε;
* *Border Point*: All points that do not exceed the threshold MinPts but within their range ε ​​have at least one Core Point;
* *Noise Point*: All the points that are not Core or Border Point.

The algorithm starts from a random point. All points included in the radius ε are calculated and if it contains a MinPts number of points, a new cluster is created otherwise it is labeled Noise-Point. The point could subsequently be found as it is included in the radius ε of a neighbor and consequently be inserted in a cluster.

If a point is associated with a cluster, the points within its radius ε are also inserted in it, and consequently also their neighbors within the radius established. This process continues until all the neighbors have been entered. Each point a cluster is associated with is marked as visited and the algorithm continues by performing the same procedure for a subsequent point that has not yet been visited.

The algorithm has *O*(*n*2) complexity but this can be reduced to O (n log n) by use of structures indexed for querying the neighborhood.

The strength of this approach is given by the good management of outliers and the consequent ability to be able to handle cluster of shapes and sizes very different. However, it is inefficient when dealing with data that are characterized by a density too variable. It is widely used to cluster for Geo-location.

### ***1.7.2 Classification and Regression Trees (CART)***

CART is a non-parametric procedure where you do not need to pre-test the normality or other assumptions concern to the statistical distribution of data. The final tree includes only the independent variables that appear to be predictive of the dependent variable; the other independent variables are not predictive have no effect on the final result; CART from this aspect differs from other traditional statistical procedures. With the classification term is refers to the process that through a collection of records, called Training Set, trying to build a model able to attribute a feature called Class attribute, based on the combination of other properties that characterize the individual of the population. Once you have the template, the structure of a classification tree includes non-terminal nodes (parent nodes), which has two direct descendants (child nodes), and the terminal nodes that do not undergo further bipartitions (terminal nodes). The first node (root node) contains all observations. From the root node descended two "child nodes." Each child node, denoted with the letter "t" contains a subsample of the original sample, in which members share the same characteristics that influence the dependent variable of interest. Each t, in turn, constitutes a parent node potential that can still be divided into two child nodes. The process continues until the tree does not stop its growth.

There are some important steps to follow when building a decision tree with the CART procedure: 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.

To choose the split criterion is generally used a technique of Recursive Binary Splitting.

The method is binary and recursive: binary, since each parent node is divided into two direct descendants, and recursive, since the nodes (non-terminal) born from the splitting of the parent node into two direct descendants tha can become, in turn, parent node divided successively into two others nodes.

A decision trees with many nodes and a huge number of divisions may lead to a data overfitting. This means that the model is difficult to interpret, as it becomes inaccurate for later forecasts and needs of the stopping rule. The methods to avoid this problem is 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.

The different existing algorithms differ depending on the strategy employed on individual nodes, for the evaluation of Split. There are in fact different indices for the validation of a classification:

* *GINI INDEX:* Identifies the quality of the split. Considering the relative frequency of the node to the class:
* *GAIN INDEX*: It is based on the concept of entropy and identified the homogeneity index relating to the node, obtained by performing a particular split on node:

#### **1.7.2.2 Other Types Of Classification**

* *Based on Instances*: It consists of a family of algorithms which, rather than performing explicit generalizations, compare new instances directly with the analyzed records and properly stored by the training set. Worthy of note, is the Nearest-Neighbor procedure that uses a particular arbitrary metric for the distance calculation and a parameter k representing the minimum number of neighbors to be extracted [15]. For each record that should be classified, it calculates the distance from the training set identifying the k closest records and using the values of their attributes, classify the records;
* *Byesian Classifier:* It consists of a probabilistic framework for solving the problem of classification. They consider the attributes and the class as random variables based in a strongly relying on the concept of Conditional Probability. Given a record with attributes (A1, A2, ..., An), the goal is to predict the class C. We want to find the value of C that maximizes the probability P (C A1, A2, ..., An). It follows the Bayes theorem:

Thanks to Bayes' theorem, you get an equivalent optimization problem that is to find C that maximizes: P (A | C) = P (A, C). There are different ways to estimate this probability based on data, such as normal distribution, density estimate and Laplace [15];

* *Support Vector Machine (SVM):* The classification is performed by finding the hyperplane that maximizes the margin between the two classes. The vectors (possible attributes of the class) that define the hyperplane are called support vectors. The advantage of this method is that if the data is linearly separable, then there exists a unique global minimum. An ideal SVM produce a hyperplane which completely separates the two non-overlapping classes. Typically, the complete separation is not always possible, but often it gets to obtain a model with too many possible cases that involves an incorrect classification [18].

