Università degli Studi di Genova

Dipartimento di Informatica, Bioingegneria,
Robotica ed Ingegneria dei Sistemi

Ph.D. Thesis in Computer Science and Systems Engineering
Systems Engineering Curriculum

Analytics and Intelligence for Smart Manufacturing

Ilenia Orlandi

May, 2018

Dottorato di Ricerca in Informatica ed Ingegneria dei Sistemi
Indirizzo Informatica
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università degli Studi di Genova

DIBRIS, Univ. di Genova
Via Opera Pia, 13
I-16145 Genova, Italy
`http://www.dibris.unige.it/`

Ph.D. Thesis in Computer Science and Systems Engineering
Systems Engineering Curriculum
(S.S.D. INF/01)

Submitted by Ilenia Orlandi
DIBRIS, Univ. di Genova

. . . . .

Date of submission: January 2018

Title:Analytics and Intelligence
for Smart Manufacturing

Advisor: Prof. Davide Anguita
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università di Genova

CoAdvisor: Prof. Luca Oneto
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università di Genova

. . . . .

Ext. Reviewers: Fabio Aiolli, Erik Cambria

Abstract

Digital transformation is one of the main aspects emerged by the current 4.0 revolution. It embraces the integration between the digital and physical environment, including the application of modelling and simulation techniques, visualization, and data analytics in order to manage the overall product life cycle.

In this thesis we want to emphasise two macro areas of analysis that can be performed thanks to the technologies provided by this manufacturing digitalization. The first macro area concerns the descriptives/diagnostic analysis that can be executed on data by means the Manufacturing Intelligence tools. Instead, the second macro area identifies the Manufacturing Analytics techniques, which enable to perform predictive/perscriptive analysis.

This thesis aims at providing several contributions to improve certain aspects of these two macro areas. In the framework of Manufacturing Intelligence, we will present a methodology that, starting from the data collected at the factory, enables to identify and then, visualize in several dashboards, a set of standard *KPIs useful to control production processes. Then, we will analyze the Sustainability as a strategic priority that emerged in this 4.0 revolution and that is becoming increasingly important among the industries' critical objectives. Our contribute is to provide a guideline that identifies the main sustainability trends and drivers in manufacturing environment and a set of KPIs that allows industries to control and monitor the reaching of these goals.*

Finally, in the framework of Manufacturing Analytics, we will present the application of several state-of-the-art Machine Learning techniques, in two manufacturing case studies related to the quality evaluation and the predictive maintenance. In both applications, we mainly focused on the performance evaluation of our models by means the exploitation of several model selection approaches that are well suited to Big Data problems. We will show how to apply these model selection strategies in order to configure such techniques achieving the best accuracy of the final models.

Table of Contents

List of Figures

List of Tables

Nomenclaure

AMR	Advanced Manufacturing Research
BI	Business Intelligence
BN	Bayesian Network
CFI	Cluster Fabbrica Intelligente
CP	Cleaner Production
CRM	Customer Relationship Management
DBMS	Data Base Management Systems
DM	Data Mining
DT	Decision Trees
DSS	Decision Support Systems
DWH	Data WareHouse
EIS	Executive Information Systems
ERP	Enterprise Resource Planning
ETL	Extract, Load, Transform
FI	Fabbrica Intelligente
ERP	Enterprise Resource Planning
GLM	Generalized Linear Model
GPR	Gaussian Process Regression
GSC	Global Supply Chain
GSCF	Global Supply Chain Forum
GUI	Graphical User Interface
HMI	Human Machine Interface
HP	High Pressure

HTML	HyperText Markup Language
ICT	Information and Communication Technology
IoT	Internet of Things
ISO	International Organization for Standardization
IT	Information Technologies
KPI	Key Performance Indicators
LP	Low Pressure
MA	Manufacturing Analytics
MDX	MultiDimensional eXpressions
MI	Manufacturing Intelligence
ML	Machine Learning
MOM	Manufacturing Operations Management
MS	Model Selection
MVMM	Manufacturing Value Modelling Methodology
NN	Neural Network
OData	OASIS Open Data Protocol
ODBC	Open DataBase Connectivity
OLAP	OnLine Analytical Processing
OLTP	OnLine Transaction Processing
RF	Random Forest
RL	Reverse Logistics
RUL	Remaining Useful Life
SCC	Supply Chain Council
SCM	Supply Chain Management
SM	Smart Manufacturing
SLT	Statistical Learning Theory
SVM	Support Vector Machine
WIP	Work In Progress

Chapter 1

Introduction

Today's industry is faced with huge challenges and issues in terms of demographic change (global population increase, ageing of society, urbanisation increase), new emerging markets (globalisation, growth in exports, growth of developing countries), scarcity of resources (energy, water, raw materials, other goods), climate change (CO_2 increase, global warming, the ecosystem at risk) and the acceleration of technological progress (exponential growth of technologies, cost reductions, greater pervasiveness)[CFI15]. These events has led to a new meaning of the concept *Manufacturing system*, no longer seen as an industry in which a set of operations brings the raw materials to the expected final product, making it more valuable than the original state.
Nowadays, this concept has become synonymous for *Supply Chain*, therefore, it has been extended to all individuals, organizations, resources, activities and technologies involved in the creation and distribution of a product. In other words, manufacturing systems covers the whole events flow that occurs from supplier to manufacturer, wholesaler, retailer and customer.
In the current competitive scenario, manufacturing companies need to transform the industrial systems engineering domain in order to meet an increasing level of variability. The variability means different sets of dimensions such as demand, volume, process, manufacturing technology, customer behaviour and supplier attitude.

This new paradigm is known as "the fourth industrial revolution" or "Industry 4.0". The term "Industry 4.0" refers to a new production patterns, including new technologies, productive factors and labour organizations, which are completely changing the production processes and the relationship between customer and company with relevant effects on the supply and value chains [Cas]. Even though most of the aforementioned innovations are in an embryonic stage, they are still an important part of research and progress. The association of these cause new "matched technologies" which could work in a physical and digital environment. Industries are changing their production and logistic operations replacing the traditional industrial systems with new emerging technologies including the "Internet of Things" (IoT), cloud computing, miniaturization and 3D printing that will enable flexible and sustainable industrial processes and intelli-

gent manufacturing.
Flexibility and sustainability are the emergent manufacturing competitive priorities/dimensions ([TM15, EBW16]). The traditional competitive priorities include quality, cost, speed, delivery and efficiency. Hayes and Wheelwright, and Leong [HW84], [LSW90] describe these competitive dimensions as “strategic priorities or goals or ways that are selected by companies to maximize competitiveness in the market”. During the last few years, various researchers have assumed additional competitive priorities and proposed some methods to evaluate their performances. Furthermore, several national strategies and new technological roadmap (e.g. the German high tech strategy “Industrie 4.0” or the US “Smart Manufacturing”) aim at approaching this transformation enhancing the sustainability, flexibility and re-configurability of current manufacturing systems among many other competitive dimensions. New emerging technologies could allow the next generation of manufacturing systems to become real smart factories [LBJ16].

Flexibility allows a rapid change in the priorities of the short-term performance (low costs rather than high quality) from the strategic point of view of achieving a trade-off between the various services [SS07]. It can be expressed as the system ability to create a new product in relation to marketing strategies, to make changes to a product in relation to market demands, to cope with fluctuations in the volume of market demand through variations in the production volume, to create different combinations of products with the same total volume, to perform machining of different products through set-up of equipment (tooling changes), to allocate the different products utilizing its own resources optimally, and to define programming changes based on the priority assignment of the work orders.

Sustainability is the ”development that meets the needs of present without compromising the ability of future generations to meet their own needs.” [McC91]. This definition embodies the concepts of resources and their narrowness in three perspectives: environmental, social and financial.
Environmental sustainability is the ability to maintain the quality and reproducibility of natural resources over time, to preserve biological diversity and to guarantee the integrity of ecosystems. Social sustainability requires that fair conditions of well-being and access to essential goods (security, health, education) are guaranteed for current and next generations.
Finally, financial sustainability has the objective of generating long-term income and work, and achieving eco-efficiency that is the rational use of the available resources and the reduction of non-renewable resources consumption.
The need to implement new innovation strategies aimed at orienting design, production and consumption approaches towards sustainable directions is a growing awareness in manufacturing environment [ACO12]. In recent decades, numerous eco-innovation methods and tools have been developed to support the design of new products, processes or services that are able to provide value to the customer, profit to the company and, at the same time, reduce the environmental impact. Companies are increasingly adopting the theme of ”green practices”, as that series of organizational and managerial tools and structures aimed at reducing the impact of the company’s activity on the ecosystem and implementing a strategy towards environmental sus-

tainability [TET13, TSM15, BME13, Rus07, LYQ+12]). These approaches are mainly aimed at energy production from alternative sources, recovery of waste and products, logistics optimization, product innovation, and efficiency in production processes and/or staff structures.

Re-reconfigurability implies the Digital transformation of manufacturing systems. Digital manufacturing allows the simulation of the whole supply chain with the idea of the virtual factory, integrating procurement, production, product logistics, service and diverse IT technologies in order to predict, solve and control problems in the virtual and physical environment. These technologies allow: i) the reduction of time to market, ii) decrease in costs and iii) and the increase of process flexibility by analysing production data.

This is the reason why, in this Industry 4.0 era, we add the adjective “Smart” to the term “Manufacturing”. This new popular expression, “Smart Manufacturing” (SM), refers to the adoption of all these new technologies and techniques within manufacturing industries in order to improve their production processes. So, in Smart Manufacturing, besides the introduction of new technologies, changes also the way by which those already existing are used. It is necessary that the different parts that make up a manufacturing system are interconnected and integrated in order to provide a clearer understanding of the overall performances. The integration of multiple and heterogeneous data sources and the convergence of process control systems and business intelligence layer such as the Enterprise Resource Planning (ERP) or Customer Relationship Management (CRM) are paving the way for important progress in plant operation optimization and management, thanks to the multitude of information they can collect.
Data play a primary role in this 4.0 revolution, up to the point that was coined the expression *Big Data* to denote this increasing amount of information produced with the digital transformation of industrial sector. Big Data are commonly described by three main characteristics:

- *Volume* indicates the exponential increasing of the available information;
- *Variety* means that the available data can be structured such as numerical and categorical, or unstructured like text, images, audio, and video data;
- *Velocity* indicates the way the data are handled and become available to be used. For instance, in many cases the processing of the Big Data is required to be performed in real time.

In SM systems, the introduction of a Big Data infrastructure implies a new way to use process data: not only for tracking purposes, but for improving operations. Hence, Big Data embodies the concepts of *Manufacturing Intelligence* (MI) and *Manufacturing Analytics* (MA).
MI and MA are referred to all those *Business Intelligence* (BI) and *Business Analytics* (BA) applications in manufacturing that enable to monitor production performances in terms of metrics. BI applications are software platforms dedicated to move data from one source to another, preprocess data (eliminate duplications, inconsistencies, missing data, incorrect values), transform

original data into aggregate ones and visualize these ”new” processed information into a single-screen display called *dashboard*.

Dashboards are reporting systems that show strategic quantifiable indicators by means graphical items (tables and charts) to give a simple and quick understanding of the current operations performances and to support final users in taking more tactical and strategic decisions, not only based on intuition but also thanks to an updated and accurate knowledge of data.

These strategic indicators, commonly called *Key Performance Indicators* (KPIs), are related to the main manufacturer’s activities and targets, at each level of the enterprise - from the macro financial level to the operation level. KPIs enable to “see” the results in an aggregated way, rather than using large amounts of information often unusable and/or unused. In a few indicators easy to understand, there are information able to provide an immediate diagnosis about the phenomenon in exam: the real time control of these indicators enables to take immediate decisions on those activities not performing well.

In addition to dashboards, that provide a description and diagnosis of a set of KPIs related to a specific scenario, more advanced techniques are exploited in Big Data perspective.

Before, getting all the useful data was a difficult task and many times it was impossible to retrieve all the information needed to perform an exhaustive analysis. Another problem is that traditional methodologies take too much time and computational cost for processing all the available data [MSC13]. Instead now, by means BA techniques, the large amount of data allow to perform exhaustive investigations in a way to generate predictions that can fit better to the population under exam.

This procedure is possible thanks to the development of several techniques such as modelling, Machine Learning (ML), Data Mining (DM) and game theory. These methods are well suited to solve complex problems in manufacturing scenarios where it is necessary process a large amount of data, take into account many constraints and variables and where, often, there aren’t mathematical equations to measure the physical models underlying the observed events ([SSA09, CSZ10, AGOR14, Bre01b, CSC17]).

These methods use computational procedures to “learn” information directly from the data and can adaptively improve their output accuracy as the number of samples available for learning increases. In short, data driven models [SSA09].

Data driven models exploit non-parametric inference, where it is expected that an effective model would stem out directly from the data, without any assumption on the model family nor any other information that is external to the data set itself [Bre01b]. With the advent of the Big Data era, this approach has gained more and more popularity, up to the point of suggesting that effective predictive models, with the desired accuracy, can be generated by simply collecting more and more data [Dha13].

These tools enable to perform predictive analysis and simulate future outputs based on previous decisions. The BA’s powerful computational features along with their integration into increasingly user-friendly and low-cost technologies has dramatically accelerated the use of these techniques in manufacturing applications.

1.1 Thesis Contributions

Thanks to the huge quantity of data provided by the integration of new technologies along with the traditional ones, and their elaboration throug Big Data approaches, it is now possible measure and manage manufacturing performances. Furthermore it is also possible understand not evident relations between different events, comprehend cause-effect relationships, and predict future events.

Despite the important opportunities offered by these technologies and techniques, industries struggle to adopt new approaches. There is still little awareness of the power provided by the data, few people know how to identify strategic KPIs to make in-depth analysis of the performance of their industry, it lacks a corporate training plan aimed to learn how to exploit these new technologies, and again, these sophisticated analytics are applied on specific problems without extending them to the whole system [DFL12].

For these reasons, with the aim to develop an interconnection between research and industry so to know the existing problems and transfer the development of new solutions within the factories, was founded, in 2012, the italian national technology Cluster “Fabbrica Intelligente” (CFI), a No-profit Association that groups 72 big and medium-small size enterprises (industrial members), 15 universities and research centres (research members) and 10 associated members.

This Ph.D. has been supported by the Italian Ministry of University and Research (MIUR), through the “SMART MANUFACTURING 2020” project which is one of the four projects submitted by the CFI.

This project aims to promote the ”development of new systems of planning and productive monitoring for the optimization of the use of resources and energy, new manufacturing solutions for the preventive and intelligent maintenance, and new advanced methodologies for the virtual prototyping of products and productive processes”[1].

In this context, this Ph.D. research focused on two main objectives.

On one side, to analyze what are the key KPIs in manufacturing that can support and improve these operations. On the other side, to apply predictive ML techniques in smart manufacturing problems in order to provide some advantages in their use.

Concerning the first topic of this research, we will present a practical methodology that, starting from the data collected at the factory, enables to identify and then, visualize in several dashboards, a set of KPIs useful to control production processes. This is the first step in Big Data analysis, the descriptive/diagnostic investigation on data available in order to extract the real useful information.

Then we will move to the analysis of the Sustainability as a strategic priority that emerged in this 4.0 revolution and that is becoming increasingly important among the industries’ critical objectives.

As discussed at the begin of this Chapter, the companies business model, traditionally structured to correlate with a small number of known subjects and with deterministically definable needs,

[1] www.fabbricaintelligente.com

is now facing a context where the counterparts rapidly change and have equally variable needs which are independent from the value creation proper of the company itself. This observation about social field, antithetical to that focused on the mere profit generation, requires to move from a measurement of value based only on monetary indicators (profits and wages), with a short-term horizon, to one that, in the long-term, allows the company to sustain the conditions underlying its prosperity. This leads to the creation of reference frameworks that base their decision-making criteria on social and environmental values and not only on organizational criteria aimed at maximizing profit, internalizing aspects so far considered ”externalities”. Every industry carries out actions that can have an economic, social, and environmental impact outside its confines and so, it is responsible for these actions. The search for sustainability as a goal to be achieved is a research field faced by many academic. However, there is not yet a structured approach able to manage sustainability performance indicators ([MML$^+$13, KMK15, JH04]).
Our contribute is to provide a guideline that identifies the main sustainability trends and drivers in manufacturing environment and a set of KPIs that allows industries to control and monitor the reaching of these goals.

The second scope of this thesis is to provide contributions in the application of several state-of-the-art ML techniques in order to adapt them to manufacturing problems.
This is the next step in Big Data analysis, the predictive/prescriptive investigation on data already processed in order to evaluate future behaviors about the phenomenon in exam.
We have already introduced the motivations that have driven the intense development and application of these techniques into complex scenarios such as factories.
The scientific literature abound of applications of data mining and machine learning techniques devoted to building predictive models for monitoring manufacturing systems [JH93, ABT02, KBT11, LLBK13, WIT14] but the assessment of the predictors themselves is quite rare [OOA15].
To this end, we will focus on the analysis of data-driven model approaches in order to evaluate the accuracy of these classifiers. In particular, we will address two manufacturing problems that are fundamental activities within a production process.
In the first one, we exploited the Support Vector Machine (SVM) algorithm to build a model for quality evaluation of final products in an assembly line of refrigerators.
Whereas, in the second one, we applied the Random Forest (RF) algorithm for creating a model able to predict future maintenance on a production line of calipers. In both applications, we mainly focused on the performance evaluation of the obtained models by means the exploitation of several model selection approaches that are well suited to Big Data problems.

1.2 Outline of the Thesis

The remainder of this thesis is structured as follows:

Chapter 2 describes the research background and identifies the key elements of Smart Manufacturing systems. Particular attention will be spent on the definition of the main kind of analysis and relative tools that can be exploited to manage and/or investigate manufacturing events.

Chapter 3 presents the design of a sustainability catalogue starting from the analysis of the existing state of the art on industrial sustainability performance management. We will re-organize these information accordingly to a hierarchy of structured KPIs, highlighting the areas that need to be further developed both in the literature and in the industrial systems.

This work have been presented in the XXI Summer School "Francesco Turco" with the title "Investigating sustainability as a performance dimension of a novel Manufacturing Value Modeling Methodology (MVMM): from sustainability business drivers to relevant metrics and performance indicators" [DOTA16]

In **Chapter 4** we will discuss several application of ML techniques in order to adapt them to Manufacturing problems. In particular, in the Section 4.1, we will review the main results of Statistical Learning Theory (SLT) for the purpose of assessing the performance and quantify the uncertainty of a predictive classification model applied to the final product quality estimation in a smart manufacturing system. Then, in Section 4.2 we will examine a model selection procedure for choosing the best set of hyperparameters that allows to build a Random Forest model characterised by the best generalisation performances.

The contribution of the Section 4.1 have been published in the IEEE Conference on Big Data with the title "Performance Assessment and Uncertainty Quantification of Predictive Models for Smart Manufacturing Systems" [OOA15].
The work discussed in Section 4.2, have been presented in the European Symposium on Artificial Neural Network, Computational Intelligence and Machine Learning with the title "Random Forest Model Selection" [OOA16].

Chapter 5 presents a practical implementation of a set of business dashboards by means the "Analytics BI" software platform. In particular, we will discuss the KPIs model design and how this BI software can be exploited in order to visualize these information.

Chapter 6 describes the development of a model for predictive maintenance in a production line of calipers. This activity has been made possible thanks to the collaboration established within the "Smart Factory 2020" project with an Italian manufacture that is a world leader in the production of braking systems for vehicles.

Finally, in **Chapter 7** conclusions and future developments are discussed.

Part I

Related Works

Chapter 2

From Manufacturing to Smart Manufacturing

Manufacturing is the transformation of input materials in output elements through the realization of technological processes (physical, chemical, mechanical) [Gro07]. At the same time, these operations makes the material more valuable than the original state. Therefore, Manufacturing can be also defined as a set of activities aimed to procure utility (needs fulfilment), or utility increase of generic stuff. These definitions highlight two aspects of manufacturing: technical and economic (Figure 2.1).

The market evolution, over the last 50 years, has led to a new meaning of the concept production, intended as a corporate function. Strictly speaking, production and production system identify the real physical transformation of goods within industrial companies (manufacturing). However, there is also a strong link between the production activities and the logistic activities. These last cover all flows of raw materials from supplier to plant, all flows of semi-finished products among the various production phases and all flows of finished products to retailers. In general, logistics can be understood as a company area that oversees the management of physical and information flows along the supply chain from suppliers, to the company and to the final customers. Logistics can be external (procurement and distribution) and internal (transformation). Furthermore, production operations are associated to design and industrialization activities. So, in conclusion, the term production has gained a broader meaning that includes logistics and product development, in addition to manufacturing ([BSW95, IOK94, ElM05]).

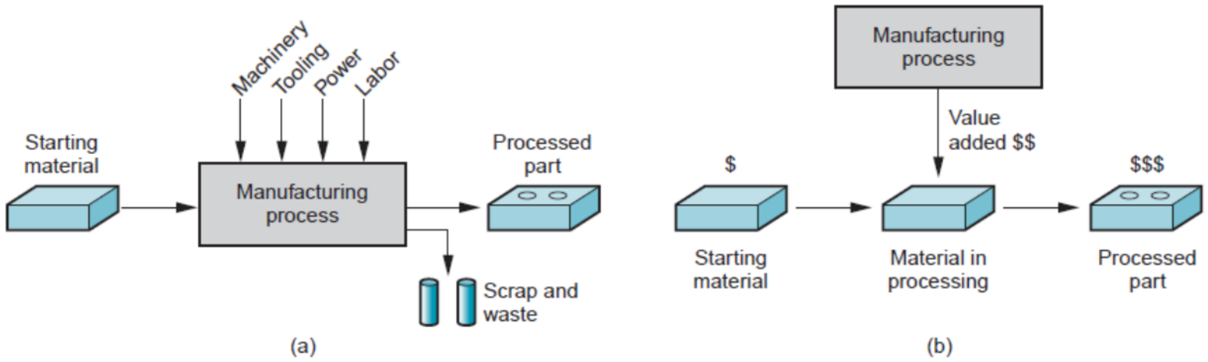


Figure 2.1: Two ways to describe manufacturing. (a) technical process, (b) economic process. Source: [Gro07]

2.1 Production Systems

Industrial plant or manufacturing system are commonly defined as a set of different elements of any kind, integrated and interacting, whose objective is the production of goods or services are expected to benefit. In particular, a production system is characterized by ([MSMC78, BCM97]):

- *Objectives*: the set of goals that justify the existence of the system;
- *Structures*: the elements that compose a production system, distributed into different sub-systems;
- *Processes*: the activities carried out by the elements of a plant, whose effects are addressed both within the system itself and outside;
- *Interrelations*: the relationships between sub-systems, processes, and between sub-systems and processes.

In general, any type of organization can be described according to a systemic approach, specifying its objectives, parts, processes and interrelations. Thus, understanding a manufacturing system corresponds to understanding the behaviour of the overall organization. The company's objectives are its institutional aims and are different according to the company type: for a public service company, the goal is to provide the best possible quality and cost conditions with a set of services (transport, communication, and so on); for a private company, the goal is to obtain profits through the design, production and sale of goods. For a production company, in particular, it can certainly be said that the main objective is to create value, both for customers that satisfy their own usefulness in the acquisition of goods, both for itself, obtaining a monetary profit from the products sales. The parts of a company are the building blocks of the company itself. It is useful to note that among these parts there are not only the physical elements (machines, plants, and so on), but also of the pure organizational units (directions, services, offices, etc.). A production system consists of interacting sets of operations (tasks), materials, resources, products, plans and events. The plans contain the process plans, i.e. the flows of elements in the system, and the production plans. Processes are all those functions necessary to achieve the operational objectives set by the company itself, and are implemented within the structure. The operative objective of a manufacturing system is the production, that is the process of transformation of raw materials into finished products. In order to better understand the processes of a company, it is necessary to deepen the concept of resource. By corporate resource we mean every entity that the company exploits in its processes to pursue its operational objectives. Resources are: the products or services offered by the company, the materials employed, goods, but also, the money and the people employed to produce. In addition to the resources mentioned, which are examples of the so-called internal resources, we must also consider external resources, on which the company operates indirectly only, such as, the social environment and economic operators. The resources acquisition, for a manufacturing company, is a key moment of the production process since that, for each resource, is associated a fee. Moreover, the choice of a technologically more advanced resource than another may mean a greater speed of execution and precision of the process to which it is associated. In the literature there are a large number of classifications of company resources ([SBJ02, YSDL96, SNW07]). Specifically, for a manufacturing system, the resources can be grouped in:

- *Proprietary productive resources*: all the plant and machine resources managed directly and used to realize the production itself. The cost of availability of these resources is fixed and it can include depreciation costs, maintenance costs and power costs. The operational scope can't significantly affect these costs, but can pursue maximum productivity; this objective is typically defined as saturation of productive resources.

- *Operational resources*: these are the necessary and not strictly productive resources (human resources). Their use is proportional with the requirements. The operational planning task is to manage human resources utilization in a flexible and consistent manner respect to the utilization productive resources. Furthermore, operating resources and productive resources make up the work centres that represent an important constraint on the production capacity itself.

- *Instrumental resources*: these are resources whose management is of strategic importance for the production capacity constraint. Operational planning must manage all situations of contemporaneity since a particular tool, of which there is only one sample, is able to penalize the production capacity of several work centres.

- *External production resources*: the reasons for adopting a third party processing can be systematic or temporary; in the case of production planning and scheduling, this type of solution is used, after having saturated the internal production resources, if the production capacity is not sufficient. At the aggregated planning level, preliminary agreements are stipulated with the contractors in order to consider them, during the operational planning phase, as machining units with capacity, average response time and costs already defined.

Starting from the resources it is possible to build a model of the production system and its processes, and then, define the structural units in which the processes take place and the resources are managed. Once the system model has been built, it is necessary to understand the interrelations between the elements, in order to understand its behaviour. In this sense, the privileged feature is the elements flow (materials, products, information, resources) within the system boundaries, and, therefore, the analysis of the system changes over time. The materials are moved through the personnel employed, the transports, the machinery, and undergo several processes that transform them into products. Resources are assigned to tasks, which need them to perform these transformations. The events are the time instants where resources begin and end a task. Information about the status of resources, materials and products indicates the status of the system at any time. The complexity of a production system often leads to break down the management of a production process into a series of sub-phases. Frequently, the individual steps in which a process can be decomposed are defined operations. An operation occurs when an object is intentionally modified in its chemical or physical characteristics, or assembled or disassembled; when it is transported, stored or inspected ([Bla04, Par05]). A first subdivision of production processes can be made by referring to the main activities carried out within a production system ([H$^+$05, CAJ98]):

- Acquisition of materials, information and production factors;
- Transformation:
  - Conversion or processing: physical creation of products through the transformation of the resources (alteration of physical characteristics, dimensions, shape or others features);
  - Waiting and storage: relate to the location and storage of an existing good (changes of materials and products availability and time);
  - Transport: moving materials and products from one place to another (change of physical position without changing the input form or state);
  - Control: activities aimed at identifying defects or preventing them.
- Distribution of the product to the customer.

Therefore, a production system is strongly conditioned by three fundamental variables ([Ack70, GVD02, ORW01, BO04]):

- Upstream interaction with suppliers;

- Downstream interaction with customers;
- Product's characteristics.

The first two points represent the interaction of the production system with the exterior and results in a change of the internal structure for better interfacing with the environment. The third point represents the specificity of the element to produce, around which the production process is built in order to achieve the product's conformity to the required specifications. The production strategic parameters are the internal variables on which the company can operate to construct an action that generates an outside competitive advantage over competitors. Hayes and Wheelwright [HW84], and later Gunn [Gun88], proposed some key factors that can give competitive advantage: Production capacity, Service level, Technological level, Degree of Integration, Workforce, Quality, Production planning, Organization, Flexibility, and Inventory turnover index.
A fundamental characteristic of these parameters is that they are quantifiable indicators, since it is possible to improve and optimize the performance only if it is possible to measure it. A common expression used for these quantifiable parameters is "Key Performance Indicators" (KPI).
In order to optimize the management of a production system and to sustain a competitive industrial context, it is necessary to plan the manufacturing operations according to a set of identified strategies.
In complex organizations like the manufacturing systems, the strategic objectives cannot be achieved immediately, instead, they need to be broken down into sub-targets more operational and detailed: formulation of company policies, operational plans, procedures and rules ([DSH89, Ste85]). In the management process it is possible to distinguish different hierarchical levels: moving from one level to the next, the higher objectives (decisions strategic) are fragmented into more sub-objectives; the degree of abstraction (abstraction from the details of objectives realization) decreases and, instead, the operational and procedural aspect (passage from what? to how?) increases [Gui00]. Every phase must pursue the objectives of the upper hierarchical level, using the tools provided by the lower one. Strategic objectives are wide-ranging instruments with a far horizon of forecast, while plans, programs, and procedures have a more detailed level. Therefore, management is a dynamic process that exploits different levels of abstraction and aggregation, both to simplify the complexity of the problem, and to cope with the different degree of information available along the time axis.

Information is available only in the immediate proximity of events; while it is very scarce during the strategies formulation, because of the temporal distance that separates them from their objective. The strategic plan is that which is more far from its objective, and the typical order of magnitude is a time horizon of 3-5 years. The realization of a strategic plan needs the formulation and achievement of several business plans: sales and marketing plan, research and development plan, financial plan, production plan. For each one of them, there is a further fragmentation: in the case of the production plan, there is an aggregated plan (1 year), a main production program (Master Production Schedule, 3-6 months), a possible assembly program (1 month) and finally an operational control (continuous). Depending on the tactical and operational levels of choice, it is possible to identify three basic moments for each phase of management:

- **Planning**. That is the management function: the selection of an organization's objectives, the establishment of the necessary strategies, policies, procedures, programs and projects at their achievement. More specifically, the production planning means the overall level of production.
- **Schedule**. That is the execution of a planned goal, exploiting the tools available to a certain hierarchical level. Therefore, Scheduling defines what must be done, in what quantity, and in which deadlines. This phase comes after the planning phase because its task is to execute a goal with known feasibility. The scheduled program must be feasible and the best among the possible alternatives, exploiting the most of the allocated resources.
- **Control**. The deviation between programs and reality is inevitable. For this reason a control phase is needed, aiming to provide the necessary information to direct corrective actions. This last phase controls the progress of activities respect to the scheduled program, timely highlights the serious divergences in order to carry out the interventions pointed out by the management.

2.2 Supply Chain and Management

Manufacturing systems are the core of a larger system that includes several specific aspects called *Supply Chain* (SC). As we already said at the beginning of this Chapter, the term *production* has acquired such a broad meaning that it is often synonymous with SC.
The Supply Chain official definition was provided by the Supply Chain Council, an independent no-profit organization founded in 1996 whose members belong to companies and organizations interested in the application of advanced Supply Chain Management techniques and systems. According to the Supply Chain Council "the Supply Chain includes all the efforts involved in the production and distribution of a finished product, from the supplier to the customer's customer" [LV99].
In literature it is possible to find innumerable definitions of the term Supply Chain. These definitions emphasize the organizations that make up the supply chain ([CM07, MMRC01, Bea98]) or the activities performed by supply chain ([Har01, HN99, LV99]). The Supply Chain concept involves all the components, directly or indirectly, interested in satisfying the customer's request: not only the manufacturer and the suppliers but also the transporters, the distributors and even the consumers themselves. It is necessary to note that the supply chain is not only a chain of companies with business relationships, but rather, it is a network of companies with multiple relationships (([LC00]). For this reason, this concept is often expressed with the most appropriate term of "Supply Network" ([WH04, NDZ02]). As stated in the previous paragraph, from the point of view of a single company, a supply chain is made up of two distinct networks:

- one upstream (upstream network), formed by the company's suppliers: it covers all flows from the manufacturing company up to the initial suppliers;
- one downstream network, formed by the enterprise's customers: it covers all flows from the manufacturing company to the final consumers.

Figure 2.2 summarizes the concept by showing the supply network of a reference company, called "focal company".

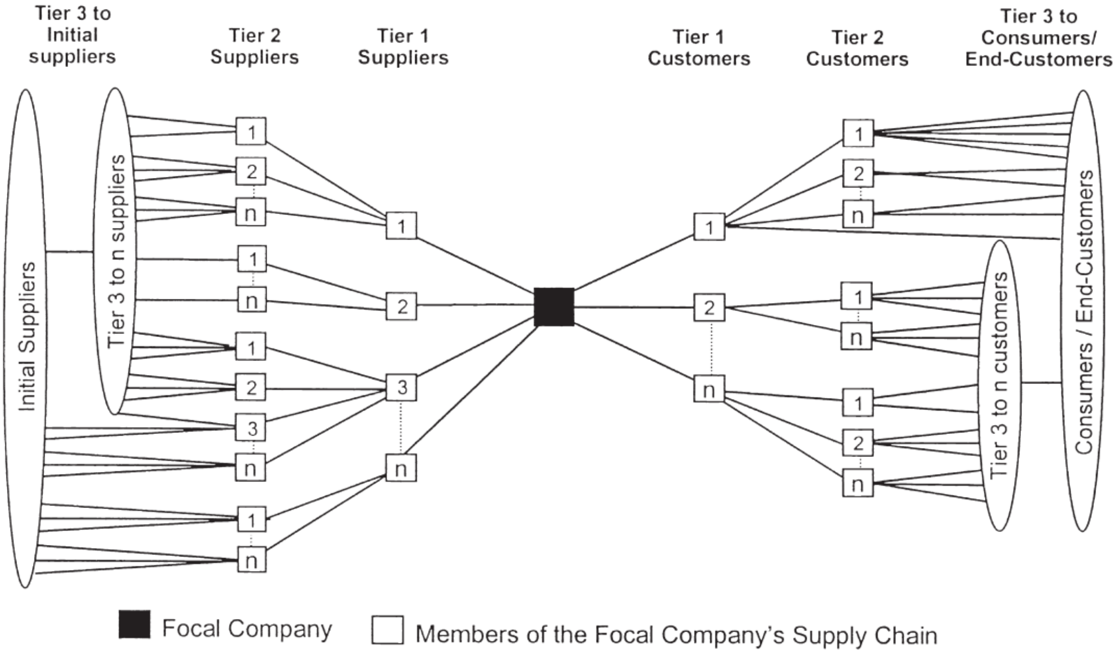


Figure 2.2: Structure of the supply chain.
Source: [LC00]

Therefore, in today's competitive environment, the success of a single business depends on the management's ability to integrate the company's complex network of business relationships. This capacity of managing multiple relationships along the supply chain refers to the term "Supply Chain Management" (SCM). This expression has an official definition, formulated in 1996 by the Global Supply Chain Forum (GSCF), a group made up of representatives of companies and academic researchers that met regularly since 1994 to improve the theories and management practices of SCM. GSCF defined the SCM as "the integration of the key business processes from the final consumer to the source suppliers that provides products, services and information that add value to consumers and other stakeholders" [FSC06].
In other words, the systematic and strategic coordination of traditional business functions within a single company, becomes transversal to the different SC businesses, in order to improve the

long-term performance of individual companies and of the entire supply chain. Starting from this definition, it is possible to highlight some fundamental aspects of Supply Chain Management:

- Materials Management, the information and financial resources flows. While the physical goods flows are mainly directed by the sources of raw materials towards the final market, the information flows (related to orders) and the financial flows (related to transactions) go back to the supply chain ([CM07]).
- Processes Management, it goes beyond the boundaries of individual company functions and individual supply chain companies. The main processes of SCM are: management of customer relations, demand management, order management, production management, purchases, product development and reverse logistics ([LC00]).
- The search for an optimization of global performance, as well as local ones. The goal of the supply chain should be, in fact, the maximization of the total value generated, not only at each individual stage. The management of sharing the value among the different members of the supply chain is one of the major challenges of Supply Chain Management ([CM07]).
- The need for coordination among the various companies, also through long-term collaborations that involve information sharing (for example demand forecasting, promotional activities, orders received, stocks of materials, components and products) or an integration of logistics and production activities (for example, through techniques such as Just in Time, Vendor Management Inventory, Collaborative Planning Forecasting and Replenishment). The development of information and communication technologies is the greatest enabling factor for the adoption of cooperative practices by an increasing number of companies ([Spi06]).

2.3 Megatrend in Manufacturing

With the ever-increasing globalization and competition, industries have had to change their operational procedures with the introduction of new technological systems in order to face sustainability issues, costs, management of activities and, in general, to properly evaluate the performance of their processes. Numerous studies have identified and analysed these megatrends of change and development. In this Section, the main trend which drive the new era of manufacturing are investigated:

- Supply chain complexity;
- Mass customization;

- Demographic Change;
- Industrial Sustainability;
- Technological Innovation;
- Globalization;
- Regulatory constraints.

2.3.1 Supply Chain Complexity

Supply chain complexity can be defined as the level of detail and dynamic complexity exhibited by the products, processes and relationships that make up a supply chain. Three drivers stand out in terms of their impact on plant performance: i) lengthy supplier lead times; ii) instability in the master production schedule; iii) variability in demand. Furthermore this complexity even increases when considering the customization as well as the digitalization and the resulting interconnectivity between the product and business processes [BWFF09].

2.3.2 Mass customization

Mass customization relates to the ability to provide bespoke products or services through flexible processes in high volumes and at reasonably low costs. It has been identified as a competitive strategy by an increasing number of companies. Agility and quick responsiveness to changes have become mandatory to most companies in view of current levels of market globalization, rapid technological innovations, and intense competition. Mass customization broadly encompasses the ability to provide individually designed products and services to customers in the mass-market economy. [DEP$^+$12]

2.3.3 Demographic Change

Over the next decades, it is estimated that the global population will grow by 18%, reaching a total of 8.4 billion people by 2040. One consequence of these changes will be an increase in the percentage of senior workers. It will therefore be necessary to find a balance between the need to allow over-65s to prolong their working life and the need to create new job opportunities for the young. Moreover, this also meets the broader need to increase the well-being of all workers in terms of increased satisfaction, safety and inclusivity (Roadmap *Cluster Fabbrica Intelligente*).

2.3.4 Industrial Sustainability

According to the traditional industrial view, product design and process technology typically determine the types of pollutants emitted, solid and hazardous wastes generated, resources harvested and energy consumed. Unfortunately, in a business environment of resource and energy supply uncertainty, the traditional view and the related business model, requires the continuous exploitation of new markets for growth, the enhancement of products to maintain demand and global sourcing to sustain margins, whilst absorbing the costs of compliance with end of life cycle legislation, is clearly unsustainable. It is generally agreed that sustainability has environmental, social and economic dimensions [REAG12].

2.3.5 Technological Innovation

Technological innovation has historically been considered the main effective source of competitive advantage among enterprises and the economic growth and social benefit of countries. Innovation has the ability to not only increase productivity but also creates processes for new types of products. Innovation can also provide the means for manufacturing flexibility [MPG14].

2.3.6 Globalization

The trend in globalization has changed the ways of connecting customers with products and therefore the factors that are analysed, be it the company, the manufacturing network, or the supply chain. Because of globalization, the vast majority of manufacturing in large companies is carried out in value networks. As globalization of markets raise competitive pressures, one essential requirement for the survival of organizations is their ability to meet competition. Market needs cause unlimited changes in the life cycle, shape, quality, and price of products [SWMK11].

2.3.7 Regulatory constraints

From a regulatory point of view quality refers to the basic objective requirements under the existing laws to assure that goods/food are safe, not contaminated or adulterated or fraudulently represented. Food quality and safety requirements are neither optional nor negotiable [LRS12].

The described mega-trends characterize and influence the global scenario of competition for the manufacturing sector. The specific challenges posed by mega-trends must be dealt with by implementing industrial strategies that follow the development of appropriate strategic actions.

2.4 Smart Manufacturing and Industry 4.0

The previous Sections highlighted the key factors that affect the actual global changes and developments in manufacturing environment. These innovations, that have a strong impact on the lives of people, companies and communities, influencing the economy and consumption, can be identified in a single expression that is “Industry 4.0”.
This definition was pronounced for the first time at the annual Hannover Fair in 2011 by a working group dedicated to Industry 4.0. When we talk about Industry 4.0, we refer to the fourth industrial revolution, the one that characterizes nowadays. The term “revolution” describes a radical and unexpected changing and this is the scenario we are assisting: this revolution supports the creation of the “smart factory” in which physical, digital and flexible production systems are integrated with the aim of reaching “mass personalization” and “faster product development”. Industry 4.0 is also identified as the “digital revolution”, focussing on all those digital technologies that are able to increase the interconnection and cooperation of resources (people or computer systems) without limiting themselves to one sector rather than another. We are witnessing changes, even radical ones, affecting the industrial sector, from the production of goods and services, to the society in all its aspects. The common factor is, indeed, the communication, or in other words, the interconnection between several elements of a system. High levels of communication and the optimal exploitation of all those services related to it, will become the main goal for anyone who wants a “4.0 perspective”.
In this scenario, the concept of *data* plays a primary role, since it underlies any operation. It went from a simple information born and dead in a small local system, to a tool able to create value. This is the reason why, in Industry 4.0 era, we talk about Big Data. This expression means all the operations related to the collection and interpretation of the enormous quantity of information extracted from several heterogeneous sources through the use of advanced computer science techniques. Big Data, in turn, introduces a new way for analysing manufacturing problems.
We are talking about the Manufacturing Intelligence and Analytics.
These new approaches are having a great success to solve complex problems in which it is necessary to manage a lot of information and constraints, and there is no knowledge about the physical models underlying the analyzed events. Although this revolution is leading to noticeable changes, there are still many open problems that require new developments and research to be solved. One example is that, despite the importance of the data, manufacturing continues to be data rich and knowledge poor, and as a result, operates with constricted decision processes, even in operations in which sophisticated modeling and control technologies are used.

Another key problem is due to the lack of a real integration of these Smart Manufacturing systems within production processes: these tools should allow the orchestration and execution of production operations with the aim of supporting individual production, providing advanced decision support. However, business plans and day-to-day management decisions are being implemented with incomplete knowledge of the relationship between product output, resource use and environmental impacts ([RB12, GPM04, DEP$^+$12]).

The 2017–2018 annual Critical Issues Agenda published by Manufacturing Leadership Council[1] provides an industry articulation of the most urgent and important issues facing the global manufacturing industry in the year ahead. This agenda identifies seven critical topics that need to be addressed by large and small manufacturers.

A **Factories of the Future**: industries must increase the use of technologies in order to improve their production processes. This implies the use of roadmap, maturity models, agile production models, end-to-end digitization and analysis of manufacturing and engineering processes and functions.

B **The Collaborative Manufacturing Enterprise**: in order to maintain a solid competitive advantage, manufacturing companies should coordinate activities and processes involving heterogeneous departments and maintaining collaborative contacts with customers, suppliers, third parties, distributors and partners.

C **Manufacturing Enterprise Innovation**: industries must use new approaches such as, best practices or collaboration with other manufacturing entities, in order to change their processes methodologies toward more innovative ones.

D **Transformative Technologies in Manufacturing**: industries need to adopt new Industry 4.0 technologies such as Internet of Things, 3D printing, advanced analytics, modelling and simulation, machine learning, collaborative robotics.

E **Next-Generation Manufacturing Leadership & the Changing Workforce**: there must be synergy between leaders and their teams. Leaders have the responsibility to promote their own and their team's continuous education, and to propose their own objectives so as these can be achieved through the activity of the team.

F **Cybersecurity in Manufacturing**: An industry that wants to develop its competitive potential and respond to the challenges of integrated, flexible and increasingly connected production can't afford to underestimate the risk related to the interconnection of its assets. Therefore it is necessary to analyze the reliability of the company systems and to increase the information security culture among companies.

G **Manufacturing 4.0 Sustainability**: Sustainable manufacturing is becoming increasingly important. This requires sustainable industrial system different from today's global industry: new business models, strategies, frameworks, and tools that allows to achieve competitive advantage through the increase of material efficiency, energy saving, closed-loop control at industrial system level.

[1] https://www.manufacturingleadershipcouncil.com/

In the next Chapters we will present several contributions and results achieved during this Ph.D. related to several points of the aforementioned agenda. In particular, we have investigated three topics:

- Development of tools to help industries to control their processes (A). An achieved result on this topic consists in the design of a set of dashboards with the use of a BI software platform. These dashboards aim to be universally valid and industry-neutral, in order to monitor several indicators, specific for the needs of manufacturing companies. More details about this application will be provided in Chapter 5.
- Development of new technologies (D). Regarding this point, in Chapter 4 and Chapter 6, we will present some contributions of this thesis related to the application of ML techniques in smart manufacturing problems.
- Holistic, sustainable manufacturing business models (G). In Chapter 3, we will discuss the development of a sustainability framework for measuring performance in Sustainable Supply Chains.

2.4.1 Manufacturing Intelligence and Manufacturing Analytics

As introduced in the previous Sections, the global evolution of the manufacturing market together with the important technological developments of the last years have led to the collection of an unprecedented amount of data. For instance, the integration of sensor systems on machineries, reporting tools that collect data coming from different sectors connected to the industry, and new hardware and software solutions designed to enable the communication between the various units of a manufacturing system.

Several terms have been coined to describe the features of these new technologies. This Section provides a detailed description of the meaning of the terms ”Manufacturing Analytics” and ”Manufacturing Intelligence” as these are the key topic of this thesis.

The expression *Manufacturing Intelligence* is coined by Advanced Market Research (AMR) and it is referred to all those Business Intelligence application in manufacturing that enable to monitor production performances in terms of metrics [Unv13].

These tools, indeed, can connect to various sources allowing to extract different kind of data (and so, different kind of information), offering several functionalities to perform data processing (e.g. mathematical calculation, aggregations) and, finally, providing a lot of graphical items to visualize these “new” information inside specific containers called “dashboards”. In this way, simple data are converted into more structured information (metrics).

Similarly, *Manufacturing Analytics* refers to the application of Business Analytics techniques in the manufacturing domain.

Now, let’s go in detail to investigate what BI and BA are.

The term *Intelligence* has been used by researchers in artificial intelligence since the 1950s. Business intelligence became a popular term in the business and IT communities only in the 1990s. In the late 2000s, *Business Analytics* was introduced to represent the key analytical component in BI ([Dav06]). Both BI and BA refer to the activity of collecting and analysing a large amount of data in order to extract knowledge about the investigated problem and about its context. Business Intelligence is the foundation of the technologies that have been developed later with the Business Analytics. This can be guessed also by the chronological order in which these two terms were introduced, since BI is about 10 years older with respect to BA. In order to understand the objectives of the BI and the BA and, also, the different technologies that derive from them, it may be useful to start from the description of the four categories of analysis commonly known in the state of the art: Descriptive, Diagnostic, Predictive, Prescriptive ([CV70, SSJ14, MSDS14, Hui15, CSM12, BBA13, EL12]).

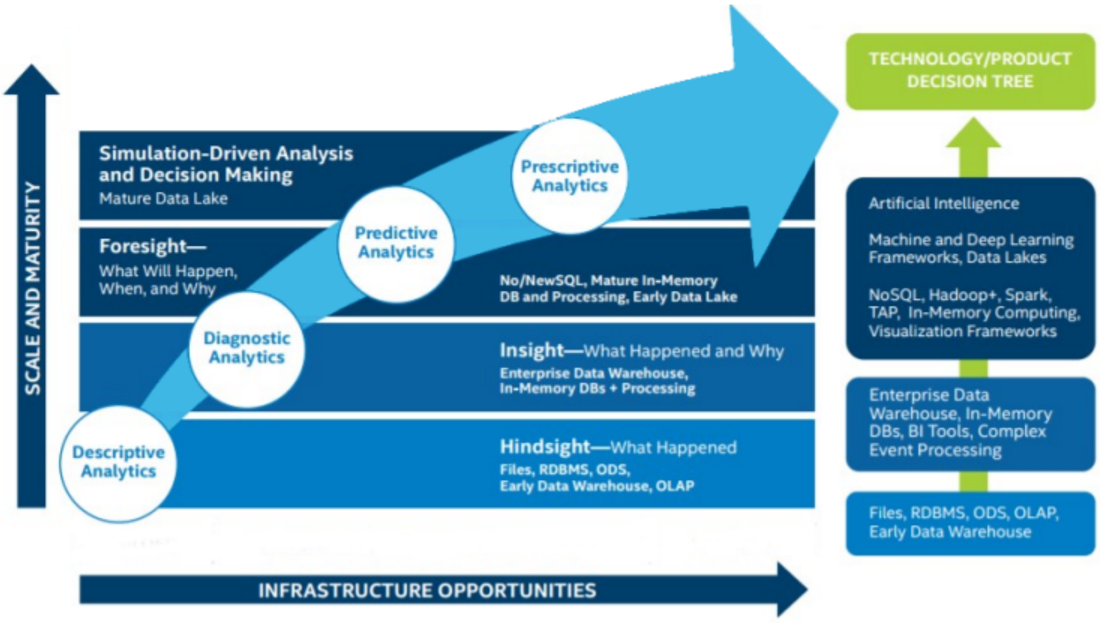

SCALE AND MATURITY
Simulation-Driven Analysis and Decision Making
Mature Data Lake
Foresight—
What Will Happen, When, and Why
No/NewSQL, Mature In-Memory DB and Processing, Early Data Lake
Insight—What Happened and Why
Enterprise Data Warehouse, In-Memory DBs + Processing
Hindsight—What Happened
Files, RDBMS, ODS, Early Data Warehouse, OLAP
Descriptive Analytics
Diagnostic Analytics
Predictive Analytics
Prescriptive Analytics
INFRASTRUCTURE OPPORTUNITIES
TECHNOLOGY/PRODUCT DECISION TREE
Artificial Intelligence
Machine and Deep Learning Frameworks, Data Lakes
NoSQL, Hadoop+, Spark, TAP, In-Memory Computing, Visualization Frameworks
Enterprise Data Warehouse, In-Memory DBs, BI Tools, Complex Event Processing
Files, RDBMS, ODS, OLAP, Early Data Warehouse

Figure 2.3: The four types of Analytics and their relative tools.

As represented in Figure 2.3, these type of analysis can be viewed as four steps that gradually lead to a deeper knowledge of the problem of interest.

The *descriptive* analysis answers the question: ”what happened”? It provides a representation of previous and current performances, through the extraction and processing of historical data. For this reason, this type of analysis provides a description of the events (”as is” analysis) but does not explain the causes that led to their occurrence. For this type of analysis, graphic visualization tools are commonly used to allow the creation of reports. These reports are intended to provide an immediate understanding about many characteristics such as the company’s production, financials, operations, sales, inventory and customers. Some examples of metrics that are commonly used in manufacturing can be the total stock in inventory, the plant OEE, the year change year in sales, and so on. In general, these tools allow users to extract data using queries executed on various data sources, to perform aggregations or other kinds of data processing, and to use a wide range of graphs in order to represent data in a dashboard.

