ISCTE IUL
Instituto Universitário de Lisboa

Department of Information Science and Technology

Explaining Portuguese's Public Administration Absenteeism Through Data Mining

Leandro Miguel Bartolomeu da Cruz Costa

A Dissertation presented in partial fulfillment of the Requirements for the Degree of

Master in Computer Science and Business Management

Supervisor: Sérgio Moro, PhD, Assistant Professor, Instituto Universitário de Lisboa (ISCTE-IUL)

Co-supervisor: Ricardo Ramos, MSc, Researcher at ISTAR-IUL, Instituto Universitário de Lisboa (ISCTE-IUL)

August, 2018

ISCTE IUL
Instituto Universitário de Lisboa

Department of Information Science and Technology

EXPLAINING PORTUGUESE'S PUBLIC ADMINISTRATION ABSENTEEISM THROUGH DATA MINING

Leandro Miguel Bartolomeu da Cruz Costa

A Dissertation presented in partial fulfillment of the Requirements for the Degree of

Master in Computer Science and Business Management

Supervisor: Sérgio Moro, PhD, Assistant Professor, Instituto Universitário de Lisboa (ISCTE-IUL)

Co-supervisor: Ricardo Ramos, MSc, Researcher at ISTAR-IUL, Instituto Universitário de Lisboa (ISCTE-IUL)

August, 2018

ISCTE IUL
Instituto Universitário de Lisboa

Explaining Portugueses̉ public administration absenteeism through data mining

Leandro Miguel Bartolomeu da Cruz Costa

August, 2018

Abstract

Portuguese Public Administration (PPA) is the largest contractor in the country, with 12.8% of the Portugal's active people working for it. Absenteeism and productivity are mutually connected. Thus, companies from public and private sector should always have it in mind, to prevent flaws in the processes and profit loss.

Effectively, the main goal of this study is to understand PPA's absenteeism, particularly the duration of the worker's next absence, what leads to it, as well as explaining it, by creating a data mining model that fits the problem.

To study PPA's absenteeism it was collected data from a Human Capital Management (HCM) system, by extracting the annual absenteeism report, for 2016, and queries to the worker's profile, absenteeism history and job characteristics, resulting in around 59,000 different absence records.

Data mining techniques were used to clean the dataset and Recency, Frequency and Monetary (RFM) value methodology to add new variables to the problematic, originating richer information about the worker and the absence itself.

Thereafter, the Support Vector Machines (SVM) algorithm was applied for modeling the absence duration in day and a 10-fold cross-validation scheme was adopted to assess and confirm the model's robustness.

Finally, major findings were revealed by this study as features related to the worker's profile are less relevant than absence related features; the influence of the RFM methodology in this study, which managed to get all its computed variables in the 25th most important features; and the discovery of the most concerning employee profile.

Keywords: Absenteeism; Portugal Public Administration; data mining; RFM

Resumo

A Administração Pública Portuguesa (APP) é o maior contratante do país, englobando 12.8% da população ativa. O absentismo e a produtividade estão mutuamente ligados, logo tanto as empresas dos vários setores devem tê-las em atenção para prevenir falhas nos processos e perda de lucro.

Efetivamente, o principal propósito deste estudo é perceber o absentismo na APP, em especial a duração da próxima ausência de um trabalhador, as suas causas e explicá-la, através da criação de um modelo adequado ao problema.

Para modelar o absentismo na APP recolheram-se dados de um sistema de gestão de recursos humanos, extraindo o relatório anual de absentismo, para 2016, e dados do perfil do trabalhador, histórico de absentismo e especificações do contrato, resultando em cerca de 59,000 ausências.

Por sua vez, foram usadas técnicas de *data mining* para limpar o conjunto de dados e a metodologia *Recency, Frequency and Monetary value* (RFM) para adicionar novas variáveis à problemática e obter mais perspetivas sobre o trabalhador e a ausência.

De seguida, foi aplicado o algoritmo *Support Vector Machines* (SVM) para modelar a duração da ausência em dias e um esquema de validação cruzada com *10 folds*, que testou e aprovou a robustez do modelo.

Por fim, este estudo revelou várias descobertas como: variáveis relacionadas com o perfil do trabalhador são menos relevantes que as relacionadas com a ausência em si; a influência da metodologia RFM neste estudo, que conseguiu ter todas as suas variáveis nas mais importantes; e a descoberta do perfil do trabalhador mais preocupante.

