POLITECNICO DI TORINO

Master's Degree in Management Engineering

Master Thesis

Machine Learning for Project Management

Relatori:
Prof. Alberto De Marco
Filippo Maria Ottaviani

Candidato:
Salvatore Pace

Academic Year 2020/2021

Table of Contents

Abstract

Given the great propensity of companies to work on projects, today project management has become a real vital pillar for all organizations. The two fundamental activities that allow us to understand and analyse the evolution of a project are Project Monitoring and Project Control. Both are essential to analyse the project in their key features, so, in terms of costs, time and quality. A fundamental aspect to consider within a project is Risk Management, which often has a considerable impact on the estimates of a project both in terms of costs, and time, which are therefore updated several times before showing up with the desired result the final one. The two typical project management disciplines mentioned so far must act in parallel in project management, so that everything remains under control. From the analysis carried out in the literature, it can be seen that there are several methods for estimating budget and time in advance, but there is no one that established a method doing it with a high degree of accuracy because ~~the~~ risks are never taken into account. It was therefore this theoretical problem, stimulated by the curiosity to evaluate and optimize the estimation models integrated with the risk analysis to give way to the drafting of this dissertation which aims to take into account the risks within a project and thus give a more accurate beforehand evaluation for the Estimate at Completion (EaC), taking into account all the surrounding factors. Subsequently, it will also be possible to reschedule the timelines to have a more approximate estimated time during the initial phase.

1 Introduction

Earned Value is a well-established method within the project management culture. It allows, through simple formulas, to control and monitor the progress of a project, making fairly accurate estimates of time and cost to completion at the beginning of the project or during a checkpoint called Time Period. Project monitoring and project controlling, therefore, become two core activities for understanding how a project evolves in terms of time and cost, so consequently it will also affect the quality related to a project. They allow to monitor the actual performance of a project by comparing it with the planned performance and evaluating the variance in order to improve the accuracy related to the forecasts of the activities at the end of the project.

In fact, the Earned Value Method (EVM) allows Project Managers to identify, in the shortest possible time, any deviations in terms of time and cost from what was initially budgeted for a given moment. It is common to divide the project in Time Periods (TPs), that are usually ~~are~~ monthly, for the continuous control of the project. So, Project Manager can check the pattern of the project during each TP and reschedule the calculations made in the previous Time Period in order to be more accurate, and from there give a new value of time and cost at the end of the project. These deviations may occur within a project for different reasons, the main one, which is the source of possible delays or increases, both in terms of cost and time, is the one related to risks, so uncertain events whose impact may lead to consequences positive or negative on the whole project. Project Monitoring and Project Control are therefore two parallel disciplines that are fundamental to project management. In literature, they have always been used together but never integrated into a single model. Only in the last few years there has ~~there~~ been

considerable interest in this direction. The first approach was carried out through Monte Carlo simulation, which, after the identification of critical activities, allowed to estimate the relative probability of delay. (Vanhoucke, 2010-2011). Others have developed methods to estimate the evolution of the risk value over time in order to calculate the possible maximum levels of over-run (Lopez-Paredes, Pajares, 2011) and still others have developed growth models capable of estimating the costs at the end of a project through linear regressions (De Marco, Narbaev, 2014). It was precisely De Marco and Narbaev who, in 2016, stimulated by the understanding of the valuable contribution that this union could generate, developed the first estimation-to-end models integrated with risk analysis. Now, with the advent of machine learning, there is this new tendency to apply the algorithms aimed at improving the calculation of EaC as De Marco and Narbaev in 2014. This new approach was computed by (Rezouki S.E., 2020) that will calculate new variables through Non-Linear Regression (NLR) and then apply a linear regression and a nonlinear one to find the optimal model. In the same mindset there is the article of (Balali A., 2020) which applied the Artificial Neural Network (ANN) to a specific project in the construction field, they proceeded with the computation of the traditional EVM, then they applied a multiple regression and at the end the ANN. As we will discuss in the following chapter, they obtained a better result through ANN, than both the traditional EVM and the output obtained by applying a multiple regression. Lastly, I can define the basis of my studies the paper “A new project scheduling control method based on activity quantities” (Chang H., 2019). Its objective is to recalculate the Earned Duration Management (EDM) through the activity quantities of the critical path, not the activity value, so that the risk can be included with greater accuracy.

Risk Management becomes therefore fundamental for the success of a project and it has received during the last decades a great consideration, which has led to the birth of numerous methodologies and techniques to manage them in the best way. The goal of this discipline is: to minimize the impact and the probability of risks that negatively affect the project, while maximizing the impact and the probability of risks that positively affect the project. This is now a prerogative, with resources and action plans dedicated to such activities. The Earned Value methodology, with its output known as EaC, relies solely on actual data to make its final estimates, so the history of the project in terms of time and cost until the TP where this analysis is performed. In these estimates, data from other different projects are not taken into account, so the risks that have occurred up to that point, and those that may yet materialize, are not considered. The challenge of this elaborate is therefore to study the current literature, with all the actual way of calculation of the Estimate at Completion and, starting from there, through the introduction of new variables that will take into account the risk part, propose a new model that allows a more accurate computation for the EaC. For the development of this model the starting point will be the study of historical data of projects already completed, the calculation of new variables that take into account the risk, and then apply machine learning algorithms to predict what are the costs at completion of a project, considering in it also the risk factor that until now has often been avoided.

2 Literature

2.1 Project Monitoring: Earned Value Management

Predicting the final results of a project has always been a critical element in business management. Understanding how a project is evolving in terms of quality, time and cost is achieved through two simultaneous and cyclical activities: Project Monitoring and Project Control. A typical Project Control process consists of monitoring the actual performance by comparing it with the planned performance, to better understand if there are deviations from the estimated values so that the forecasts can be revised more accurately, or to understand if corrective actions can be taken to get back within the budgeted time and cost. One of the objectives of this branch of Project Management is to identify potential risks, in order to take the necessary actions to avoid them or to limit their effects in case of negative impact in order to minimize or neutralize them for the success of the project. This activity must be carried out with a methodical data collection, related to the progress of the activities, the costs actually incurred, and the criticalities encountered, checked during a TP or through specific meetings or conversation with the managers that handle each work package. Obviously, the frequency of data collection is closely linked to the level of "granularity" and depth of control, considering also the cost related to each TP, so it’s impossible have a continuous flow, but usually project manager schedule it once per month. This will depend on the project’s importance, size and cost ~~of the project~~ and on the possible risks involved or detected. Often, an overly frequent data collection can be unproductive and lead to unreliable data; on the other hand, infrequent collection can lead to a lack of project control, so the project will be willing to go “out of control”. The project manager must find the right combination of detail,

frequency, and effort required of those who must provide this information, so as not to disrupt operations and to identify deviations in time to take appropriate corrective action. To understand if it is the right project management team, you can use this information to determine whether the project performance is under control and within acceptable limits. If the project is "out of control", the project manager must research the causes and problems, examining the corrective actions to be taken in the time remaining until the end of the project process. This is done with the aim of bringing the project back “under control”, or, in order to reschedule the project to take advantage of any opportunities that have arisen, as well as to reduce and minimize any cost and time expenses associated with the risks that are on the near horizon. An important issue in assessing the severity of deviations between project goals and expected performance is the inherent variability and uncertainty that is present in any type of human activity. The actual performance of any activity within the project, no matter how well the schedule proceeds according to the budgeted plans, will always be subject to variability, and therefore to minimal deviation. This is because a project's schedule in terms of cost and time is developed on the basis of assumptions regarding the availability and cost of all the resources used within the project, such as raw materials, equipment, specialized resources and available personnel. Consequently, even the implementation of a well-developed plan is subject to natural variations dependent on the specific outputs of the factors considered. In this context, the EV method will be analysed, and will be choose the one that is a very efficient and concise performance measure that encompasses important techniques for estimating costs and time at completion.

2.1.1 How to evaluate the performance

Earned Value Management (EVM) allows the project manager and the project team to gain insight into project performance by being able to modify the project strategy and adjust the schedule to bring it in line with the actual performance trend (Anbari, 2003). The four basic elements on which EVM is based to assess project performance in terms of costs are:

- Planned value (PV) or Budget Cost of Work Scheduled (BCWS): it represents the distribution of the financial budget allocated to the project, subdivided temporally, so for each TP. It allows, at any given moment in time, knowledge of the theoretical cost until the previous TP to understand if the planned activities and costs are under control and therefore in line with what was budgeted. Cumulating the PVs with respect of time leads to the classical S-shaped curve (Figure 1). This feature is characteristic of the increasing operating costs of a plan, deriving from the fact that the greater part of the costs is supported between the 30% and the 70% of completion of the plan,

that becomes therefore the more critical part with a greater variability inside the entire plan.

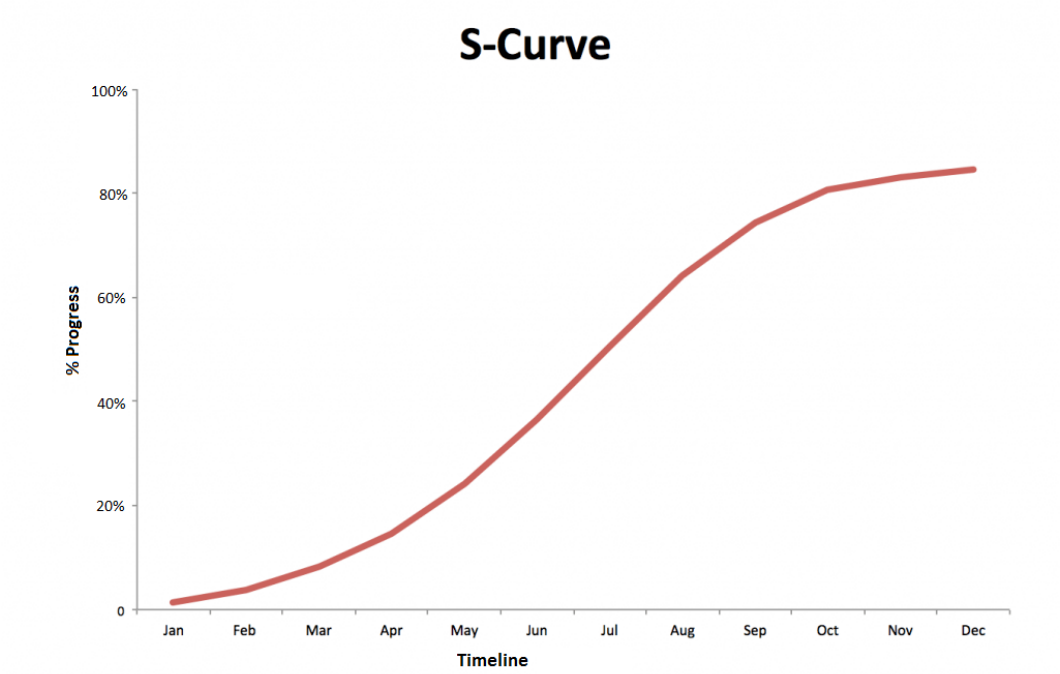

Figure 1: S Curve

- Budget at Completion (BAC): this represents the budget that will be reached, if the project follows what was estimated at the beginning of the project. It coincides with the highest value of the PV, and therefore the last point of the S curve (Figure 2), so the cumulative part of the PV.

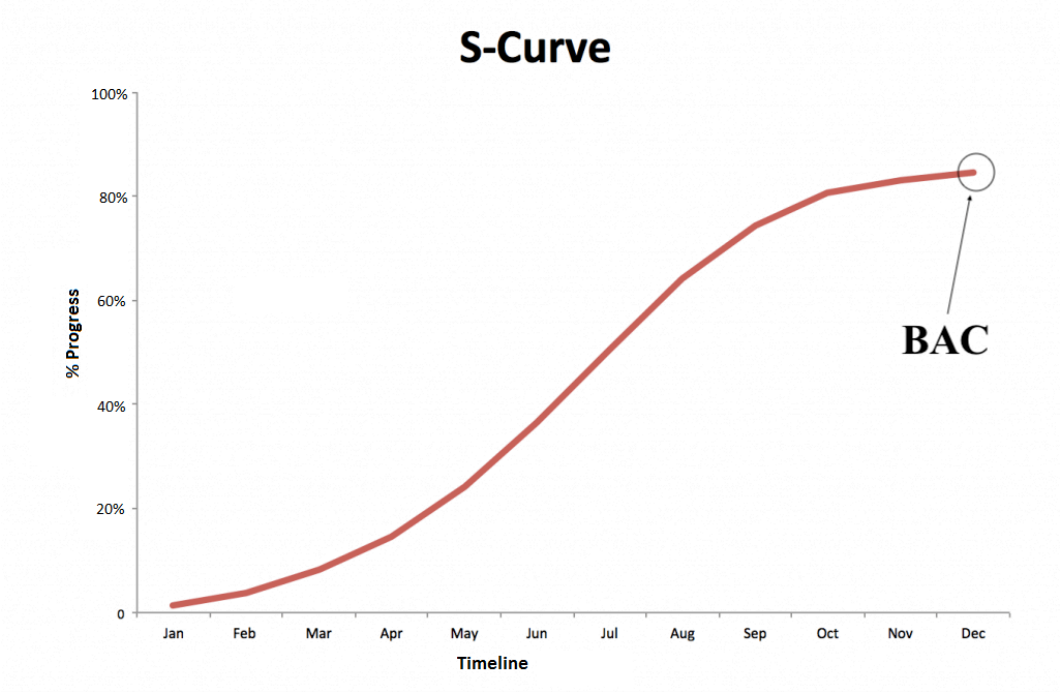

Figure 2: BAC as the last point of the PV

- Actual Cost (AC) or Actual Cost of Work Performed (ACWP): represents the actual cost incurred in a given period. Thus, the cost of each individual period accumulated to the current time of what was actually spent to perform the activities performed up to that time which may represent a different, lower or higher number than budgeted (Figure 3). If the project is under control, it is equal to the PV, else, if the project is out of control it will be higher, otherwise it will be lower.

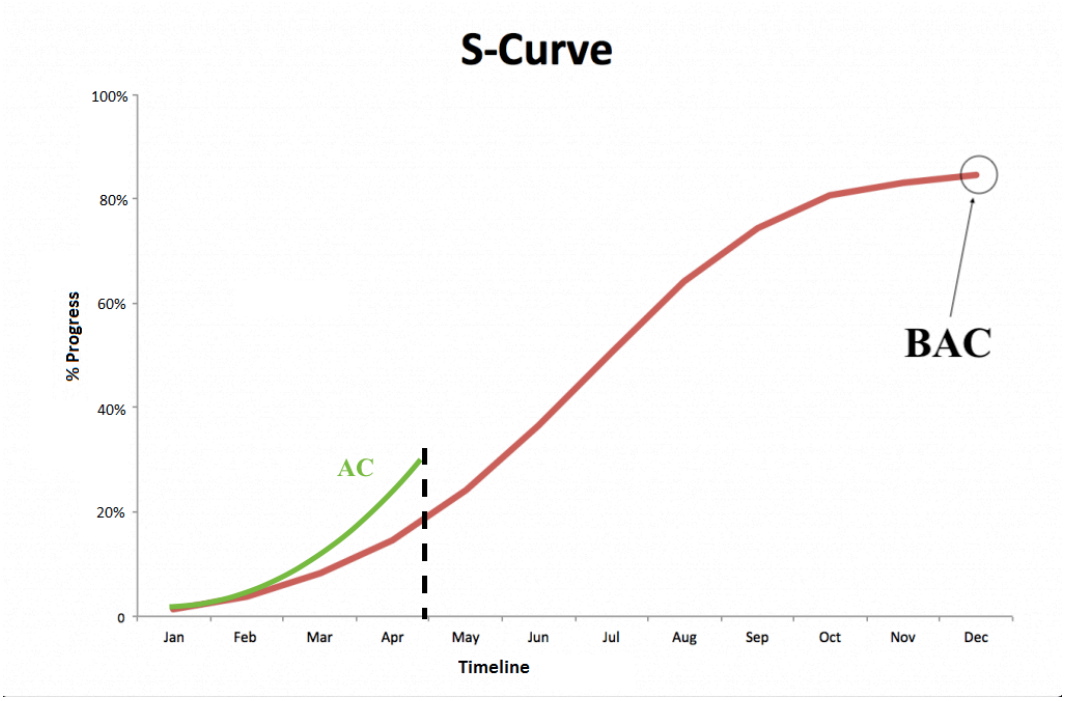

Figure 3: AC of a project

- Earned Value (EV) or Budget Cost of Work Performed (BCWP): it represents the estimated cost to budget to complete the activities concluded up to the current time (Figure 4). It is obtained therefore like the product between the activities actually concluded and the costs previewed to budget for the development of them. As for the previous one, in case the project is under control, it will be equal to the PV.

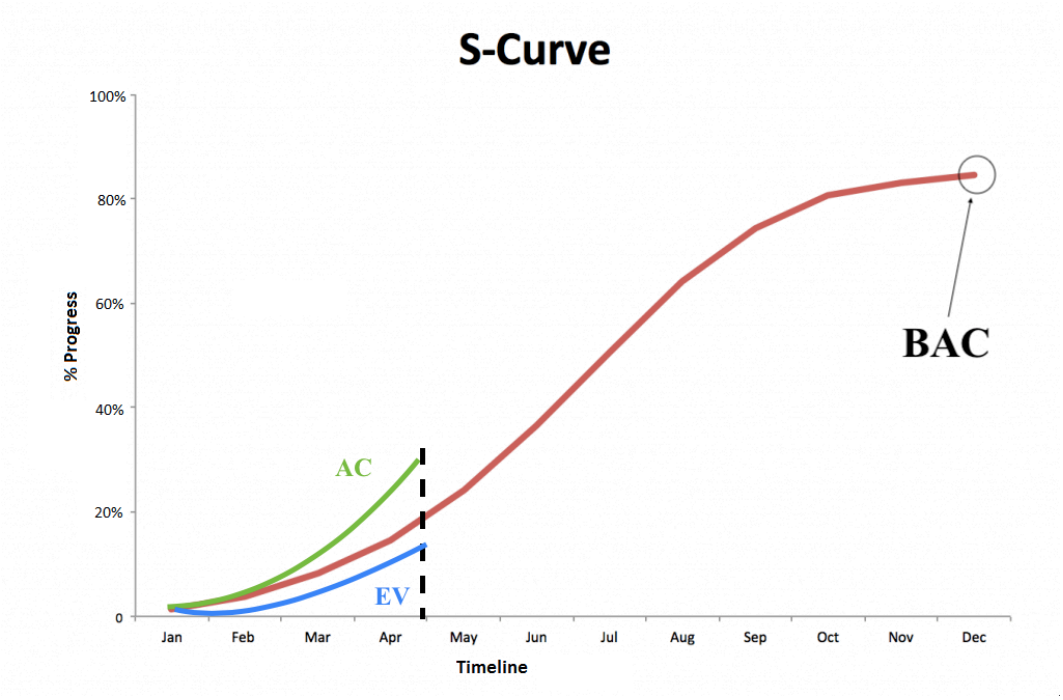

Figure 4: EV of a project

2.1.2 Performance analysis: indexes

There are several indicators aimed at evaluating the performance of a project in terms of cost and time, which will be described in detail below:

- Schedule Variance (SV): This indicator allows us to understand the current position with respect to the initial planning, whether we are therefore ahead or behind. It is calculated through the difference between what has actually been spent with the activities completed to date (EV) and what was budgeted for the achievement of the activities completed to date (PV). The Schedule Variance does not, therefore, represent the efficiency of the achieved result, but it is able to determine how much the planned work is ahead or behind schedule. Naturally a negative SV represents a delay, while a positive SV represents an advance. If the SV is equal to 0, then the project is on schedule. In formula:

$$SV = EV - PV$$

- Schedule Performance Index (SPI): This is an indicator that measures the actual percentage progress of the project compared to the budgeted schedule. In formulas:

$$SPI = \frac{EV}{PV}$$

This performance indicator represents an index of efficiency, with similar considerations to the above, it can be seen that if SPI=1, the project performance is efficient, equal to the planned one and the actual project progress follows the budgeted schedule. SPI < 1 indicates a delay in the project, while SPI > 1 indicates an advance on schedule.

