International Journal of Advanced Engineering Research and Science (IJAERS)
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-8, Issue-3; Mar, 2021
Journal Home Page Available: https://ijaers.com/
Journal DOI: 10.22161/ijaers
Article DOI: https://dx.doi.org/10.22161/ijaers.83.26

Data Driven: An Overview and Practical Measures for Organizations

Alisson Paulo de Oliveira, Hugo Ferreira Tadeu Braga

Innovation Center, Fundação Dom Cabral
Avenida Princesa Diana, 760, Alphaville, Lagoa dos Ingleses, Brazil.

Received: 25 Nov 2020;

Received in revised form:

19 Jan 2021;

Accepted: 16 Feb 2021;

Available online: 25 Mar 2021

***Keywords*—Data-Driven, Prediction Models, Analytics, Big-Data, i4.0, Internet of Things, Artificial Intelligence.**

Abstract*—*Data is considered a primary resource for innovation. The existence of a large amount of available data, as well as technological tools capable of explore them, allows companies to extract information that can be used to create and implement new ideas and new projects. To this end, the details regarding the care that organizations should have with data are explored. The difficulties regarding the adoption of data-driven approach and some measures to implement this type of decision-making approach are also discussed. Some examples of data-driven approaches for diverse industries and products are shown. A real example of prediction model for decision making that is based on industrial data is also discussed. This example shows the difficulties in the preparation of data for the building of these models, which confirms that most of the time spent in the construction of predictive models it is due to this step. The use of the data-driven approach allows organizations to obtain superior results in their processes, thus becoming a tremendous competitive advantage and a special strategic factor in a highly competitive market, regardless of the field of activity.

I. INTRODUCTION

Business decisions are usually based on the instincts of leaders. However, if they were based on data transformed into reliable and high-level information, it would certainly be more assertive decisions. Lately, there has been a revolution in data collection, processing, and use. Organizations today can gather data in a detailed way to provide knowledge to their partner, consumers, and competitors **[1]**. In addition, a huge variety of devices have been designed with the ability to continuously collect data **[2]**. The technology called Internet of Things is a revolution in this area and refers to scenarios where network connectivity and computational capacity extend to objects, sensors, and everyday items, which are not considered computers **[3]**.

The Big-Data is one of the key technologies in the so-called Industry 4.0 or Fourth Industrial Revolution (See figure 1**[4]**). This has the potential to transform the industrial production system with the introduction of new technologies, such as autonomous robots, Internet of Things, Cloud Computing, Big-Data itself, Artificial Intelligence, Additive Manufacturing and Virtual Reality. And these technologies will increase connectivity, inside and outside organizations, by creating a network of physical systems known as Cyber-Physical systems **[5]**. The digitization that proliferates in society and the economy has been causing marked changes in the conditions that prevail in this market. Consumer habits are changing rapidly, such as product customization according to consumer wishes. These changes are creating a vast number of potential opportunities in terms of new business. Organizations can discover new ways to provide, create and capture value. The vast amount of data

generated is not only the basis for an organization's existence, but this data offers organizations the chance to make sense of it, extracting information and using it for their own benefit, to simplify the decision-making process**[6]**.

Fig.1: Characteristics of each industrial revolution.

The latest innovations are based on the use of technology, most often in the form of digital innovation or digitization. Digital innovation means the combination of physical and digital components to produce new products, which offer different services, which will dissolve the frontiers of industry and products. Digital innovation not only implies a change in technology, but also indicates a change in the dynamics of relationships within businesses and markets **[7]**. In an innovative approach on the use of information and decision-making within organizational environments, carrying out activities with a higher degree of complexity requires the processing of a large amount of data to provide information. Thus, it is necessary for organizations to design information processing. The improvement of organizational performance, which includes the reduction of costs, is a function of the use of technologies that allow the massive collection of data, that generates information, in addition to facilitating the flow of information within an organization **[8]**. The dramatic improvements in data collection, storage and processing capabilities have created opportunities for businesses in recent years. In particular, the collection and reliability of data that can be used for managerial activities, like data-driven decision making. Evidence suggests that better performance in a sample of public companies is associated with the intensive use of data **[9]**. Digital technologies have stimulated an exponential growth in the availability of data, which has generated great demand for adequate processing to extract information from this data. The Big-Data were created due to this high availability of data **[10], [11]**. Evidently, the data started to have an enormous value in this new context, starting to be the new oil **[12]**.

This paper is divided into five sections. The following sections are divided as follows: Section 2 discuss the main topics relating to an organization that wants to become Data-Driven. Section 3 discuss some examples of Data-Driven Approaches for Solving real problems. Section 4 discusses the organization of data that will be used on the construction of a predictive model based on Artificial Intelligence (Artificial Neural Networks). Section 5 presents the conclusion and final considerations.

II. TOPICS TO BE CONSIDERED WHEN BUILDING A DATA-DRIVEN ORGANIZATION

The following topics, 2.1 to 2.5, aims to show some topics to be considered when an organization wants to be data-driven from a practical point of view. This type of organization is one that makes decisions based on data on an ongoing basis. It is an organization that build tools, skills and maintain a data-based culture **[13].**

2.A. Collection and accessing the Data.