The *diagnostic* analysis answers the question ”why did it happened”? The idea of this analysis is to extract more details with respect to the descriptive analysis by identifying patterns and relationships between the data. However, in general, this type of analysis is strictly correlated to the descriptive one. In many cases, we speak of a descriptive/diagnostic analysis with the aim of obtaining a cause-effect vision of the events. In fact, thanks to the analysis of historical data, it is possible to understand what happened and perform estimations about the current performances. Also in this case, it is common to perform a graphic visualization of the data. This visualization may offer the possibility to create interactive dashboards that show the correlations among the information through drill-down and linking functionality between the graphs. The descriptive and diagnostic analysis provide a deeper knowledge of past and current events but do not allow to understand future ones.

The step forward, which has driven the evolution of new techniques up to the point of transforming BI into BA, is due to the other two types of analysis: predictive and prescriptive analysis. *Predictive* analysis answers the question ”what will happen”? It enables to predict future performances based on historical data. An example of a predictive analysis in manufacturing, that will be described in details in Chapter 6, is to know in advance when it is necessary do maintenance to a machinery through the exploitation of its historical measures. More in details, with the predictive analysis, all the available data are analyzed in order to derive models (patterns) on such information that will be used to predict how they will evolve in the future. This procedure is possible thanks to the development of several techniques such as modelling, machine learning, data mining and game theory.

Finally, the *prescriptive* analysis answers the question ”what should I do”? It supports the user in making decisions based on a series of possible outputs (simulations) deriving from predictive analysis.

Prescriptive analysis, as well as predictive one, implicitly includes the concept of Big Data. In fact, for these type of analysis it is necessary to gather information from multiple sources, internal (inside the organization) and external (e.g. ERP, CRM). These data can be structured such as numerical and categorical, or unstructured like text, images, audio, and video data. All these data are processed with algorithms and tools such as business rules, machine learning and computational modelling procedures. In general, as it will be described in Section 2.4.2, the accuracy of the resulting output increases when more data are added to the model.

To summarize, both the predictive and prescriptive analysis have the objective of optimizing the forecast of future events. On one side, the predictive analysis focuses on model future events.

On the other hand, the prescriptive analysis provides the possibility to simulate the possible future outcomes based on the choice of certain decisions and actions. In general, predictive and prescriptive analysis are applied in contexts in which several variables, constraints and information must be analysed. Traditional BI tools are well suited for the descriptive and diagnostics analysis ([WW07, Neg04, CDN11, CCS12]) but, in general, do not fit well for other kind of analysis. For this reason, in the last years many tools have been proposed to perform BA and in particular predictive and prescriptive analysis ([CV70, SSJ14, MSDS14, Hui15, CSM12, EL12, BBA13, LLS^{+}11, SN87, KRS02, TMDOL10, Lie13, SS05, SSM12, SMK14]). Such tools enable the exploitation of state-of-art algorithms and techniques to extract the required insights from the available data.

The integration of BI and BA tools in manufacturing systems is one of the focal point for improving their operations and performances. In fact, the use of these technologies facilitate manufacturing actors (from managers to line operators) in monitoring, orchestrate, planning and forecast all activities involved in production (and, in general, in all the supply chain). In particular, in "Manufacturing Intelligence" can be considered all those analyses that help users to understand the past and actual performances of their processes, identify possible cause-effect correlation and give suggestion to make decisions within a short time range. Usually, MI tools deal with visual information of data collected by different systems and aggregated by the relevant aspects chosen by the user who performs the analysis. For example, in a shop floor, the real time visualization of the measures recorded by a machinery can help line operators to detect problems in the production. Whereas, "Manufacturing Analytics" involves more sophisticated techniques which enable to obtain deeper knowledge about the observed information, forecast future events based on previous behaviours, find implicit relations among data (not only clear correlations), and so on. MA tools exploit Machine Learning, Statistics, and other advanced mathematical techniques that can process a big amount of data and environmental variables and can perform prediction on future data. For example, in a shop floor scenario, having a tool that predicts when it will be necessary do maintenance on the machinery, implies to avoid many issues (e.g. production scraps, downtime).

2.4.2 Big Data

Nowadays, it does not exist a unique definition about what is Big Data. However, it is commonly accepted that Big Data was born after the 2000 because of the large amount of information and dataset available. Since the beginning, this amount of information is constantly increasing thanks to many technological progress. To better understand this phenomenon, we can introduce the main characteristics that are commonly used in literature when describing Big Data ([CML14, CZ14]). Such characteristics are known as the three V: Volume, Variety, and Velocity (Figure 2.4).

First of all, the *Volume* indicates the exponential increasing of the available information. Such information may come from different sources such as Internet, PCs, smartphones, sensors, and all the electronic devices available.

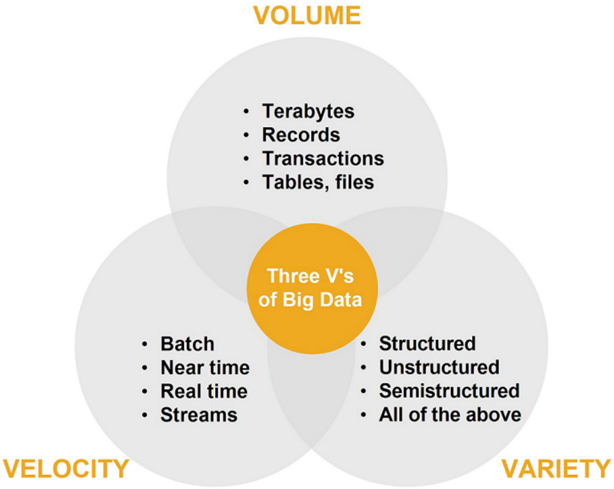


Figure 2.4: Big Data 3 Vs.

Instead, *Variety* means that the available data is heterogeneous and not organized. Many information may be incomplete and/or not correct and, although such data may bring useful information, it is required a long phase of compression and transformation in order to exploit useful and correct information. In addition, data can be generated from different sources and in different formats. For instance, it is possible that data contains movements, phone calls, photos, messages, and many more. Such heterogeneity make impossible an immediate use of such data and a pre-processing must be required in order to exploit all the different sources of data.
The last V is *Velocity*. This can be viewed under two different aspects. From one side, Velocity refers to the great speed in which the Big Data phenomenon was born. On the other side, it refers to the way the data are handled and become available to be used. For instance, in many cases the processing of the Big Data is required to be performed in real time.

There are also other ways to define Big Data. For instance, the possibility to retrieve important knowledge from the aggregated information, that can be applied in many different scenarios [R+11]. Or, the possibility of predicting new trends and indicators that can be retrieved only exploiting a large amount of data [SB12b].
There is a big difference between Big Data and the traditional approach to perform the identification of new trends. One of the major difference is the question to ask when starting an analysis: it is no more why, but what. Big Data enables to identify relationships and correlation between the data, in this way, it is possible to identify trends and facts without knowing the motivations and the structure of the data and why several events occur. However, such approach is enough to extract information from the data in order to make decisions and predictions about different phenomena.

The approach that characterize the Big Data processing is to apply techniques borrowed from Mathematics and Statistic in order to exploit the large amount of data to identify trends and perform predictions. Thanks to the Big Data revolutions we see a shift in how the processing of data is performed. Before, getting all the useful data was a difficult task and many times it was impossible to retrieve all the information needed to perform an exhaustive analysis. For this reasons, researchers and industries usually restrict the population under analysis in order to have a sufficiently representative sample of the data and then, extend the outcome. Now, the large amount of data allow to perform exhaustive searches in a way to generate predictions that can fit better to the population under exam. However, this process requires a different approach respect those traditional, and constitutes a turning point with respect to the statistical techniques are used [MSC13]. One of the problems of statistics, since the dawn of this discipline, has always been the access to data and to information that describe the population. Making use of all the available data could be too difficult, both in terms of time and costs. To this end, the sampling has been largely used, that is the selection, through appropriate criteria, of a part of the population in order to facilitate the analysis and subsequently extend the results to all the elements. The sampling method intrinsically gives rise to errors, which can be reduced by randomness of the sample, rather than by making use of more data [MSC13]. The technological improvement that has been found in recent years has made the management and processing of huge amounts of data much less difficult. It is now possible to overcome the limitations of the sampling method and to process all data. This enables to obtain a better analysis, without no approximations, able to fully capture the different aspects. In fact, some trends often remain hidden when the sampling is applied.

A consequence of the use of an increasing quantity of data is the need to handle the errors. It is very common that data contains errors and it is impossible to eliminate them altogether. Therefore, in making predictions about general trends, we are forced to accept inaccuracies in the data. In general, when processing Big Data, it is preferable the quantity with respect to data quality. This is motivated because usually the estimates obtained are more reliable if the amount is increased. This entails a further change of perspective because the absolute accuracy of the available data is no longer mandatory. The less precise data in Big Data, in general, allows, thanks to their quantity, to solve the quality problem in order to obtain precise estimates. A fundamental aspect to consider when handling estimates performed making use of Big Data, is the correlation. The correlation allows to establish if and how two statistical variables are connected to each other, based on the variations of one variable against the variations of the other. There is a lack of correlation if the change in a variable does not involve any change in the other considered variable. Big Data makes it possible to highlight all the correlation between the variables considered when making predictions. If the correlation is high, the probability that the variables are connected is high [MSC13] allowing to ”understand the present and foresee the future”. Thus, if we know that two phenomena are strongly correlated, it will be sufficient to monitor one of the variables to control the changes of the others. However, Correlations do not express causality. Correlations express what happens, but they do not explain why. This aspect is sufficient for the purposes of understanding trends and the prediction of phenomena, but a change in the statistical process used is required. The common approach to increase a knowledge is the detection of the cause-effect relationships and to ask ourselves why phenomena occur. Instead, Big Data can achieve efficient and effective forecasts, but it is not easily exploitable to understand causal relationships.

2.4.3 Descriptive and Diagnostic techniques in Manufacturing

In Section 2.4.1 we discussed the main four type of analysis. Now we will provide an overview about the techniques and tools involved for performing descriptive and diagnostic analyses. Gartner Group coined the term Business Intelligence with this definition: “Business Intelligence describes the enterprise’s ability to access and explore information, often contained in a Data Warehouse, and to analyze that information to develop insights and understanding, which leads to improved and informed decision making. BI tools includes: ad hoc query, report writing, Decision Support Systems (DDS), Executive Information Systems (EIS) and, often, techniques such as statistical Analysis and On line Analytical Processing (OLAP)” [Neg04]. By this definition it is possible to understand that the term BI includes two different aspects, business processes and technologies, that are necessarily related.

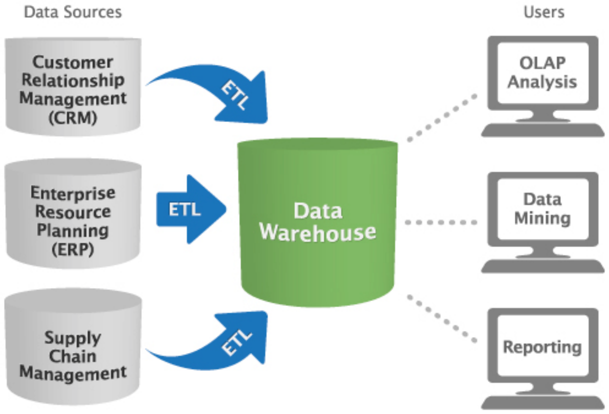


Figure 2.5: Business Intelligence processes and techniques.

In order to give an overview about what the BI tool involved in each step of BI processes, we will use the Figure 2.5.

Manufacturing organizations collect information to make assessments and estimates about their performances and use the gathered information through a BI system in order to increase their competitive advantage. The first step to perform in order to analyse any kind of problem, is to collect all those data involved in the scenario of interest, and then, execute several operations to prepare data for their last processing. These activities are essential to start any kind of analyses. If these steps are well done, it is possible to obtain right and really expendable information from data ([GJR11, Dav12]). Typically, data are extracted from various source systems to be centralized in a single system called Data Warehouse (DWH). The operation of extracting and loading data is called ETL which is the acronyms for Extract Transform Load. During this procedure, manufacturing data are taken from different sources like Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), and other systems involved for the Supply Chain Management. These data undergo several steps:

- **Cleaning**: this phase deals with improving the quality of data by eliminating ”dirty” data due to duplications, inconsistencies, missing data, incorrect values, and so on.
- **Transforming**: this is the central phase and it aims to convert data from the source operational format in order to map them into the DWH.
- **Loading**: in this phase, data are uploaded to the DW through two alternative modes: refresh (the data are rewritten replacing the previous ones completely) or update (only the changes to the data are added to the DW without overwriting all the data at each iteration).

At this step, data are collected in the DWH. The Data Warehouse is a tool that aims to store all the business interest data, in an optimized way for performing query. The DWH data structure is different respect to those of the operational databases: while these last are based on relational concepts and rules (Entity-Relationship), DWHs are generally based on the dimensional model or Star Schema, appropriate for quickly respond to queries of various kinds. The events that occur in a manufacturing system are too many to be individually analyzed and then it is usual to place them in an n-dimensional space. In this multidimensional space, the axes (called analysis dimensions) define different perspectives for the event identification. Just this dimension concept gave rise to the cube metaphor for the representation of multidimensional data. According to this metaphor, events correspond to a cube cells. Whereas, the cube edges represent the dimensions of analysis. Each cell in the cube contains a measure value (that is, the specific event occurred respect to a specific dimension). Therefore, the data modeling in a DWH, involves three key management concepts:

1. **fact**: the central analysis concept. It models an event that happens in the company;
2. **measure**: an atomic property of a fact that can be analyzed. It describes a quantitative aspect;
3. **dimension**: describes a perspective on which perform the analysis.

The two main approaches that can be performed by users for querying a DWH, are reporting and OLAP. OLAP stands for “On-Line Analytical Processing” and includes a set of software techniques for the interactive and fast analysis of large amounts of data. These techniques allow user to investigate data in a more complex way, and quickly provide answers to multidimensional analytical queries. Then, an OLAP system allows to study a large amount of data (the measures contained in each fact), analyze data from different perspectives (dimensions), and support decision-making processes. The OLAP analyses can be visualized in a report. The reporting systems exploits DWH data, OLAP aggregated data, and other sources data in order to provide an overall visual representation of performance measures (KPIs), trends and exceptions within an appropriate container called dashboard. [EALS07]. In dashboards, data can be shown respect to different and hierarchical perspectives of analysis so that, thanks to graphical drill-down capabilities, it is possible to investigate more detailed information. In Section 2.4.3.2 we will provide additional details about dashboards, being one of the main topics of this thesis.

Finally, the last step of a BI process is to perform Data Mining (DM) analysis. DM procedures aim to identify the presence of specific patterns (patterns) within large volumes of data through the application of algorithms. The term DM refers to the application of one or more techniques that allow the exploration of large amounts of data, with the aim of identifying the most significant information and making it available and directly usable in decision making. The knowledge extraction (meaningful information) takes place through the identification of associations, or ”patterns”, or repeated sequences, or hidden regularities in the data. In this context, a pattern indicates a structure, a model, or more generally, a synthetic representation of the data. Typical DM algorithms are: Anomaly Detection, Association Analysis, Clustering, Classification, Regression and Summarization.

In conclusion, BI systems allow user to set goals and compare data relevant to the achievement of their objectives and, sometimes, to make interventions in view of the information obtained so to improve future performance.

2.4.3.1 Key Performance Indicators

Key Performance Indicators (KPIs) are a very useful approach that enables to analyse the performance of manufacturing processes and to manage the Supply Chain operations ([CLXL09, RZJ04, MPMDM10, AD02]). In order to control the large and complex variety of industrial systems, it is necessary identify a set of metrics related to the main manufacturer’s activities and targets, at each level of the enterprise - from the macro financial level to the operation level. KPIs enable to ”see” the results in an aggregated way, rather than using large amounts of ”raw” data. In a few indicators easy to read, there are information able to provide an immediate diagnosis about the phenomenon under exam: users have the continuous performance control and, when the indicator is not aligned with its benchmark, they can focus on improvement actions in well-defined directions. By measuring and managing performance, organizations have more opportunities to successfully achieve operational and financial goals.

In order to achieve these goals, it is essential to identify appropriate KPIs which provide a valid support to obtain some key advantages:

- Analysis of ongoing processes;
- Evaluation of the overall company performances (not exclusively economic-financial), thanks to a direct feedback provided by those company's users who have the ability to manage critical variables;
- Company performance evaluation as a trend, thanks to the identification of flows and signals over time: the evaluation starts from the results achieved, focuses on the management operations that will be projected, in turn, into the future in order to identify those prerequisites for maintaining and improving performance;
- Acquisition of essential information in order to manage the planning and scheduling of company activities, by setting up preventive and corrective actions.

It is also possible to develop distinct KPIs representation as analysis' outputs: company dashboards, reports and graphs related to such indicators, more or less detailed, more or less sophisticated, deep in the analysis or synthetic in the exposition ([YO12, HVNK13, GHH$^+$13]). They can be weekly, or even daily, instruments for analyzing the bias between the expected (target) values of a specific indicator and its real value. Depending on the level of detail and/or aggregation, this information allows the various managerial levels to carry out their analyzes and to identify their objectives for the continuous improvement.
Since the KPIs enable to monitor the company's activities related to different specific levels, they must be balanced on different hierarchical levels of analysis:

1. Summary indicators at strategic level;
2. Indicators related to the medium-term tactical analysis and planning;
3. Specific and disaggregated indicators for the assessment of the operational performance.

In order to exploit KPIs capabilities, they must be:

- Simple and inexpensive to be detected, processed and interpreted;
- Easy to measure and, if it is possible, immediate to understand and unbiased (e.g. quantity, percentage, ratio, and so on);
- Relevant and able to respond to specific objectives;
- Comparable through defined standards (with an appropriate target value and tolerance/deviation);
- Accessible by those who must perform analyzes on them;

- Processable with mathematical or statistical tools and reproducible on tables, graphs or diagrams in a clear and immediate understanding;
- Shareable;
- Systematic, that is, punctually recorded with established regularity and immediately updated in case of uncommon/unexpected events.

It is better to choose a few key measures that has these features rather than using a complex data system whose control costs more than the benefits that can be derived from it. Furthermore, it would be useful to have a reference table containing the following information related to each indicator:

- Indicator Description;
- Reference process;
- Detection method used;
- Data source;
- Calculation method;
- Unit of measurement;
- Frequency;
- Responsibility for the management of the indicator.

This reference scheme for the KPIs identification has been defined within the ISO 22400 "Automation systems and integration — Key performance indicators (KPIs) for manufacturing operations management" standard[2].
In fact, given the importance of KPIs in the monitoring of manufacturing processes, the research of the "right" KPIs has gained importance both in academic and industrial fields. This research led to the emergence of several organizations like ISO 22400 and SCOR, which established a set of general evaluation systems to express the objectives and critical factors, and proposed a common structure for standardizing KPI definition. The International Organization for Standardization (ISO) published, in 2014, the first two parts (Part 1: Overview, concepts and terminology and Part 2: Definitions and descriptions) of the ISO 22400 standard, marking a milestone in the field of operations management. This standard aims at defining a common scheme for Manufacturing Operations Management (MOM) KPIs across industries. The innovative aspect of this

[2] www.iso.org/standard/56847.html

standard is to have rationalized the use of indicators related to manufacturing operations, already known and used in industrial processes, and to have classified them according to a standard hierarchy: company, site, process, and even users (operators, supervisor, management). Using as reference model the IEC 62264 standard, in which it is defined the functional hierarchy model of manufacturing enterprise (Figure 2.6), ISO 22400 specifies the KPIs residing at level 3 related to Manufacturing Operations Management.

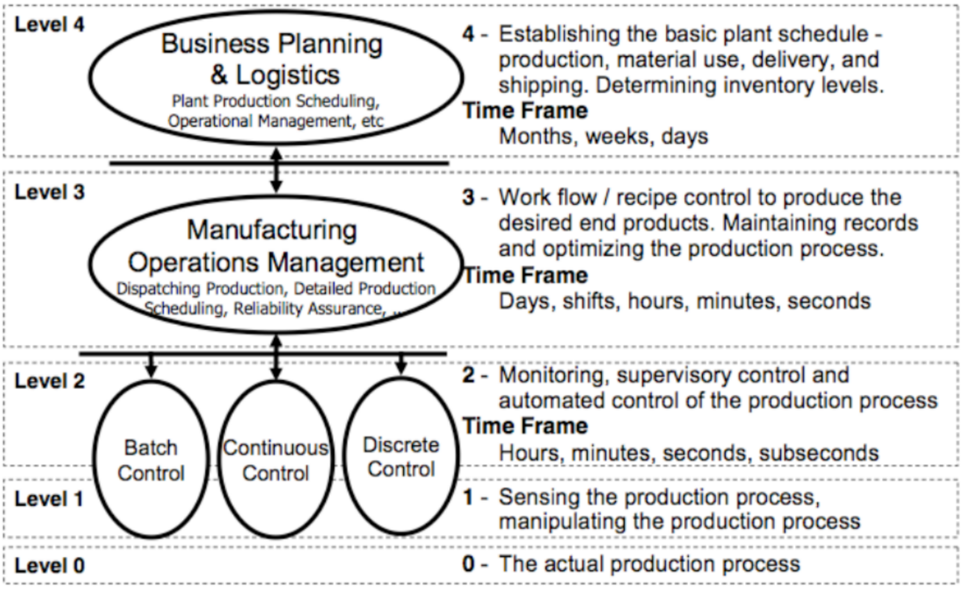

Level 4
Business Planning & Logistics
Plant Production Scheduling, Operational Management, etc
4 - Establishing the basic plant schedule - production, material use, delivery, and shipping. Determining inventory levels.
Time Frame
Months, weeks, days
Level 3
Manufacturing Operations Management
Dispatching Production, Detailed Production Scheduling, Reliability Assurance, ...
3 - Work flow / recipe control to produce the desired end products. Maintaining records and optimizing the production process.
Time Frame
Days, shifts, hours, minutes, seconds
Level 2
Batch Control
Continuous Control
Discrete Control
2 - Monitoring, supervisory control and automated control of the production process
Time Frame
Hours, minutes, seconds, subseconds
Level 1
1 - Sensing the production process, manipulating the production process
Level 0
0 - The actual production process

Figure 2.6: Functional Hierarchy in IEC 62264.
Source: IEC 62264-3

IEC 62264 standard identifies these four hierarchical levels of the plant, since they provide different functions and work in different time frames. Moreover, IEC 62264 standard also defines a hierarchical structure for the physical equipment (Figure 2.7).

Thanks to these two reference models, it is possible to "read" the ISO 22400 schema for its KPIs definition: each KPI is related to the Level 3 of the IEC functional hierarchy of manufacturing enterprise and it is calculated respect to a specific level of the IEC hierarchical structure for the physical equipment. The main categories evaluated by ISO 22400 indicators include:

- production operations;
- inventory handling operations;
- quality assurance testing operations;
- maintenance operations.

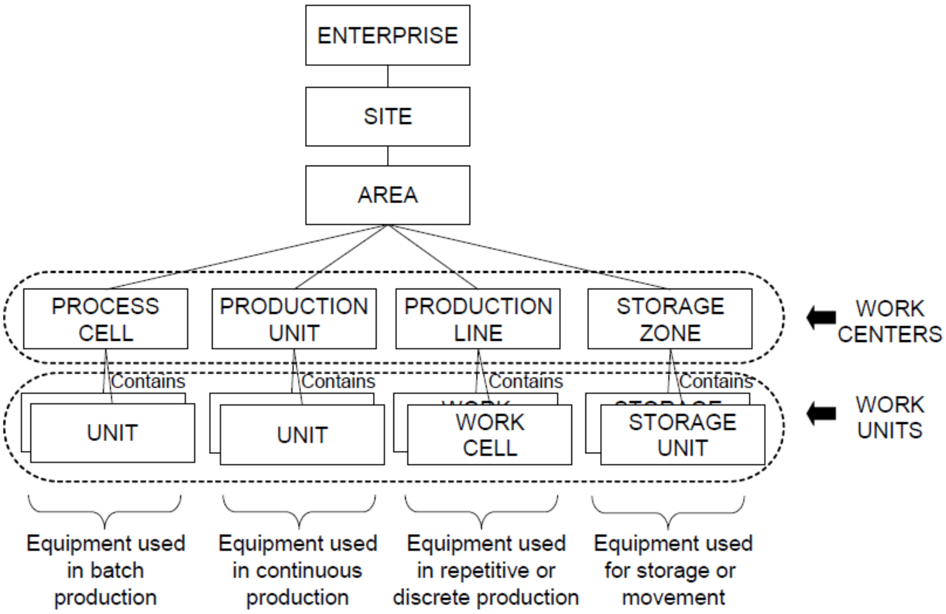

ENTERPRISE
SITE
AREA
PROCESS CELL
PRODUCTION UNIT
PRODUCTION LINE
STORAGE ZONE
WORK CENTERS
Contains
Contains
Contains
Contains
UNIT
UNIT
WORK CELL
STORAGE UNIT
WORK UNITS
Equipment used in batch production
Equipment used in continuous production
Equipment used in repetitive or discrete production
Equipment used for storage or movement

Figure 2.7: Role based equipment hierarchy in IEC 62264.
Source: IEC 62264-3

Each category can be further detailed using eight sub-categories: i) detailed scheduling, ii) dispatching, iii) execution management, iv) resource management, v) definition management, vi) tracking, vii) data collection, and viii) analysis. The first part of this standard describes the concepts and terminology used for KPIs, summarizing them a tabular structure as depicted in Table 2.1.

The second part analyses 34 KPIs related to MOM, targeting both discrete, batch and continuous production.

Another important reference model was developed by the Supply Chain Council (SCC) and Advanced Manufacturing Research (AMR) in 2000[3]. This model aims to simplify the systematic management of the supply chain performance measurement and improvement. SCOR is a cross-industry framework applied to evaluate all existing interactions within the supply chain, like:

- All customer interactions, from order entry through paid invoice;
- All product (physical material and service) transactions, from your supplier's supplier to the customer's customer, including equipment, supplies, stoks, products, software, and so on;

[3] www.apics.org/apics-for-business/products-and-services/apics-scc-frameworks/scor

Table 2.1: Tabular structure of a KPI in the ISO 22400 standard. Source: ISO 22400

Content:	
Name	Name of the KPI
ID	A user defined unique indication of the KPI in the user environment
Description	A brief description of the KPI
Scope	Identification of the element that the KPI is relevant for, which can be a work unit, work center or production order, product or personnel
Formula	The mathematical formula of the KPI specified in terms of elements
Unit of measure	The basic unit or dimension in which the KPI is expressed
Range	Specify the upper and lower logical limits of the KPI
Trend	Is the information about the improvement directions, higher is better or lower is better
Context:	
Timing	A KPI can be calculated either in: real-time, after each new data acquisition event; on-demand, after a specific data selection request; periodically, done at certain interval, e.g. once per day
Audience	Audience the user group typically using this KPI. The user groups used in this part of ISO 22400 are: Operators, personnel responsible for the direct operation of the equipment; Supervisors, personnel responsible for directing the activities of the operators; Management, personnel responsible for the overall execution of production
Production methodology	Specify the production methodology that the KPI is generally applicable for: Discrete Batch Continous
Effect model diagram	The effect model diagram is a graphical representation of the dependencies of the KPI elements that can be used to drill down and understand the source of the element value. NOTE: This is a quick analysis which supports rapid efficiency improvement by corrective actions, and thus reduces errors
Notes	Can contain additional information related to the KPI. Typical examples are: Constraints, Usage, etc.

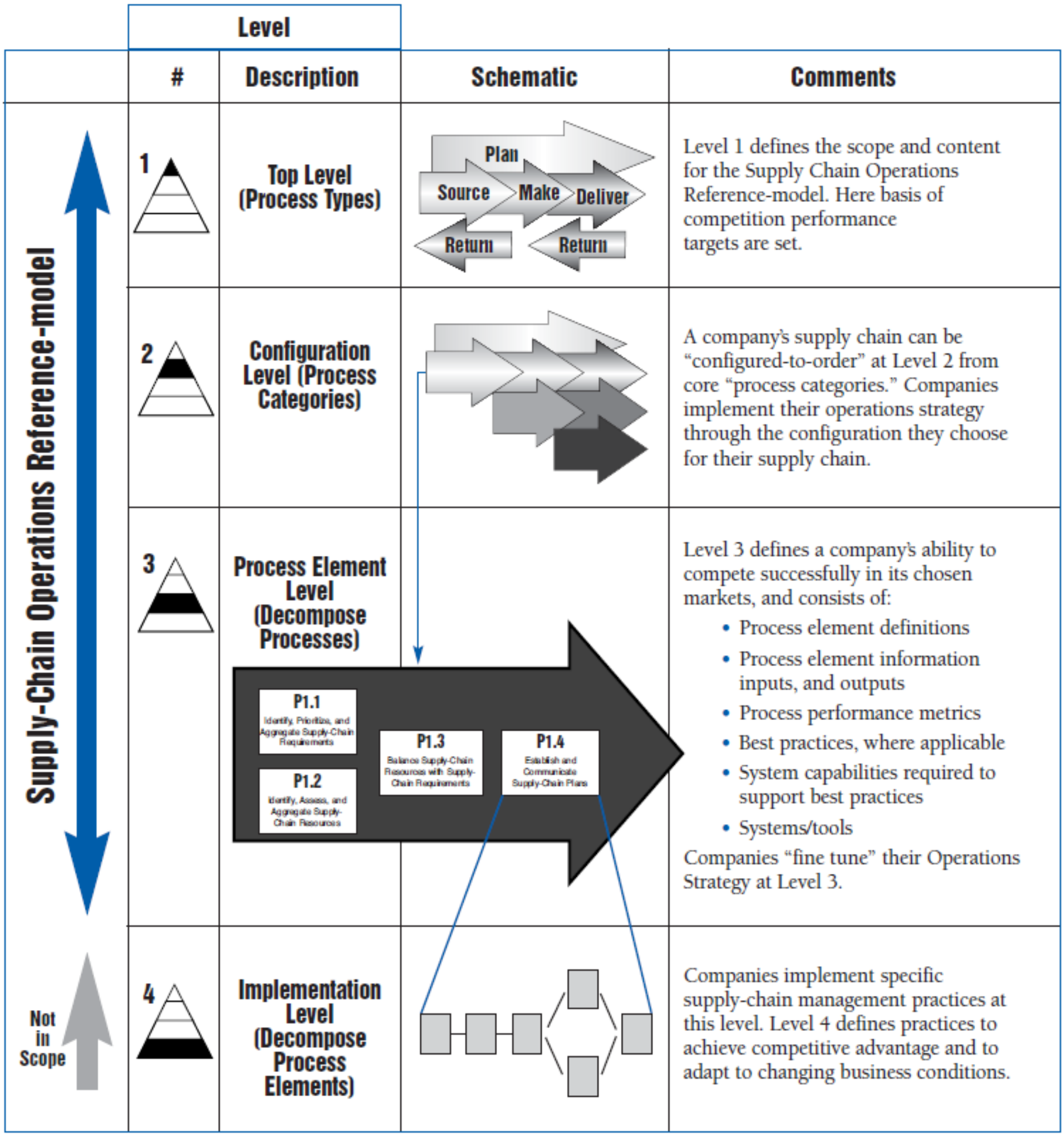
- All market interactions, from the aggregate demand to the fulfilment of each order.

This model identifies four analytical stages:

- Performance: Standard metrics to describe process performance and define strategic goals
- Processes: Standard descriptions of management processes and process relationships
- Practices: Management practices that produce significant better process performance
- People: Standard definitions for the skills required to perform supply chain processes.

organized around the five primary management processes:

- **Plan**: Includes all processes that balance aggregate demand and supply to develop a set of actions that best meets sourcing, production and delivery requirements
- **Source**: Includes all processes that provide goods and services for meeting planned or current demand.



	Level			
	#	Description	Schematic	Comments
Supply-Chain Operations Reference-model	1	Top Level (Process Types)	Plan Source Make Deliver Return Return	Level 1 defines the scope and content for the Supply Chain Operations Reference-model. Here basis of competition performance targets are set.
	2	Configuration Level (Process Categories)		A company's supply chain can be "configured-to-order" at Level 2 from core "process categories." Companies implement their operations strategy through the configuration they choose for their supply chain.
	3	Process Element Level (Decompose Processes)	P1.1 Identify, Prioritize, and Aggregate Supply-Chain Requirements P1.2 Identify, Assess, and Aggregate Supply-Chain Resources P1.3 Balance Supply-Chain Resources with Supply-Chain Requirements P1.4 Establish and Communicate Supply-Chain Plans	Level 3 defines a company's ability to compete successfully in its chosen markets, and consists of: • Process element definitions • Process element information inputs, and outputs • Process performance metrics • Best practices, where applicable • System capabilities required to support best practices • Systems/tools Companies "fine tune" their Operations Strategy at Level 3.
Not in Scope	4	Implementation Level (Decompose Process Elements)		Companies implement specific supply-chain management practices at this level. Level 4 defines practices to achieve competitive advantage and to adapt to changing business conditions.

Figure 2.8: Supply Chain operations reference model.
Source: SCOR

- **Make**: Includes all processes that transform a product to a finished state for meeting planned or current demand.
- **Deliver**: Includes all processes that provide finished goods and services for meeting planned

or current demand, typically including order management, transportation management, and distribution management.

- **Return**: Includes all processes associated with returning or receiving products. These processes cover the post-delivery customer support.

SCOR model defines more than 150 KPIs to support supply chain analysis at three different levels of process detail, as depicted in Figure 2.8.
The first level metrics help companies to gauge their performances respect to their targets within the competitive market space.
In the second level, 26 core process categories of a supply chain are defined. These processes can be seen by companies in order to identify their ideal or current operations.
The third level provides the information required for successfully planning and setting goals for supply-chain improvements.
Finally, the last level, that is not part of SCOR topics, focuses on the application of specific supply-chain improvements.
The main goal of this model is to help companies to evaluate their *as-is* (current state) state of processes in order to achieve their target *to-be* (future state) state. In order to achieve this goal, SCOR provides various different concepts: it defines the standard management processes, analysing the relationship among them; it provides a large number of standard metrics in order to measure process performances and, finally, it presents over 430 best practices for assisting companies to choose the activities to be performed in order to close the gaps among their current and target state.

2.4.3.2 Advance Human Machine Interface: the importance of dashboards

In Section 2.4.3.1 we discussed the importance of controlling those KPI that describe process performance and define strategic goals.
After collecting the data for calculating the chosen indicators, it could be appropriate to summarize these information in a periodic report. An effective way to transfer the information that emerges in the reporting system, can be the creation of a graphical dashboard, or using a synonym, a management cockpit.
The approach by which strategic information are shown ”visually” through graphical dashboards, is called “Visual Analytics” ([TC06, KMS$^+$08, WRS13]).
A dashboard is a single-screen display that shows important information about a company so that the whole situation, for example, in a factory or on a production line, can be quickly understood.
In the broader sense, digital dashboards are intuitive and easy-to-use front ends for monitoring, analysing and optimizing critical business activities by enabling users on all hierarchy levels to improve their decisions. Thanks to the exploitation of dashboard tools such as charts and graphs, the information can readily be understood, leading to a series of advantages such as:

- be aware of the current operation performances in order to understand and make strategic decisions in the future;
- evaluate the performances state respect to the prefixed goals or appropriate reference benchmark;
- take more tactical and strategic decisions, not only based on intuition but also thanks to an updated and accurate knowledge of data;
- help the process to a continuous improving: the detection of deviations, especially the negative ones, and the identification of their reasons should help business executives to understand the mistakes and to propose corrective actions.

In other words, a dashboard is a management tool with the aim to provide a sort of "indicator light" about the current state of a company: as well as a car's dashboard highlights the anomalies to driver, also this reporting tool helps company's users to highlight if their company is moving in the right direction over time, according to predefined plans. Dashboard allows to quickly know the company's state through graphs that summarize the key indicators and key success factors. In brief, the two main objectives of the dashboard can be:

- the key variables (KPI) and core processes performance monitoring to achieve business success,
- provide a synthetic and complete interpretation about the bias among company's results and expected outcomes in order to define corrective actions.

The added value of this tool lies in its ability to exploit existing business information assets with costs close to zero, by recovering and importing data recorded in company management systems and by displaying them in a clear and synthetic manner with various kind of graphs and navigable tables.
Until now it has been discussed the usefulness and advantages in the use of dashboards, but what was their evolution? and why they are taking so much interest within the manufacturing environment? Groger et. Al [GHH$^+$13] state that traditional approaches for information monitoring on the shop floor are not able to provide an overview across the entire manufacturing process. This is due to the fact that each tool records the information abstracting it from its operating context and without correlating it among the different hierarchical levels of the production process. Consider, for example, the several information flows that are generated within a shop floor when a downtime occurs: the operators who work in the previous and next steps of the production line (respect to the step where the downtime occurred), must be informed. When this scenario occurred, the production must be directed to another available machinery, and it is necessary to temporarily change the line scheduling. At the same time, the operators who are working on

the faulty machinery, must understand, as soon as possible, the failure reasons so as to open the maintenance request by defining a priority level. Furthermore, they must communicate to the other operators in order to inform them of the downtime. All this information flow among the different steps of the production phases must occur in real time and quickly, so as to speed up the decisions that must to be done for reducing waiting times, costs and waste of resources.

The need to control the process' performance - starting from an overview of a single or multiple plants behavior (managers), up to the real time monitoring of the shop floor (line supervisors, operators) - combined with the technological evolution (computers, tablets and smartphones) has led to the development and to an increasingly use of the Human Machine Interface (HMI).
The HMI is the interface between the process and the operators, in other words, an operator's dashboard (Figure 2.9).

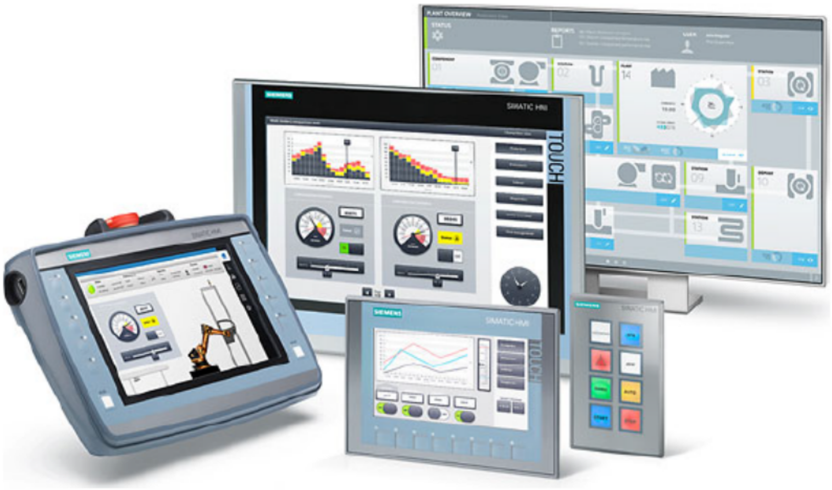


Figure 2.9: Example of Human Machine Interface.

This tool enables to display real-time operational data, helping operators and line supervisors to monitor and synchronize manufacturing processes and enabling them to obtain usable information thanks to the processing of complex variables.
In industry 4.0, designing interfaces means going beyond the interface itself. The single machine's HMI is the tile of a larger mosaic: it is necessary to consider a complex and distributed system made up of multiple actors (the operator, the technician, the production manager, etc.); multiple devices (panel for the use of the machinery, tablet for remote control, big screen for the control room); and other machines with which the one in question must interact. The role of the designer changes as well: no longer the design of a finished product, but the design of a system of components and of a set of rules that compose it.
HMIs are part of a BI tools that will be described in the next Section.

2.4.3.3 Descriptive and Diagnostic tools

Every year, an interesting overview of the trendiest BI tools in the business market, is proposed by Gartner Inc.[4], a world leader in strategic consulting, research and analysis in the field of Information Technology. The Gartner Magic Quadrant (MQ) is a qualitative research methodology that analyse and evaluate the capacities and improvements on several specific technology-based market. The MQ is a grid made up of four fields and two parameters: on the x-axis is measured the ”Completeness of vision”, whereas on the y-axis the ”Ability to execute”. Figure 2.10 depicts the MQ.

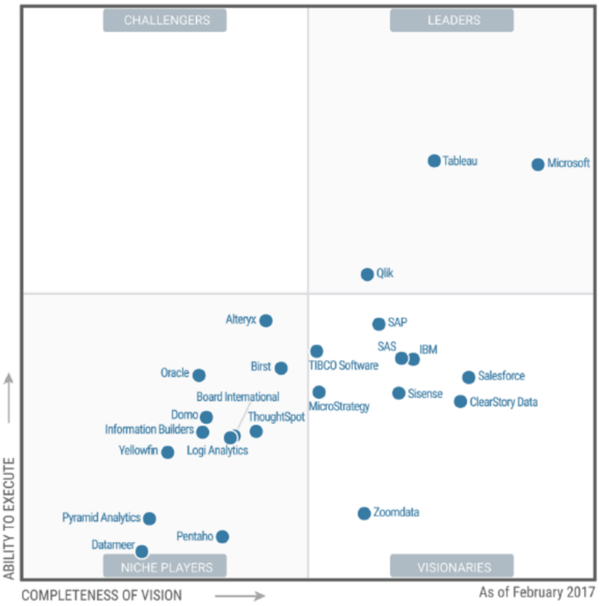


Figure 2.10: Magic Quadrant for Business Intelligence and Analytics Platform. Source:www.gartner.com

Within the matrix, four large areas are identified:

- **Leaders**. They are the players with the highest performances and completeness of vision. Usually these are very large companies that operate in mature markets where they stand out as the leaders of the sector. Generally they have a very large customer base, developed over time and thanks to the ability and continuous investments in increasing visibility on the market.
- **Niche actors**. They are in the opposite box to the leaders. These are companies that focus on a very specific market or on a very vertical market segments. Unlike leaders,

[4] www.gartner.com

the completeness of vision and the ability to execute of these players, is very limited. These are smaller companies, but with extremely specific capabilities and limited to an area of expertise. Sometimes large companies are positioned in this quadrant, due to their development difficulties or to a little ”vision” that limit themselves to decisively influence the market.

- **Challengers**. They are companies that have great capabilities, attractive offers on the market, and resources for a continuous growth, but with a more limited future vision. This is the reason why, they are not yet able to put themselves at the top of the market.

- **Visionaries**. They are those players who develop and provide products/services with a very advanced and wide vision, able to respond to important problems, even on a large scale. They are companies, often very innovative, able to understand and predict market developments, but lacks in the ability to execute. In fact, these are often smaller companies than the leaders.

The “Magic Quadrant for the Business Intelligence and Analytics platforms” structure was redesigned by Gartner in 2016 in order to reflect the last ten years deep changes in the IT sector, with the advent of new visual analysis and exploration tools (now become mainstream), interfaces easier to use even by non-IT users, and a general orientation to agile and business-centred solutions. The rules that establish the inclusion and position of IT vendor on the MQ, are based on five services (“use case”):

- **Agile Centralized BI Provisioning**: it identifies the possibility to centralize all data from other systems, inside the enterprise itself.

- **Decentralized Analytics**: it expresses the possibility to exploit BI tools also for non-IT users in order to make users autonomous in generating their analysis.

- **Governed Data Discovery**: it certifies the contents generated by users through a System of Record managed by IT

- **Embedded BI**: it indicates the possibility of helping users during their analysis, this feature is increasingly popular in business processes or applications

- **Extranet Deployment**: the same purpose of the first case, but outside the enterprise

mapped on 15 functions (“capability”):

- Infrastructure:

  - administrative, security, auditing and other functions;
  - cloud deployment and delivery capabilities;
  - connection and acquisition to data sources.

- Data Management:
  - metadata management;
  - capacity of embedded ETL and data storage;
  - data selection and preparation in a graphical (drag & drop) and self-service manner.

- Analysis and Content Creation
  - advanced and integrated analytical functions;
  - interactive dashboards;
  - interactive visual exploration;
  - 'smart' data discovery, without predefined models or algorithms and, possibly, with natural language;
  - use of mobile devices in publishing or interactive mode;

- Sharing of Findings
  - support for the inclusion of analytical content in corporate business processes;
  - make the analyzes known by various methods of distribution;

- Overall platform capabilities
  - level of 'seamless' use of the various capacities of the product (s);
  - ease of management and deployment.

In Gartner "Magic Quadrant for Business Intelligence and Analytics Platforms" 2017, the three vendors positioned at the leader area are: Microsoft with its Power BI suite, Qlik with its flagship product Qlik Sense and, finally, Tableau with its desktop, server and online solutions. All these three leaders have similar strengths:

- they provide an interactive and intuitive visualization, exploration and analysis of data, with in-memory analysis or with direct query on large data sets;
- they allow users (thanks to forums and tutorials) to create analytical contents autonomously, without the support of an IT-expert;
- they enable to perform complex analysis, thanks to the ability to manage heterogeneous data from various sources (from relational and semi-structured databases, Hadoop-based, and so on) both internal and external the company.

2.4.4 Predictive and Prescriptive techniques in Manufacturing

In Section 2.4.1 we discussed the main four type of analysis. Now we will provide an overview about the techniques and tools involved for performing predictive/prescriptive analyses.
In the traditional analysis techniques, the operations that can be performed on the data are, in a sense, pre-set: even if the data processing is automated, the inference applied to choose and relate the observed data depends on the user knowledge. The tool is limited to perform the operations that are decided by the user without going deeper on the ”methodological” accuracy of its processing. In contrast, machine learning techniques (called *algorithms*) try to construct rules to describe the data provided, and autonomously understand whether or not a new case responds to the rule they have defined. The difference between traditional programming and machine learning is therefore clear.
In the manufacturing context, characterized by a high number of information (big data), dimensions of analysis (variables), business rules, and constraints, ML techniques are playing a key role in the management and optimization of production processes. In fact, these techniques allow to analyze large, high-dimensional and multi-variate datasets.
Obviously, the output effectiveness obtained by applying a ML algorithm, always depends on the initial preparation of the data. Section 2.4.3 describes all the phases that start with the extraction of information from different data sources until their final preparation to carry out the desired analyzes. Frequently, in addition to the variable of interest, it may also be useful to consider other information involved within the context in exam. In fact, even if there are no evident correlations between the event of interest and other exogenous events, ML algorithms can find patterns and relationships between them, thanks to their nature of ”learning” from the data.

An example of this approach will be discussed in Chapter 6, concerning the creation of a model for predictive maintenance on a production line. The ability to know in advance, when it will be necessary do maintenance on a machinery so as to avoid the production of waste or downtime, is the goal we want to obtain. The KPI that need to be considered, is a distinct measure of the production items, which is detected on a specific machinery of the production line in exam. Since it is not possible to know a priori whether the value of this KPI can be influenced by other additional measures that are previously carried out during the processing, we have also considered all the other variables recorded on the production line to create our predictive model.
In conclusion, the ML algorithms are well suited to solve complex problems, such as those in the Smart Manufacturing field, where it is necessary to process a large amount of data and many variables and where, often, there aren’t mathematical equations to measure the physical models underlying the observed events ([SSA09, CSZ10, AGOR14]). In fact, ML techniques do not require in-depth knowledge of on the analyzed system by the model’s developers. These methods use computational procedures to ”learn” information directly from the data without relying on the knowledge of the physical models that underlie the phenomenon in question. Furthermore , they can adaptively improve their performance as the number of samples available for learning increases. The integration of these algorithms into increasingly user-friendly and

low-cost technologies has further accelerated the use of these techniques. However, the main problem is still that, very often, these ML techniques are applied in specific contexts rather than on the whole system [DFL12].

2.4.4.1 Machine Learning

There are many types of machine learning algorithms that can be classified into various categories depending on their use.
The main distinction is made between supervised and unsupervised algorithms, as depicted in Figure 2.11.

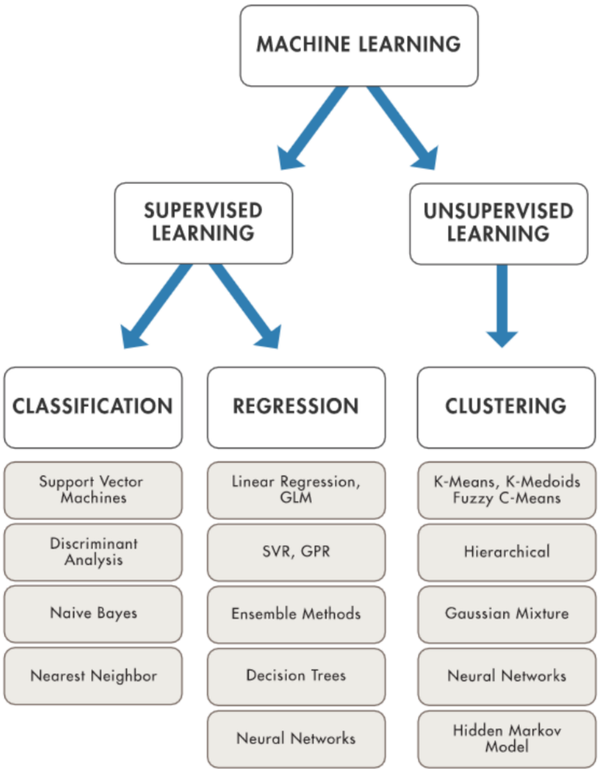


Figure 2.11: Supervised and Unsupervised Machine Learning techniques.
Source: `www.mathworks.com`

Supervised learning is the ML methodology of learning a model from labeled training data. The training data is a pair *(X,Y)*, where *X* is an input object (typically a vector) and *Y* is the label. A supervised learning algorithm learns the patterns from the training data and generate a model able to map new unknown examples. Supervised learning techniques are based on classification and regression.

Unlike supervised learning, the *Unsupervised* learning does not use previously classified and labeled data; therefore, it is not possible to know the class of data. The algorithm task is to extract a rule that groups the observed data according to their attributes. The algorithm, in this case, look for a relationship between the data in order to understand if and how they are connected to each other. Since there isn't a pre-set information, the algorithm is called to create ”new knowledge” (knowledge discovery). Unsupervised learning techniques are mostly based on clustering or association rules.

In order to start in using ML algorithms, it is necessary to identify the type of analysis you want to perform into the right category: classification, regression, clustering, or pattern recognition.

Classification: The task of developing a model, from a set of previously classified examples, that can correctly categorize new examples from the same population. Given a collection of records (training set), each record has a set of attributes, some of which express the class of the record. The aim is to assign new records to one (or more) classes as accurately as possible, identifying a model for the class attributes.
A test set is used to evaluate the accuracy of the model. The test and training sets come from the available data, which are divided into these two groups. An example of classification can be the reduction of the marketing cost by the identification of a small group of consumers who are more likely to buy a new model of cell phone.
Approach: use the data for a similar product introduced earlier and identify who bought the product or not. This attribute (Yes / No), is the attribute of the class. It is therefore necessary to gather information about these customers and use these data as input attributes to ”train” the classifier.
If the class has only two attributes, it is called *binary classification*. Otherwise, if the available class outputs are greater than two, the classification is called *multi-class*.
Among the most common classification algorithms it is possible to remark: Support Vector Machine, Decision Trees (DT) (also considering their combination in the Ensemble methods), K-Nearest Neighbour, Naïve Bayes, Discriminant Analysis, Logistic Regression and Neural Networks (NN).