- Cost Variance (CV): is an indicator that measures the expenditure incurred until the date compared to the budget. It is linked to the difference between what was budgeted to be spent on the work actually done to date (EV), and what was actually spent on that work (AC). The Cost Variance therefore focuses on the analysis of the efficiency of the project. If it is negative, it represents inefficiency of the project, otherwise it represents efficiency. In formula:

$$CV = EV - AV$$

- Cost Performance Index (CPI): is an index that measures the performance of actual project costs against budget. In formulas:

$$CPI = \frac{EV}{AC}$$

As mentioned for the SPI, the CPI is an indicator of efficiency in that a CPI=1 indicates that the actual costs incurred to date are in line with those budgeted for the scheduled activities (considering completion rates). A CPI<1 identifies inefficient performance in terms of costs as more is being spent than

budgeted, while a CPI>1 is synonymous of efficiency as costs are being saved. The CPI allows to make forecasts on the overall costs of the entire project; in fact, if the project were to continue to progress with the same trend you can easily estimate an over budget or under budget equal to (1-CPI) %, thus obtaining an estimate of the new budget at the end of the project equal to:

$$BC = \frac{BAC}{CPI}$$

This can be summarized as follows:

Table 1: Variance Index, with comparison between CI and SI

Variance Index		**Time**		
		SV>0, SPI(t)>1	SV=0, SPI(t)=1	SV<0, SPI(t)<1
Cost	CV>0, CPI>0	Schedule Advance, Budget Saving	In line on the schedule, Budget Saving	Late in the schedule, Budget Saving
	CV=0, CPI=0	Schedule Advance, in line with the budget	In line on the schedule, in line with the budget	Late in the schedule, in line with the budget
	CV<0, CPI<0	Schedule Advance, over budget	In line on the schedule, over budget	Late in the schedule, over budget

2.2 Cost and time forecast

The EVM method is not simply aimed at evaluating the performance of a project, but also finds its usefulness in linear time and cost forecasting by analysing current performance. In detail, the main performance indicators are: CEAC (Cost Estimate at Completion) and TEAC (Time Estimate at Completion).

- CEAC (Cost Estimate at Completion): The goal of this metric is to provide an estimation of the project's cost to completion by analysing the project's current and past performance. There are several project end-cost estimates that can be used, as there are different forecasting methodologies for future performance, each based on different assumptions. (Project Management Institute, 2000, Anbari, 2003). The following describes and analyses the three main methodologies for calculating end-of-pipe costs:
    - $CEAC_1$: This first method is often used when there are significant changes from the estimates made at the time of the evaluation or the previous control time, due to changes in conditions that affect the activities, the work package or the project itself. After these variations, it will then be necessary to make a new estimate, more consistent with the new project parameters, taking into account the variation that has occurred, following the following formula:

      $$CEAC_1 = AC + ETC$$

      where the ETC (Estimate to Complete) relates to the remaining portion of the cost to be incurred until completion of the project, which will therefore be different from that calculated in the estimate.
    - $CEAC_2$: The second method is applied when past performance is not a good approximation for future performance, due to problems or

opportunities that have affected the results achieved up to that point. In order not to affect future estimates, it will therefore be necessary to evaluate the new $CEAC_2$ taking these changes into account, as follows:

$$CEAC_2 = AC + BAC - EV.$$

In this case, the ETC for the remaining activities is the difference between the original budget and the work already done (EV).

- $CEAC_3$: The third and final metric is used when previous performance is particularly significant for future performance, in that the efficiencies or weaknesses recorded up to that point will continue into the future. The estimate of the CEAC can then be obtained by inserting the cost performance indicator CPI into the formula, which will allow for a more detailed calculation:

$$CEAC_3 = AC + \left(\frac{BAC - EV}{CPI}\right) = AC + \left(\frac{BAC}{CPI}\right)$$

- TEAC (Time Estimate at Completion). Following the same logic as CEAC, by making the same assumptions we can calculate this metric, which is used in estimating time to completion based on past performance (Anbari 2002). In more detail, parallel to CEAC we will look at the three main methods to calculate TEAC:
  - $TEAC_1$: As in the previous case, this metric is used when the current analysis shows that the assumptions underlying the original time estimate were flawed or not applicable due to changes in project conditions, thus some risk that occurred or some positive or negative action to complete the project. Therefore, it will be necessary to proceed with a new estimate of project duration and time to completion that will vary from that originally calculated:

$$TEAC_1 = AT + TETC$$

where AT indicates the current time and TETC indicates the evaluated time to completion at current estimate or in the previous Time Period.

- $TEAC_2$: Again, this metric is used when there is not much reliability regarding past performance, so they are not a good forecast for the future, due to issues or opportunities that have affected past performance. To prevent this from affecting current performance, the TEAC can be recalculated, taking the variance into account as well:

$$TEAC_2 = SAC - TV$$

Where SAC denotes the baseline-based estimate of duration and with TV denotes the skew.

- $TEAC_3$: Finally, this metric is used in cases where the past estimate is very significant for the calculation of the future one, as the past trend also allows us to understand the trend for the rest of the project, taking into account both positive and negative effects that will persist until the end of the project. Taking this factor into account we can invoke the SPI described above for a more accurate calculation, obtaining:

$$TEAC_3 = \frac{SAC}{SPI}$$

After having thoroughly analysed the methodologies that allow the calculation of estimated costs and time to completion, and after having chosen the most appropriate one for the project, it is necessary to discuss also two other factors that are fundamental in the analysis of the scenarios described above. They are:

- TVAC (Time Variance at Completion): This metric is used to understand how far ahead or behind schedule the project will be and is calculated as follows:

$$TVAC = SAC - TEAC$$

A TVAC = 0 indicates that the project will finish on schedule, a positive TVAC indicates an underestimate and therefore the project will finish early, and a negative TVAC indicates an overestimate and therefore the project will finish late.

- VAC (Variance at Completion): This metric indicates how much the final estimate deviates from the initial cost evaluation. This indicator provides an understanding of whether the project is on track to meet the final budget or deviate from it in both over-budget and under-budget cases. It is calculated as follows:

$$VAC = BAC - EAC$$

Similarly, to the TVAC, a VAC=0 indicates that the project will end on budget, a positive VAC indicates that less was spent than budgeted, the project will therefore end with a savings, while a negative VAC indicates that more was spent than budgeted, indicating that the project will end with an excess over budget.

As it has been possible to deduce from the definitions cited until this moment, the EVM turns out to be a very useful and efficient way in how much it succeeds to give instantaneous indications to the project manager regarding surplus or minus valences in terms of costs and times, allowing therefore to act timely through corrective actions times to increase the success of the project that comes estimated through the variation regarding how much budgeted in advance. Obviously, all these indicators, if used in the correct timeframe, can positively influence the success of the activities as they quickly allow the project manager to carry out corrective actions aimed at the right progress of the project. However, despite its high diffusion in all

industrial sectors, the EVM method has some limitations related to the indicators of which it is composed. The first factor to consider is the accuracy of the method, it depends on past performance and indicators that best allow to estimate future ones, in fact it assumes that the future budget is updated as performance changes (CPI and SPI), calculated on the basis of past performance that will not always reflect the actual one. (Christensen, Heise, 1993) (Fleming, Koppelman, 2006) (Kim, Reinschimdt, 2011). The second factor includes the outcome of the method, which is highly dependent on when the estimate is made. If you are in the early stages of the project, you have little information on which to base your estimates, and they will not be entirely reliable. Similarly, if you are in the late stages of the project, the budgeted cost to complete the activities (EV) will tend to the cost of planned work (PV), and as a result, SV will converge to zero and SPI will tend to one even if the project is significantly behind schedule. So, this implies a relevant loss of information. Finally, a last limitation is related to the inherent uncertainty that resides in the activities that make up a project. In the EVM method, the uncertainty that resides in the individual activities is not fully considered because the individual activities are considered as a completely deterministic component. Therefore, the uncertainty associated with the activity data must be evaluated in order to have more realistic measures of both project performance and actual project progress.

2.2.1 Improvements to the EVM

In order to understand and overcome these limitations, solutions have been developed over the years that lead to significant improvements. The main methods will be mentioned and analysed below:

- Lipke was the first to introduce the concept of Earned Schedule (ES) for the calculation of the SPI. the factor related to the cost in the ES was replaced with the use of time in the calculation of the performance of the schedule, used in the EVM method. In more detail, the objective was to compare the cumulative EV with the PV: the instant of time at which the actual EV should have been achieved, according to the schedule (as shown in Figure 4), is identified by projecting the cumulative EV onto the PV curve. This determines the point at which the expected value (PV) is equal to the accrued EV. Depending on whether the program is ahead or behind schedule, this point may be before or after AT. (Lipke, 2003-2004)

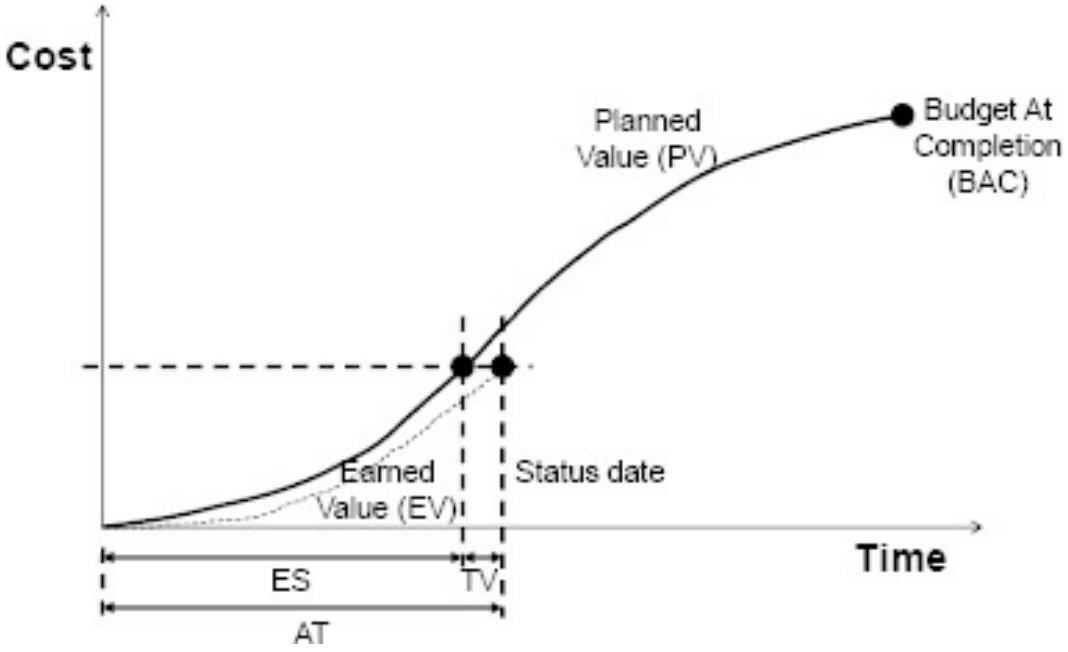

Figure 5:ES

Operationally, the following equations are used to analyse project progress and to compute estimates to finish: $ES = C + I$ where: C represents the number of times increments for which EV exceeds PV; I, on the other hand, is given by:

$$I = \frac{EV - PV_C}{PV_{C+1} - PV_C}$$

$$SPI_t = \frac{ES}{AT}$$

$$CF = \frac{TEAC}{PD} = \frac{1}{SPI_t}$$

where CF is a completion factor that indicates the estimated time to completion toward the unit. When the value of CF is 1, based on the current progress of the work, it indicates that the duration of the project is on track. If it is greater than 1, this means that the project will finish late, while if it is less than 1, the project will finish early. As can be seen, the calculation of ES is relatively simple and does not require any additional data to those used for EV analysis, other than having data collected at least monthly and accurately. Moreover, compared to the EVM method, it is more intuitive because it uses time as the unit of measure for Schedule Variance, rather than expressing it in units of cost. The relative estimate to finish of the times can therefore be estimated in the following way:

$$TEAC = \frac{PD}{SPI_t}$$

- Vandevoorde and Vanhoucke (2006), as well as De Marco and Narbaev (2013) in the construction industry, applied the two methods EV and ES for end-of-project estimation of the duration of some projects. Their results showed that the unreliability of the estimates obtained at the end of the project with the EVM method is fully overcome with the ES method, which provides more valid and reliable results throughout the duration of the projects.
- Elshser (2013) following the development of ES, proposed a method that can incorporate asset sensitivity information into the method.
- Naeni et al (2011) to overcome the intrinsic uncertainty of the activities have proposed a method based on "Fuzzy Theory" through which it is possible to

partially manage the uncertainty arising from the lack of certain information on the total workload to carry out the activities, which make up the project. In practice it is a matter of expressing the progress of the activities in words (such as "very high", "high", "low") and then transforming these expressions into numbers that can be used with the EV method.

- Lipke et al (2009) and then Kim and Reinschmidt (2010) applied real methods including EVM, ES, probabilistic methods, statistical forecasting and testing methods that can further refine a project's cost and duration predictions. All of this is accomplished by integrating uncertainty, through the use of upper and lower tolerance limits, in performance measures and variations in actual project trend.
- Caron et al.(2013) proposed a Bayesian EVM approach, using Expert Opinions in addition to project data-sets, to determine time and cost estimates at completion. In fact, expert opinion can provide useful advice and solutions regarding the detection of future threats and opportunities and methods to identify corrective actions.
- Kuhl and Graciano (2014) proposed an EVM-based simulation method to model and analyse the project using stochastic time and cost parameters.

These are just a few of the examples of additions and modifications to the EVM method found in the literature; others inherent in the introduction of risk management in the estimation of costs to completion are reported in the following sections and presented in more detail as they are of greater interest to this study.

2.3 Project Risk Management

Risk is an element that is always contained within projects, but until now has never been taken into consideration. We can also consider that the more innovative a project is, the riskier it is, as there are always different development that do not always respect the expected trend. We can consider the novelty as one of the main factors of success for a project, therefore knowing how to manage the risk dependent on these developments is a very complex job and one of the main skills, qualifying to work for projects. Risk therefore derives from the uncertainty that is characteristic and intrinsic of projects, but it is not necessarily an issue, so, a serious problem that cannot be solved through careful management. Rather, it is an uncertain event or situation that, if it occurs, has a positive or negative effect on one or more of the project's objectives and must therefore always be kept under control to avoid getting out of control. Risks, which therefore pose a threat to the project, can be accepted if they are offset by the reward of taking the risk, and they can also be well managed by the project manager if they are identified immediately. It is necessary for organizations to balance risks and opportunities, carefully planning countermeasures to deal with uncertainties through a proactive attitude and the use of robust Project Risk Management tools designed to formulate managerial responses that best manage risk. Risk management techniques can be either reactive or proactive. The proactive approach is certainly the preferable one, since it is the one that allows to prevent rather than correct the critical issues arising from the occurrence of a risk with negative impact, its limitation is that it is more expensive than a reactive approach because it is necessary to foresee the possibility of this risk. Since risk factors are an intrinsic and defining part of projects, the proactive approach is more about prevention than total elimination of risk. Project Risk

Management (PRM) is the systematic process of identifying, analysing, and responding to project risks. The objective of PRM is to try to maximize the probability of positive events occurring in the project development cycle and at the same time minimize negative events that could therefore divert the progress. Proper project risk management involves six steps (see figure 6)

1- Risk Management Planning
2- Risk Identification
3- Qualitative Risk Analysis
4- Quantitative Risk Analysis
5- Risk Response Planning
6- Risk Monitoring and Control

Figure 6: Risk Management Steps

The Risk Planning, Risk Identification and Risk Analysis phases are also known as Risk Assessment. Each of the six phases of PRM involves different parties and uses

specific techniques. In addition, it will be necessary for each process to produce its respective output as described in the table:

Table 2: Output of the processes of risk management

PROCESS	OUTPUT
Risk Management Planning	Risk Management Plan
Risk Identification	Risk Register
Qualitative Risk Analysis	Risk Priority Risk
Quantitative Risk Analysis	Analysis of the project's likelihood of meeting time and cost objectives
Risk Response Planning	Risk response (mitigation) plan
Risk Monitoring and Control	Assessment plan and corrective plan, update risk response plan checklist to identify risks in future projects

The specific steps are analysed below:

1- Risk Management Planning

The risk management planning is used to schedule how risk management will be structured and subsequently executed during the course of the project, becoming a key subset of the overall project management plan

2- Risk Identification

The goal of risk identification is to identify all potential risks inherent in the project, through group discussions, examination of similar past experiences, and formalization of project constraints/opportunities. Risks, in fact, can be thought as constraints and opportunities, both of which are affected by uncertainty. Risk identification is an activity that requires the involvement of all project team members,

but the responsibility always lies with the functional managers or team members responsible for a particular activity. It is useful to use a checklist of risks that the organization must implement over time, accompanied by supporting documentation and signed off by the project manager at the end. Risk identification can also be done using more sophisticated techniques than a checklist, such as the brainstorming, the Delphi method, personal interviews, cause-and-effect diagrams, SWOT analysis and the Risk Breakdown Structure (RBS) matrix. Just as the WBS (Work Breakdown Structure) is the project manager's primary tool in defining the scope and work of the project, similarly the RBS is one of the most developed methods for structuring and guiding the risk management process (Hillson, 2002). In addition, the identification process must not only determine what risks may affect the project but must also document the characteristics. This is all contained Risk Register. The preparation of the Risk Register should begin at the stages prior to the launch of a project, such as at the approval stage of a contract. The Risk Register, together with the checklists used to identify them or the Delphi technique questionnaires, should be contained in a risk database. The historical archive of all Risk Registers contained in the database should become an essential source of information for future projects. This archive represents a company's historical asset.

3- Qualitative Risk Analysis

The quantitative analysis estimates, for each identified risk, the probability of occurrence and the effects of the risk in terms of impact, corresponding to the missed or borrowed achievement of project performance. It should then be possible to translate, through the quantitative analysis, the risk into terms of economic impact. Quantitative risk analysis is a quick and relatively simple tool for prioritizing risk response planning. In particular, it considers the risk as a combination of two

variables: the severity of the consequences (impact) and the frequency of occurrence (probability) of the risky event. Therefore, after the classification, each risk/opportunity is assigned the respective value of impact and probability always taking into account the scope, cost, time and quality in which the project is developed. At this point, in order to make an overall assessment of each event and plan the correct preventive actions, it is necessary to construct the probability/impact matrix. In order to quantify the probability of occurrence of the event, it is possible to base the estimation on past experiences and carry out an objective analysis. The quality of information the organization has about the possible event is of paramount importance at this stage. However, if the event is non-recurring and the organization has no previous experience, the analysis can only be subjective, causing a significantly reducing of the understanding of the risk. Risk assessment can generally be performed using a qualitative approach, creating a risk ranking, semi-quantitative or quantitative. A representation of risk R, also referred to as risk exposure, common to the three methods is as follows: $R = p(En) \times I(R)$, where:

- $p(En)$ expresses the probability that the negative event En may occur on the basis of the identified causes of risk.
- $I(R)$ represents the impact/loss caused by the occurrence of negative events En in the event that the R.