Data is a key enabler for smart manufacturing and data-driven organizations. However, data in its raw form is not so useful in providing good information. These need to be "turned" into something more valuable, and this is usually done in several stages. The different stages of data collection, storage, analysis, and visualization can be referred to as the "data life cycle" **[14].**

The collection of the data is the beginning of all. The data collected must be the right data and must be relevant to the solution of the issue that the company wants to solve. Such data must be easy to obtain, must be accurate, clear, and impartial and, above all, reliable. Unfortunately, such data can lead to numerous inaccuracies and errors, which requires a previous assessment and cleaning step, to make them reliable. If used in the way they were obtained, there is a great chance of generating false information and leading to wrong actions in companies, which could result in losses. This step of cleaning and processing the data is very time consuming, difficult, and expensive, and can require 80% of the time of a data scientist, whereas the construction of the models, the analysis of the results and conclusions take only 20% of the time. The data alone is not enough. And a small amount of clean and reliable data can be more valuable than a huge amount of raw data that has not been cleaned or treated **[13], [15]**.

There are numerous pre-requisites when it comes to data quality **[13]**:

- Accessible: Data, as well as analysis tools, must be accessible to analysts;
- Accurate: Such data must represent the true value or state of the entity. Badly typed values, or badly collected, do not represent accurate data;

- Coherent: Refers to the combination of certain data with other data in a precise way, bringing a more complete and real image to a given situation. For this, its correct identification is essential;
- Complete: There can be no gaps, or data not saved in a databank;
- Consistent: The data is the same regardless of the source consulted. Otherwise, there must be a reliable source and the one with inconsistencies must be evaluated and corrected;
- Defined: There must be a well-understood and unambiguous meaning for each individual data field. If a data dictionary is available that explains the origin and meaning of that data, the quality of the data will improve;
- Relevant: The data must be relevant for the analysis of the phenomenon of interest;
- Reliable: Refers to how complete this data is as well as its accuracy, if the correct information is provided;
- Timely: Refers to the shortest possible time between data collection and its availability for analysis. The smaller the better.

Having accurate, timely and relevant data is not a sufficient condition for an organization to be considered as Data Driven. The next session will discuss about reports, analysis, the difference between them and the main professional roles necessary to an organization to be data driven.

2.B. Reports, Analysis, and people involved with Analytics.

Reports and alerts are not sufficient characteristics for an organization to be considered as Data Driven. Something more is needed. However, both reports and alerts are of great importance, especially reports. There are legal demands for reports, and they may not necessarily represent an internal aspect aimed at improving the business. Analysis, otherwise, allows organizations to find out the real causes of why a process is changing, and make some actions to know why it is changing. The definitions of reports and analysis are **[13]:**

- Reports: They are the process of preparing informative summaries from organized data. They aim to monitor the performance of several areas of an organization;
- Analysis: **It is about transforming data**, a form of assets, into information that will increase the competitiveness of organizations, which will allow better business decisions.

The reports are descriptive whereas the analysis is prescriptive. Table 1 shows the difference between reports and analyzes **[13].**

Table 1. Main attributes of reports versus analyzes.

Reports	**Analysis**
Descriptive	Prescriptive
What?	Why?
Backward looking	Forward looking
Raise questions	Answer questions
Data → information	Data + information → insights
Reports, dashboards, alerts	Findings, recommendations, predictions
No context	Context + storytelling

Analysis always seeks to answer why a problem has occurred and is not limited to just describing it, as occurs in the Reports. Instead of raising questions, the analysis already seeks to answer them and with the use of the information obtained from the data, there is an insight into the problems, which leads to their anticipation through recommendations and forecasts. Therefore, in a problem-solving context, the Analysis can solve them, the opposite of what happens with the Reports, which basically describes the problem.

A useful structure for understanding the analysis is shown in Table 2 **[13]**.Report (A) and alert (B) are not based on data. There is simply an assertion that something out of the ordinary has happened in the past or at that very moment. The reason for the event is not explained, nor its causes, and there is no recommendation on how to avoid a recurrence of the situation. (D) is close to Data-Driven since it already uses modeling and design of experiments. (E) and (F) represent what is Data-Driven, but only if the information is used since its basis is the understanding of the phenomenon and only this understanding allows the formulation of a plan or recommendations. (C) is a danger zone since the extrapolation may not allow the necessary precision. Therefore, the causal model should be pursued **[13]**.

As shown in table 1, the analysis allows anticipating problems through recommendations and forecasts. From table 2 we have that only the approaches (E) and (F) can be considered as Data-Driven and this only occurs due to the ability of prediction possible with the use of predictive models. Thus, organizations must seek predictive capacity to anticipate their problems and consequently increase their competitive potential.

Table 2. Hypothetical issues addressed by the analysis. D) is a valuable analysis, but only E) and F) are data driven and if, and only if, the information is used.

	Past	Present	Future
Information	A) What happened? **Reporting.**	B) What is happening now? **Alerts.**	C) What will happen? **Extrapolation.**
Insight	D) How and why did it happen? **Modeling, experimental Design.**	E) What is the next best action? **Recommendation.**	F) What is the best/worst that can happen? **Prediction, optimization, simulation.**

Experts who make use of data analysis tools, who know what the necessary analyzes are, and the meaning of the results from these analyzes can work in several areas, such as: Business Analysis, Data Analysis, Big-Data Analysis and Data Science **[16]**.A data-driven organization is likely to have a variety of analyst roles, usually organized into multiple teams **[13]**.

