Importance of Statistics for Data Mining and Data Science

Vitor Ribeiro, André Rocha, Rui Peixoto, Filipe Portela* and Manuel Filipe Santos
Algoritmi Research Centre, University of Minho, Portugal

***Abstract*—Knowledge has been significantly recognized by managers as an important asset for organizations. This recognition stems from the fact that knowledge is increasingly used as a strategic resource to create competitive advantage, improve organizational processes, reduce costs, and more. Data Mining (DM) is an area of study that facilitates that process, allowing you to extract useful information and predictions from the vast data sets produced by the company. With the help of statistics and their mathematical methods, DM has gradually become important and useful. Some of the main statistical metrics used to perform data analysis are mean, median, variance, standard deviation, variance analysis, correlation and regression. This study aims to highlight and prove the importance of statistics in DM, which has so much potential in terms of creating a competitive advantage on behalf of the companies. A case study using Intensive Care Medicine data were chosen to prove the importance of statistics for Data Mining.**

***Keywords*— Data Mining, Statistics, Statistical Analysis, Data Science.**

I. Introduction

The new economy poses challenges, while also offering opportunities for organizations. To overcome challenges and to take advantage of opportunities, organizations need to take an active stance and update their strategies, taking risks on new management tools. In this context, organizations need to study, evaluate and extract relevant information from the vast data sets produced by their information system.

Statistics consist of data science that involves resorting, classifying, structuring, organizing, analyzing, and interpreting numerical information [1], [2]. This area and the statistical analysis, although increasingly used to improve organizational decision-making, is beginning to be insufficient to evaluate organizational data sets in consequence of their size.

Data Mining (DM) is a relatively recent area, but capable of covering the inability to analyze huge data sets from the statistics side. DM is an area capable of performing the process of exploring and analyzing large data sets, either automatically or semi-automatically, allowing the extraction of useful information, patterns, associations or trends [3]–[6].

This article intends to emphasize the importance that statistics have on projects and on the area of DM, resorting to its origins, components and phases of its projects, using as base the methodology Cross Industry Standard Process for Data Mining (CRISP-DM). This methodology defines a standard process model, which provides a framework to help accomplish DM projects, regardless of industry and technology used [7].

In the following section, all topics related to this article will be presented. In a third section, CRISP-DM methodology will be briefly explained, followed by the phases of the methodology. In a fourth section, an explanation and proof of the importance of statistics and the analytical methods for DM area and for the support of related projects phases. In this same section, a comparison of statistics with the Data Science area as well as with decision making. In a fifth section, a tool capable of performing high quality statistical analysis and performing certain DM activities is mentioned followed by an example of how to do statistical analysis with the tool. In the sixth section is present a case study where a DM process is performed to define a better scenario of a dataset with vital signs of some patients collected in a 120 hours period. To this scenario, a statistical analysis using the tool R is realized, with the objective of evaluating the results obtained in the DM process.

II. Background

A. Data Science

According to Loukides [8], the future belongs to companies that realize how to collect and use data successfully. Taking this into account, one of the areas that has become very important is Data Science. This is an area that is associated to collection, preparation, analysis, visualization, management and preservation of large amount of information [9], [10].

The book “An Introduction to Data Science” [9] reveals that although this is an area closely connected to mathematics, statistics and computer science, also include non-mathematical skills, such as communication, ethical reasoning, and being able to think critically about how data will be used. This in consequence of majority of the data in the world is unstructured and qualitative.

According to the same author, Stanton [9], a data scientist plays a more active role in the "four A's": data architecture, data acquisition, data analysis and data archiving. As for the architecture, the value is in providing information on how the data is organized to support the visual analysis and representation. The acquisition focuses on how the data is collected, as well as how the data will be represented, transformed and connected. The analysis phase is the most engaging, requires many technical, mathematical and statistical aspects, but also excellent communication skills and ethical reasoning. These skills aim to determine what the user desires. Finally, archiving intends to allow the data to be reusable for all users in need.

Loukides [8] says that Data Science requires skills ranging from traditional computing to math and art. He argues that researchers combine entrepreneurship with patience, the will to build, and the ability to exploit and reach a solution by "think outside the box".

Another author, Baier et al. [10] states that the greatest challenge for Data Science is finding out what available evidence are useful for the task. Reaffirming that in practice finding meaningful evidence and interpreting its meaning is the key skill of this area. This same author also says that the results of this science goes through abstractions, that is, simple heuristic representations of reality.

As explained previously, statistics supports and is directly associated to Data Science. To emphasize this relationship, Loukides

[8] also says that the only difference between the two areas is that Data Science uses a holistic approach.

B. Statistic

Statistic is the root of science creation [11], recognized as science of data that involves resorting, classifying, structuring, organizing, analyzing and interpreting numerical information [1], [2]. According to Singpurwalla [2], there are two types of statistics commonly used in solving a problem or in making statistical decisions, these types are: descriptive and inferential statistics. The main purpose of descriptive statistics is to describe datasets, using numerical and graphical methods to discover patterns in the data, summarize relevant information and present it in a pleasant and perceptible way to users, so that they can improve decision making. Inferential statistics use data samples from dataset to make estimates, decisions, forecasts, and other generalizations. This article intends to focus mainly on descriptive statistics and the capabilities for data analysis and evaluation.

Nowadays, everyone deals with information obtained from data and, thus, create knowledge. This knowledge differs from person to person and is used in future interpretations and decisions, generating different meanings from the same data set. These differences between each individual creates one key concept of statistical analysis, variability [11], [12]. This variability, in statistics, can also be defined as "something that varies in some way" within a dataset, and is completely essential for quality statistical analysis [5]. The possibility to translate the variability by objective values using mathematical methods and variance [11]. This variability that exists in the data will be most important for this document.

According to Goodman et al. [1] and Steinley & Wasserman [13], statistical analysis is one of the most effective methodologies for producing standard models from complex data sets. This capability, combined with DM techniques, is quite promising for data analysis, provided that the results are guaranteed to have a solid statistical foundation.

