Professional MBA
Automotive Industry



# Role of Data Analytics as a Tool for reaching Operational Excellence a case study with integration of Data Analytics model in Aluminum foundry of battery housing production and its effect on the product quality results

A Master's Thesis submitted for the degree of
“Master of Business Administration”

supervised by
Professor Dr. Bernd Hellingrath

Peter Beno, Ing.

11929264

Vienna, 16.10.2022

ACADEMY FOR CONTINUING EDUCATION

STU
SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE



# Affidavit

I, **PETER BENO, ING.**, hereby declare

1. that I am the sole author of the present Master's Thesis, "ROLE OF DATA ANALYTICS AS A TOOL FOR REACHING OPERATIONAL EXCELLENCE A CASE STUDY WITH INTEGRATION OF DATA ANALYTICS MODEL IN ALUMINUM FOUNDRY OF BATTERY HOUSING PRODUCTION AND ITS EFFECT ON THE PRODUCT QUALITY RESULTS", 109 pages, bound, and that I have not used any source or tool other than those referenced or any other illicit aid or tool, and
2. that I have not prior to this date submitted the topic of this Master's Thesis or parts of it in any form for assessment as an examination paper, either in Austria or abroad.

Vienna, 16.10.2022

Signature

# ACKNOWLEDGEMENT

I would like to thank to Professor Bernd Hellingrath for mentoring this thesis and his support during the writing process. Parallel I would like to thank to the internal lean team under supervision of Radoslav Palacka, where they understood that the thesis topics is a foundation brick for reaching operational excellence of our company in the future.

Last by not least, I would like to give big thank you as well to my wife and daughter, my brother, and parents for their support during the time necessary for learning and working on the thesis.

# ABSTRACT

The current market situation not only in the automotive segment is very unpredictable and very volatile. The missing semiconductors, war conflict in 2022 is bringing many phenomena as a price and energy, material and personal costs increase. The business transformation in automotive segment is moving from internal combustion engines to electrifications what brings also many challenges and opportunities. This all brings the pressure on the OEMs supplier chain to produce their products effectively with high effectiveness in shortest possible time and low product price. To sustain competitiveness on the market the supplier chain has to adopt it and challenge to these changes. The new era of Industry 4.0 brings many opportunities where integration of big data analytics in real case study. The experience with proposed random forest algorithm is displayed on the real case study of battery housing production in aluminum high pressure die casting foundry. The final model precision is reached 74,9% with recall on the level 60,6%. So far there aren't similar experiences and results in the foundry business to compare on real case situations. The internal quality was significantly improved after implementation of outputs from the model results what brought the battery housing production closer to reaching operational excellence. Data analytics integration can help organization improve internal performance and financial indicators.

**Key words:** data analytics, aluminum casting, process parameter optimization, casting defects, internal quality, operational excellence

# TABLE OF CONTENTS

# 1. INTRODUCTION

The new "Era – called Digitalization Era," has begun around 2014 where digitalization of industrial processes is also called the fourth industrial revolution. Organizations in a variety of industries are starting to be more and more interested in working with large amounts of data that is related to process variables. This offers new opportunities and to get deeper insights on process understanding of parameters and their effects on quality and performance, and in the end on process parameter optimization. Besides the fact that the new available technologies have direct impact on the processes in the companies and OEMs, the innovations also have impact on customer behaviour. A significant proportion of product portfolio has changed after the "Digitalization Era" has brought new potentials on the field of digital products and services.

Business technology advancements and development are increasingly transforming the conventional operational supervision process. Data analytics (DA), machine learning models, and artificial neural networks are increasingly significant in assisting businesses in connecting workflow operations with task processing across the company to simplify complicated operations and program management procedures. Alerts can be set up as a preventative measure for manufacturing items of poor quality. Automation also contributes to standardization and uniform, high-quality delivery across all projects and procedures. Also, the data analytics technologies like machine learning along with the other methods provide a variety of benefits to the companies and to the customers. Mentioned above has an influence on the market development. Many of these technologies aid in the interpretation of data as well as automate numerous procedures and choices "Intelligent Industry" has the potential to provide additional commercial and efficiency benefits beyond those of Industry 4.0, and it can assist original equipment manufacturers (OEMs) in remaining competitive in the long run.

On the other hand, the industries are challenging to survive. Increasing of almost every cost including energy, personal costs, material, and others are with regular frequency in the year 2022 reaching maximum prices what brings another pressure to whole supply chain.

OEMs precisely watch the performance of the supplier chain and in a case of regular bad performance, for instance (f.i.) irregularity of shipments or external quality issues leads to start thinking about product relocation to another supplier - even the cost impacts on OEMs side are not small. This pressure forces supplier to invest into continuing improving of internal processes that give an opportunity and space to start using internal datasets in wider range. As another important input information for the process engineers and technicians for strengthening the process and by that become more competitive on the market. Strength and robustness of internal process towards reaching operational excellence is one of the main aspects in the strategies of companies. It can be also defined as action for overcoming the upcoming volatile years. Operational excellence (OE) – two magic words are bringing solution how to challenge up and down trends on the market. Operational excellence is mostly focusing to produce the right top quality (lowest possible internal rejects /scrap) in the right (shortest) production time with the right amount (lowest) of resources that are needed for product production. The center of everything is then producing the quality products where it is easily directed to operational excellence. If the rejection / scrap is high, the necessity of extra production hours is needed to fill the missing quantity of the product with extra resources (gas, electricity, personal costs etc.). That means that the lowest scrap belongs to main structure of operational excellence. Operational excellence is one part and evaluation of OE is the second. Commonly used in the praxis are key performance indicators (kpi) in the direction of evaluating process performance. As mentioned before – internal scrap belongs to them. Also, there are used other ones – overall equipment efficiency (OEE), productivity, on – time delivery (OTD) etc. The chosen correct kpi depends on the area of OE scope and management definition what is important to reach.

Anyway, current unpredictable market situation creates not only risks and negative impacts but also opportunities. Opportunities how to transform the companies to be more adaptive, agile, and prepared towards upcoming future. One form is using new approach which offer big data analytics. Massive data downloading, processing and evaluation can bring the missing puzzle and opportunity for the business development. Better understanding of internal processes can bring the right direction and input data for continuous process improvement and through it the robust process set up, with low variation can be reached.

## 1.1 Motivation/background

Scarcity of the resources, regulation from the governments towards $CO_2$ / $NO_X$ emission reductions which are linked to the global warming process is an increasing pressure on everybody in the supply chain. Original equipment manufacturers (OEMs) mainly in the automotive sphere are forced to bring the new solutions. Due to the environmental regulation, they must develop their product in a short time, products which are still 100% reliable, cost effective – interesting for the market, on time delivered to the final customer = products which bring expected sales volume and revenue. Nowadays a trend can be seen of increasing investments towards electrification of car fleets to reach the emissions targets set up by European Union.

New technology development in the automotive segment (but not only in automotive segment) for introduction of new products on the market sooner than competitors is only increasing the expectations to be more innovative, cost efficient, competitive, and fast. Overall, this pressure creates waves through the whole supply chain, where higher product complexity isn't leverage by higher product prices but more common in opposite way. The development time is shortened what has a significant negative effect of the product, process development .

With the war conflict in the 2022 the situation in the market has become only more unpredictable. Increasing prices for energy – gas, electricity, and material – such as aluminum, rare metals etc. led to devastate impact on the industries where energy consumption starts play an important role. The high energy prices are increasing costs of manufacturing. To this group of industry belongs also aluminum foundry business. High pressure die casting process (HPDC) is a foundry process for production of high-quality complex products. During the casting production there are many process variables which must be under control, and which interact among each other - what can cause with combination of process imperfection and very complex design product to higher internal scrap rate. The trend of high internal scrap, even with implemented lean principle for continuing improvements, is an opposite direction to using internal resources effectively in the maximum possible way with the lowest costs. By mention, as compensation, management is building several initiatives how to improve core process for reaching operational excellence in the business. To support and create more robust and productive internal environment a detailed understanding of internal processes is necessary and simply a need. The Fourth Industrial Revolution is influencing the manufacturing business and the demands related to the innovation are still increasing. Also, the machines in the manufacturing sector are being equipped and used with sensors that allow to acquire large volume of data. It is important to focus on working and using the obtained data for maintaining competitive advantage of the company. In contrast with that, to this day still a huge number of organizations don't realize the huge potential of data. On the other hand, some organizations don't know how to work the correct way with obtained data and how to capitalize it. However, one of the biggest challenges for companies today is to get consistent reports and information quickly so that the analysis can guide their goals and objectives. In this regard, HPDC process create and offer a perfect platform for structured approach to data mining processes which can be used for data analytics to solve complex quality problems.

The volume and variety of data have far outstripped the capacity of manual analysis, and in some cases have exceeded the capacity of conventional databases. At the same time, computers have become far more powerful, networking is ubiquitous, and algorithms have been developed that can connect datasets to enable broader and deeper analyses than previously possible. The convergence of these phenomena has given rise to the increasingly widespread business application of data science. Therefore, effective management of data and appropriate operations decisions based on big data analytics (BDA) has become more challenging in the Big Data era.

While the internal environment is ready for deeper analyzing of processing datasets - this is an ideal stage for start of implementing data analytics or machine learning models. But to find the correct DA model, or any other machine learning model with highest precision to real time situations there is some time necessary to invest with try and error procedure. Every process is distinguished by own key process parameters and to have one common DA model is a utopia and but also a dream of every process and operation manager. Correct DA offers a solution for complex process to set up on -time information as support for the team to faster and hopefully right decision making. Casting process belongs to the complex processes where even several simulations runs are made in yearly stage of the product development to define the correct process and casting design. In real-life production with real-life production problems and real-life process irregularities the quality results are variable and negatively differed from ideal simulated cases as was shown in the early stages of the product development. To keep the process variance of defined critical process parameters under control, several measurements are implemented as for example (f.e.). process capability indexes (cp/cpk indicators) or lean core tools as competitive tools of continuing process improvement and process robustness measurement. Increasing of cp/cpk indices belongs to the focus of team in automotive industry to create robust process set up with low variance, but with multi factoring process variables interactions as are in the HPDC process, it is a mission impossible with other support. Companies need comprehensive manufacturing strategies with support and right process set ups for making fast and correct decisions. OE supported with data analytics can bring the advantages but also many opportunities towards effectively using internal resources, systems, and organization structure. To sustain the business and its readiness for the next challenges innovations and reaching the best quality product could be possible outcome. To reach OE in manufacturing process means bringing the right direction to increase added value and react on the multitude of challenges coming from the market or more customized products. DA integration as a supportive tools of lean continuous improvement concept could be valid for all production companies and industries – production of product with the top quality at lower possible cost with maximized utilization of internal sources is one of common mantras in the manufacturing.

## 1.2 Definition of the research problem and research questions

Unfortunately, the quality level of casting in the aluminum foundry area is often oscillating. There are several reasons and impact factors why. For stabilizing it is necessary to have deep knowledge, be “lucky” in terms of robust casting design and correct process set up considering all his variability. The knowledges and deeper insight in the real-life process can bring DA model as a supportive tool for process engineers and shopfloor specialists to define desired action plan. The primary goal of engineers is to satisfy customer specification and minimize parts produced out of specification (NOK parts). The negative influence of scrap level has a significant influence on financial, performance and organizational outputs of the company. With high internal scrap the on-time deliveries are in risk due low first time true and overall installed machine capacity. The high shopfloor operative is in many cases compared to firefighting operation which is in the long-term period unsustainable. Because of capability of the manufacturing process which varies with time, tooling conditions and machine during the time is changing - the technological parameters and working environment must react to this situation. To keep everything under control is a nice wish and target of every process engineer and production. But it isn’t in human capabilities to supervise and act on-line 24/7 time.

The casting aluminum process belongs to very specific and complex technological process where various operational technologic parameters interact between others to get a product which reflects the customer specifications – dimensional, technological, and other special requirements or characteristics. All these parameters must replicate in all possible conditions of production with no negative influences on the process disturbances. For sure before the first casting is done, there are made several simulations in development and optimizing phases. Anyway, the following statement will be far away from the truth that in real life every casting is unique because it is not possible to maintain 100% of the same conditions during the production. Dealing with thermodynamics processes, hot tooling, cycle times, machine itself and product designs along with dozens of other variables and process noises is challenging to get repetitive results in the quality level. Better understanding of the process with a higher control of process variables and noise process reduction must logically lead to a higher sustainability of the process and by that improving of the process where the key input is a correct input information. Process support by integration of data analytics model can be the right missing key where more detailed data, correlations of data with the high model prediction for the real-life situation about actual manufacturing process will have positive impact on the knowledge, faster reaction and should lead to improving internal quality level and getting closer to operational excellence.

Without products produced with top quality, on time and every time the possibility to approach OE is almost equal to zero.

The main purpose of this thesis is to implement a data analytics model in the real-life environment on the real product, then test it. Firstly, to the find best model results – it is very important to understand how the DA model can be improved to get the highest possible outcome for increasing precision and recall model parameter. Secondly, to have and knowledge about the main aspect of the thesis - what possible outcome can be reached after better understanding of the effect of process variables on the chosen defect of product. This has to bring the positive impact on an internal quality level in foundry industry in similar product portfolio family. Improvement of quality level is step closer for reaching OE. A practical case will demonstrate experience with the chosen model implementation following the Cross Industry Standard Process for Data Mining (CRISP DM) scientific approach (Nodeh et al., 2020).

The thesis shall address the following research questions:

**RQ 1** Will be the chosen machine learning model suitable for casting process and defined casting defect ?

**RQ 2** Could be results from learning model used as an input for improving quality results as supportive tool for reaching OE?

**RQ 3**: What take ways could be transferred to the other industries?

## 1.3 Hypothesis statement

Application of DA machine learning model in a real case study in the aluminum casting Al-Si alloy will predict the defect formations in the casting with high precision and recall and by that reasonable action will be defined for internal quality improvement. Based on better insight of the process internal quality will improve and what cause a positive trend towards operational excellence.

## 1.4 Structure of the thesis

**Introduction**
- Motivation / background
- Definition of the research problem and research questions
- Hypothesis statement
- The structure of the Thesis

**Operational Excellence / Tools**
- Operational excellence
- Continuos improvements process

**Big Data Analytics**
- Data analytics and big data
- Methodologies applying data analytics
- Application of data analytics towards process control

**Research Assignment**
- Description of the methodical approach
- Literature study and information retrieval
- Case study definition in internal environment

**Research Results and Analysis**
- Application of DA in the area of aluminum foundry
- Possible improvements after implementation of DA and their interconnections to OE/process robustness
- Findings identified from the case study
- Deployment strategy
- Executive summary

**Discussion and Implications**
- Suggestions for solutions

**Conclusion**

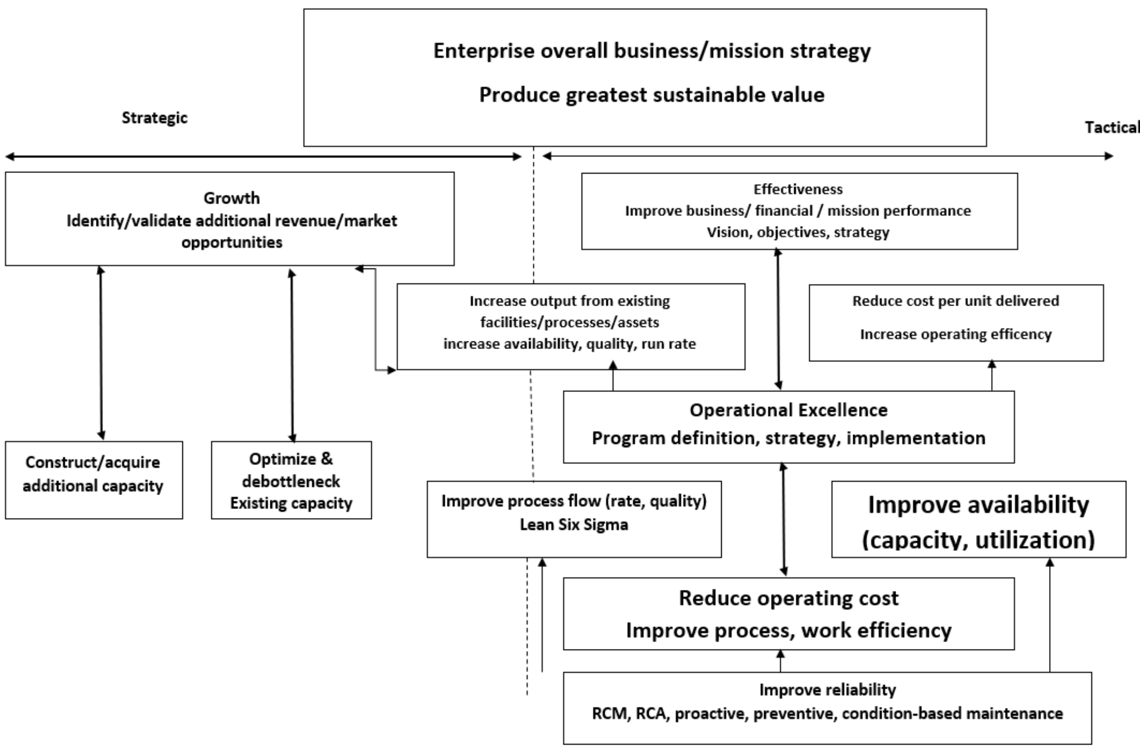
# 2. OPERATIONAL EXCELLENCE

The term Operational Excellence (OE) describe the ideal state of an operating company. It is continuing way of improvement process in the areas most critical for gaining and sustaining business and company mission success. In the manufacturing sector, the start of OE is related to using the 6S or Toyota Production Sytem (Tissir 2022). Dugan (2011) defines OE as a point where every employee can see the flow of value to the customer and fix it when the flow is interrupted or breaks down. Treacy & Wiersema (2017) describe the concept of OE like a strategy that is used by companies for maintaining quality, price, and other services that is unique when compared to other companies in the sector. Both brings own point of look on the OE. Each company or industry can identify own point of look what they understand as operational excellence (Mitchel 2014). Based on Miller (2014) an operational excellence is a mindset, not a methodology. Many organizations are implementing OE initiatives to be competitive in the market or industry. Dugan (2011) defines that a company that can be labeled as "operationally excellent" in the ideal conditions, is a company where everyone is acknowledged with their own contribution in the process of flow of value directed to the costumer. Management strategy and related processes related to OE are the main factors when defining and further working with OE.

In the operating industry where belongs aluminum foundry is OE built largely on the quality of products and productivity of the processes. Everything is focused on the efficiency (performing a given task well) and effectiveness (perform the correct task efficiently) of the processes. But without reliable, repetitive process set ups is almost impossible to reach results which can be considered and close to expectation of OE. By Chang (2010) process delays, waste, inefficiencies, and redundancies are the major barriers to reaching high levels of OE.

According to Mitchell (2014) OE is connected with an essential question: “where we must be” based on the management answer the project set up for OE is defined. The Figure 1 shows the border line between strategic and tactical scope and the position of the OE program to fulfil the companies long term strategy. The common indicators of OE (areas of the interest) are for process flow: rate and the quality, availability, and often operating costs (energy, personal, material etc.).

Figure 1: Strategic and tactical elements of OE set up



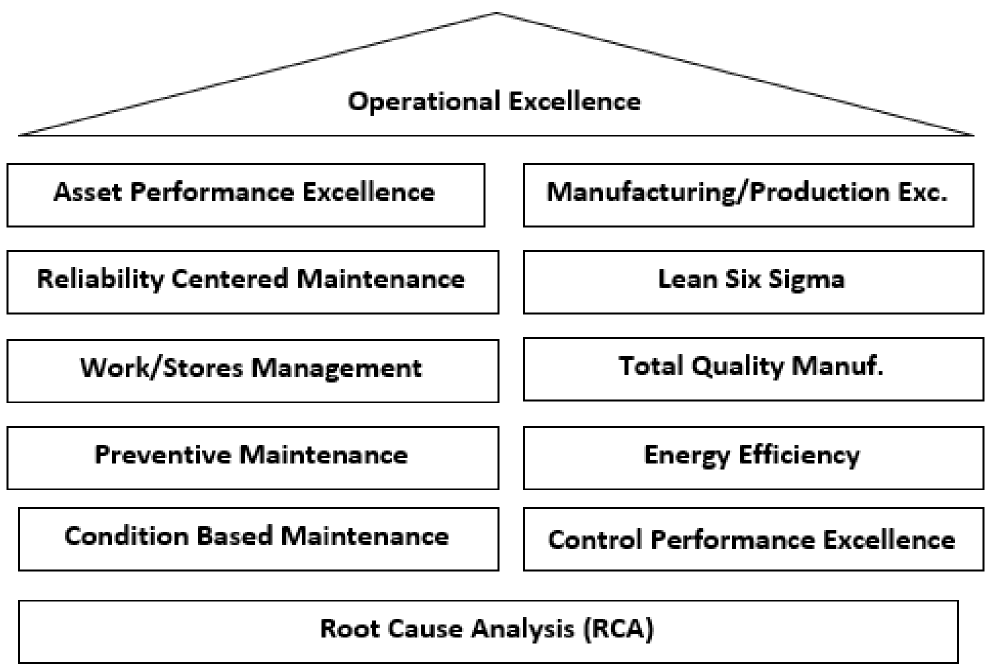

Source: Mitchell (2014), own elaboration

## 2.1 Lean principles and OE

Reaching the OE within a company is based on a variety of factors. One of the main factors that are very important to reach OE is to understand the processes of the company in the first place. There are often used tools as Value Stream Mapping (Nash & Poling, 2008) and Lean Six Sigma (George, 2003) that are proven methodologies and can be efficaciously applied for deeper process understanding.

Mitchell (2014) created a "house" with including all functional improvement tools for OE – see Figure 2.

Figure 2: House with all improvements program




Source: Mitchell (2014), own elaboration

As previously said, Lean Six Sigma is a well-known waste-reduction approach that has been used successfully in the manufacturing industry. Six Sigma was first used by

Motorola in the 1980s for high-volume, highly standardized production operations. The goal was to reduce waste by getting near-perfect results. (Biolos, 2002).

The Lean principle focuses on process speed or flow improvement. Six Sigma (6S) applies the DMAIC procedure. DMAIC is an integral part of 6S (Chang, 2010, Dale et al 2007). DMAIC is a shortcut for: Define, Measure, Analyze, Improve, and Control – these are the steps to increase the quality of current processes in the company. Along with the Lean and Six Sigma it can be used to make the whole working process as easy as possible. The concept of Lean is focused on increasing the speed of the process and quality improvement. The mention systems merged into one concept called Lean Six Sigma - that system is focuses on minimizing the complexity of related activities, but at the same time increases the quality of the products and process speed. Along with the focus on the flow of data and the relation to people, Lean Six Sigma can be also used in a variety of a processes beside manufacturing – sales, call and research and development centers etc. (Chang, 2010). OE and continuous improvement go together when it comes to achieving a lean organization; however, they are not the same thing.

**Continuous improvement process (CIP)** is the act of implementing improvements to a product, service, or process. These changes can either be incremental (over time) or breakthrough, which means all at once (Tallyfy, 2021). The key word here is continuous as CPI is not a one-time initiative. To really succeed with a process improvement, it is important to constantly analyze the variables in a defined period. This allows the company to evaluate the weak and strong aspects, and also if there are any changes that need to be done based on the findings.

**Six Sigma (6S)** is a method with the utility to manage process failures that cause a variety of defects, which means they are deviated from the stated target. The concept of 6S is based on further working with the possible variations of factors which then allows to minimize the likelihood of the occurrence of these defects in the future (Simonova, 2014).

**Lean Six Sigma** is a methodology that stands for the combination of six sigma and lean. This combination is done by merging principles to overcome weaknesses, which highlights the benefits of both programs. Lean focuses on eliminating waste of all kinds and 6S focuses on process control, but out of the process (efficiency issues) (Stankalla, 2019).