The validation of these process has a huge importance, since it allows to evaluate the performance of the constructed model and can compare it with other possible modeling. The evaluation measures are based on the Test-Set, the data partition on which to apply the predictive model.

The application of the model on the Test-Set produces the Confusion Matrix, which is a matrix indicating the incidence between the classes and their real value of the records in the Test-Set. You can then determine the following types of prediction:

* *True Positive:* Correct Predictions Positive;
* *False Positive:* Correct Predictions Negative;
* *True Negative:* Wrong Predictions Positive;
* *False Negative:* Wrong Predictions Negative.

This can be applied to any type of attribute, not only to binary classes.

The most commonly used metrics are: Accuracy, Precision, Recall, F-Measure.

**Immagine che contiene screenshot

Descrizione generata con affidabilità molto elevata**

Figure 9: Confusion Matrix

### ***1.7.3 Prediction: Association Rules***

The starting point of an association rule algorithm consists of a set of transactions. Each transaction consists of a set of items. The algorithm extracts the Association Rules to predict the occurrence of an item on the basis of other appropriate item, included also in the available transactions.

Is important to define some of the concepts behind this technique:

* *ItemSet*: A collection of one or more elements generally defined by means of the parameter k, indicative of its size in k-itemset form;
* *ItemSet Support*: Given an itemset I, the support is the fraction of transactions that contain I and denoted by supp (I);
* *Frequent itemset*: All itemsets that exceed arbitrary minimum support threshold;
* An association rule is an implication in the form: X → Y with X, Y itemset where X is called premise and Y is called consequence of the rule.

In addition to the support, seen previously, there is another form of validation rule that takes into account both the premise and the consequence: Confidance. Shows the frequency with which a given rule matches, is the ratio between the number of transactions that contain the rules and complete transactions that contain the premise:

Formally the support supp(X ∪*Y*) It can be rewritten as the joint probability *P*(*EX ∩ EY* ), where *EX* e *EY* are all the transactions that contain X or Y respectively. Thus, we can express confidence as the conditional probability *P* (*EX* | *EY*).

Given a set of transactions, the goal is the extraction of all the rules that meet the arbitrary threshold of the support and the confidence. Their extraction can not be performed with a Brute-force approach, due to the number of rules that can be generated. To reduce the number of possible rules, it uses the Apriori principle.

#### **1.7.3.1 Apriori**

This principle is based on the anti-monotonic properties of the support, which allows to establish with certainty that if an itemset is not frequent, then even all itemset that contain it will become not frequent. Such a property is so formalized, with X and Y itemsets:

*∀X, Y: (X ⊆ Y) ⇒ supp (X) ≥ supp (Y)*

This property is the basis of Apriori algorithm, where, starting from all possible items with cardinality 1, it builds all the itemset of dimension n + 1, with "n" the start dimension of the itemset; at each iteration occurs if the itemset generated is frequent or less.

The anti-monotone property allows to exclude infrequent itemset and therefore all possible itemsets arising therefrom.

The steps to which the procedure is made are as follows:

|  |  |
| --- | --- |
| 1: **function** Apriori(*T, s*) | **DESCRIPTION** |
| 2: *L*1 ← {*large* 1 − *itemsets*} | (Set T transactions, minSupport) |
| 3: k=2 | *k* = 1 and Generation itemsets with cardinality 1 |
| 4: **while** *Lk*−1 ƒ= ∅ **do** |
| 5: *Ck* ← *Generate*(*Lk*−1) | *Generation itemset with cardinality* k + 1. |
| 6: **for** *transaction t* ∈ *T* **do** |
| 7: *Ct* ← *Subset*(*Ck*1*, t*) | *Elimination itemset containing infrequent.* |
| 8: **for** *candidates c* ∈ *Ct* **do** |
| 9: *count*[*c*] ← *count*[*c*] + 1 | *Calculation support itemset generated.* |
| 10: *Lk* ← {*c* ∈ *Ck*|*count*[*c*] ≥ *s*} |
| 11: *k* ← *k* + 1 | *Elimination itemset containing infrequent.* |
| 12: **return** U*k Li* |
| 1: **function** Apriori(*T, s*) |  |