- Data analyst: They are generally more focused on collecting and preparing data. The specific roles in an organization depend on the size, maturity, domain, and market of the organization. The delivery will be a mix of reports and analysis. In addition to the breadth of the domain, analysts vary widely in their level of technical skills;
- Data engineers and analytical engineers: Responsible for obtaining, cleaning, and adjusting the data and making it available in such a way that analysts can access and analyze it;
- Business analysts: They act as the interface between business stakeholders and the technology department. Its function is to improve business processes or help to design and develop new or improved features;
- Data scientists: They are the professionals most inclined to mathematics or statistics, usually with advanced degrees (usually in quantitative disciplines, such as mathematics, science, and computer science) and developed coding skills. They divide their time working on a variety of projects, from building statistical models and algorithms;
- Statisticians: These are professionals who focus on statistical modeling across the organization. They are involved not only in the analysis, but in the research design, experiments, and collection protocols to obtain the raw data. The job may involve increasing the quality of a new data source;
- Quantitative Analysts: They are mathematically qualified professionals working in the financial services sector, securities modeling, risk management and stock movements on the buy and sell side of the market;
- Accountants and Financial Analysts: These are professionals with a focus on internal financial statements, auditing, forecasting, business performance analysis, etc.;
- Experts in data visualization: These are the professionals who create infographics, panels, and other design assets.

A data-driven organization will have a team with an emphasis on analytics with different roles, as shown above, and people with complementary skills. As in any other area of an organization, it is necessary to know the team's skills and thus strengthen the fields of knowledge that are sometimes absent and sometimes weak **[13]**. The next session will go deeper on the definitions of data-analysis and the maturity levels on this activity.

2.C. Data-Analysis and Maturity

Data analysis is the basis of what is also known as process mining. This is a field of studies that deals with the discovery of processes, verification of compliance and improvement using data from that process. By considering records (or labeled data) in large quantities, it is possible to discover an accurate representation of the model. Such a model can be used to discover possible deviations from the process from the use of new data not yet processed. Such models can be improved through the use and learning from new data **[14]**. Data-Analysis concerns the transformation of data (assets) into competitive perceptions (information) that will serve as beacons for decisions and actions. There is a total of six types of analysis, starting with the simplest to the most complex **[13]**:

- Descriptive: Describes a data set in a quantitative way, however it does not describe anything about the data population from which that set originates. Its objective is only to describe numerically the main characteristics of the sample;
- Exploratory: Use of graphics to examine and visualize data, which helps in viewing the big picture. It is possible to gain insights from exploratory analysis;
- Inferential: The purpose of this type of analysis is to infer information (Parameters, distributions, or relationships) about the broader population from which the sample originates. The hypothesis test, used

to test and analyze the understanding of the underlying mechanisms, can be started from inferential analysis;

- Predictive: This type of analysis is based on inferential analysis and aims to learn about the relationships between the variables in a training data set and thereby develop a statistical model capable of predicting output values for new data, whether these are incomplete and future. **The potential of predictive analytics is enormous with a wide range of applications**;
- Causal: Type of analysis where a series of experiments are carried out and in which the greatest possible number of variables is controlled. In these experiments, only one of the variables is changed at a time and the results are evaluated. Such experiments provide a causal understanding and with greater depth of the system being analyzed, allowing greater understanding and how to act to influence its results, optimizing a system for example;
- Mechanistic: It is a type of analysis that allows an understanding of a system at a high level of depth. It is derived from studies of a stable system with many experiments over many years.

Depending on the activity one or more types of analysis can be used within the same organization. For example, the R&D area can prefer to use a mechanistic approach instead of a predictive approach that can be more useful for a financial area.

There are 8 levels of maturity in Data Analysis **[13]**:

[1] Standard Reports: What happened? When happened?

[2] Ad-Hoc reports: How much? How often? Where?

[3] Detailed query (Or online analytical processing): Where exactly is the problem? How do the answers can be found?

[4] Alerts: When the reaction is necessary? What actions are needed now?

[5] Statistical Analysis: Why is this happening? What opportunities are missed?

[6] Forecasting: What happens if these trends continue? How much is needed? When will it be necessary?

[7] Predictive modeling: What happens next? How will it affect the business?

[8] Optimization: How do it better? What is the best decision for a complex problem?

One of the possible ways to interpret these 8 levels of maturity in data analysis is to think that the maximum level at which the organization is engaged is positively correlated with the level of commitment, investment and usefulness of the analysis and the analytical competitiveness. As the level rises, there is a greater sophistication in the use of data, with evident improvement in the causal analysis and definition of countermeasures in the solution of problems. Certainly, business results will be substantially better in a level 8 organization compared to a level 1 organization.