Regression: The task of predicting the value of a continuous variable, from other variables' values, assuming the existence of a dependency model (linear or non-linear). Regression examples include: prediction of a new product's sales based on advertising costs; prediction of wind speed as a function of temperature, humidity, pressure, and other parameters.
Among the most common regression algorithms there are: Linear Regression (and Generalized Linear Model (GLM)), SVM (and Gaussian Process Regression (GPR)), Decision Trees (also considering their combination in the Ensemble methods) and Neural Networks.

Clustering: Clustering techniques are based on measures related to the similarity between the elements. Given a set of data, each described by a set of attributes, and given a measure of similarity, identify clusters (groups) such that:

- the data, in a cluster, are more similar to each other,

- the data between two different clusters are less similar.

An example of clustering can be the activity of market segmentation: gather customer information and then identify groups (clusters) of similar customers. Among the most common clustering algorithms there are: K-means and K-medoids, hierarchical clustering, Gaussian Mixture models, Hidden Markov Models and Clustering Fuzzy C-means.

Pattern Recognition: The pattern recognition aims to learn a classifier of data (*pattern*) based on a priori knowledge, or statistical information extracted from the patterns. The patterns to classify are typically groups of measures or observations, which define the points in an appropriate multidimensional space. For example: process the data collected at cash desk through barcode scanners in order to identify products that are purchased together by a significant number of customers.

As previously stated, there are many ML algorithms and they can be applied on different scenarios on the base of their characteristics. Therefore, the first step is to choose which one best fits (i.e. that provides more accurate analysis) to the problem you want to investigate.
In the manufacturing sector, some aspects in which ML is mostly applied are, for example, the prediction of maintenance activities on machineries, fault diagnosis, quality estimation of products, and so on.
A very promising and fitting supervised ML algorithm for manufacturing research problem is **Statistical Learning Theory** [WWIT16]. SLT try to find necessary and sufficient conditions for non-parametric inference to build predictive models from data or, using the language of SLT, learn an optimal model from data [OOA15]. In other words, SLT approach is to identify that function which better estimates the output for previously unseen inputs [EPP00].
This is a huge advantage for the manufacturing environment, where often, only few labelled samples are available, while the largest part of the dataset is composed of unlabelled data [CSZ10, AGOR14]. When the collection of labels is expensive or difficult, this approach allows exploiting all the available information for building effective models. Other features of these methodologies, useful for manufacturing applications, are: their strong adaptability to different scenarios of analysis and the ability to process large amounts of data.
Neural Network, Bayesan modelling, and Support Vector Machines are some of the most popular algorithms that exploit the theory of SLT [BB05].

Neural Network algorithms try to mimic the human brain behavior. A neural network is made up of a set of processing nodes (also called *units*) connected together through one or more *output arcs* and one or more *input arcs*. Each arc is defined by a weight that must be accurately estimated in order to reduce the error between the estimated value and the real one. These algorithms are able to process high-dimensional and multi-variate data on a similar rate to the later introduced SVM [KZP07]. NN can be basically applied to all problems of supervised and unsupervised learning, proving high performance in terms of accuracy and generalization.
NNs are exploited in various applications of semiconductor manufacturing and in various problems related to process control ([HSK^{+}06, LH09a, WCL05]) like: prediction of products or

materials properties on the basis of the parameters of the technological process involved; equipment failures prediction on the basis of selected signals; identification of the reasons that cause faulty products.
Despite the high success of NN, these techniques have some limitations such as overfitting, complex and time-consuming training process and ambiguity of the built model, since it is not possible, in most neural networks architectures, to directly correlate the weights of the connections with the input's characteristics ([KZP07, PA05]).

Bayesian Networks (BN): provide a model for probability distributions, expressing conditional dependency relationships between the variables involved. In other words, BNs learn the full joint probability distributions of the attributes and class. The learning of causal relationships is important because it increases the degree of understanding of the problem's domain and it allows to make predictions about future interventions. The prior knowledge of the domain, is a Bayesian Networks feature that is important especially when the data are scarce or expensive. Other strengths of these techniques are: an efficient approach to avoid data overfitting and the ability to manage missing and hidden values without compromising the result performance. Bayesian networks can be used in any problem where it is necessary to model reality in situations of uncertainty, where probabilities are involved. Some examples of NBs application in manufacturing are: final product quality estimation [ACS$^+$11], failure prediction [HE$^+$01], identification of the defect causes in products, and damage detection in engineering materials [ASMO07]. The main manufacturing application areas of BN are: the semi-conductor industry ([YL12, NJL$^+$11]), the automobile industry [LJ13] and in machining [CBPT09].

Support Vector Machines: another technique considered one of the most effective [FDCBA14] and popular classification and regression algorithm [CV95, AGOR12c] is the SVM. Like the NN, this algorithm can analyze high-dimensional and multi-variate datasets and can provide very accurate predictions. The SVM objective is to minimize the risk associated with the classification error, by identifying the best separation limits between a series of data. These separation limits are represented through feature vectors, which are weighed by the algorithm during the training phase in order to ensure that the margins that separate different classes are maximized in the data space. In other words, given a set of training examples, each marked as belonging to one or another class (in a binary classification problem), the SVM task is to identify those hyperplane that separates the data with the maximum margin. If the data classes are not linearly separable, the algorithm nevertheless chooses the hyperplane that separate the examples as clearly as possible, ever maximizing the margin between the closest examples [NNB15]. The problem of classification therefore becomes a problem of optimization of the margin value.
The main SVM features are: excellent ability in modeling linear and non linear relationships, it works very well with high-dimensional data, the ability to create highly accurate models even exploiting a limited number of input data, and computational saving ([KC$^+$09, SKK10, ASK$^+$13]). SVM is exploited in several manufacturing applications such as, machine condition monitoring, defect and fault diagnosis ([Dem13, XZA$^+$08, WY07, HY09, Chi02]). Another application field of SVM is in semiconductor manufacturing [LH09b].

Finally, another class of ML algorithm that is important to consider in this overview, are the **Ensemble Methods**. These methods use the combination of multiple models in order to increase the performance of classification or regression problems [Zho12]. The ensemble method exploit a set of basic classifiers (for example NN, DT, nearest neighbour) that are then combined in order to obtain a minor predictive error. A typical method to combine these basic classifiers is to use a voting system that counts how many basic classifiers would assign a new (previously unseen) observation t to each of the possible classes. Obviously, t is assigned to the class that has the most votes. There are two approaches, called *Boosting* and *Bagging*, to build basic classifiers.
In the first case, the classifiers are built sequentially: the output of the first model is used to generate the next classifier. Each model is trained on the dataset of its previous classifier, giving more weight to those data that had previously been misclassified. Thus, in Boosting, instead of taking a random subset of the training set, a set of misclassified samples are used to train each new model.
Differently, in Bagging, the training set is repeatedly sampled, with replacement and with uniform probability distribution. So, each dataset is generated from a random set, sampled with replacement, of the original training data. In this way, at each data extraction from the original training set to generate a new dataset, the data is replaced in the population from which it is sampled, therefore the same data can potentially become part of the same sample several times. This procedure allow to reduce the variance of the individual classifier. All these dataset have the same size of the original training set and, each one is exploited to create a model (classifier). During the classification phase of a previously unseen record x, each model classifies x in a class $Y_i \in \{1, , n_t\}$, and then, the final classification is the weighted combination of all the answers provided by each classificator.

Random Forest: One famous example of bagging methods [WWIT16] is the Random Forest algorithm [Bre01a], which is a combination of randomly sampled tree predictors. Each tree is built independently with bagging approach. Each tree node is split using the best among a subset of predictors randomly chosen at each node. In the end, a simple majority vote is taken for the prediction of new observations.
The main features that make RF compliant for manufacturing analysis are the ability to avoid overfitting and the ability to process high multidimensional and and multi-variate dataset, without variable selection ([Bre01a, Bia12, VGB11]).
In manufacturing, RF are mainly applied for monitoring machinery conditions, the diagnosys of the failures reasons, and the prediction of the Remaining Useful Life (RUL) of mechanical systems or components ([YDH08, CSS$^+$15, CY09, WJT$^+$17]).

In the following Chapters, we will discuss several contributions of this Ph.D. thesis in relation to the analysis and application of ML techniques in manufacturing problems. In Chapter 4.1, the effectiveness of the application of SLT to manufacturing systems is exemplified by targeting the derivation of a predictive model for quality forecasting of products on an assembly line.
In Chapter 6, we will consider a manufacturing problem where the RF algorithm was applied for maintenance forecasting in a production line.

2.4.4.2 Predictive and Prespective tools

To maintain symmetry with the BI tools presented in Section 2.4.3.3, this Section exploits the Gartner 2017 ”Magic Quadrant for Data Scientist platforms” report as a reference model to describe the current state of the art tools that offer ML solutions. This report addresses the so-called ”data scientists” - all those users who analyze data - providing two advices: (i) exploit open-source software in order to spread the application of these technologies on an even wider range of users, (ii) and try to use different applications in order to find the one that best suits the techniques of data processing desired by the user (some user prefers to write their own models in Scala rather than Apache Spark, rather than using Java code). The rules that establish the inclusion and position of analytics-platform-vendors on the Magic Quadrant, are based on three services (“use case”):

- **Production refinement**: the seller’s ability to continuously improve its product’s functionalities;
- **Business exploration**: the ability to provide simple and optimized features for extrapolating, exploring and visualizing data;
- **Advanced prototyping**: the key aspect of this report. The software’s ability in providing new data analysis capabilities, such as: the connection to different datasource types, the possibility to query data through natural-language, and the ML algorithms available to process data;

mapped on 14 functions (“capability”): Data access, Data preparation, Data exploration and visualization, Automation, User interface, Machine learning, Other advanced analytics, Flexibility, extensibility and openness, Performance and scalability, Delivery, Platform and project management, Model management, Precanned solutions, Collaboration, Coherence.

As can be seen in Figure 2.12, the four vendors positioned in the leading area are: IBM with SPSS Modeler and SPSS Statistics, SAS with Enterprise Miner (EM) and the SAS Visual Analytics Suite (VAS), the open-platform KNIME and, finally, RapidMiner.
IBM is the undisputed leader thanks to the continuous innovation of its data science and machine learning capabilities like supporting open-source technologies, supporting a broad range of data types (Hadoop distributions, NoSQL DBMSs and a variety of relational databases), and providing high capabilities of models management. Differently, SAS is more focused on interactive modeling with VAS. KNIME offers the open-source KNIME Analytics Platform with powerful functionalities for advanced data scientists. This platform is strong in several industries, especially in manufacturing and life sciences[5]. Finally, RapidMiner offers Graphical User Interface (GUI) suitable for beginner and expert data scientists, and it also offers access to open-source code. Table 2.2 and 2.3 show a list of ML algorithms supported by these four software platforms.

[5] https://www.knime.com/solutions/manufacturing

Table 2.2: Machine Learning algorithm supported by magic quadrant's leaders (part 1)

MACRO CATEGORIES	ALGORITHMS	SAS EM	RAPIDMINER	IBM SPSS	KNIME
CLUSTERING	Affinity Propagation				
	Agglomerate Hierarchical Clustering	X	X	X	X
	Anomaly Detection	X			
	DBSCAN	X	X	X	
	EM	X	X	X	
	Fuzzy Clustering	X	X		
	K-Means	X	X	X	X
	K-Medoids	X			
	K-Nearest Neighbor	X	X	X	
	Self-organizing maps	X	X	X	
	SVM	X	X	X	X
CLASSIFICATION	AdaBoost	X	X		X
	Artificial Neural Network	X	X	X	X
	C4.5 Algorithm	X	X		
	CART Algorithm	X	X	X	
	CHAID Algorithm	X	X	X	X
	Feature Selection	X	X	X	
	ID3 Algorithm	X	X	X	
	K-Nearest Neighbor	X	X	X	
	Linear Discriminant Analysis	X	X	X	X
	Logistic Regression	X	X	X	
	Multiclass Classification	X	X	X	
	Multilayer Perceptron	X	X	X	X
	Multinomial Logistic Regression	X	X	X	
	Multiple Discriminant Analysis	X	X		
	Naïve Bayes Classifier	X	X		
	Nearest Centroid Classifiers	X			
	Radial Basis Function Network	X	X		
	Rules Extraction System Family	X	X	X	X
	Support Vector Machine (SVM)	X	X	X	X
	Decision Tree	X	X	X	X
	Generalized Linear Models (GLM)	X	X	X	
REGRESSION	ANOVA	X	X	X	
	BI-Variate Regression	X			
	Exponential Regression	X	X	X	
	Least Median Squares	X	X	X	
	Logistic Regression	X	X	X	X
	Multiple Adaptative Regression Splines (MARS)				
	Multiple Linear Regression	X	X	X	
	Polynomial Regression	X	X	X	
ASSOCIATION	Apriori	X	X	X	X
	FP-Growth	X			
	FP-Tree				
	K-Optimal Rule Discovery	X	X		
TIME SERIES	ARIMA	X		X	
	ARMA				
	Brown Exponential Smoothing				
	Croston's Method	X			
	Forecast Accuracy Measures				
	Forecast Smoothing				
	Linear Regression	X	X	X	
	Exponential Smoothing	X	X		
	Transfer Function				
	FFT	X			
	Forest				
PROCESSING	Binning	X	X	X	X
	Inter-Quartile Range Test				
	Principal Component Analysis (PCA)	X	X	X	X
	Random Distribution Sampling	X			
	Sampling	X	X	X	
	Scaling Range	X	X		
	Substitute Missing Value	X			
	Variance Test				

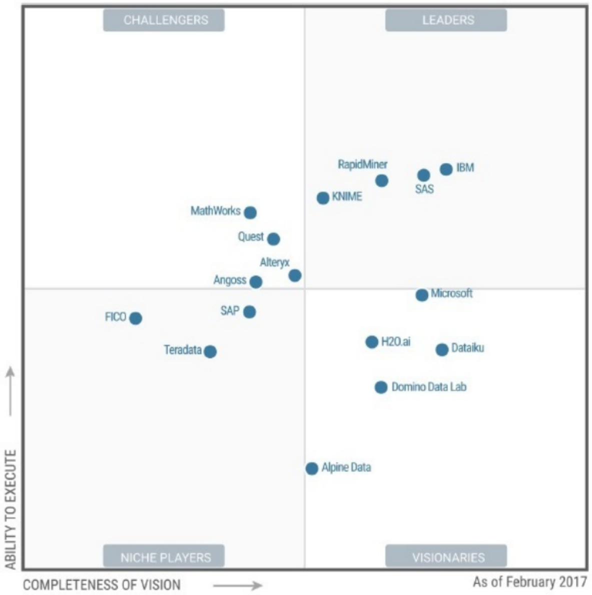


Figure 2.12: Magic Quadrant for Data Science Platforms.
`Source:www.gartner.com`

Table 2.3: Machine Learning algorithm supported by magic quadrant's leaders (part 2)

STATISTICAL INFERENCE	Bivariate Statistics	X	X	X	
	Chi-Squared Test	X	X	X	X
	Cumulative Distribution Function	X	X		
	Distribution Fitting	X	X	X	
	Multivariate Statistics	X	X		
	Quantile Function				
	Univariate Statistics	X	X	X	
MISCELLANUOUS	ABC Analysis				
	Scoring	X	X	X	X
	Weighted Score Table				
NETWORK ANALYSIS	Link Analysis	X		X	
	Social Network Analysis	X			
SURVIVAL ANALYSIS	Survival function	X		X	
OPTIMIZATION	Genetic Algorithm		X		

Part II

Manufacturing intelligence and analytics

Chapter 3

Value Mapping and Assessment Framework for Sustainable Manufacturing

A key topic during this Ph.D. was the investigation of the main KPIs and the key strategic priorities (or dimensions of analysis) by which these indicators can be investigated in Manufacturing environment. As mentioned in Chapter 1, the term "Industry 4.0" refers to a new production patterns, including new technologies, productive factors and labour organizations, which are completely changing the production processes and the relationship between customer and company with relevant effects on the supply and value chains [Cas]. Even though most of the aforementioned innovations are in an embryonic stage, they are still an important part of research and progress. At the same time, given the uncertainty and complexity of these new paradigms, several national strategies and new technological roadmap such as, the German high-tech strategy "Industrie 4.0" or the Italian cluster "Fabbrica Intelligente" (FI), aim at approaching this transformation enhancing the sustainability, flexibility and re-configurability of current manufacturing systems among many other competitive dimensions. The table 3.1 shows ten roadmap taken in consideration.

The main topics covered by these roadmaps are:

- Flexible, personalized and innovative production systems;
- Strategies, methods and tools for industrial sustainability;
- Digital transformation in the manufacturing environment.

The first two points reflect the need of change the way by which industrial processes are performed, towards new approaches and new objectives. Whereas, the third point, is focused on what are the new technologies that can help industries in achieving these new strategies.
The first research priority has as its major theme "Customization", which over the past 15 years

Table 3.1: Ten road-mapping programs

Country	Title	Date of issue
Austria	BMVIT (Austrian Ministry for Transport, Innovation and Technology) Innovation: Solutions for the future	October 2009
Denmark	Manufacturing 2025: Five future scenarios for Danish manufacturing companies	May 2010
Finland	Finland's regional development strategy 2020	September 2010
France	France Europe 2020: A strategic agenda for research, technology transfer and innovation	February 2013
Ireland	Making it in Ireland: Manufacturing 2020	2012
Latvia	Sustainable Development Strategy of Latvia until 2030	June 2010
Netherlands	Smart Industry – Dutch industry fit for the future	April 2014
Sweden	Swedish Production Research 2020	2008
UK	The future of manufacturing: a new era of opportunity and challenges for the UK	2013

has emerged as one of the strategies that allows companies to supply bespoke products that can be mass-produced in a flexible manufacturing systems. This topic is associated with different aspects of product development, such as Information and Communication Technology (ICT) solutions for the acquisition of the client's requirements, product configurators, advanced measuring systems, platforms for client monitoring and technologies for personalized production, such as additive manufacturing, micro-manufacturing, hybrid processes, and so on.
The second point focuses on "strategies, methods and tools for industrial sustainability", which has become a key issue on the agendas of industrialists and politicians [CHIS15]. The aim is for an improved understanding in environmental, social and economic challenges leading to a transformation in industrial behaviour. Three waves of change will shape the industrial system over the next two decades, these are:

- Improvements in environmental performance without changing current products and processes,
- Development and introduction of new technologies,
- Changes in the industrial system as a whole.

This requires awareness and re-engineering of the industrial processes in order to reduce carbon emissions and improve energy efficiency shifting towards the *Industrial Symbiosis*.
Industrial symbiosis is, therefore, a system in which all activities, starting from extraction and production, are organized in such a way that waste becomes a resource, unlike the linear economy, in which, a product becomes waste when consumption ends, forcing the economic chain to continually repeat the same sequence: extraction, production, consumption, disposal. These

systems must be coherent with the evolution of the markets and future technologies, using them as a competitive lever towards the three areas of sustainability (economic, environmental and social). Within industrial sustainability, specific interest in de-manufacturing has recently grown due to the rising cost of raw materials and specific laws introduced by the European Union to improve the recovery rates of the materials. Furthermore, the demand for materials critical to the manufacture of high-tech products is constantly increasing in Europe, causing huge problems in economic and strategic terms (e.g. electronic waste, which is an important source of metals for technological products).

The third and final research priority concerns the "Digital transformation of manufacturing environment". Digital manufacturing allows the simulation of the whole supply chain with the idea of the virtual factory, integrating procurement, production, product logistics, service and diverse IT technologies in order to predict, solve and control problems in the virtual and physical environment. These technologies allow the reduction of time to market, decrease in costs and the increase of process flexibility by analysing production data.

It is fundamental to have a universal vision of the current industrial scenario in order to identify the main objectives and needs.

The aim of this research was to investigate the relevance of sustainability within the smart manufacturing environment and provide a guideline for companies, in order to understand what sustainability trends and drivers are fundamental to the manufacturing environment and a set of KPIs that allows firms to control and monitor the reaching of these goals. This sustainability guideline was developed by performing a qualitative literature review on industrial sustainability performance management with the aim of analyzing the existing state of the art, re-organizing these information accordingly to a hierarchy of structured KPIs, and highlighting the areas that need to be further developed both in the literature and in the industrial systems. Related works in the context of this research consist of two research domains. On the one hand, there is the research regarding Industrial Sustainability and on the other hand the one on value modeling and mapping already addressed in ([LBJ16, YL16, KLC$^+$16, AA11]) within the Manufacturing Value Modeling Methodology (MVMM). Both domains are crucial for implementing a proper sustainability catalog; the research on industrial sustainability is important for the creation of the underlying sustainability model, while the research on manufacturing value modeling is seen as a key influencer towards constructing the framework for identifying the correct sustainability demand.

In Section 3.1 we will discuss the concept of value creation and modelling in manufacturing companies. Finally, in Section 3.2, we will present one of this Ph.D. thesis contributions: how the MVMM was used to develop the sustainability catalog.

3.1 Value Modeling in Sustainability

Sustainable manufacturing is becoming increasingly important. This requires a sustainable industrial system that differs from today's global industry with different business models, creating different products and services. The evolution towards a sustainable industrial production system requires a holistic approach, with a fundamental reassessment of value creation.
In order to achieve this target, a system design approach is required. An existing and specific MVMM is used as a value-mapping framework to help firms in creating value propositions better suited for sustainability, considering economic, environmental and social perspectives. Like any business, addressing important sustainability issues requires specific, hard-wired organizational support, capabilities, and measurement. As highlighted by Smith et al. [SB12a], achieving sustainability in manufacturing requires a holistic view spanning product design, manufacturing processes, manufacturing systems, and the entire supply chain. Such an approach must be taken to ensure the economic, environmental and societal goals of sustainability are achieved.
Related works in this context are in two domains.
On the one hand there is research regarding industrial symbiosis and on the other hand research in the field of value modelling. The value mapping model now combines the two perspectives by holding on to the hierarchical structure of value proposed by Bocken et al. [BRS15] (Figure 3.1), as well as respecting internal and external factors.

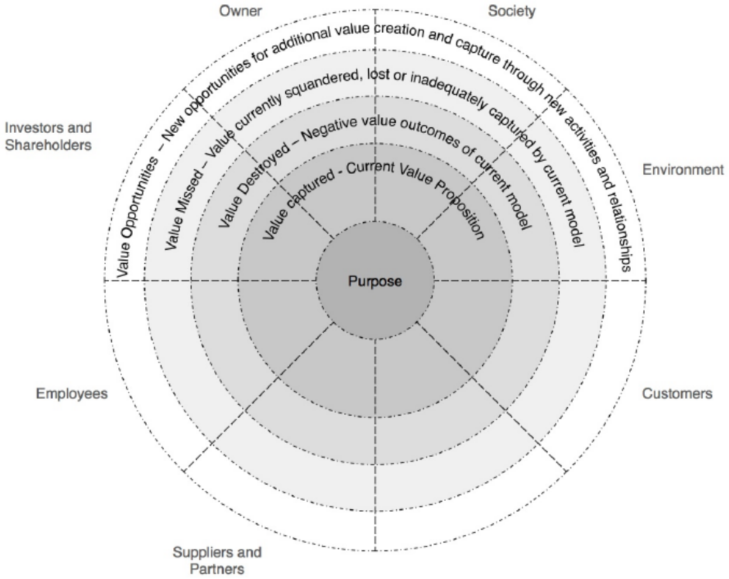

Owner
Society
Investors and Shareholders
Environment
Employees
Customers
Suppliers and Partners
Value Opportunities – New opportunities for additional value creation and capture through new activities and relationships
Value Missed – Value currently squandered, lost or inadequately captured by current model
Value Destroyed – Negative value outcomes of current model
Value captured - Current Value Proposition
Purpose

Figure 3.1: Value mapping tool proposed by Bocken et al. (2014).

This combination also leads to further review considering evaluation regarding the relationships between these factors. Thus, the purpose is to present an overview regarding sustainability trends, implications and possibilities that could affect manufacturing companies and supply chains, with the aim of creating a model that allows different dimensions of industrial sustainability (economic, environmental and social) to be mapped. The hierarchical structure of value proposed by Bocken et al.is composed of:

- Value captured: “represents the positive benefits delivered to stakeholders”;
- Value missed: “represents cases where stakeholders fail to capitalize on existing assets, capabilities and resources, are operating below best practices or fail to receive benefits they seek from the network”;
- Value destroyed: “is negative outcomes of the business and concerns the damaging social and environmental impacts of business”;
- Value opportunities: “firms will need to go beyond “damage control” and seek out new value creation opportunities to deliver novel solutions to social and environmental problems that begin to address the wider sustainability challenges directly”.

Starting from these definitions of value, we have developed a value-mapping model with the aim to provide a guide for manufacturing companies, in order to understand what sustainability trends are fundamental to the manufacturing environment.
The value-mapping model adopts the structure of the MVMM, in order to include internal and external influence factors analysing them with respect to the triple bottom line approach. The value-mapping model starts from the core concept of the MVMM, using the structure of external influence factors (*Trends*), internal influence factors (*Implications*) and *Possibilities*, and also applies the value map by using the aforementioned contents, and the concept of relationships between the value map items.

External factors The external view represents trends; they are sustainable challenges that have an impact on the manufacturing environment. This Section gives a background on the challenges associated with embedding sustainability into corporate performance management. Examples of trends could be for instance “Manage environmental changes” and/or “Reduce energy consumption”. Due to the different markets in which companies operate, the trends might vary from scenario to scenario and related industrial context. While there might be trends, which are globally valid, there are also trends which are only true for a certain branch of industry. It is important to study the environment of the company and the domain in which it operates in order to identify a valid set of trends. The external view is followed by the analysis of the internal process and strategies.

Internal factors After the market related view, the MVMM suggests reviewing the implications. The goal thereby is to identify the strategy of the company and how it is achieved. Nevertheless when analysing the company, it is mandatory to understand the business side and production side, since they should fit in the overall strategy of the company. The purpose of this step is to set up a system that identifies critical areas, which have to be addressed. After identifying the implications, it is important to further specify the context in which the sustainability demand occurs with the analysis of the Possibilities.

Possibilities The identification of the context consists of selecting the correct functional areas or practices, which need a detailed analysis. Focusing on production, these practices are the functional areas in the MOM domain. Starting with the analysis of the manufacturing strategy, and then focusing on sustainable industrial practices brings out an alignment of manufacturing strategy with business strategies.

Relationships in the value-mapping model From a value modelling point-of-view, capturing the environment of the given scenario by identifying the external and internal influence factors and mapping them is necessary to find out which domain specific market trends fit to which domain specific project targets. Besides the general description of the value-modelling model, it is mandatory to explain the application of the model itself. Since the general approach is derived from the MVMM approach it is also possible to create relationships between the different components (Figure 3.2).

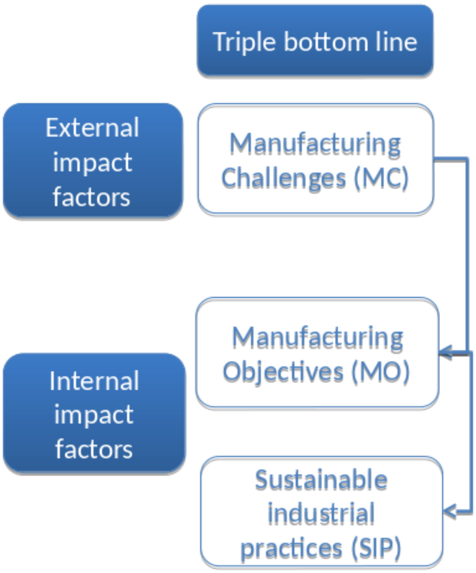

Triple bottom line
External impact factors
Manufacturing Challenges (MC)
Manufacturing Objectives (MO)
Internal impact factors
Sustainable industrial practices (SIP)

Figure 3.2: The link Internal and External impact factors in Manufacturing and their evaluation respect to the environmental, social and financial dimensions (Triple Bottom Line).

Generally speaking there is the possibility to create a relationship between external influence factors (Trends) and the business strategy (Internal factor) that is used to tackle them. This means there is a certain set of internal influence factors that fit to a certain set of external factors. Furthermore, after the assessment and the definition of external, internal impact factors and possibilities, it is possible to analyse these contents through a value analysis, which identifies:

- Positive and negative aspects of value in the company or network;
- Possible sources of conflicting value;
- Value opportunities to improve sustainable development.

3.2 Sustainability Catalogue

Company strategy to achieve competitive advantage is composed by the capacity of making decisions, the identification of business drivers and the ability to model strategic dimensions that affect the market, the environment and the company itself [BDT$^+$17]. It is in this context that Industrial Sustainability is recognized as a strategic dimension that affects the manufacturing environment such as flexibility, innovation, and quality. In particular, sustainability is a capability that allows to achieve competitive advantage through the increase of material efficiency, energy saving, closed-loop control at industrial system level [TET13], and through the increasing competitiveness by improving economic, environmental and social performance [TSM15]. Regarding industrial sustainability performance management, despite many literature contributions regarding sustainability, there is a lack of understanding of how sustainability may be effectively embedded in corporate performance management systems [BME13] especially on the manufacturing area. Hence Authors decide to focus their attention on sustainability both as a competitive and strategic dimension in the manufacturing environment with respect to value modeling and the industrial sustainability performance management. Thus the purpose of this paper is to present an overview regarding sustainability trends, implications that affect the manufacturing company and supply chains, and lead a review in order to analyze the existing body of literature on performance management, with the aim of creating a set of structured Key Performance Indicators (KPIs). As proposed by Authors [TDLT16], a Manufacturing Value Modeling Methodology (MVMM) is used to model the internal, external conditions that affect the company, organizing these contents in a set of structured information specifically named catalogue. From a value modeling point-of-view, capturing the environment of the given scenario by identifying the external and internal influence factors and mapping them is necessary to confront these evidences with preset catalogues containing suggestions about those sustainability practices that might be suitable for handling the given set of internal and external influence factors. Moreover, in order to identify important sustainability issues and address them to a large degree and change perceptions, it is important to require specific organizational support, capabilities and measurement in terms of specific indicators. Hierarchical industrial sustainability metrics and related key performance indicators (KPIs) need to be defined and developed in order to translate business goals into manufacturing strategy, and possibly improving manufacturing operational performance by adopting a set of best practices. Within this context Authors present a (Industrial) Sustainability Catalogue with an overview on implications and possibilities of Industrial Sustainability in Section 3.2.2, then a simple qualitative literature review is reported on Industrial Sustainability

Performance Management in Section 3.2.3. In Section 3.2.4 is presented the reorganization of data collection on metrics and KPIs by adopting a hierarchical structure highlighting limitations and missing areas. Finally, Section 3.2.5 shows the consequences and issues of the Authors' work and conclusions are discussed.

3.2.1 Sustainability Catalogue

The scope of the Authors' work is to provide a guide for manufacturing companies, in order to understand what sustainability trends and drivers are fundamental to the manufacturing environment and a set of KPIs that allows firms to control and monitor the reaching of these goals. The sustainability catalog combines these two approaches by using the MVMM, in order to include internal and external influence factors as well as the hierarchical KPIs. The sustainability catalogue uses the structure of external influence factors (Manufacturing Challenges), internal influence factors (Manufacturing Objectives and Sustainable Industrial Practices) from MVMM.

3.2.1.1 External Factors

The external view represents Manufacturing Challenges, this component describes sustainable challenges that have an impact on the manufacturing environment. This section gives a background on the challenges associated with embedding sustainability into corporate performance management. Examples of Manufacturing Challenges could be for instance Manage environmental changes and/or Reduce energy consumption. The external view is followed by the analysis of the internal process and strategies.

3.2.1.2 Internal Factors

The internal influence factors are used to represent the sustainable goals and strategies of the manufacturing company. Different internal influence factors could be identified as:

- *Manufacturing Objectives*: describe the company strategy in terms of sustainable opportunities and issues;
- *Sustainable Industrial Practices*: as a set of planning practices, production, purchasing and logistics aimed to incorporate a sustainable perspective in operations.

3.2.2 Sustainability Performance Management qualitative literature review

The dataset selected for this study was "Scopus", the Authors interrogated the database searching for ("Sustainable" OR "Sustainability") AND ("Supply Chain Performance") AND ("Manufacturing" OR "Management" OR "Metrics" OR "Indicators"), in the titles, abstracts and keywords of papers published between 2000 and 2015. The interrogation resulted in 175 papers published in 88 different journals that constitute the basis for further analysis. The earliest paper included in the dataset was published in 2004 and the most recent in 2015. Figure 3.3 presents the list of the first journals where research has been published. Journal of Cleaner Production, Supply Chain Management, International Journal of Production Economics and International Journal of Production Research lead the ranking with 16, 12, 11, 11 publications, respectively.

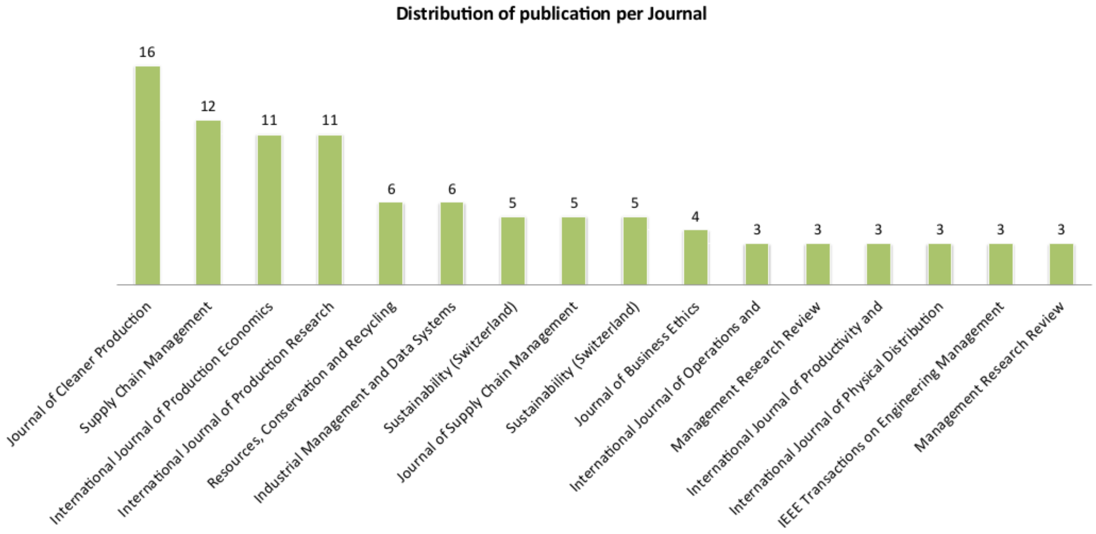


Figure 3.3: Distribution of publication per Journal.

Figure 3.4 presents the geographic diversity of scholars. In this case it is relevant to note the leadership of USA and European academic institutions that contribute substantially equally to the research field development.

Further, Figure 3.5 presents the frequency of publications over time highlighting a research field that is growing rapidly, while Figure 3.6 highlights the ranking of keywords used by authors.

The literature review on sustainable performance management has highlighted that sustainability metrics and indicators may be applied with different levels of complexity [THE07] and are increasingly being recognized as practical tools for manufacturing decision making, and communication purposes in many contexts. Previous research has shown that there is a great range in the types of metrics reported, though many metrics for all three areas of the triple bottom line are available [RS12]. The analysis of the published metrics highlights the need for the devel-

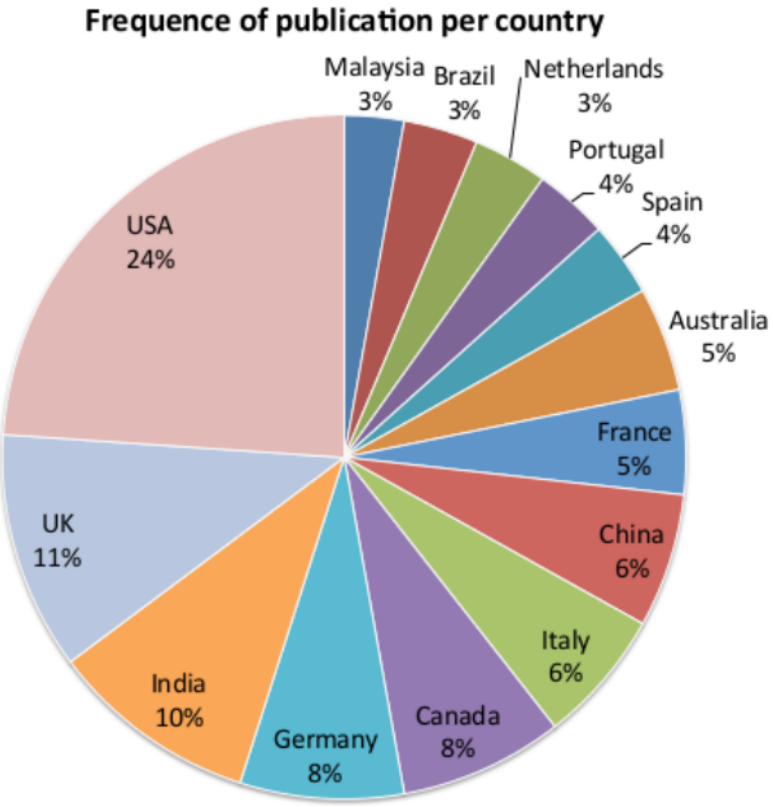


Figure 3.4: Distribution of publication per country.

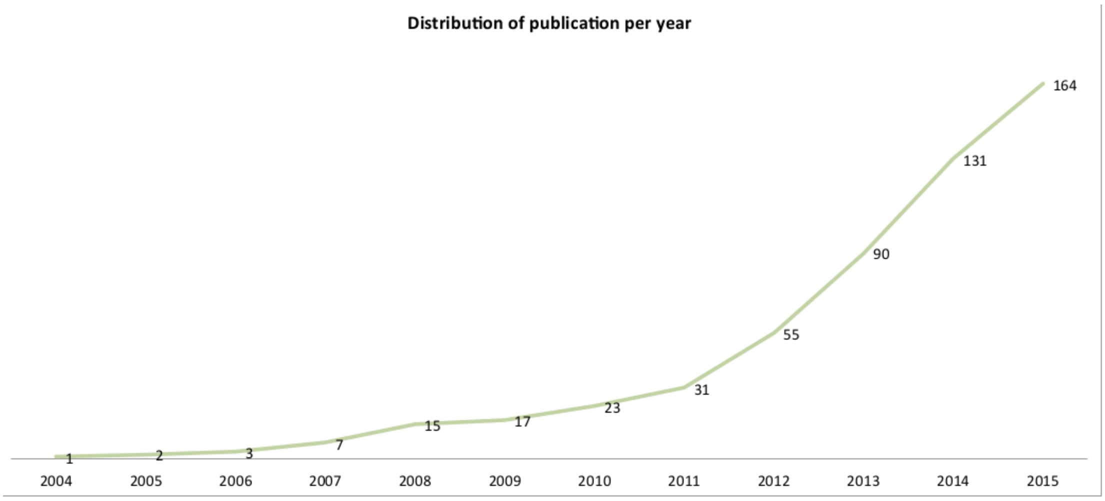


Figure 3.5: Distribution of publication per year.

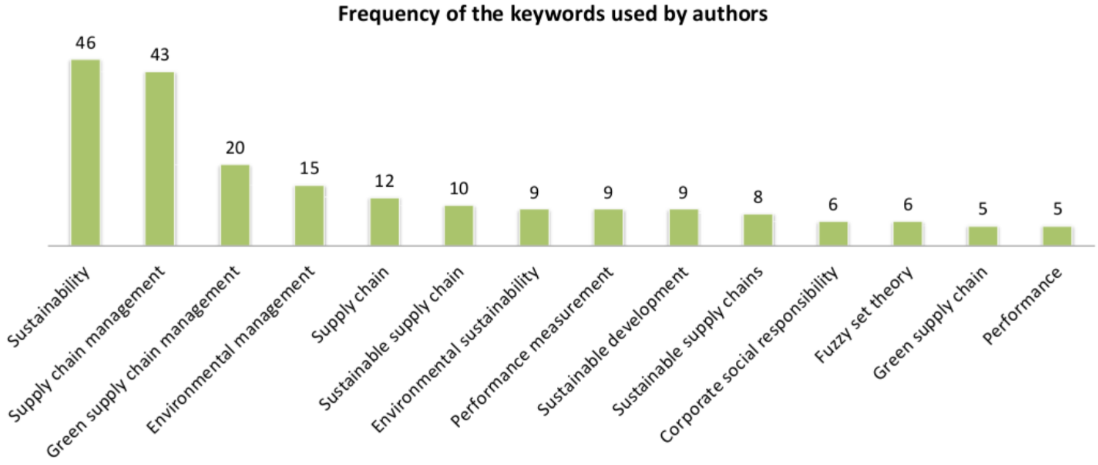


Figure 3.6: Distribution of publication per keywords.

opment of an original conceptual framework for measuring performance in Sustainable Supply Chains [AS15]. In the next section the Authors provide the first comprehensive database of metrics that have been reported in the literature, and the analysis provide a starting base for future academic and practitioner work.

3.2.3 Measuring Sustainability dimension

Sustainability even if a relatively new research area, already shows an interesting number of measures and metrics mainly de-structured and at very different levels. This variety is creating confusion among manufacturers when they attempt to select an operational set of indicators for assessing sustainability in manufacturing in practical terms [RWVD08] being the related practices interacting each other through not so trivial relationships and trade-offs. A further challenge in selecting metrics for sustainable manufacturing, in fact, is that it is not an inherently intuitive process; these metrics, in fact, are not necessarily related to the function of the part of being manufactured. Additionally, a complete picture of environmental impact and sustainability requires numerous metrics [MBST15].

To address this challenge Authors have collected almost a hundred of sustainability measures that they categorized among indicators and metrics based on the thirteen Manufacturing Objectives listed in Table 3.3. Finally, they grouped these Objectives in four main sections: General Aspects, Materials, Energy and Emissions.

General Aspects encompasses two manufacturing objectives: Safety and Climate Change. This section takes into account objectives that are not specific on how to improve the manufacturing processes but, indirectly, are likely to affect the industry's sustainability performance.

Materials covers all those metrics and indicators about the efficient and effective use of material and it is made up of five objectives: Material Efficiency, Reduce usage of raw material, Increase usage of renewable material, Minimize water usage and Waste reduction.

Energy is the most commonly used and analysed field for the assessment of the environmental performance. It covers two main aspects: Energy Efficiency and Decarbonisation of global system. For both these two categories there is a detailed KPI hierarchy that allows an accurate assessment of targets that a manufacturing industry wants to achieve.

Finally, Emissions includes the intensity of the weight of all releases to air/land/water during a reference period. Its objectives are: Minimize emissions to air, Minimize emissions to land, Minimize emissions to water and Increase recycling rates.

In Section 3.2.7 is reported the Authors' categorization of manufacturing sustainability KPIs and metrics.

3.2.4 Industrial Sustainability Performance Management: A KPIs hierarchical approach

As previously introduced, the Authors' aim is to start sketching a sustainability KPIs hierarchy which translates business goals into manufacturing strategy, and allows to improve operational performance by adopting a set of best practices. In order to align the company and manufacturing strategy, a multi-level structure of strategic and operational KPIs is developed, with the scope of mapping the Sustainability Catalogue Internal Factors' respect to metrics and indicators which have been reorganized. The framework proposed, allows to define objectives, establish goals and measure operational progress, aiming to a constantly monitor and control of relevant activities. Below two examples of the proposed framework are shown. Figure 3.7 reports the connection between manufacturing objectives, Sustainable Industrial Practices and metrics/KPIs respect to the energy section.
In particular, companies can control the achieving of the "Energy efficiency" goals through the high level indicator "Total Energy Consumption" which is composed by three more operational metrics: "Energy intensity", "Energy availability" and "Energy usage".
Similarly Figure 3.8 shows the same connection respect to the materials section with the specific sustainability goal "Minimize water consumption" whose achievement is monitored by "water consumption". As in the energy example, the high level metrics consists of "water loss", "% water loss" and "water intensity" which are a lower level metrics.

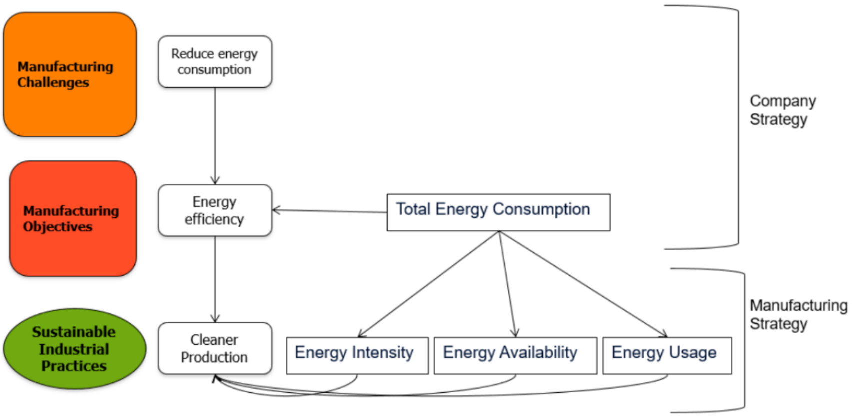


Figure 3.7: Mapping between MVMM and Energy section.

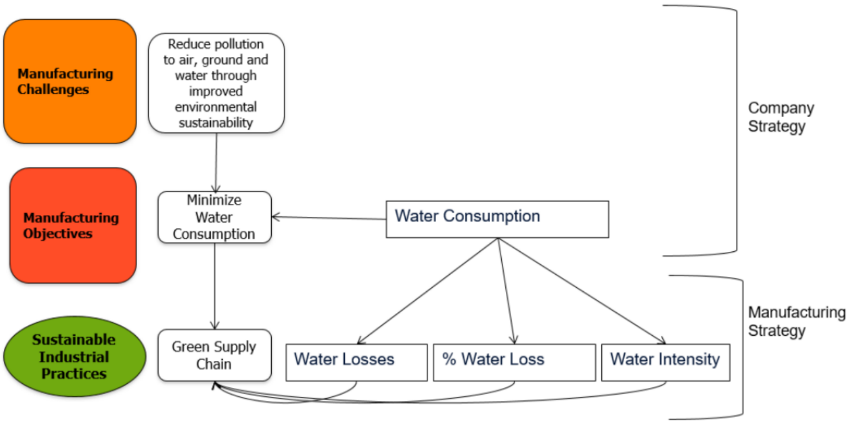


Figure 3.8: Mapping between MVMM and Material section.

3.2.5 Discussions and implications for MMVM

It is possible to note that, except for the Energy, all the other sections do not present a complete set of indicators. One limitation is due to the lack of the availability of sustainability-related data. Even if a precise calculation of all indicators can be performed, problems may occur due to granularity of available data. In fact, decision making in a manufacturing enterprise can take place at many different levels, therefore the scope of application should be understood when using the metric formulations give above [MBST15]. In the current situation, sustainability monitoring is widely done on plant level, but to really use the approach, a finer granularity on machine level is necessary. A real and complete case analysis can only be done in a time frame in which the multi-level sustainability measurement and monitoring systems allow sustainability analysis for different levels of the plant, down to machine level. However, such limitation could be dealt with by increasing the role and use of advanced data collection IT systems in manufacturing environments [MML$^+$13]. Another reason for the scarcity of indicators in some manufacturing objectives like Minimize emission to land, Minimize emission to water, Waste reduction, Material efficiency is related to the fact that these issues gained importance only in the last 10-15 years and they require a deep (re)design of the manufacturing processes in order to be taken into account (see the ambitious and innovative Zero Emission Manufacturing effort done by Toyota Motor Europe to this end)[1]. By performing a qualitative literature review on Industrial Sustainability Performance Management, Authors identified and structured different sections that can be improved in terms of industrial sustainability performance management.
As highlighted by the categorization of KPIs, the last level of the MVMM, it lacks a hierarchy of indicators for assessing sustainability in manufacturing. While it is easier to find metrics that provide an overview on the variables to consider for assessing a specific problem, it is not so easy analyze it in a timely manner through a set of indicators.

3.2.6 Conclusions

Authors' purpose was to examine evidence of industrial sustainability as a manufacturing strategic dimension and collect metrics and indicators in order to create a hierarchical KPIs with the scope of assessing the performance of the firm in terms of sustainability and suggesting related best practices. To accomplish this, Authors realized a structured Industrial Sustainability Catalogue to be complementary used within the proposed MVMM methodology, with the aim to translate sustainable trends and goals into manufacturing strategy, and suggesting a set of best practices and related KPIs. Authors identified 21 sustainability indicators, both simple and aggregated, able to cover nine on the thirteen Manufacturing objectives selected. It is noticed that Energy Efficiency and De-carbonization of the global system are the areas most explored both

[1] `http://media.toyota.co.uk/wp-content/files_mf/1323855497environnement_brochure_11.pdf`

in the literature that at industrial level, therefore, it has been possible to derive the detailed indicators. In contrast, aspects such as Waste reduction, Material efficiency, Minimize emissions to land and Minimize emissions to water are still not treated in detail, at least at industrial level, showing a relevant research gap on the characterization of sustainability metrics for manufacturing processes. These metrics and KPIs collection and analysis show the research that gaps identified in the literature review and the poor relationships with best practices suggestion and adaptation inhibit manufacturing strategy transformation at industrial sustainability level. In particular, Authors feel that more attention should be given to industry-specific research on sustainable supply chain management. Finally, it is evident that this area still requires significant investigation at the operational and strategic levels, the framework provided will guide industry and supply chain sustainable progress and improvement.

Table 3.2: Manufacturing Challenges.