## 2.2 Continuous improvement process and Industry 4.0

Sustaining continuous improvement requires a systematic process of selecting "what" and "how" to focus resources on priority challenges and opportunities, as well as investing into workers training process. The selection process is based on a variety of aspects. One of the main factors are the techniques currently used, along with the company individual factors which determine the functioning and principles of the organization. Many aspects influence the structured process selection, including the present techniques in place, the type of work to be managed, the level of readiness of the organization. Often, more than one option is a good match, or a combination of options is successful, but the company must choose a method as part of its strategic deployment. (Bach, 2021). There are many improvement methodologies available.

For instance, in the company where the case is made, they are focusing on the following methodologies:

1) **Lean:** Reduce waste and increase value to the customer.
2) **Six Sigma:** Strive for perfection. Eliminate defects. Generally, quite a data-driven process.
3) **VSM:** Find the bottleneck and eliminate waste, improve material flow and efficiency etc.
4) **Statistical Process Control:** Utilize process data to reduce variability, thus improving product quality improve cp and cpk indices.
5) **Quality standards ISO, IATF and others:** Combining the industry standards with the intention to improve the current practices and operations.
6) **Toyota KATA approach:** new systematic approach integrated within last years is becoming more popular.
7) **TWI** – Can be defined as training within the industry.

   It can also be linked to the description of the ISO standard, which is based on continuously taking steps to enhance performance. (ISO, 2014). Also, the continuous improvement can be linked to the concept of Kaizen, which is made up from two Japanese words – Kai, who stands for improvement, and Zen, which translates as "good".

There are 6 phases or steps of Kaizen: 1) Identify a problem or opportunity; 2) Analyze the process; 3) Develop an optimal solution; 4) Implement the solution; 5) Study the results and adjust; 6) Standardize the solution.

Continuous improvement activities as were already mentioned focus on reducing variation in processes to increase process capability what finally leads to satisfaction of the customer by fulfilment of their requirements. The introduction of the CIP always leads to changes in existing processes. Based on this, it is recommended that the employees are involved and acknowledged within the whole CIP process. It is therefore essential that employees are involved in the CIP process which is related to the common daily work tasks, along with the possible solutions that can be used to solve these problematic situations.

A new pattern can be identified for the companies in the context of the future Industry 4.0, which will change and improve the way of work. The question is how the Industry 4.0 will be integrated in the already used systems which were described above.

The Industry 4.0 has been described as a concept consisting mainly of digitalization and automatization. Combination of these two can make a significant improvement in productivity (Chiarini & Kumar, 2020). This is realized mostly by the connection and integration in the production along with the value chains through the new technologies and Internet of Things (IoT) (Ghobakhloo, 2018).

Industry 4.0 can be defined as a summary of various digital technologies, which are connected. Based on the cooperation, it is possible to obtain real-time data for analytics and manufacturing purpose. The most used technologies are cloud services, data analytics and others (Frank, Dalenogare & Ayala, 2019).

During the last years, Industry 4.0 has been categorized as a strategic model. This model is used mainly for increasing competitiveness and improving indicators like cost, quality, productivity and actions related to customer services (Bibby & Dehe, 2018). Similar goals are shared by OE methodologies such as Lean and Six Sigma that have supported organizations in the last three decades to achieve efficiency gains and enhance customer satisfaction that the business keeps growing.

Implementing of Industry 4.0 especially part Big Data, offers a huge amount of new data that must be somehow processed and evaluated and used as an input for the continuous improvement process or input for Lean Six Sigma activities. The availability and reliability of the data are very important for success of continuous improvement and lean 6Sigma activities. As a part of Industry 4.0 Big Data and Data Analytics offers solution of on-line ensuring real-time data what avoid shortcomings (Tissir, 2022) which could led to misunderstandings or bad decision-making during problem solving activities.

The question is, if combining traditional methods such as Lean Six Sigma with next evolution Industry 4.0, especially data analytics integration could possibly lead to effective outcome from the Industry 4.0. This issue is further investigated and it is needed to do more research for obtaining more relevant data and views on this of this issue (Antony, 2022). This thesis is focusing on the answer on this question on real-life case with implementation of data analytics model on special casting defect. The question is if the model precision will be so strong to determine the defect occurrence with significance process variables factors as main influencers on the product quality related to the observed defect. The model precision will build the foundation for the process engineers to define root cause precisely, faster and by that the impact on the product quality improvement will be visible sooner.

## 2.3 Evaluation and metric of operational excellence

Metrics and performance benchmarking are a key aspect of evaluation of OE. To optimize performance, companies must regularly analyze the performance to achieve steadily grow. Use of OE concepts in manufacturing processes allows companies to increase profits. Cost pressure has always been and will continue to be on the supplier’s agenda. Productivity improvements and reaching best-in-class quality results have become a standard expectation of customers, owners, stakeholders, management.

To measure OE effectively, an organization must be fully committed to relevant and achievable goals against which its success can be measured. OE can be evaluated through external and internal indicators. One of the most important measurements of OE program's success is external benchmarking, for example: How the organization is performing against its peers in specific key operational metrics? When comparing one company to other, it is important to keep in mind the constant improvement of the individual companies and also evolution of the whole market and industry. Company's competitors aren't frozen in time as their businesses are growing and changing as well. One of the best ways how to evaluate whether the company's pace is sufficient to build or sustain the competitive edge is through careful assessment of the leading internal indicators. Productivity and quality are the most used metrics in the context of performance manufacturing (Wilkings, 2021).

According to Garza-Reyes, 2010, **Overall equipment effectiveness** (OEE) is a metric used for evaluating productivity, but also as a sign while working with information related to process or performance. The level of OEE indicator for world -class companies is 85% (Mitchell, 2014).

In many cases to increase performance without stable quality results it is mission impossible. Same is valid with availability indicators – if the process is discontinued and there are many stop and go this has in many cases negative effect on the product quality as well.

The first concept of OEE was used for monitoring and controlling performance (Dal, 1999). He explains that OEE works with all possible indicatives. Dal (1999) suggests that the role of OEE goes far beyond the task of just monitoring and controlling. This allows the OEE to take into account process improvement, and also prevent under-optimization of individual machines or production line. It can also provide a systematic way to set production targets. It combines practical management tools and techniques to achieve a balanced view of performance and quality. OEE can also be used as a sign of improvement process, while it can also be a opportunity to achieve the intended improvement. OEE is the single best metric for identifying losses, benchmarking progress, and improving the productivity of manufacturing equipment (i.e., eliminating waste). There are several other OE indicators used in automotive industry but in foundry area standard OEE is implemented as a one of tool for the process performance evaluation. L. A. Gólcher-Barguil (2019) summarize the used metrics for evaluation of OE and define operational excellence profitability (OEP) approach. Most of the authors are evaluating the performance of the operations.

**Process capability** (PC), is mostly used for measuring quality performance (Garza-Reyes, 2010). The evaluation of PC is carried out through a framework called process capability analysis (PCA).

The selection and use of measures of performance such as OEE and PC is determined by the priorities of an organization. PCA is defined by Deleryd (1999) as an improvement method where a product characteristic is measured and analyzed.

The process capability can be evaluated also through other measures and methods. The most common way of doing it is through the capability indices (CI). CI are specific measures that compare the actual process output with the specification limits for a certain characteristic (Deleryd, 1999). In other words, they show the ability of a process to meet its numerical requirements. Among the most well-known are CI are Cp and Cpk indices.

**Process robustness** is defined as "the ability of a manufacturing process to tolerate the variability of raw materials, process equipment, operating conditions, environmental conditions and human factors" (Glodek et al., 2006, p. 3). It can be found in process and also product design.

According to Giannetti and Ransing (2016), process resilience may be accomplished by adjusting the tolerance limits to improve process performance. Achieving process robustness in manufacturing is a daunting task, because the final outcome quality is related to a number of other things (Giannetti et al., 2015).

Process performance and process variability influence how resilient the process is. Well-designed procedures decrease the possibility of human error, increasing resilience. (Glodek et al., 2006). In order to identify the crucial process inputs and outputs, their tolerances, and the most effective ways to regulate them, both inputs and outputs are examined during product and process development (Patil, 2010).

Robustness must be built into the product's design and development rather of being tested into it. Throughout scale-up, introduction, and normal manufacture, the performance of the product and process must be monitored to ensure robustness is maintained and to adapt the process and related controls as needed. The secret to creating and running a reliable process is to have a thorough grasp of how process inputs impact important product qualities. (Glodek, et al., 2006).

Principles of process robustness are as follows: 1) Critical quality attributes; 2) Critical process parameters; 3) Normal operating range and proven acceptable range; 4) Variability: source and control; and 5) Setting tolerance limits.

By addressing the tolerance synthesis problem, it is possible to select process variable tolerance limits that minimize variety in responses (Ransing, 2016). Without making any assumptions about the distributions of the variables or the linearity of the connection, likelihood ratios (LR) associated with the new tolerance limits that may be identified with quantile regression trees can be used to test the robustness of the process (Giannetti, 2017).

LR represents the ratio between conditional probabilities of obtaining optimal values and avoid values when the new tolerance limit is chosen. According to Giannetti (2017) this method is effective to isolate a small number of root causes of defects due to interactions and helps process engineers to find new tolerance limits that reduce variation in the process, hence improve its robustness.

Process robustness metrics is not a common way and standard in foundry business even the tolerance limits for many processes' variables area defined. The limits are mostly set up by product and process engineers in development phase, as an input the simulation results and own experience. To decrease this effect the modern DA platforms automatically offers and recommend to the user the best fit tolerance as kind of advice factor to determined not only critical process variable but to give higher process control for the critical process variables and by that to keep the whole process stability under higher supervision.

Most common tool in the automotive industry and even so in foundry business is SPC – statistical process control. Capability indicators are often used as indicators how the process is performing and there are direct relations to process outputs if they are low = unstable process = low quality results = low productivity levels = low profit = high costs – personal, energy, production …

**Statistical process control (SPC)** is a methodology used in quality process control. An advantage of SPC over other methods of quality control, such as "inspection", is that it emphasizes early detection and prevention of problems, rather than the correction of problems after they have occurred. SPC must be practiced in two phases: The first phase is establishment of the process, and the second phase is regular production use of the process. The most common indexes related to this topic are : cp - process capability and cpk - process performance capability.

To maintain the process under control and in the best case every process technological parameter in highest cp / cpk should lead in an ideal state to minimal waste. These indices are commonly used in the automotive segment not only for process validation before start of production, but also as an input for quality, process engineers for process improvement.

Since the quality factor is one of the elements of OEE, it demonstrates that the relationship between OEE and machine PC is linked through the quality factor and that any improvement in the capability of a machine will improve the quality factors and therefore the machine's OEE (Garza-Reyes, 2010).

Process capability is a function of the specification. The specification comes from a specification without considering the function and importance of the part, and because of that it is useless to discuss the process capabilities(Montgomery, 2005).

Montgomery is one of the experts defining and also following distribution and levels of capability indexes as shown below in the Table 1.

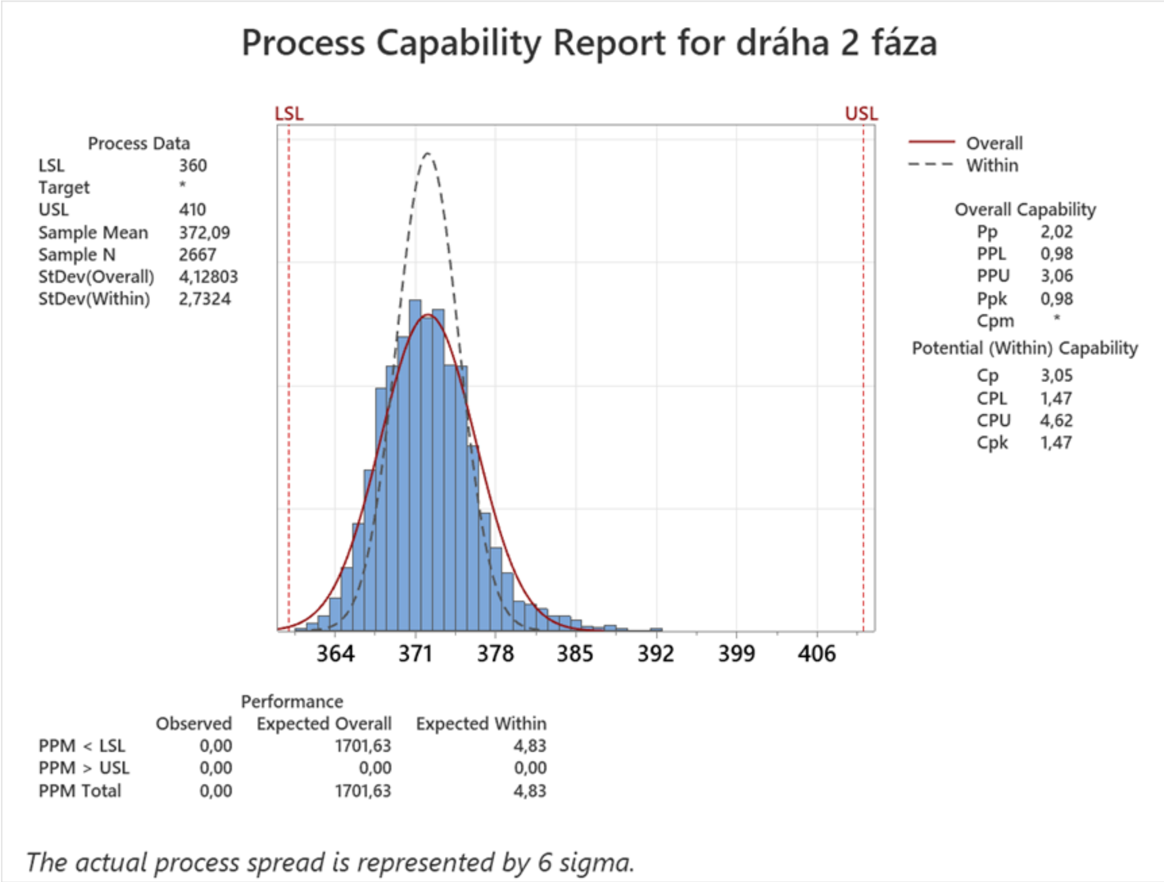
Table 1: Process capability index chart with distribution

| Situation | Recommended min process capability for two-sided specifications | Recommended min. process capability for one-sided specification |
|---|---|---|
| Existing process | 1.33 | 1.25 |
| New process | 1.50 | 1.45 |
| Safety or critical parameter for existing process | 1.50 | 1.45 |
| Safety or critical parameter for new process | 1.67 | 1.60 |
| Six Sigma quality process | 2.00 | 2.00 |

Source: Montgomery (2005)

Due to several interactions between process variables is almost impossible create an environment with zero waste and by that be closer to operational excellence where the quality levels play one of important roles. To clearly set up the process, define right technological parameters with right tolerances to keep process under control is the hardest part in the foundry business. Even the simulations in early stage of prototyping are showing some defect occurrence with some likelihood in real life in real life production problems even more factors are influencing fluent production flow. In a case study project, the company is evaluation indices cp/cpk for each critical process parameters. On Figure 3 is an example of reached results but even very high stable process set up for defined process variables is impossible to clearly define the actual root cause causing the defect formation. In other words, process is under control in terms of high cp/cpk indices with minimal variation, but total product quality doesn't reply to this process set up.

Figure 3: 2nd Phase of die filling cp/ cpk indices




Source: internal

To reach operational excellence, or to be close as much as possible for industry specific indicators f.i. quality level, without better process understanding even with the relative high process control is belonging to the mission impossible. Nakajima (Garza, Reyes 2010) mentioned that quality rate for world-class organizations should be better than 99% what correspond of cp/cpk over 1,33. What in a case of foundry business with all process variables and noises interaction without well designed casting and stable process set up hard to accomplish without integration of new tool or new approach. The lean culture with strong 6S activities which are supporting the continuous improvement process has one thing in common: quality of data, availability, and reliability of data (Tissir 2022). New revolution of Industry 4.0 offers new approach and open doors for a huge new possibility- data downloading, processing and evaluation. Unfortunately, not so many articles are written in the direction of application of data analytics with combination of 6S or lean management in real-life environment and especially in environment of the foundry business but could be stated that and integration, processing and evaluation of data's through data analytics will be the right approach where the benefits of Industry 4.0 will be combined with old known systems Lean six sigma in the operations management for bringing higher added value for the process stability and internal key indicators improvement will all other positive impacts on the overall company performance.

# 3. BIG DATA ANALYTICS

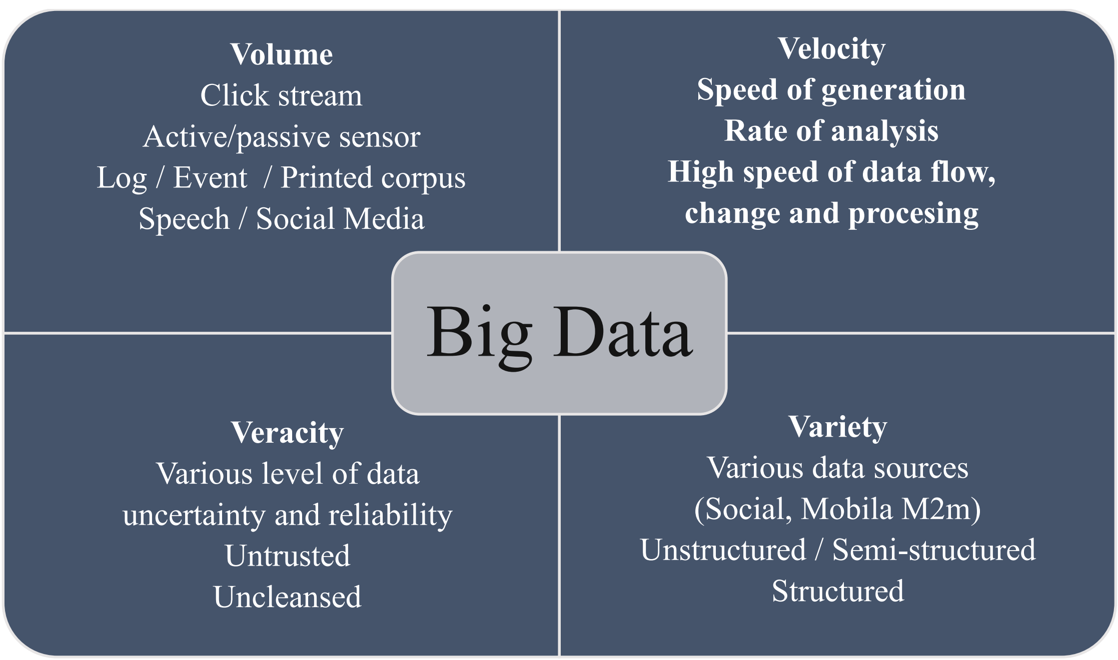
Big data analytics (BDA) is described as a holistic approach to managing, processing and analyzing large number of data's for establish potential advantages, measuring operation performance and evaluate it. BDA allows data driven decision making where together with link to the systematic approaches system like 6Sigma can be achieved the innovations and improving the operations and overall business performance. BDA is currently considered as a game changer on the field of 4$^{th}$ industrial revolution enabling improved business efficiency and effectiveness mainly due to high strategical and operational potential. In manufacturing and operation management is BDA understand as enabler of asset and business monitoring (Davenport 2014). Most companies utilize data, and they want to learn how to use and extract data from them so they can make choices more quickly and effectively. They are searching for ways to combine and correlate data from multiple areas of the company, including as finance, business development, sales, marketing, and operations, where the quality is one of the essential factors to achieving OE and business intelligence has shown to be a competitive advantage. Additionally, the companies wish to use data from other sources for additional investigation and understanding (Wauyo, 2019).

While business users have a easy to use access to data, it enables team to make data-driven decisions. When business users have access to data and the appropriate self-service analytics tools, they can turn vast volumes of data into useful knowledge. Instead of needing to write intricate queries, user can just click, choose data sets, and select choices to build their own visualizations (Shabbir & Gardezi, 2020). The trend of using big data in companies is still limited in the knowledge of defining and understanding how companies think about such innovations. This thesis will show the experience with the implementation of data analytics model and its effect on the overall performance – internal quality on special case. In following sub chapters, the short review and descriptions of big data, data analysis, data analytics, data mining approach and data algorithms will be described.

## 3.1 Big data

Big data and the way organizations use them– is changing the way the world uses business information. Big data is of vital importance everywhere, therefore numerous researchers are focusing on developing effective technologies to analyze them (Tsai, et al., 2015). Big data requires a revolutionary step forward from traditional data analysis, and it's characterized by four main components: *volume* (size of mostly in Petabytes, Exabytes and Terabytes), *velocity* (the speed of data capturing), *variety* ( can be structured, semi-structured, and unstructured datasets), and *veracity* (is related to biases, data levels, or possible abnormalities  (Russom, 2011), (Edosio, 2014), (Jelonek, 2017) (Davenport 2014) Figure 4 and Table 2.

Figure 4: Components of Big Data – 4 Vs




Source: Jelonek (2017), own elaboration

Table 2: The differences between Big Data analytics and Traditional analytics

| | **Big Data analytics** | **Traditional analytics** |
|---|---|---|
| Type of data | Unstructured formats | Formatted in rows and columns |
| Volume of data | 100 terabytes to petabytes | Tens of terabytes or less |
| Flow of data | Constant flow of data | Static pool of data |
| Analysis methods | Machine learning<br>Data analytics algorithm | Hypothesis-based |
| Primary purpose | Data-based products | Internal decision support and service |

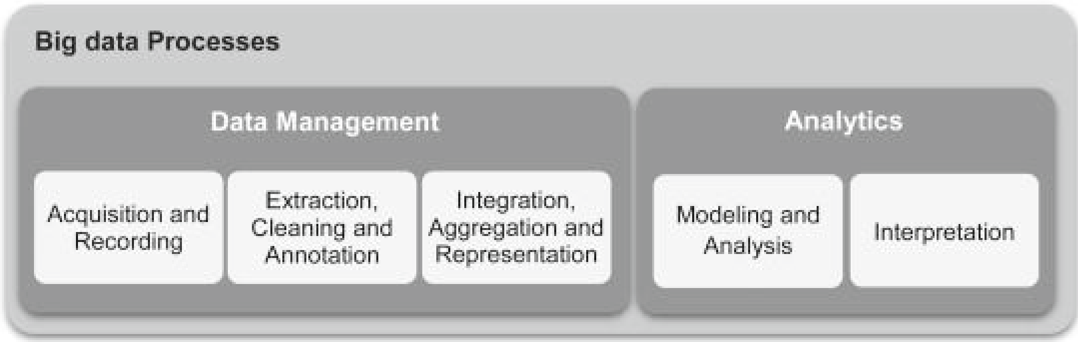
Source: Davenport (2014), own elaboration

Nowadays, the traditional data analysis methods are almost impossible to be capable of analyzing the big data as they were not designed for such large-scale and complex data. The data that need to be analyzed are not just large, but they are composed of various data types, and even including streaming data (Russom, 2011).

If we take a look at the insights from big data, we can sort them in five key stages. The five stages are divided into two main categories of sub-processes, which are data management and analytics. Data management stands for everything that is related to processes and also to the technologies, along with storing data and preparing or retrieving it for further analysis. The second category consists of analytics, which is used for all the techniques that are used for analyzing and acquiring intelligence coming from big data (Gandomi, 2015).

Below listed Fig. 5 contains processes used for extracting insights from big data.

Figure 5: Processes for extracting insights from big data.




Source: (Gandomi, 2015).

Considering the type of the industry is starting to give more attention to big data predictive analytics (BDPA) that can be a critical tool for realizing improvements in yield, particularly in any manufacturing environment in which complexity, process variability, and capacity constraints are present (Zhong et al., 2016).