2.D. Metrics and Storytelling with the data

Data-driven organizations, like any organization, need to define their business strategy and create a set of metrics, the Key Performance Indicators (KPI's). The key performance indicators (KPIs) are the set of metrics linked to the organization's strategic objectives. They help to define and track the direction the company is taking and to achieve its goals. It is extremely important that KPI's **[13]**:

- Are clearly defined;
- Be measurable;
- Have goals;
- Be visible;
- Reflect what the organization is trying to achieve.

Each company needs to choose and adapt its set of KPIs for its sector, its specific business model, its life cycle stage, and its specific objectives. KPIs will tend to cover all major areas of the business and any parts of the business that are the specific strategic focus for that period, usually one year. These KPIs are targeted at business divisions and it is possible that each division has specific additional KPIs. At the end, there are operational and diagnostic indicators that monitor tasks, programs, projects, and even strategic process variables. Thus, such indicators need to be well designed, to reflect what is really happening in organizations. There are several prerequisites when choosing or designing a metric **[13]**:

- Simple: They feature the ease of definition, the ease of disclosing to people, which allows greater understanding and less confusion. Other advantages: Simplicity in implementation, less likelihood of being incorrectly calculated and ease of comparison between teams and even between organizations;
- Standardized: Refers to the use of standard metrics. It allows greater understanding for new colleagues from other organizations, in addition to facilitating the comparison between organizations through benchmark studies;
- Accurate: The average numerical value must be like the true underlying average value. Inaccurate metrics have a bias so that their average differs from the true average in a constant or systemic way;

- Precise: Metrics must be precise. Similar values should be returned if repeated under the same conditions;
- Relative versus absolute: Absolute or relative metrics can generate a quite different image of the same scenario, so they must be chosen wisely and appropriate to what is intended to be shown;
- Robust: The metrics must be insensitive to individual extreme values, that is, there can be no significant variation of the values solely due to a single occurrence;
- Direct: Metrics should directly measure the process of interest.

Three questions must be considered about storytelling with the data **[13]**:

[1] What is the objective? One must keep in mind what is wanted with this data and what is expected to happen;

[2] Who is the audience? Is it an audience that has some knowledge about data? What is the level of expectations, interest, motivation, and availability of this audience?

[3] What is the medium to be used? How will the data be presented?

These three questions are essential to better define how the means of transmission of the message will be prepared (Type of presentation, style, depth) aiming at the greatest impact on the audience. The greatest chance of knowing the dominant patterns in the data is through a very well-planned experiment, the choice of good metrics and, mainly, through a well-defined question. And the analyst's job is to find among the data the cleanest and most important standards, interpret them and translate them in a way that has an impact on the business. However, there is more than one possible potential interpretation of these data **[13].**

The next session will discuss the characteristics and some cultural aspects of a truly data-driven organization.

2.E. Characteristics of truly data-driven organizations and decision-making

In times of Big-Data Analytics, leadership must act as an agent of change within organizations **[17]**, constantly dealing with challenges that involve understanding the benefits and availability of data, the development of analytical skills and data integration in organizational culture **[18]**. Leading means improving productivity in organizations **[19]** as well as positively connecting people **[20]**. Leaders must develop an analytical mindset to transform organizations into a decision-making environment that uses data in a very local way **[21]**. The organizations that are truly data-driven have the following characteristics **[13]**:

- Such organizations may be testing on an ongoing basis. These tests may include tests with users, where real consumers or users give feedback on new attributes or products;
- **A truly data-driven organization has a mindset of continuous improvement**. They frequently optimize its main processes. And this occurs from the realization of careful analyzes as well as the construction of mathematical or statistical models and the use of these for simulations;
- Such organizations may be involved in predictive modeling. But, even more important, it is the use of model errors as well as other lessons learned in improving the predictive capacity of these models;
- A data-driven organization will certainly guide its decisions using a set of weighted variables. The data for each set of variables that are of interest must be collected and the weights between them must be determined to allow the generation of a leadership decision that is reliable.

A truly data-driven organization will have at least one of these characteristics, looking to the future, where the data is first-class citizens. An organization that has high quality data in addition to the qualified personnel to analyze it cannot yet be considered as truly data driven. If there is no interest from people in knowing the analyzes and if the decisions of the decision makers are not influenced by these analyzes but by opinions and instinct, it cannot be said that these organizations are data driven. For organizations to be guided by data, such data must generate reports, which must influence the analysis that reaches decision makers so that they can incorporate them into their decision-making process. It is a fundamental step for an organization to be considered as data driven. **[13]**.

III. EXAMPLES OF DATA-DRIVEN APPROACHES FOR SOLVING PROBLEMS

In this section, some examples of data-driven approaches to solving problems will be explored. It is intended to give a practical view of how the information obtained from the data can provide superior results compared to other forms of decision making. Examples are related to Materials Science, Industrial Productivity, Energy (Gas, Oil and Wind) and Health (wearable devices).

On a new data-driven solution method, without the use of models, used in variational brittle fracture mechanics, the idea was to remove the assumptions of fracture modeling from the formulation and let the behavior constituting the fracture be guided exclusively by a set of material data, while maintaining the epistemic fracture laws that come from variational principles. The results obtained showed excellent agreement with those obtained via counterparts based on standard fracture mechanics. Another important point was the excellent robustness regarding noise in the data set used, despite the quality of these data being sensitive to noise**[22]**. A data-based approach was shown to diagnose production bottlenecks, using combined knowledge of maintenance and data science. Perceptions about these bottlenecks are obtained using artificial intelligence techniques such as machine learning, and real-world data sets extracted directly from the production line. The tool built in this way helps its users to plan specific maintenance actions to improve the availability of process bottlenecks and thus improve process productivity**[23]**.