External Factors: Manufacturing Challenges
Change in the interaction with the individual (customer, worker, citizen)
Exploitation of energy from waste and scrap
Growth of a new middle class at global level
Growth of emerging countries (production and consumption)
Higher production flexibility and re-configurability
Highly variable and difficult to forecast market conditions
Increase in productivity
Increase in urbanization - integration of industry in urban context
Increase the resilience of industry to global warming and climate change (on production, procurement and markets)
Increase the worker well-being in terms of high satisfaction, safety and inclusivity
Manage environmental changes due to exploitation of farmland, deconstruction of infrastructure and urbanization
Need to enhance specific competences and skills of each geographical area
Need to manage dynamic and complex business networks
Need to recycle components and products
New forms of employment
New models of collaboration reshoring-offshoring-nearshoring
New services tailored on the people
Pervasiveness of internet
Products to satisfy the demand for comfort, health and wellbeing of specific target groups
Reduce energy consumption
Reduce pollution in air, ground and water through improved environmental sustainability
Use of alternative energy sources in manufacturing

Table 3.3: Manufacturing Objectives

Internal Factors: Manufacturing Objectives	**Description**	**Sources**
Client satisfaction	Providing greater satisfaction, well-being, and added value to customers and users.	(12)
Decarbonisation of the global energy system	The CO2 emissions from materials production and processing could be reduced if the processes were powered by less carbon intensive energy.	(13)
Energy efficiency	Improving the energy efficiency of a factory as well as for calculating the embodied energy of a product, fostering the energetically optimization of its manufacturing processes.	(14)
Increase recycling rates	Extending the supply chain to include issues such as remanufacturing, recycling and refurbishing added complexity to supply chain.	(15)
Increase usage of renewable resources	Consider long term criteria in proposals, i.e. adaptability for future changes, whole life costs, select renewable resources.	(13)
Material Efficiency	Material efficiency means providing material services with less material production and processing	(13)
Minimize emissions to land	Acidification potential, land usage.	(16)
Minimize impact on species	Enhance and protect species and the natural environment	(12)
Minimize water usage	Actions were proposed to use auxiliary cleaning techniques using mechanical cleaning methods, usage of steam, etc., to minimize water usage and to preserve water for future use.	(8)
Reduce usage of raw material	Minimize consumption of raw material and natural resources.	(12)
Reduction of air emission	Optimization of process to reduce air emissions.	(17)
Safety	Enhancing site and welfare conditions, respect employees and the wider community.	(12)
Waste Reduction	Products are pulled through chains in response to demand, with minimal waste and inventory. It primarily takes a cost-based view of value, focusing on the drivers of efficiency and waste.	(18)

Table 3.4: Sustainable Industrial Practices

Internal Factors: Sustainable Industrial Practices	**Description**	**Sources**
Ecodesign	It is treated as the designing phase of product life cycle. It is based on Life Cycle Assessment that is a technique that summarizes the quantification of the environmental consequences of products and services.	(19)
Green Supply Chain (GSC)	GSC is viewed within the planning and sourcing phase of the product life cycle. GSC can be understood as sustainable operations practices together with suppliers and/or customers covering project design, selection of raw materials, selection of suppliers, green purchasing, packaging and logistics.	(20)
Cleaner Production (CP)	It refers to the production phase. It represents the application of an economic, environmental and technological strategy integrated with the processes and products in order to make them more efficient.	(21)
Reverse Logistics (RL)	RL refers to the management of waste related to the consumption of manufacturing products. Reverse logistics can be understood as the return process of moving goods in order to capture value or give the appropriate destination.	(22)

3.2.7 Additional Information regarding Manufacturing Objectives

In the following we present the Manufacturing Objectives with their category in parenthesis with the collected manufacturing Metric and KPI:

- **Safety** (General Aspects):
  - Metrics:
    1. Noise level;
    2. Impacts of activities and operations on protected and sensitive areas;
    3. Labour/ employment issues: standard issues such as health and safety, education, training, industrial relations, wages, benefits, conditions of work/employment, accountability, image/reputation and harassment;
    4. Supplier relationships: contractual agreements with suppliers, supplier diversity and company policies on the screening of suppliers.
  - KPI: Time Weighted Average = 16.61 log_{10} (D/100) + 90 where D = Noise Dose = (time actually spent at sound level $*$100%)/maximum permissible time at sound level
- **Climate change** (General Aspects):
  - KPI: GROI = Climate Change = (GHG_bau - GHG_investment) / GHG_investment where GHG_bau (Business As Usual) is the life-cycle greenhouse gas emissions of the current technology and GHG_investment is the greenhouse gas emissions from installing or utilizing a new technology
- **Material efficiency** (Materials):
  - Metrics:
    1. Total materials use other than water, by type.
    2. Percentage of materials used that are wastes (processed or unprocessed) from sources external to the reporting organisation. Refers to both post-consumers recycled material and waste from industrial sources. Report in tonnes, kilograms, or volume.
    3. Annual mass flow of different materials used (excluding energy carriers and water) in tonnels.
    4. Ratio of the used recycled input materials expressed in units % of the total input materials % of the total input materials.
    5. Weight of materials used.
    6. Percentage of materials used that are recycled. Percentage of products sold that is reclaimed at the end of the product's useful life by product category.

- **Reduce usage of raw material** (Materials):
  - Metrics:
    1. Raw material consumption.
    2. Non-renewable materials intensity.
    3. Usage of resources: water usage, material usage (overall material consumption, virgin material consumption, recycled material consumption, reused materials, remanufactured materials, other materials.
  - KPI: Years Remaining = Material Scarcity = Stock global / (Annual Use global * (1 – Recycling Rate))
- **Increase usage of renewable material** (Materials):
  - KPI: Availability Factor = Rate Of Consumption_investment / (Sustainable Available Stock – Rate Of Consumption_GeographicRegion + Rate Of Consumption_BusinessAsUsual)
- **Minimize water usage** (Materials):
  - Metrics:
    1. Water eutrophication.
    2. Water consumption Total annual water consumption m3/year.
    3. Percentage and total volume of water recycled and reused / The percentage of water recycled from the production process.
    4. Total water withdrawal by source.
    5. Total water discharge and quality.
    6. Mains water use (designed) – product Median water use m3 / 100m2 gross floor area.
    7. Mains water use - construction process Median water use m3 / £100k project value.
    8. Total level of water losses in company systems yearly. Leakage rate of the total water distribution network.
  - KPI:
    1. Water losses in cubic meter/km/day
    2. Water loss %
    3. Water Intensity = Total water intake/ Normalization factor
    4. Water Availability Factor = Water Use (local - investment) / (Renewable Water Supply_local – Water Use_local)
- **Waste reduction** (Materials):

  - Metrics:
    1. Residuals Intensity.
    2. Total amount of waste by type and destination.
    3. Weight of transported, imported, or exported waste deemed hazardous.
    4. Waste (Recycling, recovery and landfill).
    5. Radioactive waste Companies measuring radioactive waste should follow guidelines from the environmental regulators. Solid radioactive waste should be reported at the company level on an annual basis in kg or metric tons of HLW, ILW and LLW.
    6. Waste disposal to landfill expressed as percentage of production by weight.
    7. Total annual generation of waste, broken down by type in tons.
    8. Total annual generation of hazardous waste in kg or tons.
    9. Waste - construction process Median waste removed from site m3 / £100k project value.

- **Increase usage of renewable material** (Materials):
  - KPI: Availability Factor = Rate Of Consumption_investment / (Sustainable Available Stock – Rate Of Consumption_GeographicRegion + Rate Of Consumption_BusinessAsUsual)

- **Energy efficiency** (Energy):
  - Metrics:
    1. Total Energy consumption.
    2. Total direct energy use. The total energy consumption of the organization is possible to express also in tons of oil equivalent MWh, Gj. Total renewable energy use. The % of the total annual consumption of energy (electricity and heat) produced by the organization from renewable energy sources.
  - KPI:
    1. Energy Intensity = (energy consumed in production process + energy consumed in overhead)/Normalization factor
    2. Lean Energy = Valuable Energy Consumption/Overall Energy Consumption
    3. Energy Quality = Valuable Energy Consumption / Net Production Energy
    4. Energy Availability = Gross Production Energy / Net Usage Energy
    5. Energy Usage = Net Usage Energy / Energy in operating time
    6. Energy Opening = Energy Consumption in Opening Time / Energy Consumption in Theoretical Production Time

        7. Energy Scarcity = EnergyUse_investment/ (EnergyCapacity_local - EnergyUse_local + EnergyUse_BusinessAsUsual) Energy Independence = (EnergyUse_BusinessAsUsual - EnergyUse_(local-investment)) / EnergyUse_(local-investment)

- **Decarbonisation of global system** (Energy):
    - Metrics: Carbon footprint.
    - KPI: CO2 Effectiveness = Transport CO2 Effectiveness / Warehouse CO2 Effectiveness, where Transport CO2 Effectiveness = (Tonne.km * Transport CO2 Efficiency)/Total Volume Sold, and Warehouse CO2 Effectiveness = (Stok Levels* Warehouse CO2 Efficiency)/Total Volume Sold

- **Increase usage of renewable material** (Materials):
    - KPI: Availability Factor = Rate Of Consumption_investment / (Sustainable Available Stock – Rate Of Consumption_GeographicRegion + Rate Of Consumption_BusinessAsUsual)

- **Minimize emission to air** (Emissions):
    - Metrics:
        1. Air acidification.
        2. Air releases intensity.
        3. Dust and particles.
        4. Volatile organic compounds.
        5. Metal emissions to air.
        6. Transport: Road, Rail, Air, Road Freight and Other Freight. For example, conversion of miles travelled in medium-sized petrol car to tons of CO2 emitted.
        7. Total annual emissions of greenhouse gases including at least emissions of CO2, CH4, N2O, HFCs, PFCs, SF6 tons of CO2 equivalent. Total annual air emission, including at least SO2, NOx e PM Kg, tons.
    - KPI:
        1. Green House Gases Intensity = (GHGS released in energy consumption for production + GHGS released in energy consumption for overhead + GHGS released by transport used for business travel + Additional GHGs released from production process)/(Normalization factor)
        2. Intensity of pollutant releases to air = Weight of releases to air/Normalization factor

- **Minimize emission to land** (Emissions):
    - Metrics:

        1. Oil and coolant consumption.
        2. Restricted substances intensity.
        3. Metal emissions to land acid and organic pollutant emissions to land. Metal emissions to land arising from industrial activities should be reported as absolute quantities in metric tons or kilograms emitted to land. Acid and organic pollutant emissions to land: the total amount emitted in tons per annum of each type of chemical should be reported.

- **Increase recycling rates** (Emissions):
    - Metrics:
        1. Fraction of recycled materials.
        2. Light-weight design.
        3. Reducing yield losses.
        4. Diverting manufacturing scrap.
        5. Re-using components. Longer-life products.
    - KPI:
        1. Years Remaining = Material Scarcity = Stock_global / (Annual Use_global * (1 – Recycling Rate))
        2. Recycling percentage = percentage of waste recycled per total waste

Chapter 4

Models Assessment in Smart Manufacturing Systems

This Chapter focuses on another key topic of my Ph.D. that is the study of how it is possible to adapt popular machine learning techniques to several manufacturing problems.
In Chapter 2.4.4, we provided an overview of the state-of-the-art techniques and tools exploited to manage the vast amount of information that is produced in manufacturing systems. To this end, we focused on the analysis of data-driven model approaches in order to evaluate their performances and advantages in manufacturing applications. In fact, the Big Data advent and the development of several ML techniques, have led to the development of new approaches that allow to build very accurate predictive models, directly exploiting large amounts of data.
In traditional parametric statistics, that is based on the knowledge of the observed physical system, the models are generated on the basis of mathematical description of the phenomenon/system behavior. The choice of the constraints that describe the model, results into a number of free-parameters that define the complexity of the model itself. However, since there is no generalized model, the choice of these parameters is closely linked to the specific phenomenon in question. Moreover, this choice heavily depends on how much knowledge is possessed by those who build the model of the system/phenomenon.
In the ML model training algorithms that are commonly exploited for elaborations on a huge amount of data, usually there are several hyperparameters that need to be set.
The hyperparameters are those parameters, algorithm specific, which need to be set before the learning process begins. By contrast, the values of other parameters are derived via training.
Therefore, an important task to do in order to apply these models in a problem (for example, in the manufacturing context), is to identify the right settings of their hyperparameters value, in order to obtain a model that can predict outputs of new observation, with a greater accuracy.

Here, we introduce a key contribution of this thesis: the models accuracy evaluation through the use of different techniques aimed at identifying the optimal hyperparameters' values of two popular classification algorithms (Support Vector Machine and Random Forest).

As discussed earlier, the problem of the impossibility to have a model that works best for every problem, is argued by the No Free Lunch theorems and it can find an answer in Statistical Learning Theory [OOA15].
SLT is a body of analytics methods commonly used in Big Data problems, that provides a formulation for studying this problem by trying to find necessary and sufficient conditions for non-parametric inference to build effective predictive models from data (or, data driven models) [SSA09]. The main goal of SLT is to find the optimal hyperparameters in order to build the most accurate model directly from the data, without any assumption on the model family nor any other information that is external to the data set itself [Bre01b].
In SLT, the distribution probability is unknown; we can only assume that the relations between the past observations X and the future Y are independent (each new observation brings the maximum information), and identically distributed (all the observations bring informativeness about the underlying phenomenon).
The main objective of SLT is to find the function that best predicts *Y* starting from *X*, or, in other words: find the function that, after being trained on available data, is able to minimize the classification error on new data. Specifically, a learning algorithm can be made up of a set of functions $f(\alpha)$, where α represents the vector of the algorithm's hyperparameters that need to be set. Choosing a vector α, means finding the best tuning of the algorithm's hyperparameters in order to achieve the best accuracy in the prediction [VLS08].
A learning algorithm can be trained with different configurations of its hyperparameters, generating different models. The procedure of selecting the most accurate model in predicting outcome values on previously unseen data, is called Model Selection (MS) ([GE79, GS03]).
SLT provides several methodologies able to rigorously assess the performance and reliability of predictive models ([BBL02, AGOR11, AC10, McA03, K$^+$95, GSDC10]).
In Section 4.1, we will discuss some SLT techniques applied to a classification problem accomplished through the SVM algorithm, for performance assessment and uncertainty quantification of the created model.
Whereas, in Section 4.2 we will examine a model selection procedure for choosing the best set of hyperparameters that allows to build a Random Forest model characterised by the best generalisation performances.

At the end of this Section, we provide a brief description of these two algorithms and their hyperparameters.
SVM and RF are two very popular ML classification algorithms ([CV95, AGOR12c, FDCBA14, Bre01a, Bia12, VGB11]) that are exploited in various manufacturing applications like monitoring machinery conditions, defect and fault diagnosis ([YDH08, CSS$^+$15, CY09, WJT$^+$17, Dem13, XZA$^+$08, WY07, HY09]).

With the term *Support Vector Machine*, we indicate a class of methods for dealing with classification problems and, subsequently extended, to regression and identification of anomalies (outliers). The SVM task, especially as regard classification, is to minimize the risk associated with the classification error, by identifying the optimal hyperplane separator (that is the best separation limits between a series of data). The rightness related to the maximum data separation is represented by the distance that this hyperplane presents from the closest elements (support vectors) that belong to the different classes we are trying to separate.

As a result, the best separation will be associated to that hyperplane which is able to maximize the distance between the hyperplane itself and the resulting support vectors. Finally, it is necessary to consider a constant *C* (soft margin) that indicates the cost index associated with the classification error. This SVM hyperparameter controls the trade-off between maximising the margin and minimising the training error (the ability to classify the training points correctly).
Figure 4.1 depicts how the margin changes respect to different value of *C*. If *C* is high, it indicates that the hyperplane classifies well all the training data but the margin are smaller. If *C* is small, the margin are larger, allowing more errors on the classification of training data. When the observed data are not linearly separable and, therefore, the hyperplane may not exist, the kernel trick is used. This technique involves the use of specific functions (called kernel functions) in order to remap the original space of the observed data, in a larger space where they can be separated through of an appropriate hyperplane. Even the γ parameter of the kernel functions (width of a Gaussian kernel or degree of a polynomial kernel) is another SVM hyperparameter, expressing the model's flexibility.

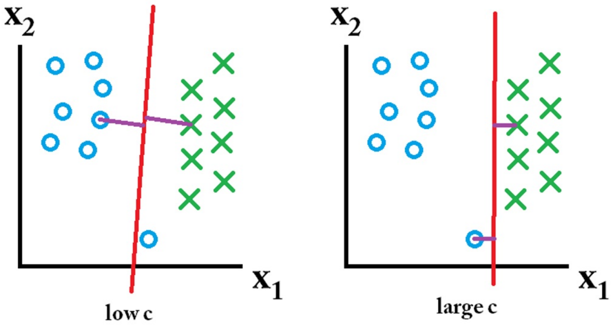


Figure 4.1: Influence of the hyperparameter *soft margin C* in SVM with linear kernel.

As depicted in Figure 4.2, if γ is small, the boundaries of the hyperplane are very smooth, classifying badly the possible non-linear relationships between features. Instead, high values of γ indicates sharpened margins that surround individual instances.

Another well-known tool for classification and regression is the Random Forest algorithm ([RGGR$^+$12, BDL08, LW$^+$02]).

RF combine bagging to random subset feature selection. In bagging, each tree is independently constructed using a bootstrap sample of the dataset. RF add an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data,

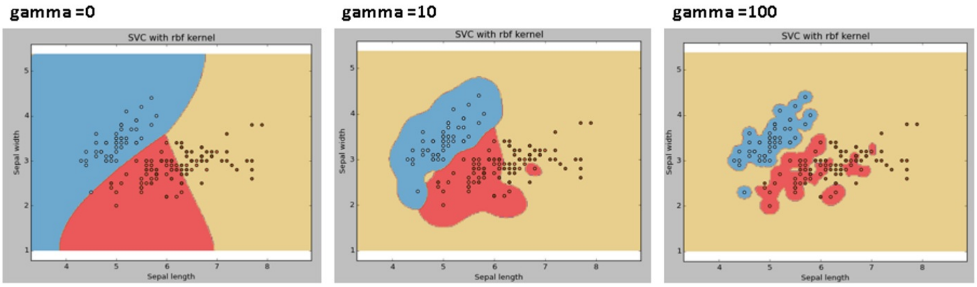
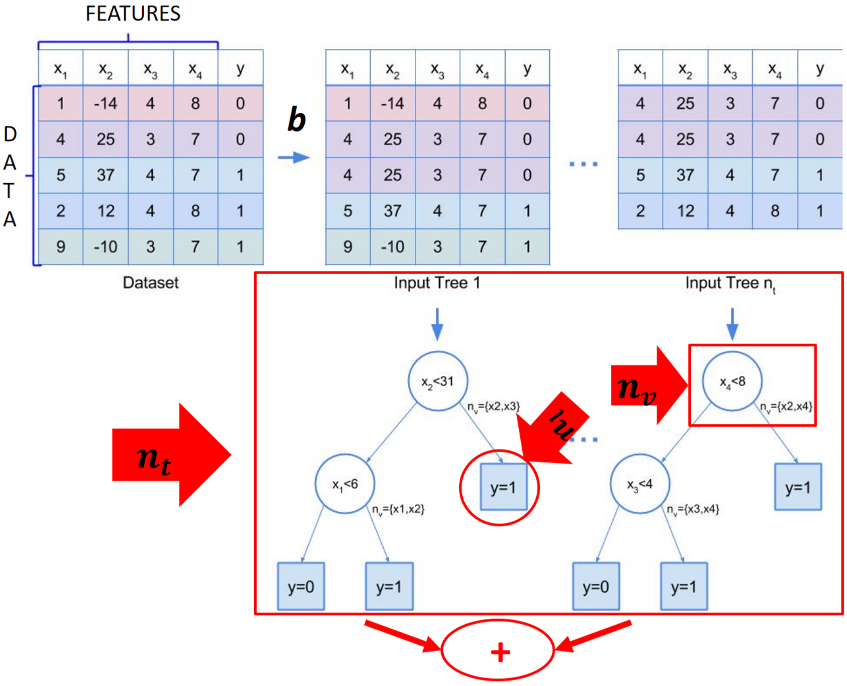


Figure 4.2: Influence of the hyperparameter *kernel coefficient* γ in SVM with non-linear separation.

RF change how the classification trees are constructed. In standard trees, each node is split using the best division among all variables. In a RF, each node is split using the best among a subset of predictors randomly chosen at that node. In the end, a simple majority vote is taken for prediction [OOA16].



x_1	x_2	x_3	x_4	y
1	-14	4	8	0
4	25	3	7	0
5	37	4	7	1
2	12	4	8	1
9	-10	3	7	1

x_1	x_2	x_3	x_4	y
1	-14	4	8	0
4	25	3	7	0
4	25	3	7	0
5	37	4	7	1
9	-10	3	7	1

x_1	x_2	x_3	x_4	y
4	25	3	7	0
4	25	3	7	0
5	37	4	7	1
2	12	4	8	1

Figure 4.3: Random Forest algorithm.

A common misconception about RF is to consider this algorithm as an hyperparameter-free learning algorithm. Examining the reference RF method called Forest-RI, introduced by Breiman, it is possible to identify several hyperparameters which characterise the performance of the final model. Figure 4.3 summarizes the chosen hyperparameters that are: the number of trees (n_t), percentage of input data to sample with replacement from the input data for growing each new tree (b), minimum number of observations per tree leaf (n_l), and the number of features to select at random for each decision split (n_v).

4.1 Performance Assessment and Uncertainty Quantification of Predictive Models for Smart Manufacturing Systems

How can we assess the performance of a predictive model and quantify its uncertainty? This question has received a solid answer from the field of statistical inference since the last century and before [DSP66]. The now classic approach of parametric statistics identifies a family of models (e.g. linear functions), a noise assumption (e.g. Gaussian) and, given some data, easily provides an assessment of the performance of the fitted model, along with a quantification of the uncertainty or, in modern terms, an estimation of the generalization error and the related confidence interval[1]. On the contrary, data-driven models exploit non-parametric inference, where it is expected that an effective model would stem out directly from the data, without any assumption on the model family nor any other information that is *external* to the data set itself [Bre01b]. With the advent of the Big Data era, this approach has gained more and more popularity, up to the point of suggesting that effective predictive models, with the desired accuracy, can be generated by simply collecting more and more data (see, for example, [Dha13] for some insights on this provocative and inexact but, unfortunately, widespread belief).

However, is it really possible to perform statistical inference for building predictive models without any assumption ? Unfortunately, the series of *no-free-lunch* theorems provided a negative answer to this question [Wol96]. They also showed that, in general, is not even possible to solve apparently simpler problems, like differentiating noise from data, no matter how large the data set is [MI00].

Statistical Learning Theory (SLT) addresses exactly this problem, by trying to find necessary and sufficient conditions for non-parametric inference to build predictive models from data [Vap99] or, using the language of SLT, *learn* an optimal model from data. The main SLT results have been obtained by deriving non-asymptotic bounds on the generalization error of a model or, to be more precise, upper bounds on the excess risk between the optimal predictor and the learned model, as a function of the, possibly infinite, family of models and the number of available samples [Vap98].

An important byproduct of SLT has been the (theoretical) possibility of applying these bounds for solving the problems raised by our first question, about the quality and the performance of

[1] We deal in this paper with a frequentist approach, which derives confidence intervals for the quantities of interest, but the credible intervals of the Bayesian approach can be addressed equally well in the parametric setting [Mac92].

the learned model. However, for a long time, SLT has been considered only a theoretical, albeit very sound and deep, statistical framework, without any real applicability to practical problems [Val84]. Only in the last decade, with important advances in this field [BBL02], it has been shown that SLT can provide practical answers, at least when targeting the inference of data-driven models for classification purposes [Lan06, AGOR12a].

We review here the main results of SLT for the purpose of assessing the performance and quantify the uncertainty of a predictive classification model applied to the final product quality estimation in a smart manufacturing system. Our purpose is to provide an overview of the advantages and disadvantages of the SLT framework, which is still an open field of research, but can be the starting point for a better understanding of the methodologies able to *rigorously* assess the performance and reliability of predictive models. We believe that SLT can be one of the core technologies for future Big Data analytics systems, especially when applied to manufacturing systems [LNR14], where, differently from other areas (e.g. social or economical sciences), dealing with uncertainty is a mandatory issue that must be rigorously addressed [NM14].

The proposed case study targets the creation of a predictive model for an assembly line, where the model is required to provide a binary value, in the shortest possible timeframe, which alerts if the final product fails to reach the required quality constraints. Quality evaluation is a fundamental activity within a production process and multi-sensing technologies embedded in production lines collect vast amounts of data, which can be analyzed to predict the final quality of the product [CTF$^+$14]. Good Quantity (i.e. GQ, the ratio between Good Parts and Inspected Parts) has been selected by the ISO 22400 standard[2] as one of the main Key Performance Indicator (KPI) for manufacturing systems, which shows the importance of this figure and the need for predicting it and to eventually start proactive actions on the production line.

The scientific literature abound of applications of data mining and machine learning techniques devoted to building predictive models for quality monitoring in manufacturing systems [JH93, ABT02, KBT11, LLBK13, WIT14], but the assessment of the predictors themselves is quite rare. We believe that SLT could be able to fill this shortcoming, which, in our opinion, is preventing a most widespread application of rigorous predictive analytics techniques to manufacturing systems.

4.1.1 Preliminary Definitions

Since in this work we deal with the problem of detecting a binary value, from now on we will focus on the standard classification problem [Vap99, Bis95, Che97]. Let $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} \in \{\pm 1\}$ be, respectively, the input and the output spaces. We consider a set of labeled independent and identically distributed (i.i.d.) data $\mathcal{D}_n : \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n\}$ of size n, where $\boldsymbol{z}_{i \in \{1,\cdots,n\}} = (\boldsymbol{x}_i, y_i)$,

[2] ISO 22400-2, Automation systems and integration - Key Performance Indicators (KPIs) for manufacturing operations management - Part 2: Definition and descriptions.

sampled from an unknown distribution μ. As we are targeting Big Data problems [WZWD14, Mad12, SS15, OPGD15, FOA15], we will focus on the case where n is very large. For later use we also define two modified training sets: $\mathcal{D}_n^{\setminus i}$ and $\mathcal{D}_n^i$, where, respectively, the i-th element is removed or replaced by another sample [BE02]:

$$\mathcal{D}_n^{\setminus i} : \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{i-1}, \boldsymbol{z}_{i+1}, \cdots, \boldsymbol{z}_n\}, \tag{4.1}$$

$$\mathcal{D}_n^i : \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{i-1}, \boldsymbol{z}_i', \boldsymbol{z}_{i+1}, \cdots, \boldsymbol{z}_n\}. \tag{4.2}$$

A learning algorithm $\mathscr{A}_{\mathcal{H}}$, characterized by a set of hyperparameters $\mathcal{H}$ that must be tuned, maps $\mathcal{D}_n$ into a function $f : \mathscr{A}_{(\mathcal{D}_n,\mathcal{H})}$ from $\mathcal{X}$ to $\mathcal{Y}$. In particular, $\mathscr{A}_{\mathcal{H}}$ allows designing $f \in \mathcal{F}_{\mathcal{H}}$ and the class of functions $\mathcal{F}_{\mathcal{H}}$, that is generally unknown (and depends on $\mathcal{H}$) [BBL02, OGRA14, AGOR12a].

The accuracy of a function $f : \mathscr{A}_{(\mathcal{D}_n,\mathcal{H})}$ in representing the hidden relationship μ is measured with reference to a loss function $\ell(f, \boldsymbol{z}) : \mathcal{F}_{\mathcal{H}} \times (\mathcal{X} \times \mathcal{Y}) \to [0, 1]$ [BE02]. In particular, since we are dealing with binary classification problems, the loss function (called the *hard* loss) simply counts the number of misclassified examples [Vap99, RDVC$^+$04]:

$$\ell_H(f, \boldsymbol{z}) = [y f(\boldsymbol{x}) \leq 0] \in \{0, 1\}. \tag{4.3}$$

The quantity which we are interested in is the generalization error [Vap99, AGOR12a], namely the error that a model will perform on new data generated by μ and previously unseen:

$$L(f) = \mathbb{E}_{\boldsymbol{z}} \ell(f, \boldsymbol{z}). \tag{4.4}$$

Unfortunately, since μ is unknown, $L(\mathscr{A}_{(\mathcal{D}_n,\mathcal{H})})$ cannot be computed and, consequently, must be estimated. Two common empirical estimators are the empirical [Vap99] ($\widehat{L}_{\text{emp}}(\mathscr{A}_{(\mathcal{D}_n,\mathcal{H})}, \mathcal{D}_n)$) and leave-one-out [DGL96] ($\widehat{L}_{\text{loo}}(\mathscr{A}_{(\mathcal{D}_n,\mathcal{H})}, \mathcal{D}_n)$) errors:

$$\widehat{L}_{\text{emp}}(f, \mathcal{D}_n) = \frac{1}{n} \sum_{\boldsymbol{z} \in \mathcal{D}_n} \ell(f, \boldsymbol{z}), \tag{4.5}$$

$$\widehat{L}_{\text{loo}}(f, \mathcal{D}_n) = \frac{1}{n} \sum_{\boldsymbol{z}_{i \in \{1, \cdots, n\}} \in \mathcal{D}_n} \ell(\mathscr{A}_{(\mathcal{D}_n^{\setminus i}, \mathcal{H})}, \boldsymbol{z}_i). \tag{4.6}$$

4.1.2 Building Predictive Models

We choose the Support Vector Machine (SVM) for building our model, since it is one of the most effective [FDCBA14] and popular classification algorithm [CV95, AGOR12a]. The SVM classifier is defined as:

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b \tag{4.7}$$

where the weights $\boldsymbol{w} \in \mathbb{R}^D$ and the bias $b \in \mathbb{R}$ are found by solving the following convex constrained quadratic programming (CCQP) problem:

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \end{aligned} \tag{4.8}$$

where $\xi_i = \max[0, 1 - y_i(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b)]$ is the hinge loss function $\ell_\xi(f, \boldsymbol{z})$, which is an upper bound of the hard loss function. Moreover the Hinge loss function is known to be the best choice in classification problems [RDVC$^+$04, AGOR12b]. The above problem is also known as the Tikhonov formulation of the SVM, because it can be seen as a regularized ill–posed problem [TA77], where the hyperparameter λ is a constant that balances the tradeoff between the accuracy on the data (leading to potential overfitting) and the complexity of the solution (leading to potential underfitting).

By introducing n Lagrange multipliers $\alpha_1, \cdots, \alpha_n$, the problem of Eq. (4.8) can be written in its dual form:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{n} \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{n} y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned} \tag{4.9}$$

where $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j)$ is a suitable kernel function [STC04]. After solving the problem (4.9), the Lagrange multipliers can be used to define the SVM classifier in its dual form:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b. \tag{4.10}$$

Unfortunately, the dual problem, even if it allows the use of the kernels and is the common approach for finding the SVM parameters, becomes intractable, in practice, if n, mostly because of the computation and storage requirements of the Hessian matrix [WHL15]. For this reason, in Big Data applications, SVM is solved in its primal form through stochastic sub-gradient descent (SGD) algorithms [SS15, STS11, ROOA15]. In particular in this paper we will make use of the Pegasos algorithm [SSSSC11] (see Algorithm 1) even if other alternatives exist [CZW$^+$07, FCH$^+$08, YL09]. Note that in the Pegasos algorithm the bias b is set to zero and one more feature to each instance $\boldsymbol{x}$ is added, thus increasing its dimension to $D+1$. The artificially added feature always takes the constant value of 1. The $D+1$ weight of $\boldsymbol{w}$ is the bias and this becomes equivalent to changing the regularization from $\|\boldsymbol{w}\|^2$ term to $\|\boldsymbol{w}\|^2 + b^2$.

Algorithm 1: Pegasos SVM.

```
Input: $\mathcal{D}_n$, $\lambda$ and number if iterations $T$
Output: $\boldsymbol{w}$
Read $\mathcal{D}_n$, $\boldsymbol{w} = 0$;
for $t \leftarrow 1$ to $T$ do
    $\mathcal{I} = \{i : i \in \{1, \cdots, n\}, y_i \boldsymbol{w}^T \boldsymbol{x}_i < 1\}$ ;
    $\eta_t = {}^1/_{\lambda t}$ ;
    $\boldsymbol{w} = (1 - \eta_t \lambda)\boldsymbol{w} + {}^{\eta_t}/_n \sum_{i \in \mathcal{I}} y_i \boldsymbol{x}_i$;
return $\boldsymbol{w}$;
```

We want to stress the fact that the search for the optimal parameters $\boldsymbol{w}$ and b does not complete the learning phase of the SVM: in fact, it is necessary to tune the set of hyperparameters as well, in order to find the SVM characterized by optimal performance. This phase relies on the performance assessment of the model (*model selection* in the machine learning jargon) and is strictly linked with the estimation of the generalization ability of a classifier (e.g. its uncertainty quantification or error estimation phase). In fact, the model selection phase consists in selecting the model characterized by the smallest *estimated* generalization error.

4.1.3 Performance Assessment and Uncertainty Quantification

The selection of the optimal hyperparameters of a predictive model is the fundamental problem of STL and is still the target of current research [AGOR12a, AC10, McA03, K$^+$95, BBL02, GSDC10]. The approaches can be divided in two large families: Out-of-Sample methods, like Hold Out, Cross Validation, and the Bootstrap[AGOR12a, AGOR11, ET93, K$^+$95], and more recent In-Sample methods, like the class of functions-based methods [AGOR12a] (based on the VC-Dimension [Vap98], Rademacher Complexity (RC) [Kol01, OGRA15a, BBM05, OGRA15b], PAC-Bayes Theory [McA98, LLST13]), Algorithm-based methods [OGRA14] (based on Compression Bounds [FW95], and Algorithmic Stability (AS) Theory [BE02, PRMN04]).

Out-of-Sample methods [IK05, AGOR12a] are favoured by practitioners because they work well in many situations and allow the application of simple statistical techniques for estimating the quantities of interest. Some examples of out–of–sample methods are the well–known k–Fold Cross Validation (KCV), the Leave–One–Out (LOO), and the Bootstrap (*BTS*) [ET93, K$^+$95, CYX10]. All these techniques rely on a similar idea: the original dataset is resampled, with or without replacement, to build two independent datasets called, respectively, the training and validation (or estimation) sets. The first one is used for training a classifier, while the second one is exploited for estimating its generalization error, so that the hyperparameters can be tuned to achieve its minimum value. Note that both error estimates computed through the training and validation sets are, obviously, optimistically biased; therefore, if a generalization estimate of the final model is desired, it is necessary to build a third independent set, called the test set, by nesting two of the resampling procedures mentioned above. Furthermore, the resampling procedure itself can introduce artifacts in the estimation process and must be carefully designed.

In-Sample methods [IK05, AGOR12a], instead, allow to exploit the whole set of available data for both training the model and estimating its generalization error, thank to the application of rigorous statistical procedures [BBL02, BE02, LLST13]. Despite their unquestionable advantage with respect to Out-of-Sample methods, their use is not widespread: one of the reasons is the common belief that In-Sample methods are very useful for gaining deep theoretical insights on the learning process or for developing new learning algorithms, but they are not suitable for practical purposes. However, recent advances and deeper insights on these methods demonstrate that this is no longer true [SSBD14].

For more details about the advantaged and disadvantaged of the different methods one can refer to [AGOR12a, OGRA14, Lan06].

4.1.3.1 Out-of-Sample methods

As we described before these techniques rely on a similar idea: the original dataset $\mathcal{D}_n$ is resampled once or many (n_r) times, with or without replacement, to build three independent datasets called, training, validation, and test sets, respectively $\mathcal{L}_l^r$, $\mathcal{V}_v^r$, and $\mathcal{T}_t^r$, with $r \in \{1, \cdots, n_r\}$. Note that $\mathcal{L}_l^r \cap \mathcal{V}_v^r = \oslash$, $\mathcal{L}_l^r \cap \mathcal{T}_t^r = \oslash$, and $\mathcal{V}_v^r \cap \mathcal{T}_t^r = \oslash$. Then, in order to select the best set of hyperparameters $\mathcal{H}$ in a set of possible ones $\mathfrak{H} = \{\mathcal{H}_1, \mathcal{H}_2, \cdots\}$ for the algorithm $\mathscr{A}_{\mathcal{H}}$ or, in other words, to perform the performance assessment, we have to apply the following procedure:

$$\mathcal{H}^* : \min_{\mathcal{H} \in \mathfrak{H}} \frac{1}{n_r} \sum_{r=1}^{n_r} \widehat{L}_{\text{emp}}(\mathscr{A}_{(\mathcal{L}_l^r, \mathcal{H})}, \mathcal{V}_v^r). \tag{4.11}$$

Since the data in $\mathcal{L}_l^r$ are i.i.d. from the one in $\mathcal{V}_v^r$ the idea is that $\mathcal{H}^*$ should be the set of hyperparameters which allows to achieve a small error on a data set that is independent from the training set.

The uncertainty quantification, instead, is performed as follows:

$$\begin{aligned} L(\mathscr{A}_{(\mathcal{D}_n, \mathcal{H}^*)}) \leq & \\ & \frac{1}{n_r} \sum_{r=1}^{n_r} \widehat{L}_{\text{emp}}(\mathscr{A}_{(\mathcal{L}_l^r \cup \mathcal{V}_v^r, \mathcal{H}^*)}, \mathcal{T}_t^r) + \sqrt{\frac{\log(\frac{1}{\delta})}{2t}}, \end{aligned} \tag{4.12}$$

where the bound holds with probability $(1 - \delta)$. Note that after the best set of hyperparameters is found, one can select the best model by training the algorithm with the whole data set $\mathscr{A}_{(\mathcal{D}_n, \mathcal{H}^*)}$ [AGRS09] and since the data in $\mathcal{L}_l^r \cup \mathcal{V}_v^r$ are i.i.d. respect to $\mathcal{T}_t^r$ we have that

$\widehat{L}_{\text{emp}}(\mathscr{A}_{(\mathcal{L}_l^r \cup \mathcal{V}_v^r, \mathcal{H}^*)}, \mathcal{T}_t^r)$ is and unbiased estimator of $L(\mathscr{A}_{(\mathcal{D}_n, \mathcal{H}^*)})$. Then, we can use any concentration result, like the Hoeffding inequality [Hoe63], for bounding the bias between the expected value and its empirical estimator. Note that an approximation has been made: for classifying a new sample we use the function retrained with the best set of hyperparameter over the whole set of data. Theoretically speaking, the rigorous approach would be to randomly select one of the classifiers $\mathscr{A}_{(\mathcal{L}_l^r \cup \mathcal{V}_v^r, \mathcal{H}^*)}$ with $r \in \{1, \cdots, n_r\}$ in order to classify a new sample, but this procedure is usually not taken into account for practical reasons [AGOR12a, AC10, AGRS09]

Note that if $r = 1$, if l, v, and t are aprioristically set such that $n = l + v + t$ and if the the resample procedure is performed without replacement we get the hold out method [AGOR12a]. For implementing the complete k-fold cross validation, instead, we have to set $r \leq \binom{n}{k}\binom{n-\frac{n}{k}}{k}$, $l = (k-2)\frac{n}{k}$, $v = \frac{n}{k}$, and $t = \frac{n}{k}$ and the resampling must be done without replacement [K$^+$95, AC10, AGOR12a]. Finally, for implementing the bootstrap, $l = n$ and $\mathcal{L}_l^r$ must be sampled with replacement from $\mathcal{D}_n$, while $\mathcal{V}_v^r$ and $\mathcal{T}_t^r$ are sampled without replacement from the sample of $\mathcal{D}_n$ that have not been sampled in $\mathcal{L}_l^r$ [ET93, AGOR12a]. Note that for the bootstrap procedure $r \leq \binom{2n-1}{n}$.

It is worthwhile noting that the only hypothesis needed in order to rigorously apply the Out-of-Sample technique is the i.i.d. hypothesis on the data in $\mathcal{D}_n$ and that all these techniques work for any deterministic algorithm.

4.1.3.2 In-Sample methods

For what concern the In-Sample methods, there are two subfamilies of techniques: the class of functions-based ones and the algorithm-based ones [OGRA14]. The difference between the two is that the first ones require the knowledge of $\mathcal{F}_{\mathcal{H}}$ and so they cannot be applied to some algorithms (e.g. the k-nearest neighbour algorithm) while the algorithm-based ones, can be applied to any deterministic algorithm without any additional knowledge. Both subfamilies, like the Out-of-Sample methods, require the i.i.d. hypothesis.

One of the most powerful class of functions-based methods is based on the Rademacher Complexity [Kol01, AGOR12a]. In particular, for any bounded loss function $\ell_b(f, \boldsymbol{z}) \in [0, 1]$ it is possible to prove that the following bound holds with probability $(1 - \delta)$ [BM03]:

$$\begin{aligned} &L(\mathscr{A}_{(\mathcal{D}_n, \mathcal{H})}) \leq \\ &\quad \widehat{L}_{\text{emp}}(\mathscr{A}_{(\mathcal{D}_n, \mathcal{H})}, \mathcal{D}_n) + \widehat{R}_n(\mathcal{F}_{\mathcal{H}}) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}, \end{aligned} \tag{4.13}$$

where

$$\widehat{R}_n(\mathcal{F}_{\mathcal{H}}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell_b(f, \boldsymbol{z}_i), \tag{4.14}$$

$$\sigma_{i \in \{1,\dots,n\}} \in \{\pm 1\}, \quad \mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = \frac{1}{2}.$$

Therefore, based on the Structural Risk Minimization principle [Vap99], one can design a series of function classes of increasing size, $\mathfrak{F} = \{\mathcal{F}_{\mathcal{H}_1}, \mathcal{F}_{\mathcal{H}_2}, \cdots\}$ with $\mathcal{F}_{\mathcal{H}_1} \subseteq \mathcal{F}_{\mathcal{H}_2} \subseteq \cdots$, so to compute at the same time both the performance assessment and the uncertainty quantification:

$$f^*, \mathcal{F}_{\mathcal{H}^*} : \quad L(f^*) \leq \tag{4.15}$$

$$\min_{\mathcal{F}_{\mathcal{H}} \in \mathfrak{F}} \left[\widehat{L}_{\text{emp}}(\mathscr{A}_{(\mathcal{D}_n, \mathcal{H})}, \mathcal{D}_n) + \widehat{R}_n(\mathcal{F}_{\mathcal{H}}) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}} \right].$$

This approach is useful when it is possible to create the hierarchical set of class of functions [AGOR12a]. When this knowledge is not available a different procedure must be adopted [FW95]. In this case to each class of function a probability $p_{\mathcal{F}_{\mathcal{H}}}$ (where $\sum_{i=1}^{\cdots} p_{\mathcal{F}_{\mathcal{H}_i}} = 1$) must be aprioristically assigned. Consequently, the procedure of Eq. (4.15) becomes:

$$f^*, \mathcal{F}_{\mathcal{H}^*} : \quad L(f^*) \leq \tag{4.16}$$

$$\min_{\mathcal{F}_{\mathcal{H}} \in \mathfrak{F}} \left[\widehat{L}_{\text{emp}}(\mathscr{A}_{(\mathcal{D}_n, \mathcal{H})}, \mathcal{D}_n) + \widehat{R}_n(\mathcal{F}_{\mathcal{H}}) + 3\sqrt{\frac{\log(\frac{2}{\delta p_{\mathcal{F}_{\mathcal{H}^*}}})}{2n}} \right].$$

Unfortunately, the quantity of Eq. (4.14) cannot be computed if we do not know $\mathcal{F}_{\mathcal{H}}$ and, moreover, for many algorithms is not even possible to define $\mathcal{F}_{\mathcal{H}}$ [OGRA14]. Algorithmic-based methods circumvent this problem through the concept of Algorithmic Stability [BE02, RMP05, OGRA14]. In particular for any bounded loss function ℓ_b it is possible to prove that the following bounds hold with probability $(1 - \delta)$ [BE02]:

$$L(\mathcal{A}_{(\mathcal{D}_n, \mathcal{H})}) \leq \widehat{L}_{\text{emp}}(\mathcal{A}_{(\mathcal{D}_n, \mathcal{H})}, \mathcal{D}_n) + \sqrt{\frac{1}{2n\delta} + \frac{3\beta_{\text{emp}}}{\delta}}, \tag{4.17}$$

$$L(\mathcal{A}_{(\mathcal{D}_n, \mathcal{H})}) \leq \widehat{L}_{\text{loo}}(\mathcal{A}_{(\mathcal{D}_n, \mathcal{H})}, \mathcal{D}_n) + \sqrt{\frac{1}{2n\delta} + \frac{3\beta_{\text{loo}}}{\delta}}, \tag{4.18}$$

where

$$\beta_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, n) = \mathbb{E}_{\mathcal{D}_n, \boldsymbol{z}'_i} \left| \ell_b(\mathcal{A}_{(\mathcal{D}_n, \mathcal{H})}, \boldsymbol{z}_i) - \ell_b(\mathcal{A}_{(\mathcal{D}^i_n, \mathcal{H})}, \boldsymbol{z}_i) \right|,$$

$$\beta_{\text{loo}}(\mathscr{A}_{\mathcal{H}}, n) = \mathbb{E}_{\mathcal{D}_n, \boldsymbol{z}} |\ell_b(\mathcal{A}_{(\mathcal{D}_n, \mathcal{H})}, \boldsymbol{z}) - \ell_b(\mathcal{A}_{(\mathcal{D}^{\setminus i}_n, \mathcal{H})}, \boldsymbol{z})|.$$

The bounds of Eqns. (4.17) end (4.18) are polynomial bounds in n (so not very tight when n is small) while β_{emp} and β_{loo} are two versions of Hypothesis Stability which are able to take into account both the properties of the algorithm and the property of the distribution that has generated the data $\mathcal{D}_n$ [OGRA14, BE02]. It is possible to improve the bounds of Eqns. (4.17) and (4.18) by exploiting a stronger notion of algorithmic stability, known as the Uniform Stability. In particular, the following bounds hold with probability $(1-\delta)$:

$$L(\mathcal{A}_{(\mathcal{D}_n,\mathcal{H})}) \leq \tag{4.19}$$

$$\widehat{L}_{\text{emp}}(\mathcal{A}_{(\mathcal{D}_n,\mathcal{H})}, \mathcal{D}_n) + 2\beta^i + (4n\beta^i + 1)\sqrt{\frac{\log(\frac{1}{\delta})}{2n}},$$

$$L(\mathcal{A}_{(\mathcal{D}_n,\mathcal{H})}) \leq \tag{4.20}$$

$$\widehat{L}_{\text{loo}}(\mathcal{A}_{(\mathcal{D}_n,\mathcal{H})}, \mathcal{D}_n) + \beta^{\setminus i} + (4n\beta^{\setminus i} + 1)\sqrt{\frac{\log(\frac{1}{\delta})}{2n}},$$

where

$$\beta^i(\mathscr{A}_\mathcal{H}, n) = \left|\ell(\mathcal{A}_{(\mathcal{D}_n,\mathcal{H})}, \cdot) - \ell(\mathcal{A}_{(\mathcal{D}_n^i,\mathcal{H})}, \cdot)\right|_\infty, \tag{4.21}$$

$$\beta^{\setminus i}(\mathscr{A}_\mathcal{H}, n) = |\ell(\mathcal{A}_{(\mathcal{D}_n,\mathcal{H})}, \cdot) - \ell(\mathcal{A}_{(\mathcal{D}_n^{\setminus i},\mathcal{H})}, \cdot)|_\infty. \tag{4.22}$$

Unfortunately the Uniform Stability (β^i or $\beta^{\setminus i}$) is not able to take into account the properties of the distribution that has generated the data $\mathcal{D}_n$ and sometimes in not even able to capture the properties of the algorithm because it deals with a worst–case learning scenario [OGRA14]. Nevertheless, all of the four stability-based bounds of Eqns. (4.17), (4.18), (4.19), and (4.20) can be used to select the best set of hyperparameters $\mathcal{H}$ in a set of possible one $\mathfrak{H} = \{\mathcal{H}_1, \mathcal{H}_2, \cdots\}$ for the algorithm $\mathscr{A}_\mathcal{H}$. In particular all the bounds are in the form: $L(\mathcal{A}_{(\mathcal{D}_n,\mathcal{H})}) \leq \epsilon(\mathscr{A}_\mathcal{H}, \mathcal{D}_n, n, \delta)$ so in order to perform the performance assessment procedure and uncertainty quantification we have to aprioristically assign to each set of hyperparameters a probability $p_\mathcal{H}$ (where $\sum_{i=1}^{\cdots} p_{\mathcal{H}_i} = 1$) of being chosen during the performance assessment procedure. The algorithmic stability-based performance assessment and uncertainty quantification procedure can then be summarized as follows:

$$\begin{aligned} &\mathcal{A}_{(\mathcal{D}_n,\mathcal{H}^*)}, \mathcal{H}^* : \\ &L(\mathcal{A}_{(\mathcal{D}_n,\mathcal{H}^*)}) \leq \min_{\mathcal{H}\in\mathfrak{H}} \epsilon(\mathscr{A}_\mathcal{H}, \mathcal{D}_n, n, \delta p_{\mathcal{H}^*}) \end{aligned} \tag{4.23}$$

The procedure of Eq. (4.23) can be exploited with any algorithm for which it is possible to compute one of the notions of stability.

4.1.4 Computational Issues for Big Data Analytics

Both naive Out-of-Sample and In-Sample methods are computationally expensive when the number of samples is large [GSDC10, AGOR12a]. For this reason, we will focus here on adapting these techniques in the Big Data context.

4.1.4.1 Bag of Little Bootstraps

The standard bootstrap procedure requires, $\forall \mathcal{H} \in \{\mathcal{H}_1, \mathcal{H}_2, \cdots\}$, to train many ($n_r$) models, which is computationally very expensive if n is large. For this reason the Bag of Little Bootstraps approach [KTSJ14, KTSJ12, KTSJ11, OPGD15] represents an alternative to standard Bootstrap, by considering in turn only $b = n^\gamma$ data, with $\gamma \in [{}^1\!/_2, 1]$, in place of the whole dataset during the creation of the Train, Validation and Test sets. Note that $\gamma \in [{}^1\!/_2, 1]$ is necessary to maintain the statistical property of the procedure. In particular, Bag of Little Bootstraps [KTSJ14] consists in sampling $n_r^{\text{no-rep}}$ times from $\mathcal{D}_n$ without replacement, several datasets $\mathcal{B}_b^i$ with $i \in \{1, \cdots, n_r^{\text{no-rep}}\}$ consisting of $b \in [\sqrt{n}, n]$ samples. Then, each $\mathcal{B}_b^i$ is sampled with replacement $n_r^{\text{yes-rep}}$ times, so to derive $\mathcal{L}_n^{i,j}$ datasets with $j \in \{1, \cdots, n_r^{\text{yes-rep}}\}$, each consisting of n samples. All the samples of $\mathcal{D}_n$, or part of them, that have not been sampled in $\mathcal{L}_n^{i,j}$ are used as validation set and test set $\mathcal{V}_v^{i,j} \subseteq \mathcal{D}_n \setminus \mathcal{L}_n^{i,j}$, $\mathcal{T}_t^{i,j} \subseteq \mathcal{D}_n \setminus \mathcal{L}_n^{i,j}$ and $\mathcal{V}_v^{i,j} \cap \mathcal{T}_t^{i,j} = \oslash$. Finally, the models are trained on the sets $\mathcal{L}_n^{i,j}$ and tested on the corresponding $\mathcal{V}_{n_v}^{j,k}$, so to define the following performance assessment procedure:

$$\mathcal{H}^*\colon \min_{\mathcal{H} \in \mathfrak{H}} \frac{1}{n_r^{\text{no-rep}}} \frac{1}{n_r^{\text{yes-rep}}} \sum_{i=1}^{n_r^{\text{no-rep}}} \sum_{j=1}^{n_r^{\text{yes-rep}}} \widehat{L}_{\text{emp}}(\mathscr{A}_{(\mathcal{L}_n^{i,j}, \mathcal{H})}, \mathcal{V}_v^{i,j}). \tag{4.24}$$

Note that, in order to find $\mathcal{H}^*$ with the procedure of Eq. (4.24) we have to train a series of models over sets composed by a maximum of n^γ distinct samples. This means that the model selection strategy, if n is large with respect to $n_r^{\text{no-rep}} n_r^{\text{no-rep}}$ scales with n^γ. Therefore, the procedure scales sub-linearly respect to n, an in the best case scenario scales with $O(\sqrt{n})$. Analogously to the usual Bootstrap procedure the uncertainty quantification is performed as follows:

$$L(\mathscr{A}_{(\mathcal{D}_n, \mathcal{H}^*)}) \leq \tag{4.25}$$
$$\frac{1}{n_r} \sum_{r=1}^{n_r} \widehat{L}_{\text{emp}}(\mathscr{A}_{(\mathcal{L}_n^{i,j}, \mathcal{H}^*)}, \mathcal{T}_t^{i,j}) + \sqrt{\frac{\log(\frac{1}{\delta})}{2t}}.$$

where the best model is obtained by training the algorithm with the whole dataset [AGRS09]:

$$f^* = \mathscr{A}_{(\mathcal{D}_n, \mathcal{H}^*)}. \tag{4.26}$$

Finally, we would like to underline that γ balances the tradeoff between accuracy and computational requirements of the statistical procedure [KTSJ12, OPGD15]. The more $\gamma \to 1$ the better will perform the model selection strategy. Since in this paper we deal with Big Data, we set $\gamma = \frac{1}{2}$. The application to SVM of this approach is straightforward by noting that the only hyperparameter of SVM is $\lambda \in [0, \infty)$.