At the end of this procedure, we get all itemsets that have stood the support threshold. We must proceed with the extraction of association rules from itemsets obtained. The generated rules will be evaluated according to their Confidence (arbitrary threshold), that, generally, does not enjoy the anti-monotonic properties.

Indicating with Conf (X⇒ Y) the confidence of the rule X ⇒ Y will get:

The algorithm proceeds generating the rules that possess only an item in the result, eliminating all the rules that do not exceed the minimum confidence threshold. On the basis of the remaining rules, it generates and evaluates rules with an additional item until all possible rules have been generated.

The extracted rules are subjected to a further step of post-processing, because the confidence can sometimes be misleading as to the validity index for a rule. This aspect emerges for itemset that are part of the premise of a rule, characterized by high support.

A very frequent itemset tends to raise the confidence index of the rules of which it constitutes the premise, regardless of the fact that the rule is contextually valid.

To have an excellent validation us is based on the following indices:

Immagine che contiene oggetto

Descrizione generata con affidabilità molto elevata

Figure 10: Indicators Of Association Rule

### ***1.7.4 Artificial Neural Networks & Deep Learning***

The Deep Learning is a specific method of machine learning that incorporates a large number of neural networks together in various layers to learn from the data iteratively.

Neural networks and deep learning are often used in image recognition applications, in speech and in computer vision.

A neural network is particularly useful when trying to study the pattern of unstructured data and are designed to emulate how, through artificial intelligence, computers can be trained to deal with problems that are not well defined [21].

It consists of three or more layers: an input layer, one or more hidden layers and an output layer. Data is ingested through the input level. Then, the data are edited and processed in the hidden layer, obtaining different output levels on the basis of weights applied to individual hidden nodes.

The typical neural network can consist of thousands, or even millions, of simple processing nodes that are densely interconnected..

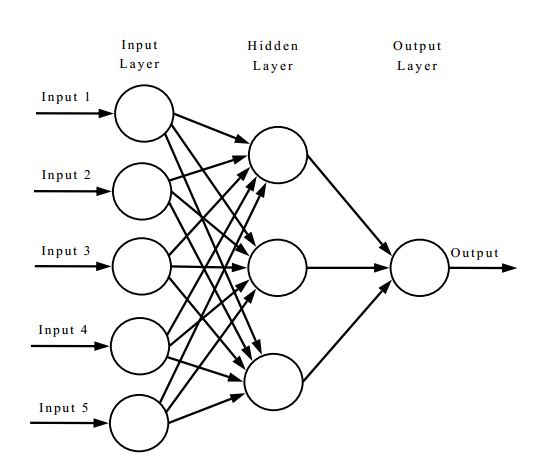


Figure 11: Artificial Neural Network Diagram

In a neural network, the input layer are constituted by the value of the attributes that must be analyzed. The output of this first level of the network remains unchanged, since the output from the input nodes are the same values that are provided for the analysis. In each node belonging to the next levels, hidden layer and output layer, it occurs the actual computation. In fact, the inputs of these levels correspond to the output of previous levels in which, however, must consider the weight associated to the connection between the two nodes and a characteristic value of the node, the offset. Considering a 𝑛 node between the hidden nodes or between those of output, its 𝐼𝑛 input is given by the following relationship:

where 𝑤𝑖, 𝑛 is the weight of the connection between node 𝑖 of the previous level and node taked into account, 𝑂𝑖 is the output of node 𝑖 of the previous level and 𝑜𝑓𝑓𝑠𝑒𝑡𝑛 is the offset associated with node 𝑛 considered.