Statistical analysis aims to answer specific questions, which directly influences the process of collecting information, making it necessary to create effective strategies for data collecting [6]. When the collected and analyzed data set is substantially large, the statistical area provides mathematical methods so that, when selecting a random sample, this sample allows properties to be inferred from the data set. For this to happen, the samples must be representative and big enough [14].

The results of statistical analysis continue to be data, but with value and useful to the end user. In order to provide this data as best as possible to the user, graphical representations are often used which assembly and summarize data based on patterns. These graphical representations display the most important characteristics and relations of the data [11]. Graphic representation means converting the data into a simple visual or tabular format [5], making the results easier to interpret and understand, more appealing and universal (dependent on the user's language).

C. Statistical Methods

A distribution of the most common statistical metrics can be illustrated through the diagram presented in Figure 1.

Initially, there are two types of variables used for statistical analysis: qualitative and quantitative variables. According to Longnecker & Ott [12] and Singpurwalla [2], qualitative variables correspond to data that cannot be measured on a natural numerical scale and are classified by one or more categories. In other words, they are non-numeric values that can be grouped into categories, such as Male and Female. Quantitative variables, in contrast, have real units of measure and can be recorded on a natural numerical scale, as they correspond to numerical data, such as 1, 2, and 3.

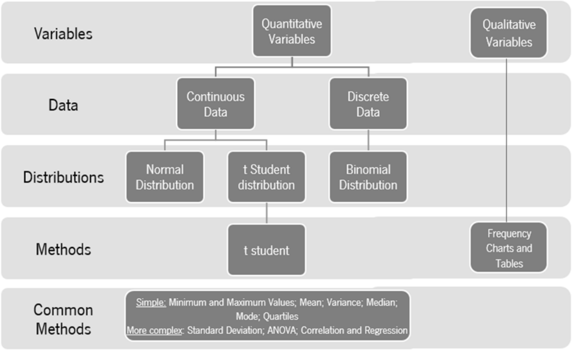


Figure 1 - Representative Diagram of Methods Used in Data Analysis

As for qualitative variables, one of the analysis consists of analyzing the number of occurrences of each existing category. For this, it is possible to use tables and graphs of frequencies, such as pie graphs and bar graphs, since they are the most used and easy to interpreter [2], [12].

Relative to quantitative variables, they can be divided into two types of data: discrete data and continuous data. Based on Casella et al. [11], Gorunescu [5] and Singpurwalla [2], discrete data can be defined as data represented by integers, which correspond to a countable number of distinct values. The distance between these values can be completely arbitrary. Examples of discrete data can be the number of children in a family as it will always be an integer. Continuous data, unlike discrete data, are usually obtained through measurements rather than counts. They are expressed in real numbers that can assume any value within one or more ranges. Some examples of continuous data can be heights, weights and time since they are all measurement results and can assume any real value.

In statistics, when it comes to discrete data, the binomial distribution is the most commonly found [11]. The binomial distribution consists of a sequence of an identical attempts, in which all these attempts have only two possible outcomes, success or failure. Each attempt is still independent of all others and the probability of success remains constant in all of them. The variable of interest in this distribution is the number of successes in n attempts [2], [11].

As for continuous data, one of the distributions most commonly used in statistics is the normal distribution [2], [14]. Many distributions found in natural sciences are modeled by this distribution, which facilitates approximations for the calculation of other distributions when they begin to have a very large number of observations [11], [14]. The main characteristics of normal distribution is: bell-shaped curve, which is denser in the center and less in the tails; always defined by its own mean and standard deviation; presents the mean, median and mode always the same [2].

Still in the continuous data, another distribution that is significantly used is the distribution t of student. This distribution is usually used when the sample size is small, to the point of disabling the use of the normal distribution. According to Casella et al. [11] and Singpurwalla [2], the distribution t is very similar to the normal, varying the density, being lower in the center and denser in the tails, thus allowing to produce values farther from the mean. It also varies in the fact that, for this type of distributions, only its mean is known, and the standard deviation unknown.

It should be noted that this last distribution, t of student, introduces a method called t of student, or t test, used to compare two samples of different sizes, helping to infer population means and coefficients of regression analysis [11], [15].

Lastly, the common methods that are applied to the three distributions referred previously to perform an analysis and understanding of the data are:

- Minimum and Maximum Values - Relative to the minimum and maximum value of the sample under analysis;
- Mean - Indicates where the data of a distribution is centralized. Calculated by dividing the sum of all values and the number of occurrences [2], [14];

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Median - Corresponds to the numerical value that separates the distribution into two equal halves in number of occurrences, in other words, the value that is exactly in the middle when ordering a sample [14];
- Mode – The value in the sample with the highest number of observations, which is, the most frequent or most common [2];
- Variance - Measures the propagation of a distribution, represents the quadratic mean distance at which the values are from the mean, in other words, the dispersion of the sample values from the mean [2], [14];

$$var(x) = \sigma^2 = E\,[(x-\mu)^2]$$

- Quartiles - In statistics, there are three main quartiles that divide the ordered data set into four equal parts. The first quartile corresponds to the value in the middle of the first half of the total sample, which is 25%. The second quartile corresponds to the median, the value that is in the middle. And the third quartile corresponds to the value found in the middle of the second half of the total sample, which is 75% [16];
- Standard Deviation - Measure from how much each distribution value deviates from the mean [14]. It can be calculated through the square root of the variance, and, unlike the variance, it's presented in the same unit as the dataset. According to Singpurwalla [2], the standard deviation has several properties that deserve to be referenced, such as: if the standard deviation equals 0 that means all values in the dataset are the same; is influenced by extreme values very deviated from the mean values, called outliers;

$$var(x) = \sigma^2 = E\,[(x-\mu)^2]$$

- ANOVA - Consists on the analysis of the variance between two or more samples taken from the same population, comparing means or medians. Similar to *t-test*, but safer when comparing more than two samples [11], [17];

$$F_{(a-1),(\sum n_i)-a} = \frac{\frac{SS_{treatment}}{df_{treatment}}}{\frac{SS_{residual}}{df_{residual}}} = \frac{MS_{Tr}}{MS_{Res}}$$