A capacity of an organization is BDPA, which shows how to use BDPA to improve organizational performance (Gupta & George, 2016; Gunasekaran et al., 2017). A big data culture and human skills may be combined with strategic resources like connection and information exchange. Connectivity and information exchange are positively correlated with BDPA acceptance, which is positively correlated with BDPA assimilation under the mediation impact of BDPA routinization, under the mediation effect of top management commitment. Next important thing to mention is that connectivity and information exchange along with (Gunasekaran et al., 2017) human skills and big data culture (Gupta & George, 2016) which can enhance operational performance (Srinivasan & Swink, 2018). Providing human capital with highly analytical skills to work with big data (BD) enables more efficient and effective solutions and will undoubtedly improve the overall performance of the organization. Despite the enormous potential of new analytics technologies, tools and applications, the biggest challenge professionals are facing when using these technologies is finding employees with the necessary skills (Mikalef et al., 2018).

Dubey, et al (2019) concluded that BDPA has significant and positive effects on cost and operational performance. In BDPA, connectivity has the role of a tangible resource and data sharing is an intangible resource, which will facilitate to create BDPA capability. The study further derived that the cognitive part of institutional pressures plays a big role and from this angle, the institutional pressures square measure positive and helpful to the resource choice call, that plays a important role in building big knowledge capability to those producing organizations that square measure troubled to reap edges from investment and existing market opportunities.

## 3.2. Data analysis

Data analysis is the process of analyzing raw data to draw out meaningful insights. The information's are then used to set up next actions. Ultimately, data analysis is a crucial driver of any successful business strategy. There are a range of methods and techniques that data analysts use depending on the type of data (Stevens, 2021).

We know several methods used for data analysis – few of them are listed below:

**Regression analysis** - utilized to estimate a set of variables relationship. Estimation of how one or more variables can have effect the dependent variable, as well as the identification of trends and patterns, are the goals of regression analysis. This is especially useful for making predictions and predicting future trends. Regressions, on the other hand, can only be used to determine if there is a relationship between a set of variables because they do not reveal cause and effect. This is an important point to keep in mind.

**Factor analysis** - is a method for reducing a large number of variables down to a smaller number of factors. It works on the premise that a number of distinct, observable variables interact with one another because they are all connected to the same underlying concept. This makes it easier to find hidden patterns.

**Cohort analysis** is a subset of behavioral analytics (see chapter 3.3.2) that breaks down user data into related groups for analysis rather than looking at all users as a whole. These related groups, or cohorts, usually share common characteristics or experiences within a defined timespan. With cohort analysis, for example, customers are divided into groups, and it is observed how these groups behave over time. As a result, it is possible to identify patterns of behavior at various points in the customer journey. As such, cohort analysis is dynamic, allowing its user to uncover valuable insights about the customer lifecycle. Companies can tailor their services to specific customer segments (or cohorts) through cohort analysis, which also enables them to optimize their service offerings and marketing in order to provide a more specialized and individualized experience.

**Cluster analysis** –is a method that tries to find patterns in a dataset. The sorting of various data points into groups (or clusters) that are both internally and externally homogeneous is the objective of cluster analysis. This indicates that data points in one cluster are identical to each other but distinct from those in other clusters. Clustering can be used as a pre-processing step for other algorithms or to learn more about the data distribution in each dataset. However, it is important to note that, while cluster analysis may reveal structures within analyzed data, it will not explain why those structures exist.

**Time series analysis** – is a technique which analyze trends and cycles over time. Time series data is a series of data points that measure the same variable at different times. By examining trends over time, analysts can predict how variables of interest may change in the future. When performing time series analysis, the main patterns to look for in your data are

a) Trends: steady linear increase or decrease over a period.

b) Seasonality: Predictable fluctuations in data due to short-term seasonal factors.

c) Circular pattern: An unpredictable cycle in which data fluctuates. Cyclic trends are not seasonal and can be driven by economies and industries.

**Sentiment analysis** – is a broad category of techniques that can be used to sort and understand textual data. With sentiment analysis, the goal is to interpret and classify sentiment conveyed in textual data. This technique allows businesses to assess how customers feel about various aspects of a brand, product, or service. We recognize several sentiment models. Some models focus on detecting positive or negative sentiment in text, while others focus on detecting overall sentiment in a given piece of content. It's important to choose the right sentiment analysis model for the task at hand, as each has its own benefits and limitations.

## 3.3 Data mining, data analytics, machine learning and data algorithms

### 3.3.1 Data mining methodology CRISP-DM

At first it is necessary to define the Data Mining (DM) objectives. DM is the search for new knowledge in data. This knowledge often comes in the form of rules that were previously unknown to the user and may be useful in the future. These rules can be in the form of specific rules inferred using rule derivation algorithms, or they can be more general statistical rules, such as those found in the prediction model. The origin of these rules is determined based on the data mining tasks. In this case, typical tasks involve categorizing or grouping data. A highly desirable characteristic of data mining is having a high-level user interface that allows the end user to identify problems and retrieve results in the most user-friendly way possible. While experts can perform data mining and interpret the results to users, it is also desirable to allow users to perform their own data mining and draw their own conclusions from the data. new material. Therefore, a corresponding user interface is very important (McClean, 2003).

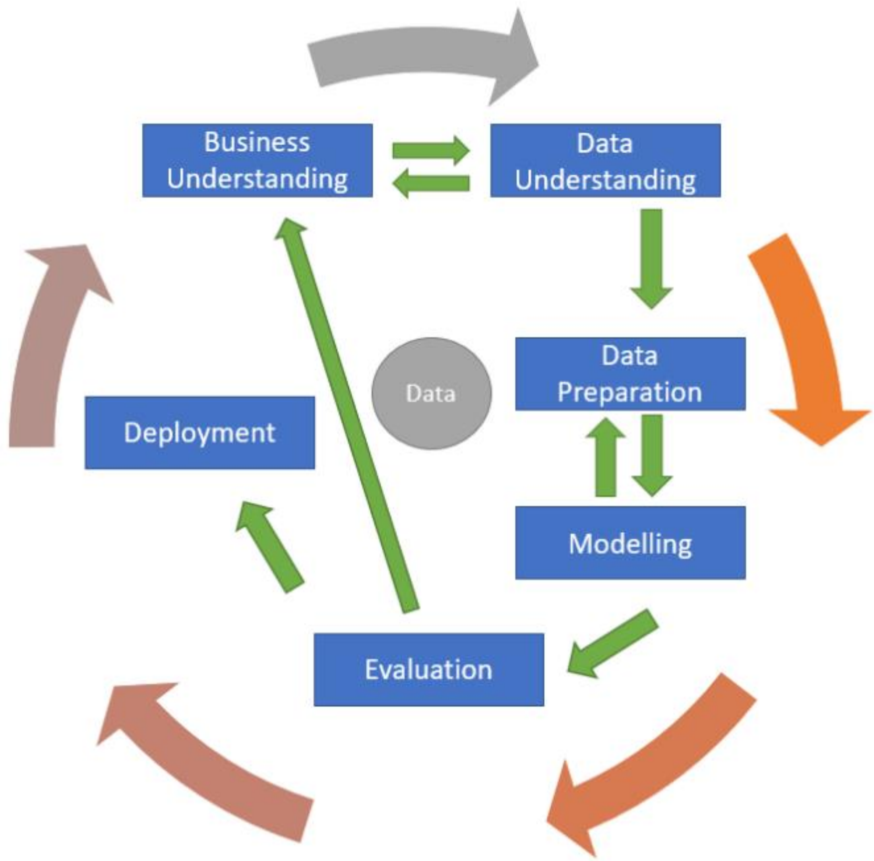
**Cross Industry Standard Process for Data Mining** (CRISP-DM) is a methodology that is independent of the particular industry in which it is being used, and it is generally considered to be a standard approach. (Azevedo & Santos, 2008). CRISP-DM presents a hierarchical and iterative process model and provides an extendable framework that can be tailored depending on the DM context. (Niakšu, 2015).

CRISP-DM defines following data mining context dimensions: a) application domain; b) problem type; c) technical aspect; and d) tools & techniques. It is considered the most referenced and practiced of DM methodologies for defining the phases, tasks, activities and outcomes of these tasks. CRISP-DM is an informal methodology as it does not provide a rigid framework, metrics, or standards of accuracy. However, it offers the most comprehensive set of tools ever for DM practitioners.

CRISP-DM proposes an iterative process flow with a loosely defined loop between phases and the overall iterative cyclical nature of the DM project itself. The outcome of each phase determines which phase should be executed next.

The six phases of CRISP-DM are as follows (Niakšu, 2015): 1) Business understanding, 2) Data understanding, 3) Data preparation, 4) Modelling, 5) Evaluation, 6) Deployment.

Figure 6: CRISP-DM – model




Source: Wijaya (2021)

The order of phases is not strict. Going back and forth between different phases is always necessary. The next phase or specific task in a phase depends on the outcome of each phase. Arrows indicate the most important and most common dependencies between phases. The outer circle in Figure 6 symbolizes the cyclical nature of DM itself. DM is not finished once a solution is deployed (Chapman, et al., 2000).

### 3.3.2 Data Analytics

Big data analytical techniques for structured and unstructured data can be described by a very long list of the techniques, so the following big data analytics present a subgroup of tools which can be used.

**Text analytics** can also be called text mining and is based on retrieving information from textual data. Textual data can be retrieved f.i. from social media and related networks (He Wu, 2013), from communication (emails, call center logs and other), from the news or from some on-line blogs, forums, and a variety of other sources. One of the biggest pros of text analytics methods is the fact that companies can transfer big volume of data. These can be than used for summaries, reports, and further decisions in the company (Gandomi, 2015). One of the methods of text analytics is information extraction which is based on retrieving structured data from unstructured text. These methods can be used f.i. in algorithms that are developed for extracting large amount of data – one of the examples in which this can be used is product name and product number along with related data (Jiang, 2012). Other methods include text summarization (producing a summary out of several resources), question answering (QA) which provide answers to natural given questions, and others (Gandomi, 2015).

**Audio analytics** is a method based on analyzing and retrieving information from unstructured audio sources. Audio analytics can be used in a variety of fields, including healthcare, customer centers and others.

**Video analytics** includes a variety of techniques which make it possible to record, monitor, analyze and extract useful information from video. Compared to the other methods of analytics, video analytics are not so much developed (Gandomi, 2015).

**Behavior analytics** is s a recent advancement in business analytics that reveals new insights into the behavior of consumers on eCommerce platforms, online games, web and mobile applications, and IoT.

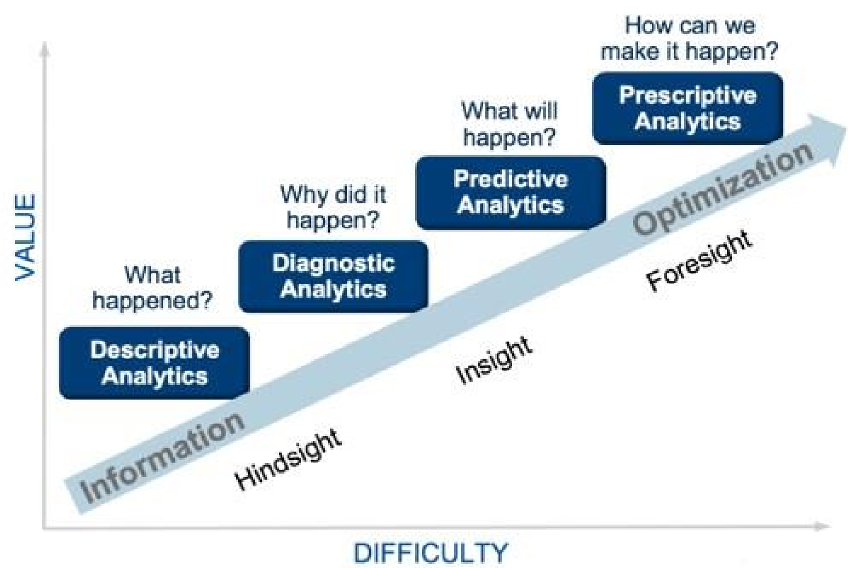
**Social media analytics** is done through analyzing structured and structured data on social media networks. Social media is a term used for describing a wide category of social media platforms that are based on sharing some type of content between users. One of the biggest categories include social networks like Facebook and Twitter, blogs (WordPress), forums (Reddit, Quora), media sharing platforms (Youtube and Instagram), wikis, reviews sites and also traveling and other interests focused platforms (Gandomi, 2015). Especially in the last years, social media analytics has become a very important source of data.

Obtained data from these sources can be used afterwards in several processes and answering questions in a number of research fields.

### 3.3.3 Data science and machine learning

Evolution of decision-making using analytics can be expressed through the Garner analytic ascendency model – Figure 7 - that depicts four levels of data analysis usage within a company. According to Vashisth, (2019) innovations as auto-ML, machine learning operationalization (MLOps), and edge analytics will have a significant impact on data science programs over the next two to five years. Predictive analytics and prescriptive analytics will assist organizations in maturing beyond descriptive and diagnostic analytics.

Figure 7: Gartner Analytic Ascendency Model




Source: Sentance, (2017)

Strategic initiatives like digital business transformation with intelligence at its core are seen as driving the highest priorities for data science and machine learning.

Generating predictive insights from asynchronous and highly heterogeneous time-series data at scale is a significant technical challenge for any successful ML approach. Creating business value from connected process data (Minevich, 2020).

Four key trends impacting the data science and ML market, such as: 1) Augmented ML; 2) Data readiness and management; 3) Scaling and operationalization; and 4) Decision management. (Vashisth, 2019).

If we take a look on the relation between ML and data science, machine learning allows artificial intelligence to enable computers to learn and anticipate in the possible outputs without being taught to do so. As a result, they become more human. Machine learning models continually learn and enhance their performance by utilizing the necessary data. Compared to that, data science is a wide, interdisciplinary area that employs ML methods when precise estimations on a particular data collection are required. While data science assists us in focusing on and analyzing the topic at hand, ML assists us in developing real-world applications to tackle the problem we have found. They cannot function independently, and these two sectors must be combined to give the best outcomes (Nicosia et al., 2020).

By utilizing automation and embedded machine learning to speed up the creation, tuning, and deployment of models, innovations with "transformational" benefits will alter the field of data science. Digital ethics thinking will also be used by businesses to anticipate whether applying data science to particular use cases is the right thing to do (Vashisth, 2019).

It is also important to mention the DA types which can be sometimes divided into a little bit different categories depending on the author. But mostly the types of DA are containing the following types:

**Descriptive**: the output is to organize large amounts of data into a data set, can be done by asking, "What is happening?"

**Diagnostic**: Analyses data from the past, These information's will be used as a base for improvements in business or product. It helps answering the question, why something happened.

**Prescriptive**: Is mostly combined by AI and big data, which helps to predict the future outcomes. This category can be further divided into part of optimization and part of random testing. It answers the question “What happens if this action happens”?

**Predictive**: The most used category, it helps in the area of identifying current trends, the two sub-categories of this part can be further divided into predictive modelling and statistical modelling. As it was stated before, the Table 3 below shows the relation between the dimensions of maturity and business impact through the years.

Table 3: Priority Matrix for Data Science and ML

| **Benefit** | **Less than 2 years** | **2 to 5 years** | **5 to 10 years** | **More than 10 years** |
|---|---|---|---|---|
| Transformational | | Advanced video / image analytics<br>Augmented analytics<br>Citizen data science<br>Deep neural networks (Deep learning)<br>Event stream processing | AI cloud services<br>Continuous intelligence<br>Conversational user interfaces<br>Generative adversarial networks | Quantum ML |
| High | Notebooks<br>Traditional model management | Augmented data management<br>AutoML<br>MLOps<br>Predictive analytics<br>Prescriptive analytics | Adaptive ML<br>Decision management<br>Digital ethics<br>Explainable AI<br>Federated machine learning<br>Graph analytics<br>Reinforcement learning<br>Synthetic data | |
| Moderate | | Data labelling and annotation services<br>Python<br>Spark<br>Text Analytics | Advanced anomaly detection<br>Transfer learning | |
| Low | | | | |

Source: Vashisth, (2019), own elaboration

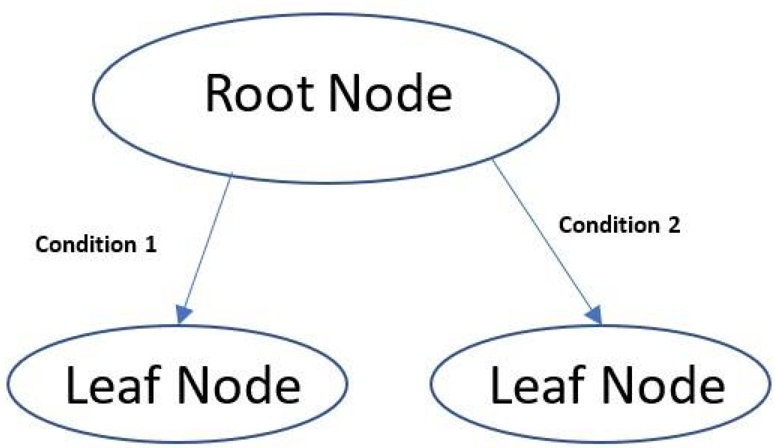
### 3.3.4 Data Analytics algorithms

The world of data analytics is constantly changing and evolving. Data analytics is the science of analyzing raw data to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption. Below some examples of most common used data analytics algorithms :

**Decision Trees**

Decision tree (DT) is a method used both for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The description of the model is visible on the Figure 8.

Figure 8: Decision tree model




Source: Pinto, Own elaboration

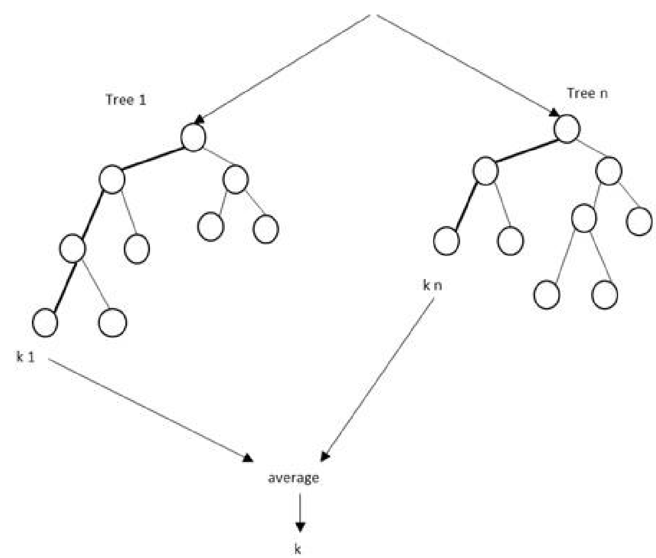
The most fundamental algorithm for decision tree learning is a top-down approach that begins by selecting the best attribute to test at the tree's root node.

**Random Forest (RF)**

The term "forest" refers to the collection of decision trees that make up the random forest. To classify a new object based on its attributes, each tree provides a classification, and that classification is chosen by voting for that tree. In addition, since the forest selects the classification with the most votes overall, each tree votes in the truest sense for a classification. A group of B trees is Random Forest. B outputs, one for each tree, are created in order to produce an output for an object. Each tree votes, and the classifier with the most votes is chosen if it is a classification algorithm. An average of the set of output values is used to calculate the final output in a regression algorithm. (Breiman 2001).

A scheme of a random forest may be observed on Figure 9.

Figure 9: Random Forest model




Source: Pinto, Own elaboration

Breiman (2001) and other authors like Kuhnlein (2014) describe RF algorithm like n tree bootstrap samples are randomly selected from the data set.

RF is commonly used where combined data sets is necessary to use. HPDC process variables is exact that case. Its regression analysis is a combination of results of averages each trees / node where AI engineer is setting up the:

- Amount of decision trees (the higher number of trees = higher complexity and probability that model will find the “right “combination” of searching defect against process variables)
- Depth of the trees (deeper tree = algorithm separate defect against non-defect data more precisely what is necessary factor for model training. There is a risk by very

complex tree design that model will be “very educated” from training data set what can cause -overfitting effect – effect that while he gets new data set it will not recognize results.

- Minimum amount of data in one leaf
- Maximum amount of data in one decision node. Parameters are chosen randomly

**Partial Least Squares (PLS)**

Partial least squares are a linear regression technique. Use principal component analysis to decompose the prediction and target matrices into principal components that maximize the variance of the data. Previously correlated variables are replaced with a smaller set of uncorrelated variables that contain almost the same information. In this way the data is projected into a smaller dimensional subspace while preserving most of the information. A regression is then performed between the principal components of each variable (Pinto, 2016).

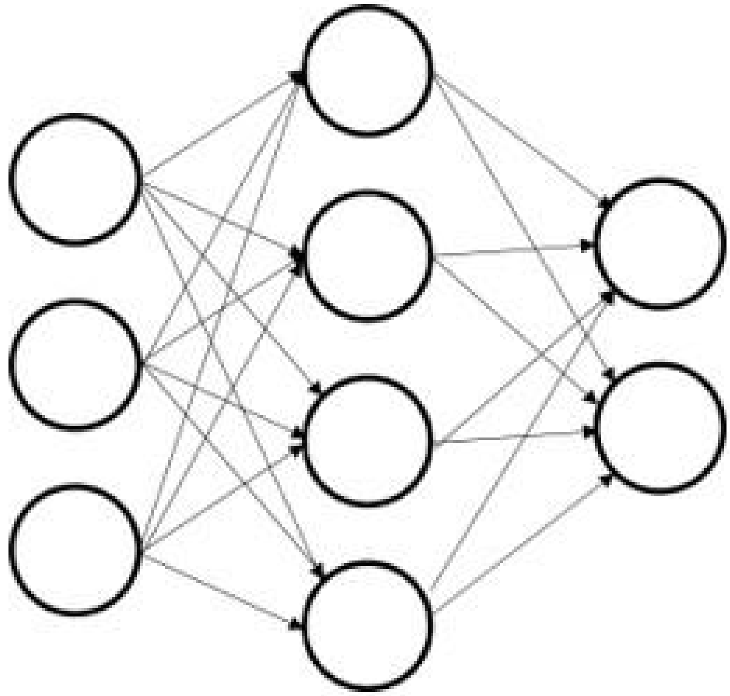
**K-Nearest Neighbors**

algorithm is often implemented and used mainly for solving classification related problems. Besides that, it can be used for solving regression problems. The algorithm works based on saving available cases and classify all new cases by voting from K neighbors. New cases are assigned to the class with the most frequent attributes.

**Artificial Neural Networks (ANN)**

Artificial neural networks work in similar way and on the same principle as the human brain where the neuron receives a signal and sends electrical signals to other neurons. This process is also known as synapse. ANNs try to copy paste this the neural activity. Most of the functions are on the field of scheme and pattern recognition. Information is processed through a parallel network of neurons.

Figure 10: Model of neural connections



Source: Pinto, Own elaboration

Depending on the information that passes through them, the nodes alter their internal information structure. Weights – a number associated with each connection that controls the signal between two neurons—are manipulated during learning. Weights need not be changed if the ANN produces the desired output. On the other hand, the weights are adjusted if the error is significant (Shiffmann 2012).

There are many advantages of ANN as well used in manufacturing area and by Huang and Zhang (1993) we can recognize benefits like a) adaptive learning, b) self-organization, c) real time operation.

## 3.4 Data analytics towards process control in manufacturing process of high pressure die casting

Large data sets make it possible to improve quality improvement efforts by better controlling processes and making real-time adjustments based on the analysis of continuous data streams (Manyika, 2012).

In 2017 Khan et al. investigated some of the difficulties that are associated with big data analysis in manufacturing as well as pertinent solutions to two major computational difficulties: memory and computing capacity limitations. He came to the conclusion in the study that the majority of small and medium-sized businesses should look into cloud computing because it can be less expensive than making a one-time investment in acquiring computational capabilities and expertise. Another important conclusion was that achieving scalability is considered a key programming objective in the field of big data analysis. Moreover, on-line information processing or real time data analysis is indeed a crucial concern in a smart factory.