Palavras-chave: Absentismo; Administração Pública Portuguesa; *data mining*; RFM

Acknowledgements

I am thankful to my supervisors, whose continuous support, concern and motivation have allowed this study to be carried out, despite the most difficult moments, due to the lack of time and the very specificities of the data. I learned a lot from this project, in addition to the deepening of my knowledge in this area, the ability to manage my time well, between work and the thesis, was key to achieve the final version of this study, but also with them, like the importance of the literature review and the ability to state our ideas in the most perceivable way for the greatest number of people, changed the way I used to think and expressed myself.

The institution should also be thanked to, because of its role in making available very good and professional supervisors and, also, the resources I needed to make the best thesis I could.

I am also grateful to my coworkers for sharing their ideas with me, which improved the quality of this study exponentially, and I am deeply thankful to my project managers that provide me access to the data, without it, this study would not be possible.

Last but not least, I want to thank my family and friends, which encouraged and motivated me in the most difficult times. Without them, it would have been way harder and as most students that work at the same time they are doing their thesis, I would have delayed the delivery or even given up.

Thank you all.

Table of contents

Index of figures

Index of tables

List of abbreviations

CE	Contract of Employment
CID	Contract of indefinite duration
CRISP-DM	CRoss-Industry Standard Process for Data Mining
DSA	Data-based Sensitivity Analysis
EU	European Union
ERP	Enterprise Resource Planning
GLPAE	General Law for Public Administration Employment
GM	General Mobility
HCM	Human Capital Management
HR	Human Resources
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
OECD	Organization for Economic Co-operation and Development
PA	Public Administration
PPA	Portuguese Public Administration
RFM	Recency, frequency and monetary value
SA	Sensitivity analysis
SM	Special Mobility
SRM	Structural Risk Minimization
SVM	Support vector machines
TEC	Term employment contract

1. Introduction

1.1 Portuguese Public Administration

Portugal is an EU member since 1986, where around 4.8 million Portuguese males and 5.4 million Portuguese females live in (PORDATA, 2017a), from which 6.5% works in the public administration, resulting in 12.8% of the active people and 14.0% of employed workers, by the end of the second trimester of 2017 (DGAEP - Direção-Geral da Administração e do Emprego Público, 2017b), close to the 18.1% general government employment as a percentage of total employment across OECD countries, in 2015 (OECD, 2017). Effectively, at 30th of June of 2017, the public employment was distributed as: 76.2% of the workers were working in the central administration (e.g., Tax Authority, Social Security, Police); 16.7% of the workers were working in the local administration (e.g., City councils, Hospitals, Trains); 5.6% of the workers were working in the autonomous regions' administrations (e.g., Regional Governments) (DGAEP - Direção-Geral da Administração e do Emprego Público, 2017a).

As it is written in the Portuguese Constitution's 266th article, the Public Administration shall seek to pursue the public interest, with respect for all those citizens' rights and interests that are protected by law. Administrative organs and agents are subject to the Constitution and the law, and in the exercise of their functions must act with respect for the principles of equality, proportionality, justice, impartiality and good faith (Portuguesa, 2005).

Although in 2008 Portugal got caught in the global financial crisis that led to years of austerity where public debt, unemployment and taxes were the predominant subject of the daily news. Unemployment rates reached historical levels, 16.2% in 2013 (PORDATA, 2017b), deficit high as never seen before, around 20,000 million euros in 2010 (PORDATA, 2017c) and the constant raise of income from taxes, 21.6% of Portugal's GDP in 2015 (PORDATA, 2017d) pictured Portugal for the last years.

However, Portugal is turning around as the results of 2017 show, the GDP per capita had a positive variation of 4.3% over the last year (AICEP, 2018), and the public debt dropped to 125.7% of the GDP, from 129.9% registered in 2016 (AICEP, 2018). In order to extend the prosperity of the country, this study will cover the case of PPA's absenteeism, as explained in the next section.

1.2 Motivation and contribution

As stated in Perry's (2010) article, it was found that those who wanted to work for government attached more importance to the social significance of a job as well as to job characteristics related to quality of life. In addition, PPA is the largest Portuguese contractor when talking about employees, including a total of 669.331 workers in 31st December of 2017 (DGAEP - Direção-Geral da Administração e do Emprego Público, 2018), with different personalities and characteristics.

Thus, it is important to be able to find patterns to comprehend why a PPA's employee do not show up to work, as well as drawing a profile of the most absent worker, in order to provide the human resources department substantiated information so that they can accomplish the efficiency desired.