Shale gas is a natural source of unconventional gas with immense reserves. Due to its ultra-low porosity and permeability, its extraction requires special drilling techniques. Accurate forecasting of its production is crucial for the reasonable design of the development plan. However, due to the complex hydraulic fracture network and the gas flow mechanism, the physics-based forecasting model does not yet exist and therefore the data-based model provides an alternative way of dealing with the production forecasting problem. The tool thus obtained, based on the random-forest method, showed the ability to make reasonable predictions of gas production when the physical model is not yet fully available**[24]**. Dealing with the problem of optimization of the lay-out of wind power generation parks to maximize the generation of electric energy. Due to the complexity of the lay-out problem, calculating the cost function takes a lot of time. To reduce the high computational cost of this calculation, while maintaining the performance of the obtained solution, an adaptive data-driven differential evolution algorithm was proposed. In most of the tested cases, a better or competitive performance was obtained with other algorithms in terms of output strength**[25]**.In the work carried out by **[26]** research is discussed where an integrated data-driven framework assisted by machine learning algorithms acquires signals based on personalized characteristics of elderly patients. Such work, based on wearable devices with a focus on patients' health, aims to improve the detection performance, based on reduced quantity of measurements, aiming at better model accuracy. With the introduction of a statistical framework, better results were obtained. And laboratory simulations showed that the performance of the system, in addition to user satisfaction, is superior to that of other conventional systems.

IV. ORGANIZING A DATABANK FOR A DATA-DRIVEN PREDICTIVE MODEL

This section aims to illustrate the process of organizing a databank for the development of an empirical predictive model for decision making and the process used to configurate a prediction model. Thus, it explores the search for data, its evaluation in seek of noise, the development of the criteria for the elimination of these noises and the treatment performed for their elimination. It can be said that most of the time spent to build the model is related to the previous treatment of the data. With this it is expected that the reader will be aware of the previous stepsfor the building of a data-driven predictive model and all the difficulties inherent to an industrial process since this example comes from a real project, developed for the steelmaking industry. The objective of these steps is to build a data-driven Artificial Neural Networks(ANN) model that is capable to provide predictive capacity for quality parameters of steel beams in a steelmaking industry. This model was built from several databanks and the steps used were as follows on the section 4.1 to 4.5**[27]**.Such a model allows to know in advance the quality results (mechanical properties) of the steel beam. In addition, it allows decision-making to adjust rolling processes in view of the initial chemical composition characteristics obtained at the melting-shop.A tool like that (**Prediction, optimization, simulation)** can answer the question, according to the table 2: “F) What is the best/worst that can happen?”. This is an approach that can be considered as Data-Driven because it allows prediction. Models like that are a particularly useful tool for quality checking and Research and Development of new products.

4.A. The Databank

The databank, which was used for training and validating the artificial neural network, contained data from some steps of the production process. Data that allowed the identification of the campaign and the type of steel were also used. From this bank, data with the following characteristics were selected **[27]**:

- Numerous technical standards;
- Numerous standardized chemical compositions;
- Total occurrences equal to 461.

Due to limitations of the Integrated Rolling Control System, only the following information can be used in the process of implementing the prediction model **[27]**:

- Chemical composition of steel, with analysis of the contents of 24 different chemical elements present in steel;
- Final Rolling Temperature;
- Thickness of the test specimen;
- Tensile Strength, Yield Strength and Elongation.

After analyzing the data recovered, it was observed that 11 of the available chemical elements did not present all the values of results in the databank. These elements were discarded as input for the model. They were Ni, Co, Ca, Ti, B, W, Zr, As, Sb, Te and Pb.The data about mechanical properties, the objective of the model, came from another databank. As this secondary databank has the same identification information as the databank originating from the Integrated Rolling Control System, it was possible to build a third, final databank, which aggregated all the information necessary for the building of the models **[27]**.

4.B. Statistical Analysis of Databank Variables

The databank obtained from the considerations of the previous steps has a high number of variables. However, it is expected that a large part of them will be strongly correlated or have a minor influence on the mechanical properties of the steel beam. For this last situation it is expected that they do not need to be included in the prediction model. To analyze the relationship between the various input variables with the mechanical properties, a statistical analysis of the data obtained was carried out to determine which variables would be used in the training of the Artificial Neural Network and in the final prediction model. The following analyzes were carried out **[27]**:

- Graphical, Scatter, analysis of the data;
- Correlation analysis between the various input variables (Chemical Composition, Temperatures, Reductions) and the outputs (LE, LR and A);
- Determination of average, minimum, maximum, and standard deviations of the input data;
- Histograms;
- Data Treatment: statistical analysis of the variables involved was carried out through the MINITAB Statistical Software. The data used were only those that were included within the range +/- 3 standard deviations to decrease the total variability of the databank;
- Elimination of outliers: Elimination of data that was not considered to be representative of the process. In the case of mechanical properties (LE, LR), the maximum difference of 20MPa was used as an acceptance criterion within the same production order (same steel heat, rolled in the same batch), for both LE and LR. Events with differences greater than 20MPa were excluded. The 20MPa criterion was adopted because production orders normally have lower values than this. Higher values are usually linked to some process abnormality, such as sampling and testing.