- Correlation - According to Singpurwalla [2], correlation measures the strength and direction of a linear association between two quantitative variables, that is, identifies and measures the dependence between variables;

$$\rho_{x,y} = \frac{cov(X,Y)}{\sigma_x\sigma_y} = \frac{E((X-\mu_x)-(Y-\mu_y))}{\sigma_x\sigma_y}$$

- Regression - Usually accompanies the use of correlation, adding the ability to distinguish between the dependent variable and the independent variable of a relation. It also adds the ability to describe how a dependent variable varies depending on the change of an independent variable. It also allows the use of a regression line to predict values of an independent variable for a given value of the dependent variable [2].

$$\hat{\beta}_1 = \frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

D. *Data Mining*

The amount of data that is generated and stored has been growing exponentially, which makes the human being unable to understand and extract useful information from these enormous data sets, just by looking or with simple statistical methods. However, in order to solve this problem, and by resorting to the increase of storage capacity and the possibility of processing data of all types, from incomplete data to data with errors, Data Mining (DM) tools appeared [3], [4].

In 2008, Goodman, Kamath & Kumar [1] admitted that DM was defined as the use of the power of computational technology to perform statistical analysis on a huge data set. Nowadays it is much more than that, being defined as the set of methods and techniques to explore and analyze large data sets, automatically or semi-automatically, with the purpose of extracting useful information, patterns, associations or trends [3]–[6]. Summarizing, DM consists on the art of extracting and generating knowledge from large data sets.

Similar to statistics, DM is composed of two types of techniques, descriptive and predictive. The descriptive analysis go on to obtain useful information that is in the data, but buried in the immense amount of records [3]. Predictive analysis is used to generate new information based on the data present [4], that is, to create predictions based on recorded events. According to Gorunescu [5], DM methods are a mixture of statistics, artificial intelligence and searches in databases (Figure 2). Statistic brought well defined techniques to identify systematic relations between different variables. Artificial intelligence contributes with information processing techniques, based on some form of human reasoning. Database systems have provided data storage that will later be used to extract information.

According to Hand et al. [6], DM approaches deal with data that has already been collected for some other purpose then DM. This allows such practices not to influence collecting data, contrary to statistical procedures, where data is collected according to a strategy, to answer specific questions.

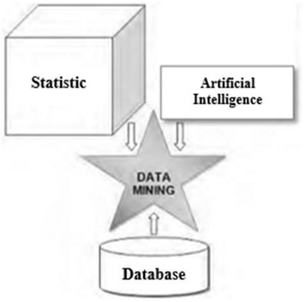


Figure 2 - Data Mining Components (Adapted from [5]*)*

To conclude, as we have already mentioned, DM's main goal is to extract useful information hidden in big data sets, considerably improving decision-making, and creating predictions based on existing data.

III. Cross industry standard Process for data Mining

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology defines a standard process model, which provides a framework to help execute DM projects, regardless of the industry and technology used. This support goes through trying to make these projects more reliable, faster, more controlled and less expensive [7].

This methodology is organized through a hierarchy, composed of sets of tasks divided by four levels of abstraction, being: phases, generic tasks, specialized tasks and process instances [18].

According to Chapman et al. [19], the widest level consists of six phases, containing several generic tasks grouped together. These tasks are quite generalized, complete and stable as they aim to cover all the possible circumstances of a DM project.

The third and fourth level, is the specialized tasks and the process instances, being more individualized for each project. The third level describes how generic tasks should be achieved and, on the fourth level, actions, decisions and results obtained during the DM project are recorded [19].

This process model provides an overview of the life cycle of a DM project. This life cycle is infinite, does not end when a solution is implemented, as new business issues arise and new projects need to be carried out. The life cycle consists of six phases that are dependent, but do not have a specific sequence [19]. Figure 3 shows the six most important and frequent phases and dependencies.

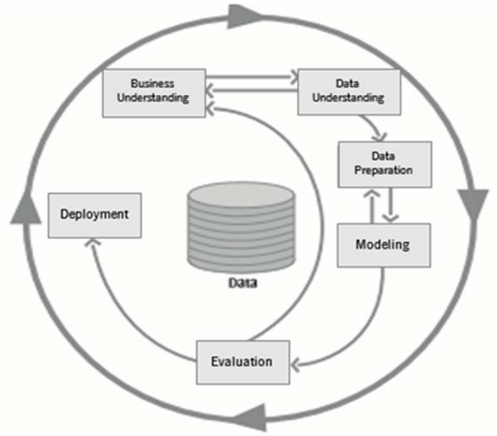


Figure 3 - CRISP-DM Lifecycle (retired from [30]*)*

According to several authors [19], [20], the phases can be described as the following:

- Business Understanding - Where a problem is created for the DM project, focusing on understanding business goals and requirements;
- Data Understanding - Consists of the initial data collected, and a first analysis of that data for understanding and identifying quality problems that may exist. This data analysis can be facilitated and even performed using classical statistics and its mathematical methods;
- Data Preparation - Aims to build a final data set from the data initially obtained. For this purpose, a selection, transformation and cleaning of the data for the modeling tools takes place. This task can also be supported by statistics and its methods;
- Modeling - Selection and application of various modeling techniques, and adjustment of the parameters to obtain optimized results;
- Evaluation – This stages ensures that the defined business objectives are achieved, the models generated in the previous phase are evaluated and the executed steps reviewed. In this activity, it is also possible to use statistical analysis to evaluate the results obtained;
- Deployment - This phase depends on the requirements previously defined for the project, either can be the creation of a report or the implementation of a DM process. Nevertheless, the main objective is always to present results and knowledge acquired, in an understandable way to the client.