4.1.4.2 Simplified Rademacher Complexity

We will show here that the Rademacher Complexity of SVM can be easily upper bounded. In particular, let us truncate the hinge loss function such that $\ell_T(f,\boldsymbol{z}) = \min[1, \ell_\xi(f,\boldsymbol{z})]$. It is easy to see that $\ell_H(f,\boldsymbol{z}) \leq \ell_T(f,\boldsymbol{z}) \leq \ell_\xi(f,\boldsymbol{z})$. Consequently, the generalization error computed with $\ell_H(f,\boldsymbol{z})$ is equal or less to the one computed with $\ell_T(f,\boldsymbol{z})$. By exploiting the bound of Eq. (4.13) for $\ell_T(f,\boldsymbol{z})$ the computation of the empirical error is straightforward and it is possible to prove that the Rademacher Complexity can be upper bounded as follows [BM03]:

$$\begin{aligned} \widehat{R}_n(\mathcal{F}_\mathcal{H}) &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell_T(f, \boldsymbol{z}_i) \\ &\leq \frac{1}{n} \sqrt{\sum_{i=1}^{n} (\|\boldsymbol{w}_\lambda^*\|^2 + (b_\lambda^*)^2) \|\boldsymbol{\phi}(\boldsymbol{x}_i)\|^2}, \end{aligned} \tag{4.27}$$

where $(\boldsymbol{w}_\lambda^*, b_\lambda^*)$ is the solution to the SVM problem of Eq. (4.8) for a given λ. Note that λ defines the size of the class of functions in SVM [AGOR12a] and so we can plug this result in the procedure of Eq. (4.15) and obtain a computationally efficient way of assessing the performance and quantify the uncertainty of SVM. In fact, in order to exploit the procedure of Eq. (4.15) we just need to train, for each value of λ, the SVM model and to compute the quantity of Eq. (4.27) which is computationally inexpensive once the SVM has been trained.

4.1.4.3 Simplified Uniform Stability

In this section we want to show how to apply the bound based on the Uniform Stability in the Big Data scenario. The bound of Eq. (4.20), which takes into account the leave-one-out error, is too computational expensive to compute. Instead, we take into account the one of Eq. (4.19). As in Section (4.1.4.2), we will make use of the truncated hinge loss function since for the hard loss function we have trivially that $\beta^i(\mathscr{A}_\mathcal{H}, n) = 1$. Consequently, once the SVM has been trained we can compute the empirical error with the truncated hinge loss. Computing $\beta^i(\mathscr{A}_\mathcal{H}, n)$ is not easy but, thanks to the result of [BE02], it is possible to upper bound it in the case of SVM as follows:

$$\begin{aligned} \beta^i(\mathscr{A}_\mathcal{H}, n) &= \left| \ell(\mathcal{A}_{(\mathcal{D}_n,\mathcal{H})}, \cdot) - \ell(\mathcal{A}_{(\mathcal{D}_n^i,\mathcal{H})}, \cdot) \right|_\infty \\ &\leq \frac{\max\{\|\boldsymbol{\phi}(\boldsymbol{x}_1)\|^2, \cdots, \|\boldsymbol{\phi}(\boldsymbol{x}_n)\|^2\}}{n\lambda}. \end{aligned} \tag{4.28}$$

Then the application of the procedure of Eq. (4.23) to SVM becomes straightforward and computationally inexpensive. From Eq. (4.28) it seems that the Uniform Stability does not take into account the distribution of the data but only the property of SVM through λ. However, Eq. (4.28) holds for most of regularization based algorithms even if the loss is not the hinge one. In other words the Uniform Stability upper bound of Eq. (4.28) is not able to capture the effect of changing the loss.

4.1.4.4 Bag of Little Hypothesis Stabilities

In order to overcome the issues of the Uniform Stability we exploit the proposal of [OGRA14] in order to estimate the Hypothesis Stability instead. As we will see, this proposal is also well suited for Big Data applications. As for the Uniform Stability, we do not consider the bound of Eq. (4.18) since it is too computationally expensive. Consequently we take into account the bound of Eq. (4.17). In this case we do not need to exploit the truncated hinge loss function, but we can exploit directly the hard loss function once the SVM model has been trained with a fixed value of λ. In order to compute the bound of Eq. (4.17), and being able to perform the procedure of Eq. (4.28), we need to compute $\beta_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, n)$. We start by making an assumption on the learning algorithm $\mathscr{A}_{\mathcal{H}}$. In particular, we suppose that the Hypothesis Stability does not increase with the cardinality of the training set:

$$\beta_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, n) \leq \beta_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, \sqrt{n} - 1). \tag{4.29}$$

We point out that this property is a desirable requirement for any learning algorithm since in order to be abel to prove the learnability in the stability framework we need that:

$$\lim_{n \to \infty} \beta_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, n) = 0, \tag{4.30}$$

or, in other words, that the impact on the learning procedure of removing or replacing one sample from $\mathcal{D}_n$ should decrease, on average, as n grows. Note also that this property has already been studied by many researchers in the past. In particular, this property is related to the consistency concept [DGL96]. However, connections can also be identified with the trend of the learning curves of an algorithm [DOS99]. Moreover, such quantity is also strictly linked to the concept of Smart Rule [DGL96]. It is worth underlining that, in many of the above-referenced works, the property of Eq. (4.29) is proved to be satisfied by many well known algorithms (Support Vector Machines, Kernelized Regularized Least Squares, k–Local Rule with $k > 1$, etc.).

Let us define $\widehat{\beta}_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, n, \mathcal{D}_n)$ which is an unbiased estimators of $\beta_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, n)$:

$$\begin{aligned} &\widehat{\beta}_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, \sqrt{n} - 1, \mathcal{D}_n) \qquad (4.31) \\ &= \frac{1}{\sqrt{n}} \sum_{k=1}^{\sqrt{n}} \left| \ell(\mathscr{A}_{\breve{\mathcal{D}}^k_{\sqrt{n}-1}}, \breve{\boldsymbol{z}}^k_i) - \ell(\mathscr{A}_{(\breve{\mathcal{D}}^k_{\sqrt{n}-1})^i}, \breve{\boldsymbol{z}}^k_i) \right|, \end{aligned}$$

note that $\forall k \in \{1, \dots, \sqrt{n}\}$ while $i \in \{1, \dots, \sqrt{n} - 1\}$:

$$\breve{\mathcal{D}}^k_{\sqrt{n}-1} : \left\{ \boldsymbol{z}_{(k-1)\sqrt{n}+1}, \dots, \boldsymbol{z}_{(k-1)\sqrt{n}+\sqrt{n}-1} \right\}, \tag{4.32}$$

$$(\breve{\mathcal{D}}^k_{\sqrt{n}-1})^i : \left\{ \boldsymbol{z}_{(k-1)\sqrt{n}+1}, \dots, \boldsymbol{z}_{(k-1)\sqrt{n}+\sqrt{n}-1} \right\}, \tag{4.33}$$

$$\breve{\boldsymbol{z}}^k_i : \boldsymbol{z}_{(k-1)\sqrt{n}+i}. \tag{4.34}$$

Clearly:

$$\mathbb{E}_{\mathcal{D}_n}\widehat{\beta}_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, \sqrt{n}-1, \mathcal{D}_n) = \beta_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, \sqrt{n}-1). \tag{4.35}$$

Since all the quantities in the summations of Eq. (4.31) are $\{\pm 1\}$ valued i.i.d. random variable (since they are computed over different sets of data) extracted from a Bernoulli distribution of mean $\beta_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, \sqrt{n}-1)$, we have that the following bound holds [Hoe63] with probability $(1-\delta)$:

$$\begin{aligned}&\beta_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, \sqrt{n}-1) \leq \\ &\quad \widehat{\beta}_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, \sqrt{n}-1, \mathcal{D}_n) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2\sqrt{n}}}.\end{aligned} \tag{4.36}$$

Note that plugging Eq. (4.36) into the bound of Eq. (4.17) gives rise to a fully empirical bound where all the quantities can be computed from the data [OGRA14]. In particular, once the SVM is trained for a given λ, the empirical error, computed with the hard loss function, is trivially computable while $\widehat{\beta}_{\text{emp}}(\mathscr{A}_{\mathcal{H}}, \sqrt{n}-1, \mathcal{D}_n)$ require to train many SVM on a small subset of the data ($\sqrt{n}$) which is computationally inexpensive. Moreover, all these SVMs can be trained in parallel (see Eq. (4.31)). Then the application of the procedure of Eq. (4.23) to SVM becomes straightforward. Note that, from Eq. (4.31), the hypothesis stability is able to capture both the property of the algorithm and the property of the distribution that has generated the data [OGRA14].

4.1.5 A Case Study on a Manufacturing Assembly Line

We target an assembly line for refrigerators, where the last steps of the assembly process are: the manual mounting of the compressor, the welding of the corresponding input and output pipes of the refrigerating system, and the gas loading. There are several ways that this process can potentially result in a defective product: a wrong gas load, the mounting of a wrong compressor, the mounting of an incorrect evaporator, or the creation of an obstruction at the welding sites.

In order to check for the quality of the final product, it would be necessary to let it switched on for a long period of time until it reaches the expected temperature. This is, however, a very time consuming process, that can be circumvented by building a predictive model able to discriminate between good and defective products. The input to the model is simply the measurements, during the test time interval, of the input and output temperatures of the gas flowing in the compressor and the power consumption of the refrigerator. Obviously, the more accurate is the model, the less time is needed for obtaining a correct estimate of the absence or presence of a mounting defect.

The dataset consists in more than hundred thousand measurement taken from compressors for 30 seconds with a time interval of 1 second. Consequently we have that $d = 90$. We split the

Table 4.1: Result of the performance assessment, uncertainty quantification and computational requirements of the different methods.

T	5 sec				10 sec				15 sec				20 sec				25 sec				30 sec			
n	BLB	RC	US	HS	BLB	RC	US	HS	BLB	RC	US	HS	BLB	RC	US	HS	BLB	RC	US	HS	BLB	RC	US	HS
Results of the performance assessment for the different methods.																								
10^1	$68_{\pm 14}$	$80_{\pm 16}$	$81_{\pm 16}$	$\mathbf{58_{\pm 13}}$	$53_{\pm 14}$	$63_{\pm 16}$	$64_{\pm 18}$	$\mathbf{48_{\pm 13}}$	$\mathbf{40_{\pm 12}}$	$53_{\pm 13}$	$52_{\pm 14}$	$42_{\pm 11}$	$40_{\pm 11}$	$47_{\pm 13}$	$44_{\pm 14}$	$\mathbf{37_{\pm 9.3}}$	$40_{\pm 8.5}$	$40_{\pm 9.6}$	$42_{\pm 11}$	$\mathbf{35_{\pm 7.8}}$	$31_{\pm 8.3}$	$41_{\pm 9.4}$	$41_{\pm 9.1}$	$\mathbf{28_{\pm 7.1}}$
10^2	$23_{\pm 7.5}$	$29_{\pm 8.5}$	$29_{\pm 9.3}$	$\mathbf{22_{\pm 7.4}}$	$23_{\pm 7.8}$	$26_{\pm 9.5}$	$26_{\pm 8.7}$	$\mathbf{20_{\pm 7.2}}$	$18_{\pm 7.1}$	$22_{\pm 8.7}$	$24_{\pm 9.0}$	$\mathbf{18_{\pm 6.4}}$	$16_{\pm 6.6}$	$19_{\pm 7.3}$	$19_{\pm 7.6}$	$\mathbf{14_{\pm 5.6}}$	$13_{\pm 5.6}$	$16_{\pm 6.9}$	$16_{\pm 7.0}$	$\mathbf{11_{\pm 5.6}}$	$13_{\pm 5.3}$	$15_{\pm 7.2}$	$14_{\pm 6.6}$	$\mathbf{12_{\pm 5.2}}$
10^3	$16_{\pm 5.0}$	$17_{\pm 5.5}$	$17_{\pm 5.7}$	$\mathbf{14_{\pm 4.4}}$	$15_{\pm 4.6}$	$17_{\pm 5.4}$	$16_{\pm 5.3}$	$\mathbf{13_{\pm 4.6}}$	$\mathbf{11_{\pm 4.3}}$	$14_{\pm 4.9}$	$14_{\pm 4.8}$	$11_{\pm 3.7}$	$8.7_{\pm 3.4}$	$10_{\pm 4.0}$	$11_{\pm 3.9}$	$\mathbf{8.3_{\pm 2.9}}$	$\mathbf{6.7_{\pm 3.4}}$	$7.9_{\pm 4.2}$	$7.9_{\pm 3.9}$	$\mathbf{6.7_{\pm 3.1}}$	$5.9_{\pm 3.5}$	$6.6_{\pm 4.2}$	$7.0_{\pm 4.5}$	$\mathbf{5.0_{\pm 3.4}}$
10^4	$8.7_{\pm 3.5}$	$9.9_{\pm 4.2}$	$9.5_{\pm 4.0}$	$\mathbf{7.3_{\pm 3.1}}$	$7.7_{\pm 3.1}$	$8.9_{\pm 3.6}$	$8.5_{\pm 3.9}$	$\mathbf{7.4_{\pm 2.9}}$	$6.7_{\pm 2.7}$	$7.7_{\pm 3.2}$	$7.5_{\pm 3.5}$	$\mathbf{5.6_{\pm 2.6}}$	$4.8_{\pm 2.5}$	$6.3_{\pm 2.8}$	$5.6_{\pm 3.0}$	$\mathbf{4.6_{\pm 2.2}}$	$4.8_{\pm 2.1}$	$5.7_{\pm 2.9}$	$5.1_{\pm 2.8}$	$\mathbf{4.3_{\pm 2.1}}$	$\mathbf{3.1_{\pm 2.3}}$	$3.8_{\pm 2.7}$	$4.0_{\pm 2.6}$	$3.2_{\pm 2.0}$
10^5	$5.0_{\pm 2.4}$	$5.7_{\pm 2.9}$	$5.8_{\pm 2.6}$	$\mathbf{4.2_{\pm 2.2}}$	$4.6_{\pm 2.4}$	$5.3_{\pm 2.9}$	$5.4_{\pm 3.0}$	$\mathbf{3.9_{\pm 2.0}}$	$3.6_{\pm 2.2}$	$4.3_{\pm 2.4}$	$4.5_{\pm 2.6}$	$\mathbf{3.1_{\pm 2.1}}$	$3.0_{\pm 2.2}$	$3.5_{\pm 2.5}$	$3.6_{\pm 2.4}$	$\mathbf{2.5_{\pm 2.0}}$	$\mathbf{2.7_{\pm 2.1}}$	$3.4_{\pm 2.3}$	$3.3_{\pm 2.4}$	$2.8_{\pm 1.9}$	$2.4_{\pm 1.9}$	$2.9_{\pm 2.1}$	$2.7_{\pm 2.3}$	$\mathbf{1.9_{\pm 1.7}}$
Results of the uncertainty quantification for the different methods. Blank space indicates that the estimation is over 100%																								
10^1	-	-	-	-	-	-	-	-	$\mathbf{95_{\pm 28}}$	-	-	-	$\mathbf{93_{\pm 25}}$	-	-	-	$\mathbf{95_{\pm 20}}$	-	-	$99_{\pm 22}$	$\mathbf{74_{\pm 20}}$	-	-	$80_{\pm 20}$
10^2	$\mathbf{29_{\pm 9.5}}$	$58_{\pm 18}$	$59_{\pm 19}$	$41_{\pm 13}$	$\mathbf{28_{\pm 9.7}}$	$52_{\pm 20}$	$53_{\pm 18}$	$36_{\pm 13}$	$\mathbf{22_{\pm 8.7}}$	$44_{\pm 18}$	$49_{\pm 19}$	$32_{\pm 12}$	$\mathbf{19_{\pm 8.1}}$	$38_{\pm 15}$	$39_{\pm 15}$	$26_{\pm 10}$	$\mathbf{16_{\pm 7.2}}$	$32_{\pm 14}$	$32_{\pm 14}$	$21_{\pm 10}$	$\mathbf{16_{\pm 6.6}}$	$31_{\pm 15}$	$29_{\pm 13}$	$22_{\pm 9.6}$
10^3	$18_{\pm 5.4}$	$23_{\pm 7.4}$	$23_{\pm 7.4}$	$\mathbf{18_{\pm 5.3}}$	$17_{\pm 4.9}$	$23_{\pm 7.3}$	$22_{\pm 7.2}$	$\mathbf{16_{\pm 5.8}}$	$\mathbf{12_{\pm 4.5}}$	$18_{\pm 6.6}$	$18_{\pm 6.6}$	$14_{\pm 4.6}$	$\mathbf{9.6_{\pm 3.7}}$	$14_{\pm 5.4}$	$15_{\pm 5.4}$	$10_{\pm 3.7}$	$\mathbf{7.2_{\pm 3.7}}$	$11_{\pm 5.5}$	$11_{\pm 5.4}$	$8.5_{\pm 3.8}$	$6.3_{\pm 3.7}$	$8.9_{\pm 5.5}$	$9.2_{\pm 6.3}$	$\mathbf{6.2_{\pm 4.3}}$
10^4	$9.3_{\pm 3.6}$	$11_{\pm 4.7}$	$11_{\pm 4.6}$	$\mathbf{8.3_{\pm 3.5}}$	$\mathbf{8.1_{\pm 3.3}}$	$9.9_{\pm 4.2}$	$9.9_{\pm 4.4}$	$8.7_{\pm 3.2}$	$7.3_{\pm 2.8}$	$8.7_{\pm 3.6}$	$8.3_{\pm 4.3}$	$\mathbf{6.3_{\pm 3.0}}$	$5.4_{\pm 2.7}$	$7.2_{\pm 3.2}$	$6.8_{\pm 3.5}$	$\mathbf{5.1_{\pm 2.4}}$	$5.2_{\pm 2.4}$	$6.7_{\pm 3.3}$	$5.9_{\pm 3.2}$	$\mathbf{5.0_{\pm 2.4}}$	$\mathbf{3.5_{\pm 2.4}}$	$4.5_{\pm 3.0}$	$4.5_{\pm 3.0}$	$3.7_{\pm 2.3}$
10^5	$5.5_{\pm 2.4}$	$6.5_{\pm 3.2}$	$6.8_{\pm 3.1}$	$\mathbf{4.4_{\pm 2.3}}$	$4.8_{\pm 2.5}$	$5.9_{\pm 3.4}$	$6.0_{\pm 3.6}$	$\mathbf{4.3_{\pm 2.2}}$	$3.7_{\pm 2.3}$	$4.8_{\pm 2.9}$	$5.2_{\pm 2.9}$	$\mathbf{3.4_{\pm 2.3}}$	$3.1_{\pm 2.2}$	$4.0_{\pm 2.8}$	$4.0_{\pm 2.8}$	$\mathbf{2.9_{\pm 2.1}}$	$\mathbf{2.9_{\pm 2.2}}$	$3.8_{\pm 2.6}$	$3.9_{\pm 2.7}$	$3.0_{\pm 2.1}$	$2.5_{\pm 2.0}$	$3.4_{\pm 2.6}$	$3.0_{\pm 2.6}$	$\mathbf{2.2_{\pm 1.9}}$
Computational requirements of the different methods (in minutes). Blank space indicates that the time is negligible																								
10^1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
10^2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$
10^3	-	-	-	-	-	-	-	-	-	-	-	-	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$0.2_{\pm 0.1}$	$\mathbf{0.1_{\pm 0.1}}$	$0.2_{\pm 0.1}$	$0.2_{\pm 0.1}$
10^4	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$\mathbf{0.1_{\pm 0.1}}$	$0.5_{\pm 0.1}$	$\mathbf{0.4_{\pm 0.1}}$	$\mathbf{0.4_{\pm 0.1}}$	$0.5_{\pm 0.1}$	$0.9_{\pm 0.2}$	$\mathbf{0.7_{\pm 0.2}}$	$\mathbf{0.7_{\pm 0.2}}$	$0.9_{\pm 0.2}$	$1.1_{\pm 0.3}$	$1.2_{\pm 0.2}$	$\mathbf{1.0_{\pm 0.3}}$	$1.4_{\pm 0.3}$	$2.1_{\pm 0.3}$	$1.9_{\pm 0.3}$	$\mathbf{1.7_{\pm 0.3}}$	$2.3_{\pm 0.4}$	$2.2_{\pm 0.4}$	$\mathbf{2.0_{\pm 0.3}}$	$2.2_{\pm 0.3}$	$2.9_{\pm 0.4}$
10^5	$2.7_{\pm 0.4}$	$\mathbf{2.3_{\pm 0.3}}$	$2.4_{\pm 0.3}$	$2.7_{\pm 0.4}$	$5.0_{\pm 0.5}$	$\mathbf{4.2_{\pm 0.4}}$	$4.4_{\pm 0.5}$	$5.1_{\pm 0.5}$	$7.1_{\pm 0.4}$	$\mathbf{6.5_{\pm 0.3}}$	$6.9_{\pm 0.4}$	$7.6_{\pm 0.5}$	$8.8_{\pm 0.5}$	$\mathbf{7.5_{\pm 0.5}}$	$7.7_{\pm 0.4}$	$8.8_{\pm 0.6}$	$10_{\pm 0.6}$	$9.8_{\pm 0.5}$	$\mathbf{9.0_{\pm 0.5}}$	$13_{\pm 0.6}$	$15_{\pm 0.5}$	$13_{\pm 0.4}$	$\mathbf{12_{\pm 0.5}}$	$14_{\pm 0.6}$

data in two different sets, one (of size $n \in \{10^1, 10^2, 10^3, 10^4, 10^5\}$) for performing the learning, the performance assessment and uncertainty quantification and one (with all the remaining samples) as a reference set for testing the quality of the final model by computing its empirical error with the hard loss function. Since it is very important to detect as soon as possible the presence of faulty pieces we built models which take into account different time horizons $T \in \{5, 10, 15, 20, 25, 30\}$ seconds. This basically means that we have 6 different learning problems with six different dimensionalities $d \in \{15, 30, 45, 60, 75, 90\}$. We exploit the SVM with a linear formulation (or in other words $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$) and we chose $\lambda \in [10^{-4}, 10^4]$ through 60 points equally spaced on a logarithmic scale. Then we exploit the four procedures depicted in Section 4.1.4:

- Bag of Little Bootstraps (BLB) with $\gamma = \frac{1}{2}$, $n_r^{\text{no-rep}} = 30$ and $n_r^{\text{yes-rep}} = 30$
- Simplified Rademacher Complexity (RC),
- Simplified Uniform Stability (US),
- Bag of Little Hypothesis Stabilities (HS).

We run the experiment through the use of the Google Cloud Platform, in particular we used the Google Compute Engine Platform with 4 machines with 4 cores and 26GB of RAM (machine type n1-highmem-4) [Goo15] equipped with the Hadoop and Spark [SS15] software platforms. We performed the experiment 30 times in order to average the results. In Table 4.1 we report several quantities for each procedure and by varying n and T:

- the error of the final model over the reference set,
- the estimated accuracy of the model,
- the time needed to build the model.

The best result for each n and T is marked in bold. From the results it is possible to note that:

- HS is the best method for assessing the performance of the model since it is the one that more often selects the most accurate model according to the reference set. Note also that BLB performs well, while RC and US are the worst ones.
- BLB is the best method for predicting the uncertainty of the model since it is the one that is able to better estimate the true accuracy of the final model. The other methods are much more conservative.
- RD and US are the most computationally saving methods, while the method that is more computational demanding is HS (which is also the most accurate one).

Note that all the methods perform quite well in practice and reach similar performance when n is large and, at the same time, they are almost equally computational expensive.

4.1.6 Conclusion

We reviewed in this work some methods from Statistical Learning Theory for the performance assessment and uncertainty quantification of predictive models. Differently from other statistical inference frameworks, SLT implements a worst-case approach to these problems, which allows obtaining rigorous and consistent generalization bounds. Thank to recent advances, as presented in this paper, the computational requirements of these methods have been improved so to allow scaling to the large datasets, which are typical of Big Data applications. We focused here on the application of SLT to a predictive model for a smart manufacturing system consisting of a binary classifier, but SLT has proved to be an effective approach in many other areas like multi-class classification, regression, density estimation, etc. Furthermore, SLT can be applied to other settings, which are interesting for manufacturing systems. One, for example, is the semi-supervised setting, where only few labelled samples are available, while the largest part of the dataset is composed of unlabelled data [CSZ10, AGOR14]. When the collection of labels is expensive or difficult, this approach allows exploiting all the available information for building effective models. Another interesting setting is when physical models of the phenomenon under exam are available: SLT can provide rigorous results about the final predicting model, even when the physical model and the data-driven one are combined [COBA15a]. In summary, we believe that SLT can provide a large set of effective tools to the field of smart manufacturing systems for building effective predictive models and rigorously assessing their quality.

4.2 Random Forest Model Selection

It is well known that combining the output of several classifiers results in a much better performance than using any one of them alone. In fact many state-of-the-art algorithms search for a weighted combination of simpler classifiers [LLST13]: Bagging [Bre01a], Boosting [SFBL98] and Bayesian approaches [GCSR14] or even Neural Networks (NN) [Bis95] and Kernel methods as Support Vector Machines (SVM) [Vap98]. Optimising the generalisation performance of the final model represents still an outstanding unsolved problem: How do we build these simple classifiers? How many simple classifiers do we have to combine? How can we combine them? Is there any theory which can support us in making a choice?

In [Bre01a] Breiman has tried to give an answer to these questions by proposing the Random Forests (RF) of tree classifiers, one of the state-of-the-art algorithm for classification which has shown to be probably the most effective tool in this context [FDCBA14]. RF combine bagging to random subset feature selection. In bagging, each tree is independently constructed using a bootstrap sample of the dataset. RF add an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, RF change how the classification trees are constructed. In standard trees, each node is split using the best division among all variables. In a RF, each node is split using the best among a subset of predictors randomly chosen at that node. In the end, a simple majority vote is taken for prediction. In [Bre01a] it is shown that the accuracy of the final model depends mainly on three different factors: how many trees compose the forest, the accuracy of each tree and the correlation between them. The accuracy for RF converges to a limit as the number of trees in the forest becomes greater, while it rises as the accuracy of each tree increases and the correlation between them decreases. RF counterintuitive learning strategy turns out to perform very well compared to many other classifiers, including NN and SVM, and is robust against overfitting [Bre01a, FDCBA14].

A common misconception about RF is to consider this algorithm as an hyperparameter-free learning algorithm [Bia12, BHA09]. In fact, there are several hyperparameters which characterise the performance of the final model: how many trees do we learn; how many data do we sample with replacement during the bootstrap resampling procedure; how is the growth of each tree in the forest; how many predictors do we use in each subset during the splitting procedure and how do we weigh each classifier? For this reason we will show that a Model Selection (MS) procedure is needed in order to select the set of hyperparameters [AGOR12a] which allows to built a RF model characterised by the best generalisation performances. Results on a series of benchmark datasets show that an accurate MS procedure over these hyperparameters can remarkably improve the accuracy of the final RF model.

4.2.1 Hyperparameters in Random Forest

Let us recall the multi-class classification problem [Bis95] where a set of labeled samples $\mathcal{D}_n = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ draw i.i.d. according to an unknown probability distribution μ over $\mathcal{X} \times \mathcal{Y}$ are available and where $X \in \mathcal{X}$ and $Y \in \mathcal{Y} = \{1, 2, \cdots, c\}$. A learning algorithm $\mathscr{A}$ maps $\mathcal{D}_n$ into a function belonging to a possibly unknown set of functions $f \in \mathcal{F}$ according to some criteria $\mathscr{A} : \mathcal{D}_n \to \mathcal{F}$. The error of f in approximating μ is measured with reference to a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Since we are dealing with classification problems we chose the loss function which counts the number of misclassified samples $\ell(f(X), Y) = [f(X) \neq Y]$. The expected error of f in representing μ is called generalization error [Vap98] and it is defined as $L(f) = \mathbb{E}_{(X,Y)}\ell(f(X), Y)$. Since μ is unknown $L(f)$ cannot be computed, but we can compute its empirical estimator, the empirical error, defined as $\widehat{L}(f) = {}^1\!/_n \sum_{i=1}^{n} \ell(f(X_i), Y_i)$. The RF learning and classification phases are reported in Algorithm 2. The learning phase of each of the n_t tree composing the RF is quite simple. From $\mathcal{D}_n$ $\lfloor bn \rfloor$ samples, these are sampled with replacement and $\mathcal{D}'_{\lfloor bn \rfloor}$ is built. A tree is constructed with $\mathcal{D}'_{\lfloor bn \rfloor}$ but the best split is chosen among a subset of n_v predictors over the possible d predictors randomly chosen at each node. The tree is grown until the node contains a maximum of n_l samples. During the classification phase of a previously unseen X, each tree classifies X in a class $Y_{i\in\{1,\cdots,n_t\}}$, and then the final classification is the $\{p_1, \cdots, p_{n_t}\}$-weighted combination of all the answers of each tree of the RF. If $b = 1$, $n_v = \sqrt{n}$, $n_l = 1$ and $p_{i\in\{1,\cdots,n_t\}} = 1$ we get the original RF formulation [Bre01a] where n_t is usually chosen to tradeoff accuracy and efficiency [HLMMS13] or based on the out-of-bag estimate [Bre01a] or according to some consistency result [HLMMS13].

In this paper we argue that performing a MS procedure over b, n_v and n_l and choosing a different weighting procedure can remarkably improve the accuracy of the final model. In particular we exploit the Bootstrap (BOO) MS procedure [AGOR12a] in order to select the best values for b, n_v and n_l. The smaller b, n_v and n_l are the more independent are the trees in the RF but also the lower will the accuracy be of each one of the trees in the RF. Obviously there is a tradeoff which produces and optimal RF classifier. Besides b, n_v and n_l the weight $p_{\{i\in 1,\cdots,n_t\}}$ are of paramount importance for the accuracy of an ensemble classifier [LLST13, GLLF15] and for this reason we will compare the original choice of [Bre01a] (W_1) with other two state of the art alternatives. One (W_2) is due to [NP82] and recently studied in [BK14] while the other one (W_3) has been proposed in [Cat07] and recently further developed in [LLST13]. For what concerns W_2 the proposal is to set $p_i = \ln\left({(1 - L(T_i))}/{L(T_i)}\right)$ where $L(T_i)$ is the accuracy of the tree T_i. Since $L(T_i)$ is unknown, we substitute the empirical error $\widehat{L}(T_i)$ over the out of bag estimate (which is an unbiased estimator of $L(T_i)$). Instead, for what concerns W_3, we have that $p_i = e^{-\gamma\widehat{L}(T_i)}$ where γ is an hyperparameter which must be set as b, n_v and n_l according to the BOO MS procedure. Note that the study of n_t has been deeply investigated in [HLMMS13] so in this paper we consider it fixed and we study the effect of the other hyperparameters for a given value of n_t.

4.2.2 Results and Discussion

Let us consider a series of biclass and multiclass problems from [Lic13]: BanknoteAuth (D1), Anneal (D2), Parkinson (D3), Wine (D4), Seed (D5), Tic-tac-toe (D6), Car (D7), LSVT (D8), Fertility (D9), Horse (D10), Blogger (D11), Nursey (D12), Segment (D13), HAR (D14), Mice-Proteins (D15), Audiology (D16), CNAE (D17), Glass (D18), SensorlessDrive (D19), Optdigits (D20), BreastTissue (D21), MovementsLibras (D22), PittsburgBridges (D23), Bach (D24), Cmc (D25), and Yeast (D26). These datasets are commonly used as a benchmark for learning algorithms. For each dataset, analogously to [GLLF15], up to a maximum of 500 samples are randomly chosen to be in the training set $\mathcal{D}_n$, and the remaining examples are kept as test set if it is not already available. We set $n_t = 100$ and for the BOO MS procedure we set the number of bootstrap resamples $n_b = 100$ [AGOR12a]. We compare the following RF models:

- RF with the weighting strategy W_1 ($p_{i\in\{1,\cdots,n_t\}} = 1$ which is the majority vote [Bre01a]) in the following cases:
  - STD: we set b, n_v, and n_l at the standard values $b = 1$, $n_v = \sqrt{n}$, and $n_l = 1$ [Bre01a];
  - O(b): we set $n_v = \sqrt{n}$ and $n_l = 1$ while $b \in \{0.20, 0.22, \cdots, 1.20\}$ is optimised

Algorithm 2: RF learning and classification phases.

```
/* Learning phase */
Input: $\mathcal{D}_n$, $n_t$, $b$, $n_v$ and $n_l$
Output: A set of tree $\{T_1, \cdots, T_{n_t}\}$
for $i \leftarrow 1$ to $n_t$ do
    $\mathcal{D}'_{\lfloor bn \rfloor}$ sample with replacement $\lfloor bn \rfloor$ sample from $\mathcal{D}_n$;
    $T_i = \mathbf{DT}(\mathcal{D}'_{\lfloor bn \rfloor}, n_v, n_l)$;
/* Classification phase */
Input: $X$, $n_t$, $\{p_1, \cdots, p_{n_t}\}$
Output: $Y$
for $i \leftarrow 1$ to $n_t$ do
    $Y_i = T_i(X)$;
$Y = \arg\max_{j \in \{1, \cdots, c\}} \sum_{i \in \{1, \cdots, n_t\} : Y_i = j} p_i$ ;
/* Functions */
function $T = \mathbf{DT}(\mathcal{D}_n, n_v, n_l)$;
if $d <= n_l$ then
    $T.l = \text{mode}(\{Y \in \mathcal{D}_n\})$ ;
else
    Split $\mathcal{D}_n$ in $\mathcal{D}'_{n'}$ and $\mathcal{D}''_{n''}$ based on the best predictor $s$ over the $n_v$ ones sampled from the whole $d$
      predictors ;
    $T.s = s;\ T.T' = \mathbf{DT}(\mathcal{D}'_{n'}, n_v, n_l);\ T.T'' = \mathbf{DT}(\mathcal{D}''_{n''}, n_v, n_l)$;
```

  based on BOO MS procedure;
  - O(n_v): we set $b = 1$ and $n_l = 1$ while $n_v \in d^{\{0.00, 0.02, \cdots, 1.00\}}$ is optimised based on BOO MS procedure ;
  - O(n_l): we set $b = 1$ and $n_v = \sqrt{n}$ while $n_l \in n \cdot \{0.00, 0.01, \cdots, 0.50\} + 1$ is optimised based on BOO MS procedure;
  - ALL: $b \in \{0.20, 0.22, \cdots, 1.20\}$, $n_v \in d^{\{0.00, 0.02, \cdots, 1.00\}}$, $n_l \in n \cdot \{0.00, 0.01, \cdots, 0.50\} + 1$ are optimised based on BOO MS procedure;
- RF with the weighting strategy W_2 ($p_i = \ln\left((1 - \widehat{L}(T_i))/\widehat{L}(T_i)\right)$ [NP82, BK14]) in the same sub-configuration depicted from W_1: STD, O(b), O(n_v), O(n_l), and ALL;
- RF with the weighting strategy W_3 ($p_i = e^{-\gamma \widehat{L}(T_i)}$ [Cat07, LLST13]) where $\gamma \in 10^{\{-6.0, -5.8, \cdots, 4\}}$ is optimised based on BOO MS procedure, in the same sub-configuration depicted from W_1: STD, O(b), O(n_v), O(n_l), and ALL;

Table 4.2 reports the error on the test set for all the datasets and for all the experimental settings that we have just described while Table 4.3 reports the corresponding value of the optimised hyperparameter which have been selected during the BOO MS procedure.

From Tables 4.2 and 4.3 we can draw the following observations:

- optimising the hyperparameters (even just one of them) mostly leads to model characterised by higher accuracy
- the hyperparameters that have the more remarkable impact on the accuracy of the model

is the weighting strategy followed by the n_v then b and the less important hyperparameter resulted to be n_l

- the usual choice $W1$ plus $b = 1$, $n_v = \sqrt{n}$, and $n_l = 1$, which is obviously the most computational inexpensive choice but resulted to be a competitive one. However if one has to chose the best tradeoff, we would suggest $W3$ plus $b = 1$, $n_v = \sqrt{n}$, and $n_l = 1$
- the weighting strategy $W2$ resulted to be the more inaccurate way of weighting the different classifiers even if this weighting strategy has very strong theoretical properties [BK14]
- the weighting strategy $W3$ resulted to be really effective in this context and this supports all its theoretical properties studied in [Cat07, LLST13]
- by observing Table 4.3, it is possible to note that even if the value chosen as hyperparameters in the MS phase differs a lot from the standard one ($b = 1$, $n_v = \sqrt{n}$, and $n_l = 1$), the corresponding accuracy in Table 4.2 may not vary so much and this means that there are a lot of combinations of the hyperparameters which give good results. This is good news when it comes to performing a MS phase since the task of selecting good hyperparameters for this algorithms results in being simplified.
- For what concerns the choice of the hyperparameters, b results to be the one that is more often different from the conventional choice of $b = 1$ while n_l is mostly near to $n_l = 0$ which is the conventional choice. Note that having $b < 1$, $n_l > 0$ it increases also the speed of the training and classification phase so in the future it will be interesting to check how a computational budget can influence the accuracy of the model.

In conclusion, we can state that we have empirically proved that the different hyperparameters characterising the RF learning algorithm have a remarkable impact on the accuracy of the final model, even the ones that are commonly not considered at all as hyperparameters (b and n_l). Most of all, choosing a good weighting strategy (in particular W_3), results in being one of the most effective ways to improve the generalization performance of the final model. This work is a step forward in understanding the learning properties of the RF and more work is needed in this direction. In particular, a deeper empirical analysis is needed and a theoretical characterisation of the different phenomena should be provided. In any case, from the results of this paper, we can definitely state that a MS phase should always be performed when RF is adopted and this MS is always beneficial with just a computational overhead on the learning phase and no overhead over the classification phase.

	W1					W2					W3				
$\mathcal{D}_n$	STD	O(b)	O(n_v)	O(n_l)	ALL	STD	O(b)	O(n_v)	O(n_l)	ALL	STD	O(b)	O(n_v)	O(n_l)	ALL
D1	0.33	0.33	0.33	0.33	**0**	0.33	0.33	0.33	0.33	**0**	0.33	0.33	0.33	0.33	**0**
D2	3	2	1	4	**0**	3	2	1	4	**0**	2	1	**0**	2	**0**
D3	3.39	1.69	1.69	3.39	1.69	3.39	1.69	1.69	3.39	1.69	3.39	1.69	1.69	3.39	**0**
D4	3.77	3.77	1.89	5.66	1.89	3.77	3.77	1.89	5.66	**0**	3.77	1.89	1.89	1.89	**0**
D5	4.76	3.17	3.17	4.76	3.17	7.94	3.17	3.17	4.76	3.17	4.76	**1.59**	3.17	3.17	**1.59**
D6	8	8	5.67	7.67	5.33	8	7.67	5.33	8.33	5	7.67	7	4.33	7.67	**3.67**
D7	10.67	9.33	5.33	9.67	5.67	10.67	9.33	5.33	9.33	5.33	10	7.33	4.67	8.33	**4**
D8	15.79	13.16	7.89	13.16	5.26	15.79	10.53	10.53	13.16	5.26	15.79	7.89	7.89	7.89	**2.63**
D9	13.33	10	13.33	13.33	10	13.33	10	13.33	13.33	10	13.33	10	10	10	**3.33**
D10	22.06	17.65	14.71	22.06	11.76	22.06	16.18	14.71	20.59	13.24	16.18	14.71	13.24	11.76	**10.29**
D11	20	16.67	16.67	20	10	20	20	20	23.33	10	20	13.33	16.67	16.67	**6.67**
D12	10.67	10.67	11.33	11.67	10	93.67	10.67	12	74	9.33	10.67	9	9.67	10.33	**8.67**
D13	10.33	7.33	5.67	8.67	4.33	100	8.67	100	97	6.33	10	6.33	5.67	8	**3.67**
D14	12.33	11	11.33	12	**9.67**	100	11.67	100	99	11	11.67	10.33	10.67	12	**9.67**
D15	11.67	13	**0**	29	**0**	100	100	87.33	87.33	**0**	6	**0**	**0**	2.67	**0**
D16	23.08	11.54	11.54	15.38	**3.85**	100	100	100	92.31	7.69	23.08	7.69	11.54	15.38	**3.85**
D17	15	14	14	16.67	13.33	99.67	98.33	89.33	86	16.33	14.67	13.33	13	14.33	**12**
D18	12.5	7.81	7.81	9.38	**6.25**	100	100	100	92.19	62.5	9.38	7.81	7.81	9.38	**6.25**
D19	15	14	13.67	15.67	12.67	100	100	99	94	19.67	14	12.67	12.67	14.33	**11.33**
D20	11.33	11	10	11.67	9.67	100	100	99.67	95.33	88.33	11.33	10.33	10.33	11.67	**9.33**
D21	28.13	28.13	25	31.25	**15.63**	100	96.88	96.88	96.88	75	25	25	21.88	25	**15.63**
D22	26.85	26.85	25	27.78	24.07	100	100	100	97.22	87.96	26.85	26.85	25.93	27.78	**23.15**
D23	34.38	34.38	34.38	34.38	31.25	100	90.63	93.75	81.25	68.75	34.38	31.25	31.25	34.38	**28.13**
D24	29.33	27.67	27.33	29.67	27.33	100	99.67	99.67	97	88.67	29.33	27	27.33	29.67	**26.67**
D25	44	42.33	41.67	44.67	40.67	80.67	77.33	78	76	59	43.67	42.67	41	43.33	**39.33**
D26	43.67	41.33	41.67	42.33	39.67	99.67	99.33	99.33	95.33	72.67	43.33	40	41.67	42	**38.67**

Table 4.2: Errors (in percentage) on the test set of the different RF models.

	W1				W2				W3				
$\mathcal{D}_n$	b	n_v	n_l	b, n_v, n_l	b	n_v	n_l	b, n_v, n_l	γ	γ, b	γ, n_v	γ, n_l	γ, b, n_v, n_l
D1	0.28	0.04	0	0.34,0.24,0	0.22	0.04	0	0.34,0.24,0	-6	0.72,0.22	1.72,0	-6,0	-6,0.34,0.24,0
D2	0.64	0.62	0	0.78,0.76,0	0.64	0.62	0	0.78,0.76,0	0.92	2.34,0.7	1.94,0.66	1.52,0	2.14,0.58,0.62,0
D3	0.56	0.44	0.01	0.4,1,0	0.56	0.44	0.01	0.4,1,0	-6	1.32,0.4	-6,0.44	-6,0.01	1.94,0.88,0.14,0.02
D4	0.54	0	0	0.2,0.14,0	0.54	0	0	0.7,0.1,0.05	-6	2.14,0.98	-6,0	1.52,0	1.72,0.36,0.58,0
D5	0.5	0	0	0.2,0.1,0	0.5	0.52	0	0.22,0.44,0	-6	2.34,0.8	-6,0	1.94,0.19	2.34,0.4,0.06,0.02
D6	0.76	0.8	0	0.76,0.72,0	0.76	0.96	0	1,0.94,0	1.12	1.52,0.78	2.14,0.96	1.32,0	1.72,0.88,1,0
D7	0.64	1	0	0.8,1,0	0.64	1	0	0.8,1,0	1.94	1.94,1	1.72,1	1.52,0	2.14,0.94,0.96,0
D8	0.42	0.62	0	0.32,0,0.01	0.94	0.82	0	0.84,1,0.2	-6	1.72,0.5	-6,0.62	1.72,0.14	1.72,0.2,0.8,0
D9	0.44	0	0	0.32,1,0	0.66	0	0	0.34,0.44,0	-6	2.34,0.34	1.32,0.24	2.34,0.03	2.54,0.44,0.28,0.05
D10	0.54	0.62	0.01	0.7,0.8,0.02	0.54	0.62	0.01	0.7,0.8,0.02	1.94	1.72,0.44	1.94,0.82	2.54,0	1.94,0.2,0.62,0
D11	0.42	0.58	0	0.2,0.82,0.01	0.4	0.18	0	0.34,0.8,0.02	-6	1.94,0.88	1.94,0.24	2.14,0.17	-6,0.4,1,0.03
D12	0.86	0.32	0.01	0.7,0.52,0	0.42	0.76	0.3	0.62,0.48,0	0.92	1.94,0.58	1.72,0.44	1.52,0	2.14,0.84,0.68,0.01
D13	0.44	0.18	0	0.54,0.14,0	0.28	0	0.29	0.32,0.32,0	1.52	1.94,0.92	-6,0.18	1.72,0	1.12,0.54,0.14,0
D14	0.88	0.52	0	0.84,0.56,0	0.54	0	0.08	0.4,0.56,0	0.72	1.12,0.54	1.32,0.52	-6,0	-6,0.84,0.56,0
D15	0.22	0.58	0	0.2,0.58,0	0.2	0.04	0.23	0.2,0.62,0	0.92	0.92,0.2	1.32,0.48	1.52,0	-6,0.2,0.58,0
D16	0.34	0.82	0	0.2,0.82,0	0.2	0	0.2	0.2,0.96,0	-6	1.52,0.48	-6,0.82	-6,0	-6,0.2,0.82,0
D17	0.84	0.58	0	0.4,0.56,0	0.36	0	0.17	0.36,0.82,0	0.72	1.12,0.94	1.72,0.62	1.52,0	1.94,0.84,0.62,0
D18	0.98	0.34	0	0.94,0.44,0	0.2	0	0.19	0.2,0,0.1	1.32	-6,0.98	1.12,0.04	-6,0	1.32,0.84,0.48,0
D19	0.64	0.56	0	0.86,0.62,0	0.2	0.06	0.2	0.28,0.86,0	1.52	1.32,0.78	1.72,0.72	1.52,0.01	0.92,1,0.52,0
D20	0.78	0.38	0	0.94,0.38,0	0.2	0	0.25	0.2,0.28,0.22	-6	0.92,0.72	-6,0.38	0.5,0	0.72,0.94,0.38,0
D21	0.62	0.8	0.05	0.86,0.76,0	0.2	0.04	0.02	0.92,0,0.3	0.72	1.72,0.8	1.32,0.68	1.52,0	1.72,0.8,0.38,0.02
D22	0.72	0.38	0	0.62,0.68,0	0.2	0	0.27	0.2,0.34,0.2	-6	-6,0.72	-6,0.04	-6,0	-6,0.94,0.14,0
D23	0.28	0.38	0	0.26,0.52,0.01	0.72	0.76	0.29	0.5,0.28,0.16	-6	1.52,0.58	1.52,0.44	-6,0	2.14,0.42,0.8,0.01
D24	0.54	0.18	0	0.44,0.42,0	0.88	0.04	0.13	0.28,0.34,0.14	-6	1.52,0.54	1.32,0.38	-6,0	1.72,0.64,0.2,0
D25	0.32	0.8	0.01	0.36,1,0	0.5	0.04	0.21	0.2,0.1,0.28	-6	-6,0.32	0.72,0.8	1.94,0.08	1.94,0.94,0.62,0.04
D26	0.5	0.86	0.01	0.32,0.94,0	0.2	0.24	0.29	0.2,0,0.1	0.92	1.94,0.5	1.12,0.52	1.32,0	1.72,1,1,0

Table 4.3: Value of the hyperpararameters selected with the BOO MS procedure which have produced the model characterized by the error of Table 4.2.