The techniques mentioned above were used to eliminate the presence of discrepant data, measurement errors, in short, noises that could compromise the reliability of the databank**[27]**.

4.C. Data Graphical Analysis

Once the data for model development was defined, MINITAB was used to graphically analyze the relationship of the output variables (LE, LR and A) with the input variables. The purpose of this procedure was to verify the impact of the variation of the input data on the mechanical properties **[27].**

It was observed that the chemical composition data of the databank were truncated, that is, they presented the decimal places lower than that obtained by the chemical analysis on the lab. As an example, the carbon content results (% by weight) in the databank had only two decimal places while the chemical analysis process provided results with greater precision. This loss of information could reduce the learning of the ANN model that was intended to be developed and, even, hinder the process of minimizing its error. Thus, it was decided to replace the chemical composition information originally obtained with the primary correspondents. Figure 2 (a) shows the carbon content histograms for the truncated data (originally contained in the databank). Truncating the input values is a considerable problem for the process of training and validating an ANN, as it hides important characteristics of the variable's behavior, which are indispensable for minimizing the model's prediction error. After the recovery of the original data, all the indicated analyzes were performed again. Figure 2 (b) shows the Histogram for the carbon content after updating the bank with the revised data **[27]:**

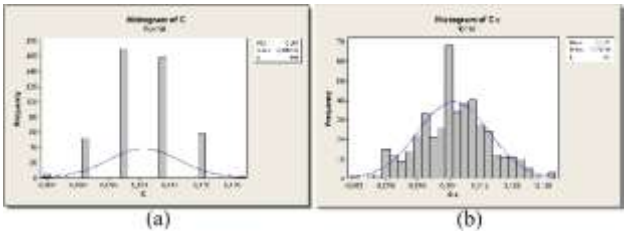

Fig.2: Grouped data (a) and non-grouped data (b) for the carbon content.

4.D. Graphical analysis of the databank

The figure 3 illustrates the dependence of the Yield Strength in relation to some process variables, Carbon, Manganese, Phosphorus and Silicon **[27]**.

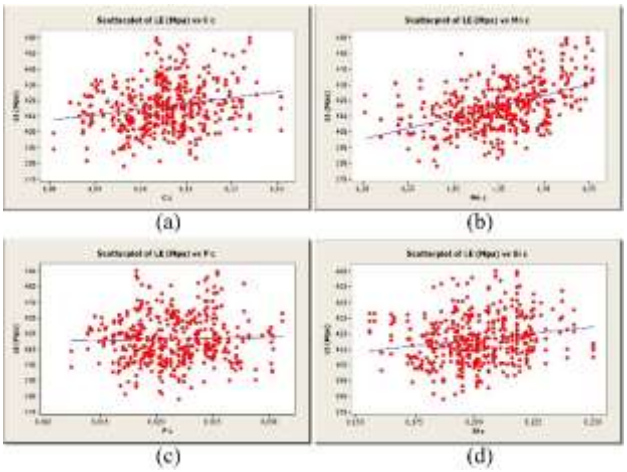

Fig.3: Yield Strength (LE) dependence due to some process variables. (a) Carbon, (b) Manganese, (c) Phosphorus and (d) Silicon.

This figure illustrates the dispersion found in the data. In this data set there may be discrepant data, data with measurement errors or even registration errors. To eliminate such noises, these data will be submitted to a previous evaluation stage to reduce the incidence of these noises as much as possible, as discussed on sections 4.1 and 4.2. What is aimed at is the development of a model that provides high precision, and, for that, it is necessary to work with the minimum possible noise.

4.E. Summary of Databank with Statistical Treatment

After performing all the procedures described in sections 4.1and 4.2, a databank with 444 occurrences was obtained. Table 3 summarizes the statistical information from the original and the final databank that will be used in the modeling **[27]**.

Table 3: Statistical Summary –Original and Final Databank.

Variable	Description	Original databank		Final databank	
		Average	Standard Deviation	Average	Standard Deviation
Y1	LE (MPa)	416.93	15.36	416.49	15.23
Y2	LR (MPa)	507.92	15.58	507.45	15.53
Y3	A (%)	28.84	2.03	28.94	1.88
X1	C (%)	0.1052	0.0089	0.1051	0.0090
X2	Mn (%)	1.3521	0.0508	1.3511	0.0510
X3	Si (%)	0.2011	0.0174	0.2012	0.0175
X4	S (%)	0.0075	0.0026	0.0073	0.0023
X5	Cr (%)	0.0302	0.0073	0.0303	0.0073
X6	Nb (%)	0.0360	0.0035	0.0278	0.0035
X7	N_2 (%)	0.0044	0.0009	0.0044	0.0009
X8	Final Rolling Temperature (FRT) (ºC)	972.50	12.45	972.29	12.37
X9	Thickness Reduction (%)	85.58	0.31	85.58	0.31

For most of the variables the standard deviation of the final databank was smaller if compared with the original databank. An additional reduction of the standard variation could be possible, but the learning of the ANN could be compromised. After choosing and defining the databank, the process of adjusting the ANN began with the objective of obtaining the best configuration, that is, the optimal number of neurons in the hidden layer. It was observed during the training process that the simulations could present different performances for the same network configuration. An explanation for this fact would be the existence of biased data in the sample set and, as they are selected randomly, depending on the distribution, the impacts would be perceived in the training phase or in the validation phase. Another explanation would be the random choice of initial weights, which would lead to different results in each simulation **[27]**.