The tasks that can be supported by statistics and statistical analysis are the phases of data comprehension, data preparation and evaluation. These phases include tasks such as description, exploration and verification of data quality, cleaning and data construction, and evaluation of results.

IV. Statistic importance

Statistics is a very important field of study, or even essential, for the existence of Data Mining (DM) and the proper effectiveness of your projects. Many of the activities of these projects are supported and facilitated by statistical methods and analysis. This is an important support for Data Science.

According to the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, one of the first steps in a DM project is data understanding, which incorporates the task of exploring data. According to Gorunescu [5], this exploration requires the analysis that resorts to the human capacity to recognize patterns using knowledge gained from previous experiences. Most of the time these patterns are not only found with our eyes, but mostly by using statistical analysis techniques. These techniques allow to gather and summarize a high number of data characteristics, highlighting the main and most influential data. Another two phases of DM projects that can be supported with the use of statistics are the data preparation phases, which include tasks such cleaning, constructing, and evaluating data, which includes evaluating results as well.

The same book, Gorunescu [5], says that exploratory analysis is the statistical part that deals with reviews, communication and use of data. Using statistics to analyze the variability of data, verify dependencies between variables and analyze censored data, namely, data that, for some reason, cannot be clearly stated. Also, used to perform regression analysis that uses mathematical models to connect between variables of response / result and variables of predictive / independent.

Rygielski, Wang, & Yen [21] and Tufféry [4] say that DM is a sophisticated data research feature that uses statistical algorithms to discover patterns in data, and is based on inferential statistics. The author Gorunescu [5] even goes so far as to say that without statistics, DM would not exist, justifying that classical statistics have brought well-defined techniques to identify systematic relations between different variables when there is insufficient information about them. He also added, computational methods and data visualization techniques. Computational methods, such as descriptive statistics (distribution, mean, median, standard deviation, etc.), frequency tables, multivariate exploratory techniques (cluster analysis, factorial analysis, etc.), among others. And visualization techniques, such as histograms and graphs of all kinds. With all this, it is possible to understand that statistics and statistical analysis are directly related to DM techniques and those directly influence the results of their projects. Another area with a similar goal is Data Science. This is an area related to the gathering, preparation, analysis, visualization, management and preservation of large information sets [9], [10].

With the description made in the chapter above of Data Science it is very easy to see that statistics and decision making are directly related by the organizations. Statistics are part of their roots, that is, mathematical and statistical capabilities are crucial for Data Science researchers. As for the relation with decision making, according to Stanton [9], projects of this type aim to build data architectures that bring the data initially collected to the managers. To do this, the data must undergo a series of analysis and transformations, becoming rigorous and with high quality information, helping managers in decision making.

From these three areas the first to emerge was statistic. According to [22], although statistical methods have been used for quite long time, the term "statistics" only originated around the year 1749. The second term to arise was Data Science, which, taking into account the Forbes website [23], appeared around the year 1960 in a book published by Peter Naur. Finally, the term Data Mining appeared in the scientific community, in the 1990s [24]. As we can see in Figure 4, the statistical area is one of the roots for the remaining areas presented and the DM field of study, although it was the last one to appear, it has integrated into Data Science.

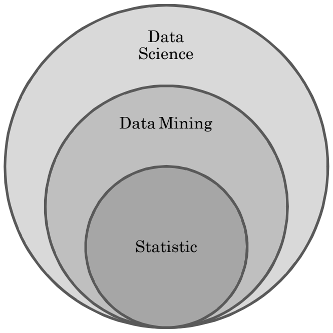


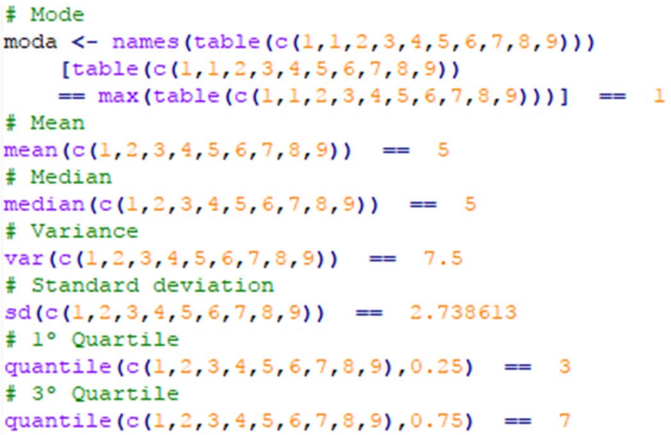
Figure 4 - Data Science Components

V. Tools for Statistical analysis and data mining

Nowadays there are several tools capable of supporting Data Mining (DM) projects and at the same time performing statistical analysis of the data. One of these tools, which according to KDnuggets [1] [25] is the most used and popular, is the tool R.

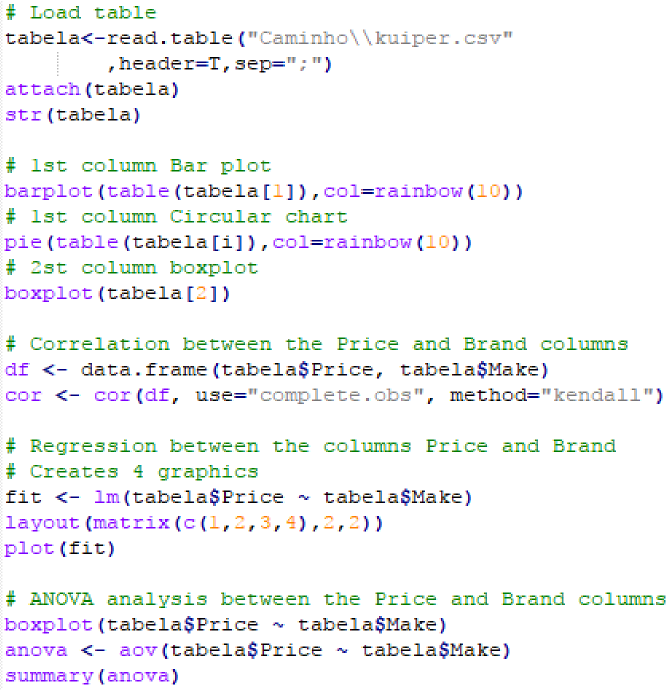
R is a free platform for data analysis, calculation and display of graphics and development activities of DM software. R is a language optimized primarily for calculations of matrix-based, as well as implementations of many machine learning algorithms [26]. According to Rangra & Bansal [27] and Zupan & Demsar [28], R is a tool widely used by professional statistics to perform complex data analysis, since R has a very extensive statistical library that covers all analytical needs and produces graphs with good quality.