Part III

Applications

Chapter 5

KPIs Modeling and Visualization

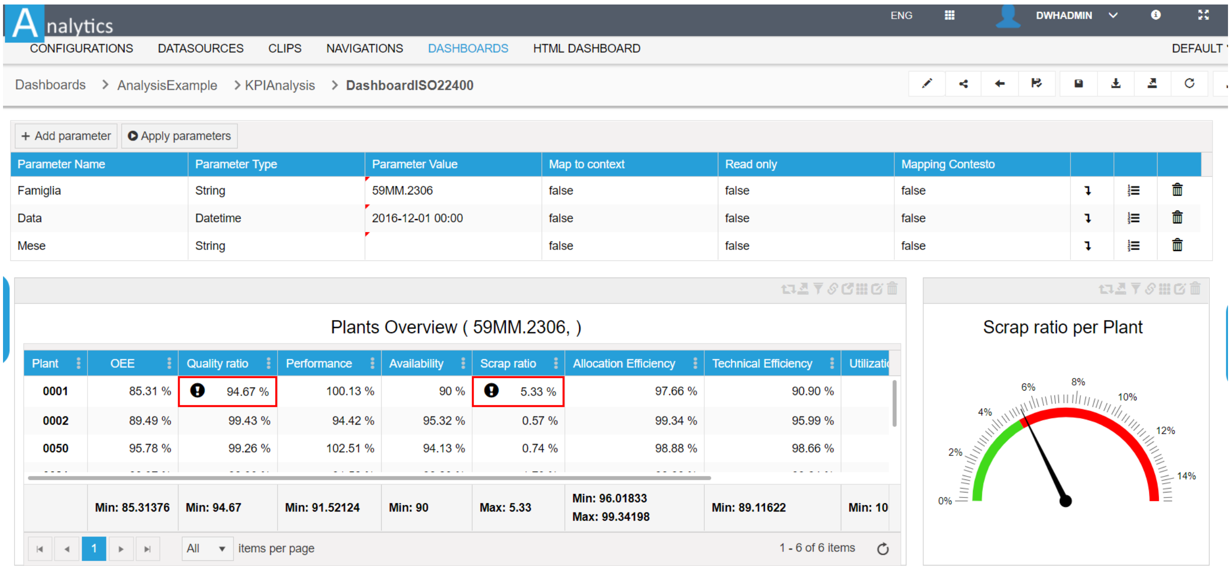
In this Section we present a practical implementation of the first key topic of this thesis: KPI analysis in manufacturing. This is a typical example of descriptive/diagnostic analysis performed through the creation of a set of business dashboards by means the ”Analytics BI” software platform.
In particular, we will present the design of the KPIs model and how this BI software can be exploited in order to visualize these information. This activity has been made possible thanks to the collaboration established within the *Smart Factory 2020* project with “SedApta s.r.l”, a Genoese company software operating in the field of Manufacturing IT & Supply Chain Management.
These dashboards aim to help manufacturers to monitor a set of KPIs, universally valid and industry-neutral, compliant with the ISO 22400 standard and the SCOR reference model. Dashboards design is the last step of a series of activities that start from the identification of the main cross-industry manufacturing processes and go through the:

1. Identification of the users for which dashboards are intended,
2. Identification of the type of information required: choose which KPIs to monitor and the data needed in order to compute them,
3. Definition of the methods needed to gather information from the industry’s data sources,
4. Data Warehouse design in order to systematize the collected data within appropriate tables,
5. Identification of the KPIs reference standards,
6. Definition of the analysis dimensions in order to query and drill-down these aggregated data,
7. Design of graphical interfaces so to have a quick and clear grasp of these KPI’s trend.

In order to build these dashboards, a toy dataset with information related to the manufacturing context, was used. The Data Warehouse and the related OLAP cubes have been designed in collaboration with the Analytics BI development team.

5.1 Analytics BI software

Analytics BI is a web tool providing business intelligence features for business process analysis. This software platform enables to create and configure analysis dashboards composed by a set of graphical representation of data for analysis purpose as depicted in Figure 5.1. Analytics BI allows to retrieve data from different sources, both relational and not. Data can be filtered and aggregated in different ways and, then, they can be visualized by various charts that range from simple grids, pivot tables, gauges, histograms, radars, pie charts and so on.



Parameter Name	Parameter Type	Parameter Value	Map to context	Read only	Mapping Contesto
Famiglia	String	59MM.2306	false	false	false
Data	Datetime	2016-12-01 00:00	false	false	false
Mese	String		false	false	false

Plants Overview (59MM.2306,)

Plant	OEE	Quality ratio	Performance	Availability	Scrap ratio	Allocation Efficiency	Technical Efficiency	Utilizati
0001	85.31 %	94.67 %	100.13 %	90 %	5.33 %	97.66 %	90.90 %	
0002	89.49 %	99.43 %	94.42 %	95.32 %	0.57 %	99.34 %	95.99 %	
0050	95.78 %	99.26 %	102.51 %	94.13 %	0.74 %	98.88 %	98.66 %	
	Min: 85.31376	Min: 94.67	Min: 91.52124	Min: 90	Max: 5.33	Min: 96.01833 Max: 99.34198	Min: 89.11622	Min: 10

Figure 5.1: Example of Analytics BI dashboard.

In summary, the Analytics BI main technical features are[1]:

- Web solution, allows users to access their dashboards wherever they are, with the only requirement of having a web browser and access to the internet.
- Multi data source, the possibility to access different data sources, from multidimensional cubes to simple excel sheets.

[1] www.sedapta.com/product/analytics

- Cube analyzer, the possibility to explore the multidimensional cubes in order to create custom analysis in autonomy.
- Mobile-ready interface,
- Monitoring processes KPI also in real time, if integrated into the production workflow.

5.2 KPIs Modeling in Analytics BI

The selection of a proper data source constitutes the basis for building up effective analysis dashboards. Thus, the first step of our KPIs model was the design of a back-end architecture to support the management of manufacturing data and the KPIs calculation. Since Analytics BI supports the connection to OLAP cubes, where it is possible to aggregate data in a more complex way respect to their query in a relational database, we choose to develop an OLAP architecture which inherently provides a Data WareHouse and an analysis cubes as depicted in Figure 5.2.

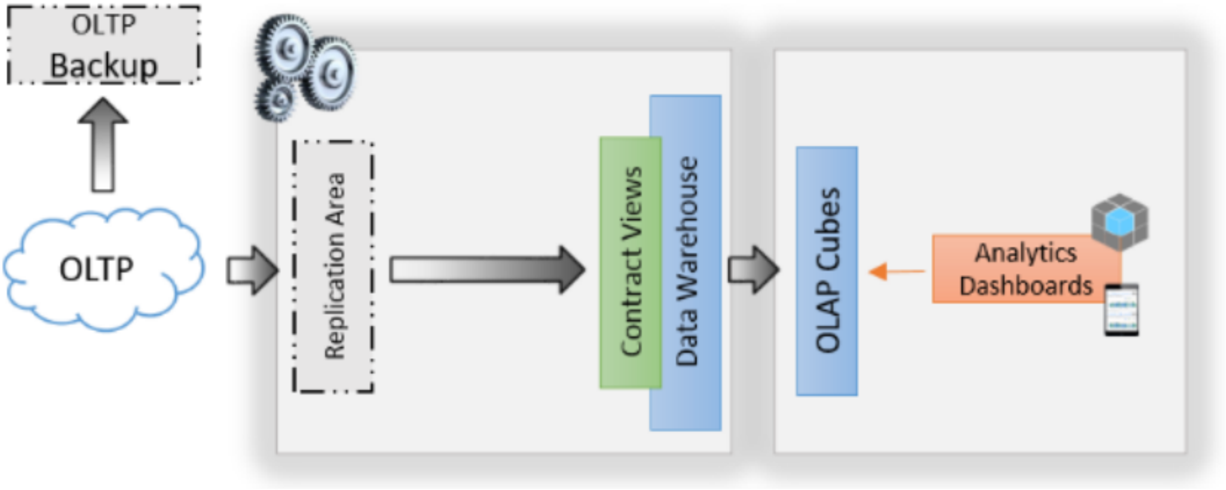


Figure 5.2: Reference back-end architecture for KPIs modelling.

This reference architecture, is composed by a set of ETLs that, by means of contract views, are able to load data into the Data Warehouse on which OLAP cubes operate. OLAP analysis provides new consolidated information (obtained by data aggregation) that will be exploited to create dashboards. The OnLine Transaction Processing (OLTP) is the set of data sources from which it is possible to extract data. Typically, an OLTP source is a transactional system able to manage information related to a process (like management or production platforms).

Once identified the data sources that will be part of the Data Warehouse, it is possible to store OLTP backups. In this way, the timing in which the analysis is performed doesn't matter, we

will always have the data available to reconstruct the events.
The DWH is where all information is merged, ”cleaned”, reprocessed and consolidated. Information is managed in new business entities in order to represent all areas of a manufacturing company.
ETLs do not access tables directly, but by view-based Contract Layers. This guarantees the best compromise between data maintainability and future DWH developments.
OLAP cubes are the end point for the information and the starting point for the analysis. Through the OLAP sources it is possible to measure the company KPIs and visualize their trend into a dashboard.

The Data Warehouse represents the logical model of our domain and, for this reason, it complies with the guidelines of the ISA-95[2] standard for Enterprise-Control System Integration, which provides a standard terminology and information models for the development of an automated interface between business and control systems.

In particular, this DWH has been designed to be fully exploited in manufacturing sector. We prepared more than 70 normalized tables in order to cover, as much as possible, the flow of activities and the main aspects that characterize a manufacturing system.

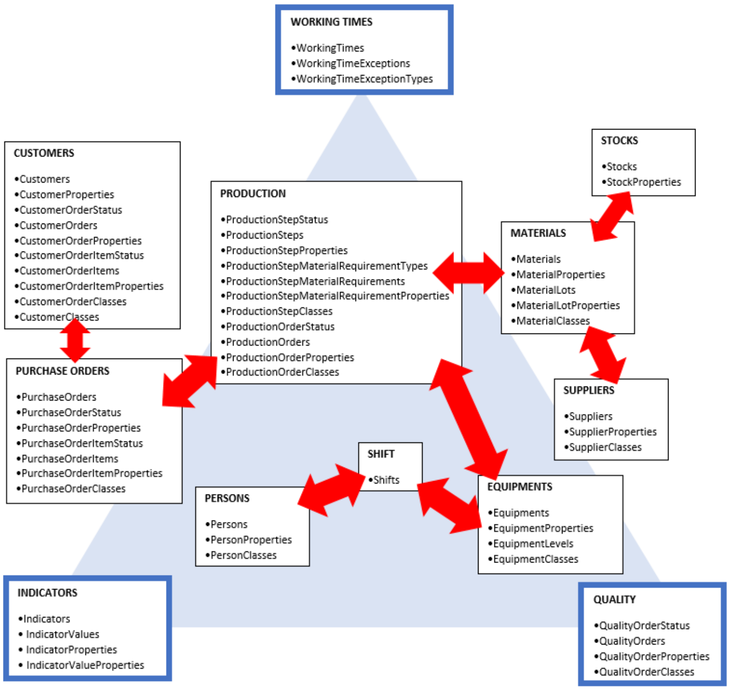


Figure 5.3: Schematic representation of Data Ware House’s main semantic areas for KPIs modelling in Manufacturing.

[2] www.isa.org/isa95

Figure 5.3 summarizes some of these tables that represent the key entities (and their related activities) of a production process. Each rectangle represents a set of DWH tables referring to the same semantics. A manufacturing system consists of different production phases that refer to an order which is connected to a customer. From a ”spatial” point of view, a work phase is connected to a work unit in which teams of operators work according to their shifts. Moreover, a work phase is linked to the materials that are spent and, consequently, to material’s suppliers and stocks. Finally, during each production phase, various measurable variables are recorded. These measures give information about the product’s quality, the processing times or other indicators related to the performance of the production.

This DWH is the basis from which we developed three OLAP cubes with a set of ”out-of-the-box” KPIs that are ISO 22400 compliant. Each cube define a specific analysis and, for each analysis, is possible to aggregate the measures by different scenarios. Note that, in all of these three cubes, the Time dimension and the hierarchical structure for the physical equipment (enterprise - site – area – work centres – work units) are always considered in addition to each cube specific dimensions.
The first cube we present, is focused on the resources (machineries, operators, equipment) analysis and is made up of 23 KPIs that ca be aggregated into four dimensions: Equipment/Shift, Work Orders, Product Family and Operators. Some of these KPIs are specific for machineries and equipment whereas others are specific for operators:

- Capacity = the resources availability, in hours, calculated on the basis of the shift and of the resources’ inventory.
- Yield = the ratio, in percentage, between the theoretical working time and the final processing time balanced.
- Efficiency = the ratio, in percentage, between the final hours balanced, net of downtimes, and capacity.
- Saturation = the resource saturation. It is calculated as 1 minus the ratio between the unsaturated hours and the capacity. The unsaturated hours are calculated as the sum between the hours of inactivity (= availability – hours of activity), the set up hours and the hours of maintenance.
- Good parts = the stated good quantity.
- Scrap parts = the stated scrap quantity.
- Scrap ratio= the ratio, in percentage, of the total scraps. It is calculated as the ratio between scrap quantity and good parts.
- Inactivity Time = the difference between the capacity and its activity time.

- Hypothetical hourly production.
- Actual gross hourly production.
- Final set up time balanced.
- Final working time balanced.
- Final start up time balanced.
- Final downtime balanced.
- Final maintenance time balanced.
- Final reworking time balanced.
- Planned working time.
- Human set up time.
- Human working time.
- Human start up time.
- Human downtime.
- Human maintenance time.
- Human reworking time.

The second cube is related to the scraps and downtimes evaluation. We calculated 5 indicators respect to three hierarchical dimensions:

- Reasons (of scraps/downtimes): it is possible to drill-down through machinery, production order and the process step in which the downtime occurred.
- Equipments/Shifts: starting from the machinery in which downtime/scraps were stated, it is possible to drill-up the shift, the production order and the production step in which this event occurred.
- Work Orders: it enables to investigate downtimes/scraps occurred on production orders

These KPIs are:

- Downtime hours = number of downtimes hours.

- Downtime occurrences = the number of times in which a downtime occurred.
- Downtime Incidence = the percentage of times in which a certain cause of downtime occurs with respect to the total causes recorded.
- Scraps quantity = final scraps quantity balanced.
- Scraps Incidence = the percentage of times in which a certain cause of scraps occurs with respect to the total causes recorded.

Finally, the last cube is focused on the inventory (stock) management. Here we consider the materials (production orders) movements among inventories, and also, the delivery and the supply of materials. The analysis dimension is the plant's inventories. The measures considered in order to calculate KPIs are:

- Timestamps of the material to move.
- Source stock code.
- Source sub-stock code.
- Source stock type (a detailed definition about this point will be provided in Section 5.4).
- Material position in the x-axis of the source warehouse space.
- Material position in the y-axis of the source. warehouse space
- Material position in the z-axis of the source. warehouse space
- Production order stored in the source warehouse.
- Production order phase (stored in the source warehouse).
- Production order part stored in the source warehouse.
- Sublot of the production order stored in the source warehouse.
- Quantity delivered of the material stored in the source warehouse.
- Quantity supplied of the material stored in the destination warehouse.
- Destination stock code.
- Destination sub- stock code.

- Destination stock type (a detailed definition about this point will be provided in Section 5.4).
- Material position in the x-axis of the destination warehouse space.
- Material position in the y-axis of the destination warehouse space.
- Material position in the z-axis of the destination warehouse space.
- Production order stored in the destination warehouse.
- Production order phase (stored in the destination warehouse).
- Production order part stored in the destination warehouse.
- Sublot of the production order stored in the destination warehouse.
- Quantity delivered of the material stored in the destination warehouse.
- Supply reasons.
- Delivery reasons.
- Materials movement types among warehouses (on hold / delivered / to be deliver).

Analytics BI dashboards represent the final result in the reporting data analysis. In the next two Sections we will discuss four dashboards developed through the exploitation of the information provided by these cubes.

5.3 First case study of KPIs visualization: OEE and production loss analysis

This Section presents two dashboards for the monitoring of the Overall Equipment Effectiveness (OEE) and other related measures, developed by means the Analytics BI software. In particular, it is described a scenario related to a multi-plant manufacturing company. The considered dashboard shows a site level aggregated OEE calculations.

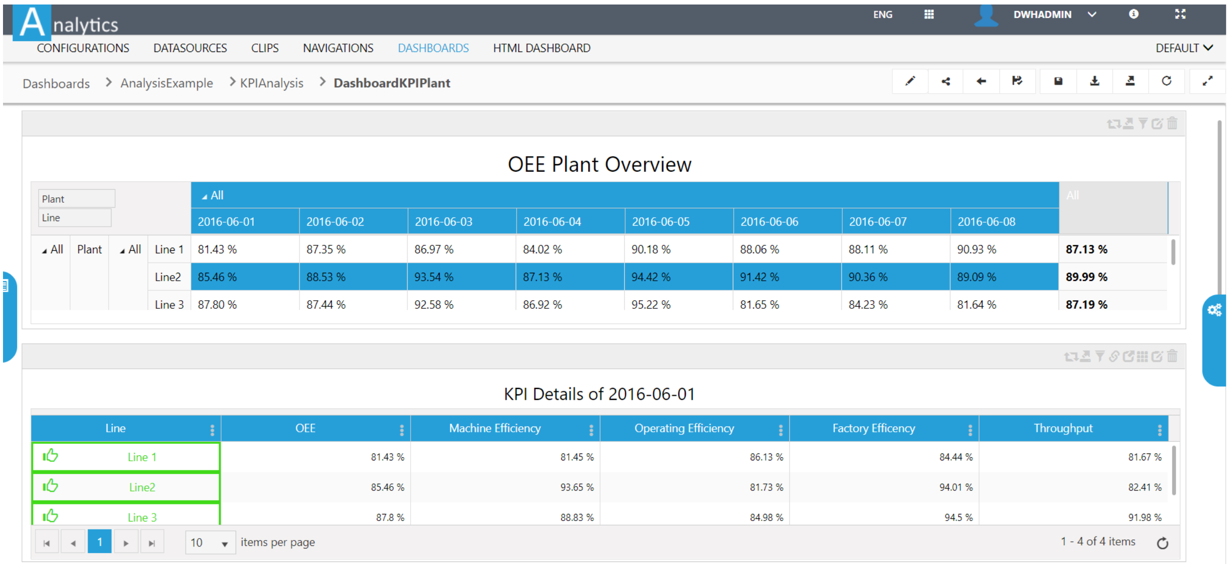
The OEE is the reference indicator for the analysis of the production process. It allows the calculation of the global efficiency of a plant, classifying the different production losses according to three factors: availability, performance and quality.
ISO 22400 standard defines:

- *Availability*: the percentage of operating time with respect to the planned production time.

- *Performance*: the percentage of parts produced compared to what is programmed.
- *Quality*: the percentage of non-defective parts compared to the total parts produced.

These are the three fundamental factors that influence global efficiency and productivity. The accuracy of the OEE value is connected to how the data are collected during production. This task must take place automatically and in real time and it is of huge importance because without a precise and timely measurement of production data, it is not possible to identify the adjustments necessary to improve and streamline the production process.

In the given scenario, at the highest level, manager observes the plant overall OEE trend on different days and, in order to investigate in detail, he can drill down to line level. At this level, it is possible to evaluate the four performing line in a specific selected day over the time range considered. The details in the table in Figure 5.4, explore the OEE value of each line respect to five efficiency indicators:



OEE Plant Overview

Plant				All								All
Line				2016-06-01	2016-06-02	2016-06-03	2016-06-04	2016-06-05	2016-06-06	2016-06-07	2016-06-08	
All	Plant	All	Line 1	81.43 %	87.35 %	86.97 %	84.02 %	90.18 %	88.06 %	88.11 %	90.93 %	**87.13 %**
			Line2	85.46 %	88.53 %	93.54 %	87.13 %	94.42 %	91.42 %	90.36 %	89.09 %	**89.99 %**
			Line 3	87.80 %	87.44 %	92.58 %	86.92 %	95.22 %	81.65 %	84.23 %	81.64 %	**87.19 %**

KPI Details of 2016-06-01

Line	OEE	Machine Efficiency	Operating Efficiency	Factory Efficency	Throughput
Line 1	81.43 %	81.45 %	86.13 %	84.44 %	81.67 %
Line2	85.46 %	93.65 %	81.73 %	94.01 %	82.41 %
Line 3	87.8 %	88.83 %	84.98 %	94.5 %	91.98 %

Figure 5.4: First dashboard prototype for Plants' OEE monitoring (part I).

- *Machine efficiency*: identifies the productivity of work units. High values of this indicator express a positive trend [SHYH08];
- *Throughput*: is calculated as the ratio between the produced quantity of an order and the actual execution time of an order. High values of this KPI highlight a good efficiency in production performance[3];

[3] www.iso.org/obp/ui/#iso:std:iso:22400:-2:ed-1:v1:en

- *Factory efficiency*: measures how effectively the production processes use its available resources. This KPI is also an indicator of the manufacturing process capacity to complete an order using the least amount of inventory [OPP$^+$02];

- *Operational efficiency*: is calculated as the actual processing time respect to the planned processing time of a line[4].

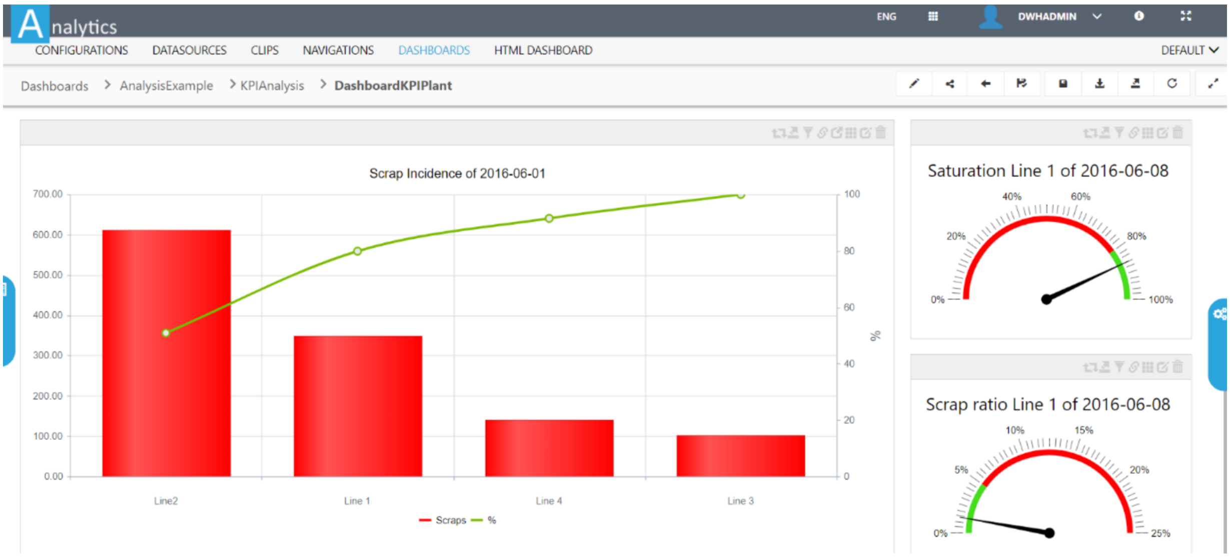
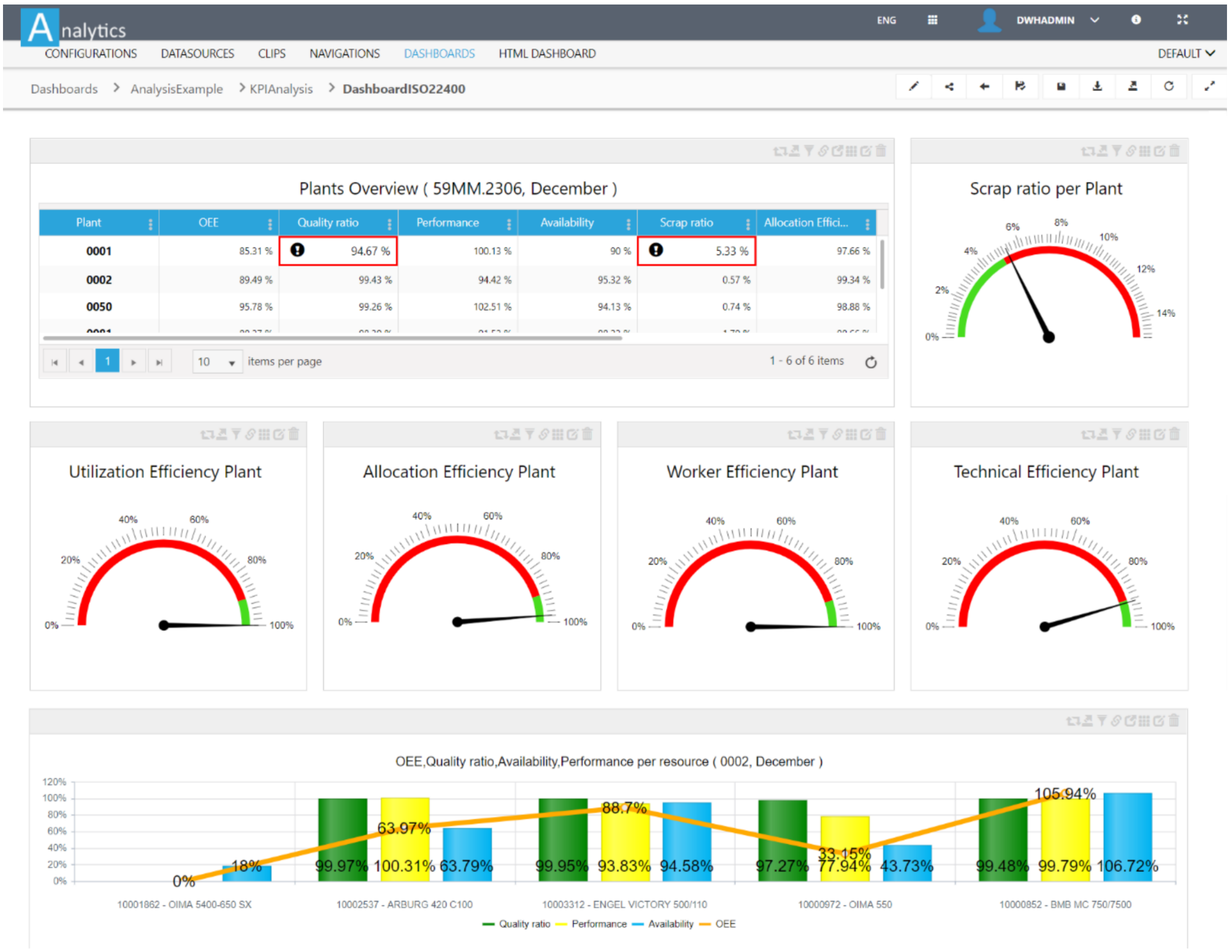


Figure 5.5: First dashboard prototype for Plants' OEE monitoring (part II).

As presented in Figure 5.5, plant manager can further drill down in order to investigate other indicators that affect the performance of a specific production line. An example can be the evaluation of the ratio between the actual production level of a line respect to its theoretical maximum (saturation). Another possible investigation can be the comparison between how much scrap was actually produced in a single line respect to the expected (scrap ratio) or ,otherwise, the evaluation of the scrap ratio among all the production line in exam. In particular, in this dashboard, the two gauge show the scrap ratio and saturation values projected on the week following the one in exam. Another example related to the monitoring of a manufacturing systems performance, is presented in the second dashboard depicted in Figure 5.6. In this case, the variable of interest regards the production of a particular family of products between the different parts of a plant. The first table provides an overview of the status of the various plants highlighting those values that diverge from the planned result for each KPI. The OEE of each plant is measured against its basic indicators and is correlated with other efficiency indicators; in this way a deeper knowledge on the motivation of its value according to the other analysis variables is given. The positive trend of the OEE is also reflected in other indicators such as:

[4] www.iso.org/obp/ui/#iso:std:iso:22400:-2:ed-1:v1:en



Plants Overview (59MM.2306, December)

Plant	OEE	Quality ratio	Performance	Availability	Scrap ratio	Allocation Effici...
0001	85.31 %	94.67 %	100.13 %	90 %	5.33 %	97.66 %
0002	89.49 %	99.43 %	94.42 %	95.32 %	0.57 %	99.34 %
0050	95.78 %	99.26 %	102.51 %	94.13 %	0.74 %	98.88 %

Figure 5.6: Second dashboard prototype for Plants' OEE monitoring.

- the productivity of all work units (utilization efficiency and technical efficiency)
- the machinery utilization (allocation efficiency)
- the performance of employees during working hours (worker efficiency)

By selecting a row in the first table, the last chart will be filtered respect to the specific plant listed in the (master) table. This last chart analyze how efficient a plant is in using the resources related to a specific product family. For each resource used in production (listed in x-axis), it is possible to investigate the plant availability, quality and performance. The OEE trend is caused by the combination of these three indicators value.

5.4 Second case study of KPIs visualization: Inventory Management analysis

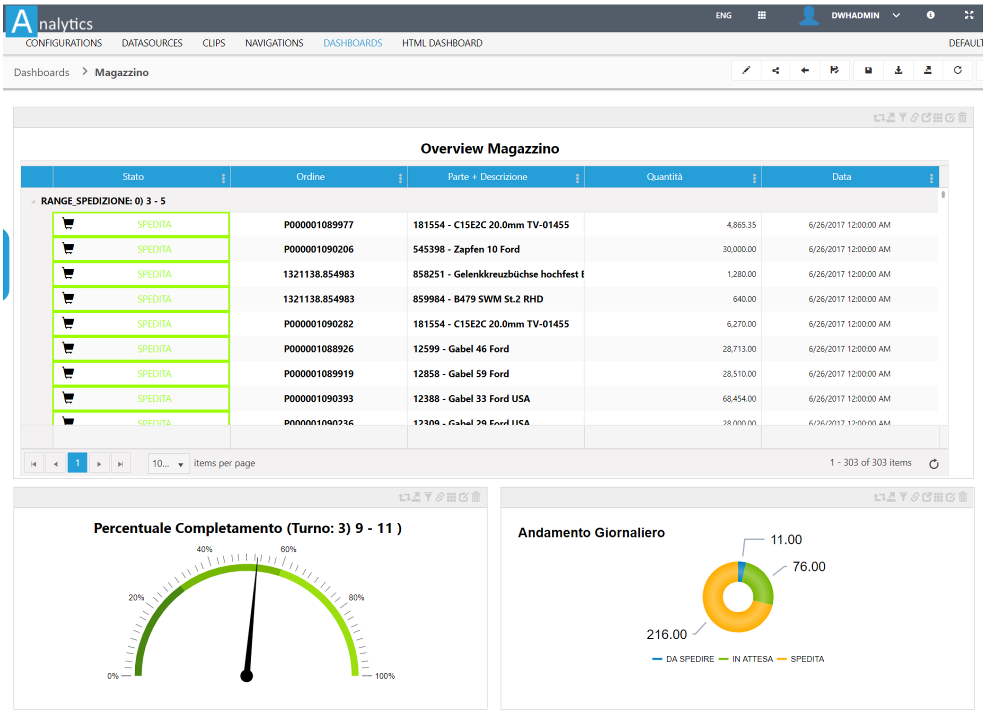
In the previous Section we described two examples of dashboards for monitoring production processes. Taking as reference the IEC 62264 functional hierarchy (Figure 2.6 in Section 2.4.3.1) of manufacturing systems, the indicators that have been used in these dashboards measure the activities that are located at the third level of the hierarchy, related to Manufacturing Operations Management.
On the other hand, this second case study presents a prototype of a dashboard to monitor an aspect more related to the context of the supply chain: the inventory management. In this case, we investigated the indicators provided by SCOR reference model. In fact, SCOR considers the Inventory Management as one of the main activities performed in "Source" area.

Inventory management covers all the activities aimed at minimizing inventory management costs, while maintaining an adequate supply of goods that is able to meet customer needs [BCC$^+$02]. The stock can be defined as a certain stored quantity of an article that must be provided to users in order to be consumed according to their needs. It is important to note that material stocks can be classified into four categories [SH07, CK00]:

- Supply stocks, are allocated to be used in the administrative sectors, production departments and maintenance operations;
- Raw materials stocks, are purchased from external suppliers and are intended for production;
- Semi-finished or Work In Progress (WIP) stocks, have their origin and destination in the various production departments;
- Stocks of finished products.

For these reasons, the inventory management process has a direct impact on a company's profit margins and cash flow. In particular, when a company does not have adequate stocks, "stock-outs" occur, with a consequent loss of customers and a decrease in turnover. In the opposite, when a company has excess inventory or a low turnover, the costs increase, and also the risk of obsolescence of the products, as well as those related to the storage of goods in the warehouse.

Figure 5.7 shows the first example of a dashboard for the management of stocks of finished products (the fourth category aforementioned). The dashboard offers a general and real-time view of the "movement-status" of finished products (on hold / shipped / to be shipped) for a specific plant's warehouse. This information is also summarized by the donut chart in the bottom right of the Figure. The first table groups all the products managed by the warehouse on the basis



Overview Magazzino

Stato	Ordine	Parte + Descrizione	Quantità	Data
RANGE_SPEDIZIONE: 0) 3 - 5				
SPEDITA	P000001089977	181554 - C15E2C 20.0mm TV-01455	4,865.35	6/26/2017 12:00:00 AM
SPEDITA	P000001090206	545398 - Zapfen 10 Ford	30,000.00	6/26/2017 12:00:00 AM
SPEDITA	1321138.854983	858251 - Gelenkkreuzbüchse hochfest E	1,280.00	6/26/2017 12:00:00 AM
SPEDITA	1321138.854983	859984 - B479 SWM St.2 RHD	640.00	6/26/2017 12:00:00 AM
SPEDITA	P000001090282	181554 - C15E2C 20.0mm TV-01455	6,270.00	6/26/2017 12:00:00 AM
SPEDITA	P000001088926	12599 - Gabel 46 Ford	28,713.00	6/26/2017 12:00:00 AM
SPEDITA	P000001089919	12858 - Gabel 59 Ford	28,510.00	6/26/2017 12:00:00 AM
SPEDITA	P000001090393	12388 - Gabel 33 Ford USA	68,454.00	6/26/2017 12:00:00 AM
SPEDITA	P000001090236	12309 - Gabel 29 Ford USA	28,000.00	6/26/2017 12:00:00 AM

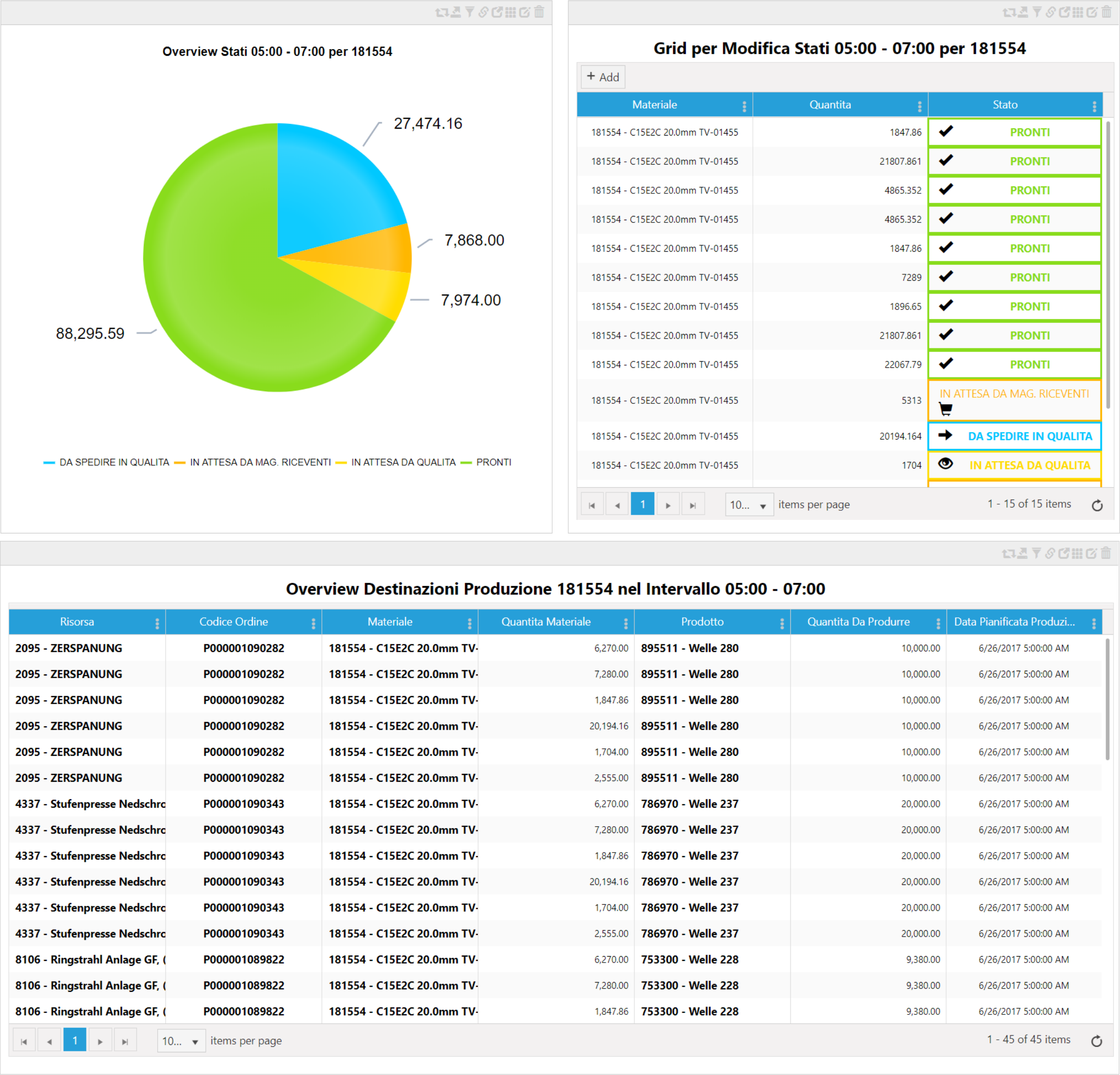
Percentuale Completamento (Turno: 3) 9 - 11)

Andamento Giornaliero

Figure 5.7: First dashboard prototype for Plants' Inventory Management.

of different time ranges within which these orders must be sent. For each finished product, the quantity and the delivery date are indicated. Similarly, but with a more immediate visual effect, the gauge shows the progress states of moved stocks for each time range selected in the table.

In addition, in monitoring the warehouse, it is important to consider the management of stock of materials (the second category aforementioned). This is depicted in Figure 5.8. In this case, the variable of interest is the quantity of material available and needed for the production of a particular family of products. The first table shows the list of materials, grouped over different time ranges, their availability trend and their request date. It may be noted that, unlike the previous dashboard, time ranges are divided into hours instead of days. This highlights the difference in timing, and also in the management, that distinguishes the permanence of a finished material in the warehouse respect to the stock of those materials that must be employed for production. Similarly to the first dashboard, it was used a pie chart to have an immediate visual effect on the movement-progress of those materials (needed for the production of a chosen products family) that are in the warehouse. This chart updates each time a user selects a different time range in the master table. So, by choosing a specific time range and a specific products family, it is possible to have a more detailed view about the required quantity of each material and its availability in warehouse. Finally, resources need to be allocated in order to transform these materials. At the same time, resources are associated to a work order that, in turn, is linked to a product which belong to the chosen family. The management of all this flow is shown in the last table.



Grid per Modifica Stati 05:00 - 07:00 per 181554

+ Add

Materiale	Quantita	Stato
181554 - C15E2C 20.0mm TV-01455	1847.86	✓ PRONTI
181554 - C15E2C 20.0mm TV-01455	21807.861	✓ PRONTI
181554 - C15E2C 20.0mm TV-01455	4865.352	✓ PRONTI
181554 - C15E2C 20.0mm TV-01455	4865.352	✓ PRONTI
181554 - C15E2C 20.0mm TV-01455	1847.86	✓ PRONTI
181554 - C15E2C 20.0mm TV-01455	7289	✓ PRONTI
181554 - C15E2C 20.0mm TV-01455	1896.65	✓ PRONTI
181554 - C15E2C 20.0mm TV-01455	21807.861	✓ PRONTI
181554 - C15E2C 20.0mm TV-01455	22067.79	✓ PRONTI
181554 - C15E2C 20.0mm TV-01455	5313	IN ATTESA DA MAG. RICEVENTI 🛒
181554 - C15E2C 20.0mm TV-01455	20194.164	→ DA SPEDIRE IN QUALITA
181554 - C15E2C 20.0mm TV-01455	1704	👁 IN ATTESA DA QUALITA

1 | 10... items per page | 1 - 15 of 15 items

Overview Destinazioni Produzione 181554 nel Intervallo 05:00 - 07:00

Risorsa	Codice Ordine	Materiale	Quantita Materiale	Prodotto	Quantita Da Produrre	Data Pianificata Produzi...
2095 - ZERSPANUNG	P000001090282	181554 - C15E2C 20.0mm TV-	6,270.00	895511 - Welle 280	10,000.00	6/26/2017 5:00:00 AM
2095 - ZERSPANUNG	P000001090282	181554 - C15E2C 20.0mm TV-	7,280.00	895511 - Welle 280	10,000.00	6/26/2017 5:00:00 AM
2095 - ZERSPANUNG	P000001090282	181554 - C15E2C 20.0mm TV-	1,847.86	895511 - Welle 280	10,000.00	6/26/2017 5:00:00 AM
2095 - ZERSPANUNG	P000001090282	181554 - C15E2C 20.0mm TV-	20,194.16	895511 - Welle 280	10,000.00	6/26/2017 5:00:00 AM
2095 - ZERSPANUNG	P000001090282	181554 - C15E2C 20.0mm TV-	1,704.00	895511 - Welle 280	10,000.00	6/26/2017 5:00:00 AM
2095 - ZERSPANUNG	P000001090282	181554 - C15E2C 20.0mm TV-	2,555.00	895511 - Welle 280	10,000.00	6/26/2017 5:00:00 AM
4337 - Stufenpresse Nedschro	P000001090343	181554 - C15E2C 20.0mm TV-	6,270.00	786970 - Welle 237	20,000.00	6/26/2017 5:00:00 AM
4337 - Stufenpresse Nedschro	P000001090343	181554 - C15E2C 20.0mm TV-	7,280.00	786970 - Welle 237	20,000.00	6/26/2017 5:00:00 AM
4337 - Stufenpresse Nedschro	P000001090343	181554 - C15E2C 20.0mm TV-	1,847.86	786970 - Welle 237	20,000.00	6/26/2017 5:00:00 AM
4337 - Stufenpresse Nedschro	P000001090343	181554 - C15E2C 20.0mm TV-	20,194.16	786970 - Welle 237	20,000.00	6/26/2017 5:00:00 AM
4337 - Stufenpresse Nedschro	P000001090343	181554 - C15E2C 20.0mm TV-	1,704.00	786970 - Welle 237	20,000.00	6/26/2017 5:00:00 AM
4337 - Stufenpresse Nedschro	P000001090343	181554 - C15E2C 20.0mm TV-	2,555.00	786970 - Welle 237	20,000.00	6/26/2017 5:00:00 AM
8106 - Ringstrahl Anlage GF, (	P000001089822	181554 - C15E2C 20.0mm TV-	6,270.00	753300 - Welle 228	9,380.00	6/26/2017 5:00:00 AM
8106 - Ringstrahl Anlage GF, (	P000001089822	181554 - C15E2C 20.0mm TV-	7,280.00	753300 - Welle 228	9,380.00	6/26/2017 5:00:00 AM
8106 - Ringstrahl Anlage GF, (	P000001089822	181554 - C15E2C 20.0mm TV-	1,847.86	753300 - Welle 228	9,380.00	6/26/2017 5:00:00 AM

1 | 10... items per page | 1 - 45 of 45 items

Figure 5.8: Second dashboard prototype for Plants' Inventory Management.

Chapter 6

Predictive Maintenance on a production line of calipers

In this Section we present a practical implementation of the second key topic of this thesis: the application of machine learning techniques to manufacturing problems. This is a typical example of predictive analysis performed in a critical problem within a production process: predictive maintenance.

In particular, we discuss the development of a predictive model able to forecast the next decay state of the machinery component for the High Pressure (HP) test.

This activity has been made possible thanks to the collaboration established within the ”Smart Factory 2020” project with an Italian manufacture that is a world leader in the production of braking systems for vehicles.

The project is very important for this Company because the goal is to tackle the problem of analyzing an issue that frequently occurs on the high-pressure calipers test machinery. The machinery in charge of performing the high pressure test may cause anomalous recording. The idea is to create a tool able to predict when it will be necessary to perform maintenance interventions on the machinery of the high pressure test, in order to avoid the faults recording.

The knowledge, in advance, of the future problems that can occur in the line, enables to obtain too many advantages. For instance, to avoid the production of waste or the occurrence of downtime. In order to create such tool, it has been necessary to understand the production process of calipers. In addition, it has been required to collect, in collaboration with the Company, all the data required to perform this analysis. Therefore, this model is developed using all the variables that can be collected in the production line in question, in order to:

- Forecast after how many measures, or, how much time, it is necessary to do maintenance of the HP test machinery component;
- Find non trivial correlations with respect to other measures recorded on the previous and

next steps.

6.1 Problem description

The context of this work concerns a particular production line of calipers. This production line includes two main parts. On one side, in this line enters machined parts that are already partially processed. Such parts require a *First Assembly* that could be the applications of screws, seals, serial number molding, the performing of various tests, etc. The second part of this line instead perform the *Final Completion*. For instance, the insertion of pads, and the quality control. The processing flows that take place on this line can be of three types:

1. Complete processing: a piece enters from the *First Assembly* and exits from *Final Completion*;
2. Partial processing for coating: a piece enters from the *First Assembly* and exits after this processing. Next, it is sent to the coating;
3. Partial processing by coating: a piece that has been already assembled and coated enters for the *Final Completion*.

For each of the previously defined processes there are several stations in which the piece undergoes a processing phase and is subjected to appropriate tests. We call *station step* a processing phase or each of the scheduled tests. In this line, there are five station steps dedicated to test the calipers:

- High pressure;
- Low pressure;
- Depression;
- Press-fitting of bushings;
- Strain (Deformation).

For each product, there are three chances for passing each single station step. If a caliper fails a step more than three times, it is definitively discarded, otherwise, if it fails any step less than three times, it is recorded as “temporarily discarded”.
Sometimes it can happen that a definitively discarded caliper come back in the production line, after an appropriate reworking. In this case, the caliper serial number (that is the unique identifier

of the single product) is different from the first time the caliper come in the production line, and therefore, this product will belong to a different production order.

In this context we analyze the test of the high pressure. Every day, the line records 250 measures for each caliper and the scraps number related to the high pressure proof is to the order of 2-4 times (a day). Observing the timing chart of this test, the signal shows several peaks that exceed the threshold tolerable (i.e. 1 bar) for the high pressure. When this value goes over the threshold, the caliper is deemed a scrap. However, it can happen that the problem is not caused by the caliper but, rather, by an equipment faulty. Even in this case, the caliper is recorded as defective and, if the problem occurs for three times, the product is removed from the line (although it is not really faulty). Currently, the only way to detect this problem, is to control if there have been a lot of consecutive faults in a short time interval (causing a *drift* behavior during the recording of new calipers). In that case, the problem will be most likely related to an equipment faulty.

6.2 Model Development

The dataset provided by the Company includes the data collected on a particular production line from July 2014 to September 2015. The data is relative to about 140K calipers. The steps required to build the predictive model are the following:

1. Identification of interest data for the analysis;
2. Preliminary analysis and data cleaning;
3. Model development for the automatically detection of the faulty machinery.

6.2.1 Identification of interest data for the analysis

The case study under exam focuses on understanding when the value of high pressure (the measurement HP DELTA-P VALUE) presents an abnormal behavior (i.e. how often this measurement exceeds the tolerance threshold of 1 bar related to a malfunction of the machinery). In addition to this variable (HP DELTA-P VALUE), we considered other measurements that are collected on the line during the calipers processing.
We want to evaluate if, in correspondence of abnormal behavior of the calipers during the HP DELTA-P VALUE test, there also other correlated measures. These additional measures are related to the low pressure test, the strain test, and the vacuum test. Therefore, for the analysis of the problem, the following 13 measurements were considered:

- LP INITIAL VALUE: initial value of the low pressure test;

- LP DELTA-P VALUE (PRESSURE-DROP): value of the pressure loss during the test of low pressure;
- LP EFFICIENCY: efficiency of the low pressure test (it is a test that is performed before the coating of the calipers so, in this step the calipers are all oxidized);
- VACUUM TEST VALUE (LEVEL): value of the vacuum test;
- VACUUM DELTA-P VALUE-LP: value of the loss during the vacuum test;
- VACUUM VACUUM EFFICIENCY: efficiency of the vacuum test;
- HP DEFORM.VALUE CALIPER PRESSURE-1: caliper deformation at pressure 1 during the high pressure test;
- HP DEFORM.VALUE CALIPER PRESSURE-2: caliper deformation at pressure 2 during the high pressure test;
- HP TOTAL DEFORM.VALUE CALIPER-PRESS: sum of the previous two deformations;
- HP TOTAL DEFORM.VALUE CALIPER-RELEASE: residual deformation after the high pressure test;
- HP INITIAL VALUE: initial value of the high pressure test (this measure changes average and variance exactly as it happens in HP DELTA-P VALUE);
- HP DELTA-P VALUE (PRESSURE-DROP): high pressure test loss value (our target variable);
- HP EFFICIENCY: high pressure test efficiency. Test carried out at the end of the high pressure test.

In addition to the above measures listed, we have examined other factors related to the calipers:

- Calipers coated or oxidated and caliper's side (right, left). The possible values are: Oxidate left, Oxidate right, Coated red left, Coated red right, Coated blue left, Coated blue right, Coated black left, Coated black right;
- Maintenance recorded on the high test machinery, low pressure, depression and deformation. Regarding this information we have 50 dates only in which maintenance activities were performed. This was a critical point during our analysis, since this information (useful for validating the model) is not automatically recorded but is manually marked at the discretion of the operators, causing a great loss of information;
- Calipers that has been discarded from the line because they had not passed the HP DELTA-P VALUE test but were checked again on other lines to make sure the problem was related to the HP test machinery (that did not work well) rather than the failed calipers.

6.2.2 Preliminary analysis and data cleaning

In this section we provide some insights about the characteristics of the data and how we prepared them for their processing. Both historical and recent data of the calipers have been considered in the analysis. We have created 13 datasets, each referring to one of the 13 variables described in the previous Section. Each dataset contains the value of the measurement and the timestamps of the event.

6.2.2.1 Analysis of the HP DELTA-P VALUE measure and its different behaviours

In all datasets, the reference timestamps is that related to the HP DELTA-P VALUE measure. We also considered (using the same timestamps) the caliper's coating, the maintenance operations' dates, and if the caliper has been mark as faulty.
In this way we have that, for each time corresponding to the calipers analyzed during the HP test, are associated the coating code, the maintenance code and the code of those "calipers marked as scraps".
Figure 6.1 depicts the values of the HP DELTA-P VALUE of all the calipers recorded over the whole analysis time range. The red line represents the average pressure value (measured in bar) of each day considered.

In Figure 6.1 is evident a change in the average value of about a factor of 2 in the middle of the year. During this preliminary analysis we displayed this behavior to the production managers whose confirmed that there have been problems with the high pressure station in correspondence with that time range. In fact, by dividing all the time space into two ranges (before and after this event), we note that the distribution of the measurements presented in the two histograms in Figure 6.2 is different.

Moreover, Figure 6.1 provides additional information. For instance, it is evident that there are many calipers whose HP DELTA-P VALUE exceeds the maximum acceptability threshold of 1 bar. Finally, we noted that after a certain trend in which the values go above the 1 bar threshold, these measurements suddenly fall back into the correct range. This fast change in the values trend suggests that an external factor has been introduced and it could be motivated by a maintenance activity performed by the line operators. In the following of this analysis, we will compare such scenarios with the effective dates in which maintenance has been performed on the machinery and we will show that a correlation between these two information exists.

Going into more details, we highlighted four different behaviors of the measurements recorded in this period. The first behavior has been called*flat*. In the flat behavior all the measures remains in a certain range constantly. Figure 6.3 shows this is correct behavior.