There is research about human resources management (Armstrong, 2014), the absenteeism (Schaufeli, Bakker, & van Rhenen, 2009) and even the Portuguese public sector (Carvalho & Bruckmann, 2014), as well as, data mining (Wu, Zhu, Wu, & Ding, 2014) and its application in various areas (Moro, Rita, & Vala, 2016); however, there is a lack of studies about predicting the absenteeism in the PPA, while comparing the ideas supported by human resources studies' authors with the results obtained. Consequently, this research aims to fill in this gap as well as to provide public organizations some indicators that might identify a more likely longer absent worker based on his/her profile.

Effectively, through this study, understanding PPA's absenteeism, what leads to it, as well as explaining it, by creating a data mining model that fits the problem, is the main objective and what determines if it a was successful dissertation or not.

This study focuses on modeling the number of consecutive days a worker is absent, thus aiming to predict it based on the motive of the absence, the worker's absenteeism history, its profile and its job specificities. Through data mining, models are tested to best fit the data collected, around 59,000 records of absences.

The second section of this study is focused on the theoretical background that supports all the assumptions and comparisons of this research. Data mining models as well absenteeism causes and consequences are explained. After that, the spotlight is on the collected dataset and its attributes along with the obtained results. To conclude, a comparison between the results found and absences causes identified in the theoretical background is made, to provide evidences for a better decision making.

1.3 Review method

To determine the quality of references and state of the art concerning the object of study, the protocol for the systematic review developed by Dybå & Dingsøyr’s (2008) was undertaken. The criteria for considering studies for review, were as follows:

- Types of studies: Articles to be included in the review must present empirical data and pass the minimum quality threshold described below.
- Types of participants: Studies of both students and recognized figures on the scientific community will be included.
- Types of intervention: Inclusion of studies will not be restricted to any specific type of intervention.

The search strategy for identification of studies will include articles published on the primary journals, primarily preference for Scopus Q1 journals, followed by Q2, Q3 and Q4, in this order, and textbooks, from the last five years, except for very specific question, which might imply then need to go back some years in order to get a valid insight for the study.

Scopus was selected to find relevant literature, since it is generally accepted as bibliographic database to find indexed publications, and queries were applied to conduct and narrow the search. The queries included the Boolean expression AND implying that any article must contain all the expressions used. If the outcome revealed few results, the Boolean expression OR will substitute ANDs to increase the spectrum of search. The OR expression implies that any article should contain at least one of the expressions. The terms used to conduct this search were absenteeism, public sector, data mining, stress, recency frequency monetary value and public administration. Some variations were used to find a satisfactory number of relevant articles needed for this study.

It was also followed some conditions to cover the three main quality issues (rigor, credibility and relevance) that need to be considered when appraising the studies of this research (Dybå & Dingsøyr, 2008). Each article was assessed according to whether:

i. the aims and objectives are clearly reported;
ii. there is an adequate description of the context in which the research was carried out;
iii. there is an adequate description of the sample used and the methods for how the sample was identified and recruited;
iv. appropriate data collection methods were used and described;

v. the study provides clearly stated findings with credible results and justified conclusions;

vi. the study provides value for research or practice.

Together, these criteria should provide a measure to extend the confidence on a specific study's findings as a valuable contribution to this review.

2. Theoretical background

Following the main objectives, theoretical background encompasses themes such as human resources and absenteeism, public services, where it will be specified the reforms that Portugal has been through, as well as some public labor characteristics, and data mining literature review, applied to the human resources and to the public sector.

2.1 Human resources and absenteeism

Employee absences are both costly and disruptive for business, and the trend has been increasing steadily over the years. Personal illness and family issues are cited as the primary reason for unplanned absences. Illness, family responsibilities, personal issues and stress all take a toll on the worker which in turn affects morale, absences and productivity in the workplace (Kocakülâh, Kelley, Mitchell, & Ruggieri, 2016). Employers have been attempting to determine the validity of these illnesses, incentives and propose possible solutions to mitigate these absences, including those caused by family issues.

The U.S. service sector alone loses 2.3% of all scheduled labor hours to unplanned absences, but in some industries, the total cost of unplanned absences approaches 20% of payroll expense. The principal reasons for unscheduled absences (personal illness and family issues) are unlikely to abate anytime soon (Eastont & Goodale, 2005). In 2002, personal illness accounted for 33 percent of unscheduled absences. The three most common reasons for unscheduled absences are personal illness (33%), family issues (24%), and personal needs (21%). Stress as a reason for absenteeism has increased over 300 percent since 1995 (Kim & Garman, 2003).