A two-layer ANN architecture was used (one hidden layer and one output layer). The optimal number of neurons in the hidden layer was defined using the trial-and-error method, and configurations with at least 4 neurons were tested. The output layer was built for just one neuron. Three networks were developed, one for each output variable (Tensile Strength, Yield Strength and Elongation).The variation observed in the training of the networks created difficulties to define the optimal number of neurons in the hidden layer. Thus, the procedure started with a statistical approach where 10 simulations were performed for each number of neurons. The results were analyzed using the following performance parameters in the simulation **[27]**:

- Minimum and maximum percentage error in the validation;
- Average percentage error in the validation;
- Linear correlation, R^2, between estimated and measured output values.

The table 4 below show an example for the results of the adjustment of the ANN for the Yield Strength, number of neurons equal to 4**[27]**.

Table 4: Simulations for Artificial Neural Network with 4 neurons in the hidden layer, Yield Strength.

Sim.	Number of epochs	Training SSE	Min. Error, (%)	Max. Error, (%)	AverageError,(%)	R^2, LE's (%)
1	300	6.7677	0.02	6.81	2.24	66.4
2	173	4.4042	0.01	7.94	2.23	64.4
3	152	4.2218	0.03	6.90	2.26	62.9
4	188	4.4674	0.06	6.77	2.27	71.1
5	358	6.4770	0.09	6.95	2.44	66.1
6	119	4.6452	0.04	7.95	1.94	70.2
7	153	4.1389	0.11	8.28	2.41	66.0
8	233	4.5037	0.02	6.27	2.20	65,1
9	183	5.3227	0.00	8.19	2.16	58.9
10	516	6.7560	0.02	7,71	2.26	66.3

An Analysis of Variance (ANOVA) was performed through MINITAB Statistical Software for the results obtained for the 10 simulations performed for each network configuration. From the results obtained, the ideal network configuration was defined, in terms of number of neurons on the hidden layer**[27]**. Table 5 shows the summary of the characteristics of the Artificial Neural Networks used in this work:

Table 5: Summary of Characteristics of Artificial Neural Networks.

Characteristic	Criteria	MATLAB
Partition of data set	Training set = 75%, Validation set = 25%.	RANPERM
Net weigh initialization	-	INITNW
Net learning ratio	-	TRAINGDX
Transfer Function	-	TANSIG
Convergence Criteria	-	
Minimum error aimed	0,001	-
Number of training cycle	700	-
Training mode	BT	-
Number of hidden layers	1	-
Size of hidden layer, LE Model	6	-
Net training mode	-	TRAINBR

The same process was repeated to LR and A to define the best configuration for the prediction model. For the LE model (6 neurons on the hidden layer) the correlation between the measured LE value and the predicted was equal to 0.65; the average error was equal to 2.27%, the minimum error was equal to 0,00% while the maximum error was equal to 7.77%. Most of the established metallurgical trends was confirmed and the model could be considered a reliable prediction tool to calculations of scenarios, decision making, and to optimize the steelmaking process, from the melting shop to the rolling mill.

V. CONCLUSION

This paper discussed the theoretical background related to Data-Drive Decision Making, its correlation with the increasing availability of data, and how organizations that use this strategy can deliver better results. Practical examples of earnings gains from data-driven organizations were also given. The details necessary for an organization to be in fact data-driven were discussed in detail, the necessary aspects related to basic characteristics: Collection and access to data; Reports, Alerts, Teams, Data Analysis; Data Analysis Maturity; Metrics; Telling Stories with Data; Information Delivery and Decision Making. It also explores the characteristics of organizations that are truly data-driven, as well as what factors can influence decision-making, factors that would prevent organizations from being data-driven. Some examples of data-driven approaches to solving problems on Materials Science, Industrial Productivity, Energy (Gas, Oil and Wind) and Health (wearable devices) was discussed. An example of Data-Driven prediction model was discussed in detail: The development of a real predictive model that enables the decision-taking in a steel industry with minimum error and metallurgically accurate. This last case is discussed with a focus on the noise reduction of the data used and the methodology used to adjust the Artificial Neural Network. This article is expected to contribute to the growth of the technical knowledge of its readers.