Figure 5 and Figure 6 illustrate some of the code needed to perform certain statistical methods, nevertheless, how to do a quick and simple analysis of the data using the tool R.



```
# Mode
moda <- names(table(c(1,1,2,3,4,5,6,7,8,9)))
    [table(c(1,1,2,3,4,5,6,7,8,9))
    == max(table(c(1,1,2,3,4,5,6,7,8,9)))]  ==  1
# Mean
mean(c(1,2,3,4,5,6,7,8,9))  ==  5
# Median
median(c(1,2,3,4,5,6,7,8,9))  ==  5
# Variance
var(c(1,2,3,4,5,6,7,8,9))  ==  7.5
# Standard deviation
sd(c(1,2,3,4,5,6,7,8,9))  ==  2.738613
# 1° Quartile
quantile(c(1,2,3,4,5,6,7,8,9),0.25)  ==  3
# 3° Quartile
quantile(c(1,2,3,4,5,6,7,8,9),0.75)  ==  7
```

Figure 5 - Statistical Analysis (1)



```
# Load table
tabela<-read.table("Caminho\\kuiper.csv"
        ,header=T,sep=";")
attach(tabela)
str(tabela)

# 1st column Bar plot
barplot(table(tabela[1]),col=rainbow(10))
# 1st column Circular chart
pie(table(tabela[i]),col=rainbow(10))
# 2st column boxplot
boxplot(tabela[2])

# Correlation between the Price and Brand columns
df <- data.frame(tabela$Price, tabela$Make)
cor <- cor(df, use="complete.obs", method="kendall")

# Regression between the columns Price and Brand
# Creates 4 graphics
fit <- lm(tabela$Price ~ tabela$Make)
layout(matrix(c(1,2,3,4),2,2))
plot(fit)

# ANOVA analysis between the Price and Brand columns
boxplot(tabela$Price ~ tabela$Make)
anova <- aov(tabela$Price ~ tabela$Make)
summary(anova)
```

Figure 6 - Statistical Analysis (2)

VI. Case Study

The objective of this case study is to apply the statistics to the best scenario resulting from a Data Mining (DM) process, and read this same scenario, making an evaluation of the results using the statistic and tool R. To perform this case study a dataset extracted from the Intensive Care Units (ICU) of the hospital Santo António in Porto was used. The dataset contains information about the vital signs of some patients collected over a period of 120 hours. Table 1 presents some information about the dataset used.

Table 1 - Data Information

Columns	Type	Data
HORA	Integer	Time of collection. Between 2 and 120.
ADMINT	Factor	Patient admission type. Value of p (Programmed) or u (Urgent).
AGE	Integer	Age of patient in classes. Between 1 and 4.
ADMINF	Integer	Origin of patient admission. Between 1 and 7.
RESPIRAT_MIS	Integer	Classified value by expert. Between 0 and 1.
COAGULAT_MIS	Integer	Classified value by expert. Between 0 and 1.
RENAL_MIS	Integer	Classified value by expert. Between 0 and 1.
HEPATIC_MIS	Integer	Classified value by expert. Between 0 and 1.
CARDIO_MIS	Integer	Classified value by expert. Between 0 and 1.
EC_AC_TOT_BIN	Integer	Critical event in real time from the patient. Between 0 and 67.

[1] http://www.kdnuggets.com/

EC_AC_BP_BIN	Integer	Critical event in real time from the patient. Between 0 and 61.
EC_AC_HR_BIN	Integer	Critical event in real time from the patient. Between 0 and 12.
EC_AC_TOT_EC_MAX_BIN	Numeric	Max critical events real time. Between 0 and 1.053.
EC_AC_TOT_HORAS_BIN	Numeric	Event per hour in real time. Between 0 and 1.5.
EC_AC_BP_EC_MAX_BIN	Numeric	Max critical events real time. Between 0 and 0.062.
EC_AC_BP_HORAS_BIN	Numeric	Event per hour in real time. Between 0 and 1.333.
EC_AC_HR_EC_MAX_BIN	Numeric	Max critical events real time. Between 0 and 0.032.
EC_AC_HR_HORAS_BIN	Numeric	Event per hour in real time. Between 0 and 0.5.
EC_AC_O2_HORAS_BIN	Numeric	Event per hour in real time. Between 0 and 0.75.
EC_AC_O2_BIN	Integer	Critical event in real time from the patient. Between 0 and 47.
OUTCOME_MIS	Integer	Classified value by expert. Between 0 and 1.
EC_AC_HR2	Integer	Accumulated critical event. Between 0 and 5.
EC_AC_HR_HORAS2	Integer	Accumulated critical events by hour. Between 0 and 7.
EC_AC_BP2	Integer	Accumulated critical event. Between 0 and 7.
EC_AC_BP_HORAS2	Integer	Accumulated critical events by hour. Between 0 and 7.
EC_AC_O2	Integer	Accumulated critical event. Between 0 and 7.
EC_AC_O2_HORAS2	Integer	Accumulated critical events by hour. Between 0 and 7
EC_AC_TOT2	Integer	Accumulated critical event. Between 0 and 7.
EC_AC_TOT_EC_MAX2	Integer	Max critical events. Between 0 and 6.
EC_AC_TOT_HORAS2	Integer	Accumulated critical events by hour. Between 0 and 7.