Furthermore, each node applies an activation function on the value that receives as input and sends the output to the next level. When the output from the nodes is generated and if during the learning phase an error occurs between the value of the calculated class and that expected for an instance, this error must be propagated to the previous levels, where the weight and offsets values ​​will be arranged of all the nodes of the layers that make up the neural network..

The term Deep Learning is used when there are multiple levels hidden within a neural network. Using an iterative approach, a neural network adapts and continually makes inferences until a specific stopping point is reached. Practically, it is a machine learning technique that uses the hierarchy of neural networks to learn from untagged and unstructured data through a combination of unsupervised algorithms and supervised algorithms.

Deep Learning is used in Internet of Things (IoT) applications or to predict when a machine will malfunction and is often referred to as a sub-discipline of Machine Learning.

### ***1.7.5 Linear Regression***

The regression analysis is a statistical technique used to determine a relationship between a dependent variable and a set of explanatory factors. The dependent variable, referred to as variable Y, is the value we are trying to determine on the basis of the indipendent variables.

The explanatory factors, referred to as X variables are also called independent variables or simply model factors. Regression analysis helps analysts to find out the sensitivity of the dependent variable to changes in explanatory factors. These feelings are essential for proper risk management.

There are three types of data commonly used in the regression analysis: time series, cross-sections and grouped data.

* Time series: Data collected for a period of time. Are economic and financial data series refer to market returns, prices and asset values, GDP, unemployment rates, interest rates, etc. These data are collected at equal time intervals such as daily, monthly, quarterly, etc.;
* Cross section: Data collected for a family of variables at the same time. For example, fundamental analysis often collects company-specific information such as price / earnings ratio, the book value, the net debt / capital ratio, or the average daily turnover;
* Grouped Data: combination of time series and cross section data.

If we have more explanatory factors, the analysis is called multiple regression model has the form:

Y = b 0+ b 1X+ b 2X2 + ⋯ b kXk + ε

where Y is the dependent variable (what we are trying to predict), X is the indipendent variable (what we are using to predict), and ε is the random noise (error). In addition, the dependent variable Y, the indipendent variable X and the error ε are vectors of values ​​of columns.

Immagine che contiene testo, mappa

Descrizione generata con affidabilità molto elevata

Figure 12: Linear Regession

In the previous equation, b 0 and b 1 are the parameters of the current model which define the exact sensitivity of the dependent variable to the explanatory factors, and ε is the amount of variability that is not explained by the model.

In practice, these exact values ​​are not known with certainty, and must be estimated from the data. To do this you use:

Variance =

Expected Value =

The goal of regression analysis is to determine the set of explanatory factors and corresponding sensitivity that explain as much as possible the employees observed values.

**Metrics and Evaluation Assumptions of the model.**

When performing regression analysis, the main metrics to analyze are:

* b k = Parameter of the model refers to the estimated sensitivity of the Y k factor;
* R2 = Goodness of fit (the percentage of the total variance explained by the model). The linear determination index is defined as the ratio of composition between deviance regression and total deviance measuring in the interval [0,1], explaining how much of the total deviance have the regressors of the model. If we consider the decomposition of the total SST deviance (Sum of Squares for Total Variation) in deviance regression SSR (Sum of Squares two to Regression) and deviance residual SSE (Sum of Squares two to Residual), we show that, with increasing the number of explanatory variables, the deviance of residues decreases and thus the linear determination index increases. Therefore, a high R2 value is not a good fit indicator because it also depends on the number of covariates included in the model.
* F-stat= Critical value for the entire model. Most of the F test arises considering a decomposition of the variability in a data collection in terms of sums of squares. The F-test is the relationship between two scaled sums of squares that reflect different sources of variability. These sums of squares are built in such a way that the statistic tends to be greater when the null hypothesis is not true. Following the F-distribution under the null hypothesis, the data values ​​are independent and normally distributed with a common variance, so, the sum of the squares should be statistically independent, and each should follow a distribution χ² scale [29 ];
* T-stat= Critical value for the estimated parameter. Thisstatistic is used in a T test when deciding whether to support or to reject the null hypothesis. The higher the T, the greater the evidence that the values ​​are significantly different from the mean. Conversely, a lower value indicates that T is not significantly different from the average [30].