In 2013 Perzyk, Kochanski, Kozlowski and team evaluate of various methodologies used to determine relative significance of input variables in data driven models. Authors used decision trees (DT), Rough Set Theory model (RST) as newer alternative to DT and Naïve Bayesian Classifier (NBC) classification system. Support Vector Machines (SVMs) is competitive to ANNs model and compare these two models. The best accuracies of input variable significances tested on large data sets were obtained using RST based methods. In a case when only small and noisy data are available the best results in identifying the most significant variables were reached by DT model. For industrial data sets practically, all model appeared to be satisfactory. The work dealt with significance analysis for single variables only.

In 2006 Hovhannes Sadoyan and his team presented data mining algorithm based on RST – Rough Set Theory which against decisions trees and neural networks seems better in its accuracy and user-friendly usage.

More details about actual trends are mentioned in literature retrieval in chapter 5.1.

## 3.5 Interconnection between Big Data analytics and operational excellence

Gunasekaran et al.(2017) describe BDA tools as support for a significant business benefit to drive an organizational improvements. BDA elicits two major viewpoints to achieve the OE of the organizations. *First,* the collection of big data from the firm and external environment. When compared to the current form of traditional data processing systems, this type of data suggests a high volume and velocity of data processing that can provide numerous benefits (Frank et al., 2019). *Second,* how big data is used in business analytics to help make decisions and run operations better. The ability to evaluate an organization's strategic move in order to achieve successful business planning is at the heart of business analytics. Business analytics strategic enhancements, such as forecasting, statistical analysis, and operational analysis through optimization techniques, significantly contribute to the improved OE (Mathivathanan et al., 2018). A complete concept of BDA emerges when these two aspects are combined. Companies gain a competitive advantage from BDA, which also leads to timely and accurate decisions. (Bag et al., 2020). Gains in operational efficiency and sound business decisions are often the foundation of increased competitiveness (Olugu et al., 2011). The BDA phenomenon's contribution to the organizations' strategic goals is widely acknowledged by corporations. In this fourth industrial revolution, BDA can be used to enable plant automation (Tseng et al., 2018).

As was mentioned earlier there are not many available articles of direct application data analytics in operations of HPDC and linking to operational excellence. From the known literature review of few authors are visible the benefits of the DA towards stabilization and improvement of performance.

Perzyk, Dybowski and Kozlowski (2019) executed a development project related to Industry 4.0 in Poland's high pressure die casting foundry. The developed data analytics system aimed at predicting the casting quality basing on the production data. They conclude that these data’s can be used for optimizing process parameters to raise the products’ quality as well as to improve the productivity. The system incorporated advanced data analytics and computation tools based on the analysis of variance (ANOVA) and applied an MS Excel platform. To guarantee the aluminum engine block castings high mechanical properties. Finally it helps  foundry engineers and operators to identify the most effective process variables.

Linking BDA and operational sustainability also mentioned R.D. Raut and team (2019) where he suggest that BDA can transform the manufacturing industry to apply sustainable practices more efficiently.

The Dubey, Rameshwar, Gunasekaran and team (2019) put together importance of big data and predictive analytics in supply chain for its improvement and operational performance. Based on their research confirm the effect on cost and significant effect on the operational performance.

The Surajit Bag, et al. (2019) research shows the big data analytics managements had a strong and significant effect on the innovative green product development and sustainable supply chain outcome with effect on the employee development).

All authors mentioned a positive impact of BDA on the operational performance, unfortunately there are not many articles with direct application of DA in case studies – except Perzyk and effect on operational excellence in practical real case studies especially in HPDC industry. Based on the experience of authors with different DA applications it can be considered positive impact of the focused kpi.

# 4. RESEARCH ASSIGNMENT

## 4. 1 Literature study and information retrieval

In foundry operation only few applications of DA were applied. The majority are focused on the semiconductors industry, iron and steel industry, metal processing industries etc.

Literature and other relevant sources in the theoretical part were mostly new articles in books, because the main theoretical topics that are listed in this topic are very fast improving and coming up with new data and information.

For better orientation in the information about literature review and information retrieval, the category will be divided into the same sections as the categories of this thesis.

Operational excellence

The introduction of operational excellence (OE) was described by Tissir (2022) and Dugan (2011). Both authors defined OE as the flow of value to the customer. Treacy & Wiersema (2017) added to this statement, that OE is specific only for one company, and what works in one, cannot fully match in other company because of the individual culture, needs and visions of every company.

Mitchell (2014) defined the main questions of OE as “where we must be” which is based on the management answer. Wilkings (2021) then defined the most important metrics of performed, which in his words are productivity and quality. Of course, the most important things can vary from author to author, which is affected by his area of interest and further activities related to his publishing.

After a few more theoretical issues we moved onto process capability, which was defined in 2005 by Montgomery. Montgomery belongs to one of these authors that are expert in this area of research. This thesis also demonstrated his process capability index with distribution as an example.

Next and one of the primary parts of this thesis was dedicated to data analytics. At first, the concept of data analytics along with big data was demonstrated, which was then followed by specific information related to big data processing and analyzing. In this part it's important to say that most of the authors here give kind of similar theoretical definition, which are different in some wider or on the opposite less detailed characterizes.

Afterwards, the most used methods of data analysis and data analytics were described. As it was stated before, some of the authors like Stevens (2021) provide a very well-structured definition of theoretical aspects, and compared to that for e.g. McClean (2003) gives a more detailed definition. After describing the 6S, a detailed description of the six phases of CRISP-DM by Niakšu (2015) was presented.

If we move from theoretical part to real life improvements, analysis, implementations, and other action, we can point out these authors and their activities:

Chen, Wang, and Kuo (2021) applied a data-driven approach to improve the furnace zone of a foundry in Taiwan. Improvements were based on historical production records, order plans, and work plan data. To address the bottlenecks provided by the company, an analysis of historical data revealed significant variations in the process. Statistical analysis was performed to identify the main factors causing deviations, suggestions were made and implemented in the production line. Significant improvements in production have been achieved, primarily through input standardization and reduced process variability. In summary, data analysis helps identify key factors causing bottlenecks in traditional industries. Afterwards the subject of data analysis types was described, which was followed by the priority matrix.

The main goal of the theoretical part was to bring up different opinions to present a complex view on the whole problematic. Along with that, for better and easier understanding a variety of figures and tables were attached in the text. In the following part of this chapter the actual situation in HPDC industry will be shown with application of data analytics. The order is from the last published articles to the oldest ones.

In 2019 Perzyk, Dybowski and Kozlowski executed a development project related to Industry 4.0 a leading manufacturer of the aluminum castings for the automotive industry, in its high pressure die casting foundry in Poland. The developed data analytics system aimed at predicting the casting quality basing on the production data

In 2019 A.Sata, B.Ravi tested data analytics based on Bayesian inference on investment casting focusing to calculate posterior probability for defined process parameters. Parallel to posterior probability also involves computation of joint probability, prior odd, likelihood ratio. This approach according to authors is relatively easy to implement against ANN networks

In 2018 Raza et al. defined those variations in casting parameters can lead to undesirable casting conditions that lead to the formation of defects, so we observed how sensitive the castability of thin-walled castings is to variations in casting parameters. In this work, variations in casting parameters due to human involvement in the process were investigated. The casting practices of different groups of casters were evaluated and the resulting variations in casting parameters were discussed. The impact of these variations was assessed by comparing the denial statistics of each group. It was concluded that variations in casting parameters due to differences in casting practices in different groups have a significant impact on casting quality. Due to variations in handling times and procedures, component quality varied from batch to batch. This was caused by variations in the temperature of the mold, the temperature of the melt, and the speed of the injection. The implementation of optimized casting instructions has significantly improved quality and process reliability.

In 2017 Raza et al. used simulation tools in predicting mold filling behavior. The outcome of the prediction was mostly based on the material and on the individual set-ups of simulations. The measurement was to replicate the process casting through the simulation. Evaluation of the findings was done by comparing the outcome of simulation with the original cast samples. The differences between the simulation outcomes and the original cast samples were discussed in a number of process variations. Based on these findings, it was deduced that the simulation was able to replicate the asymmetries in the filling, but it was not able to predict the final value of the area.

In 2017 E. Battaglia, F. Bonollo and his team applied a variety of differences between process, microstructure, and properties on experimental die horseshoe. Differences between process parameters and casting quality have been done by analyzing the obtained data. It has shown the strong impact the second phase plunger velocity had, along with temperature and the pressure on defect foof second phase plunger velocity, temperature, and pressure on defect formation.

In 2016 A. Sata tried to find correct ranges of defects' values to ensure casting quality by using data analytics based on Bayesian inference. Afterwards, the methodology was defined and programmed using Microsoft Excel. The spreadsheets were perfect for engineers, because they had the opportunity in practice implementation. The successful test was done a steel valve body, which was complemented in investment casting. The findings were afterwards used for optimization and controlling, which had also helped reducing rejection parameters of the main defects. The approach was easier to implement, and use compared to process simulation. There was opportunity for building other models (sand casting and die casting) along with other processes.

J. Kittur et al. (2015) chose for his work response surface methodology focused on process parameters – fast shot velocity, injection pressure, phase change over point and holding time as input. Porosity, surface roughness and hardness were used as response -output parameters. Two non-linear models were developed testing on ideal rectangle shape model using Central Composite design and Box-Behnken design. These two models were tested for their statistical adequacy and prediction accuracy through analysis of variation (ANOVA) and some practical test cases.

In 2014 Sushovan et al. defined framework for improving the quality in the steel foundry. The DMAIC framework was proposed, driven by data mining techniques for defect diagnosis and quality improvement with used two algorithms – classification and regression tree and chi -squared automation interaction detection. The results demonstrated significant reduction of casting defects.,

Jian Zheng, et al. in 2009 focused in their work on optimization of HPDC parameters using ANN network. They focused on evaluation system for the surface defects where ANN network generalizes the correlation between surface defects (hot cracks, cold shot, misrun) and die casting parameters such mold temperature, pouring temperature and injection velocity. Based on experiment with mentioned parameters on magnesium alloy and simple design real life product was observed that ANN model had great forecast ability (BP learning system was used) and with optimal parameters set up the acceptable casting surface quality was reached.

In 2008 Ko-Ta Chiang, et al. have done a model analysis of effect processing parameters. The already mentioned response surface methodology was used in this case. The design was done on Al-Si alloys on ideal experimental rectangular cast with the basis of the central composite design. The average values of mean particle size of primary silicon and the material hardness was analyzed and conducted the mathematical model can be used to predict for mentioned values.

In 2007 D. B. Karunakar and G.L. Datta focused with their research on how the back propagation neural network can be align as prevention casting defect in green sand steel casting. The neural network was trained with defined parameters which has major influence on the casting quality – like cracks, misrun, scabs. The ANN could successfully predict another defect. A very important feature of network is ability to learn from environment and adapt to it. In study was presented back -propagation training algorithm where the weights initially are calculated randomly.

By authors Jitender et al. in 2007 the selection of optimal casting parameters is recognized as a complex non-linear problem involving a large number of interrelated process variables, each of which influences the flow behavior of the molten metal and thus part quality and productivity. They proposed a neural network casting process model (NN-CastPro) for estimating optimal HPDC parameters in the real world. Values for melt temperature, initial mold temperature, first stage velocity, and second stage velocity were used as input parameters. A neural network was trained using data from ProCast (simulation software) to obtain predictive accuracy, which showed improved performance over other models available. Simulations were done on simple part and proposed ANN model was trained using back propagation algorithm for determine error derivate of weight.

In 2004 authors around A. Krimpenis, P.G. Benardos used neural nets and genetic algorithm in HPDC process using knowledges gained from casting simulation software focus and connecting process variables such as gating velocity, die temperature against output variables such a filling time. For defect prediction was used learning vector quantization model. For a genetic algorithm was used stochastic universal sampling. Generally according to authors ANN models can substitute simulations software in HPDC what shorted time for development can be due to less simulation runs.

In 1999 Prasad K. D. V. Yarlangadda et. al. presents the ANN system in generation of process parameters for HPDC process. Based on the results the neural network was able to select the process parameters much simpler as before a standard user.

As we see as the results of above-mentioned overview there are several experiences with implementation of variety models and artificial networks in casting industry. There is no common approach which could be considered as benchmark for next researchers what was considered in chapter 4.2.

## 4. 2 Thesis Novelty

There is no casting process which doesn't produces a certain level of scrap. This rejection or quality nonconformities is closely related to the type of casting, the process used and the equipment available. Casting quality depends on many process parameters. Each parameter affects the types of defects displayed. However, in most foundries, a significant portion of the failures are due to poor control of critical quality (CTQ) parameters commonly found in non-automated casting processes. (Raza, 2017). These variations in CTQ parameters are often overlooked while analyzing the cause of rejections. Rejected products, resulting from casting defects, are recycled a re-melted (Shivappa et al., 2012).

In HPDC gravity die casting molten aluminum is injected into chamber and by metallic piston - plunger by defined parameters shot into steel tool as shown in the figure 12 below.

HPDC process is set up from different phases. The casting process parameters are a group that can have impact on quality or also productivity of the machine (Jitenger et al. ,2007) that can be afterwards divided into four main categories– a) parametrs like plunger velocity, injection pressure...b) cast metal related- melt temperature, c) die parameters –

venting/vacuuming, , die initial temperature etc, d) parameters such as shot sleeve filling level.

However, it is necessary to mentioned that casting technology is sensitive to any changes in casting parameters. Variations in casting parameters can lead to undesirable casting conditions that lead to the formation of defects. Batch-to-batch variations in rejection rates due to casting defects are a common problem in foundries and the sources of these variations usually remain unknown due to the complexity of the process. In this situation it a mix of factors like f.e. mold temperature is dependent on the : cycle time, design of the casting itself, size of the water flow, amount of cooling circuits, initiation of cooling (which should be specified in the program), on the water temperature in the system, on the cooling design, on the spraying program etc.

Fluctuations in the target values of casting parameters also limit the use of casting process simulations. Casting process simulations has become an important tool the help designers and product engineers to determine critical zones in the design of casting. By that, sooner that die is manufactured, the appropriate actions are defined. In a case of incorrect simulation set up was used, the whole simulated conditions of the casting process will lead into incorrect decisions.

By relating variations in casting parameter target values to scrap/rework rates, the importance of variations in casting operations is determined (Raza, 2018). To minimize the process variations is need kind of standardization process, focused on casting parameters appeared to be critical in terms of their effect on quality of castings. The modern casting machines allows track and evaluation capability indices of critical technological process parameters and on-line evaluating process capability.

Based on the research and evaluation of articles of authors - researchers mainly described above, the scientific community was focused to define the best machine learning and tested their effects but on the ideal cases or ideal castings = the design of the casting is close to perfection. They have implemented and successfully tested several models not only based on ANNs or genetic algorithm, Bayesian models, Rough Set Theory models or setting up the likelihood ratios. Only few of them were used their model in real case study. Few studies were focusing on aluminum HPDC foundry area where the results can be different due to effect of another material as steel or manganese.

The novelty of this study will be in an experiment in real life product in the environment of best-in-class high pressure aluminum foundry of well-known producer in middle of Slovakia. The novelty is also based on the experiment which will be focused to test the **model** based on data analytics technique **Random Forrest (RF)** which has been proposed. Random forest algorithm is one the most common methods for classification in the machine learning field, which was mentioned and described in detail by Hatwell (2020). It is one of the most accurate learning algorithms available and it offers specific features that make it attractive for remote sensing applications. It runs efficiently on large data sets; it is simple and can easily be applied to parallel computing platforms and it can capture non-linear association patterns between predictors and response. (Kuhnlein 2014). RF is more commonly applied in other disciplines as in medical health, bioinformatics, weather prediction – rainfalls

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. This algorithm is most frequently used for the extraction of decision rules from the dataset and often used in the manufacturing processes where belongs also casting processes.

So far in the practical cases in the literature review wasn't used as well as in the environment of the aluminum foundry with experience in real product from product family – battery housing.

On chosen product from portfolio battery housing - the work will be focused on internal defect issue – **shrinkage porosity** in define area where so far, no author tried to find determining model for such a defect formation. The model results will be compared against ANOVA regression analysis to see if the effect of significance will be confirmed. It will be used one - way significance analysis. The methodologic scientific approach with implementation DA model will follow the CRISP- DM methodology as a common practice in such a case.

## 4. 3 Case study - description of internal environment

### 4.3.1 High pressure die casting process description

Casting technology is often used to produce complicated product designs. Through its combination of high performance and very economical production of aluminum or magnesium alloy components it is possible to obtain light weighting castings with high standards for mechanical, dimensional / superficial requirements. The high-pressure die casting (HPDC) had in last decades high development and implementation in modern metal industry precisely because of development in the automotive industry. Due to an environmental regulation the automotive industry had to adapt and started reducing weights of the final products - cars. Through it many car steel components were replaced by lights metal alloy.

In HPDC process there are many parallel and sequential actions – see process variables on figure 11, but generally the whole process can be described as follows: 1. injection of molten metal into the die by piston during phase of 1st and 2nd with vacuuming in parallel, 2. pressure increase phase to improve homogeneity, and solidification, 3. part extraction, 4. lubrication / spraying of die surface.

There are two kinds of process variables that need to be distinguished (Perzyk, Dybowski & Kozlowski, 2019):

**1) Fixed process parameters** – are define as a major process parameter. If a major process defect is not occurring, their variations are negligible for the part quality. These are, amongst others, $1^{st.}$ and $2^{nd.}$ phase profile, applied pressure, etc.

**2) Dependent parameters** – these parameters can be influenced by process environment such as humidity, temperature, number of parts casts in series,

K. D. V. Yarlangadda define HPDC process parameters that are constantly in conflict in a complex way. HPDC has a considerably shorth cycle time against other casting processes, but there is no common rule or “one process standard” which can be unified mainly due to the complexity of the processes and number of variables needed to be controlled. There are issues related to die temperature control, solidification of the product. (L. X. Kong, et.al 2007). The combination of metal and die temperature, casting design, mold filling, cooling rates and cooling designs during the casting interact each other and influence the integrity of the casting and final quality results of the product.

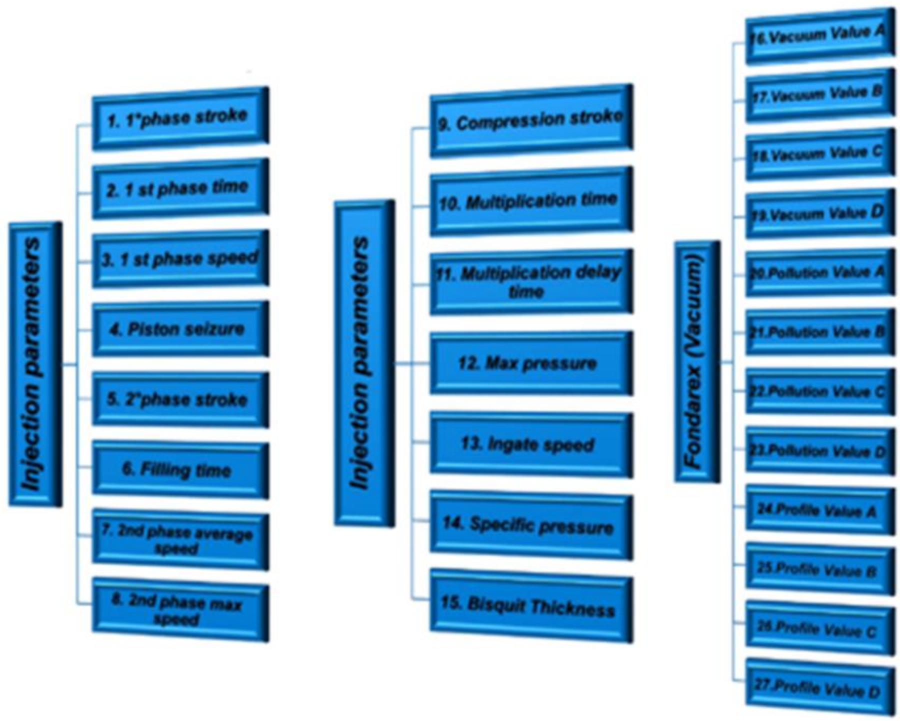
The pressure increasing applied during the casting solidification is also important and has a final effect on the integrity and microstructure homogeneity. There are a lot of studies where authors were looking for dependencies between micro porosity as output parameter and casting velocity as input (Perzyk 2019).

The thermal profile of the die as was mentioned in chapter 4.2 is another important factor and has many depended variables which interact among each other. For example, high temperature of the die will lead to longer solidification time and prolong the cycle time and could have effect on die filling as well or hot spot creation in the casting if the die doesn't have enough cooling spots. Cold die in opposite, will contribute to several surface casting defects. All in all, the HPDC casting process is a complicated manufacturing process with many sub-processes and variables which interact with each other. To find a correct set up of the process and technological parameters (even with a lot of information's about process parameters) without additional technological and analytical support is a long term run for the process engineers or production specialist. Under customer serial demands of product deliveries after start of production (SOP) the possibilities for dramatic tooling change, completely other set up of technological parameters is very limited due to the limited time and customer product validation what is necessary if any process, design change is occurred.

In a case study the correlation of in total 37 parameters is done where 27 parameters are direct controlled parameters (Figure 11) and additional 10 parameters such as: lubricant amount, pressure of water a mixture, metal temperature, sequence number of casting, evidence of cooling flow rates, evidence of die temperature before and after.

All parameters are saved into internal database.

Figure 11: Process Variables in HPDC process




Source: internal

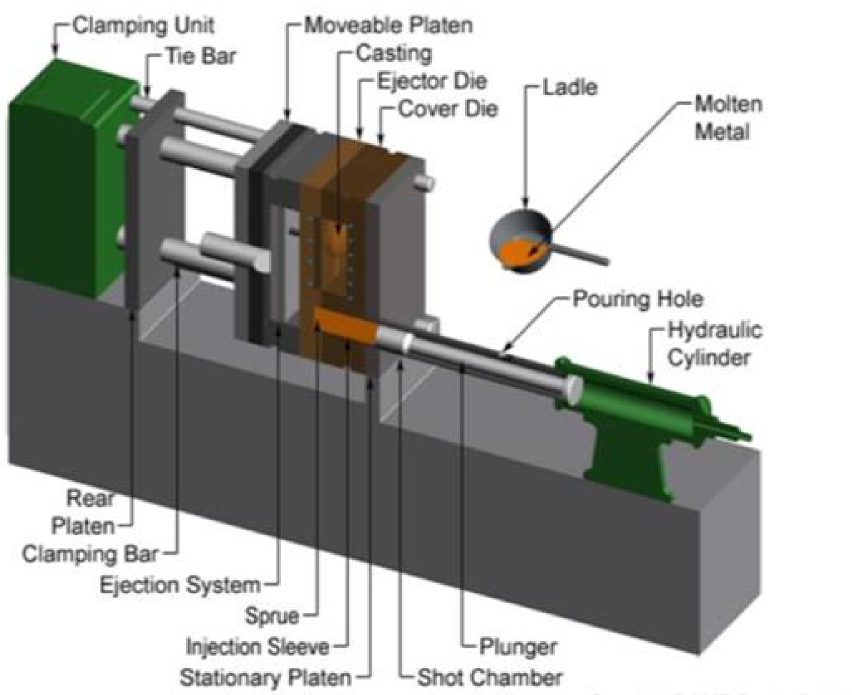
### 4.3.2 Business and problem understanding

The internal foundry where a case study was made belongs to number one supplier of aluminum components in Slovakia from point view e. g. size, amount of personal and revenue, investment, and level of technology in use.

The chosen product belongs to battery housing family (Figure 13). Due to market transformation in last 5 years mostly due to reduction of emission is visible grow of the battery housing (BH) in last years. Despite of already several suppliers is delivering the BH to their customer still there are limited information’s and experiences with BH and no long-term best practices as f.e. in gravity casting cylinder head production. The battery housings products are used as a protection of batteries in the electrical vehicles or plug in hybrids. Batteries inside stores the energy which is convert into motion as is needed. Battery housing must comply to highest quality standards where stiffness, tightness a dimensional stability belongs to the main attributes which the final product must achieve. From foundry - metallurgical point of view the tightness requirements is define as the amount of leak media (for testing is used air or helium) per define unit of time (minute). Tightness very often reflects to the homogeneity of the microstructure of the casting.