For more information about the dataset, you can review the article written by Portela, Santos, Silva, Abelha e Machado [29].

During the research process on the DM several feature select algorithms were used, using caret package. These models were C.50, Nnet, JRip, Earth, GcvEarth, RF and Rpar. Because these models use built in feature selection its variable importance's were retrieved from the models and several scenarios were created. The rules for creating the scenarios were:

For each model per attribute:

2 → 90% importance High

Among all models per attribute:

1 → 70% importance per attribute on any model Medium

The scenarios generated were:

- High Rpart – ADMINF, AGE, RENAL_MIS and HORA;
- High Gcvearth – ADMINF, AGE, HEPATIC_MIS and HORA;
- High Earth – ADMINF, AGE and HORA;
- High JRip – ADMINF, ADMINT, AGE and HORA;
- High – ADMINF, AGE, COAGULAT_MIS, HEPATIC_MIS, RENAL_MIS and HORA;
- Medium all – ADMINF, ADMINT, AGE, CARDIO_MIS, COAGULAT_MIS, EC_AC_BP2, EC_AC_BP_BIN, EC_AC_BP_HORAS2, EC_AC_BP_HORAS_BIN, EC_AC_HR2, EC_AC_HR_BIN, EC_AC_HR_HORAS2, EC_AC_HR_HORAS_BIN, EC_AC_O22, EC_AC_O2_BIN, EC_AC_O2_HORAS2, EC_AC_O2_HORAS_BIN, EC_AC_TOT2, EC_AC_TOT_BIN, EC_AC_TOT_EC_MAX2, EC_AC_TOT_EC_MAX_BIN, EC_AC_TOT_HORAS2, EC_AC_TOT_HORAS_BIN, HEPATIC_MIS, HORA, RENAL_MIS and RESPIRAT_MIS.

The first scenario includes all attributes of the dataset. All scenarios were recorded on the database for further use.

For each of these scenarios the following algorithms were used:

- C50 – C5.0
- Nnet – Neural Network
- J48 – C4.5
- LMT – Logistic Model Trees
- JRip – Rule-Based Classifier
- E1071 RandomForest – RandomForest
- Earth – Multivariate Adaptive Regression Spline
- GcvEarth – Multivariate Adaptive Regression Splines
- Caret RF – randomForest
- Gbm – Stochastic Gradient Boosting
- E1071 SVM – Support Vector Machines
- Adaboost – AdaBoost Classification Trees

Outcome was the target used to predict if a patient will or will not die. 84 models were created each using 10Fold CV (7 scenarios x 12 algorithms). In order to determine the best model, the following fitness function was defined: ((specificity + sensibility + accuracy)/3) were all the measures needs to be higher to 80%.

The best model (C5.0) achieved a result of 89.5 ((85.3 + 93.1 + 90.1) /3) having the all the features >90% in scenario medium all.

After the choice of the scenario through the DM was made the analysis of the same through the statistic and the tool R. The first step was to analyze three attributes, which were more relevant to the final result of the DM process, taking into account the hours and number of observations. These attributes were ADMINF, ADMINT e AGE. These attributes have an impact of 100% in all the scenarios, i. e., all the scenarios need these variables to provide the result above mentioned.

As we can see in Figure 7, the analysis of the attribute ADMINF considering the hours tells us that the patients coming from service 2 do not stay 120 hours hospitalized, that is, it is the only service where the patients finish the internment sooner. Based on statistical analysis we can still say that most patients come from service 1, as you can see in Table 2.

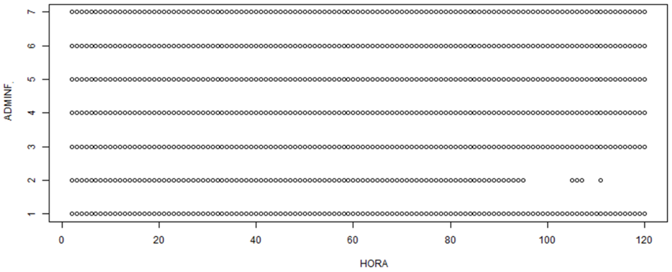


Figure 7 - Analysis of the ADMINF Attribute per Hour

Table 2 - Frequency Table of the ADMINF Attribute

Classes	Cases	Percentage (~)
1	11127	45.45%
2	158	0.65%
3	4004	16.4%
4	2876	11.75%
5	1499	6.1%
6	830	3.4%
7	3980	16.25%

In general de analysis are quite similar. Analyzing Figure 8 we can see that the urgent cases are in greater numbers, and we might think that they are also the ones that leave the intensive care system earlier. But if we analyze in percentage terms, in the first 48 hours 20.2% (44 cases) of the patients admitted to the emergency and 30.8% (20 cases) of the programed cases leave the intensive care unit, that is, although the 20.2% correspond to a larger number of patients, in percentage terms, patients in programed cases leave intensive care earlier.

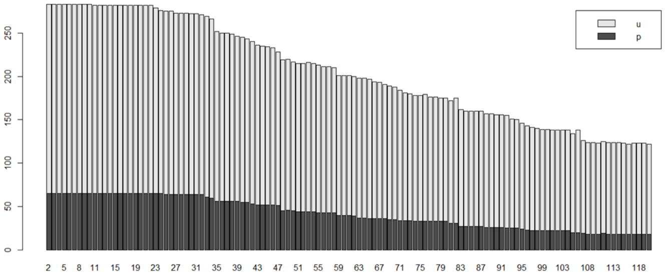


Figure 8 - Analysis of the ADMINT Attribute per Hour

Analyzing now the AGE attribute, corresponding to the age of the patients, it was possible to verify that the class 2 was verified a greater number of times (Table 3). Taking into account

Table 4 (retired of [29]) it is verified that the age range of the largest number of patients is between 47 and 65 years.