Although personal illness is a common reason for not coming to work, as Kocakülâh et al. (2016) stated, companies are now attempting to determine whether the employee is actually sick when they call in. Effectively, there are studies such as the one conducted by Hughes and Bozionelos (2007) on a sample of male workers employed as bus drivers, where interviewees confessed that many bus drivers, including themselves, consistently fake illness or sickness in order to be able to participate in and take care of important non-work activities and obligations. For example, one driver noted: "*There would be no* *chance [to be granted permission for time off] ... even for a funeral... people end up going on the sick.*".

On the other hand, as McHugh (2001) stated in his study, interviewees attributed the low absence statistics within their organizations to the fact that the organization is located in a small town where employees are exposed. This tends to make it more difficult for individuals to fake sickness.

Similarly, family issues also play a crucial role in absenteeism. Balancing work and family life can be a struggle in a way that there are plenty of studies about work–family conflicts and how it affects absenteeism and job satisfaction (AlAzzam, AbuAlRub, & Nazzal, 2017; Anafarta, 2011; Vignoli, Guglielmi, Bonfiglioli, & Violante, 2016). Childcare is often a major issue, as Payne et al. (2012) pointed, without childcare, most parents are unable to work, so if parents are not satisfied with their childcare arrangements, they may not be able to fully engage in their work, which should cause concernment in employers about the adequate childcare and parents' satisfaction.

In addition, as Calvano stated (2013), who has cared for frail, ill, or disabled elders while working knows that juggling the two roles can be complicated, daunting, and laden with practical and emotional challenges. Katz and colleagues (2011) found that caregivers experienced a disruption in their work functioning due to absenteeism. Interestingly, eldercare negatively affects women's employment across the European Union, with the greatest effect in Southern Europe and the least in Nordic countries and Central Europe in between (Kotsadam, 2011).

When talking about personal needs, it is critical that individuals seek periodic medical attention, including regularly scheduled "healthy" checkups, as it's proven that maintaining health behaviors and having good health status were associated with less absenteeism (Yun, Sim, Park, Park, & Noh, 2016). Even though, the law (Portuguesa, 2017a) establishes that the public employer has a duty to provide good working conditions and to prevent occupational risks and diseases, taking into account the protection of workers' safety and health, there is no obligation to schedule these checkups. So, it is often difficult for individuals to work a full day and be able to schedule a needed doctor's appointment that does not interfere with their scheduled work day, costing both worker and employer much needed time, which can never be recovered (Kocakülâh et al., 2016).

Finally, it is needed to refer stress as a cause to absenteeism. Employees who are suffering from stress at work are less likely to be productive (Kim & Garman, 2003). The causes of stress, or the stressors, are numerous and can be found anywhere in the workplace. Psychosocial work stressors such as job strain, low decision latitude, low social support, high psychological demands, effort-reward imbalance, and high job insecurity have all been implicated as causes of work stress-related anxiety and depressive illnesses (Stansfeld & Candy, 2006). Furthermore, stress at work also can lead to physical illness, psychological distress and illness, and sickness absence (Jordan et al., 2003). Alongside these, Hauge et al. (2010) demonstrated that workplace bullying is a potent social stressor that is negatively associated with both individual well-being and work-related outcomes, with consequences similar to and even stronger than those of other more frequently studied job stressors. All of this culminates into lowering morale among employees and lowering productivity because of poor performance (Haswell, 2003). Interestingly, burnout, resulted by stress, might explain the relationship between job satisfaction and absenteeism, when referring to total time lost, but not for absence frequency (Ybema, Smulders, & Bongers, 2010). Consequently, stress, depression, or anxiety accounts for 46% of days lost due to illness, according to Cooperand are the single largest cause of all absences attributable to work-related illness. Likewise, stress can, in its turn, lead to seeking alternate employees, which costs the employer money in the form of recruitment and training. What's more, stress can overburden co-workers with additional responsibilities while replacement personnel are hired. This can lead to a heavier workload for already distraught employees, which affects their health and eventually results in even more absenteeism (Haswell, 2003).

Effectively, absenteeism has a large effect either directly or indirectly on a company's bottom line. The costs associated with dealing with absenteeism are significant when everything involved is taken into account. One cannot look at just what it costs to replace the employee for a day. It is necessary to look at what it is going to cost to lighten the load and attempt to attack these ongoing ever-increasing problems in the workplace. One also has to look at the increase in corporate health benefit costs that will result if a hands-off policy is adopted and absenteeism is not taken seriously (Quinley, 2003).