The casting process is displayed on figure 12.

Figure 12: High pressure die casting process




Source: internal

If casting has many internal defects such a bubble, micro / macro – porosities or oxides the tightness of the product will be not achieved= part will be evaluated as defective part and scrapped. Tightness quality check is standardly done after machining process almost at the end of the process flow. For sure some of the superficial defect can be observed direct after casting but mostly are visible after CNC machining process. Scrapped parts or internal quality level are one of main indicator in the foundry and belongs to the main influencers on economic health of the company. Internal quality level is crucial indicator and shows how robust and effectively are set up the processes. If the process has high variation and technical parameters aren't set up correctly there is almost impossible in repeatable way produced sound castings. A good process understanding and understanding of inter -relationship among process variables is a must for process and product specialist. To precisely analyses the root cause of the defect and define robust action is a key aspect of the success. The case study foundry has high - class monitoring and evidence of technological parameters. Almost every process parameter (example in table 1) is assigned to the special data matrix code (DMC, Figure 13) which is printed direct after casting on the product surface. Rest parameters are also available but must assign manually to the DMC because location of these date ae in Excel or direct in the machine.

Figure 13: Battery housing product / Data matrix code

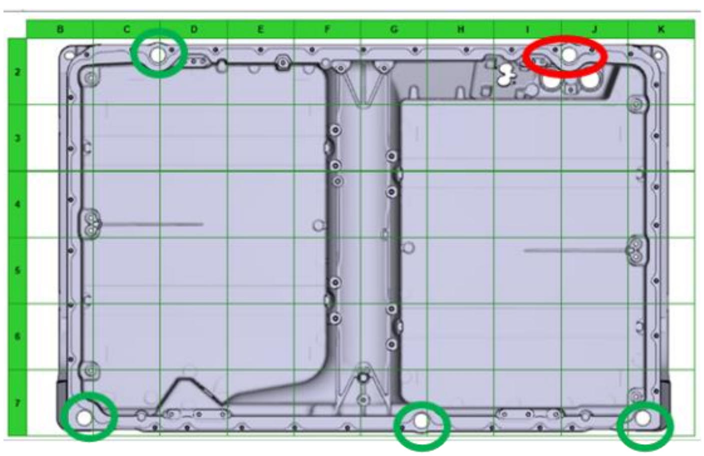

Source: internal

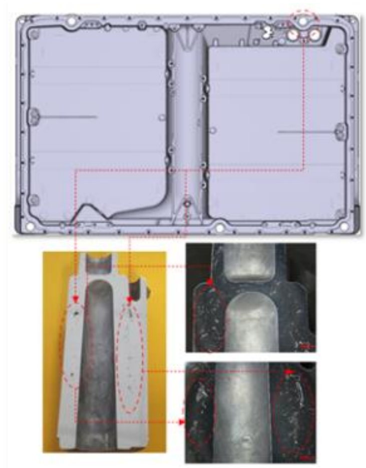
During the whole process flow through all process and quality operations is this DMC used as life book of the product. At every quality control station, the DMC is red and quality result is saved and stored in the system again.  Every DMC has information in the internal database in which parameters was produced and what was the quality results: OK or NOK part (rework part as well if is permit). The database consists also other sub-information's regarding values of specific characteristics if deeper insight or analysis is needed (time of machining, dimensional status, value of leakage etc).

As was mentioned before shrinkage defects formation in the casting belongs to the main "enemies" of process engineers and development centers in aluminum foundry business. Even of dozens of simulations are made in pre -serial phase with high prediction - only real-life casting in real life process with all process imperfections will confirm the simulation results and shows the reality quality of the casting. The thesis is not focused on elimination of the shrinkage as a priority number one but mainly on looking for testing model efficiency and his predictability of which process variables has a major impact on the shrinkage formation.

The defect formation is in the area Figure 14, for following text for defect location only zone TIJ 2 will be used in next writing.

Figure 14: Battery housing product with critical area TJI2

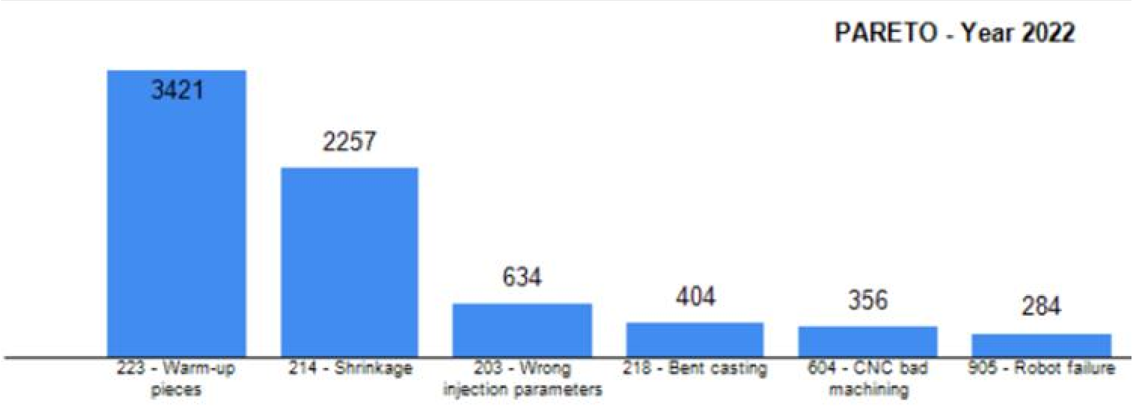




Source: internal

Figure 15 shows pareto defects distribution, shrinkage defect belongs to TOP 2 defect with major scrap reason in year 2022 what represent 3% internal scrap. Have to be mentioned that not all parts were scrap on area TIJ2 but this area is a majority from the total (50%). TOP1 from the pareto represents warm up pieces = parts scrapped directly in the casting machine after downtime longer than define time interval. This kind of defect is depended on the process continuity disruption related to the machine or tooling downtimes.

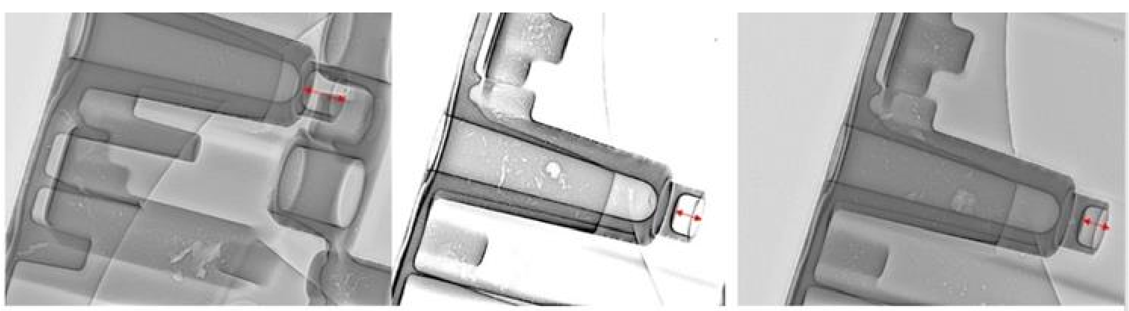
Figure 15: Battery housing – defect PARETO




Source: internal

In the critical area TIJ2 a squeeze pin is located. Figure 16 represents example of different squeeze set ups and defect formation.

Figure 16: X-ray defect formation in the critical area



Source: internal

The defects formation – size and localization are random in TJI2. X-raying of casting as a qualitative control method for casting quality determination is used randomly in range 5% castings from the casting shift batch. Only in special cases 100% X-ray operation is used (Special trials, customer request etc). X-ray set up is according to customer qualitative standard – casting with no defect is evaluated as OK, defect within specification is evaluated as letter P and HP, defect above specification as NOK. Example of Xray results for the TIJ2 area is on Table 4.

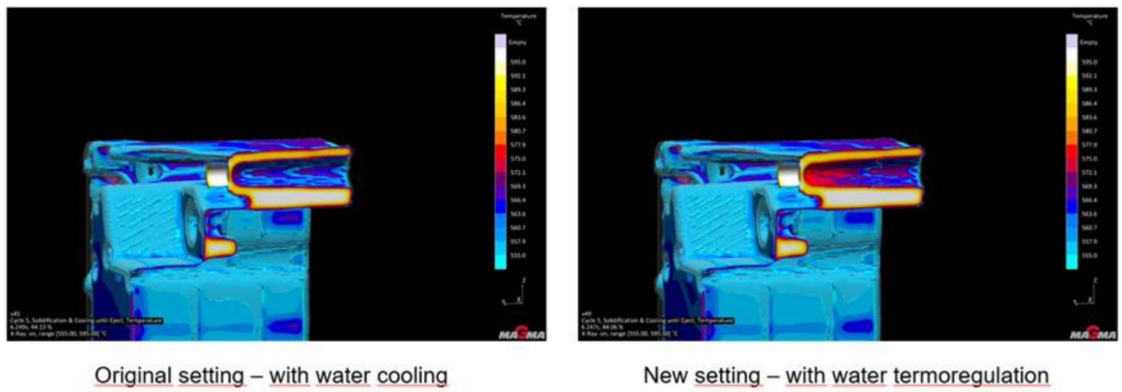
Table 4: X-ray defect results in the critical area

| Poznámka | RTG | Dátum RTG | Zmena RTG | Dátum odl. | Zmena | Stroj | Forma | Porad. č. | OK/NOK /H/P | 10 | 11 | 11 | 11,12 | 13,14 | 13,14 | 13,14 | 13,14 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | 15 | 16 | 16 | 16,17 | 18,19 | 18,19 | 18,19 | 18,19 | 20 | 21 |
| | | | | | | | | | | TK5-10 | TK23-30 | TK23-3 | TK23-10 | TIJ2-2 | TIJ2-10 | TIJ2n-10 | TJ2-18 | THI2-10 | TGH2- |
| | | | | | | | | | | 10 | 30 | 3 | 10 | 2 | 10 | 10 | 18 | 10 | 3 |
| SKRATENY | 4 | 6.12.2021 | N | 25.11.2021 | A | | D04 | 50 | H-P | | | | | P | | H-P | H-P | | |
| SKRATENY | 4 | 6.12.2021 | N | 25.11.2021 | A | | D04 | 49 | H-P | | | | | P | | P | H-P | | |
| | 3 | 7.12.2021 | R | 5.12.2021 | C | OB5 | D02 | 13 | H-P | | | | | P | | | | | |
| | 3 | 7.12.2021 | R | 4.12.2021 | C | OB5 | D02 | 49 | NOK | | | | | | | N | N | | |
| Sk 75-2021 | 3 | 7.12.2021 | R | 7.12.2021 | A | OB5 | D01 | 52 | NOK | | | | | | | | | | |
| Sk 75-2021 | 3 | 7.12.2021 | R | 7.12.2021 | B | OB5 | D01 | 38 | NOK | | | | | | | | N | | |
| Sk 75-2021 | 3 | 7.12.2021 | R | 7.12.2021 | B | OB5 | D01 | 37 | P | | | | | | P | | P | | |
| Sk 75-2021 | 3 | 7.12.2021 | R | 7.12.2021 | B | OB5 | D01 | 36 | H-P | | | | | | | | H-P | | |
| Sk 75-2021 | 3 | 7.12.2021 | N | 7.12.2021 | B | OB5 | D01 | 35 | P | | | | | | | P | P | | |
| Sk 75-2021 | 3 | 7.12.2021 | N | 7.12.2021 | B | OB5 | D01 | 34 | H-P | | | | | | | P | H-P | | |
| Sk 75-2021 | 3 | 7.12.2021 | N | 7.12.2021 | B | OB5 | D01 | 33 | P | | | | | | | P | P | | |
| | 3 | 8.12.2021 | R | 8.12.2021 | A | OB5 | D04 | 83 | NOK | | | | | | | | H-P | | |
| | 3 | 8.12.2021 | R | 8.12.2021 | A | OB5 | D04 | 82 | H-P | | | | | | | | H-P | H-P | |
| | 3 | 8.12.2021 | R | 8.12.2021 | A | OB5 | D04 | 81 | H-P | | | | | | | H-P | H-P | | |
| | 3 | 8.12.2021 | R | 8.12.2021 | A | OB5 | D04 | 80 | H-P | | | | | | | | H-P | | |
| | 3 | 8.12.2021 | R | 8.12.2021 | A | OB5 | D04 | 79 | H-P | | | | | | | H-P | H-P | H-P | |
| | 3 | 8.12.2021 | R | 8.12.2021 | A | OB5 | D04 | 78 | H-P | | | | | | | H-P | H-P | H-P | |
| | 3 | 8.12.2021 | R | 8.12.2021 | A | OB5 | D04 | 77 | NOK | | | | | | | | | N | |
| | 3 | 8.12.2021 | R | 25.11.2021 | A | OB5 | D04 | 71 | NOK | | | | | | | | | | |
| | 3 | 8.12.2021 | R | 25.11.2021 | A | OB5 | D04 | 70 | NOK | | | | | | | N | | | |
| | 3 | 8.12.2021 | R | 25.11.2021 | A | OB5 | D04 | 63 | NOK | | | | | | | N | | | |
| | 4 | 15.12.2021 | N | 10.12.2021 | C | OB5 | D04 | 31 | H-P | | | | | | | P | H-P | | |
| | 4 | 15.12.2021 | N | 13.12.2021 | A | OB5 | D04 | 8 | P | | | | | | | | P | | |
| | 4 | 15.12.2021 | N | 10.12.2021 | A | OB5 | D04 | 18 | P | | | | | | | P | P | | |
| | 4 | 15.12.2021 | N | 9.12.2021 | B | OB5 | D04 | 54 | H-P | | | | | | | H-P | P | | |
| | 4 | 15.12.2021 | N | 9.12.2021 | B | OB5 | D04 | 30 | NOK | | | | | | N | | | | |
| | 4 | 15.12.2021 | N | 9.12.2021 | B | OB5 | D04 | 37 | H-P | | | | | | | H-P | P | | |
| | 4 | 15.12.2021 | N | 9.12.2021 | B | OB5 | D04 | 32 | P | | | | | | | P | P | | |
| | 4 | 15.12.2021 | N | 9.12.2021 | B | OB5 | D04 | 31 | P | | | | | | | P | P | | |
| | 4 | 15.12.2021 | N | 9.12.2021 | B | OB5 | D04 | 26 | P | | | | | | | P | P | | |
| | 4 | 15.12.2021 | N | 9.12.2021 | B | OB5 | D04 | 35 | P | | | | | | | P | P | | |
| | 4 | 15.12.2021 | N | 9.12.2021 | B | OB5 | D04 | 27 | P | | | | | | | P | P | | |
| | 4 | 15.12.2021 | N | 9.12.2021 | B | OB5 | D04 | 36 | P | | | | | | | P | P | | |

Source: internal

During the development phase even in the serial phase there were made dozens of simulations for minimalization of defect occurrence – example is shown in figure 17. However even short-term results from the simulation trials were positive, during the standard casting process and multi variation influence of process and technological factors cause unstable quality results.

Figure 17: Simulation results from new cooling conditions in TJI2 zone




Source: internal

### 4.3.2 Data understanding

As was mentioned in chapter 4.3.1 on Figure 11, the casting machine is controlling 27 parameters + additional 10 parameters such as AED, lubricant amount, pressure of water of lubricant, water flows, cycle time, die temperature, metal injections conditions cooling flow rates, die temperature before and after spraying are giving the whole picture about process. Not all data are stored on-line or in one single source. Some of data are stored in Excel or in other SQL database and for analysis purposes must incorporate in one big dataset. Data are stored on the internal server, where the downloading frequency is for all 37 parameters saved for every casting production (depend on the product cycle time). Data inputs are coming from various sources as Excel, internal SQL databases, direct information from casting machine – thermographs before and after spraying operation. Based on the mention above all data set had to be incorporated in one sheet and paired to the DMC casting code. For model analysis the data set was used from beginning of the production = from 2022 in Stage 1, in Stage 2 were also added data from year 2021 to increase dataset (stage 2a) and by that see the effect on model precision if the amount of process variables will be focused directly only on that they have major impact on defect formation (Stage 2b). For continuing testing increasing of model precision only data from X-ray focused only on the critical area TIJ2 was used as an input for the model (Stage 3).

Table 5 shows mentioned overviews:

Table 5: Model Dataset overview

| Stage | | 1 | 2 | | 3 |
|---|---|---|---|---|---|
| | | | A | B | |
| Results | precision | DATASET 2022 + all process variables | DATASED second half year 2021+2022 + all process variables | DATASED second half year 2021+2022 + Chosen 10 process parameters | DATASET - only Xray results 04-07/2022 + all process parameters |
| | recall | | | | |

Input data's: shift, casting date, die number, DMC code, process parameters : T1 - time of $1^{st}$ phase, velocity of $1^{st}$ phase distance $1^{st}$ stroke, T2 - dosing time of $2^{st}$ phase, velocity of $2^{st}$ phase, distance $1^{st}$ stroke, maximum velocity of $2^{nd}$ phase (V2), distance 2nd stroke, compress stroke, multiplication stroke (CC) , multiplication time (T3), delay of multiplication time, maximum pressure, maximum velocity (VM), average velocity in gating (VA), filling pressure, specific pressure, metal temperature in die casting machine, metal temperature in holding furnace, dosing time from holding furnace, cycle time, biscuit thickness, vacuum for every channel A,B,C,D, spraying information about amount of flow of AED mixing device - water, ration – water/ lubricant, cooling flows and cooling circuits, time start of multiplication, specific pressure, dosing pressure Stotek, max. pressure Stotek.

Output data's: Quality results on focus defect – OK / NOK parts, X ray results for TIJ2 area– OK results – no defect, P – small defect in specification,

**Stage 1:** The data set 2022 means information about 58 847 produced castings in the 2022. The whole process variables (37 parameters) were used for finding correlation as input. As output was used the final parts quality of the part OK/NOK.

**Stage 2 A:** The data set 2022 means information about 58 847 produced castings in the 2022 + extra 25 000 casting. The whole process variables (37 parameters) were used for finding correlation as input. As output was used the final parts quality of the part OK/NOK.

**Stage 2 B:** The data set 2022 means information about 58 847 produced castings in the 2022 + extra information about 25 000 casting. The process variables were reduced only on the most critical ones based on best experience of the process engineers = 10 process parameters were used for finding correlation as input. As output was used the final parts quality of the part OK/NOK to optic of chosen critical parameters was reduced from 37 to 10 critical process parameters: dosing time of 2$^{nd}$ phase, velocity of 2$^{nd}$ phase, maximum velocity of 2$^{nd}$ phase, compress stroke, multiplication time, delay of multiplication time, average velocity in gating, maximum pressure, specific pressure, metal temperature in the holding furnace.

**Stage 3**: The data set 2022 means information about 3347 produced castings in the 2022 which went through Xray control and could be assigned process parameter to DMC code. The whole process variables (37 parameters) were used for finding correlation as input. As output was used quality information from the Xray where we directly distinguish the defect size formation as was already described. We can consider that this kind information is very precise linked to casting which went through Xray operation.

Table 6 shows the process data processing into overview.