Table 3 - Frequency Table of the AGE Attribute

Classes	Cases	Percentage (~)
1	2907	11.9%
2	9032	36.9%
3	5229	21.4%
4	7306	29.8%

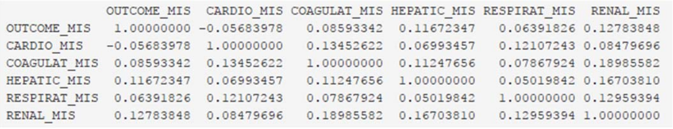
Table 4 - Age Classes

Classes	Min	Max
1	18	46
2	47	65
3	66	75
4	76	130

Taking a last attribute, CARDIO_MIS, which gives us information about cardiovascular signs through values 0 and 1, we can only say that the number of occurrences of the value 0 were 15189 (62.06%) and of the value 1 were 9285 (37.94%).

Turning now to the analysis of more than one attribute together, the attributes selected for this analysis were the RESPIRAT_MIS, COAGULAT_MIS, RENAL_MIS, HEPATIC_MIS, CARDIO_MIS and OUTCOME_MIS.

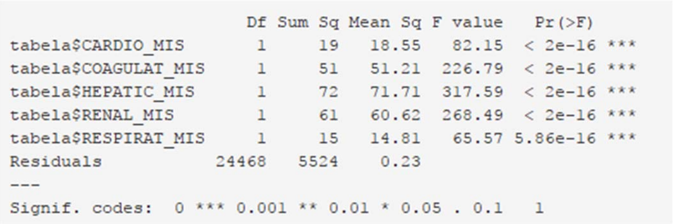
The first analysis is correlation, as we can see in Figure 9. This analysis tells us which attributes are more closely related to each other, more dependent, that is, 1 means that there is a lot of correlation and -1 means no correlation, 0 is neutral. Taking this into account, we can verify that the most correlated attributes are COAGULAT_MIS and RENAL_MIS, following the attributes RENAL_MIS and HEPATIC_MIS.



```
             OUTCOME_MIS  CARDIO_MIS COAGULAT_MIS HEPATIC_MIS RESPIRAT_MIS  RENAL_MIS
OUTCOME_MIS   1.00000000 -0.05683978   0.08593342  0.11672347   0.06391826 0.12783848
CARDIO_MIS   -0.05683978  1.00000000   0.13452622  0.06993457   0.12107243 0.08479696
COAGULAT_MIS  0.08593342  0.13452622   1.00000000  0.11247656   0.07867924 0.18985582
HEPATIC_MIS   0.11672347  0.06993457   0.11247656  1.00000000   0.05019842 0.16703810
RESPIRAT_MIS  0.06391826  0.12107243   0.07867924  0.05019842   1.00000000 0.12959394
RENAL_MIS     0.12783848  0.08479696   0.18985582  0.16703810   0.12959394 1.00000000
```

Figure 9 - Correlation Analysis

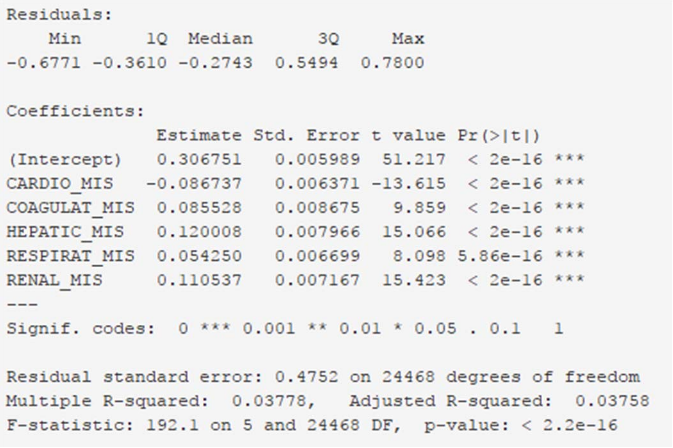
A second analysis was the analysis of variance (ANOVA). This analysis verifies the significance of each attribute and sees if it is the largest that has more importance. Considering the results, visible in the Figure 10, it is possible to verify that all attributes are quite significant for the result, with RESPIRAT_MIS being the least significant attribute because it has a higher p-value.



```
                        Df Sum Sq Mean Sq F value   Pr(>F)    
tabela$CARDIO_MIS        1     19   18.55   82.15  < 2e-16 ***
tabela$COAGULAT_MIS      1     51   51.21  226.79  < 2e-16 ***
tabela$HEPATIC_MIS       1     72   71.71  317.59  < 2e-16 ***
tabela$RENAL_MIS         1     61   60.62  268.49  < 2e-16 ***
tabela$RESPIRAT_MIS      1     15   14.81   65.57 5.86e-16 ***
Residuals            24468   5524    0.23                     
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Figure 10 - ANOVA Analysis

A final analysis of the various attributes was linear regression (Figure 11). From this analysis, we can extract the estimates for the model intercept that is 0.306751 and, through estimate column, the coefficient measuring the slope of the relationship with all used attributes. It's still possible obtain information about standard errors of these estimates in the Std. Error column. Finally, we can also verify the significance of each attribute as in the ANOVA analysis.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-0.6771 -0.3610 -0.2743  0.5494  0.7800 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.306751   0.005989  51.217  < 2e-16 ***
CARDIO_MIS   -0.086737   0.006371 -13.615  < 2e-16 ***
COAGULAT_MIS  0.085528   0.008675   9.859  < 2e-16 ***
HEPATIC_MIS   0.120008   0.007966  15.066  < 2e-16 ***
RESPIRAT_MIS  0.054250   0.006699   8.098 5.86e-16 ***
RENAL_MIS     0.110537   0.007167  15.423  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4752 on 24468 degrees of freedom
Multiple R-squared:  0.03778,   Adjusted R-squared:  0.03758 
F-statistic: 192.1 on 5 and 24468 DF,  p-value: < 2.2e-16
```

Figure 11 - Linear Regression Analysis

VII. Conclusion

Classical statistic has been around for a quite a long time and are still used every day in a variety of situations, being very important when it comes to organizational decision-making. Data Mining (DM), by contrast, is a technological development still emerging and with a long way to go, but already with a recognition and importance for organizations well known.