In the context of industrial projects, typically, it is possible to represent economic damage, or time value, or even a change in performance. In addition, by assessing in detail the impacts of high-priority risks on specific project objectives, it is suggested to get guidance on what activities to address to reduce relative uncertainty. If the qualitative approach is used, scales of levels can be applied for

both probability (such as very high, high, medium, low, very low) and impact (catastrophic, critical, medium, marginal, negligible), as follows:

Probability:

Very High	High	Medium	Low	Very Low

Impact:

Catastrophic	Elevate	Medium	Low	Negligible

Once the probability of all events' occurrence ~~of all events~~ has been defined, it is cross-referenced with the level of impact they have on the project, obtaining the result R which allows to define a scale of overall riskiness of the events and therefore of priority. An example of the probability/impact matrix is as follows:

Table 3 Probability over Impact Matrix

Probability/Impact	Negligible	Low	Medium	Elevate	Catastrophic
Very High					
High					
Medium					
Low					
Very Low					

Where risks in the red area represent high priority risks that require immediate action and aggressive response strategies. The threats in the low-risk area (green area) represent the low priority risks for which monitoring is sufficient, while the risks in the intermediate area (yellow area) represent the risks with medium priority that must be kept under observation, because they could become red area risks but must still be adopted mitigation actions less than the threats in the red area. If a semi-quantitative

technique is used instead, the same principle is applied but the levels described with the qualitative approach are valorised with arbitrary numerical classes that vary for example from 1 to 5; the valorisation must be chosen in a univocal way for a project and must be reported in the Risk Management Planning. Taking a cue from the example used in the qualitative analysis, a subdivision could be done as follows:

Probability:

Very High	High	Medium	Low	Very Low
5	4	3	2	1

Impact:

Catastrophic	Elevate	Medium	Low	Negligible
5	4	3	2	1

Using qualitative or semi-quantitative techniques, it is possible to generate the risk/activity matrix, also known as the RBM (Risk Breakdown Matrix), a useful tool for classifying the most influential risks at project level and those that are individually more critical, as well as the activities most exposed to risk. In fact, this matrix associates the risks identified in the RBS (Risk Breakdown Structure) and their respective probability of occurrence with the critical activities (or activities close to criticality) of the WBS and their respective impacts. An example of RBM is shown in the figure:

	EXTERNAL SOURCES								CONTINGENCY BUDGET PER TASK
		HUMAN	POLITICAL		COMMERCIAL		SUPPLYING	OTHERS	
	Flood	Health	Regulations compliance	Environmental impact	Exchange rate variability	Contractual adequacy	Entirety	Thefts	
Patient Transport System	p11	p12	p13	p14	p15	p16	p17	p18	
Compose survey for the staff	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	10,0%	
	11,0%	4,0%	8,0%	4,0%	9,0%	10,0%	15,0%	12,0%	**€ 25,77**
	1,38	**€ 0,50**	**€ 1,00**	**€ 0,50**	**€ 1,13**	**€ 1,25**	**€ 1,88**	**€ 1,50**	
Brainstorming and filling in the survey	5,0%	5,0%	5,0%	5,0%	5,0%	5,0%	5,0%	5,0%	
	11,0%	4,0%	8,0%	4,0%	9,0%	10,0%	15,0%	12,0%	**€ 529,14**
	28,26	**€ 10,27**	**€ 20,55**	**€ 10,27**	**€ 23,12**	**€ 25,69**	**€ 38,53**	**€ 30,82**	
Collecting and analyzing the surveys	7,0%	7,0%	7,0%	7,0%	7,0%	7,0%	7,0%	7,0%	
	11,0%	4,0%	8,0%	4,0%	9,0%	10,0%	15,0%	12,0%	**€ 36,08**
	1,93	**€ 0,70**	**€ 1,40**	**€ 0,70**	**€ 1,58**	**€ 1,75**	**€ 2,63**	**€ 2,10**	
Get-together PC, ICT committee & board	20,0%	20,0%	20,0%	20,0%	20,0%	20,0%	20,0%	20,0%	
	11,0%	4,0%	8,0%	4,0%	9,0%	10,0%	15,0%	12,0%	**€ 995,52**
	53,16	**€ 19,33**	**€ 38,66**	**€ 19,33**	**€ 43,49**	**€ 48,33**	**€ 72,49**	**€ 57,99**	
Enrollment of the meeting	15,0%	15,0%	15,0%	15,0%	15,0%	15,0%	15,0%	15,0%	
	11,0%	4,0%	8,0%	4,0%	9,0%	10,0%	15,0%	12,0%	**€ 19,33**
	1,03	**€ 0,38**	**€ 0,75**	**€ 0,38**	**€ 0,84**	**€ 0,94**	**€ 1,41**	**€ 1,13**	
Registering the technical requirements	3,0%	3,0%	3,0%	3,0%	3,0%	3,0%	3,0%	3,0%	
	11,0%	4,0%	8,0%	4,0%	9,0%	10,0%	15,0%	12,0%	**€ 58,24**
	3,11	**€ 1,13**	**€ 2,26**	**€ 1,13**	**€ 2,54**	**€ 2,83**	**€ 4,24**	**€ 3,39**	
Set-up a detailed budget study	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	12,0%	
	11,0%	4,0%	8,0%	4,0%	9,0%	10,0%	15,0%	12,0%	**€ 123,72**
	6,61	**€ 2,40**	**€ 4,80**	**€ 2,40**	**€ 5,41**	**€ 6,01**	**€ 9,01**	**€ 7,21**	
Putting together the formal document	2,0%	2,0%	2,0%	2,0%	2,0%	2,0%	2,0%	2,0%	
	11,0%	4,0%	8,0%	4,0%	9,0%	10,0%	15,0%	12,0%	**€ 30,93**
	1,65	**€ 0,60**	**€ 1,20**	**€ 0,60**	**€ 1,35**	**€ 1,50**	**€ 2,25**	**€ 1,80**	
Research suppliers transport systems	7,0%	7,0%	7,0%	7,0%	7,0%	7,0%	7,0%	7,0%	
	11,0%	4,0%	8,0%	4,0%	9,0%	10,0%	15,0%	12,0%	**€ 108,25**
	5,78	**€ 2,10**	**€ 4,20**	**€ 2,10**	**€ 4,73**	**€ 5,26**	**€ 7,88**	**€ 6,31**	
Meeting with board for final proposal	4,0%	4,0%	4,0%	4,0%	4,0%	4,0%	4,0%	4,0%	
	11,0%	4,0%	8,0%	4,0%	9,0%	10,0%	15,0%	12,0%	**€ 61,86**
	3,30	**€ 1,20**	**€ 2,40**	**€ 1,20**	**€ 2,70**	**€ 3,00**	**€ 4,50**	**€ 3,60**	

Figure 7: Risk Breakdown Matrix (RBM)

With regard to this matrix, on the last line it is possible to see the most influential risks, that are those with a higher R-value (obtained as $P * I$) while in the last column it is possible to obtain a sorting of the activities considering as priority those burdened by higher risk.

4- Quantitative Risk Analysis

The purpose of quantitative analysis is to quantify the economic impact of the possible occurrence of adverse events on the Project's costs, after including the benefits of corrective actions in the assessment. The quantitative approach is a useful tool to support decision making, but because it requires a high level of effort in some circumstances, it is only applied to risks that have been assigned a high priority in the probability/impact matrix described above. These events must be considered a threat to the continuation of the project and to its completion within costs and timescales considered acceptable by management. The quantitative approach then evaluates the possible benefits of mitigation actions through "ex ante" and "ex post" comparisons of the effect of corrective actions on the expected value of the risk. The most common quantitative risk assessment techniques are those that

use simulation techniques such as the Monte Carlo method and deterministic or probabilistic models to determine project costs such as CPM (Critical Path Method), PERT (Project Evaluation and Review Technique), GERT (Graphical Evaluation and Review Technique). The availability of time and budget and the need for quantitative descriptions of risks and impacts should guide the project manager in choosing to use quantitative techniques. The literature on project risk management is full of techniques that claim to be statistically and from an engineering perspective, extremely competitive and effective for risk analysis. However, the choice of the most suitable method of analysis is fundamental to the success of the project, so it is important to know its characteristics and applicability. To do this, there have been many studies that can propose a general framework of methodologies, so as to make it easier to select a risk analysis technique (qualitative or quantitative) under certain characteristics of the project to be managed (De Marco, Jamaluddin Thaheem, 2014). In general, quantitative techniques (and particularly simulation-based techniques) require more effort to collect and process data than qualitative analysis techniques. As a result, quantitative techniques tend to be applied in projects with a higher level of risk.

5- Risk Response Planning

Risk response planning encompasses all those operational actions designed to bring risks back within limits that are acceptable to the organization. The inputs to planning should be the results of qualitative analysis, but also a part of risk monitoring and control; monitoring the progress of the project may, in fact, indicate the need to increase or decrease risk management action. An organization can adopt four different strategies to address threats or risks:

- Avoiding risks by anticipating a change in the project management plan to remove possible threats.
- Transferring risks, by assigning responsibility for managing them to a third party; this strategy is most effective where there is exposure to a financial risk. The transfer depends on the type of contract entered into.
- Mitigating risks by taking preventive action to reduce the likelihood and/or impact of a risk that may occur on the project.
- If the project is to be completed in the first year of operation, the project manager should be able to determine whether the project has been completed. The project manager should have the ability to determine the extent to which the project will be able to deliver the project. These are included in the forecasted income statement. Already with a semi-quantitative technique, it is possible to define for each project risk a Contingency to put in reserve, as follows:

$$Cost\ Contingency\ (R_i) = P_i \times I_i \times C_i$$

where C_i represents the estimated cost of mitigation activities for each identified risk. The ratio of the total expected possible damage (total EVM) to the contingency values gives an assessment of the level of risk acceptance, also referred to as the level of risk exposure. It is useful to assess the overall risk exposure to determine whether it is appropriate to continue with the project without changing the environment, in the case of low exposure, and also to compare the risk levels of multiple projects within the same portfolio to enable statistics to be collected or high-level business strategies to be implemented.

6- Risk Monitoring and Control

Risk monitoring and control must continually keep risks under control, scale back some areas of risk and place greater emphasis on others. As mentioned, even a risk-acceptance approach that involves simply paying constant attention to the occurrence of risks, such as the "time-out" required in basketball games, can guard against and contain any risks. The monitoring and control of risks in the broadest sense must include the continuous identification, analysis and planning of new risks and their registration in the risk database, as well as the monitoring of residual risks and the review of the execution of responses to risks while evaluating their effectiveness. The risk monitoring and control process extends throughout the project life cycle and must apply techniques, such as variance and trend analysis, that involve the use of data on the performance achieved by project execution. We will now look at methodologies for integrating end-to-end cost estimation with risk analysis and related contingency.

2.4 Risk Management related to cost

In this paragraph, the main studies aimed at analysing the possible link between the estimate of costs to completion and the possible inherent risks of a project are discussed in order to overcome the limitations presented by the EVM method. In fact, in addition to the limitations previously exposed, the EVM method, is based strictly on the detection of time and costs and the related estimate at the end of a project, without taking into account the reasons why often this evaluation is not in line with the original. In particular, it does not analyse whether such deviation is actually intrinsic to the project or if it comes from outside. In this regard, some researchers have done some deep-dive by analysing EVM and risk analysis in parallel. Others have studied whether the possible limitations of EVM could be

solved by developing more complex methods based on linear or non-linear regressions. And finally, still others have tried to integrate contingency cost management into the CEAC formula as an integral part of project monitoring. Let's see below the various developments, focusing mainly on studies concerning the use of non-linear regressions and the inclusion of contingency costs in the CEAC estimation, as they are of greater interest for the study presented below.

2.4.1 Literature behind risk analysis

In the literature, no model has ever been developed to deal jointly with the estimation of costs to finish and the assessment of risks. The two typical disciplines of the project management (estimates to finish and risk management) are in fact parallel, used jointly in order to study the course of the plan but never integrated in a single procedure. However, it is easy to see the enormous advantages that such an overview can generate in terms of accuracy and precision of time and cost estimates. These last ones would be in fact modernized not only regarding the schedule or regarding the uncertainty on the percentages of completion of the activities, like some of the extensions to the method EVM introduced in the preceding chapters already do, but also regarding the happened risky events and for which the due mitigating actions have been activated. The research carried out in this direction was as follows:

- Vanhoucke (2011) developed a Monte Carlo simulation-based approach to predict the total duration of a project by combining two approaches: the top-down and the bottom-up methods. While the former uses the EVM strategy, the bottom-up approach is based on the Schedule Risk Analysis method.

- Pajares-Lopez- and Paredes (2011) integrated the variability and risk analysis methodology within the basic EVM approach. They obtained the information regarding the probability and the distribution function of duration and cost, through quantitative risk analysis to calculate maximum overrun levels (with a certain level of confidence). In other words, they obtained a measure of the "expected" or "planned" variability of the project (assuming the probabilistic nature of costs and durations of activities) that allows them to be able to control the overruns experienced during the project.
- Acebes et al (2015) proposed an approach based on Monte Carlo simulation and statistical learning techniques to integrate risk analysis within the EVM method, thereby analysing whether current costs and any deviations from planned values stayed within the expected variability limits.
- Kim (2014, 2015) introduced a model to facilitate project monitoring. Specifically, they used an algorithm that is able to estimate the duration of a project based on the possible risks associated with false alarms in the early stage and on incorrect trends that affect the duration of a project. All of this using probabilistic methods.
- Due et al (2016) by using the Markov chain simulation method applied to the cost indicators for each period, developed a method for predicting the cost-to-end estimate based solely on the sum of each cost for each simulated period.

2.4.2 An alternative to the EVM

After having thoroughly analysed all the researchers who compute the EVM through the methods dictated by the literature, we move our attention to examine instead

those who in recent times have begun to apply the Machine Learning (ML) models aimed at optimizing the calculation of EAC.

We can find through the article "Improving the results of the earned value management technique using artificial neural networks in construction projects" (Balali A., Valipour A., Antucheviciene J., Šaparauskas, J., 2020), how they applied both linear regression and ANN to a project. To perform ANN, the variables they considered with relative importance are the following:

Table 3. Prioritization and importance coefficients of the study factors using ANN.

Priority	Sign	Factor	Factor's Coefficient
1	F1	Project Schedule	0.81
2	F2	Payment status	0.65
3	F3	Inflation rate	−0.58
4	F4	Fortuitous events	0.42
5	F5	Qualification of project management team	0.4
6	F6	Delivery of land	0.38
7	F7	Conflicts	−0.33
8	F8	Climate	0.24
9	F9	Minor contractors	0.21
10	F10	Plans	0.20
11	F11	Relationship among project's parties	0.14
12	F12	Risk management	0.12
13	F13	Accessibility of materials and appliances	0.1
14	F14	Initial geotechnical studies	−0.017

Figure 8 Table of Importance from the paper

While by applying the linear regression the results obtained are slightly different. Below there is the table with the outputs:

Table 6. Multiple regression model's results.

Factor	Sign	Unstandardized Coefficients		Sig	Standardized Coefficients
		B	Std. Error		β
(Constant)		−4.999	1.154	0.000	
Payment status	F2	0.155	0.065	0.022	0.230
Climate	F8	0.098	0.098	0.324	0.105
Conflicts	F7	0.201	0.084	0.023	0.254
Plans	F10	0.263	0.081	0.003	0.321
Accessibility of materials and appliances	F13	−0.031	0.105	0.766	−0.032
Fortuitous events	F4	−0.031	0.113	0.784	−0.026
Delivery of land	F6	0.062	0.096	0.523	0.062
Minor contractors	F9	0.208	0.072	0.007	0.283
Project schedule	F1	0.238	0.084	0.008	0.311
Qualification of project management team	F5	0.060	0.089	0.505	0.072
Inflation rate	F3	−0.029	0.014	0.040	−0.206
Risk management	F12	0.259	0.081	0.003	0.333
Relationship among project's parties	F11	0.222	0.082	0.011	0.297
Initial geotechnical studies	F14	0.061	0.072	0.400	0.085

Figure 9 Results from the regression

Comparing the results obtained with the traditional EVM, we can note that:

MSE	R	Model
0.0152	0.727	Traditional EVM
0.00206	0.896	ANN
0.012	0.864	Multiple regression

Figure 10 Comparing the results

ANN both in terms of MSE and R obtains the best value among the three approaches analysed, while multiple regression still remains better than the traditional method in both parameters calculated.

Another interesting article is: "The Factors Affecting on Earned Value Management" (Rezouki S.E., 2020).

In this paper, the calculation of EVM through both linear and non-linear regression is carried out, but the most important part is the calculation of the new variables.

In fact, in addition to the use of the classic variables that have already been mentioned over and over again such as CPI and SPI, all the following indexes are computed using the Leven Berg-Marquardt technique to develop the NLR equations. This technique is based on the insertion of new variables in a nonlinear equation built according to some values of equation parameters and checked by the "coefficient of determination" test. The new variables are:

- **COI** → Cost Index,
- **TII** → Time Index
- **QUI** → Quality Index
- **RSI** → Risk Index
- **SAI** → Safety Index
- **SOI** → Social Index

All this parameters applied for 15 different projects presented the following results:

Project	CPI	SPI	COI	TII	QUI	RSI	SAI	SOI	ETC*	EAC*	VAC*
P.1	0.86	0.99	0.83	0.79	0.66	0.64	0.15	0.83	69,895,054	2,940,241,084	-76,471,084
P.2	0.92	1.05	0.85	0.94	0.77	1.00	0.44	0.85	80,884,348	6,610,675,348	-110,675,348
P.3	0.77	0.86	0.74	0.25	0.16	0.24	0.38	0.74	206,385,189	3,686,555,999	-214,211,189
P.4	0.91	0.72	0.65	0.16	0.22	0.20	0.30	0.65	1,947,198,512	21,536,650,285	3,578,031,475
P.5	0.82	0.23	0.48	0.28	0.20	0.23	0.19	0.48	7,818,581,904	14,858,637,292	12,618,734,708
P.6	0.70	0.58	0.80	0.32	0.39	0.30	0.35	0.80	522,731,053	4,122,876,335	388,110,665
P.7	0.83	0.90	0.38	0.84	0.70	0.64	0.29	0.38	170,775,490	5,069,775,490	-170,775,490
P.8	0.97	0.93	0.72	0.94	0.83	1.00	0.62	0.72	265,419,075	7,098,197,727	1,337,331,473
P.9	0.87	0.57	0.71	0.82	0.66	0.20	0.24	0.71	198,841,110	1,178,966,940	557,393,060
P.10	0.85	0.83	0.29	0.79	0.77	0.57	0.51	0.29	720,489,510	14,341,244,310	616,755,690
P.11	0.72	0.73	0.81	0.94	0.77	0.19	0.99	0.81	15,815,861	696,934,271	-124,565,860
P.12	0.86	0.61	0.60	0.25	0.16	0.30	0.35	0.60	94,635,017	484,780,297	178,716,623
P.13	0.66	0.41	0.90	1.01	0.63	0.93	0.40	0.90	72,232,356	373,490,718	216,701,202
P.14	0.90	0.82	0.46	0.72	0.94	0.59	0.13	0.46	2,182,024,039	22,146,739,737	7,453,260,263
P.15	0.66	0.68	0.22	0.17	0.25	0.19	0.13	0.22	26,129,712,162	116,910,652,662	8,788,856,138

Figure 11 Results with new variables

That compared to the EAC estimated by the classical method gives the following output:

No.	Project	EAC	EAC*	Residual Value
P.1	Construction of gate in Basra Sports City	3,318,319,110	2,940,241,084	-378,078,026
P.2	Construction of Olympic stadium in Najaf	7,097,598,913	6,610,675,348	-486,923,565
P.3	Construction of a closed hall in Saydiya	4,490,542,981	3,686,555,999	-803,986,982
P.4	Yarmouk Intersection/Mosul	27,590,777,145	21,536,650,285	-6,054,126,860
P.5	Muthanna Intersection	33,684,475,541	14,858,637,292	-18,825,838,248
P.6	Pavement the entrance of Rabia Nineveh	6,486,748,256	4,122,876,335	-2,363,871,921
P.7	Construction of a closed sports hall in Kirkuk	5,938,181,818	5,069,775,490	-868,406,328
P.8	Construction of Schools 12 classes in Salah Al din	8,726,409,517	7,098,197,727	-1,628,211,790
P.9	Construction of the General commission Taxes/Karbala	2,000,256,796	1,178,966,940	-821,289,855
P.10	Construction of a water purification unit for Basra Sports City	17,689,291,948	14,341,244,310	-3,348,047,638
P.11	Adnan Ghazwan School 18 classes	799,153,360	696,934,271	-102,219,090
P.12	Zuhawr Al-Iraq School 18 classes	774,097,778	484,780,297	-289,317,481
P.13	Sibawayh Al-Obeidi School 18 classes	888,667,735	373,490,718	-515,177,017
P.14	Al Shaiba Intersection/Basra	32,729,042,128	22,146,739,737	-10,582,302,391
P.15	Basra Sport City Street 20 km length	189,126,959,375	116,910,652,662	-72,216,306,713

Figure 12 Comparison to the EVM Model

2.4.2 The Gompertz Model

The most relevant alternatives to the limitations highlighted by the EVM method, based on cost and time indices, are those based on linear and non-linear regression. By using these mathematical methodologies, the application boundaries of traditional cost/time estimation methods can be extended. These methods turn out to be more complex than the previous ones, but they are able to generate a better forecast at an early stage (3 2005). In fact, nonlinear formulas better describe the nonlinear relationships between input and output variables and are used to build the cost growth model. In addition, the S-curve, representative of the growth model, is generated by sigmoid models or also referred to as growth models. These models describe situations in which the data follow a growth path, with a growth rate that

increases monotonically to a maximum, before gradually decreasing to zero. The figure shows: the Gompertz curve.