REFERENCES

[1] Brynjolfsson, E. Hitt, L. M., Kim, H. H. (2011). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? http://dx.doi.org/10.2139/ssrn.1819486;

[2] Pentland, A., Pentland, S. (2008). Honest Signals: How They Shape Our World. The MIT Press;

[3] Rose, K. Eldridge, S., Chapin, L. (2015). The Internet of Things: An overview, understanding the issues and challenges of a more connected world. The Internet Society (ISOC);

[4] Industry 4.0 and how smart sensors make the difference. https://www.spectralengines.com/articles/industry-4-0-and-how-smart-sensors-make-the-difference;

[5] Hajoary, P. K. (2020). Industry 4.0 Maturity and Readiness Models: A Systematic Literature Review and Future Framework. International Journal of Innovation and Technology Management. https://doi.org/10.1142/S0219877020300050;

[6] Mosig, T., Lehmann, C., Neyer, A-K. (2019). Data-Driven Business Model Innovation: About Barriers and New Perspectives, International Journal of Innovation and Technology Management. https://doi.org/10.1142/S0219877020400179;

[7] Dutta, D. Sarma, M. K. (2020). Adoption of Digital Innovation-Formulating Adopter Categories and Levels of Adoption in a Digital Sphere in an Emerging Economy. International Journal of Innovation and Technology Management. https://doi.org/10.1142/S0219877020500595;

[8] Galbraith, J. R. (1974). Organization Design: An Information Processing View. Interfaces, vol. 4, no. 3, pp. 28–36. JSTOR, www.jstor.org/stable/25059090;

[9] Brynjolfsson, E., McElheran, K. S. (2019). Data in Action: Data-Driven Decision Making and Predictive Analytics in U.S. Manufacturing. Rotman School of Management Working Paper No. 3422397, http://dx.doi.org/10.2139/ssrn.3422397;

[10] Chen, H., Chiang, R. H. L., Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, vol. 36, no. 4, pp. 1165–1188. www.jstor.org/stable/41703503;

[11] Tambe, P. (2014). Big Data Investment, Skills, and Firm Value. Management Science, vol. 60, no. 6, pp. 1452–1469., www.jstor.org/stable/42919614;

[12] Acito, F., Khatri, V. (2014). Business analytics: Why now and what next? Business Horizons. 57. 10.1016/j.bushor.2014.06.001;

[13] Anderson, C. (2015). Creating a Data-Driven Organization : Practical Advice from the Trenches / Carl Anderson. First edition. Beijing, China: O'Reilly, Print;

[14] Farooqui, A., Bengtsson, K., Falkman, P., Fabian, M. (2020). Towards data-driven approaches in manufacturing: an architecture to collect sequences of operations, International Journal of Production Research, 58:16, 4947-4963, DOI: 10.1080/00207543.2020. 1735660;

[15] Sun, Y., Haghighat, F., Fung, B. C. M. (2020). A review of the-state-of-the-art in data-driven approaches for building energy prediction, Energy and Buildings, Volume 221, 110022, ISSN 0378-7788;

[16] Power, D., Heavin, C., McDermott, J., & Daly, M. (2018). Defining business analytics: An empirical approach. Journal of Business Analytics, 1(1), 40–53;

[17] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: The management revolution. Harvard Business Review, 90(10), 60–68;

[18] Cosic, R., Shanks, G., & Maynard, S. (2015). A business analytics capability framework. Australasian Journal of Information Systems, 19, S5–S19;

[19] Koohang, A., & Hatch, M. (2017). Leadership effectiveness in IT-centered organizations: Gender and levels of management. Journal of Computer Information Systems, 57(4), 385–391;

[20] Northouse, P. G. (2010). Leadership: Theory and practice (5th edition). Thousand Oaks, CA: Sage;

[21] Carillo, K. (2017). Let's stop trying to be "sexy" – preparing managers for the (big) datadriven business era. Business Process Management Journal, 23(3), 598–622;

[22] Carrara, P., De Lorenzis, L., Stainier, L., Ortiz, M. (2020). Data-driven fracture mechanics, Computer Methods in Applied Mechanics and Engineering, Volume 372, 113390, ISSN 0045-7825;

[23] Subramaniyan, M., Skoogh, A., Muhammad, A. S., Bokrantz, J., Johansson, B., Roser, C. (2020). A data-driven approach to diagnosing throughput bottlenecks from a maintenance perspective, Computers & Industrial Engineering, Volume 150, 106851, ISSN 0360-8352;

[24] Xue, L., Liu, Y., Xiong, Y., Liu, Y., Cui, X., Lei, G. (2021). A data-driven shale gas production forecasting method based on the multi-objective random forest regression, Journal of Petroleum Science and Engineering, Volume 196, 107801, ISSN 0920-4105;

[25] Long. H., Li, P., Gu, W. (2020). A data-driven evolutionary algorithm for wind farm layout optimization, Energy, Volume 208, 2020, 118310, ISSN 0360-5442;

[26] Ba, T., Li, S., Wei, Y. (2021). A data-driven machine learning integrated wearable medical sensor framework for elderly care service, Measurement, Volume 167, 108383, ISSN 0263-2241;

[27] Oliveira, A. P. (2008). Prediction Model of Mechanical Properties of Hot-Rolled Structural Beams: An Approach in Artificial Neural Networks. (Dissertation, Master's in Metallurgical and Mining Engineering). Digital Library of Universidade Federal de Minas Gerais, UFMG. http://hdl.handle.net/1843/MAPO-7RLKBJ.