Table 6: Main Process parameters data- example

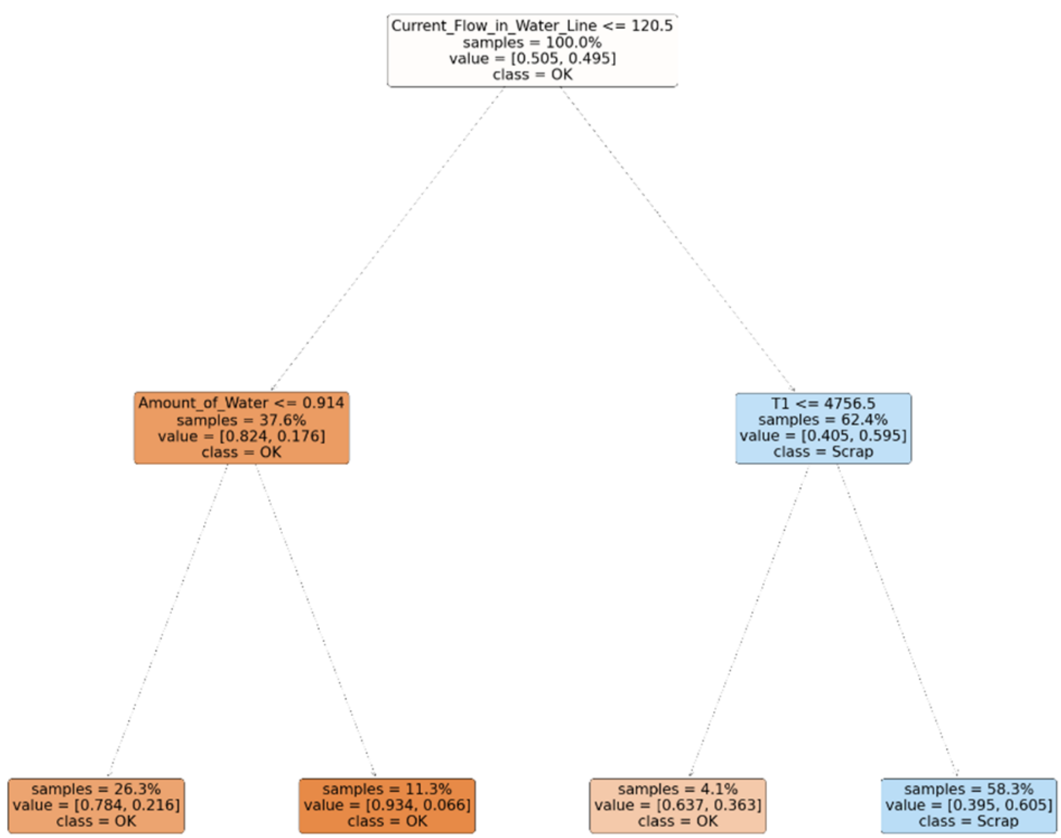
| celý názov | názov v dátach | počet vyplnených polí | priemer | štandardná odchýlka | min | 25% percentil | 50% percentil | 75% percentil | max |
|---|---|---|---|---|---|---|---|---|---|
| | **forma** | 58847 | 2,87786973 | 1,223872159 | 1 | 2 | 3 | 4 | 5 |
| | **Zmena** | 58847 | 2,080360936 | 0,80642676 | 1 | 1 | 2 | 3 | 3 |
| | **Injection_Number_of_Shift** | 58847 | 85,08695431 | 60,16670811 | 1 | 34 | 73 | 128 | 266 |
| **dráha 1. fáza** | **C1** | 58847 | 738,000034 | 3,481858539 | 576 | 737 | 738 | 738 | 1123 |
| **čas 1. fáza** | **T1** | 58847 | 4835,249477 | 107,4056611 | 990 | 4770 | 4825 | 4915 | 8700 |
| **rýchlosť 1. fáza** | **ID1st_Phase_Speed** | 58847 | 0,151597363 | 0,005962771 | 0,13 | 0,15 | 0,15 | 0,15 | 0,58 |
| **zadretie piestu** | **GP** | 58847 | 214,5696467 | 3678,13385 | 0 | 5 | 7 | 9 | 65535 |
| **dráha 2. fáza** | **C2** | 58847 | 374,0547691 | 6,640393958 | 0 | 370 | 373 | 378 | 539 |
| **čas naplnenia 2. fáza** | **T2** | 58847 | 98,39130287 | 20,69021284 | 0 | 91 | 94 | 101 | 1688 |
| **priemerná rýchlosť 2. fázy** | **ID2nd_Phase_Speed** | 58847 | 3,910309447 | 1,40348426 | 0,23 | 3,77 | 3,98 | 4,09 | 327,67 |
| **maximálna rýchlosť 2. fázy** | **VM** | 58847 | 5,314267507 | 0,604796269 | 1,64 | 5,36 | 5,43 | 5,52 | 6,21 |
| **dráha stlačenia** | **CC** | 58847 | 10,17103676 | 2,734261117 | 0 | 10 | 11 | 11 | 337 |
| **čas multiplikácie** | **T3** | 58847 | 54,31598892 | 94,14443466 | 0 | 51 | 54 | 57 | 4259 |
| **čas oneskorenia multiplikácie** | **TD** | 58847 | 261,0680918 | 945,4021555 | 0 | 50 | 56 | 59 | 8466 |
| **maximálny tlak** | **PM1** | 58847 | 210,7520689 | 14,41099327 | 136 | 214 | 214 | 214 | 487 |
| **rýchlosť kovu v záreze** | **VA** | 58847 | 44,72275358 | 5,108041209 | 2,62 | 43,21 | 45,57 | 46,82 | 327,67 |
| **tlak doplnenia** | **PR** | 58847 | 124,5197716 | 20,32195939 | 0 | 121 | 128 | 133 | 389 |
| **špecifický tlak** | **PS** | 58847 | 548,0368923 | 36,27557243 | 359 | 556 | 556 | 556 | 1128 |
| **teplota kovu DCM** | **Metal_Temperature** | 58847 | 534,2556035 | 237,692975 | 0 | 639 | 641 | 644 | 681 |
| **čas cyklu** | **Cycle_Time** | 58847 | 108,4381277 | 11,8294631 | 0 | 108,4 | 108,8 | 109,4 | 357,6 |
| **výška tablety** | **Bisquit_Thickness** | 58847 | 38,3355991 | 4,832463613 | 16 | 35 | 38 | 42 | 59 |
| **vákuum kanal A** | **Vacuum_Value_A** | 58847 | 27,35736741 | 9,997911758 | 0 | 23 | 26 | 30 | 235 |
| **vákuum kanal B** | **Vacuum_Value_B** | 58847 | 44,55889 | 13,94648571 | 0 | 40 | 44 | 48 | 284 |
| **vákuum kanal C** | **Vacuum_Value_C** | 58847 | 30,63957381 | 13,1138711 | 0 | 25 | 28 | 34 | 240 |
| **vákuum kanal D** | **Vacuum_Value_D** | 58847 | 28,83052662 | 13,43463157 | 0 | 24 | 27 | 31 | 279 |
| **teplota kovu Stotek** | **Metal_Temperature_STOTEK** | 58847 | 642,8637824 | 3,929246449 | 632 | 640 | 642 | 644 | 681 |
| **množstvo kovu v peci** | **Metal_Level_STOTEK** | 58847 | 73,48038133 | 11,47836065 | 24 | 66 | 74 | 82 | 107 |
| **plniaci tlak stotek** | **Min__Pressure_STOTEK** | 58847 | -89,83173314 | 7,926794582 | -129 | -95 | -89 | -84 | 0 |
| **dávkovací tlak stotek** | **Max__Pressure_STOTEK** | 58847 | 777,4731762 | 19,29013546 | 0 | 769 | 775 | 782 | 831 |
| **prietok mixéru realne AED** | **Current_Flow_in_Water_Line** | 58847 | 120,629174 | 0,483241115 | 119 | 120 | 121 | 121 | 122 |
| **tlak na mixér AED** | **Current_Pressure** | 58847 | 5,750184319 | 0,15124866 | 4,86169 | 5,656829 | 5,766782 | 5,847222 | 8,097222 |
| **celková dávka vody AED** | **Amount_of_Water** | 58847 | 0,916134628 | 0,202900653 | 0 | 0,80928695 | 0,9276982 | 1,036253 | 29,09662 |
| **tlak AED** | **SetPoint_for_the_Pump** | 58847 | 44,32145224 | 12,47984713 | 10,5 | 50 | 50 | 50 | 50 |
| **pomer mazadlo/voda AED** | **Ratio** | 58847 | 122,4253063 | 0,498501201 | 117 | 122 | 122 | 123 | 123 |
| **meraná dávka AED** | **Amount_of_Lubricant** | 58847 | 213,6783343 | 23,96942771 | 0 | 202,1 | 212 | 224,3 | 655,3 |
| **dávka na pohyblivú časť AED** | **Amounth_of_water___moving_side** | 58847 | 0,772013031 | 0,352956213 | 0 | 0,67927165 | 0,8925659 | 1,028296 | 2,19049 |
| **bod otvorenia DCM** | **Value_of_mold_opening** | 58847 | 2231,14937 | 3,4106873 | 2200 | 2230 | 2232 | 2233 | 2238 |
| **celkový čas zalisovania** | **T3S___Time_start_multiplication** | 58847 | 5376,463116 | 148,5729916 | 1500 | 5280 | 5382 | 5467 | 9999 |
| | **Metal_dosing_time_from_Stotek** | 58847 | 8,276622598 | 0,151863614 | 7,95 | 8,17 | 8,26 | 8,35 | 9,56 |
| | **Shot_Sleeve** | 58847 | 738,0363995 | 303,6144955 | 12 | 865 | 865 | 865 | 865 |
| | **vsetky_vady** | 58847 | 0,116029704 | 0,320263259 | 0 | 0 | 0 | 0 | 1 |
| | **vada_214** | 58847 | 0,03097864 | 0,173261287 | 0 | 0 | 0 | 0 | 1 |

### 4.3.3 Data preparation

Once the dataset is retrieved and organized, data pre-processing techniques are used not only for the purposes learning algorithm purposes but also for data cleaning and clustering. Due to using different data inputs – Excel, internal SQL databases, data must be to match to DMC code. Once are the data gathered and understood, next step is the phase of data preparation. It includes checking of missing values, transforming the format of specific columns if needed, filtering unwanted rows from dataset and choosing which parameters to use. In our case we deleted the data without results from Xray and omitted columns such as date and time, DMC code and information about location of scrap. Then the dataset is split to training and testing data - normally in the ratio 80% training, 20% testing. Training data is then used for training of the model (model is learning on them). After the model is trained, model is evaluated on testing data (20%) to find out the performance on data, which are new to the model and have not been seen during the training.

#### 4.3.4 Modelling

Figure 18: Random Forest model




Source: Cognexa, internal

Figure 18 shows example of random forest (RF) algorithm which was used. First row refers to question that divide training data with respect to the answer (is current_flow_in_water_line less or equal to 120.5? yes - go left to another node, no - go right to another node).

Second row express percentage of all the training dataset that goes to that node.

Value refers to the ratio of data according to the real results (for example in the first node 50.5% of the dataset was OK and 49.5% are scrap). Class represents final class that would be predicted according to class (=real result) of majority of the training data in the node.

RF belongs to machine learning algorithms used both for classification and regression. The aim of the model is to predict the value of target variable / variables by learning. As is visible on figure 18 there are root node, branches, and leaf nodes. Branches represent conditions on process parameters and paths between root and leaf nodes are chosen by if-then rules. Each node further conditions and leads to other nodes until it reaches a leaf node.

For the learning - training purposes as was mentioned above 80% of data was used and on 20% data was tested the model predictability. The total grow till 1000 trees were used during the modelling in this case study. There are two main indicators as precision and recall, where the target is to reach the highest number as possible. Precision is defined as percentual ration of relevant analysis results to all = that means how many % NOK pieces from the model prediction will be really NOK parts (how is model correct). On other hand the recall is defined as ratio of relevant results of the analysis to all relevant occurrence in the sample size = how many % from NOK pieces model really marked as NOK pieces.

### 4.3.5 Evaluation of Random Forest Model

In accordance with the definitions provided in the preceding chapter, the three stages of this study were tested using three distinct datasets to determine the effect of dataset size on the chosen model's predictability in relation to the entire process parameters or the chosen parameters as the primary process variables, based on the experience of process engineers (Stage 2b). The final target is to reach as highest possible number for prediction and recall.

From the first glance of simple correlation used on Figure 19 is visible some correlations of process parameters.

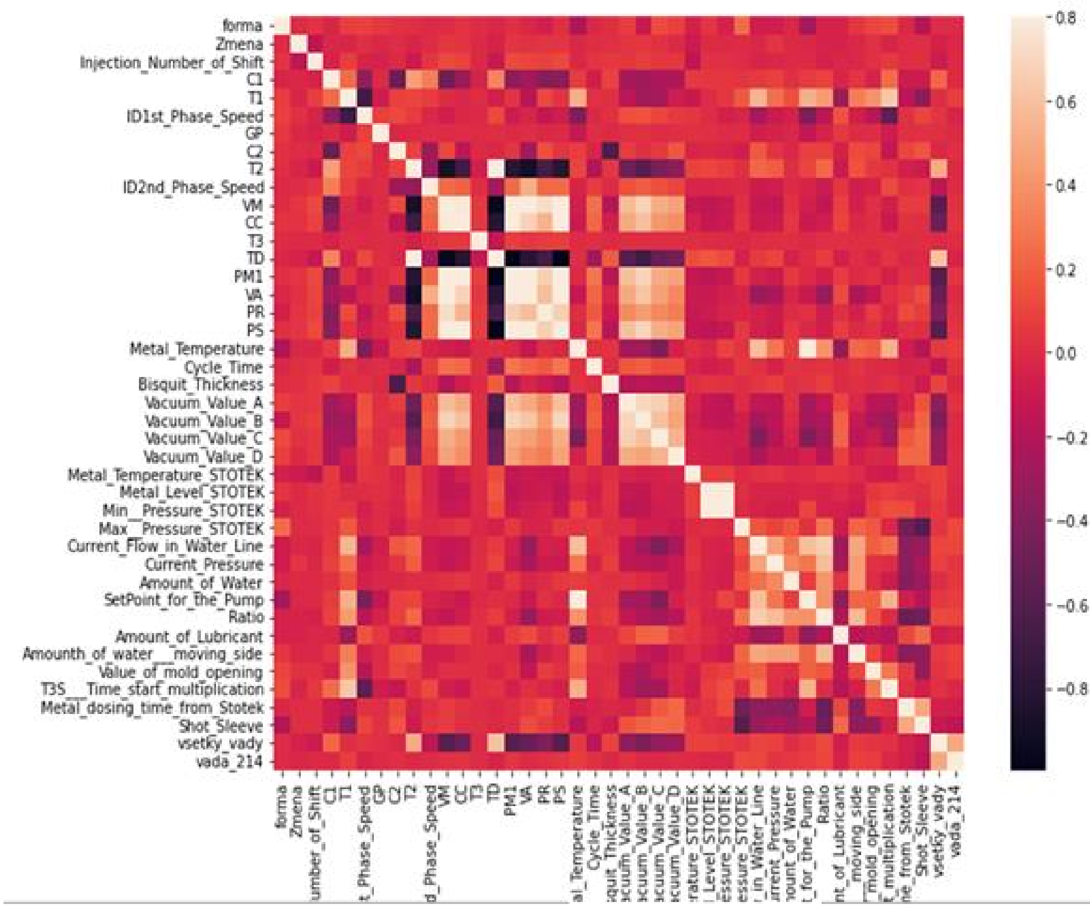
The distribution of colors means:

By brighter color shades = positive correlation = with one parameter increasing second has tendency to increase as well

Red – correlation close to Zero – two paramters doeant show any correlation

By darker color shades = higher negative correlation - with one parameters is increasing the second has tendency to decrease

Figure 19: Correlation Matrix of each process parameters against target defect




Source: internal (Cognexa)

Table 7: Correlation of main process parameters

| Parameter | Correlation with defect | Parameter | Correlation with defect |
|---|---|---|---|
| forma | -0,001951914 | Vacuum_Value_A | 0,028130571 |
| Zmena | -0,023533853 | Vacuum_Value_B | -0,008754594 |
| Injection_Number_of_Shift | -0,008971512 | Vacuum_Value_C | -0,007875089 |
| C1 | -0,012705934 | Vacuum_Value_D | 0,00733669 |
| T1 | 0,072369359 | Metal_Temperature_STOTEK | 0,000906724 |
| ID1st_Phase_Speed | -0,040990204 | Metal_Level_STOTEK | -0,02090685 |
| GP | -0,010027965 | Min__Pressure_STOTEK | -0,017826771 |
| C2 | 0,01923316 | **Max__Pressure_STOTEK** | **0,11471799** |
| T2 | -0,008880437 | **Current_Flow_in_Water_Line** | **0,106964911** |
| ID2nd_Phase_Speed | -0,003853658 | Current_Pressure | 0,042637085 |
| VM | 0,028921578 | Amount_of_Water | 0,05125877 |
| CC | 0,030604999 | SetPoint_for_the_Pump | 0,073285954 |
| T3 | 0,001239699 | **Ratio** | **0,111098301** |
| TD | -0,039262832 | Amount_of_Lubricant | -0,046162661 |
| PM1 | 0,040284164 | Amounth_of_water___moving_side | 0,083142793 |
| VA | -0,010271648 | Value_of_mold_opening | 0,073407259 |
| PR | 0,021797855 | T3S___Time_start_multiplication | -0,015165104 |
| PS | 0,038936079 | Metal_dosing_time_from_Stotek | -0,06627847 |
| Metal_Temperature | 0,073415883 | **Shot_Sleeve** | **-0,164687582** |
| Cycle_Time | 0,023136489 | vsetky_vady | 0,493513161 |
| Bisquit_Thickness | -0,013289839 | vada_214 | 1 |

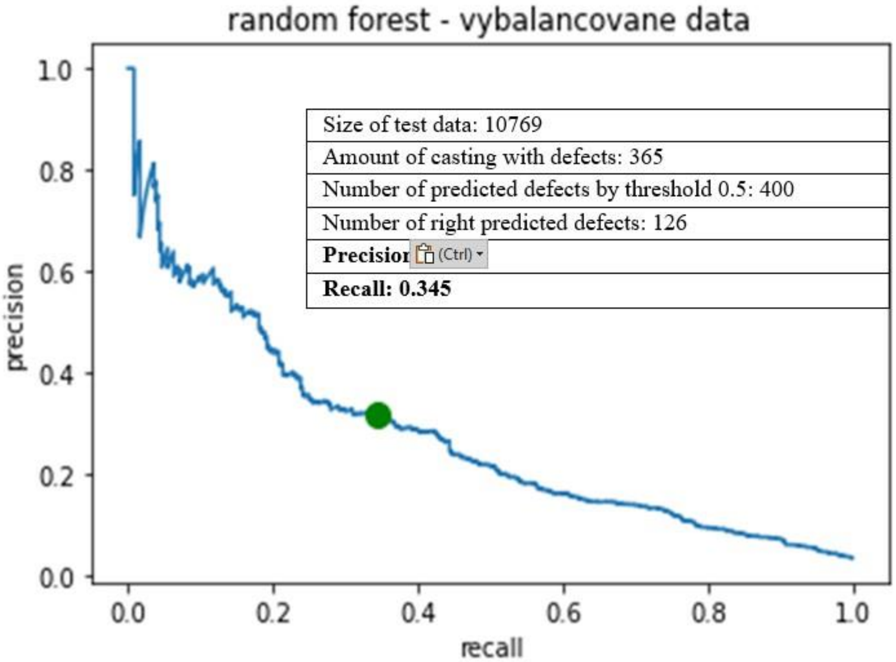
Source: internal

We can observe from first simple analysis in table 7 that time delay of pressure increase has highly negative correlation with maximum pressure (PM1) and gating velocity (VA), filling pressure (PR) and specific pressure (PS). The highest positive correlation we can observe by: dosing pressure Stotek (max. pressure Stotek) and flow of AED mixing device. The highest negative correlation shot sleeve parameter. The results from the model first test shows some higher correlations but based on the experience of the team - shot sleeve parameter doesn't have direct impact on the quality in are TIJ2. Opposite statement is for dosing pressure and AED mixing device parameters what must be investigated deeper.

On following figures 20 till figure 24 is shown results from the different model set up as was mentioned in the table 5. Information about data quantity is shown in the figures – size of test data. The final evaluation of reached results for precision and recall is presented in the table 8.

**Dataset 2022 – Model results STAGE 1:**

Figure 20: Random Forest model results – dataset year 2022

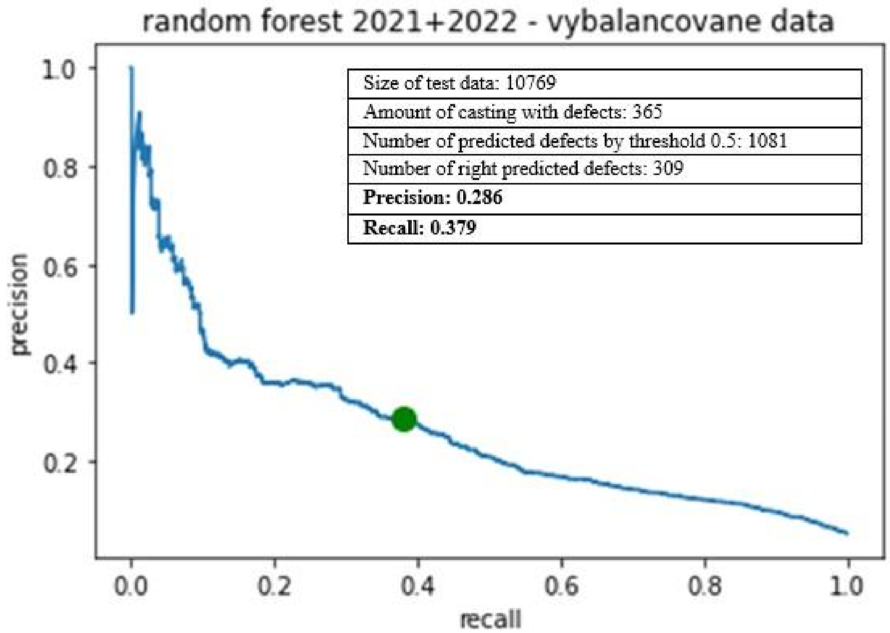



Source: internal (Cognexa)

The results from first stage shows precision on level 0,315 and recall on 0,345. These numbers means that the model accuracy is low and in a case of application the likelihood of determine defect occurrence in shopfloor will be low.

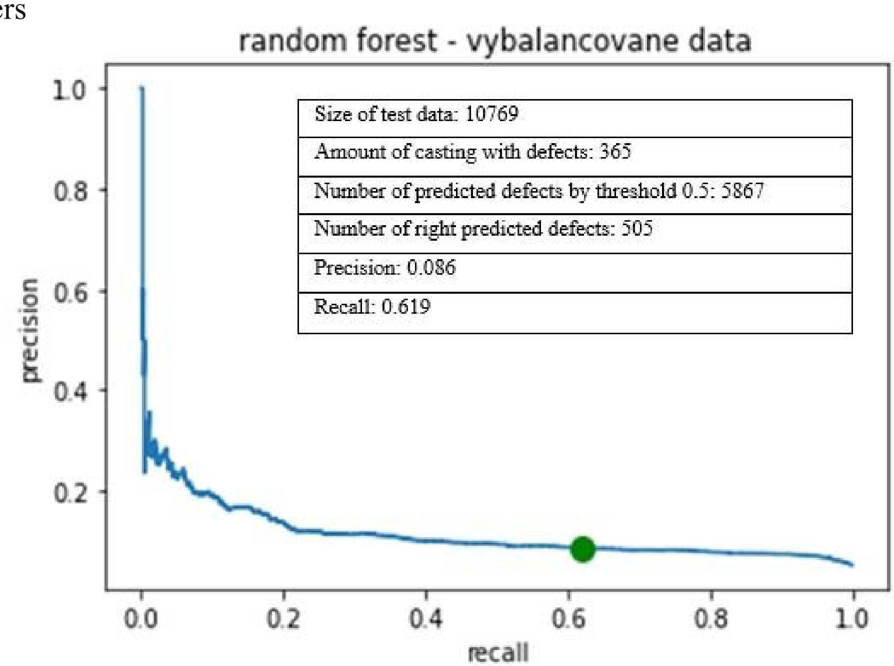
**Dataset 2022+2 HY 2021 – STAGE 2:**

Figure 21: Random Forest model results dataset year 2022+2021




Source: internal (Cognexa)

Figure 22: Random Forest model results dataset year 2022+2021 + only defined process parameters

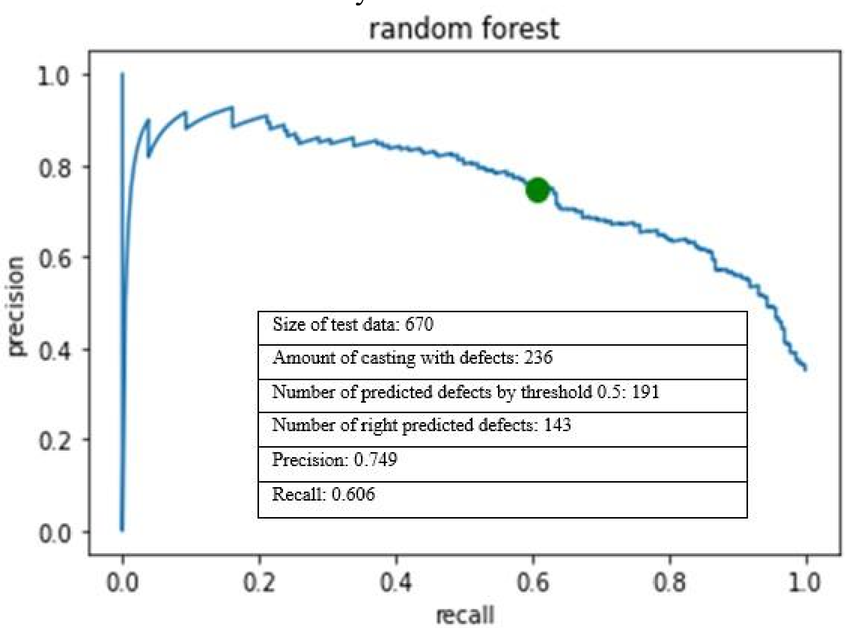



Source: internal (Cognexa)

For stage 2 A, B we increase amount of dataset up to 83 847 pcs. The stage 2A was used all 37 casting parameters as an input. The results show increasing dataset for the RF didn't bring increasing of precision and recall – the level is almost same. The 2B stage has used decreased amount of process parameters what in theory could help the model reduce variance. The results show worsening of precision almost to zero, what means that is incorrect even recall level on 0,619 is high = defect part will be marked as defective part. Have to noticed that for both stages 1 and 2 was model using information about NOK parts at the end of the process – final control, what is manual and visual control of the product operated by human. So human factor is in this case very high. This can be the reason why the results are very low.