To conclude, since every company is different, it will require various levels of analyses to identify the factors that impact absenteeism for a specific employer. If absenteeism is identified as a significant problem, the company will need to take a hard look at the cause of the problem and begin to consider strategies to recapture lost revenues. Furthermore, as the economy tightens and the related financial stress increases for most employees, it is very likely that employers may see an increase in absenteeism due to stress related issues. The more aware a company is of issues related to employee absenteeism, the more successful they will be in implementing strategies to reduce the related cost and increase productivity (Kocakülâh et al., 2016).

2.2 Public services

2.2.1 Reforms

For the past forty years PPA has seen its rulers proclaiming reforms in all its areas of action (Madureira, 2015), from employment to public debt, going by taxes and its own structure (DGAEP - Direção-Geral da Administração e do Emprego Público, 2013).

From 2005 to 2009, Portugal was subject of a wide PA reform focused mainly on the: Structural Reorganization of State Central Administration; Reform of the Civil Service Regime; Modernization and Administrative Simplification; Modernization of Public Management, as well as the Development of e-Administration (Presidency, 2010).

The reform of the civil service regime presents the following features (Presidency, 2010):

- Progressive convergence of the Civil Service Social Protection Scheme with the General Social Security Scheme;
- Abolishment of the former general mobility mechanisms and their replacement by two new mechanisms: assignment of public interest and internal mobility. Establishment of the special mobility regime;
- Reform of the attachment, careers and remuneration scheme of staff fulfilling public functions, from which should be highlighted the following:
    - Alignment with the private sector with regard to the legal employment relationship;
    - The status of "civil servant" is assigned to a few, special functions related to the exercise of powers conferred by public law that safeguard the general interests of the state: Military, Foreign Affairs, State Security, Criminal Investigation, Public Protection and Inspection Activities;

  - Reduction in the number of general and special regime careers. Establishment of only 3 general regime careers;
  - Establishment of a single pay scale made of 115 pay-steps to be used in setting workers' basic remuneration, replacing the 22 existing pay scales with a total of 522 pay-steps.
  - Replacement of a "career system" by a "position system";
  - Progressions are no longer based on seniority and career advancement and change of pay step is based on performance assessment according to available budget appropriations;
  - Introduction of performance bonuses related to assessment;
- Establishment of the employment contract in public functions scheme aiming at bringing the labor legislation of PA closer to the labor regime of the private sector, highlighting collective bargaining (signing of the first two collective agreements in PA). Together with this new employment contract scheme a new disciplinary statute has entered into force;
- Establishment of an Integrated Public Administration Management and Assessment System (SIADAP) that, for the first time, is applicable to the assessment of services of respective managers and remaining staff. A percentage system (quotas) was set up for the differentiation of performance, including managers: 25% for relevant performance and, within this percentage, 5% for excellent performance. In case of the service itself obtaining the classification of excellent, the percentages for workers increase respectively to 35% and 10%.

In 2008 a financial crisis strikes in and the government sees its ideals being forced by external entities which demanded measures to be taken and directives to be followed as mentioned in Memorandum of Understanding (Troika, 2011).

Effectively, the goals set in the Memorandum, by the IMF, European Central Bank and European Commission colligation, also referred as Troika, resulted in reform measures. A lot of them being applied between 2011 and 2013 without any studies to back up their implementation (Madureira, 2015). As matter of fact, there were reforms like on the mobility regime and HR valorization that were adopted and should be explained, despite some adjustments recently (INA - Direção-Geral da Qualificação dos Trabalhadores em Funções Públicas, 2017a).

Although, not only is needed to understand the influence of these reforms but also why are they needed, i.e. the nature of the crisis itself. The perception of the causes of failure and the responses to those failures will reflect a number of factors (Peters, Pierre, & Randma-Liiv, 2010). One way of understanding the different responses is that they will reflect the interaction of national patterns of governance (Painter, 2010) with the real and perceived nature of the crisis within that country. Thus, the nature of the crisis may be very different in different countries. A comparative analysis of these differences is crucial for understanding the nature of the crisis and to learn from it (Peters et al., 2010).

To conclude his study, Peters et al. (2010) stated, the various responses to the crisis across the range of industrialized countries make the point that there has been no new paradigm for governance emerging as a result of the crisis. This absence of a paradigm appears to result from at least three broad factors: the crisis has been perceived differently in different settings; the different starting points for the different regimes have meant that the policy and governance options available to them were very different, having a new paradigm requires new ideas and those appear to have been in rather short supply.