We have called *drift* the behavior where there is a trend with many close peaks and holes. This is the expected anomalous behavior, on the base of the information provided by the production

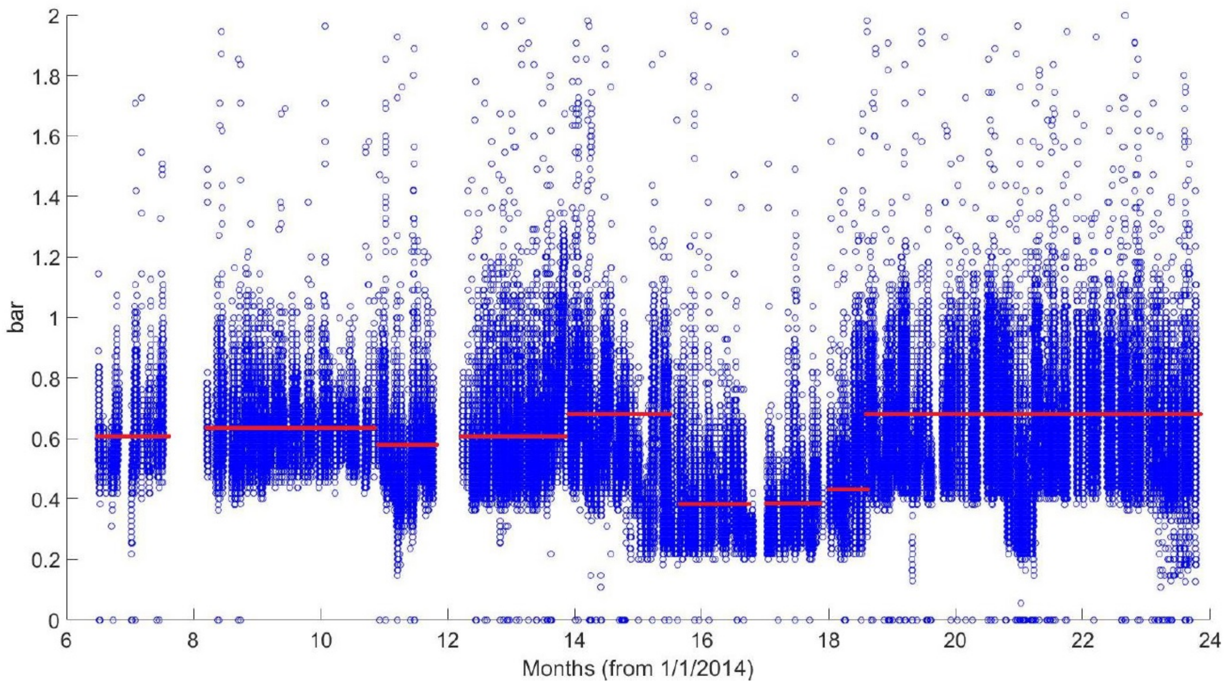


Figure 6.1: HP DELTA-P VALUE measures behaviour within the whole time range considered.

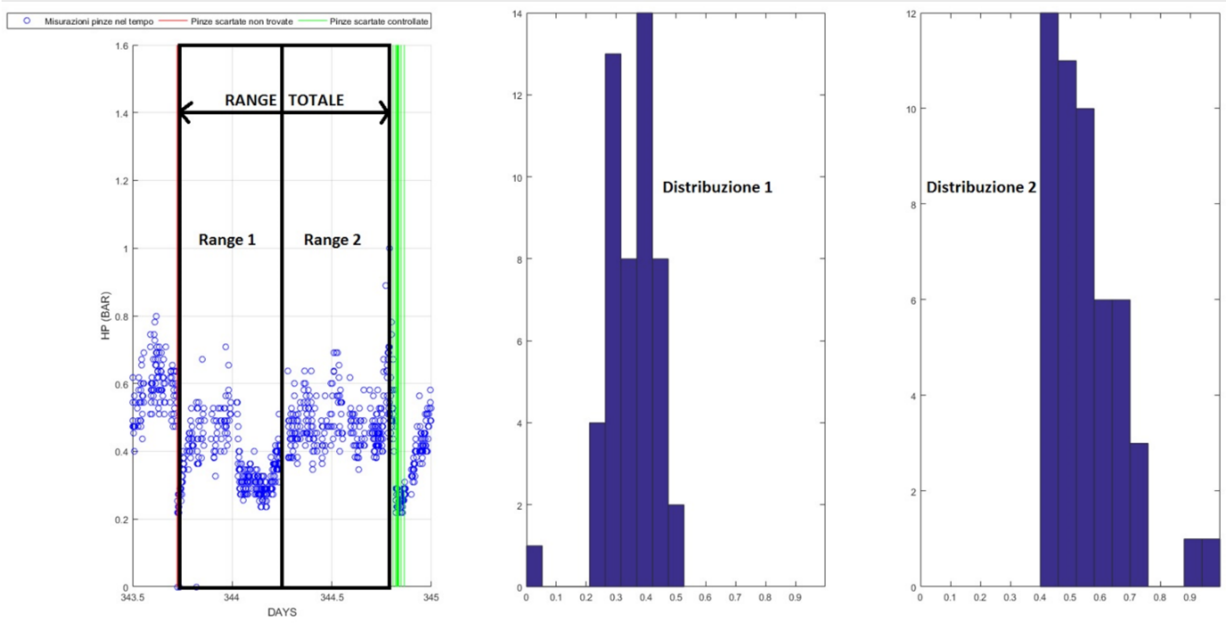


Figure 6.2: Different HP DELTA-P VALUE data distributions before and after a specific event detected in to the production line.

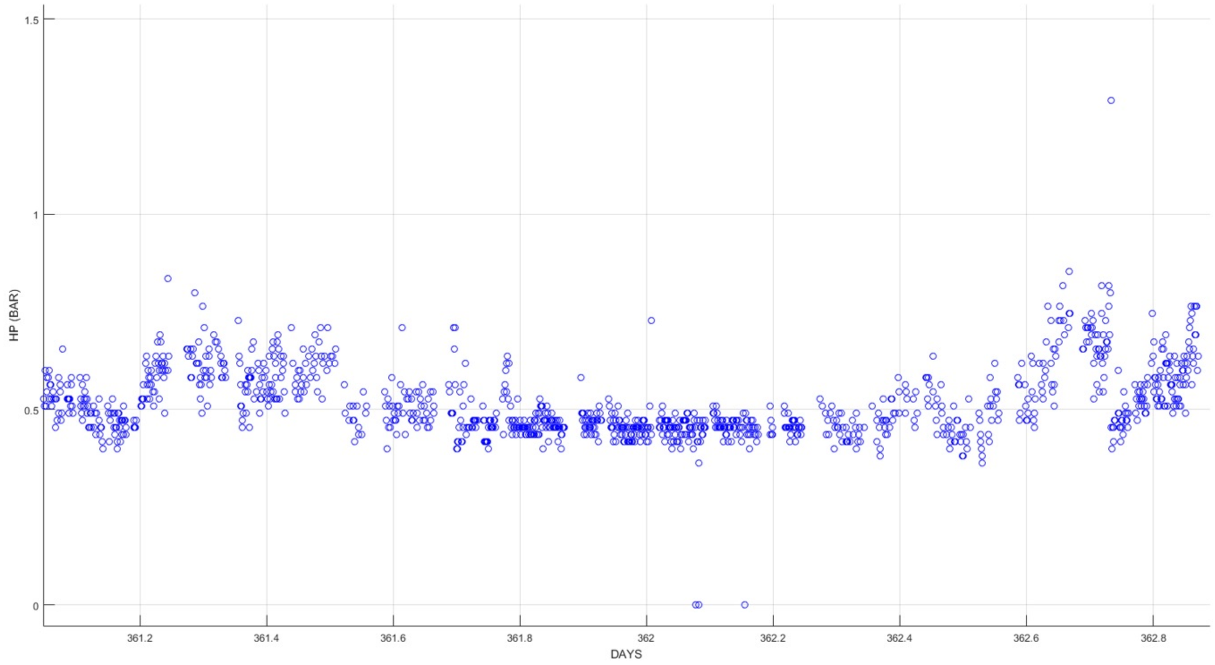


Figure 6.3: The *flat* behavior of the HP DELTA-P VALUE measure.

managers. An example is presented in Figure 6.4.

On the other hand, we encountered also the behavior we called *inverse drift* presented in Figure 6.5. This behavior suggests that as new caliper is recorded, as the machinery performs well. In this case, the machinery, which was behaving abnormally, tends to return within the correct range of measurements.

One last behavior that we found by analyzing the measurements of the HP DELTA-P VALUE, shows a *chaotic* behavior of the values. This is presented in Figure 6.6.

6.2.2.2 Analysis of the other type of measures

Similarly to the analysis of the HP DELTA-P VALUE, we analyzed the other 12 measurements performed on the line. The goal of this analysis is to check if the other measures have similar behaviors in the same time frames where we highlighted the four behaviors on HP DELTA-P VALUE measure. As a result of this preliminary analysis, there seems to be a not clear correlation between the different measures. In Figure 6.7 is depicted an example of comparison between HP DELTA-P VALUE and one of these 12 measures, the LP DELTA-P VALUE.

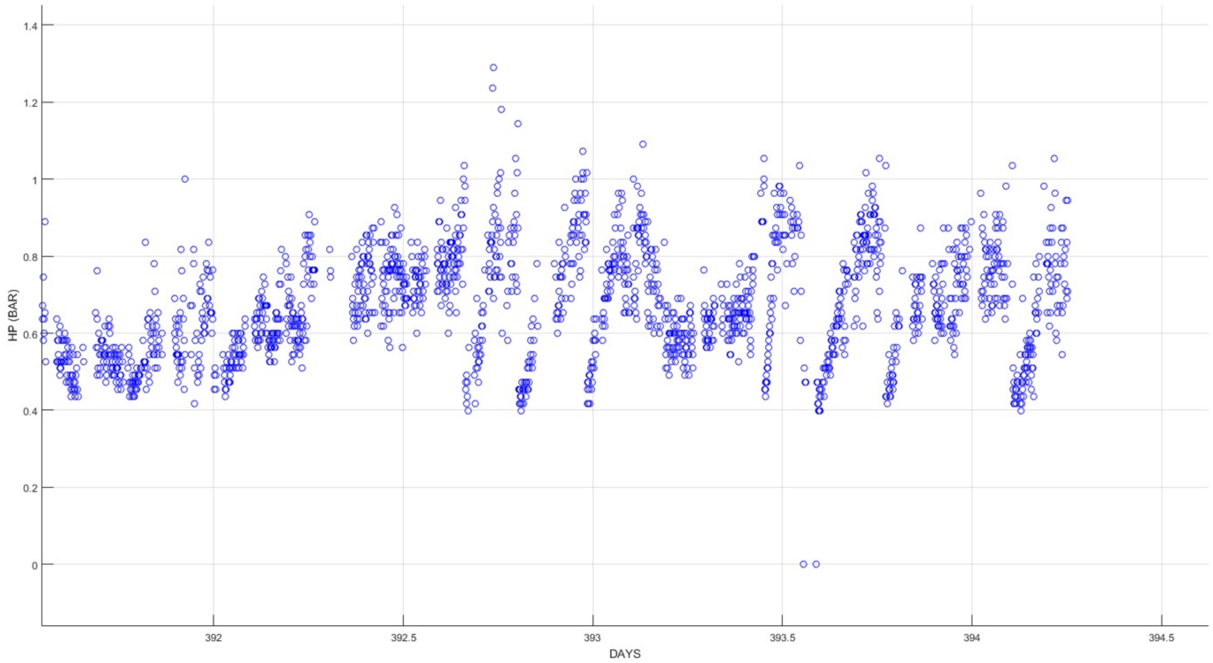


Figure 6.4: The *drift* behavior of the HP DELTA-P VALUE measure.

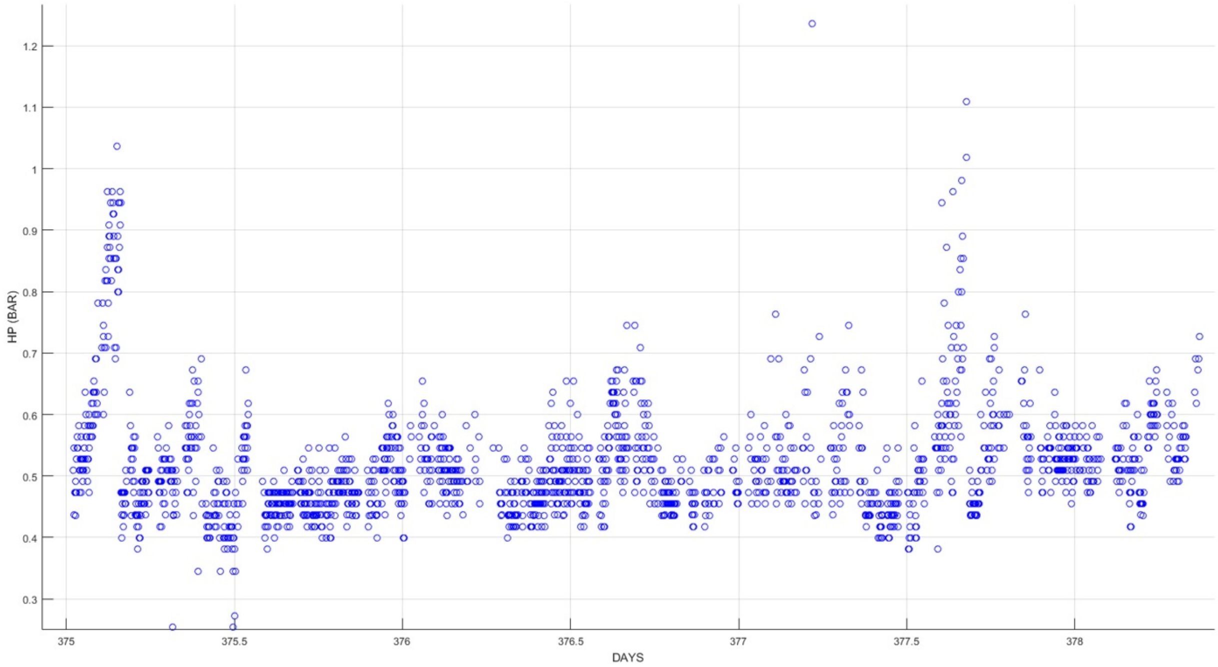


Figure 6.5: The *inverse drift* behavior of the HP DELTA-P VALUE measure.

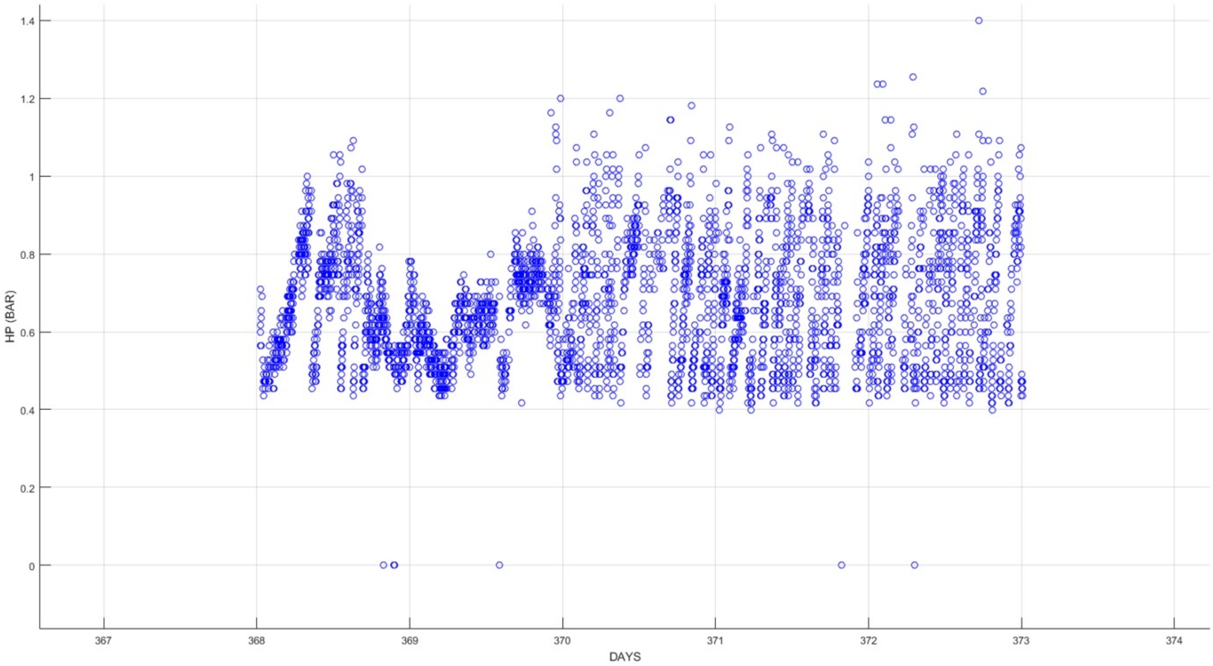


Figure 6.6: The *chaotic* behavior of the HP DELTA-P VALUE measure.

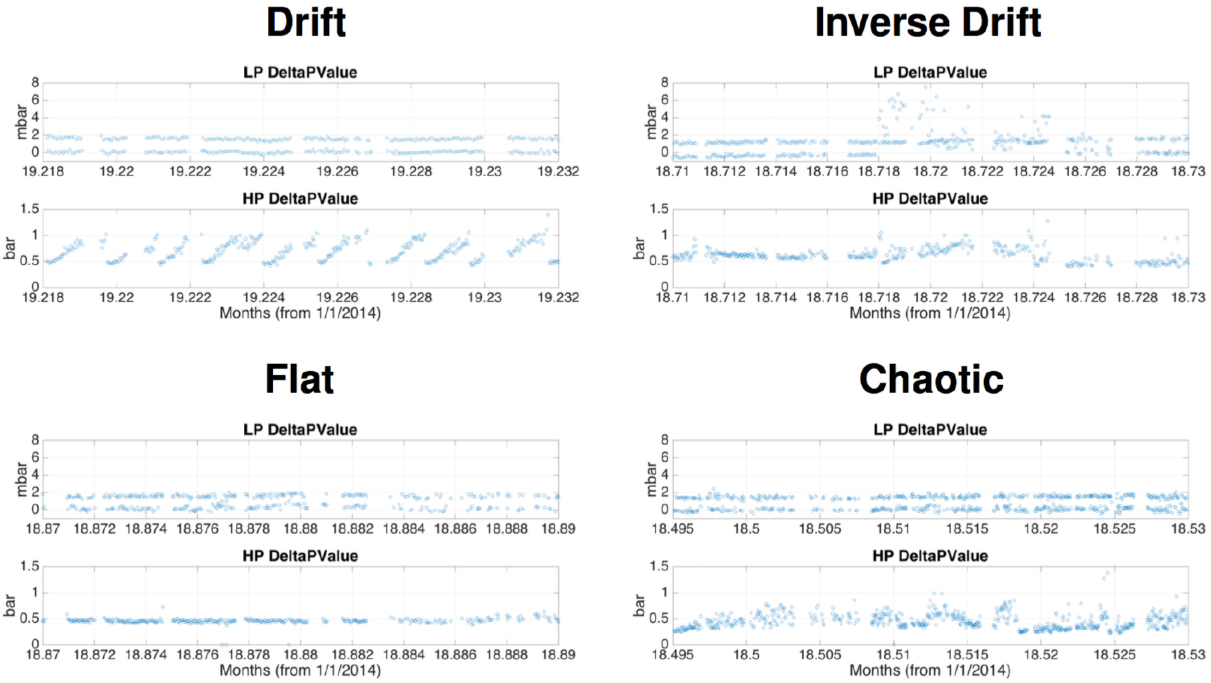


Figure 6.7: Comparison between the four HP DELTA-P VALUE behaviors and the LP-DeltaPValue behaviour, mapped in the same time range.

6.2.2.3 Analysis of the oxidation and coating color

We performed an additional analysis to verify if the behavior of the HP DELTA-P VALUE measure is influenced by oxidation or coating. Furthermore, we differentiate between oxidation or coating on the right or left side of the caliper. For each of these calipers we considered the combination of these two factors (oxidized / coated and right side / left side) in order to check if the measure of high pressure are correlated to these factors. Figure 6.8 and 6.9 presents the result of this analysis. Unfortunately, also in this case, we don't detected an immediate correlation between the different values of the HP DELTA-P VALUE measure and the oxidation/coating of the right/left caliper side.

6.2.2.4 Analysis of the manual maintenance

We evaluated whether the information related to the "corrective manual maintenance" interventions on the different test machineries of the line could have a correlation with the HP DELTA-P VALUE measurement trends. As shown in Figure 6.10, we mapped the dates of the maintenance activities in correspondence with the HP DELTA-P VALUE measurements. The two figures represent different instants in which a maintenance activity has been activated (green line) and terminated (red line).

In correspondence with the data *drift* behavior has been performed a manual maintenance. At the end of this maintenance, the HP measurement fall within the correct range. When a manual maintenance is performed, it is required to stop the machinery from its activity. For this reason, in the first figure there are no measurements between the beginning and the end of this activity. On the other hand, this is not the case for the second figure. The reason is due to the fact that, often, the line operators realize that the measurements are having a strange behavior. Thus, the operators may perform a small maintenance without stopping the production.

Finally, we analyzed the calipers that "strangely" failed the high pressure test and then, checked again on another line, passed the test. We mapped the dates in which these false rejects were recorded to the time range of the HP DELTA-P VALUE's measurements. In Figure 6.11 is reported this analysis.

It is possible to note that, in correspondence of abnormal behavior in the data, there are signals of double checked calipers.

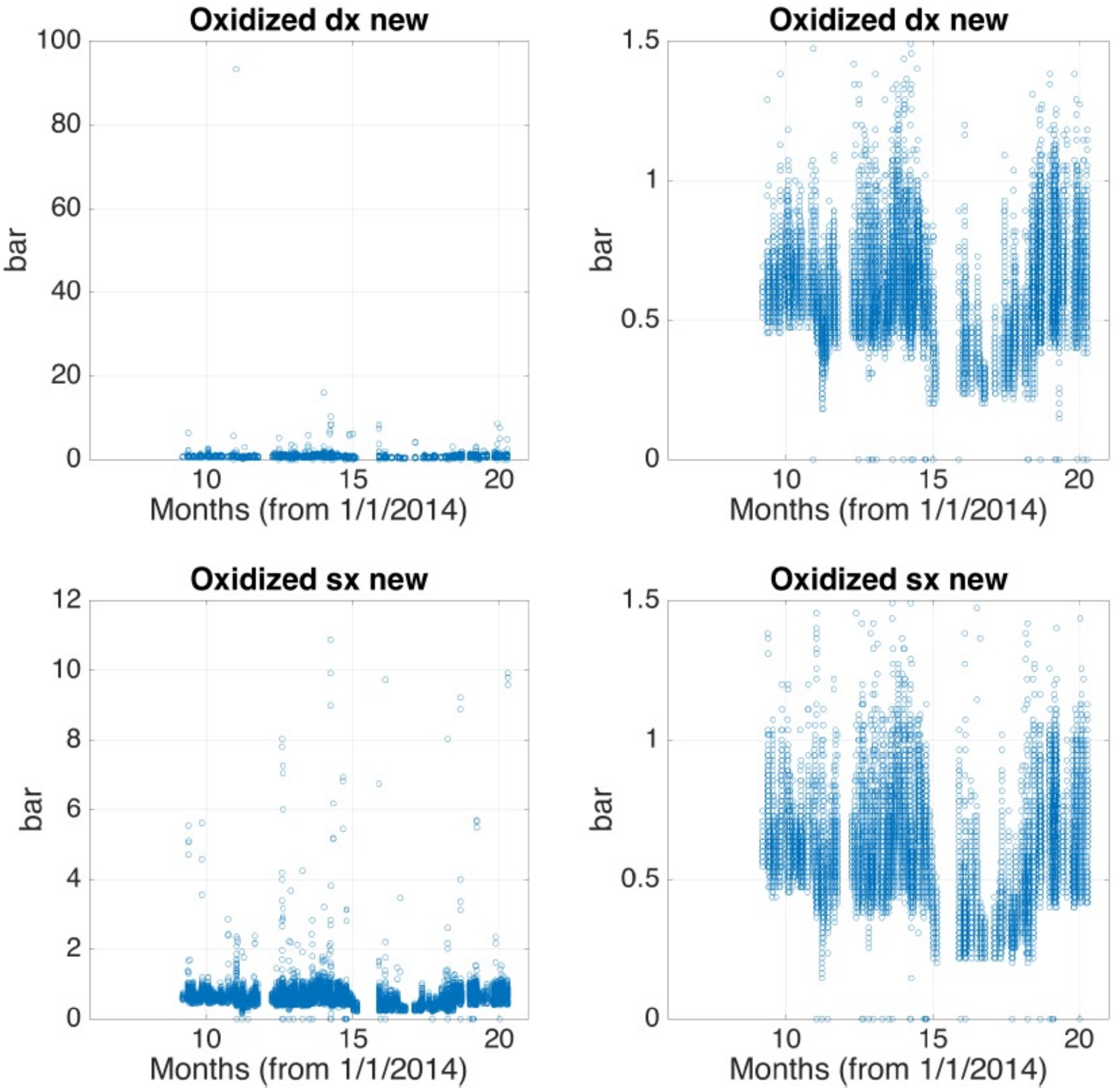


Figure 6.8: Investigation of oxidated (right/left side) calipers respect to the HP DELTA-P VALUE behaviours.

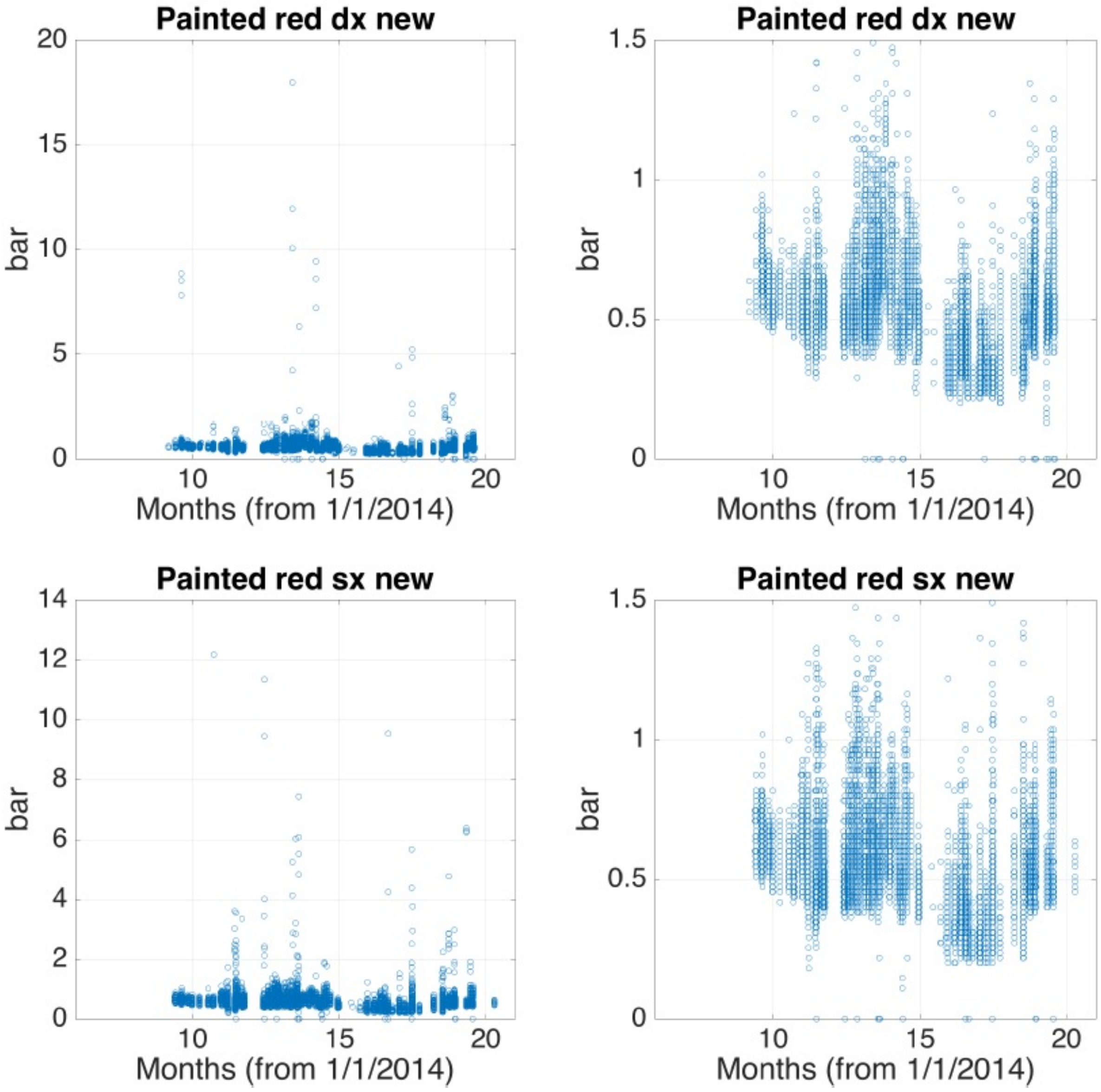


Figure 6.9: Investigation of coated (right/left side) calipers respect to the HP DELTA-P VALUE behaviours.

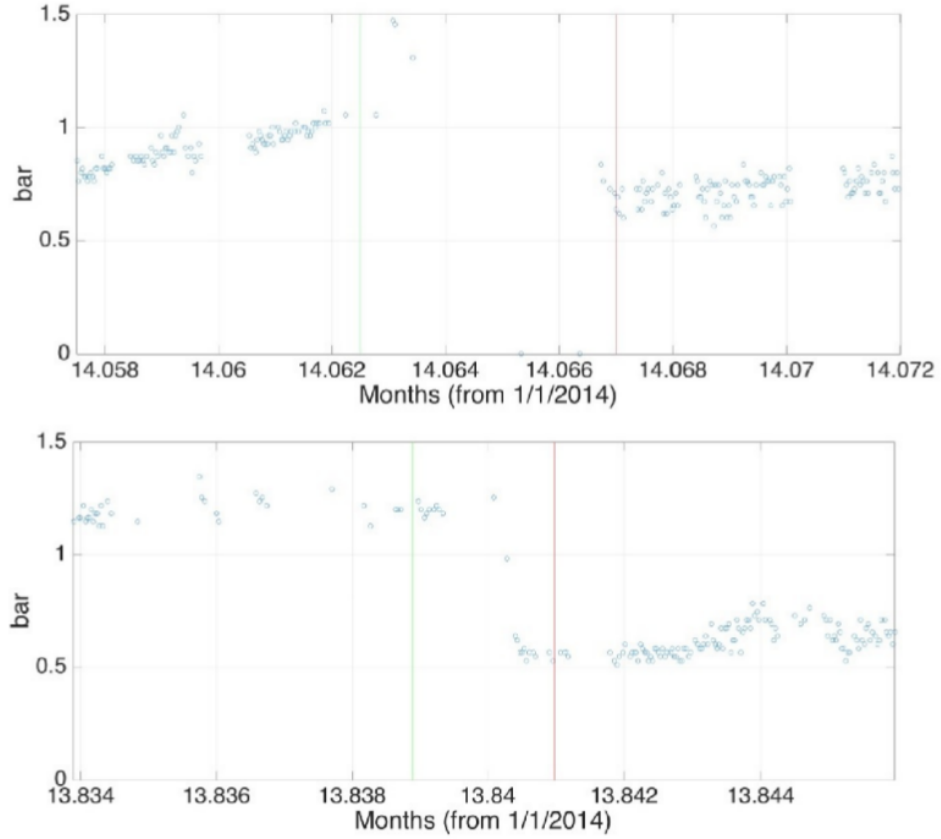


Figure 6.10: Two zoomed time frame where maintenance has been carried out on the HP test machinery. Start (green line) and end dates (red line) of maintenance are mapped to the HP DELTA-P VALUE measures.

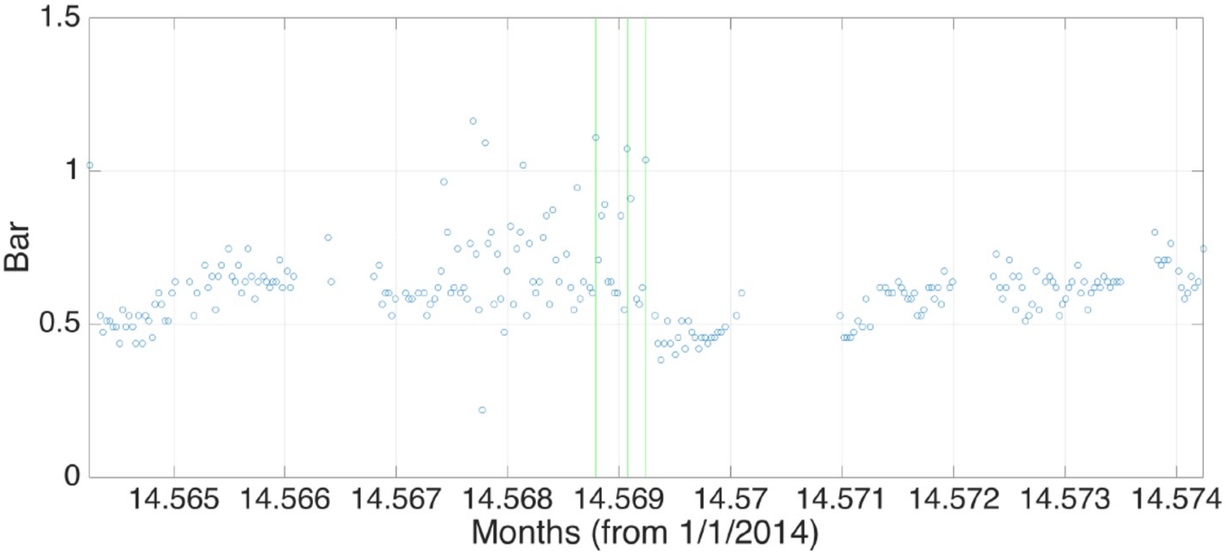


Figure 6.11: Comparison between the double checked calipers and the HP DELTA-P VALUE measures.

6.2.3 Model development for the automatically detection of faulty machinery

Now we present how we created a model able to detect when the HP test machinery is faulty. As a result of the preliminary analysis presented in Subsection 6.2.1 and Subsection 6.2.2, we outline the following four main considerations:

- The high pressure test showed four different behaviors within the time interval considered for the analysis;
- There is no evidence correlation between the anomaly behavior of the high pressure test and the other tests performed on the same assembly line;
- There is a correlation between the anomaly behavior of the high pressure test and the maintenance activities carried out on the production line.
- There is a correlation between the anomaly behavior of the high pressure test and double-checked calipers.

To create the predictive model, we make use of a multivariate and autoregressive approach. We started building a basic model with the variable of interest, the HP DELTA-P VALUE. After, we enlarge our model adding also the other variables (i.e. the other 12 measurements described in Subsection 6.2.1).

These could give indirect information about the operation of the high pressure test machinery. The model has been trained on historical data and it is able to update and perform prediction in real time while new data are recorded.

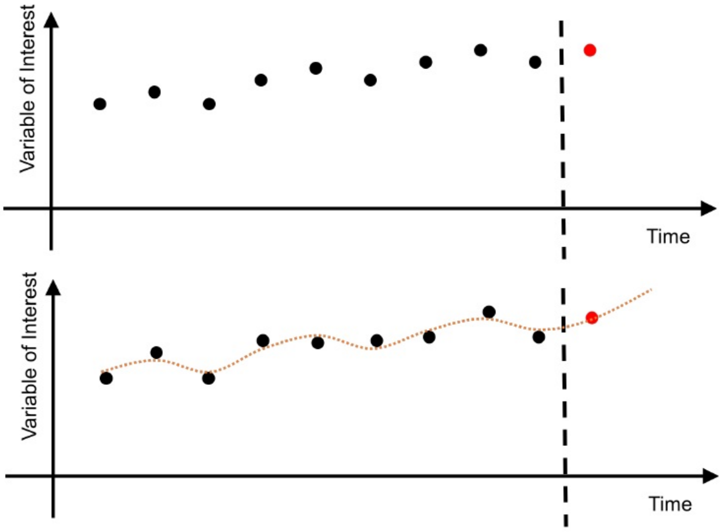


Figure 6.12: First modelling step: transformation of HP DELTA-P VALUE measures in a time series by means a moving average filter.

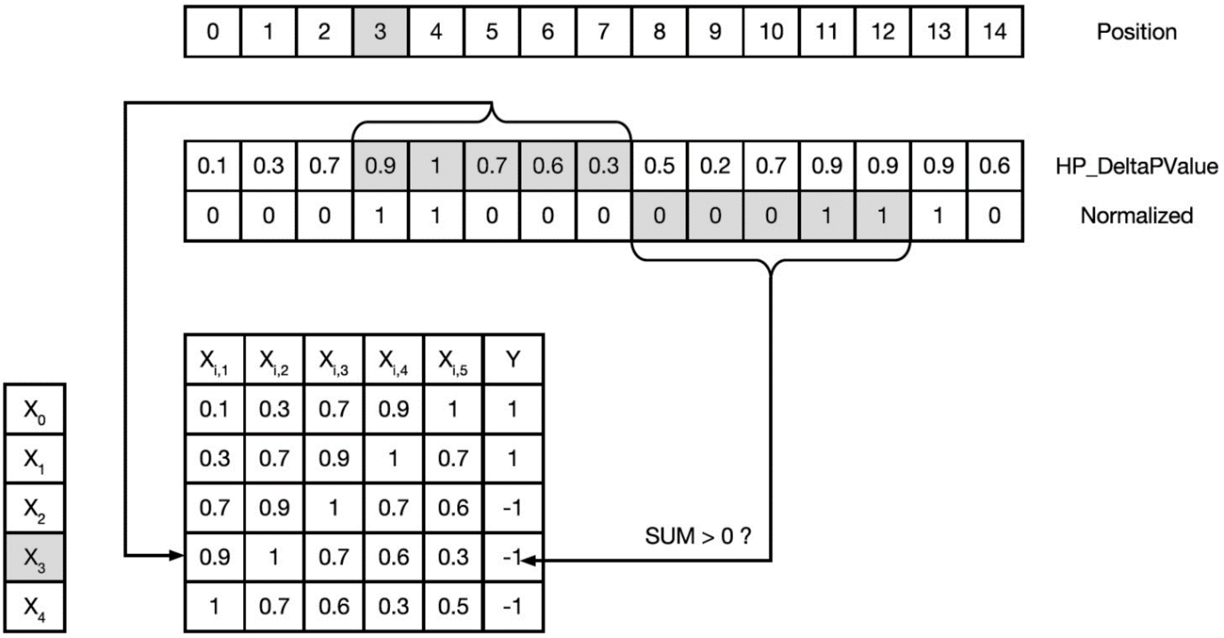
As outlined in Figure 6.12, the first step is to analyze the variable of interest HP DELTA-P VALUE over time. To remove noise and dampen data fluctuations, we constructed a moving average filter with a time window of $T = 10$ samples. More in details, every time a new observation is read, we will use this new information to update the model by applying the average on all the data contained in the time window (the new record plus the 9 previous data). Thus, a time series of the variable of interest (HP DELTA-P VALUE) is obtained in which, each point derived from the average of the previous T values. This procedure was repeated even on the other 12 variables related to the one of interest.

The second step is to transform the time series into a linear regression problem. The goal is to build a matrix similarly to the one presented in Figure 6.13.

First of all, the data of the time series were normalized. A threshold $\Theta = 0.8$ has been defined (since the maximum tolerance is of 1 bar). For each value of the time series we have associated:

- 0, if the value is below or equals to the threshold;
- 1, if the value is above the threshold.

Thus, a pair of values is obtained where, the first is the value of the time series and the second is the normalized value. We can now generate the matrix $Z = [X_{i,j} Y_i]$. Making use of the



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Position	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
HP_DeltaPValue	0.1	0.3	0.7	0.9	1	0.7	0.6	0.3	0.5	0.2	0.7	0.9	0.9	0.9	0.6
Normalized	0	0	0	1	1	0	0	0	0	0	0	1	1	1	0

	$X_{i,1}$	$X_{i,2}$	$X_{i,3}$	$X_{i,4}$	$X_{i,5}$	Y
X_0	0.1	0.3	0.7	0.9	1	1
X_1	0.3	0.7	0.9	1	0.7	1
X_2	0.7	0.9	1	0.7	0.6	-1
X_3	0.9	1	0.7	0.6	0.3	-1
X_4	1	0.7	0.6	0.3	0.5	-1

Figure 6.13: Second modelling step: from a time series to an auto-regressive model.

model, we want to understand if the machinery is faulty looking at the values of the HP DELTA-P VALUE in a short period of time. For this reason, the idea is to observe 5 consecutive records of the time series, and predict if the threshold will be exceeded at least once in the successive 5 records.

In detail, the $X_{i,j}$ are the values of the time series, where:

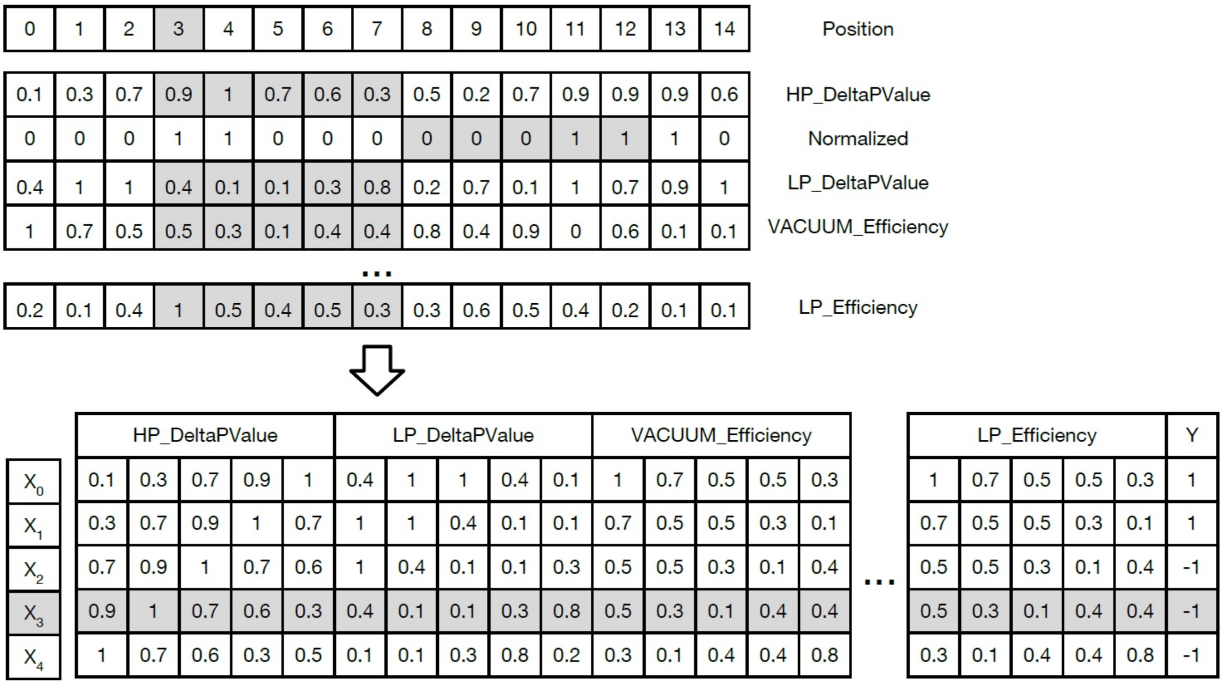
- i ranges from 0 to the total number of points in the time series;
- j varies from 1 to 5, as we have chosen to use 5 observations.

For instance, taking in consideration the values in Figure 6.13 we have that $X_{1,1} = 0.1$ is the first point of the time series; $X_{1,2} = 0.3$ is the second point, and so on. Moreover, in the second line $X_{2,1} = 0.3$ corresponds to $X_{1,2}$.

The column Y maintains the prediction (label) and it can have the following 2 values:

- 1, if the sum of the normalized values of the time series from the position $i + 5$ to $i + 9$ (inclusive) is zero;
- -1, if at least one of these values is greater than zero.

For instance, in the first row of the matrix there are the first 5 values of the time series $(0.1; 0.3; 0.7; 0.9; 1)$. Hence, the value $Y_1 = 1$ because the sum of the values normalized, from position 5 to position 9, is 0. In this way the matrix $Z = [X_{i,j}, Y_i]$ was obtained using only the data of the variable of interest HP DELTA-P VALUE. Finally, the matrix Z has been extended to take into account also the other 12 exogenous variables. Figure 6.14 summarizes the final matrix.



0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Position

0.1	0.3	0.7	0.9	1	0.7	0.6	0.3	0.5	0.2	0.7	0.9	0.9	0.9	0.6	HP_DeltaPValue
0	0	0	1	1	0	0	0	0	0	0	1	1	1	0	Normalized
0.4	1	1	0.4	0.1	0.1	0.3	0.8	0.2	0.7	0.1	1	0.7	0.9	1	LP_DeltaPValue
1	0.7	0.5	0.5	0.3	0.1	0.4	0.4	0.8	0.4	0.9	0	0.6	0.1	0.1	VACUUM_Efficiency

...

0.2	0.1	0.4	1	0.5	0.4	0.5	0.3	0.3	0.6	0.5	0.4	0.2	0.1	0.1	LP_Efficiency

	HP_DeltaPValue					LP_DeltaPValue					VACUUM_Efficiency						LP_Efficiency					Y
X_0	0.1	0.3	0.7	0.9	1	0.4	1	1	0.4	0.1	1	0.7	0.5	0.5	0.3		1	0.7	0.5	0.5	0.3	1
X_1	0.3	0.7	0.9	1	0.7	1	1	0.4	0.1	0.1	0.7	0.5	0.5	0.3	0.1		0.7	0.5	0.5	0.3	0.1	1
X_2	0.7	0.9	1	0.7	0.6	1	0.4	0.1	0.1	0.3	0.5	0.5	0.3	0.1	0.4	...	0.5	0.5	0.3	0.1	0.4	-1
X_3	0.9	1	0.7	0.6	0.3	0.4	0.1	0.1	0.3	0.8	0.5	0.3	0.1	0.4	0.4		0.5	0.3	0.1	0.4	0.4	-1
X_4	1	0.7	0.6	0.3	0.5	0.1	0.1	0.3	0.8	0.2	0.3	0.1	0.4	0.4	0.8		0.3	0.1	0.4	0.4	0.8	-1

Figure 6.14: Third modelling step: multivariate auto-regressive model.

Thus, each exogenous variable contributes to extending each row $X_{i,j}$ of the matrix with the values of the 12 variables. Also in this case, in each row, for each of the variable, we considered the 5 values registered in the same timestamp of the 5 values considered for the HP DELTA-P VALUE.

Finally, the column Y does not change because it constitutes the variable of interest, intended as an indicator of exceeding the HP DELTA-P VALUE threshold in the 5 times after the ones considered. Such matrix is split in two sets, the 50% of the data is used to generate the training set and the remaining 50% constitutes the test set.

The matrix contains $140K$ rows and 66 columns (65 features plus the prediction column).

To create the predictive model, we applied the RF learning algorithm. Based on the outcome discussed in Chapter 4.2, we preformed a model selection making use of several values for the hyperparameters. We get, as a result, that the best performing combination of hyperparameters is the following:

- Percentage of input data to sample with replacement from the input data, for growing each new tree $= 1$;

- Number of trees $= 500$;
- Number of features to select at random for each decision split $= 21.66$. This values has been computed as $\frac{number of features}{3}$;
- Trees are grown to the maximum possible.

Finally, to each tree has been assigned the same weight in the final voting.
The model accuracy in predicting the average of the HP DELTA-P VALUE is approximately $95 - 98\%$.
Figure 6.15 and Figure 6.16 depict two examples of the model simulated in real time: the blue points represent the real data of the HP DELTA-P VALUE, the red line is the time series obtained by applying the moving average filter, while the green line is the forecast obtained by the model. As it is possible to note, the model is able to predict the progress of the HP DELTA-P VALUE in an almost correct way both in normal cases (Figure 6.15) and in the case of anomalous behaviors (Figure 6.16).

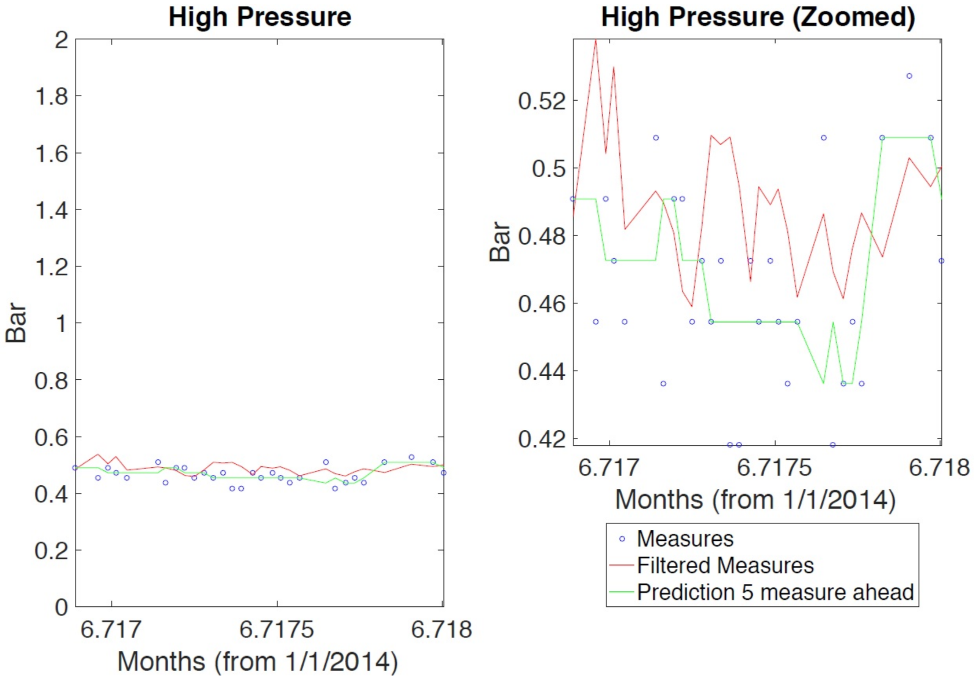


Figure 6.15: Still image of the real time model simulation: time frame in which the model predicts normal behaviour.