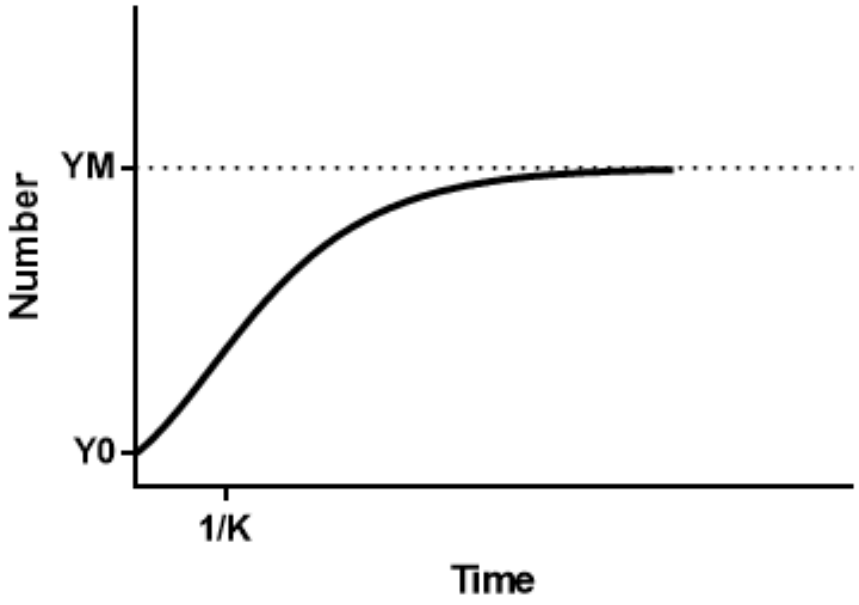

Figure 13: Gompertz Curve

In the literature and in many other fields, growth models have been widely applied combined with linear regression to study cumulative cost growth. However, little investigation has been done to combine these two techniques in order to obtain cost-to-finish estimates in complex projects. In this regard, De Marco and T.Narbaev (2014), developed an estimation method by integrating the ES method with a growth model using nonlinear regression. The methodology was mainly developed to more quickly and accurately provide CEAC in the early, middle and final stages of a project. The main objectives that were set in order to derive the best results were as follows:

1- Develop a formula based on integrating the ES method across four growth curves (Logistic, Gompertz, Bass, and Weibull), selecting the one that best represents the progress curve of the projects (S-curve)
2- Validate the methodology across nine projects in the construction industry
3- Select the equation that best represents project progress (S-curve) through statistical testing and compare the accuracy of the different CEAC estimates.

To achieve these goals, the authors divided the methodology into three basic steps:

1- Development of growth model equation and determination of its parameters.

The parameters of the different growth models (α = asymptote representing the cost to end when time tends to infinity, β = intercept on the y-axis i.e., the cost at time t = 0, and γ = cost growth rate) were estimated using nonlinear regression (using statistical software, e.g., Minitab) with the following input data:

- the AC and PV values of the project normalized with respect to the BAC
- time instants normalized with respect to PD

Their analysis showed that the Gompertz-based growth method appears to be the most valid method capable of generating the most accurate estimate of CEAC compared to the other models. The generic Gompertz function is given by.:

$$GGM(x) = \alpha e^{[-e^{(\beta - yx)}]}$$

Where α, β, and γ are the model parameters (described above).

2- Calculation of CEAC using the growth model estimated parameters of.

The evaluation of the costs to finish has been determined in the same way analysed in the EVM; so by inserting only the values of the growth model calculated in the previous point, in the following way:

$$CEAC(x) = AC(x) + [growth\ model(1.0) - growth\ model(x)] \times BAC$$

where:

- AC(x) is the cost currently incurred at time x (instant in which you want to evaluate the cost to finish).
- BAC is the estimate made at the beginning of the project of the cost to finish.
- GGM is the model of growth of Gompetz: $GGM(x) = \alpha e^{[-e^{(\beta - yx)}]}$ whose parameters have been previously described.

3- Integration of ES within the equation of CEAC

In this step, the main objective of the authors was to evaluate how much the work progress can influence the estimation of CEAC through the integration of ES. This is achieved by substituting in the growth model the completion factor (CF) instead of the value 1.0, as follows: $CEAC(x) = AC(x) + \{GGM[CF(x)] - GGM(x)\} \times BAC$

where:

- CF(x) is the completion factor at time x:

  $$CF = \frac{1}{SPI}$$

  calculated using the ES method. If CF(x) = 1 the project is on track. If CF(x) > 1 the project is behind schedule while if CF(x) < 1 the project is ahead.

- The difference of two points on the S-curve of the growth model:

  $$GGM[CF(x)] - GGM(x)$$

  the point corresponding to the percentage of project completion at time x and the point corresponding to the current time x. Doing this is is possible to adjust the BAC, which multiplies this expression, not by an indicator of past performance (CPI or CR) but by the non-linear progress modelled by the GGM.

As previously described the GGM brings to better results especially in the initial and intermediate phases of the project and its integration with the progress of the work , contributes to improve the precision of the CEAC estimation. Therefore, overall the model turns out to be more accurate and precise regarding the EVM standard, In particular, it is highlighted how the progress of the works is also a factor that significantly affects the costs behavior and the relative estimates to finish. The following methodology developed by Narbaev and De Marco (2013-2014) leaves considerable room to expand its theoretical framework and application methods. Indeed, we will analyse in the next section its evolution through the integration of risk analysis.

2.4.3 Cost Contingency

This section will examine in detail at how, not only risk analysis but also contingency cost can be considered an intrinsic factor in a project's performance and how its status can affect the CEAC estimate. As described earlier, the calculation of end-point estimates and risk management are two parallel disciplines that should be considered during the monitoring phase of a project. In particular, during the evolution of a project, contingency costs are used to cover any uncertainties and risks in order to bring the project back in line with the pre-established time and cost targets. Contingency therefore, as Barraza and Bueno (2007) state, just as it may not be correctly allocated to the initial project cost estimate, it must still be properly controlled and consumed during project execution. Some studies have analysed in more detail the intrinsic evolution of the risk contingency cost, examining its possible integration in the CEAC estimation. By the way, Touran (2003) presented a probabilistic model based on the uncertainties in the costs of a project that can

calculate the contingency costs corrected for the statistical confidence level of the project. Cioffi and Khamooshi (2009), sticking to project risks and their impacts and probabilities, devised a method to estimate with certainty the possible impact of risks so that the project manager can set aside the correct amount of contingency. Xie et al (2012) presented a method for forecasting and updating contingency costs based on risk values at a certain confidence level, which may occur during project execution. However, no single method has ever been developed with the ability of capturing the impact of risk and the related contingency in estimating the costs at the end of a project. This large gap present in the research, prompted De Marco in collaboration with Narbaev (2017) and later in partnership with M. Rosso (2016), to develop two method studies capable of evaluating CEAC taking into account both performance related influences on progress and contingency cost utilization during project execution. The first study by De Marco, Nabaev (2017), thus stems from the need to have a single algorithm that generates an end-to-end estimate of costs updated not only regarding the percentage of project completion but also relative to burned-in and remaining contingency. The idea is based on the model previously proposed by the same authors (Narbaev and De Marco, 2014), in which they estimated CEAC using nonlinear regression, represented by the Gompertz growth model. In particular, this study goes to modify and extend the estimation of CEAC, going to appropriately include the factor related to contingency, so as to integrate future risk within the estimates based on past performance. In particular, the purpose of this study is actually twofold: on the one hand, it integrates within the estimate of costs to end through the EVM, the management of contingency, arriving at a single forecasting model, but on the other hand, the study presents different estimates of the CEAC "adjusted" with the curve of consumption of contingencies in three

different forms, each aimed at representing the possible logic of expenditure of the risk reserve adopted by project managers. The study is then articulated as follows:

1- Estimation of the parameters related to the original CEAC formula

As stated earlier, the study starts from the basic model of CEAC estimation, set on the Gompertz growth model (GGM) integrated with ES, according to the following formula:

$$CEAC(x) = AC(x) + \{growth\ model[CF(x)] - growth\ model(x)\} \times BAC$$

where:

- AC(x) is the cost currently sustained to the time x (instant in which it is wanted to estimate the cost to finish).
- BAC is the estimate made at the beginning of the project of the cost to finish.
- GGM is the growth model of Gompetz:

  $$GGM(x) = \alpha e^{[-e^{(\beta-\gamma x)}]}$$

  whose parameters have been previously described.
- CF(x) is the completion factor at time x:

  $$CF = \frac{1}{SPI}$$

2- Development of the new CEAC estimation model integrated with the risk contingency.

In this step the authors, after introducing the new notation $\Omega(x) = GGM\big(CF(x)\big) - GGM(x)$, have supplemented the formula by including the contingency cost component, as follows:

$$CEAC_{risk}(x) = AC(x) + \Omega(x) \times BAC + \Omega(x) \times CC$$

where CC means the initial budget prepared for the contingency which represents a predetermined portion of the BAC:

$$CC = BAC \times k$$

In this way, in the final cost formula, in addition to the estimate of the remaining BAC, there is also the remaining contingency factor which, since it is closely related to the BAC and its relative consumption, will also be linked to the Gompertz growth curve, as we will see later. This implies that, as the project proceeds, the total amount of contingency is gradually consumed by the project team to initiate corrective actions due to the emergence of risks, until the CC is completely exhausted (Ford 2002).

3- Application of the CEAC estimation method to three different contingency consumption logics and on three distinct categories of projects

The last investigation carried out by the authors in order to deepen the criteria of use of the contingency during the evolution of a project, has led to develop three different formulas of $CEAC_{risk}$, based on three possible logics of consumption of the contingency of risks, adoptable by the project manager. In particular, as the BAC is spent in a cumulative way in accordance with the S-curve better estimated by the GGM, in the same way the cumulative curve over time of the contingency of risk, behaves as an inverted S as shown in the figure:

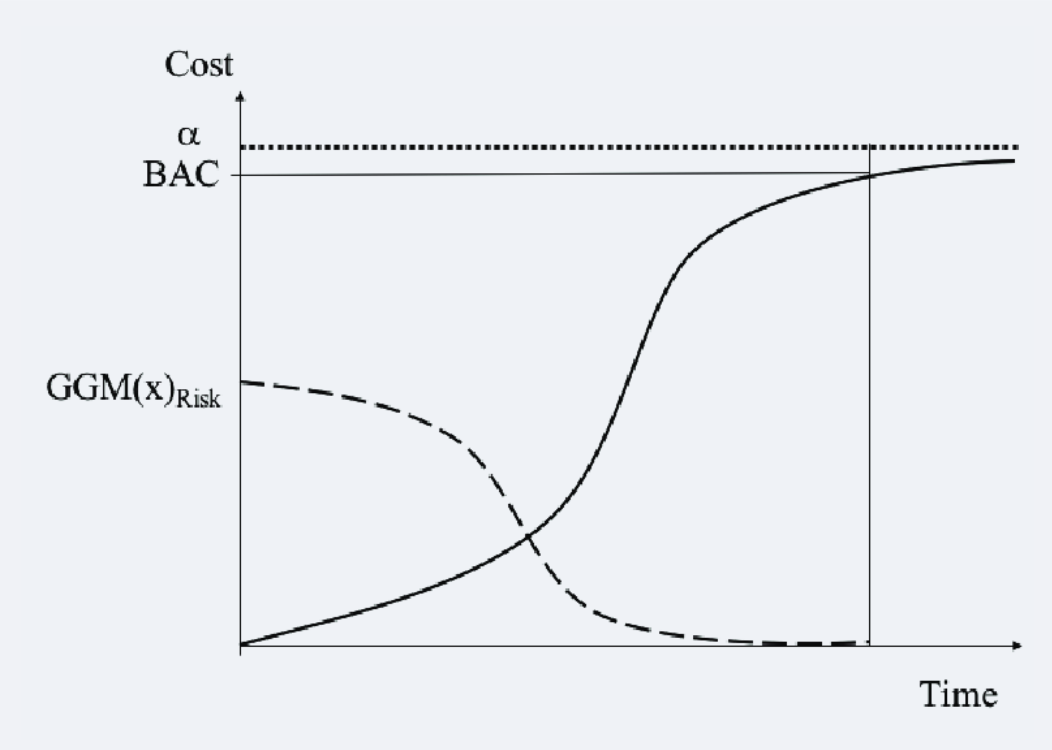

Figure 14: Contingency over time

The symmetrically opposite trend of this curve with respect to the curve of the costs is easily explained thinking about the erosion of the budget to cover the risks in the time: it starts with a total value of "full" contingency at the beginning of the project that is eroded as the work proceeds and the assumed risky events occur, until arriving at the complete exhaustion of the amount set aside in the budget at the end of the project when, by then, the work has been concluded. Going therefore ideally to superimpose the two accumulated curves of the costs of the Project and the contingency of the risks on an only diagram it can be noticed like, in the simplifying hypothesis that the contingency is a percentage of the estimated BAC, that the two curves have a specular and symmetrically opposite behaviour, assuming that the initial budget in order to cover the risks is equal to $C_0 = k * \%BAC$.

In this way, just as the risk contingency curve depends on the BAC, similarly its consumption will depend on the Gompertz growth curve. In addition to the influence of the GGM itself, the contingency is then shaped by the factor Ω(x) which will vary

according to the three predicted consumption logics. Specifically, contingency consumption can be spent proportionally to the BAC with a constant rate or non-proportionally with a variable rate, dependent on Ω(x). Thus, the authors determined three different scenarios of end-to-end cost estimates:

- Consumption of risk contingency proportional to BAC

In this scenario, the contingency consumption rate is constant, equal to 1, and linear with the BAC. This logic represents the project manager's decision to spend the contingency budget over time following the same trend as the S-curve of project budgeted costs. The CEAC estimate is then calculated as follows:

$$CEAC_{risk}(x) = AC(x) + \Omega(x) \times BAC + \Omega(x) \times CC$$

- Risk contingency consumption with decreasing rate relative to BAC (reactive approach)

Contingency consumption in this scenario is not proportional to the decrease in BAC but depends on the factor $\Omega(x)^2$. In this way, the authors represent the logic of project managers aimed at consuming all contingency in the early stages of the project, decreasing utilization as the project progresses. The relative estimate of the cost to finish is as follows.:

$$CEAC_{risk}(x) = AC(x) + \Omega(x) \times BAC + \Omega(x)^2 \times CC$$

- Consumption of risk contingency at an increasing rate relative to BAC (proactive approach)

In the latter scenario, the consumption of the risk contingency increases nonlinearly with respect to the BAC and its change is governed by the factor $\sqrt{\Omega(x)}$. In particular, it is a very proactive approach in which the project manager decides in advance not to consume all the contingency at the beginning of the project, but to

keep a low rate of consumption in the early stages, and then increase it in the final stages of the project if unexpected risks arise.

$$CEAC_{risk}(x) = AC(x) + \Omega(x) \times BAC + \sqrt{\Omega(x)} \times CC$$

The authors then applied the different formulas to data from three different type of project according to the distribution of workload:

- Front Loaded: projects where most of the work is planned in the first half of the project
- Mid Loaded: projects where the workload is greatest in the middle phase of the project
- Back Loaded: projects where the main activities planned for the second half of the project

It has therefore emerged that, in the case of Front Load and Mid Loaded projects, the best estimate of the costs to finish is the one in which a greater consumption of contingency is considered, above all in the initial phases (reactive approach). This is also confirmed by the fact that, above the Front Loaded projects, the load of the activities and therefore the probability of the emergence of new risks is greater in the early stages of the project. As could be expected, for Back-Load projects, since the workload is higher in the final stages of the project, the occurrence of risks is also higher at this stage. Thus, the best estimate of CEAC is the one where an increasing rate of contingency consumption is considered compared to BAC. Therefore, it can be concluded that this new model on one hand provides a finite cost estimate using non-linear regression and the Gompetz growth model, and on the other hand provides a procedure for the analysis and evaluation of project risks that leads to the quantification of a contingency to be budgeted according to three different consumption logics. With the aim of integrating and expanding the developed model,

De Marco, Narbaev and Rosso, continued the study, going to analyse in more detail the evaluation of risks contingency of . In particular, they resumed the analysis starting from the conclusion previously exposed based on the fact that, the trend of the contingent risk curve was nothing more than an inverted S-curve from the plan cost modelling curve, shaping with a Gompertz growth curve, which is described as follows:

$$GGM(x)_{risk} = \alpha - GGM(x)$$

Placing the total project cost curve and the cumulative contingency consumption curve on the same plane, as seen above, results in a mirror-image trend. Assuming that the total contingency is a percentage k of the BAC the authors modelled a completion factor $CF(x)$, similar to the cost factor, for the risk contingency as well. In this case, $CF(x)_{risk}$ represents the erosion of the budgeted contingency at a given time instant x during the project life cycle. Given the diametrically opposed trends of the two cumulative curves of project cost and risk contingency, they modelled the completion factor for risk as follows:

$$CF(x)_{risk} = 1 - CF(x)$$

In this way, at each instant of time x, corresponding to a certain completion factor $CF(x)$, the value of the current costs and the corresponding remaining contingency are obtained. Thanks to the introduction of the factor of completion $CF(x)_{risk}$ (that it would be more opportune to call "factor of erosion of the contingency") it is succeeded therefore to have an idea of the course of the risks, parallel to the schedule of plan. The use of this factor leads, therefore, to adjust the BAC estimated at the beginning not only with respect to the trend (in terms of completion) of the schedule, but also with regard to the residual contingency at time x. The BAC is then increased by the residual contingency, which changes at each instant of time x,

following a trend symmetrically opposite to that of the cumulative project costs. Applying this type of reasoning, an adjusted BAC (BAC_{adj}) is obtained with respect to the risk analysis conducted in parallel with the process of estimating costs to finish:

$$BAC_{adj} = BAC\{1 + k[\alpha - GGM([CF(x)])]\}$$

Placing this formula within the cost-to-finish estimate developed earlier by Narbaev and De Marco (2014-2017), we obtain:

$$CEAC(x) = AC(x) + \{GGM[CF(x)] - GGM(x)\} \times BAC_{adj}$$

$$= AC(x) + \{GGM[CF(x)] - GGM(x)\} \times BAC\{1 + k[\alpha - GGM([CF(x)])\}$$

As you can easily see there is no additional computational effort compared to the basic method of estimating costs to finish, if not to make the estimated contingency (with one of the qualitative methods in the literature) as a function of the BAC, calculating a reference k for the project under consideration (or in an inverse but coarser way, deciding a k a priori). After testing the model on 9 construction projects the results were of considerable interest: in fact, it was found that especially in the initial and intermediate stages of monitoring a project, the model provides very accurate estimates and with a very low percentage error. Starting therefore from the gaps present in the research and from the awareness that the dynamism and unpredictability of the project environment must be as much as possible controlled and monitored as a factor constituting the final cost of the project, a very effective estimation algorithm has been devised, able to update the CEAC, considering also the status of the project risks. To this day, the model remains very compelling and leaves ample room to expand its applicability. With the aim of further understanding the behaviour of the model in a dynamic environment, it would be interesting to evaluate how the method behaves in other sectors different than construction, and possibly what potential improvements could be made to try to better capture the

large risk events that affect cost and time to completion. As we will see later, such is the main intent of this study.