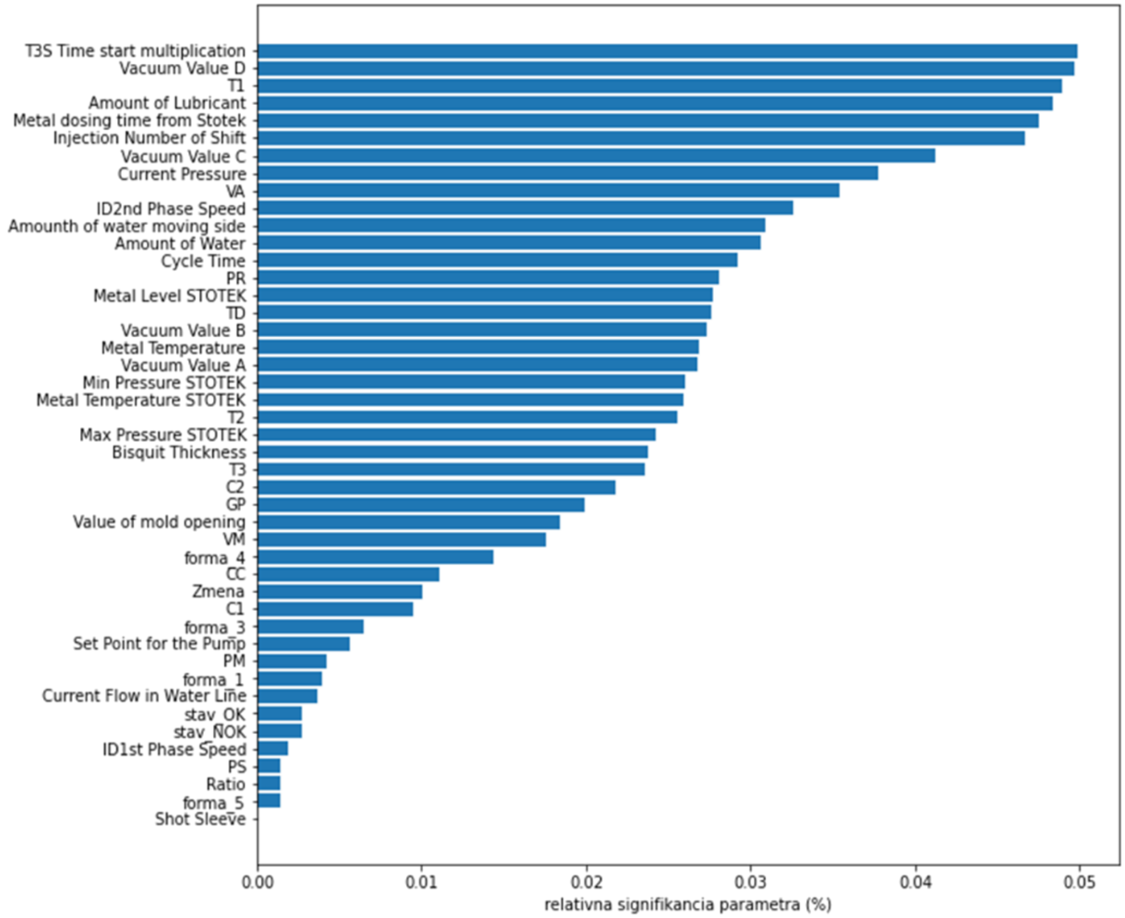
**Dataset 2022 Xray results STAGE 3:**

Figure 23: Random Forest model X ray results




Source: internal (Cognexa)

Figure 24: Random Forest model results – significance of process variables




Source: internal (Cognexa)

According to model set up information about 3347 produced castings in the 2022 were used for the stage 6. These parts went through Xray control and give deeper insight about level of the defect. All parts were aligned to the process parameters through DMC code. The reached precision – 0,749 and recall – 0,606 level is the highest reached number from all 3 stages what means that by adding more detailed information the model works better than expected.

In comparison with Wang, 2022 where RF model was used for forecasting ameloblastoma in medical industry and the results was reached 0,77 and 0,825. The author recognizes that the reached results of model precision for detection health issue is very good. The Al-Quraishi,2018 for breast cancer predictions, compare several data analytics algorithms and finally RF result accuracy was the highest 0,98 what compared to our case is almost ideal results.

Because of no similar case in foundry area where RF model was used there can be concluded that precision and recall number in stage 3 are high. Have to be mentioned that used Xray data set is much lower than in previous stages what can have impact on the results. For confirming the results must be increasing batch size which have to go through Xray operation.

Table 8 shows summary of all three stages.

Table 8: Summary of Random Forest Model Results – precision and recall

| **Stage** | | **1** | **2** | | **3** |
|---|---|---|---|---|---|
| | | | **A** | **B** | |
| **Results** | **precision** | 0,315 | 0,286 | 0,086 | 0,749 |
| | **recall** | 0,345 | 0,379 | 0,619 | 0,606 |

### 4.3.6 Evaluation of significance through ANOVA

Table 9: Statistical results – ANOVA

| **significance variable** | **process variable** | **ANOVA** | | | | |
|---|---|---|---|---|---|---|
| | | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| **1** | **Time T3 - start of multiplication (ms)** | 3 | 1478607 | 492869 | 18,82 | 0,000 |
| **2** | **Vacuum channel D (mbar)** | 3 | 39311 | 13103,8 | 50 | 0,000 |
| **3** | **Time T1 - start of filling (ms)** | 3 | 998160 | 332720 | 63,34 | 0,000 |
| **4** | **Amount of Lubricant (ml)** | 3 | 26392 | 8797,4 | 30,18 | 0,000 |
| **5** | **Dosing Time (s)** | 3 | 0,8563 | 0,28545 | 14,22 | 0,000 |
| **6** | **Injection number of casting** | 3 | 31741 | 10580 | 2,9 | 0,034 |
| **7** | **Vacuum channel C (mbar)** | 3 | 11998 | 3999,2 | 24,3 | 0,000 |
| **8** | **Current pressure (mixture device )** | 3 | 0,3759 | 0,12529 | 9,18 | 0,000 |
| **9** | **VA - velocity in gating (m/s)** | 3 | 467,2 | 155,726 | 23,48 | 0,000 |
| **10** | **2nd phase speed** | 3 | 3,524 | 1,1746 | 23,25 | 0,000 |

Source: internal

The top 10 significances from figure 24 marked by RF were put in one-way ANOVA where the focus was on the confirmation of hypothesis if the process has a significance influence on the Xray results. Based on the results in the table 9 all of listed parameter has a major influence due to p value is less than 0,05. Deeper investigation two or three-way parametric testing is recommended.

# 5. RESEARCH RESULTS AND ANALYSIS

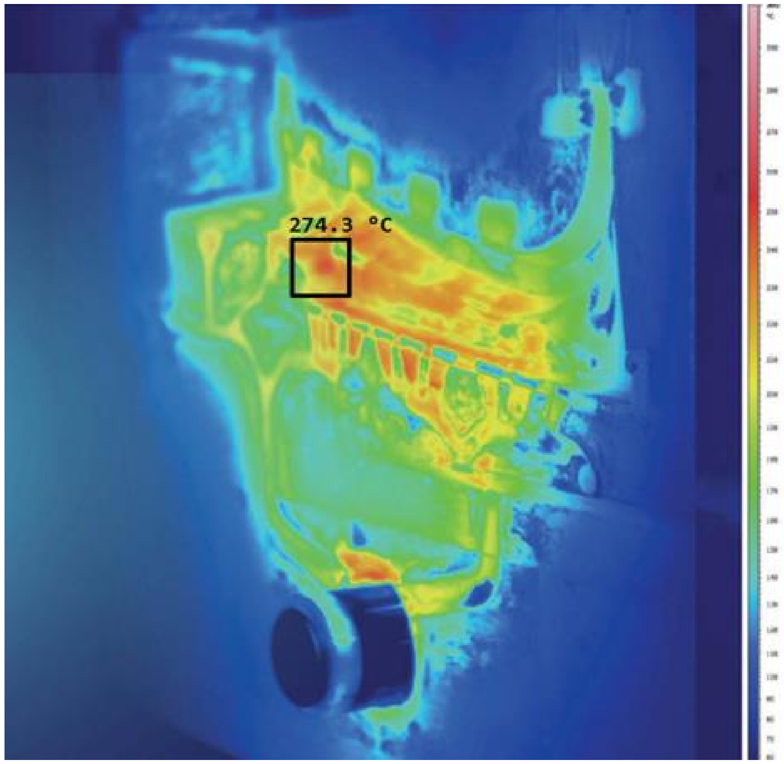
## 5.1. RQ 1: Will be the chosen machine learning model suitable for casting process as supportive tool for reaching OE?

The last results from the last random model were following precision 0,749 and recall 0,606. What means that the model has high predictability on 74,9% and the NOK parts will be marked correctly in 60,6% cases. Results were highly improved, but the question stands if there are potentials to improve it over 90%. For recommendation in the context of improving the precision and recall there are some options which are necessary to keep in mind and focus on them.

As it was shown on Figure 23 by implementation of more detailed dataset about casting quality in this case from X-ray the results were almost double improved. By that it can be assumed that if we increase quantity of castings which will go through Xray we can obtain more precise data for the RF model for learning and testing period.

Another possibility for increasing of model results is to add additional process parameter which wasn't so far implemented in the model. Figure 25 shows results of casting die thermogram which can be used as another input into the model. Each casting process is characterized by two thermograms before spraying (direct after part removing) and after spraying procedure. Currently the data are stored in the casting machine but due to complicated process wasn't so far implement in the RF model. The data must be transfer to server and align to each DMC code of the casting and as was discussed the inputs weren't implemented in this thesis.

Figure 25: Thermogram of the die after spray operation




Source: Internal

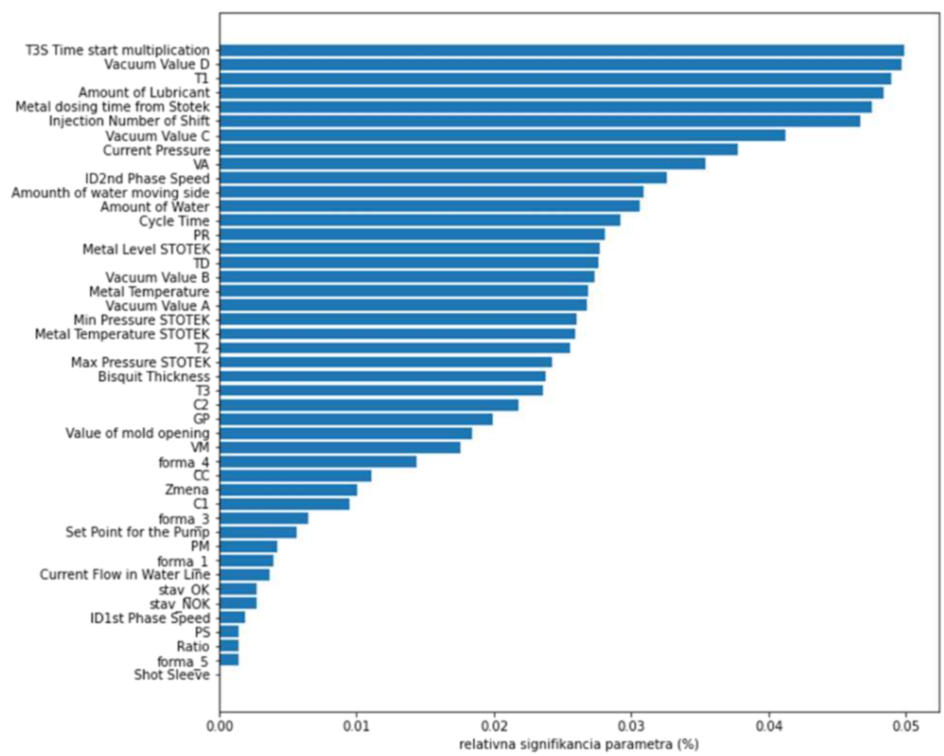
High model prediction can be used as perfect additional tool for reaching excellence in real time operations. They could be used several options how to predict the defect formation where the machine itself can be stopped if some defined critical parameters will exceed the tolerance limit. On another hand, the model output shows some significance process parameters, and this can be evaluated simultaneously during the production and process parameters can be adjusted in reasonable and technical possible way on -line. In other words, implementation of some kind machine learning algorithm into casting machine.

In following chapter 5.2 will be visible trend of internal scrap as an answer on the of research question if the model results will be supportive tool for reaching OE in the quality indicator. As is shown the quality trend was improved mainly due to information basis from the model and implemented actions.

## 5.2 (RQ 2) Could be results from learning model used as an input for improving quality results as supportive tool for reaching OE?

Based on the model result on the Figure 26 the significance of each process parameter was put in the chart. The highest values are around 5% of significance which seems that from the first glance is very low number. But after analysis of each process variable, it can be observed that filling times and vacuum process parameters and lubrication (amount of water) has major influence on the results. Also, can be used other data analysis systems for confirming the results from RF output or deeper investigation by interaction of two or three process variables on shrinkage formation. In many cases happed that model shows the process variation dependency f.e. 2$^{nd}$ casting velocity 50 milliseconds and model shows that by 55 milliseconds there is a significance dependency. Technically is not possible reach so small set up of the machine due the technical limitations of the machine response time.

Figure 26: Significance of process parameters




Source: Internal

Based on the results described above the process engineers defined the trial loops. The reached total quality results are shown on the figure 27. The results from the calendar week 32 after implementation of actions show the good performance in the critical of TIJ2 zone was improved. Still the 100% quality results are not fulfilled or defect completely solved but we can understand this results as very successful and be more under control from point of view variance. The total quality product results were improved below 5% after implementations of action in calendar week 32. The shrinkage development (TOP 2 defect) was improved in average below 1%.

Figure 27: Internal quality results TOTAL and shrinkage development

**Defect:** multi filter

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Produced (pcs) | 0 | 2 167 | 2 344 | 2 421 | 2 098 | 1 203 | 168 | 122 | 2 930 | 2 984 | 73 | 1 690 | 1 477 | 2 622 | 1 234 | 1 173 | 2 920 | 4 074 | 2 923 | 2 564 | 3 570 | 7 411 | 3 242 | 6 673 | 3 100 | 1 499 |
| Scrap (pcs) | 0 | 470 | 592 | 423 | 451 | 178 | 51 | 55 | 493 | 669 | 44 | 199 | 204 | 186 | 196 | 147 | 251 | 365 | 373 | 230 | 557 | 766 | 283 | 551 | 263 | 119 |
| % | N/A | 21,69 | 25,26 | 17,47 | 21,50 | 14,80 | 30,36 | 45,08 | 16,83 | 22,42 | 60,27 | 11,78 | 13,81 | 7,09 | 15,88 | 12,53 | 8,60 | 8,96 | 12,76 | 8,97 | 15,60 | 10,34 | 8,73 | 8,26 | 8,48 | 7,94 |

| | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 2022 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Produced (pcs) | 2 798 | 2 920 | 508 | 0 | 0 | 2 630 | 3 997 | 4 008 | 4 501 | 1 786 | 1 864 | 3 505 | 3 663 | 1 463 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 96 325 |
| Scrap (pcs) | 269 | 272 | 159 | 0 | 0 | 228 | 201 | 178 | 163 | 145 | 92 | 161 | 304 | 87 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 375 |
| % | 9,61 | 9,32 | 31,30 | N/A | N/A | 8,67 | 5,03 | 4,44 | 3,62 | 8,12 | 4,94 | 4,59 | 8,30 | 5,95 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 10,77 |


**Defect:** 214 - Shrinkage

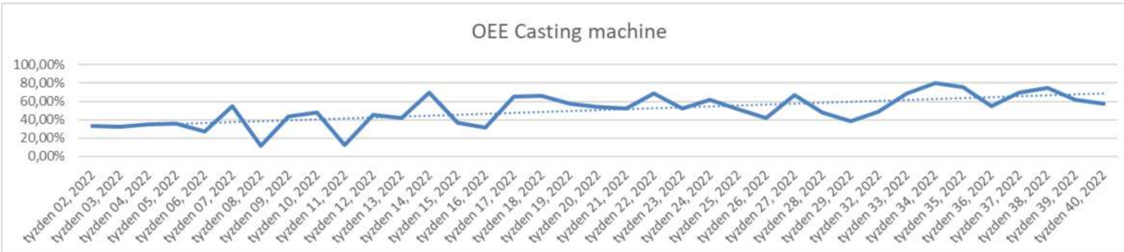
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Produced (pcs) | 0 | 2 167 | 2 344 | 2 421 | 2 098 | 1 203 | 168 | 122 | 2 930 | 2 984 | 73 | 1 690 | 1 477 | 2 622 | 1 234 | 1 173 | 2 920 | 4 074 | 2 923 | 2 564 | 3 570 | 7 411 | 3 242 | 6 673 | 3 100 | 1 499 |
| Scrap (pcs) | 0 | 186 | 337 | 195 | 171 | 86 | 12 | 2 | 171 | 285 | 2 | 71 | 12 | 31 | 11 | 13 | 51 | 149 | 165 | 49 | 44 | 98 | 28 | 21 | 27 | 18 |
| % | N/A | 8,58 | 14,38 | 8,05 | 8,15 | 7,15 | 7,14 | 1,64 | 5,84 | 9,55 | 2,74 | 4,20 | 0,81 | 1,18 | 0,89 | 1,11 | 1,75 | 3,66 | 5,64 | 1,91 | 1,23 | 1,32 | 0,86 | 0,31 | 0,87 | 1,20 |

| | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 2022 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Produced (pcs) | 2 798 | 2 920 | 508 | 0 | 0 | 2 630 | 3 997 | 4 008 | 4 501 | 1 786 | 1 864 | 3 505 | 3 663 | 1 462 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 96 324 |
| Scrap (pcs) | 82 | 14 | 10 | 0 | 0 | 29 | 16 | 36 | 33 | 1 | 0 | 8 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 470 |
| % | 2,93 | 0,48 | 1,97 | N/A | N/A | 1,10 | 0,40 | 0,90 | 0,73 | 0,06 | 0,00 | 0,23 | 0,16 | 0,00 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 2,56 |


Source: internal

Improvement of internal scrap has overall positive effect on the operational excellence metric OEE of the casting machine – see Figure 28.

Figure 28: OEE development of the die casting machine




Source: internal

For sure well-designed model will bring many other opportunities and as was shown above the quality as factor for reaching operational excellence was improved dramatically. There is clear based on the results above that big data analytics has a positive effect on overall foundry performance.

## 5.3 (RQ 3) What take ways could be transferred to other industry?

It is necessary to noticed that rightly adjusted model along with trying to get the highest possible numbers along with implementation, which is the number one priority in every pilot phase of a implementation of a data analysis. Next application relies on the reached number of precision and recall.

Data analysis offers many possibilities. First and foremost is the processing of the extracted data. Here, process engineers and other technical department heads get information as quickly as possible. The early knowledge of needed information and data is the main precondition of understanding the processes and the process variables, and that means the best possible measure should be done based on that.

Secondly, the right model that is connected online with the engine means minimalizing the possibilities of NOK as much as possible.

The function "Watch Dog" also sometimes called "Alert system" can be used in determining variables with more process parameters in the final quality indicator.

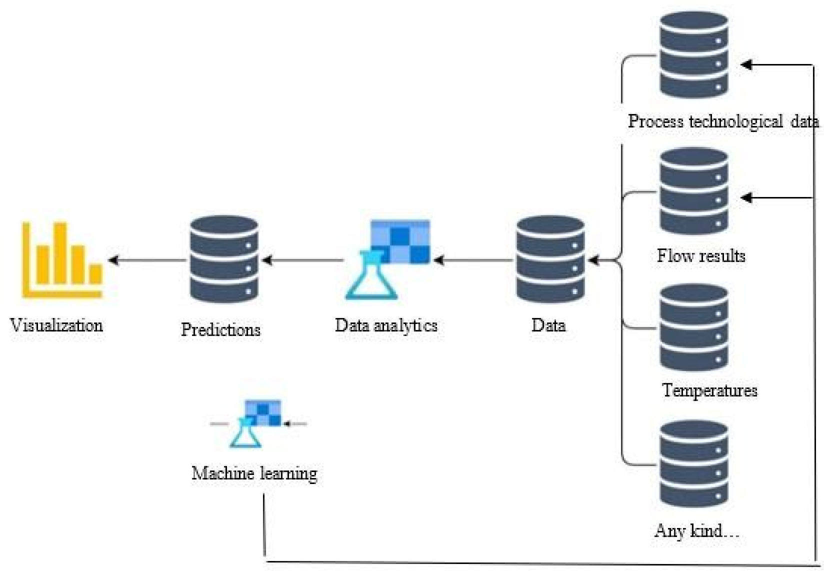
The third option is that the machine is adjusted automatically based on the parameters and according to the real state of process parameter like temperature, time of filling etc. That means, this process will automatically be adjusted without any further operator assistance from other persons and this process will lead to some kind autopilot. The further development of machine learning models in casting area has to be analyzed and develop for sure.

As it was mentioned in the beginning, the most important and the key part is the first phase. Finding of the right data analytic model will be done along with a lot of testing, creating new possible models (which means there is a special innovative technique related to it).

In the thesis used Random forest models can be one of the options which will show how to adjust the process and to be closer to operational excellence in the quality indicator.

Practically, the model showed on the Figure 29 below can be used for every industry. The main part of the model along with data about the process is done and evaluated through data model. The outputs of the data model will be then used as a prediction for further steps of innovation or as an output data for visualization or in a case and possible application of machine learning automated adjustment of process parameters according to output from data analytics (f.e. water flow regulation in critical area based.

Figure 29: Implementation model for predictions and visualization




Source: internal

The question which has been arisen is how the whole organization is prepared for such a change from organizational point of view. From practical point of view, it will be recommended that in a case adoption of BDA into company the new job position have to be created – data analyst. The position will take care of monitoring data, correlating them, and supporting process engineer and specialist towards faster data processing and faster data evaluating.

## 5.4 Deployment strategy

Based on the theoretical and practical research that has been done in this thesis, it is possible to upgrade the existent application the following way:

- We can use the application on different kinds of possible defect that can occur, because the mentioned defect is one of the most frequent seen defects in the industry. If we would like to investigate the cause of the defect, basically it is caused by a variety of reasons and also conditioned by a number of parameters.

- every application and also defect is unique which is important to keep in mind and act according to it during the deployment strategy.

If we would like to analyze the defects further, one possible way of analyzing them may contain these steps:

- First step is based on formulating and describing what we wish to examine and study further
- Secondly it is necessary to choose the right methods to implement and realize the intended investigation.
- Afterwards, the model and parameters should be defined.
- Next, the analytical model or for example previously mentioned Random Forest should be set up.
- After all the previous steps, it's possible to further examine using other data analytics models or algorithms

## 5.5 Executive summary

The key part of the thesis are the own findings, which are demonstrated in tables with results. One of the key findings is mentioned on figure 26, which contains information about the parameter's significance, and their calculated correlation with defect (order from highest to lowest significance). The further use of these values can be used to examine further and also to evaluate the risk of having similar defects in the future. If we talk about using defect models for further work, it is mainly dedicated to analytics. That means there can be done variety of steps which are focused on formulating and testing hypotheses. Along with that, new findings and knowledge is acquired through the process – f.i. in the developer activity and experience, along with new developer knowledge.

Next key result of the thesis is the ANOVA analysis based on the X-ray information, which was done to confirm or reject the results obtained from the Random Forest model. The ANOVA regression analysis was then compared to the model results. After evaluating all the available data, we came to an end that the p value is smaller than 0,05 – which means, the stated hypothesis is confirmed, and the final statement is that process factors have an impact on the final quality results. However, these findings are limited by the fact, that there was done only one regression analysis, but it is recommended to do at least one or three analyses of parameters. Also, as it was mentioned before, it is important to do a detailed analyze to define the characterizes needed for setting the trial plan, which enables to obtain, determine, and confirm the results.

Also, it is possible to adjust the model to increase the values of prediction and recall parameters to reach closer to the value of 1. This process also includes adjusting the Xray model and adding parameters related to the temperature before and after the spraying is done.

In my opinion, to comparing the results and findings from this thesis to the findings of other authors is not relevant. This is caused by the fact, that most of the mentioned authors have used ANN models in the perfect conditions – and that can't be relevant to compare with my findings that are based on real, non-ideal conditions. The final output and answer on the question if DA has connection for improving of kpi – in our case internal scrap – this answer is relevant also to other authors who confirms also positive impact of DA on the overall performance results.

# 6. DISCUSSION AND IMPLICATIONS

If we take a look at the future of data analytics, it is obvious that data analytics will make a huge impact on the whole populations, including individuals, groups, and companies. Data analytics has a huge potential in changing the business sector, including finance, manufacturing, and also healthcare. The future role and development of data analytics in my opinion is also related to the use and emergence of systems that can improve the balance between scaling ability and ability to analyze relevant data. Related to that, analytic techniques can abstract much of the effort that goes into choosing the best algorithms for data modelling and visualization. This allows data analysts to focus on other, maybe higher questions. This also plays a role in the way people query from the data, how to merge and group them, along with many other requests related to this.

And last, but not least important point in my opinion is the increased speed, scope, and reliability of analytical techniques, which is important in improving connectivity all over the world. Also, the offline world cannot be forgotten to mention, because it is possible to assume that collecting and processing data in the offline world will be done very similarly in my opinion compared to the on-line world as it's known now.

Also, it is possible to assume that in the very near future there will be a connection between the machines and real time results, which will be evaluated by the machine and based on that, the machine will make appropriate measures. This will mean a huge change, where the human factor won't be needed as much as today, and everything will be automatized.