These two areas are directly related and statistics represents a strong root for DM and for support of some of the phases of their projects. This article consists of highlighting, explaining and proving this relation and the importance that statistics have for the area of DM and, consequently, projects. To do so, it uses a methodology that provides a framework to help execute DM projects, the Cross Industry Standard Process for Data Mining (CRISP-DM), verifying that some of the phases are used for statistical methods to achieve their results. This paper also uses Data Science and decision making to reinforce the importance of statistics in general.

This article also offers a quick study of one of the most commonly used tools for statistical analysis of data, R. Together with this brief study, we present an example of how to perform a simple statistical analysis to a dataset using code in R.

To conclude, with regards to the evolution and developments of the DM area and related tools, a very significant growth is expected, with more and better tools, used by organizations and successful cases. The evolution of technology and computational power will also favor the area of DM, allowing the incorporation of new capabilities and concepts to related tools.

ACKNOWLEDGES

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013

VIII. References

[1] A. Goodman, C. Kamath, and V. Kumar, “Data analysis in the 21st century,” *Stat. Anal. Data Min.*, vol. 1, no. 1, pp. 1–3, 2008.

[2] D. Singpurwalla, *A Handbook of Statistics An Overview of Statistical Methods*. 2013.

[3] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining - Practical Machine Learning Tools and Techniques*. ELSEVIER, 2011.

[4] S. Tufféry, *Data Mining and Statistics for Decision Making*, WILEY. WILEY SERIES IN COMPUTIONAL STATISTICS, 2011.

[5] F. Gorunescu, *Data Mining - Concepts, Models and Techniques*, Springer. Intelligent Systems Reference Library, 2011.

[6] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. Cambridge, 2001.

[7] R. Wirth and J. Hipp, “CRISP-DM : Towards a Standard Process Model for Data Mining,” *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000.

[8] M. Loukides, “What is Data Science?,” *O’Reilly Radar Rep.*, 2010.

[9] J. Stanton, *An Introduction to Data Science*. 2013.

[10] D. Baier, R. Decker, and L. Schmidt-Thieme, *Data analysis and decision support*. Springer, 2005.

[11] G. Casella, S. Fienberg, and I. Olkin, *Springer Texts in Statistics*, vol. 102. Springer, 2006.

[12] M. Longnecker and R. Ott, *An introduction to statistical methods and data analysis*. Brooks/Cole, Cengage Learning, 2010.

[13] D. Steinley and S. Wasserman, “Introduction: special issue of statistical analysis and data mining on networks,” *Stat. Anal. Data Min.*, vol. 4, no. 5, pp. 459–460, 2011.

[14] G. Bohm and G. Zech, *Introduction to statistics and data analysis for physicists*. 2010.

[15] D. W. Zimmerman and B. D. Zumbo, “Rank Transformations and the Power of the Student T-Test and Welch T-Test for Nonnormal Populations With Unequal Variances,” *Can. J. Exp. Psychol.*, vol. 47, no. 3, pp. 523–539, 1993.

[16] P. C. de Araújo and C. A. A. P. Abar, “Sobre o Boxplot no GeoGebra,” *1ª. Conferência Lat. Am. GeoGebra*, pp. 13–21, 2012.

[17] O. Hammer, “PAleontological STatistics - Reference Manual,” 2016.

[18] IBM, “IBM SPSS Modeler CRISP-DM Guide,” p. 53, 2011.

[19] P. Chapman *et al.*, “Crisp-Dm 1.0,” *Cris. Consort.*, p. 76, 2000.

[20] A. Azevedo and M. F. Santos, “KDD, SEMMA and CRISP-DM: a parallel overview,” *IADIS Eur. Conf. Data Min.*, no. January, pp. 182–185, 2008.

[21] C. Rygielski, J.-C. Wang, and D. C. Yen, “Data mining techniques for customer relationship management,” *Technol. Soc.*, vol. 24, no. 4, pp. 483–502, 2002.

[22] H. Walker, *Studies in the History of the Statistical Method*. 1929.

[23] G. Press, “A Very Short History Of Data Science,” 2013. [Online]. Available: https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#4c40ae3b55cf.

[24] Rayli, “History of data mining,” 2015. [Online]. Available: https://rayli.net/blog/data/history-of-data-mining/.

[25] KDnuggets, “R, Python Duel As Top Analytics, Data Science software – KDnuggets 2016 Software Poll Results,” 2016. [Online]. Available: http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html.

[26] W. N. Venables and D. M. Smith, “An Introduction to R,” *R. Gentlem. R. Ihaka Copyr. c*, vol. 3.3.2, p. 105, 2016.

[27] K. Rangra and K. L. Bansal, “Comparative Study of Data Mining Tools,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 4, no. 6, pp. 2277–128, 2014.

[28] B. Zupan and J. Demsar, “Open-Source Tools for Data Mining,” *Clin. Lab. Med.*, vol. 28, no. 1, pp. 37–54, 2008.

[29] F. Portela, M. . Santos, A. Silva, A. Abelha, and J. Machado, “Towards Pervasive and Intelligent Decision Support in Intensive Medicine – A Data Stream Mining Approach,” *Elsevier Sci.*, p. 23.

[30] Rui Peixoto, Filipe Portela and Manuel Filipe Santos. Towards a Pervasive Data Mining Engine - Architecture overview. Advances in Intelligent Systems and Computing (WorldCist 2016 - Pervasive Information Systems Workshop). Volume 445, pp 557-566. ISBN: 978-3-319-31306-1. Springer. (2016).