3 The introduction of new Variables

3.1 Starting point

As highlighted in the literature chapter, it is only in recent years that methods have been developed to integrate the cost-to-finish estimation and the risk analysis with related risk contingency management (De Marco-Narbaev, 2016-2017). Previously, there have been early approaches towards such correspondence aimed at improving the related estimation algorithms, but they have never been so relevant to the estimation of the end costs of a project. Thus, methods have arisen to estimate the total project duration with Monte Carlo simulations in order to identify the most critical activities and thus the relative probability of delay (Vanhoucke, 2010- 2011). In addition, methods have emerged to estimate the evolution of the risk value over time in order to calculate possible maximum levels of overrun (Pajares and Lopez-Paredes, 2011) and growth models capable of estimating the costs at the end of a project through linear regressions (De Marco-Narabev,2014). It is precisely from these studies aimed at filling the gap in the literature related to the limitations of the EVM method that the growing interest in integrating risk management into the calculation of end-to-end estimates has arisen. In particular, two leading scholars, Narbaev and De Marco (2017), have moved in that direction, developing algorithms capable of generating an updated end-to-end cost estimation not only with respect to the percentage of project completion, but also with respect to the residual risk contingency.

3.2 Quantity-based project schedule

The traditional Earned Value Management (EVM) discussed at length during the literature section, suffers from many theoretical weaknesses, which over time has been tried to improve, as we can see from the studies of De Marco and Nerbaev and later from the model of Gompertz. They still adopt the basis of Planned Value (PV), Earned Value (EV) and Actual Cost (AC) to evaluate the global performance of the project, not taking into account everything that has happened in the past and therefore historical data, which could allow us to estimate the Budget At Completion (BAC) with greater accuracy, since the risks that have occurred can be analyzed in greater detail. They tend to focus the manager's conception on the performance of the schedule through the various current performance indicators, rather than studying historical trends to apply a more careful analysis on the project schedule. To solve the problems of traditional schedule control methods, we can see that in “The 10th International Conference on Engineering, Project, and Production Management” (Kriengsak Panuwatwanich, Chien-Ho Ko), it was proposed a new method, namely the quantity-based project schedule control model (Q-PSCM), rather than individual values. It thus adopts the concepts of EDM, but calculates the project schedule performance based on the "activity quantities" of the critical path instead of the overall "activity values", as usual in the traditional EVM. The quantity information is used to calculate the 'Estimate to Complete (ETC)' project duration using the critical path method. The overall project performance index of the project is evaluated based on the current project information, the ETC duration of the project and the planned duration of the project. The result of the case study shows that the proposed Q-PSCM can evaluate the project schedule performance more effectively and provide a more useful and effective tool for project schedule control. Another

approach that differs from the traditional EVM method comes from people like Rezouki S.E., Morthada S.B., Balali A., Valipour A., Antucheviciene J., Šaparauskas J., who have laid the foundations through their articles for a new type of approach, namely the unification of a mathematical-informatics approach to what is project management. They are the ones who algorithms have enabled a new approach to this study through the application of different kinds of machine learning.

3.3 Definition of the dataset

These described methods lay the foundation for the study presented in the following paper, so the combination of the branch of project management with machine learning and the introduction of new variables, are the concepts used in order to optimize the variables described in literature. To do this, we can divide the initial work into several steps:

1. The first step focused on finding the datasets to be used in applying the algorithms.
2. Subsequently, the variables considered relevant by the literature were calculated, such as CPI, SPI.
3. Finally, the third and last step of the data collection part was the creation of new variables.

The variables, drawn from the literature and used up to now are:

- Project: assigns a number to each project, to keep track of the change
- Category: 4 different categories to allocate projects in order to add a percentage of contingency

	Contingency%
Aerospace	0.150
Construction	0.075
Manufacturing	0.100
Information System	0.200

Figure 15: Contingency for category

- TP: Time Period of the periodic check
- Start Tracking Period: Start of TP
- Status date: End of TP
- %WS: Percentage of work schedule
- d%WS: Variation of work schedule
- %WP: Percentage of work planned
- d%WP: Variation of work planned
- BAC: Budget at Completion, computed through the EVM method
- PV: Planned Value
- dPV: Variation of Planned Value
- EV: Earned Value
- dEV: Variation of Earned Value
- AC: Actual Cost
- dAC: Variation of Actual Cost
- ES: Earned Schedule
- Contingency: different from each project as described above
- SV: Schedule Variance
- SPI: Schedule Performance Index
- CV: Cost Variance

- CPI: Cost Performance Index
- SV(t): Schedule Variance over time
- SPI(t): Schedule Performance Index over time
- All the different calculation of the EaC known from literature:
    - EAC(t)-PV (PF=1)
    - EAC(t)-PV (PF=SPI)
    - EAC(t)-PV (PF=SCI)
    - EAC(t)-ED (PF=1)
    - EAC(t)-ED (PF=SPI)
    - EAC(t)-ED (PF=SPI)
    - EAC(t)-ES (PF=1)
    - EAC(t)-ES (PF=SPI(t))
    - EAC(t)-ES (PF=SCI(t))
    - EAC.1
    - EAC.CPI
    - EAC.SPI
    - EAC.SPI(t)
    - EAC.SCI
    - EAC.SCI(t)
    - EAC.0802
    - EAC.0802t

The first dataset that will be used for the application of ML algorithms will be composed of the variables defined above that are taken from the literature. This dataset will be called "*DB.float*".

3.4 Implementation of the new variables

The new variables calculated allow to take into account new factors in order to create algorithms that can give solutions more similar to reality. So, the goal was to compute the variance of the total project duration. To do that we can consider that the total duration of the project is the same of the duration of the Critical Path. The risk is that a sub-critical path with a higher variance can became the critical path itself in the worst case. We can summarize the steps that allow us to compute our new variables as follows:

1. For each activity estimate the variability through the PERT model, so Optimistic (O), Most Likely (ML) and Pessimistic (P).
2. Take as assumption a Beta Distribution for all the activities
3. Computing the variability:

$$(\frac{O-P}{6})^2$$

4. Divide each project into Time Periods (TP)
5. Computing Active Tasks and the Active Critical Tasks
6. Computing through the process described above, the Variability of the Active Tasks and Active Critical Tasks for each TP as a composition of critical and sub-critical activities in order to avoid the problem relative to the sub-critical path.

So, following these steps it was possible to compute all the new variable in order to take in consideration the risks inside a project. The variables are:

- Number of Active Tasks ($PI1a$): this was useful for understanding whether active tasks were influencing (either negatively or positively) the calculation of EaC.
- Variability in each TP ($PI2a$): having divided the project into TP and having calculated the number of Active Tasks, it was also possible to calculate the variability in each TP, thanks also to the contribution of the PERT method.
- Number of Active Critical Tasks ($PI1b$): the last variable calculated was the number of active critical tasks in each TP. To do this, it was necessary to calculate the critical path in advance and then, see in each TP how many of these were active. The reason for this last variable is certainly because the deviation of the project is often due to a variation within the critical path.
- Variability of Critical Tasks ($PI2b$): as for the previous variability, it was computed by considering the PERT (optimistic, most likely and pessimistic) to obtain the variability of each task, and then it was calculated by considering only the Critical Tasks active for each TP.

The new variables defined, which are linked to the project's intrinsic risk, appear to be significant already in the planning part. Particularly, the Variability in each TP, was calculated as the composition of the duration variances in the active activities of the TP. The other 3 variables are also of particular relevance, contributing to a better understanding of how the risk can be explained and mitigated. In detail, the number of active tasks and the number of active tasks in the critical path allows to easily understand how many of the current variables could negatively influence the positive outcome of the project, estimated during the planning phase, while the variability of active tasks in the critical path allows to understand how much risk one actually faces. In the following phases, the previous variables will be analyzed in detail and

inserted into the dataset, in order to understand if they are significant for the application of the algorithms, which will be described in more detail in the following chapters. More specifically, the dataset will consist of:

- dAC
- BAC
- %WS
- Number of Active Tasks
- Number of Active Critical Tasks
- Variability in each TP
- Variability of Critical Tasks
- EMV (Expected Monetary Value)

This dataset will be called "*dAC.crit*".

After describing the new variables, and present the new dataset, they will be introduced within the different algorithms in order to more accurately recalculate the EaC. The algorithms used to find the optimal model will be described in the following chapter.

4 Algorithms

In order to be able to calculate an optimal value of the EaC, after the first part of computation of the variables used by the EVM method and having calculated the new variables defined in the previous paragraph, there was a study related to the machine learning algorithms that will be used. After a careful study, it was decided to apply both white-box models and black-box models as the former would be preferred given their transparency in terms of both process and result, but the latter, as we know from literature, gives a more accurate outcomes, although unfortunately lacking in transparency. The objective was to test most of the algorithms that could have given an optimal solution and to do so, the application of the algorithms has been carried out through the use of Python, more in detail there is a library fundamental to use: scikit-learn. It allows both the application of the different algorithms, and the subsequent calculation of metrics to understand which are the most accurate among those tested. In the following we will describe the algorithms used for the analysis, making the very important distinction that was considered earlier: white-box and black-box models.

4.1 White-Box Model

White-box models are called this way because of their transparency. They are the type of models in which the explanation of the behavior is clear and well defined; in fact, one can perfectly understand how they make predictions and which variables have the greatest influence. The characteristic elements that allow the distinction between white and black-box algorithms are that the characteristics must be understandable and the process transparent.

- **Linear regression** (see figure 16)**:** in statistics, linear regression allows to explain the relationship between dependent and independent variables. It is an approach to model the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables) in a linear manner. In the case where the explanatory variable is only one, we are dealing with a case of linear regression, whereas the more common case with several explanatory variables is called multiple linear regression. Multiple linear regression differs from multivariate linear regression, in which the dependent variables under analysis are correlated with each other. In linear regression, relationships are modelled using linear predictive functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression calculates the conditional probability distribution of the response, taking into account the values of the predictors, rather than the joint probability distribution of all these variables. Linear regression was the first approach to this new type of analysis to be studied in greater detail and subsequently to be used in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine. (https://en.wikipedia.org/wiki/Linear_regression, s.d.). Linear Regression fits a linear model with coefficients $w = (w_1, \dots, w_p)$ to minimize

the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

- **OLS:** Ordinary Least Squares Linear Regression.
- **Elastic Net:** Linear Regression with combined L1 and L2 priors as regularize
- **Lasso:** Linear Model trained with L1 prior as regularize
- **LARS:** Least Angle Regression Model
- **Ridge:** Linear Model trained with L2 prior as regularize
- **LARS Lasso:** Lasso Model implemented using the LARS algorithm
- **Bayesian Ridge:** is a linear regression model that assumes w following a Gaussian and is explained by the following formula:

$$p(w|\lambda) = \mathcal{N}(\mathrm{w}|0, \lambda^{-1}\mathbb{I}_{\mathrm{p}})$$

- **Automatic Relevance Determination:** is similar to Bayesian Ridge, but can lead to sparser coefficients w. The distribution of w is supposed to be an axis-parallel, elliptical Gaussian distribution.

$$p(w|\lambda) = \mathcal{N}(\mathrm{w}|0, \mathrm{A}^{-1})$$

- **Generalized Linear Model:** extend linear model in two ways. First, the predicted value $\hat{y}$ are linked to a linear combination of the input variable X via an inverse link function h as: $\hat{y}(w, X) = h(Xw)$. Secondly, the squared loss function is replaced by the unit deviance d of a distribution in the exponential family. The minimization problem becomes:

$$\min_{w} \frac{1}{2n_{samples}} \sum_{i} d(y_i, \hat{y}_i) + \frac{a}{2} \left|\left|w\right|\right|_2$$

- **Stochastic Gradient Descent (SGD):** is a simple and efficient approach to fit linear models. Very useful when the number of samples is very large.
- **Passive Aggressive:** it is part of the family of algorithms for large scale learning. It does not require the learning rate.
- **Robustness**
    - **Random Sample Consensus (RANSAC):** fits a model from a random subset of inliers from the complete dataset
    - **Huber:** it applies a linear loss to the sample that are classified as outliers. A sample is classified as an inlier if the absolute error of the sample is less than a certain threshold.
    - **Theil Sen:** is comparable to the OLS in terms of asymptotic efficiency and as an unbiased estimator. Differing from OLS, Theil-Sen is a non-parametric method, which means that it makes no assumptions about the distribution of the data. It is therefore not necessary for the data to follow a specific distribution; it can be applied all the time.

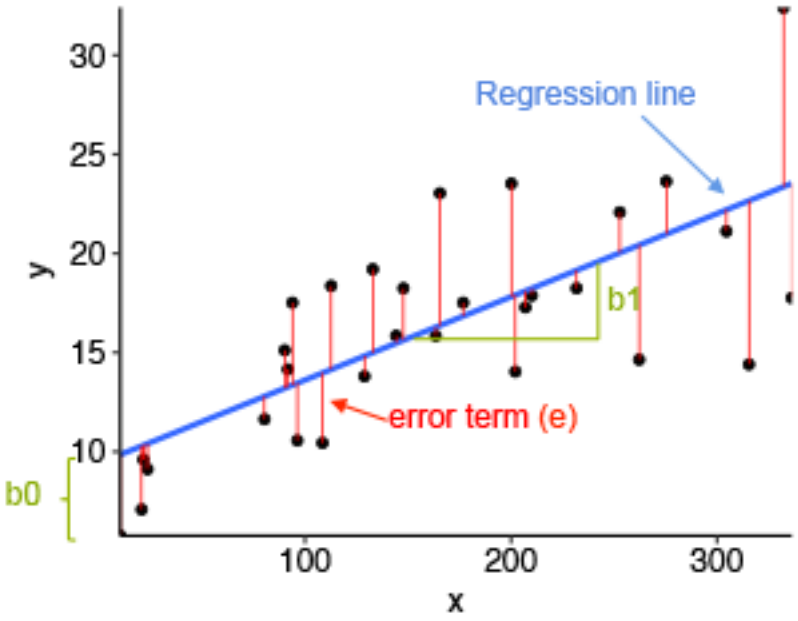

Figure 16: Linear Regression

4.2 Black-Box Model

On the other hand, black-box models, such as deep-learning (deep neural network), boosting and random forest models, are highly non-linear by nature and are more difficult to explain in general. With black-box models, users can only observe the input-output relationship, without any transparency about the process. Black-box models often result in better accuracy than white-box models, but they sacrifice transparency and accountability, which remain two important factors, particularly in this field.

- **Kernel Ridge** (see figure 17)**:** Kernel Ridge Regression (KRR) performs a combination of kernel makeup and ridge regression, which, as described in the previous section, uses linear least squares with l2-norm regularization. it is similar to the Support Vector Regression (SVR), the only difference being in the loss function taken into analysis. KRR uses squared error loss combined with l2-regularization. It was estimated that fitting KRR is typically faster for

medium-sized dataset, while the learned model is non-sparse, so lower than the SVR. (Murphy, Machine Learning: A probabilistic perspective, 2012)

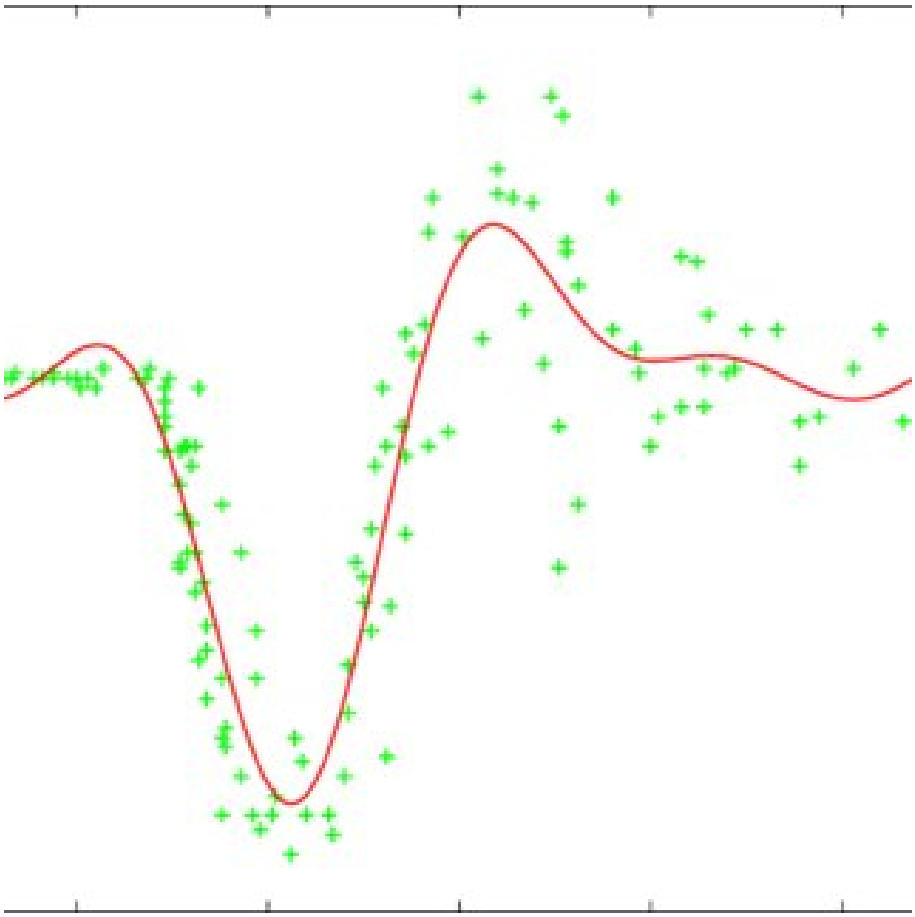

Figure 17: Kernel Ridge

- **Support Vector Machine** (see figure 18)**:** The goal of Support Vector Machine (SVM) is to find a hyperplane in an N-dimensional space (with N equal to the number of different features) that distinctly classifies the data points. It is commonly used for classification. To separate the two classes there are many possible hyperplanes that could be chosen. The goal of the algorithm is to find the one that maximize the margin distance, so that future points can be classified with more confidence. It can be extended to solve regression problems.
    - **SVR:** Epsilon-Support Vector Regression, with free parameters epsilon and C. The implementation is based on libsvm, with the fit time

complexity higher than quadratic. It is suggested for small-medium datasets.