The huge potential of data analytics is also related to the changes that must be done by the companies in order to maintain its place in the market and also to keep up with the trends and work as much effectively and precisely as possible.

In the company that was analyzed in the thesis, it is possible to assume that a variety of changes is needed in the near future. One of the main changes related to DA implementation is impact on the organizational structure of the company - this should be subject for an analysis. Related to that, one of the most likely scenarios is that there will open a work position dedicated for the role of data analytic. His responsibilities will contain creating and watching trends, along with providing relevant inputs for process engineers and technologists. Parallel to that it is important to suggest a training process of process engineers, specialist towards six sigma – green or yellow belt level to better understand statistical results and outcomes from the data analytics.

# 7. CONCLUSION

The primary objective of this thesis was understanding the role of data analytic in an application on specific case study – battery housing production. Second important area of investigation was what effect of data analytics will bring on the operational excellence – chosen kpi internal quality.

The casting process belongs to process where a lot of process parameters varies among each other and to obtain stable quality results long development time is needed. This time can be reduced by effective implementation of data analytics model. In our case the random forest model was used. Experience with the model integration shows model limitations and possibilities how the model accuracy can be improved. The quality of datasets shows significance differences in the model outputs. The highest and best results from model test so far were reached in the stage 3. The precision parameter was on level of 74,5% and recall 60,6%. Compared to experience with random forest model in the other industries especially in medical sphere, the reached parameters aren't so high. The next model test is recommended to run with integration of more casting data from the Xray or adding information's about die temperature. The model output – significances of the parameters was used for the process improvement. In this specific case the significant positive results were reached in the internal quality which is directly visible in the performance and brought overall improvement on OEE as well. The answer on the question – if the data analytics will have a positive impact towards reaching operational excellence is clear - in shown case of battery housing production reached results belong to the internal benchmark results.

# BIBLIOGRAPHY

A.Krimpenis, P.G. Benardos, G.C.Vosniakos, A.Koukouvitaki : Simluation based selection of optimum pressure die-casting process parameters using neural nets and genetic algorithms 2004

Al-Quraishi, Tahsien, et al. "Breast Cancer Recurrence Prediction Using Random Forest Model." Advances in Intelligent Systems and Computing, vol. 700, Springer International Publishing, 2018, pp. 318–29, https://doi.org/10.1007/978-3-319-72550-5_31.

Amit Sata, B.Ravi, 2019 : Foundry Data Analytics to identify critical parameters affecting quality of investment casting,Mechanical engineering Department, India, DOI: 10.1115/1.4041296

ANTONY, Jiju, Olivia MCDERMOTT, Daryl POWELL a Michael SONY, 2022. The evolution and future of lean Six Sigma 4.0. The TQM Journal [online]. [cit. 2022-09-26]. ISSN 1754-2731. Retrieved from: doi:10.1108/TQM-04-2022-0135

Azevedo A. / Santos M. (2008): KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European conference data mining, pp. 182–185. [Online] Available at http://recipp.ipp.pt/handle/10400.22/136. Viewed 2021-09-01.

Bach Nancy (2021): Choosing Operational Excellence Methodology, a Required First Step. [Online] Available at https://www.eonsolutions.io/blog/choosing-opex-methodology-first-step. Viewed 2021-08-27.

Bag, S. et al. (2020): Big data analytics as an operational excellence approach to enhance sustainable supply chain performance. Resources Conservation and Recycling, Vol. 153.

Bibby L. / Dehe B. (2018): Defining and Assessing Industry 4.0 Maturity Levels-Case of Defence Sector. Production Planning & Control 29 (12): 1030–1043. DOI: 10.1080/09537287.2018.1503355.

Biolos Jim (2002): Six Sigma Meets the Service Economy, Harvard Management Update (November 2002).

Breiman, Leo. "Random Forests." Machine Learning, vol. 45, no. 1, 2001, pp. 5–32, https://doi.org/10.1023/A:1010933404324.

Chang C. M. (2010): Operational Excellence - Lean Six Sigma, Web-Based Applications, and SOA. DOI:10.1002/9780470900208.ch13, [Online] Available at https://www.researchgate.net/publication/319532504_Operational_Excellence-Lean_Six_Sigma_Web-Based_Applications_and_SOA. Viewed 2021-08-26.

Chapman Pete, et al. (2000): CRISP-DM 1.0: Step-by-step data mining guide. Printed in the U.S.A © Copyright 2000 SPSS Inc. CRISPWP-0800.

Chen, Kuentai / Wang Chien-Chih / Kuo, Chi-Hung. (2021): Empirical Study of Foundry Efficiency Improvement Based on Data-Driven Techniques. Processes 2021, 9, 1083. [Online] Available at https://doi.org/10.3390/pr9071083 and https://www.mdpi.com/journal/processes. Viewed 2021-09-23.

Chiarini Andrea / Kumar Maneesh (2020): Lean Six Sigma and Industry 4.0 integration for Operational Excellence: evidence from Italian manufacturing companies. ISSN: 0953-7287 DOI: 10.1080/09537287.2020.1784485. [Online] Available at https://doi.org/10.1080/09537287.2020.1784485. Viewed 2021-08-26.

D.Benny Karunakar, G. L. Datta : Prevention of decets in casting using back propagation neural networks, DOI 10.1007/s00170-007-1289-0

Dal Bulent. (1999): Audit and review of manufacturing performance measures at Airbags International Limited. MSc dissertation, Manchester School of Management, UMIST, Manchester.

Davenport T. (2014): Big Data @ Work: Dispelling the Myths, Uncovering the Opportunities, Harvard Business School Press, Boston. ISBN:978-1-4221-6816-5. Figure 3 retrieved from http://www.creativante.com/new/index.php/2013-02-03-19-36-05/2014/210-big-data-at-work-dispelling-the-myths-and-uncovering-the-opportunities. Viewed 2021-08-10.

Deleryd, M. (1999), "A pragmatic view on process capability studies", International Journal of Production Economics, Vol. 58 No. 3, pp. 319-30.

Dubey, et al (2019): Big data and predictive analytics and manufacturing performance: integrating institutional theory, resource-based view and big data culture. British Journal of Management, 30 (2-3). pp. 341-361. ISSN 1045-3172, DOI: 10.1111/1467-8551.12355.

Dugan Kevin J. (2011): Design for Operational Excellence: A Breakthrough Strategy for Business Growth, McGrow-Hill Education, ISBN-13: 9780071768245.

E.Battaglia, F.Bonollo, G.Timelli, e.Fiorese and G.Kral : Correlation between process, microstructure and properties die casting alumunium – silicon alloys, Advances in Materials and Processing Technologies, DOI 10.1080/2374068X.2016.1250206

Edosio Uyoyo Zino (2014): Big Data Analytics and its Application in E-Comerce [Online] Available at https://www.researchgate.net/publication/264129339_Big_Data_Analytics_and_its_Application_in_E-Commerce. Viewed 2021-08-16.

Frank A. G. / Dalenogare L. S. / Ayala N. F. (2019): Industry 4.0 Technologies: Implementation Patterns in Manufacturing. International Journal of Production Economics 210: 15–26. DOI: 10.1016/j.ijpe.2019.01.004.

GANDOMI, Amir a Murtaza HAIDER, 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management [online]. 35(2), 137-144 [cit. 2022-09-28]. ISSN 02684012. doi:10.1016/j.ijinfomgt.2014.10.007

Garza-Reyes Jose Arturo et al. (2010): Overall equipment effectiveness (OEE) and process capability (PC) measures: A relationship analysis. International Journal of Quality & Reliability Management. Volume 27, No. 1/2010, pp 48-62. DOI: 10.1108/02656711011009308.

Ghobakhloo M. (2018): The Future of Manufacturing Industry: A Strategic Roadmap toward Industry 4.0. Journal of Manufacturing Technology Management 29 (6): 910–936. DOI: 10.1108/JMTM-02-2018-0057.

Giannetti Cinzia (2017): A framework for improving process robustness with quantification of uncertainties in Industry 4.0. Swansea University. DOI: 10.1109/INISTA.2017.8001155.

Giannetti Cinzia / Ransing Rajesh S. (2016): Risk based uncertainty quantification to improve robustness of manufacturing operations. Computers & Industrial Engineering, vol. 101, pp. 70 – 80.

Giannetti Cinzia et al. (2015): Organisational knowledge management for defect reduction and sustainable development in foundries. International Journal of Knowledge and Systems Science (IJKSS), vol. 6, no. 3, pp. 18–37.

Glodek M. et al. (2006): Process robustness a PQRI white paper. [Online] Available at https://www.researchgate.net/publication/238660855_Process_robustness-A_PQRI_white_paper pp. 1–32. Viewed 2021/08/29.

Gunasekaran A. et al. (2017): Big data and predictive analytics for supply chain and organizational performance. Journal of Business Research, 70, pp. 308-317.

Gunasekaran, Angappa. et al. (2017): Big data and predictive analytics for supply chain and organizational performance. Jornal of Business Research 70, pp. 308-317.

Gupta M. / George J. F. (2016): Toward the development of a big data analytics capability. Information & Management, 53(8), pp. 1049-1064.

Hatwell, Julian, et al. "CHIRPS: Explaining Random Forest Classification." The Artificial Intelligence Review, vol. 53, no. 8, 2020, pp. 5747–88, https://doi.org/10.1007/s10462-020-09833-6.

He Wu / Zha Shenghua / Li Ling (2013): Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33(1), pp. 464-72.

Hovhannes Sadoyan, A.Zakarian, P.Mohanty : Data mining algorithm for manufacturing process control, Int J Adv Manufacturing Technology 2006, DOI 10.1007/s00170-004-2367-1

Huang and Zhang 1993: Neural Networks in Manufacturing. A survey. In. IEEE/CH?T Int. Electronics Manufacturing Technology Symposium

International Standard Organization (ISO) (2014): Draft BS EN ISO 9001 Quality Management Systems - Requirements. International Standard Organization, Geneva, Switzerland, pp. 16.

J. Mitchell (2014): Operational Excellence, journey for creating sustainable value - book

Jayant K.Kittur, M.N.Choudhari, M.B.Parappagoudar : Modelling and multi-response optimization of high pressure die casting process using response surface methodology, Int J Adv Manufacturing Technology (2015) 77:211-22č, DOI 10.1007/s00170-014-6451-x

Jelonek Dorota (2017): Big Data Analytics in the Management of Business, Czestochova University of Technology, Poland. MATEC, Web of Conferences 125, DOI: 10.1051/matecconf/20171250.

Jian Zhen, Quodong Wang, Peng zhao, Congbo Wu: Opmitization of high pressure die casting process parameters using aritificial neural network, Int. J Adv Manufactoring Technology (2009), 44:6667-674, DOI: 10.1007/s00170-008-1886-6

JIANG, Jing, 2012. Information Extraction from Text. In: AGGARWAL, Charu C. a ChengXiang ZHAI, ed. Mining Text Data [online]. Boston, MA: Springer US, 2012-1-7, s. 11-41 [cit. 2022-09-28]. ISBN 978-1-4614-3222-7. doi:10.1007/978-1-4614-3223-4_2

Jitender K.Rai, Amir M. Lajimi, P.Xirouchakis : An intelligent system for predicting HPDC process variables in interactive environment, DOI 10.1016/j.jmatprotec.2007.10.011

Khan Abdul Rauf, et. al. (2017): Big Data analytics for Industrial Process Control. 22nd IEEE International Conference on Emerging Technologies and Factory Automation. [Online] Available at https://www.researchgate.net/publication/320211699_Big_Data_Analytics_for_Industrial_Process_Control/link/5abb5c3f45851563660afedf/download. Viewed 2021-09-27.

Kühnlein, Meike, et al. "Improving the Accuracy of Rainfall Rates from Optical Satellite Sensors with Machine Learning — A Random Forests-Based Approach Applied to MSG SEVIRI." Remote Sensing of Environment, vol. 141, 2014, pp. 129–43, https://doi.org/10.1016/j.rse.2013.10.026.

L.X.Kong, F.H.She, W.M.Gao, S. Nahavandi, P.D. Hodgson: Integrated optimization system for high pressure die casting processes, journal of materials processing technology 2008

Manyika J. (2012): Manufacturing the future: the next era of global growth and innovation. McKinsey Global Institute, [Online] Available at https://www.mckinsey.com/business-functions/operations/our-insights/the-future-of-manufacturing. Viewed 2021-08-28.

Marcin Perzyk, A.Kochanski, J.Kozlowski, A.Soroczynski, R.Biernacki, 2013 : Comparison of data mining tools for significance analysis of process parameters in application to process fault diagnosis, Institute of Manufacturing Technology, Faculty of Production engineering, Warsaw University

Mathivathanan, D. / Kannan, D. / Haq, A. N. (2018): Sustainable supply chain management practices in Indian automotive industry: A multi-stakeholder view. Resources Conservation and Recycling, Vol. 128. DOI: 10.1016/j.resconrec.2017.01.003.

MCCLEAN, Sally I., 2003. Data Mining and Knowledge Discovery. In: Encyclopedia of Physical Science and Technology [online]. Elsevier, 2003, s. 229-246 [cit. 2022-09-28]. ISBN 9780122274107. doi:10.1016/B0-12-227410-5/00845-0

Mikalef, P., Giannakos, M. N., Pappas, I. O., & Krogstie, J. (2018). The human side of big data: Understanding the skills of the data scientist in education and industry. In 2018 IEEE Global Engineering Education Conference (EDUCON) (pp. 503-512). IEEE. https://doi.org/10.1109/EDUCON.2018.8363273

Miller Andrew (2014): The Future of Operational Excellence for Manufacturing, [Online] Available at: https://www.industryweek.com/operations/article/21963001/the-future-of-operational-excellence-for-manufacturing. Viewed 2021-08-12.

Minevich Mark (2020): The automotive industry and the data driven approach. [Online] Available at https://www.forbes.com/sites/markminevich/2020/07/13/the-automotive-industry-and-the-data-driven-approach/?sh=5c464696f9a5. Viewed 2021-08-15.

MONTGOMERY, Douglas C., c2005. Introduction to statistical quality control. 5th ed. Hoboken: John Wiley. ISBN 0-471-66122-8.

Nash Mark / Poling Sheila (2008): Mapping the Total Value Stream: The Complete Guide to Production and Transactional Mapping. New York: Productivity Press. ISBN-10: 1563273594.

Niakšu Olegas (2015): CRISP Data Mining Methodology Extension for Medical Domain. Baltic J. Modern Computing, Vilnius, Lithuania, volume 3 (2015), No. 2, pp. 92-109.

NICOSIA, Giuseppe, Varun OJHA, Emanuele LA MALFA, Giorgio JANSEN, Vincenzo SCIACCA, Panos PARDALOS, Giovanni GIUFFRIDA a Renato UMETON, ed., 2020. Machine Learning, Optimization, and Data Science [online]. Cham: Springer International Publishing [cit. 2022-09-28]. Lecture Notes in Computer Science. ISBN 978-3-030-64582-3. doi:10.1007/978-3-030-64583-0

Nodeh, M. J., Calp, M. H., & Şahin, İ. (2020). Analyzing and Processing of Supplier Database Based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) Algorithm. In D. J. Hemanth & U. Kose (Eds.), Artificial Intelligence and Applied Mathematics in Engineering Problems (pp. 544-558). Springer International Publishing. https://doi.org/10.1007/978-3-030-36178-5_44

Olugu, E. U. / Wong, K. Y. / Shaharoun, A. M. (2011): Development of key performance measures for the automobile green supply chain. Resources Conservation and Recycling, Vol. 55, Issue 6, pp. 567-579.

Patil Rajkumar P. (2010): Process Robustness in Pharmaceutical Manufacturing. [Online] Available at https://learnaboutgmp.com/good-manufacturing-practices-cgmp/process-robustness-in-pharmaceutical-manufacturing/. Viewed 2021-08-28.

J.C.Dias Pinto (2016) Performance management Analytics for automotive industry :An empirical study

Perzyk M. / Dybowski B. / Kozlowski J (2019): Introducing Advanced Data Analytics in Perspective of Industry 4.0 in a Die Casting Foundry. Archives of Foundry Engineering. ISSN (2299-2944), Volume 19, Issue 1/2019. DOI: 10.24425/afe.2018.125191, pp. 53-57.

Prasad K.D.V.Yarlangadda, Eric Chenge Wie Chiang : A neural network system for prediction of process parameters in pressure die casting, School of Manufacturing and Medical Engineering, Queensland University of Technology, PII: S0924-0136 (99)00171-0

R.D.Raut, S.K.Mangla, V.S.Narwane, B.B.Gardas (2019): Linking big data analytics and operational sustainability practices for sustainable business management, . Dep. Pf Operations and Supply Chain Management, Vihar Lake, Mumbai,

Raza, M. et al. (2017): Effects of Process Related Variations on Fillablity Simulation of Thin-Walled IN718 Structures. International Journal of Metalcasting. [Online] Available at https://link.springer.com/content/pdf/10.1007/s40962-017-0189-9.pdf. Viewed 2021-09-23.

Raza, M. et al. (2018): Effects of Process Related Variations on Defect Formation in Investment Cast Components: Volume 18, Issue 1/2018, pp. 103-108. ISSN: 1897-3310.

Russom Philip (2011): Big Data Analytics. The Data Warehousing Institute, 4 (1), pp.1-40. [Online] Available at https://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx. Viewed 2021-08-15.

S.Tissir, A,Cherrafi : Lean Six Sigma and Industry 4.0 combination:scoping review and perspective (2022): DOI: 10.1080/14783363.2022.2043740

Sata, Amit. (2016): Foundry Data Analytics to Prevent Defects in Investment Casting. Conference Paper. [Online] Available at https://www.researchgate.net/publication/295731855_Foundry_Data_Analytics_to_Prevent_Defects_in_Investment_Castings. Viewed 2021-09-23.

Sentance Rebecca (2017): Gartner Analytics Ascendancy Model. [Online] Available at https://www.clickz.com/how-can-ai-allow-marketers-to-predict-the-future/112268/gartner-analytic-ascendancy-model/. Viewed 2021-08-18.

Shabbir, M. Q., & Gardezi, S. B. W. (2020). Application of big data analytics and organizational performance: the mediating role of knowledge management practices. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00317-6

Shiffman, Daniel. The Nature of Code : [simulating Natural Systems with Processing]. Version 1.0, generated December 6, 2012., [Selbstverl.], 2012.

Shivappa, D.N. / Rohit / Bhattacharya, A. (2012): Analysis of Casting Defects and Identification of Remedial Measures – A Diagnostic Study. International journal of engineering inventions. Volume 1, issue 6, pp. 1-5. ISSN: 2278-7461.

SIMONOVA, Stanislava a Hana KOPACKOVA, 2014. Utilization of Six Sigma for data improvement. In: The 10th International Conference on Digital Technologies 2014 [online]. IEEE, 2014, s. 305-310 [cit. 2022-09-26]. ISBN 978-1-4799-3303-7. Retrieved from: doi:10.1109/DT.2014.6868732

Srinivasan R. / Swink M. (2018): An investigation of visibility and flexibility as complements to supply chain analytics: An organizational information processing theory perspective. Production and Operations Management, 27(10), pp. 1849-1867.

STANKALLA, Roland a Felicita CHROMJAKOVÁ, 2019. A STRUCTURED REVIEW OF THE SIX SIGMA BELT SYSTEM. In: DOKBAT 2019 - 15th International Bata Conference for Ph.D. Students and Young Researchers [online]. Tomas Bata University in Zlín, 2019, s. 970-982 [cit. 2022-09-26]. ISBN 9788074548932. Retrieved from: doi:10.7441/dokbat.2019.095

Stevens Emily (2021): The 7 Most Useful Data Analysis Methods and Techniques, [Online] Available at https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/. Viewed 2021-08-11.

Surajit Bag, L.Wood, L.Xu, P.Dhamija, Y.Kayikci (2019) : Big data analytics as an operational excellence approach to enhance sustainable supply chain performance, Post Graduate School of Emgineering Managementm University of Johannesburg, Soufth Aufrica,

Sushovan Ghish, J. Maiti: Data mining driven DMAIC framework for improving foundry quality – a case study, Production Planing and Control, DOI:10.1080/09537287.2012.709642

Tallyfy (2021): Continuous Process Improvement (CPI): Definition and Techniques, [Online] Available at https://tallyfy.com/continuous-process-improvement/. Viewed 2021-08-26.

Treacy M. / Wiersema, F. (2007): The discipline of market leaders: Choose your customers, narrow your focus, dominate your market. New York, NY: Basic Books. ISBN-10: 0201407191.

Tsai Chun-Wei, et al. (2015): Big data analytics: a survey. Journal of Big Data, 2:21 DOI: 10.1186/s40537-015-0030-3.

Tseng, M. L. et al. (2018): Circular economy meets industry 4.0: can big data drive industrial symbiosis? Resources Conservation and Recycling, Vol. 131, pp. 146-147.

Vashisth Shubhangi et al. (2019): Hype Cycle for Data Science and Machine Learning. Gartner, 2019 - ID G00369766.

Wang, R., et al. "Prediction of Ameloblastoma Recurrence Using Random Forest—a Machine Learning Algorithm." International Journal of Oral and

Maxillofacial Surgery, vol. 51, no. 7, 2022, pp. 886–91, https://doi.org/10.1016/j.ijom.2021.11.017.

WAUYO, Fred, Edwin OMOL a James OKUMU. EFFECTIVENESS OF BUSINESS INTELLIGENCE T. European Journal of Technology. 2019, 1.(2.), 55-73. ISSN 2520-0712.

Wijaya Cornellius Yudha (2021): CRISP-DM methodology for your first data science project. [Online] Available at https://towardsdatascience.com/crisp-dm-methodology-for-your-first-data-science-project-769f35e0346c. Viewed 2021-09-01.

Wilkings Brian (2021): How to measure operational excellence. [Online] Available at https://www.eonsolutions.io/blog/how-to-measure-operational-excellence. Viewed 2021-09-02.

Zhong R. et al. (2016): Visualization of RFID-enabled shopfloor logistics Big Data in Cloud Manufacturing. The International Journal of Advanced Manufacturing Technology, 84(1-4), pp. 5-16. DOI 10.1007/s00170-015-7702-1.

# LIST OF ABBREVIATIONS

| | |
|---|---|
| 6S | six sigma |
| AI | Artificial intelligence |
| ANN | Artificial neural network |
| BD | Big data |
| BDA | Big data analytics |
| BDPA | Big data predictive analytics |
| BH | Baterry housing |
| CI | Capacity indices |
| $CO_2$ | Carbon dioxide |
| CIP | Continuous improvement process |
| CPK | Process performance capability |
| CP | Process capability indices |
| CRISP-DM | Cross Industry Standard Process for Data Mining |
| CTQ | Critical quality |
| DA | Data Analytics |
| DM | Data mining |
| DMC | data matrix code |
| DMAIC | Define, measure, analyse, implement, control |
| DT | Decision tree |
| Etc. | Et cetera |
| f.e. | For example |
| f.i. | For information / for instance |
| HPDC | High pressure die casting |
| IoT | Internet of Things |
| ISO | International Standard Organisation |
| KPI | Key performance indicator |
| LR | Likelihood Ratio |

| | |
|---|---|
| ML | Machine learning |
| MLOps | Machine learning operationalization |
| NBC | Naïve Bayesian classifier |
| NOK | Not Okay |
| $NO_X$ | Nitrogen oxides |
| OE | Operational excellence |
| OEE | Overall equipment effectiveness |
| OEM | Original equipment manufacturer |
| OEP | Operational excellence profitability |
| OK | part fulling quality specification |
| OTD | On – time delivery |
| PLS | Partial leaves square |
| PC | Process capability |
| R&D | Research and development |
| RF | Random forest |
| RSM | Response surface model |
| RST | Rough set theory |
| SOP | Start of production |
| SPC | Statistical process control |
| SVM | Support vector machine |

# LIST OF FIGURES/CHARTS/DIAGRAMS

# LIST OF TABLES