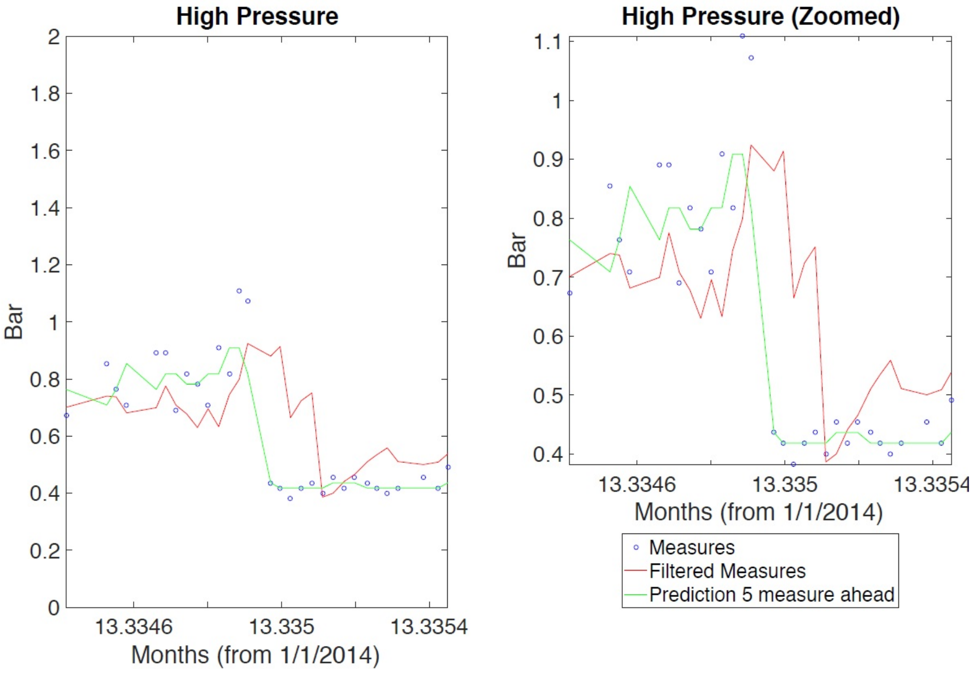


Figure 6.16: Still image of the real time model simulation: time frame in which the model predicts drift behaviour.

Chapter 7

Conclusion and Further Developments

Is it possible to use the data collected by the manufactures in a different way from the one for which they were conceived?

It is very common that manufactures collect data within their production processes, but until recent years, the potential advantages deriving from their integration and use has not been exploited ([LLBK13, SS13, DD13, LKY14]). Thanks to the development of new technologies and techniques, manufactures have understood the relevance of the interconnection between the different components of their systems and the amount of information that derives from them. The acquisition of these new techniques within their production processes could allow the next generation of manufacturing systems to become real smart factories. Digital transformation is one of the main aspects emerged by the current 4.0 revolution. It embraces the integration between the digital and physical environment, including the application of modelling and simulation techniques, visualization, and data analytics in order to manage the overall product life cycle.

In this thesis we wanted to emphasise two macro areas of analysis that can be done thanks to the technologies provided by this manufacturing digitalization.
The first macro area concerns the *descriptives/diagnostic* analysis that can be performed on data by means the *Manufacturing Intelligence* tools ([WW07, Neg04, CDN11, CCS12]).
Instead, the second macro area identifies the *Manufacturing Analytics* techniques, which enable to perform *predictive/perscriptive* analysis ([CV70, SSJ14, MSDS14, Hui15, CSM12, EL12, BBA13, LLS$^+$11, SN87, KRS02, TMDOL10, Lie13, SS05, SSM12, SMK14]). The latter macro area gives added value to manufacturing industries since these techniques enables to extract new insights from big data, finding new pattern within the data themselves that were not possible to identify with the previous approaches ([R$^+$11, SB12b, MSC13]).
In this thesis we proposed several contributions to improve certain aspects of these two macro areas.

In the framework of Manufacturing Intelligence, whose objective is to perform analysis aimed to manage and optimize the production systems, we presented two achievements.

The first one concerns the application of a methodology that, starting from the data collected at the factory, enables to identify a set of *standard* KPIs (*standard* means that these indicators are defined in literature) useful to control production processes. This methodology develops from the definition of a data architecture in which to systematize the main work flows and entities (operators, resources, work orders, equipments, and so on) of a manufacturing system. Then, we exploited this data base in order to calculate a set of strategic KPIs related to the production operations assessment, and the inventory management. Finally, we have made this information more usable by displaying it in a set of interactive dashboards. This contribute can be viewed, for instance, as a support for manufacturing companies that need to organize their big and scattered data.

Moreover, another strategic priority that emerged in this 4.0 revolution and that is becoming increasingly important among the industries' critical objectives, is the sustainability ([TET13, TSM15, BME13, LYQ$^+$12]).
Manufacturing industries are increasingly integrating the so-called "green practices" into their production processes. This implies that, in addition to evaluating the indicators (already known and defined by many standards) aimed at measuring the production performances, it is necessary to consider also other indicators that meet the sustainability constraints (for example energy efficiency, material efficiency, waste reduction, and so on). Since the sustainability is a new strategic dimension, there is a lack of understanding of how it may be effectively embedded in corporate performance management systems especially on the manufacturing area ([MML$^+$13, KMK15, JH04]).
Our contribute was to provide a guideline that identifies the main sustainability trends and drivers in manufacturing environment and a set of KPIs that allows industries to control and monitor the reaching of these goals. This sustainability guideline was developed by performing a qualitative literature review on industrial sustainability performance management with the aim of analyzing the existing state of the art, re-organizing these information accordingly to a hierarchy of structured KPIs, and highlighting the areas that need to be further developed both in the literature and in the industrial systems.

The second group of results presented in this thesis, concerns two contributions in the field of Manufacturing Analytics.
The first one is the application of two state-of-the-art ML techniques, Support Vector Machine and Random Forest algorithms, respectively for quality estimation of the final products in an assembly line of refrigerators and for predictive maintenance of a machinery in a production line of calipers.
Quality evaluation and predictive maintenance are two fundamental activities within a production process and multi-sensing technologies embedded in production lines collect vast amounts of data, which can be analyzed. Good Quantity (the ratio between good parts and inspected parts) and Corrective Maintenance Ratio (the ration between corrective maintenance time and

total maintenance) have been selected by the ISO 22400 standard as two of the main KPIs for manufacturing systems, which shows the importance of these figures and the need for predicting them and to eventually start proactive actions on the production line.
As a result, these two applications show the ability of the selected ML algorithms in building effective models for the problems in question.
Finally, the second contribution concerns the way by which we developed the SVM and the RF models for these two manufacturing applications. The assessment of the predictive models exploited for manufacturing analysis, is a quite rare topic in the scientific literature [OOA15]. To this end, we applied some methods of Statistical Learning Theory to the SVM classification model. We chose these techniques because are suitable for big data problems ([WWIT16, KAAD01, EPPV02]). Our purpose was to analyse the advantages and disadvantages of the SLT framework, which is still an open field of research. As a result we observed that these approaches are able to rigorously assess the performance and reliability of our predictive model. This is an important aspect, especially when applied to manufacturing systems, where, differently from other areas (e.g. social or economical sciences), dealing with uncertainty is a mandatory issue that must be rigorously addressed.
Instead, in order to build the RF predictive model, we performed a model selection procedure. A common misconception about RF is to consider this algorithm as an hyperparameter-free learning algorithm ([Bia12, BHA09]). Examining the reference RF method introduced by Breiman, we identified several hyperparameters which characterise the performance of the final model and we verified that, by performing a model selection procedure on these hyperparameters, it can improve the accuracy of our final model.

Finally, starting from the results achieved in this thesis, there exist many researches that can be conducted:

- Although manufactures are sensible to the new possibility and benefits of the Industry 4.0, we think that the majority of them still lack in the application of these methodologies. At the same time, usually it can happens that factories invest on powerful visualization and analysis tools in-house, but their staff is not trained so to use these tools to their full potential. For this reason, the first step is to increase knowledge in the use of these analytics tools through company training. Then, it could be helpful to define a set of *standard* (i.e. industry-neutral) methodologies in order to support manufacturers to independently carry out their analyzes. These methodologies will have to provide several guidelines in order to support users both in the systematization of their factory data and in the selection of those Manufacturing Analytics techniques that are most appropriate for the problem in question.

- Methodologies can provide a reference model in supporting manufacturers to define the theoretical steps to produce for their analysis. In addition, these analytics tools should provide special features in supporting the management of big data in manufacturing. In fact, there are a lot of Analytics platforms that provide powerful features for managing data but they are general purpose [TK09]. Therefore, a future innovation of these tools in the manufacturing field, could be to provide specific features in order to support users in their data pre-processing phase: understand which data are really useful to collect, how to prepare them for subsequent queries, and understand the meaning of this information in order to apply the most appropriate methods of analysis. These specific features will allow to know what are the useful information to work on, avoiding wasting time for thinking about how to use the large amount of information collected. Moreover, in order to guide manufactures during these steps, it is necessary that these "future" tools will include semantic features in order to identify associations and complex relationships among data extracted from heterogeneous sources. In the field of bioinformatics, financial services, national security there are already tools that allow automatic extraction of semantic meta data [KMSZ06]. In the same way, these tools could be improved in order to provide advanced features able to extract semantics among heterogeneous data coming from the manufacturing domain.

- Concerning Sustainability, there isn't an analytical method to measure these indicators in manufacturing. We provide a preliminary review on sustainability trends and implications that affect the manufacturing company, creating a set of structured KPIs on which we highlighted limitations and missing areas. A future development of this work can be to take into account also other analysis dimension during the modelling of the KPIs back-end architecture, such as resources in addition to raw materials only (e.g. sheets, foils, etc.). Otherwise, instead of considering data related to the production or financial only, it could be useful to take into account also data related to the recycling processes. Or even, consider metrics that measure the environmental and social impact as well as the aspect purely related to the production. Metrics aimed at measuring the aforementioned aspects already exist, but the next step will be the integration of all these scattered information within the same analysis.

- About the topic of model evaluation in manufacturing applications, we focused on the application of SLT to a predictive model for a smart manufacturing system consisting of a SVM binary classifier. SLT has proved to be an effective approach in many other areas like multi-class classification, regression, density estimation, and so on. Furthermore, SLT can be applied to other settings, which are interesting for manufacturing systems. One, for example, is the semi-supervised setting, where only few labelled samples are available, while the largest part of the dataset is composed of unlabelled data ([CSZ10, AGOR14]). When the collection of labels is expensive or difficult, this approach allows exploiting all the available information for building effective models. Another interesting setting is when physical models of the phenomenon under exam are available: SLT can provide rigorous results about the final predicting model, even when the physical model and the data driven one are combined [COBA15b]. In summary, we believe that SLT can provide a large set of effective tools to the field of smart manufacturing systems for building effective predictive models and rigorously assessing their quality.

- Finally, since the amount of data handled by the manufactures is always increasing, it could be interesting to exploit other recent techniques that are able to process large amount of data. From the many alternatives, the distributed processing of the data allows to exploit cloud infrastructure and to speed-up the computation dividing the analysis on many machines. In this context, recently it have been proposed some popular projects like MLlib but many problems are still open, for instance, the porting of additional techniques or the adaptation of these techniques to a real time processing.

Bibliography

[AA11] Américo Azevedo and António Almeida. Factory templates for digital factories framework. *Robotics and Computer-Integrated Manufacturing*, 27(4):755–771, 2011.

[ABT02] D. Anguita, A. Boni, and L. Tagliafico. Svm performance assessment for the control of injection moulding processes and plasticating extrusion. *International Journal of Systems Science*, 33(9):723–735, 2002.

[AC10] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[Ack70] Russell Ackoff. A concept of corporate planning. *Long Range Planning*, 3(1):2–8, 1970.

[ACO12] Peter O Akadiri, Ezekiel A Chinyio, and Paul O Olomolaiye. Design of a sustainable building: A conceptual framework for implementing sustainability in the building sector. *Buildings*, 2(2):126–152, 2012.

[ACS$^+$11] Asrul Adam, Lim Chun Chew, Mohd Ibrahim Shapiai, Lee Wen Jau, Zuwairie Ibrahim, and Marzuki Khalid. A hybrid artificial neural network-naive bayes for solving imbalanced dataset problems in semiconductor manufacturing test process. In *Hybrid Intelligent Systems (HIS), 2011 11th International Conference on*, pages 133–138. IEEE, 2011.

[AD02] M Munir Ahmad and Nasreddin Dhafr. Establishing and improving manufacturing performance measures. *Robotics and Computer-Integrated Manufacturing*, 18(3):171–176, 2002.

[AGOR11] Davide Anguita, Alessandro Ghio, Luca Oneto, and Sandro Ridella. In-sample model selection for support vector machines. In *International Joint Conference on Neural Networks*, 2011.

[AGOR12a] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. In-sample and out-of-sample model selection and error estimation for support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, 23(9):1390–1406, 2012.

[AGOR12b] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. In-sample model selection for trimmed hinge loss support vector machine. *Neural processing letters*, 36(3):275–283, 2012.

[AGOR12c] Davide Anguita, Alessandro Ghio, Luca Oneto, and Sandro Ridella. In-sample model selection for trimmed hinge loss support vector machine. *Neural processing letters*, 36(3):275–283, 2012.

[AGOR14] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. Unlabeled patterns to tighten rademacher complexity error bounds for kernel classifiers. *Pattern Recognition Letters*, 37:210–219, 2014.

[AGRS09] Davide Anguita, Alessandro Ghio, Sandro Ridella, and Dario Sterpi. K-fold cross validation for error rate estimate in support vector machines. In *International Conference on Data Mining*, 2009.

[AS15] Payman Ahi and Cory Searcy. An analysis of metrics used to measure performance in green and sustainable supply chains. *Journal of Cleaner Production*, 86:360–377, 2015.

[ASK$^+$13] Ali Azadeh, Morteza Saberi, Ahmad Kazem, Vahid Ebrahimipour, A Nourmohammadzadeh, and Zahra Saberi. A flexible algorithm for fault diagnosis in a centrifugal pump with corrupted data and noise based on ann and support vector machine with hyper-parameters optimization. *Applied Soft Computing*, 13(3):1478–1485, 2013.

[ASMO07] O Addin, SM Sapuan, E Mahdi, and M Othman. A naive-bayes classifier for damage detection in engineering materials. *Materials & design*, 28(8):2379–2386, 2007.

[BB05] Mauro Brunato and Roberto Battiti. Statistical learning theory for location fingerprinting in wireless lans. *Computer Networks*, 47(6):825–845, 2005.

[BBA13] Arindam Banerjee, Tathagata Bandyopadhyay, and Prachi Acharya. Data analytics: Hyped up aspirations or true potential? *Vikalpa*, 38(4):1–12, 2013.

[BBL02] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113, 2002.

[BBM05] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[BCC$^+$02] Donald J Bowersox, David J Closs, M Bixby Cooper, et al. *Supply chain logistics management*, volume 2. McGraw-Hill New York, NY, 2002.

[BCM97] Umit S Bititci, Allan S Carrie, and Liam McDevitt. Integrated performance measurement systems: a development guide. *International journal of operations & production management*, 17(5):522–534, 1997.

[BDL08] Biau, Luc Devroye, and Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(Sep):2015–2033, 2008.

[BDT$^+$17] Niklas Burger, Melissa Demartini, Flavio Tonelli, Freimut Bodendorf, and Chiara Testa. Investigating flexibility as a performance dimension of a manufacturing value modeling methodology (mvmm): a framework for identifying flexibility types in manufacturing systems. *Procedia CIRP*, 63:33–38, 2017.

[BE02] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[Bea98] Benita M Beamon. Supply chain design and analysis:: Models and methods. *International journal of production economics*, 55(3):281–294, 1998.

[BHA09] S. Bernard, L. Heutte, and S. Adam. Influence of hyperparameters on random forest accuracy. In *International Workshop on Multiple Classifier Systems*, 2009.

[Bia12] G. Biau. Analysis of a random forests model. *JMLR*, 13(1):1063–1095, 2012.

[Bis95] B. M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[BK14] D. Berend and A. Kontorovitch. Consistency of weighted majority votes. In *NIPS*, 2014.

[Bla04] Benjamin S Blanchard. *System engineering management*. John Wiley & Sons, 2004.

[BM03] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[BME13] Nancy Bocken, David Morgan, and Steve Evans. Understanding environmental performance variation in manufacturing companies. *International Journal of Productivity and Performance Management*, 62(8):856–870, 2013.

[BO04] Bilge Bilgen and Irem Ozkarahan. Strategic tactical and operational production-distribution models: a review. *International Journal of Technology Management*, 28(2):151–171, 2004.

[Bre01a] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[Bre01b] L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[BRS15] NMP Bocken, P Rana, and SW Short. Value mapping for sustainable business thinking. *Journal of Industrial and Production Engineering*, 32(1):67–81, 2015.

[BSW95] Jim Browne, PJ Sackett, and J Cl Wortmann. Future manufacturing systems—towards the extended enterprise. *Computers in industry*, 25(3):235–254, 1995.

[BWFF09] Cecil C Bozarth, Donald P Warsing, Barbara B Flynn, and E James Flynn. The impact of supply chain complexity on manufacturing plant performance. *Journal of Operations Management*, 27(1):78–93, 2009.

[CAJ98] Richard B Chase, Nicholas J Aquilano, and F Robert Jacobs. *Production and operations management*. Irwin/McGraw-Hill,, 1998.

[Cas] L Cassettari. Digitalization of manufacturing execution systems: the core technology for realizing future smart factories.

[Cat07] O. Catoni. *Pac-Bayesian Supervised Classification*. Institute of Mathematical Statistics, 2007.

[CBPT09] Maritza Correa, C Bielza, and J Pamies-Teixeira. Comparison of bayesian networks and artificial neural networks for quality detection in a machining process. *Expert systems with applications*, 36(3):7270–7279, 2009.

[CCS12] Hsinchun Chen, Roger HL Chiang, and Veda C Storey. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4), 2012.

[CDN11] Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya. An overview of business intelligence technology. *Communications of the ACM*, 54(8):88–98, 2011.

[CFI15] CFI. Roadmap fabricaintelligente. `www.fabbricaintelligente.it/wp-content/uploads/Booklet-Fabbrica-Intelligente-2015-PAGINE-SINGOLE.pdf`, 2015.

[Che97] V. Cherkassky. The nature of statistical learning theory. *IEEE Transactions on Neural Networks*, 8(6):1564–1564, 1997.

[Chi02] Ratna Babu Chinnam. Support vector machines for recognizing shifts in correlated and other manufacturing processes. *International Journal of Production Research*, 40(17):4449–4466, 2002.

[CHIS15] Danfang Chen, Steffen Heyer, Suphunnika Ibbotson, and Thiede Sebastian Salonitis, Konstantinos. Direct digital manufacturing: definition, evolution, and sustainability implications. *Journal of Cleaner Production*, 107:615–625, 2015.

[CK00] Charu Chandra and Sameer Kumar. Supply chain management in theory and practice: a passing fad or a fundamental change? *Industrial Management & Data Systems*, 100(3):100–114, 2000.

[CLXL09] Jian Cai, Xiangdong Liu, Zhihui Xiao, and Jin Liu. Improving supply chain performance management: A systematic approach to analyzing iterative kpi accomplishment. *Decision support systems*, 46(2):512–521, 2009.

[CM07] Sunil Chopra and Peter Meindl. Supply chain management. strategy, planning & operation. *Das summa summarum des management*, pages 265–275, 2007.

[CML14] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.

[COBA15a] A. Coraddu, L. Oneto, F. Baldi, and D. Anguita. A ship efficiency forecast based on sensors data collection: Improving numerical models through data analytics. In *IEEE OCEANS*, 2015.

[COBA15b] Andrea Coraddu, Luca Oneto, Francesco Baldi, and Davide Anguita. Ship efficiency forecast based on sensors data collection: Improving numerical models through data analytics. In *OCEANS 2015-Genova*, pages 1–10. IEEE, 2015.

[CSC17] Xu Chi, Tan Puay Siew, and Erik Cambria. Adaptive two-stage feature selection for sentiment classification. In *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*, pages 1238–1243. IEEE, 2017.

[CSM12] Ranko Cosic, Graeme Shanks, and Sean Maynard. Towards a business analytics capability maturity model. In *ACIS 2012: Location, location, location: Proceedings of the 23rd Australasian Conference on Information Systems 2012*, pages 1–11. ACIS, 2012.

[CSS$^+$15] Diego Cabrera, Fernando Sancho, René-Vinicio Sánchez, Grover Zurita, Mariela Cerrada, Chuan Li, and Rafael E Vásquez. Fault diagnosis of spur gearbox based on random forest and wavelet packet decomposition. *Frontiers of Mechanical Engineering*, 10(3):277–286, 2015.

[CSZ10] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 2010.

[CTF$^+$14] M. Colledani, T. Tolio, A. Fischer, B. Lung, G. Lanza, R. Schmitt, and J. Vancza. Design and management of manufacturing systems for production quality. *CIRP Annals-Manufacturing Technology*, 63(2):773–796, 2014.

[CV70] Larry J Connor and Warren H Vincent. A framework for developing computerized farm management information. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, 18(1):70–75, 1970.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[CY09] Jin-Ding Cai and Ren-Wu Yan. Fault diagnosis of power electronic circuit based on random forests algorithm. In *Natural Computation, 2009. ICNC'09. Fifth International Conference on*, volume 2, pages 214–217. IEEE, 2009.

[CYX10] F. Cheng, J. Yu, and H. Xiong. Facial expression recognition in jaffe dataset based on gaussian process classification. *IEEE Transactions on Neural Networks*, 21(10):1685–1690, 2010.

[CZ14] CL Philip Chen and Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

[CZW$^+$07] E. Y. Chang, K. Zhu, H. Wang, H. Bai, J. Li, Z. Qiu, and H. Cui. Psvm: Parallelizing support vector machines on distributed computers. In *Neural Information Processing Systems*, 2007.

[Dav06] Thomas H Davenport. Competing on analytics. *harvard business review*, 84(1):98, 2006.

[Dav12] Thomas H Davenport. Business intelligence and organizational decisions. *Organizational Applications of Business Intelligence Management: Emerging Trends: Emerging Trends*, page 1, 2012.

[DD13] Thomas H Davenport and Jill Dyché. Big data in big companies. *International Institute for Analytics*, 3, 2013.

[Dem13] M Demetgul. Fault diagnosis on production systems with support vector machine and decision trees algorithms. *The International Journal of Advanced Manufacturing Technology*, pages 1–12, 2013.

[DEP$^+$12] Jim Davis, Thomas Edgar, James Porter, John Bernaden, and Michael Sarli. Smart manufacturing, manufacturing intelligence and demand-dynamic performance. *Computers & Chemical Engineering*, 47:145–156, 2012.

[DFL12] Stefanos Doltsinis, Pedro Ferreira, and Niels Lohse. Reinforcement learning for production ramp-up: A q-batch learning approach. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 1, pages 610–615. IEEE, 2012.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer, 1996.

[Dha13] V. Dhar. Data science and prediction. *Communications of the ACM*, 56(12):64–73, 2013.

[DOS99] Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. *Physical review letters*, 82(14):2975, 1999.

[DOTA16] Melissa Demartini, Ilenia Orlandi, Flavio Tonelli, and Davide Anguita. Investigating sustainability as a performance dimension of a novel manufacturing value modeling methodology (mvmm): from sustainability business drivers to relevant metrics and performance indicators. *XXI Summer School "Francesco Turco*, pages 262–270, 2016.

[DSH89] Peter Duchessi, Charles M Schaninger, and Don R Hobbs. Implementing a manufacturing planning and control information system. *California Management Review*, 31(3):75–90, 1989.

[DSP66] N. R. Draper, H. Smith, and E. Pownell. *Applied regression analysis*. Wiley New York, 1966.

[EALS07] Turban Efraim, Jay E Aronson, Ting-Peng Liang, and Ramesh Sharda. Decision support and business intelligence systems. *Pearson Prentice Hall, New Jersey*, 2007.

[EBW16] Behzad Esmaeilian, Sara Behdad, and Ben Wang. The evolution and future of manufacturing: A review. *Journal of Manufacturing Systems*, 39:79–100, 2016.

[EL12] James R Evans and Carl H Lindner. Business analytics: The next frontier for decision sciences. *Decision Line*, 43(2):4–6, 2012.

[ElM05] Hoda A ElMaraghy. Flexible and reconfigurable manufacturing systems paradigms. *International journal of flexible manufacturing systems*, 17(4):261–276, 2005.

[EPP00] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in computational mathematics*, 13(1):1–50, 2000.

[EPPV02] Theodoros Evgeniou, Tomaso Poggio, Massimiliano Pontil, and Alessandro Verri. Regularization and statistical learning theory for data analysis. *Computational Statistics & Data Analysis*, 38(4):421–432, 2002.

[ET93] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[FCH$^+$08] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[FDCBA14] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *JMLR*, 15(1):3133–3181, 2014.

[FOA15] Emanuele Fumeo, Luca Oneto, and Davide Anguita. Condition based maintenance in railway transportation systems based on big data streaming analysis. *Procedia Computer Science*, 53:437–446, 2015.

[FSC06] Andrew Feller, Dan Shunk, and Tom Callarman. Value chains versus supply chains. *BP trends*, pages 1–7, 2006.

[FW95] S. Floyd and M. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine learning*, 21(3):269–304, 1995.

[GCSR14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Taylor & Francis, 2014.

[GE79] Seymour Geisser and William F Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

[GHH$^+$13] Christoph Gröger, Mark Hillmann, Friedemann Hahn, Bernhard Mitschang, and Engelbert Westkämper. The operational process dashboard for manufacturing. *Procedia CIRP*, 7:205–210, 2013.

[GJR11] MJSRM Ghazanfari, M Jafari, and S Rouhani. A tool to evaluate the business intelligence of enterprise systems. *Scientia Iranica*, 18(6):1579–1590, 2011.

[GLLF15] P. Germain, A. Lacasse, M. Laviolette, A. ahd Marchand, and Roy J. F. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *JMLR*, 16(4):787–860, 2015.

[Goo15] Google. Google compute engine, google cloud platform. `https://cloud.google.com/`, 2015.

[GPM04] Angappa Gunasekaran, Christopher Patel, and Ronald E McGaughey. A framework for supply chain performance measurement. *International journal of production economics*, 87(3):333–347, 2004.

[Gro07] Mikell P Groover. *Fundamentals of modern manufacturing: materials processes, and systems*. John Wiley & Sons, 2007.

[GS03] Carl Gold and Peter Sollich. Model selection for support vector machine classification. *Neurocomputing*, 55(1):221–249, 2003.

[GSDC10] Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin Cawley. Model selection: Beyond the bayesian/frequentist divide. *The Journal of Machine Learning Research*, 11:61–87, 2010.

[Gui00] V Daniel R Guide. Production planning and control for remanufacturing: industry practice and research needs. *Journal of operations Management*, 18(4):467–483, 2000.

[Gun88] Thomas G Gunn. *Manufacturing for competitive advantage: becoming a world class manufacturer*. Ballinger Pub Co, 1988.

[GVD02] Marc Goetschalckx, Carlos J Vidal, and Koray Dogan. Modeling and design of global logistics systems: A review of integrated strategic and tactical models and design algorithms. *European journal of operational research*, 143(1):1–18, 2002.

[H$^+$05] Terry Hill et al. Operations management. 2005.

[Har01] Terry P Harrison. Global supply chain design. *Information Systems Frontiers*, 3(4):413–416, 2001.

[HE$^+$01] Greg Hamerly, Charles Elkan, et al. Bayesian approaches to failure prediction for disk drives. In *ICML*, volume 1, pages 202–209, 2001.

[HLMMS13] D. Hernández-Lobato, G. Martínez-Muñoz, and A. Suárez. How large should ensembles of classifiers be? *Pattern Recognition*, 46(5):1323–1336, 2013.

[HN99] Robert B Handfield and Ernst L Nichols. *Introduction to supply chain management*. PhD thesis, Univerza na Primorskem, Znanstveno-raziskovalno središče, 1999.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[HSK$^+$06] JA Harding, M Shahbaz, A Kusiak, et al. Data mining in manufacturing: a review. *Journal of Manufacturing Science and Engineering*, 128(4):969–976, 2006.

[Hui15] Dennis Oliver Huisman. To what extent do predictive, descriptive and prescriptive supply chain analytics affect organizational performance? B.S. thesis, University of Twente, 2015.

[HVNK13] S Hesse, V Vasyutynskyy, D Nadoveza, and D Kiritsis. 15.1 visual analysis of performance indicators and processes in modern manufacturing. 2013.

[HW84] Robert H Hayes and Steven C Wheelwright. Restoring our competitive edge: competing through manufacturing. 1984.

[HY09] Yao-Wen Hsueh and Chan-Yun Yang. Tool breakage diagnosis in face milling by support vector machine. *Journal of materials processing technology*, 209(1):145–152, 2009.

[IK05] Atsushi Inoue and Lutz Kilian. In-sample or out-of-sample tests of predictability: Which one should we use? *Econometric Reviews*, 23(4):371–402, 2005.

[IOK94] K Iwata, M Onosato, and M Koike. Random manufacturing system: a new concept of manufacturing systems for production to order. *CIRP Annals-Manufacturing Technology*, 43(1):379–383, 1994.

[JH93] B. Joseph and F. W. Hanratty. Predictive control of quality in a batch manufacturing process using artificial neural network models. *Industrial & engineering chemistry research*, 32(9):1951–1961, 1993.

[JH04] Xun Jin and Karen A High. A new conceptual hierarchy for identifying environmental sustainability metrics. *Environmental Progress & Sustainable Energy*, 23(4):291–301, 2004.

[K$^+$95] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, 1995.

[KAAD01] Vladimir Koltchinskii, Chaouki T Abdallah, Marco Ariola, and Peter Dorato. Statistical learning control of uncertain systems: theory and algorithms. *Applied mathematics and computation*, 120(1):31–43, 2001.

[KBT11] G. Köksal, I. Batmaz, and M. C. Testik. A review of data mining applications for quality improvement in manufacturing industry. *Expert systems with Applications*, 38(10):13448–13467, 2011.

[KC$^+$09] Reshma Khemchandani, Suresh Chandra, et al. Knowledge based proximal support vector machines. *European Journal of Operational Research*, 195(3):914–923, 2009.

[KLC$^+$16] Hyoung Seok Kang, Ju Yeon Lee, SangSu Choi, Hyun Kim, Jun Hee Park, Ji Yeon Son, Bo Hyun Kim, and Sang Do Noh. Smart manufacturing: Past research, present findings, and future directions. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 3(1):111–128, 2016.

[KMK15] Deogratias Kibira, K Morris, and Senthilkumaran Kumaraguru. Methods and tools for performance assurance of smart manufacturing systems. *National Institute of Standards and Technology, NISTIR*, 8099, 2015.

[KMS$^+$08] Daniel A Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. Visual analytics: Scope and challenges. In *Visual data mining*, pages 76–90. Springer, 2008.

[KMSZ06] Daniel A Keim, Florian Mansmann, Jörn Schneidewind, and Hartmut Ziegler. Challenges in visual data analysis. In *Information Visualization, 2006. IV 2006. Tenth International Conference on*, pages 9–16. IEEE, 2006.

[Kol01] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[KRS02] Ron Kohavi, Neal J Rothleder, and Evangelos Simoudis. Emerging trends in business analytics. *Communications of the ACM*, 45(8):45–48, 2002.

[KTSJ11] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. Bootstrapping big data. In *Advances in Neural Information Processing Systems, Workshop: Big Learning: Algorithms, Systems, and Tools for Learning at Scale*, 2011.

[KTSJ12] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. The big data bootstrap. In *International conference on Machine learning*, 2012.

[KTSJ14] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

[KZP07] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques, 2007.

[Lan06] J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6(1):273, 2006.

[LBJ16] Jay Lee, Behrad Bagheri, and Chao Jin. Introduction to cyber manufacturing. *Manufacturing Letters*, 8:11–15, 2016.

[LC00] Douglas M Lambert and Martha C Cooper. Issues in supply chain management. *Industrial marketing management*, 29(1):65–83, 2000.

[LH09a] Jang Hee Lee and Sung Ho Ha. Recognizing yield patterns through hybrid applications of machine learning techniques. *Information Sciences*, 179(6):844–850, 2009.

[LH09b] Te-Sheng Li and Cheng-Lung Huang. Defect spatial pattern recognition using a hybrid som–svm approach in semiconductor manufacturing. *Expert systems with Applications*, 36(1):374–385, 2009.

[Lic13] M. Lichman. UCI machine learning repository, 2013.

[Lie13] Jay Liebowitz. *Big data and business analytics*. CRC press, 2013.

[LJ13] Yinhua Liu and Sun Jin. Application of bayesian networks for diagnostics in the assembly process by considering small measurement data sets. *The International Journal of Advanced Manufacturing Technology*, pages 1–9, 2013.

[LKY14] Jay Lee, Hung-An Kao, and Shanhu Yang. Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia Cirp*, 16:3–8, 2014.

[LLBK13] J. Lee, E. Lapira, B. Bagheri, and H. A. Kao. Recent advances and trends in predictive manufacturing systems in big data environment. *Manufacturing Letters*, 1(1):38–41, 2013.

[LLS$^+$11] Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S Hopkins, and Nina Kruschwitz. Big data, analytics and the path from insights to value. *MIT sloan management review*, 52(2):21, 2011.

[LLST13] G. Lever, F. Laviolette, and F. Shawe-Taylor. Tighter pac-bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

[LNR14] D. Lechevalier, A. Narayanan, and S. Rachuri. Towards a domain-specific framework for predictive analytics in manufacturing. In *IEEE International Conference on Big Data*, 2014.

[LRS12] Richard J Lehmann, Robert Reiche, and Gerhard Schiefer. Future internet and the agri-food sector: State-of-the-art in literature and research. *Computers and Electronics in Agriculture*, 89:158–174, 2012.

[LSW90] G Keong Leong, David L Snyder, and Peter T Ward. Research in the process and content of manufacturing strategy. *Omega*, 18(2):109–122, 1990.

[LV99] Rhonda R Lummus and Robert J Vokurka. Defining supply chain management: a historical perspective and practical guidelines. *Industrial Management & Data Systems*, 99(1):11–17, 1999.

[LW$^+$02] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[LYQ$^+$12] Xianbing Liu, Jie Yang, Sixiao Qu, Leina Wang, Tomohiro Shishime, and Cunkuan Bao. Sustainable production: practices and determinant factors of green supply chain management of chinese companies. *Business Strategy and the Environment*, 21(1):1–16, 2012.

[Mac92] D. J. C. MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

[Mad12] S. Madden. From databases to big data. *IEEE Internet Computing*, (3):4–6, 2012.

[MBST15] Gökan May, Ilaria Barletta, Bojan Stahl, and Marco Taisch. Energy management in production: A novel method to develop key performance indicators for improving energy efficiency. *Applied Energy*, 149:46–61, 2015.

[McA98] D. A. McAllester. Some pac-bayesian theorems. In *Computational Learning Theory*, 1998.

[McA03] David A McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

[McC91] Ian McChesney. *The Brundtland report and sustainable development in New Zealand.* Lincoln University and University of Canterbury. Centre for Resource Management., 1991.

[MI00] M. Magdon-Ismail. No free lunch for noise prediction. *Neural computation*, 12(3):547–564, 2000.

[MML$^+$13] Mahesh Mani, Jatinder Madan, Jae Hyun Lee, Kevin W Lyons, and Satyandra K Gupta. Review on sustainability characterization for manufacturing processes. *National Institute of Standards and Technology, Gaithersburg, MD, Report No. NISTIR*, 7913, 2013.

[MMRC01] John A Muckstadt, David H Murray, James A Rappold, and Dwight E Collins. Guidelines for collaborative supply chain system design and operation. *Information systems frontiers*, 3(4):427–453, 2001.

[MPG14] Ruchi Mishra, Ashok K Pundir, and L Ganapathy. Manufacturing flexibility research: A review of literature and agenda for future research. *Global Journal of Flexible Systems Management*, 15(2):101–112, 2014.

[MPMDM10] Peter N Muchiri, Liliane Pintelon, Harry Martin, and Anne-Marie De Meyer. Empirical analysis of maintenance performance measurement in belgian industries. *International Journal of Production Research*, 48(20):5905–5924, 2010.

[MSC13] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà*. Garzanti, 2013.

[MSDS14] Hajar Mousannif, Hasna Sabah, Yasmina Douiji, and Younes Oulad Sayad. From big data to big projects: A step-by-step roadmap. In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, pages 373–378. IEEE, 2014.

[MSMC78] Raymond E Miles, Charles C Snow, Alan D Meyer, and Henry J Coleman. Organizational strategy, structure, and process. *Academy of management review*, 3(3):546–562, 1978.

[NDZ02] Anna Nagurney, June Dong, and Ding Zhang. A supply chain network equilibrium model. *Transportation Research Part E: Logistics and Transportation Review*, 38(5):281–303, 2002.

[Neg04] Solomon Negash. Business intelligence. *The communications of the Association for Information Systems*, 13(1):54, 2004.

[NJL$^+$11] TA Nguyen, P-Y Joubert, Stéphane Lefebvre, G Chaplier, and L Rousseau. Study for the non-contact characterization of metallization ageing of power electronic semiconductor devices using the eddy current technique. *Microelectronics Reliability*, 51(6):1127–1135, 2011.

[NM14] S. Nannapaneni and S. Mahadevan. Uncertainty quantification in performance evaluation of manufacturing processes. In *IEEE International Conference on Big Data*, 2014.

[NNB15] Janmenjoy Nayak, Bighnaraj Naik, and H Behera. A comprehensive survey on support vector machine in data mining tasks: applications & challenges. *International Journal of Database Theory and Application*, 8(1):169–186, 2015.

[NP82] S. Nitzan and J. Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, 23(2):289–97, 1982.

[OGRA14] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Fully empirical and data-dependent stability-based bounds. *IEEE Transactions on Cybernetics*, 2014.

[OGRA15a] Luca Oneto, Alessandro Ghio, Sandro Ridella, and Davide Anguita. Global rademacher complexity bounds: From slow to fast convergence rates. *Neural Processing Letters*, pages (in–press), 2015.

[OGRA15b] Luca Oneto, Alessandro Ghio, Sandro Ridella, and Davide Anguita. Local rademacher complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, pages (in–press), 2015.

[OOA15] Luca Oneto, Ilenia Orlandi, and Davide Anguita. Performance assessment and uncertainty quantification of predictive models for smart manufacturing systems. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 1436–1445. IEEE, 2015.

[OOA16] Ilenia Orlandi, Luca Oneto, and Davide Anguita. Random forrest model selection. ESANN, 2016.

[OPGD15] L. Oneto, B. Pilarz, A. Ghio, and Anguita D. Model selection for big data: Algorithmic stability and bag of little bootstraps on gpus. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2015.

[OPP^{+}02] Richard Oechsner, Markus Pfeffer, Lothar Pfitzner, Harald Binder, Eckhard Müller, and Thomas Vonderstrass. From overall equipment efficiency (oee) to overall fab effectiveness (ofe). *Materials Science in Semiconductor Processing*, 5(4):333–339, 2002.

[ORW01] Jan Olhager, Martin Rudberg, and Joakim Wikner. Long-term capacity management: Linking the perspectives from manufacturing strategy and sales and operations planning. *International Journal of Production Economics*, 69(2):215–225, 2001.

[PA05] DT Pham and AA Afify. Machine-learning techniques and their applications in manufacturing. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 219(5):395–412, 2005.

[Par05] YB Park*. An integrated approach for production and distribution planning in supply chain management. *International Journal of Production Research*, 43(6):1205–1224, 2005.

[PRMN04] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.

[R^{+}11] Philip Russom et al. Big data analytics. *TDWI best practices report, fourth quarter*, 19:40, 2011.

[RB12] Geary A Rummler and Alan P Brache. *Improving performance: How to manage the white space on the organization chart*. John Wiley & Sons, 2012.

[RDVC$^+$04] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

[REAG12] Suresh Renukappa, Charles Egbu, Akintola Akintoye, and Jack Goulding. A critical reflection on sustainability within the uk industrial sectors. *Construction Innovation*, 12(3):317–334, 2012.

[RGGR$^+$12] Victor Francisco Rodriguez-Galiano, Bardan Ghimire, John Rogan, Mario Chica-Olmo, and Juan Pedro Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93–104, 2012.

[RMP05] A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

[ROOA15] Jorge L Reyes-Ortiz, Luca Oneto, and Davide Anguita. Big data analytics in the cloud: Spark on hadoop vs mpi/openmp on beowulf. *Procedia Computer Science*, 53:121–130, 2015.

[RS12] Laurence Clément Roca and Cory Searcy. An analysis of indicators disclosed in corporate sustainability reports. *Journal of Cleaner Production*, 20(1):103–118, 2012.

[Rus07] Cathy Rusinko. Green manufacturing: an evaluation of environmentally sustainable manufacturing practices and their impact on competitive outcomes. *IEEE Transactions on Engineering Management*, 54(3):445–454, 2007.

[RWVD08] Corinne Reich-Weiser, Athulan Vijayaraghavan, and David A Dornfeld. Metrics for sustainable manufacturing. In *Proceedings of the 2008 International Manufacturing Science and Engineering Conference, Evanston, IL*, 2008.

[RZJ04] Andrej Rakar, Sebastjan Zorzut, and Vladimir Jovan. Assesment of production performance by means of kpi. *Control 2004*, pages 6–9, 2004.

[SB12a] Leigh Smith and Peter Ball. Steps towards sustainable manufacturing through modelling material, energy and waste flows. *International Journal of Production Economics*, 140(1):227–238, 2012.

[SB12b] Srinath Srinivasa and V Bhatnagar. Big data analytics. In *Proceedings of the First International Conference on Big Data Analytics BDA*, pages 24–26. Springer, 2012.

[SBJ02] Roger G Schroeder, Kimberly A Bates, and Mikko A Junttila. A resource-based view of manufacturing strategy and the relationship to manufacturing performance. *Strategic management journal*, 23(2):105–117, 2002.

[SFBL98] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[SH07] William J Stevenson and Mehran Hojati. *Operations management*, volume 8. McGraw-Hill/Irwin Boston, 2007.

[SHYH08] SKAL Subramaniam, Siti Huzaimah Binti Husin, Yusmarnita Binti Yusop, and Abdul Hamid Bin Hamidon. Machine efficiency and man power utilization on production lines. In *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*, number 7. World Scientific and Engineering Academy and Society, 2008.

[SKK10] Karim Salahshoor, Mojtaba Kordestani, and Majid S Khoshro. Fault detection and diagnosis of an industrial steam turbine using fusion of svm (support vector machine) and anfis (adaptive neuro-fuzzy inference system) classifiers. *Energy*, 35(12):5472–5482, 2010.

[SMK14] Rajeev Sharma, Sunil Mithas, and Atreyi Kankanhalli. Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems*, 23(4):433–441, 2014.

[SN87] Paul M Swamidass and William T Newell. Manufacturing strategy, environmental uncertainty and performance: a path analytic model. *Management science*, 33(4):509–524, 1987.

[SNW07] Morgan Swink, Ram Narasimhan, and Cynthia Wang. Managing beyond the factory walls: effects of four types of strategic integration on manufacturing plant performance. *Journal of Operations Management*, 25(1):148–164, 2007.

[Spi06] G Spina. La gestione dell'impresa. *Organizzazione, processi decisionali, marketing, acquisti e supply chain, Etas, Milano*, 2006.

[SS05] Andreas Seufert and Josef Schiefer. Enhanced business intelligence-supporting business processes with real-time business analytics. In *Database and Expert Systems Applications, 2005. Proceedings. Sixteenth International Workshop on*, pages 919–925. IEEE, 2005.

[SS07] Mark Stevenson and Martin Spring. Flexibility from a supply chain perspective: definition and review. *International Journal of Operations & Production Management*, 27(7):685–713, 2007.

[SS13] Seref Sagiroglu and Duygu Sinanc. Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, pages 42–47. IEEE, 2013.

[SS15] A. G. Shoro and T. R. Soomro. Big data analysis: Apache spark perspective. *Global Journal of Computer Science and Technology*, 15(1), 2015.

[SSA09] D Solomatine, Linda M See, and RJ Abrahart. Data-driven modelling: concepts, approaches and experiences. In *Practical hydroinformatics*, pages 17–30. Springer, 2009.

[SSBD14] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[SSJ14] Guodong Shao, Seung-Jun Shin, and Sanjay Jain. Data analytics using simulation for smart manufacturing. In *Proceedings of the 2014 Winter Simulation Conference*, pages 2192–2203. IEEE Press, 2014.

[SSM12] Marten Schläfke, Riccardo Silvi, and Klaus Möller. A framework for business analytics in performance management. *International Journal of Productivity and Performance Management*, 62(1):110–122, 2012.

[SSSSC11] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

[STC04] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[Ste85] Kathryn E Stecke. Design, planning, scheduling, and control problems of flexible manufacturing systems. *Annals of Operations research*, 3(1):1–12, 1985.

[STS11] J. Shawe-Taylor and S. Sun. A review of optimization methodologies in support vector machines. *Neurocomputing*, 74(17):3609–3618, 2011.

[SWMK11] Sandeep Salunke, Jay Weerawardena, and Janet R McColl-Kennedy. Towards a model of dynamic capabilities in innovation-based competitive strategy: Insights from project-oriented service firms. *Industrial Marketing Management*, 40(8):1251–1263, 2011.

[TA77] A. N. Tikhonov and V. I. Arsenin. *Solutions of ill-posed problems*. Vh Winston, 1977.

[TC06] James J Thomas and Kristin A Cook. A visual analytics agenda. *IEEE computer graphics and applications*, 26(1):10–13, 2006.

[TDLT16] F Tonelli, M Demartini, A Loleo, and C Testa. A novel methodology for manufacturing firms value modeling and mapping to improve operational performance in the industry 4.0 era. *Procedia CIRP*, 57:122–127, 2016.

[TET13] Flavio Tonelli, Steve Evans, and Paolo Taticchi. Industrial sustainability: challenges, perspectives, actions. *International Journal of Business Innovation and Research*, 7(2):143–163, 2013.

[THE07] Esther Turnhout, Matthijs Hisschemöller, and Herman Eijsackers. Ecological indicators: between the two fires of science and policy. *Ecological indicators*, 7(2):215–228, 2007.

[TK09] Jim Thomas and Joe Kielman. Challenges for visual analytics. *Information Visualization*, 8(4):309–314, 2009.

[TM15] Loukas K Tsironis and Panagiotis Petros Matthopoulos. Towards the identification of important strategic priorities of the supply chain network: An empirical investigation. *Business Process Management Journal*, 21(6):1279–1298, 2015.

[TMDOL10] Peter Trkman, Kevin McCormack, Marcos Paulo Valadares De Oliveira, and Marcelo Bronzo Ladeira. The impact of business analytics on supply chain performance. *Decision Support Systems*, 49(3):318–327, 2010.

[TSM15] Marco Taisch, Bojan Stahl, and Gokan May. Sustainability in manufacturing strategy deployment. *Procedia CIRP*, 26:635–640, 2015.

[Unv13] Hakki Ozgur Unver. An isa-95-based manufacturing intelligence system in support of lean initiatives. *The International Journal of Advanced Manufacturing Technology*, pages 1–14, 2013.

[Val84] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vap98] V. N. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[Vap99] V. N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

[VGB11] Antanas Verikas, Adas Gelzinis, and Marija Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330–349, 2011.

[VLS08] Ulrike Von Luxburg and Bernhard Schölkopf. Statistical learning theory: models, concepts, and results. *arXiv preprint arXiv:0810.4752*, 2008.

[WCL05] K-J Wang*, JC Chen, and Y-S Lin. A hybrid knowledge discovery model using decision tree and neural network for selecting dispatching rules of a semiconductor final testing factory. *Production planning & control*, 16(7):665–680, 2005.

[WH04] Kenneth H Wathne and Jan B Heide. Relationship governance in a supply chain network. *Journal of marketing*, 68(1):73–89, 2004.

[WHL15] C. C. Wang, C. H. Huang, and C. J. Lin. Subsampled hessian newton methods for supervised learning. *Neural computation*, 8(27):1738–1765, 2015.

[WIT14] T. Wuest, C. Irgens, and K. D. Thoben. An approach to monitoring quality in manufacturing using supervised machine learning on product state data. *Journal of Intelligent Manufacturing*, 25(5):1167–1180, 2014.

[WJT$^+$17] Dazhong Wu, Connor Jennings, Janis Terpenny, Robert X Gao, and Soundar Kumara. A comparative study on machine learning algorithms for smart manufacturing: Tool wear prediction using random forests. *Journal of Manufacturing Science and Engineering*, 139(7):071018, 2017.

[Wol96] D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[WRS13] Matthew J Woodruff, Patrick M Reed, and Timothy W Simpson. Many objective visual analytics: rethinking the design of complex engineered systems. *Structural and Multidisciplinary Optimization*, 48(1):201–219, 2013.

[WW07] Hugh J Watson and Barbara H Wixom. The current state of business intelligence. *Computer*, 40(9), 2007.

[WWIT16] Thorsten Wuest, Daniel Weimer, Christopher Irgens, and Klaus-Dieter Thoben. Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research*, 4(1):23–45, 2016.

[WY07] Achmad Widodo and Bo-Suk Yang. Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical systems and signal processing*, 21(6):2560–2574, 2007.

[WZWD14] X. Wu, X. Zhu, G. Q. Wu, and W. Ding. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):97–107, 2014.

[XZA$^+$08] Xiuqiao Xiang, Jianzhong Zhou, Xueli An, Bing Peng, and Junjie Yang. Fault diagnosis based on walsh transform and support vector machine. *Mechanical Systems and Signal Processing*, 22(7):1685–1693, 2008.

[YDH08] Bo-Suk Yang, Xiao Di, and Tian Han. Random forests classifier for machine fault diagnosis. *Journal of mechanical science and technology*, 22(9):1716–1725, 2008.

[YL09] Jing-Lin Yang and Han-Xiong Li. A probabilistic support vector machine for uncertain data. In *Computational Intelligence for Measurement Systems and Applications, 2009. CIMSA'09. IEEE International Conference on*, 2009.

[YL12] Lei Yang and Jay Lee. Bayesian belief network-based approach for diagnostics and prognostics of semiconductor manufacturing systems. *Robotics and Computer-Integrated Manufacturing*, 28(1):66–74, 2012.

[YL16] Xifan Yao and Yingzi Lin. Emerging manufacturing paradigm shifts for the incoming industrial revolution. *The International Journal of Advanced Manufacturing Technology*, 85(5-8):1665–1676, 2016.

[YO12] Ernie Mazuin Mohd Yusof and Mohd Shahizan Othman. A review on the dashboard characterisics for manufacturing organizations. *Journal of Information Systems Research and Innovation*, pages 2289–1358, 2012.

[YSDL96] Mark A Youndt, Scott A Snell, James W Dean, and David P Lepak. Human resource management, manufacturing strategy, and firm performance. *Academy of management Journal*, 39(4):836–866, 1996.

[Zho12] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.