- **NuSVR:** Nu-Support Vector Regression, similar to the previous, Nu replace the epsilon. The implementation is based on libsvm.
- **LinearSVR:** Linear Support Vector Regression, similar to SVR with kernel = '*linear*', but implemented in terms of liblinear rather than libsvm. (Chang, Ling)

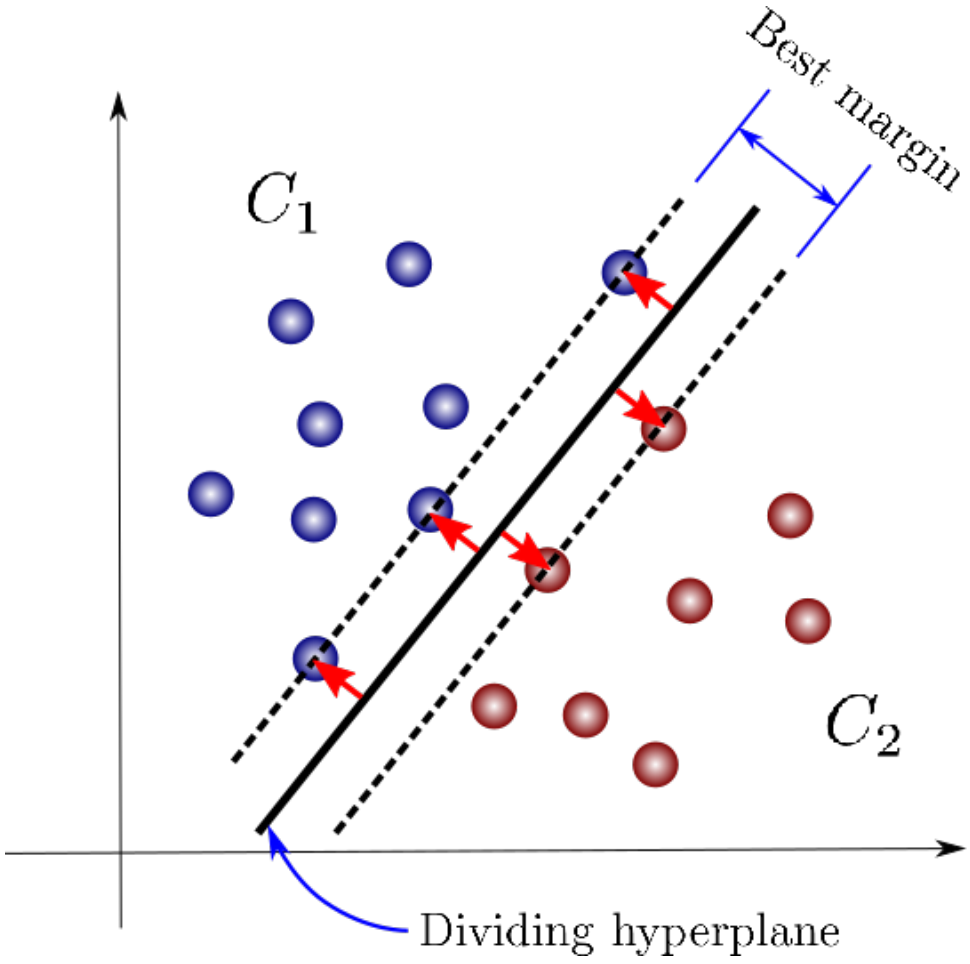

Figure 18: Support Vector Machine

- **K-Nearest Neighbors** (see figure 19)**:** the k-nearest neighbors (KNN) algorithm is an easy-to-implement supervised machine learning algorithm that can be used both for classification and regression. It uses features similarity to predict the values of new data, so to the new point is assigned a value

based on how closely it resembles the points in the training set. The first step of the algorithm is to compute the distance, there are three different methods:

- Euclidean Distance: $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
- Manhattan Distance: $\sum_{i=1}^{k}|x_i - y_i|$
- Hamming Distance, used for classification: $D = 0\ if\ x = y$ $D = 1\ if\ x \neq y$

Then we choose the number of k, that represent the neighbors we look at when we assign the value of any observation. The basic KNN uses equal weight for each class, but sometimes can be better to assign different weight in order to improve the performance. (Scikit-Learn, s.d.)

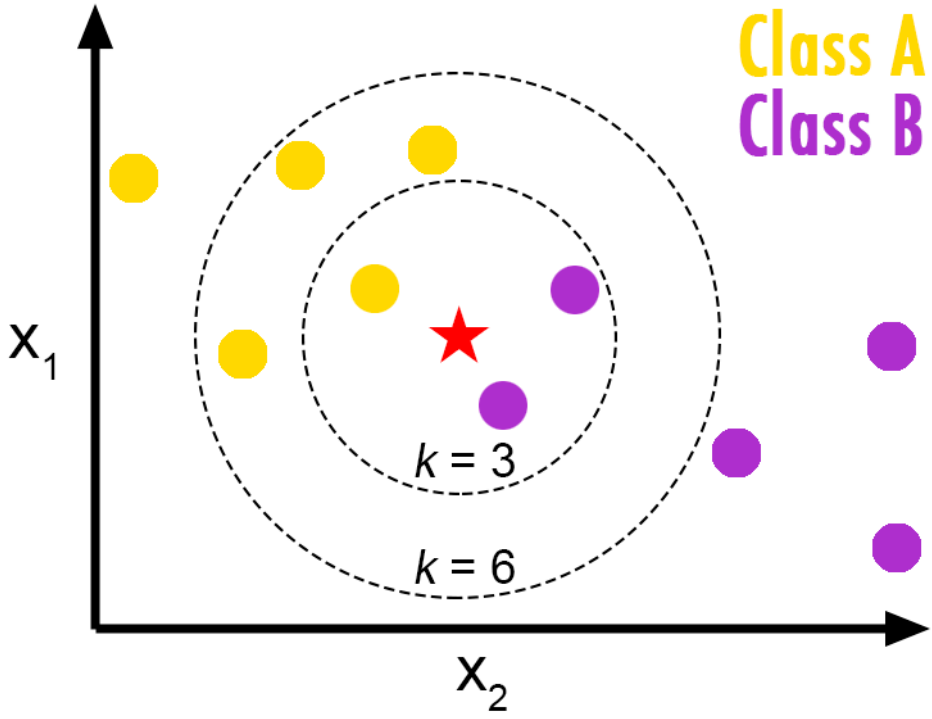

Figure 19: KNN

- **Gaussian Processes** (see figure 20)**:** Gaussian processes (GP) is a generic supervised learning method that aims to solve probabilistic regression and

probabilistic classification problems. The key points that allow its use are the following:

- In the case of regular kernels, the prediction interpolates the observations.
- The prediction follows a normal trend, which simply allows the calculation of the confidence interval, in order to decide whether it is consistent with the dependent variable or needs to be readjusted.
- Versatile: high possibility of customization, there are kernels already in place, but it is possible to set other custom ones.

There are also disadvantages to using GP, which are the following :

- They always use the whole training set for prediction, it is not possible to use only a part of it.
- In case the number of features is high, it has low efficiency. (scikit-learn, s.d.)

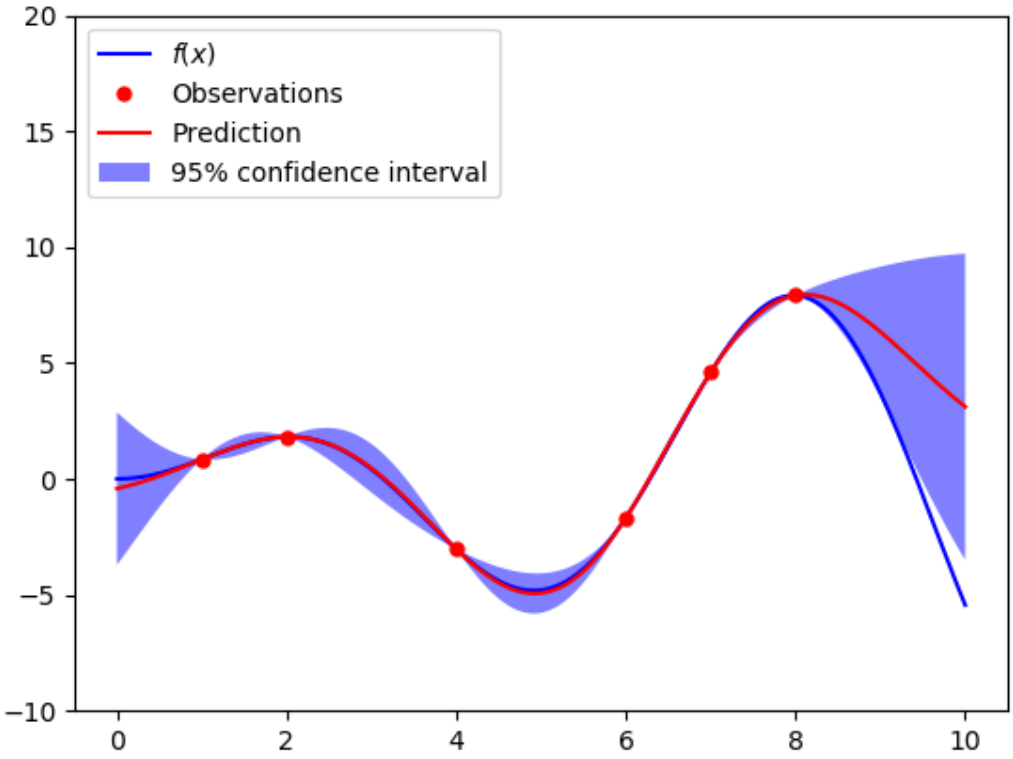

Figure 20: Gaussian Process

- **Cross Decomposition** (see figure 21)**:** Cross-decomposition contains supervised estimators that aim to reduce dimensionality as well as apply a regression model, known as Partial Least Square (PLS). In more detail, these algorithms take as input two matrices, the first containing X (the independent variables) and the second y (the dependent variable) and look for the relationships existing between them. They try to find the multidimensional direction in the X-space that explains the maximum multidimensional variance direction in the Y-space.
    - **PLS:** Partial Least Square Regression.
    - **PLS Canonical:** Partial Least Square Transformer and Regressor.
    - **CCA:** Canonical Correlation Analysis

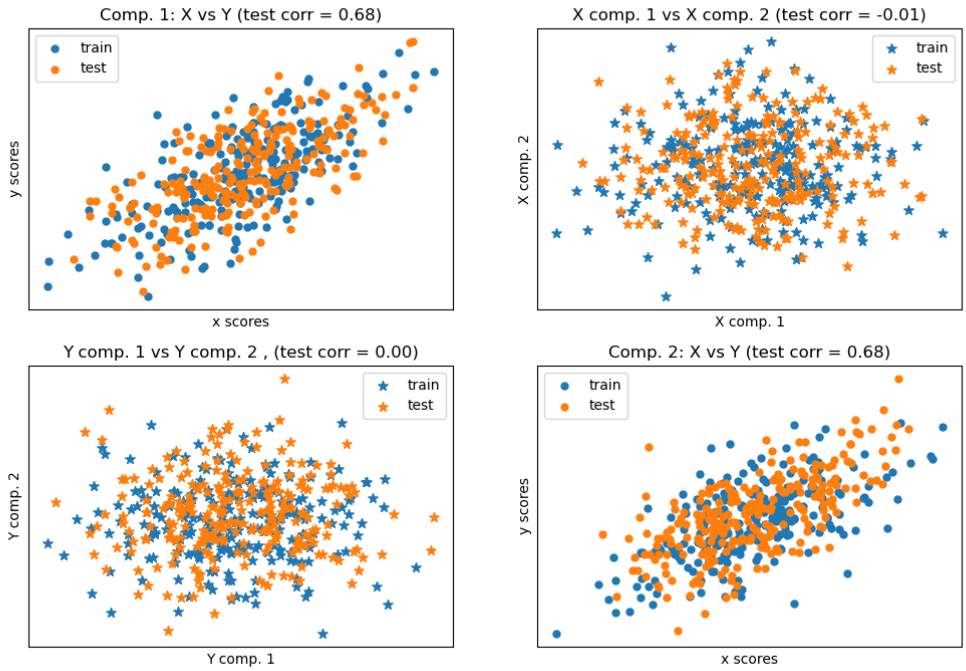

Figure 21: Cross Decomposition

- **Naïve Bayes** (see figure 22)**:** Naïve Bayes methods are a set of supervised learning algorithms that apply Bayes' theorem in a "naïve" mode, assuming conditional independence between each pair of features given the value of the class variable. Bayes' theorem states:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

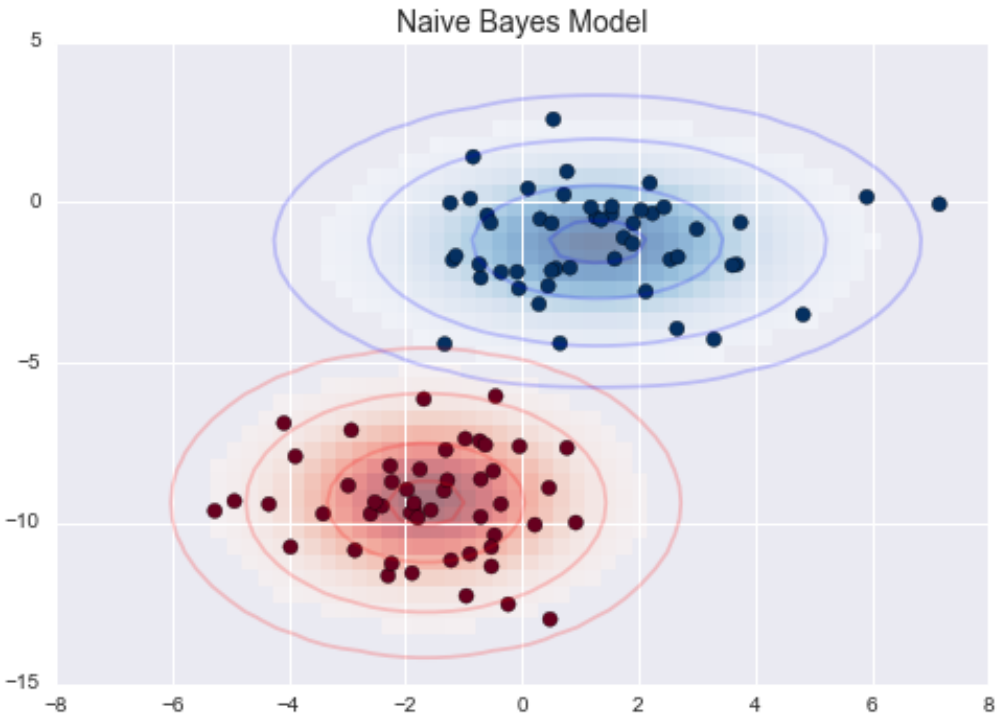

Figure 22: Naive Bayes

- **Trees** (see figure 23)**:**
    - **Extremely Randomized Trees** implements a meta estimator that fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. A random subset of candidate features is used, randomly extracting thresholds for each candidate feature and the best among them is set as the division rule. This usually allows to reduce the variance of the model a bit more, at the expense of a slightly greater increase in bias (Breiman, "Random Forest", 2001)
    - **Decision Trees (DTs)** are a non-parametric supervised learning method used for classification and regression. The DT learns from decision rules, which are identified by studying the characteristics of the data, and then aims to create a model that can accurately predict

the value of the target variable. It can be seen as a constant approximation, and the result is dependent on the depth the tree reaches.

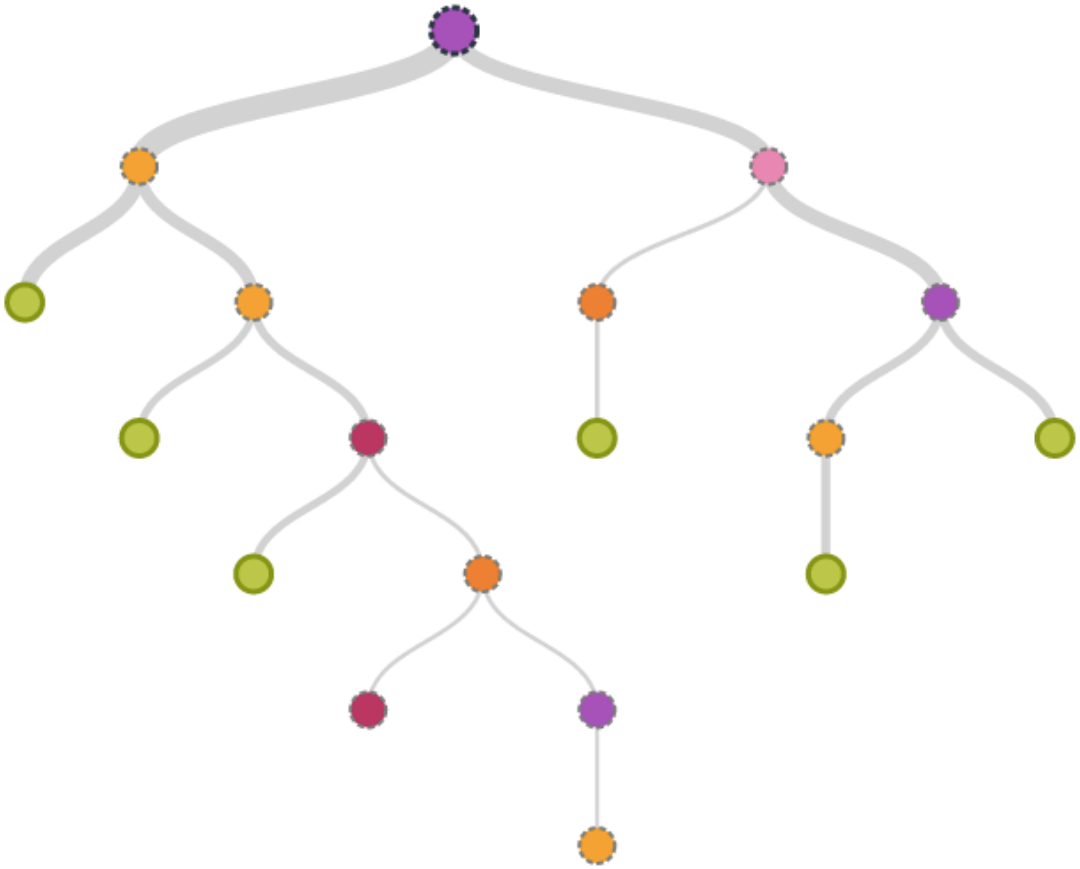

Figure 23: Trees

- **Ensemble** (see figure 24)**:**
  - **Gradient Tree Boosting** or Gradient Boosted Decision Trees (GBDT) is a generalization of boosting to arbitrary differentiable loss functions. GBDT is an accurate and effective off-the-shelf procedure that finds application in both regression and classification problems. (Friedmann, 1999)
  - **XGBoost** is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine

learning algorithms under the Gradient Boosting framework. XGBoost is able to create several gradient tree boosting (GBDT) in parallel, so it is very fast in execution and accurate. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples. (XGBoost)

- **Random Forest** In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features or a random subset of size max_features. The two sources of randomness just described are applied with the aim of decreasing the variance of the forest estimator. This is because decision trees typically show high variance and a tendency to overfit the model. Due to this randomness factor added within the forests, the decision trees produced contain several prediction errors different among them. By averaging these predictions, most of the errors will tend to cancel out among themselves. Random forests achieve reduced variance by combining several trees, sometimes at the cost of a slight increase in bias. In practice, the variance reduction is often significant, resulting in an overall better model. A random forest is a meta-estimator that fits a number of decision trees classifiers on various subsamples of the dataset and uses the average to improve predictive accuracy and control overfitting. (Breiman, "Random Forest",Machine Learning, 2001)

- **ADA Boost's** core principle is to fit a sequence of weak learners on repeatedly modified versions of the data. The forecasts of all of them are then combined using a weighted sum to produce the final prediction. The data modifications at each so-called boosting iteration consist of applying weights to each of the training samples. Initially, those weights are all set, so that the first step simply trains a weak learner on the original data. For each successive iteration, the sample weights are individually modified, and the learning algorithm is reapplied to the reweighted data. At a given step, those training examples that were incorrectly predicted by the boosted model induced at the previous step have their weights increased, whereas the weights are decreased for those that were predicted correctly. As iterations proceed, examples that are difficult to predict receive ever-increasing influence. Each subsequent weak learner is thereby forced to concentrate on the examples that are missed by the previous ones in the sequence. (Drucker, 1997)
- **Voting:** Voting combines several machine learning regressors, combines their results and returns the average between them. Useful for balancing the individual weaknesses of individual algorithms.
- **Stacking:** Stacking is a method of combining estimators to reduce their biases. The predictions of each estimator are then combined together and used as input for the final calculation of the independent variable (y) trained through cross-validation.

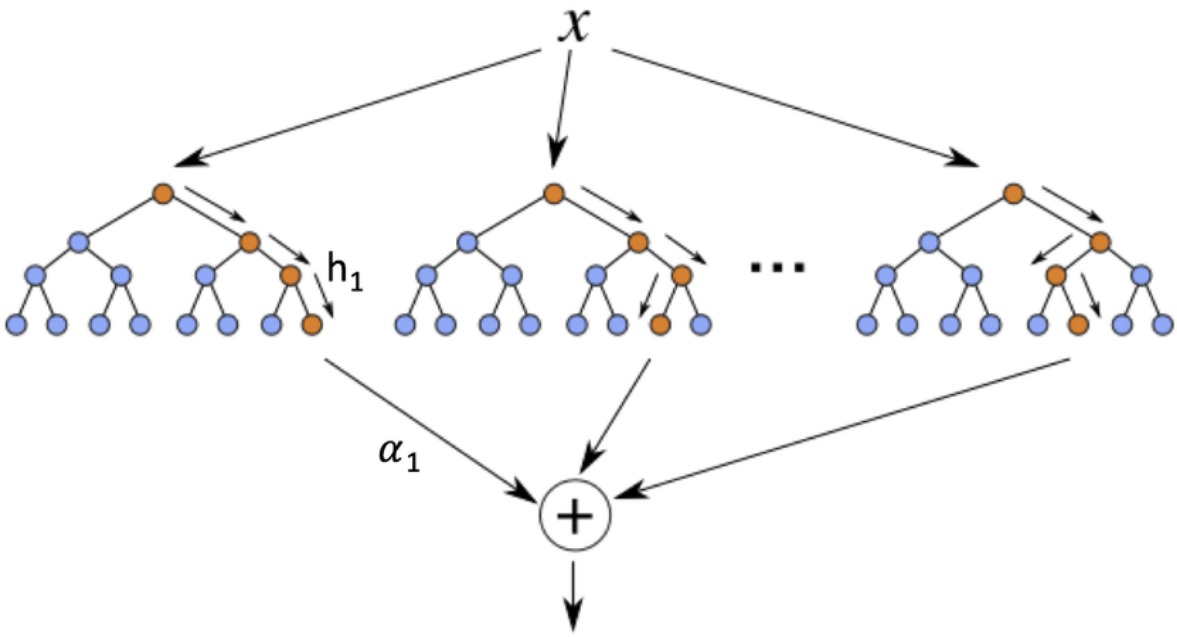

Figure 24: Ensemble

- **Isotonic Regression** (see figure 25)**:** it reduces the function to 1 dimensional data. It solves:

$$\min \sum_i w_i (y_i - \hat{y}_i)^2$$

$$\hat{y}_i \leq \hat{y}_j \; when \; X_i \leq X_j$$

Where, w_i are the weight strictly positive.

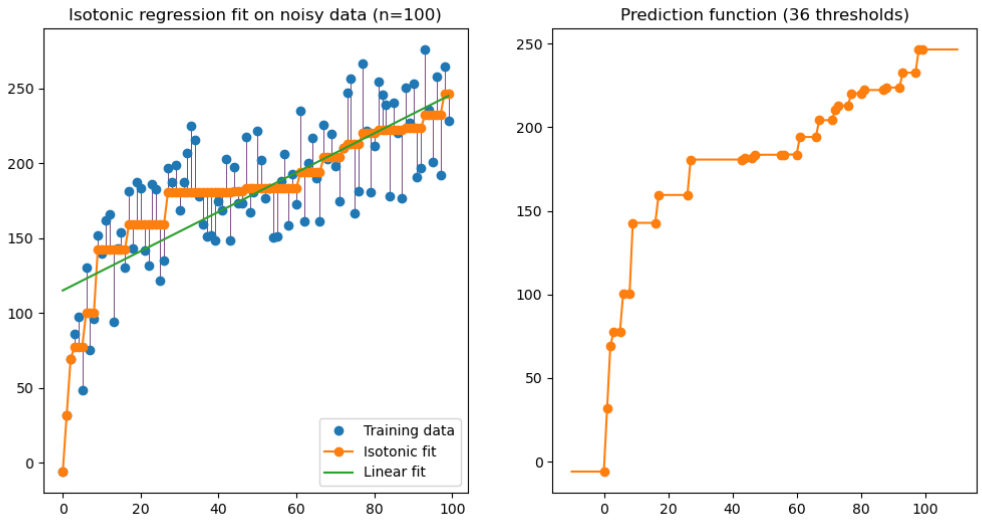

Figure 25: Isotonic Regression

- **Neural Network** (see figure 26)**:** ANN is a system made up of interconnected units that tend to resemble the neurons of a human brain. The units, called neurons, receive an input signal and pass it to the other units connected to it. Within these neurons the input is processed through non-linear functions and finally, after the last unit, an output is released.
    - **Multi-layer Perceptron** trains iteratively since at each time step the partial derivatives of the loss function with respect to the model parameters are computed to update the parameters. Usually, a regularisation term is added to the loss function to prevent and avoid overfitting. (Kingma, Diederik, Jimmy Ba., 2014)

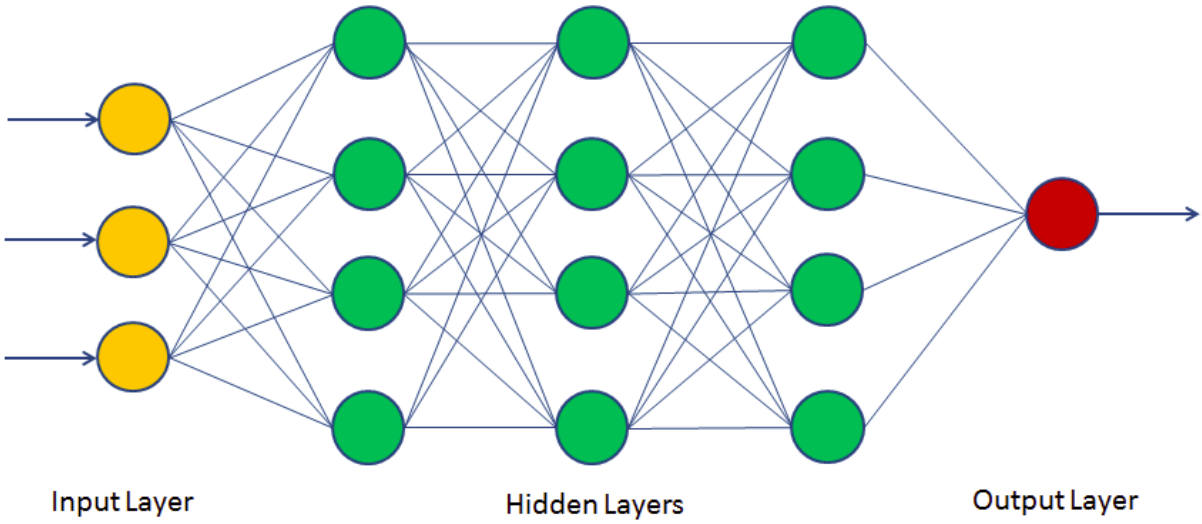

Figure 26: Neural Network

5 Model's Description

After a preliminary study of the dataset and after making the appropriate modifications, the next steps for applying the algorithms are as follows (see Figure 16):

1- Data Loading

2- Data Transformation

3- Machine Learning Algorithm application

4- Metrics computation

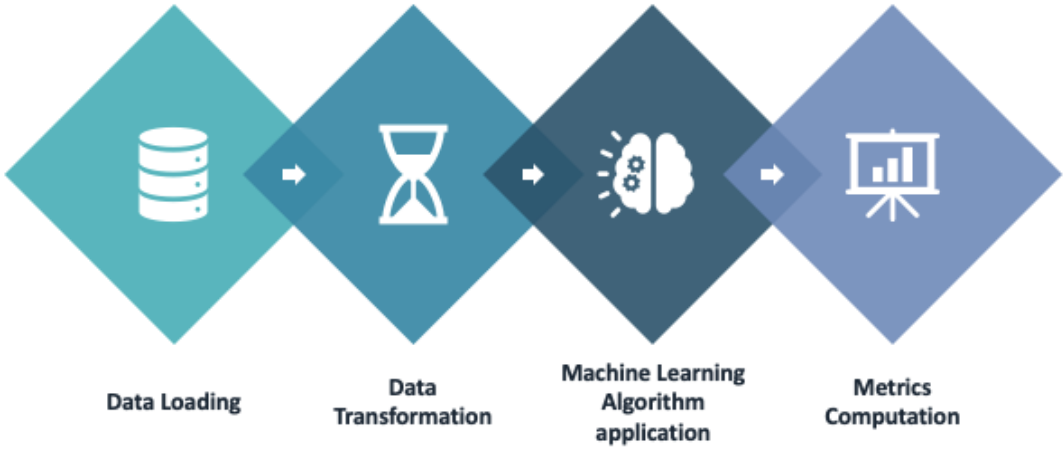

Figure 27: Step to apply the algorithm

5.1 Data Loading

The first step in applying machine learning algorithms to the dataset is the loading of the dataset. After modifying it, by adding the new variables, it was loaded into the environment used for the subsequent development of the model. It was previously loaded into a Google Worksheet, and then imported into the Colab Notebook interface. The variables names were separated from the rest of the column, and both the input variable (X) and the output variable (y) were defined representing the

parameter to use and the final objective to be optimized by the algorithm. More in detail, the dataset that we are going to take into analysis will be "Patient Transport System" regarding the dataset “*dAC.crit*”, while for the dataset “*DB.float*”, I take into account a list of 108 different dataset. This represents the basis for the studies, with both the datasets described in the previous chapter used as input variables and the dAC of each dataset as output variable.

5.2 Data Transformation

After loading the dataset into the development environment, the first step is to transform the data so that the algorithms can be later applied. The first step is to transform all the variables into categorical ones. In our case, this was not necessary as it was a procedure already carried out in the preliminary part of the creation of the dataset, although it is usually advisable to do this at this early stage.

5.2.1 Normalization

Normalization allows the dataset to be read by the algorithm in a uniform way, thus transforming the available raw data into a format that is more understandable by the algorithm. In fact, rather than having values in the range $(-\infty, +\infty)$ normalization reduces this interval to $[0, +1]$. The aim is to eliminate data redundancy in order to avoid anomalies. More specifically, the formula that is applied in order to normalize the data is as follows:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Where:

- z_i represent the normalized value,
- x_i represent the initial value,
- $\min(x)$, $\max(x)$ represent the maximum and minimum values respectively within the column under analysis

This practice is very useful, particularly since the algorithms do not have their own intelligence, and so tend to give more weight to the values they identify as greater. In this way, by placing all values within the same, narrower range, the algorithm will not make any initial distinction and will consider the variables as being at the same level.

5.2.2 Polynomial features

After the normalization, a function called "polynomial features" was created within the program. The aim of this function is to multiply the rows of the dataset in order to find variables with a greater impact in the result. In fact, the first step is to select n, which will correspond to the degree of the polynomial, and then the nth products and the various double products will come out of the rows. For example, if we select $n = 2$, and apply the function, we get the following output for each pair of rows $(a, b) = (1, a, b, ab, a^2, b^2)$. Of course, the term known was removed, as it would have been redundant and unnecessary for the subsequent application of machine learning algorithms. This feature is very important, as it allows the use of the variables taken into consideration not only in linear form, but also in different forms, represented by the value raised to the nth power, and the different combined products between the variables. It is important not to select a large value for the variable n because high degrees can cause overfitting. In statistics, overfitting means that a model is perfectly calibrated to solve the dataset provided as a test, but then turns out to be poor if a different dataset is used. So, in case of overfitting, there are too many parameters in

the model and a high variability. The model is too complex and sensitive to training data. Conversely, there is also underfitting, which occurs when there are too few parameters in the model and a high bias, which causes the learning process to be too simple. So, the ultimate goal of polynomial features is to slightly complicate the algorithm in order to avoid underfitting, but at the same time not overcomplicate it to prevent overfitting. The figure below shows an example of an underfitted model, an optimal model and an overfitted model.

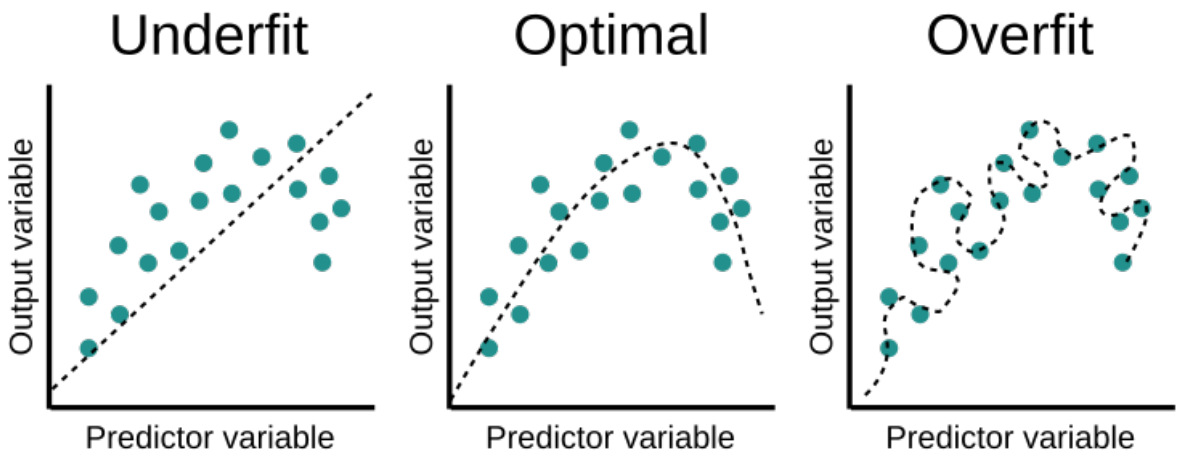

Figure 28 Model underfit or overfit

5.2.3 Select min/max

The next step after the initial data transformation, is the selection of the minimum and maximum values. These values represent the minimum and maximum number of variables that will be considered as significant by the algorithms. This feature was added later because, after having carried out the polynomial features (described in the previous paragraph), it was also necessary to add a column filtering method to avoid redundancy in the model, as there were repeated columns, and also to simplify the model and therefore avoid overfitting.

5.2.4 Split

Finally, the last step of the data transformation is the division between training dataset and testing dataset.

- **Training Dataset:** The training dataset represents the part of the original dataset that is used to train the machine learning model that will be applied. The model will learn from the data that are provided within the training dataset and will map a function F(x) where "x" represents the input variable described above. The result of the function instead will represent the output variable or "y". Through the training dataset, both the X and the y will be supplied to the algorithm so that it can carry out its own training using both the input and the output values. Therefore, the objective of the training dataset will be to being able to train the model in such a way that it will be capable of predicting with the greatest possible accuracy the output.
- **Testing Data:** The testing data, on the other hand, is used to validate the dataset, representing a part of the original dataset used for checking and estimating the accuracy of the model. The testing phase uses only the input variables ("X"), while the output variable ("y") will be used later, comparing it with the result obtained. From this, the performance measures of the model can be deduced.

This function of select training and testing before applying algorithms is necessary, otherwise the model would carry out both phases (training and testing) using the same data, and this could easily cause overfitting (e.g. the model would be fictitious for that data on purpose, and as soon as a new project is added the values would all be wrong).

To apply this, given the complexity of the dataset under analysis, we decided to use 70% of the dataset as the training dataset, and the remaining 30% for the testing part.

5.3 Machine Learning algorithms

After loading and transforming all the available data, we finally moved on to the application of the models. The first study to be carried out is to understand what kind of algorithm can be applied to our case. A first distinction of algorithms is the one between supervised learning and unsupervised learning.

5.3.1 Supervised Learning

Supervised Learning refers to the process of creating a machine learning model, based on already labelled training data (each column refers to a category). In this way, we tell our algorithm what the mapped values correspond to and based on this mapping the algorithm will return an output. Supervised learning algorithms then analyze the training dataset in order to create a function capable of mapping the values corresponding to the output variable. There are six steps in a supervised learning algorithm:

1- Determine the type of training examples. Then understand what type of data will be used within the training dataset.

2- Gather a training set. The training set must represent the function. In fact, both input variables and output variable data are gathered. This data can be collected in different ways, it can be through measurements (quantitative data) or also through experts (qualitative data).

3- Determine the input variables. After collecting the data, it is important to understand which of them will be used as input for the subsequent creation of the function to be applied and which of them will be the desired outcome.
4- Decide which algorithm is the most appropriate. Next it is right to decide which algorithm or set of algorithms is the most appropriate for the measurement that will be performed. It is important to understand the nature of these because within supervised learning there are two major groups of algorithms, classification and regression
5- Complete the design. Run the algorithm or the algorithms using the training dataset defined above.
6- Evaluate the accuracy. The last point is the one that finally allows us to make distinctions between the different algorithms, because it permit to better understand which one is the most suitable for our data set, and to do this we use metrics that therefore allow us to ultimately choose the most suitable one.

As discussed above, the major supervised learning algorithms are classification and regression algorithms.

- **Classification:** Classification is the process of finding a pattern that helps separate data into several categorical classes. The algorithm processes the data and returns an output that belongs to a class. The greatest fields of application are found when the output is binary (0, 1) or when the output is represented by a class. The function therefore allows a mapping into classes. The prediction is then that of discrete values.
- **Regression:** Regression is the process of finding a model or function to distinguish data into continuous real values instead of using classes. Mathematically, the goal of a regression problem is to find the function that

best approximates the model while minimizing the error. In this case, the algorithm creates a function that allows the mapping of values in a continuous way. The prediction is of continuous values.

5.3.2 Unsupervised Learning

Unsupervised learning (UL) consists of a series of algorithms that learn to create models from unlabeled data. The main goal is for the machine to learn through experience and construct its own representation of the data. The UL exhibits self-organization that captures patterns as neuronal predilections or probability densities. Other levels in the spectrum of supervision are reinforcement learning in which the machine is given only a numerical performance score as a guide, and semi-supervised learning in which a smaller portion of the data is labelled. Two major methods in UL are neural networks and probabilistic methods.

- **Neural Networks:** Artificial neural networks are mathematical models composed of several hidden layers that are inspired by the functioning of the human brain. They therefore find application in solving artificial intelligence engineering problems.
- **Probabilistic Methods:** are probability-based design tools.

5.3.3 The choice of the model

The models applied take into account the fact that the desired output in this type of analysis is a continuous value, so classification algorithms were omitted. A first approach was made by means of regression algorithms (linear and non-linear) and then the analysis was carried out by means of unsupervised learning algorithms of both types described in the previous paragraph.

5.4 Metrics computation

After the application of the various algorithms, a method was needed to understand which algorithm provides the best approximation for the model. After several studies, it was decided that the metrics to be taken into account are the R^2 and the R^2_{adj}. The R^2, or coefficient of determination, indicates the percentage of variance in the dependent variable that is explained by the independent variable. It measures how strong the relationship between the model and the dependent variable is on a percentage scale, so, between 0 and 100%. The formula for determining it is as follows:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where:

- RSS is the sum of squares of residuals
- TSS is the total sum of squares.

The R^2_{adj} is a modified version of R^2, adjusted for the number of predictors in the model. Its formula is as follows:

$$R^2_{adj} = 1 - \frac{(1 - R^2) \times (N - 1)}{N - p - 1}$$

Where:

- R^2 is the sample R-square
- N is the total sample size
- p is the number of predictors

Other parameters were also taken into account in the analysis which allowed for greater model accuracy:

- MAE: Mean Absolute Error, the formula is:

$$MAE = \frac{\sum_1^N |y_i - x_i|}{N}$$

- RMSE: Root Mean Square Error, the formula is:

$$\sqrt{\frac{\sum_1^N (x_i - \widehat{x_i})}{N}}$$

- Time: the time it takes for the model to run
- Number of parameters represents the number of parameters used for the model.

6 Results interpretations

6.1 Results without new variables

All that has been described up to this point has been applied to the "*DB.float*" dataset, a set of 108 completed project, containing different TPs within which some rescheduling has been performed, as can be seen by observing it. So, it was a huge dataset that contain a lot of variables. Fortunately, it was possible in this paper to propose a statistical validation of the model with the respective comparison between the various models used. Therefore, it has been verified the goodness of the algorithms applied in quality of described metrics, $R^2 and\ R^2_{adj}$, in order to carry out the estimate of the Estimate at Completion.

Table 4: Summary of results for DB.float

Model	R^2	R^2_adj	Variable Important
Ordinary Least Square	0,4078	0,4057	BAC*%WS; %WS*PV; %WS*dPV
Ridge	0,3806	0,3785	dPV; %WS*dPV; d%WS*dPV
Lasso	0,4003	0,3982	BAC*%WS; %WS*PV; %WS*dPV
Elastic Net	0,3995	0,3974	BAC*%WS; %WS*PV; %WS*dPV
LARS	0,4078	0,4057	BAC*%WS; %WS*PV; %WS*dPV
LARS Lasso	0,4078	0,4057	BAC*%WS; %WS*PV; %WS*dPV
Bayesian Ridge	0,4071	0,405	BAC*%WS; %WS*PV; %WS*dPV
Automatic Relevance Determination	0,407	0,405	BAC*%WS; %WS*PV; %WS*dPV
Stochastic Gradient Descent	0,0026	0,0018	%WS; %WS^2
Passive Aggressive Regressor	0,3754	0,4351	dPV; d%WS^2; d%WS*dPV
Random Sample Consensus	0,013	0,3624	BAC*%WS; %WS*PV; d%WS^2
Theil Sen	0,3592	0,3895	d%WS; PV; d%WS*PV
Kernel Ridge	0,3958	0,3937	dPV; BAC*%WS; d%WS*dPV
Support Vector Regression	-0,0762	-0,0774	BAC*%WS
Nu Support Vector Machine	-0,0475	-0,0487	BAC
Linear Support Vector Machine	-2,517	0,4421	dPV; %WS*dPV; d%WS*dPV
Knearest Neighbors	0,6074	0,606	dPV; %WS^2; %WS*PV
Gaussian Process	0,2458	0,245	BAC
Decision Tree	0,7275	0,7265	BAC*%WS; BAC*dPV; %WS*d%WS

Extremely Randomized Trees	0,7256	0,7247	BAC; %WS^2;PV*dPV
Random Forest	0,7533	0,7524	BAC*PV; BAC*dPV; %WS*d%WS
ADA Boost	0,7091	0,7081	%WS*d%WS;d%WS^2; PV*dPV
Gradient Boosting	0,7176	0,7176	BAC*PV; %WS*dPV; PV*dPV
XGB	0,7254	0,7244	BAC*PV; %WS*dPV; PV*dPV
Multi-layer Perceptron	0,5027	0,5009	dPV; BAC*%WS; %WS*PV

As it can be seen from the summary table (see table 4), the application of the algorithms in $DB.float$, containing 108 different projects, is satisfactory and with prospects for improvement, with R^2 and R^2_adj reaching 0.7533 and 0.7524 respectively when applying the Random Forest algorithm. While the black box algorithms average 0.7 for both metrics considered, the white boxes average 0.4. In general, it can be seen that the results of the black box algorithms are more significant than those of the white box algorithms. As far as the significant variables are concerned, it can be immediately noted that those that most influence the calculation of the dAC (delta Actual Costs) are:

- BAC
- %WS
- d%WS
- PV
- dPV

For a better application, the new variables defined in the previous chapter could be introduced.

For the complete table see Annex.

6.2 Results with new variables

The same algorithms were also applied to the "Patient Transport System" dataset called $dAC.crit$, a completed project, containing 23 TPs within which some rescheduling has been performed, as can be seen by observing it. Also in this case it has been verified the goodness of the algorithms applied in quality of described metrics, $R^2 and\ R^2_{adj}$, in order to carry out the evaluation of the Estimate at Completion. In order to have a more accurate estimation it would have been necessary to have a dataset composed of many projects, similar to $DB.float$, but unfortunately it was complex because the projects available needed a manual calculation of all the variables subsequently introduced through the studies performed. In spite of this, it is possible to affirm that the introduction of new variables in this dataset proved to be very useful at the time of writing, allowing the statistical validation of the project. Below, we can see how the different algorithms have performed with respect to the dataset (see table 5).

Table 5:Summary of Results for dAC.crit

Model	R^2	R^2_adj	Variable Important
Ordinary Least Square	0,9889	0,9861	PI1a^2; PI2a*PI2b; PI2b^2
Ridge	0,9863	0,9828	%WS^2; PI2a*PI2b; PI2b^2
Lasso	0,9012	0,8764	%WS*PI2a; PI1b^2; PI2a^2
Elastic Net	0,884	0,855	%WS^2; PI2a^2; PI2b^2
LARS	0,991	0,9888	PI2a*PI2b; PI2b^2; PI2b*EMV
LARS Lasso	0,991	0,9888	PI2a*PI2b; PI2b^2; PI2b*EMV
Bayesian Ridge	-0,0704	-0,1469	BAC*%WS
Automatic Relevance Determination	-0,0704	-0,1469	BAC*%WS
Stochastic Gradient Descent	-8,2562	0,988	%WS*PI2b; PI1a*PI2b; PI1b*PI2b
Passive Aggressive Regressor	-0,0471	0,9894	%WS; PI2b*EMV
Random Sample Consensus	0,9977	0,9971	EMV; BAC*PI1b; PI1b*EMV
Theil Sen	0,9942	0,9928	%WS; PI2b^2; PI2b*EMV

Kernel Ridge	0,9875	0,9844	PI1b; PI2a*PI2b; PI2b^2
Support Vector Regression	-0,145	-0,2267	PI1a*PI1b
Nu Support Vector Machine	-0,1258	-0,2063	PI2a^2
Linear Support Vector Machine	0,9955	0,9914	PI1a; %WS*EMV; PI2b*EVM
Knearest Neighbors	0,0193	-0,0508	BAC*%WS
Gaussian Process	-0,0487	-0,1236	%WS
Decision Tree	-0,0746	-0,1513	BAC^2
Extremely Randomized Trees	0,0918	-0,0479	BAC*%WS; %WS^2
Random Forest	-0,0746	-0,1513	BAC^2
ADA Boost	-0,0831	-0,1605	PI1a^2
Gradient Boosting	-0,0746	-0,1513	BAC^2
XGB	-0,0746	-0,1513	BAC^2
Multi-layer Perceptron	0,9953	0,9941	PI1a*EMV; PI1b*BAC; PI2a*PI2b

As it can be seen from the summary table (see Table 5), the application of the algorithms in dAC.crit, containing the Patient Transport system project, is very significant and with prospects for improvement related to the size of the dataset. The values of R^2 and R^2_{adj} reach 0.9977 and 0.9971 respectively when applying the Random Sample Consensus algorithm.

In general, we can say that the white box algorithms perform better than the black box ones, probably due to the few columns contained within the dataset. The only black box algorithm that performs optimally is the neural network. As regards the significant variables are concerned, it is immediately noticeable that the new variables are highly significant within the dataset. For a better application, the new variables could also be introduced within the remaining projects and the algorithms reapplied.

For the complete table see Annex.

7 Conclusions

Projects are a constant in the life of every company and as such need to be managed in the best possible way. Incurring in "overcost" and "out of control", as well as leading to economic damage, can in fact also cause problems to the image of the company. For this reason, being able to accurately estimate time and costs at the end of a project in progress with great precision proves to be a very important tool in the activities of a project manager. As seen in this paper, in literature there are several methodologies and schools of thought for the calculation of EaC, but only a minimal fraction of them partially takes into account the importance of risks within the estimation of time and costs at the end of a project. Risks, on the other hand, are a fundamental aspect of any project and are often one of the main causes of time and cost variations within it. In addition, we can also consider that only the latest studies take into account historical data, which instead be crucial to the success of a project. There are a number of risk-related variables in the project environment that are not monitored, some of which are analyzed in this paper. If they were all considered and monitored, they would be very useful to predict fluctuations or at least an estimate of the variance of costs, through Machine Learning algorithms. The impossibility of calculating the new variables on a sufficient number of projects unfortunately made the conclusions, even if valid, not entirely reliable, since a more intensive study would be needed to statistically validate the model. But the first study carried out on 108 projects shows us the reliability and the possibility of improvement in the application of the ML algorithms. Further studies on the subject should therefore continue first of all by introducing a greater number of datasets within it, and then, if it is deemed necessary, by focusing on the definition of new variables that could be

more significant than those defined in this paper. The objective of this paper turns out to be satisfied, since the applied models seem to give a good approximation.

Annex

Table 6: Full Table DB.float

Model	R^2	R^2_adj	MSE	RMSE	Explained Variance	Bias	MAE	Max Error	Variable Important
Ordinary Least Square	0,4078	0,4057	3,89*e+11	6,24*e+5	0,4147	67263	204853	8,86*e+6	BAC*%WS; %WS*PV; %WS*dPV
Ridge	0,3806	0,3785	4,07*e+11	6,38*e+5	0,4096	138043	189625	8,11*e+6	dPV; %WS*dPV; d%WS*dPV
Lasso	0,4003	0,3982	3,94*e+11	6,28*e+5	0,4079	70606	205609	8,94*e+6	BAC*%WS; %WS*PV; %WS*dPV
Elastic Net	0,3995	0,3974	3,95*e+11	6,28*e+5	0,4071	70851	205791	8,95*e+6	BAC*%WS; %WS*PV; %WS*dPV
LARS	0,4078	0,4057	3,89*e+11	6,24*e+5	0,4147	67263	204853	8,86*e+6	BAC*%WS; %WS*PV; %WS*dPV
LARS Lasso	0,4078	0,4057	3,89*e+11	6,24*e+5	0,4147	67263	204853	8,86*e+6	BAC*%WS; %WS*PV; %WS*dPV
Bayesian Ridge	0,4071	0,405	3,90*e+11	6,24*e+5	0,414	67537	204896	8,86*e+6	BAC*%WS; %WS*PV; %WS*dPV
Automatic Relevance Determination	0,407	0,405	3,90*e+11	6,24*e+5	0,414	67718	204814	8,87*e+6	BAC*%WS; %WS*PV; %WS*dPV
Stochastic Gradient Descent	0,0026	0,0018	6,54*e+11	8,09*e+5	0,0074	41432	304570	1,05*e+7	%WS; %WS^2
Passive Aggressive Regressor	0,3754	0,4351	1,78*e+11	4,22*e+6	-25,7538	458208	500290	7,87*e+7	dPV; d%WS^2; d%WS*dPV
Random Sample Consensus	0,013	0,3624	7,07*e+11	8,41*e+5	0,0077	233861	245685	1,08*e+7	BAC*%WS; %WS*PV; d%WS^2
Theil Sen	0,3592	0,3895	5,33*e+11	7,30*e+5	0,1907	35910	297206	7,11*e+6	d%WS; PV; d%WS*PV
Kernel Ridge	0,3958	0,3937	3,97*e+11	6,30*e+5	0,4031	69684	194523	8,81*e+6	dPV; BAC*%WS; d%WS*dPV
Support Vector Regression	-0,0762	-0,0774	7,07*e+11	8,41*e+5	0	223806	246397	1,07*e+7	BAC*%WS
Nu Support Vector Machine	-0,0475	-0,0487	6,88*e+11	8,30*e+5	0	176654	249593	1,06*e+7	BAC
Linear Support Vector Machine	-2,517	0,4421	6,36*e+11	7,97*e+5	0,0908	195782	232049	1,13*e+7	dPV; %WS*dPV; d%WS*dPV
Knearest Neighbors	0,6074	0,606	2,58*e+11	5,08*e+5	0,6084	25779	158480	6,15*e+6	dPV; %WS^2; %WS*PV
Gaussian Process	0,2458	0,245	4,95*e+11	7,04*e+5	0,2459	5146	234887	0,29*e+6	BAC
Decision Tree	0,7275	0,7265	1,79*e+11	4,23*e+5	0,7275	3654	186214	2,91*e+6	BAC*%WS; BAC*dPV; %WS*d%WS
Extremely Randomized Trees	0,7256	0,7247	1,80*e+11	4,25*e+5	0,7265	24468	169737	4,54*e+6	BAC; %WS^2;PV*dPV
Random Forest	0,7533	0,7524	1,62*e+11	4,03*e+5	0,7534	8239	176749	3,71*e+6	BAC*PV; BAC*dPV; %WS*d%WS
ADA Boost	0,7091	0,7081	1,91*e+11	4,37*e+5	0,7206	86885	249784	4,78*e+6	%WS*d%WS;d%WS^2; PV*dPV
Gradient Boosting	0,7176	0,7176	1,85*e+11	4,30*e+5	0,7186	1383	172955	5,09*e+6	BAC*PV; %WS*dPV; PV*dPV
XGB	0,7254	0,7244	1,80*e+11	4,25*e+5	0,7255	8622	173035	5,05*e+6	BAC*PV; %WS*dPV; PV*dPV
Multi-layer Perceptron	0,5027	0,5009	3,27*e+11	5,72*e+5	0,5039	28927	182688	8,00*e+6	dPV; BAC*%WS; %WS*PV

Table 7: Full Table dAC.crit

Model	R^2	R^2_adj	MSE	RMSE	xlained Varia	Bias	MAE	Max Error	Variable Important
Ordinary Least Square	0,9889	0,9861	1,19*e+7	3451	0,9889	239	2238	7826	PI1a^2; PI2a*PI2b; PI2b^2
Ridge	0,9863	0,9828	1,47*e+7	3834	0,9871	967	2673	8524	%WS^2; PI2a*PI2b; PI2b^2
Lasso	0,9012	0,8764	1,06*e+8	10289	0,9087	2844	6082	24328	%WS*PI2a; PI1b^2; PI2a^2
Elastic Net	0,884	0,855	1,24*e+8	11147	0,886	1487	6869	22318	%WS^2; PI2a^2; PI2b^2
LARS	0,991	0,9888	9,61*e+6	3100	0,9913	589	2304	6003	PI2a*PI2b; PI2b^2; PI2b*EMV
LARS Lasso	0,991	0,9888	9,61*e+6	3100	0,9913	589	2304	6003	PI2a*PI2b; PI2b^2; PI2b*EMV
Bayesian Ridge	-0,0704	-0,1469	1,15*e+9	33860	0,0176	9711	16006	88775	BAC*%WS
Automatic Relevance Determination	-0,0704	-0,1469	1,15*e+9	33860	0,0176	9711	16006	88775	BAC*%WS
Stochastic Gradient Descent	-8,2562	0,988	3,88*e+10	197101	-29,6262	77742	77742	521220	%WS*PI2b; PI1a*PI2b; PI1b*PI2b
Passive Aggressive Regressor	-0,0471	0,9894	6,29*e+9	79322	-3,9872	30823	32745	209415	%WS; PI2b*EMV
Random Sample Consensus	0,9977	0,9971	1,25*e+9	35356	-0,0102	12960	14696	93450	EMV; BAC*PI1b; PI1b*EMV
Theil Sen	0,9942	0,9928	6,16*e+6	2482	0,9974	1845	2091	4241	%WS; PI2b^2; PI2b*EMV
Kernel Ridge	0,9875	0,9844	1,34*e+7	3658	0,9875	34	2402	8528	PI1b; PI2a*PI2b; PI2b^2
Support Vector Regression	-0,145	-0,2267	1,23*e+9	35020	0	12463	14676	92556	PI1a*PI1b
Nu Support Vector Machine	-0,1258	-0,2063	1,21*e+9	34726	0	11610	14798	91705	PI2a^2
Linear Support Vector Machine	0,9955	0,9914	5,48*e+8	23400	0,5611	8803	10759	61333	PI1a; %WS*EMV; PI2b*EVM
Knearest Neighbors	0,0193	-0,0508	1,05*e+9	32411	0,0657	7049	17411	83655	BAC*%WS
Gaussian Process	-0,0487	-0,1236	1,12*e+9	33514	0,0466	10100	15732	88014	%WS
Decision Tree	-0,0746	-0,1513	1,15*e+9	33926	0	8937	16500	89033	BAC^2
Extremely Randomized Trees	0,0918	-0,0479	9,73*e+8	31190	0,1545	8193	15595	81039	BAC*%WS; %WS^2
Random Forest	-0,0746	-0,1513	1,15*e+9	33926	0	8937	16500	89033	BAC^2
ADA Boost	-0,0831	-0,1605	1,16*e+9	34061	0,0599	12378	14266	90025	PI1a^2
Gradient Boosting	-0,0746	-0,1513	1,15*e+9	33926	0	8937	16500	89033	BAC^2
XGB	-0,0746	-0,1513	1,15*e+9	33926	0	8938	16499	89033	BAC^2
Multi-layer Perceptron	0,9953	0,9941	5,06*e+6	2249	0,9965	1138	1742	3450	PI1a*EMV; PI1b*BAC; PI2a*PI2b

Bibliography

Breiman. (2001). *"Random Forest",Machine Learning.*

Friedmann. (1999). *"Stochastic Gradient Boosting".*

XGBoost. (n.d.). Retrieved from https://xgboost.readthedocs.io/en/latest/

(Vanhoucke, 2.-2. (n.d.).

Vanhoucke. (2010-2011).

Rezouki S.E., M. S. (2020).

Balali A., V. A. (2020).

Chang H., Y. W. (2019).

Anbari. (2003).

Christensen, Heise. (1993).

Lopez-Paredes, Pajares. (2011).

De Marco, Narbaev. (2014).

Fleming, Koppelman. (2006).

Kim, Reinschimdt. (2011).

Fleming, Koppelman. (2006).

Lipke. (2003-2004).

De Marco, Jamaluddin Thaheem. (2014).

Balali A., Valipour A., Antucheviciene J., Šaparauskas, J. (2020).

Rezouki S.E., M. S. (2020).

Drucker. (1997). *Improving Regressor using Boosting Techniques.*

Breiman. (2001). *"Random Forest".*

Kingma, Diederik, Jimmy Ba. (2014). *"Adam: A method for stochastic optimization".*

XGBoost. (n.d.).

Murphy. (n.d.).

Murphy. (2012). *Machine Learning: A probabilistic perspective.*

Chang, Ling. (n.d.). *LIBSVM: A Library for Support Vector Machines.*

Scikit-Learn. (n.d.). *KNN*. Retrieved from https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-regression

scikit-learn. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/gaussian_process.html

Kriengsak Panuwatwanich, Chien-Ho Ko. (n.d.). The 10th International Conference on Engineering, Project, and Production Management.

https://en.wikipedia.org/wiki/Linear_regression. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Linear_regression

Figure and Tables