2.2.2 Labor mobility

As the Ministry of the Presidency (2010) mentioned, the mobility regime includes two patterns, General Mobility (GM) and Special Mobility (SM), also known as professional valorization (INA - Direção-Geral da Qualificação dos Trabalhadores em Funções Públicas, 2017a).

The GM regime consists of transitional modification of functional situation of the worker, within the same body or service, or between different bodies or services, based on grounds of public interest, targeting to increase effectiveness of services by way of a rational use and valuing of Public Administration human resources (Presidency, 2010).

The SM regime has been defined to frame processes for the abolishment, merger and restructuring of public services. Likewise, a general regime for the staff rationalization process has been laid down in situations where human resources assigned to some services are mismatched with regard to permanent needs and the pursuit of the objectives (Presidency, 2010). In order to maximize the full productivity of the workers, it seeks to meet the needs identified by the different bodies and services, promoting the professional valorization of the workers, through standardized training and ensure the resumption of functions by integration, in a period of 3 months (INA - Direção-Geral da Qualificação dos Trabalhadores em Funções Públicas, 2017b).

Additionally, mobility can be external or internal, always presupposing the existence of a public interest. On one hand, external mobility, also called as a public interest cession, occurs when mobility is from public entities to private entities, or by public entities not covered by GLPAE, or vice-versa. On the other hand, internal mobility covers workers who have a legal relationship of public use, for an indefinite period, and can be operated between public entities not belonging to the GLPAE. In this case, there are two situations to consider: mobility in the same category and mobility intercarrier or categories (between different ones) (INA - Direção-Geral da Qualificação dos Trabalhadores em Funções Públicas, 2017c).

To conclude, these regimes had, and still have, an important role in the PPA, noticing that, according to public employment statistics, there were 1176 workers, in December of 2011, placed in the valorization regime, and, in September of 2017, that number dropped to 323 workers (DGAEP - Direção-Geral da Administração e do Emprego Público, 2017b), meaning that there were a significant amount of workers who did not get unemployed, which has been, as referred before, a big deal in Portugal, because of these reforms. Additionally, as Boswell et al. (2005) proved that, for a 5-year period, low satisfaction would precede a voluntary job change, which is followed by an immediately increase in job satisfaction (the honeymoon effect), even though there is a decline in job satisfaction after (the hangover effect), and as demonstrated in Ybema's et al. (2010) study, individuals who are dissatisfied with their jobs are more likely to be frequently absent in future. It's possible to infer that mobility might influence positively absenteeism.

2.2.3 Contract of Employment characterization

According to the 11th article, Contract of Employment (CE) is the contract in which a person undertakes, through retribution, to provide his or her activity to another person or persons, within the organization and under their authority (Portuguesa, 2016);

However, it is presumed that the parties have concluded an employment contract if it has some of the following characteristics, according to the 12th article (Portuguesa, 2016):

- activity carried out in a place belonging to or by the beneficiary has been determined;
- the equipment and work tools used belong to the beneficiary of the activity;
- the provider of the activity complies with times determined by the beneficiary of the activity;
- a certain amount is paid to the provider of the activity as a consideration thereof;
- the provider of the activity performs managerial or managerial functions.

Effectively, according to the 6th article, the work in public functions may be provided through the CE, being this the one by which a natural person provides his activity to a public employer, in a subordinate manner and for remuneration. The public employment link has the following modalities (Portuguesa, 2017b):

a) Contract of employment in public functions, which is, usually, the main public employment link, according to the 7th article;

b) Appointment;

c) Service commission.

Thus, the bond of public employment may be constituted for an indefinite period or for a definitive term (Portuguesa, 2017b).

On one hand, according to the 8th article, the public employment link is constituted by appointment in the cases of exercise of the following functions (Portuguesa, 2017b):

a) Generic and specific missions of the Armed Forces in permanent staff;

b) External representation of the State;

c) Safety information;

d) Criminal investigation;

e) Public security, both in the free environment and in an institutional environment;

f) Inspection.

On the other hand, the public employment link is constituted by service commission in the following cases, as stated in the 9th article (Portuguesa, 2017b):

a) Positions not included in careers, namely positions of leadership;

b) Functions exercised with a view to acquiring specific training, academic qualification or professional title per employee with public employment bond of indefinite duration.

One empirical variation of these contracts was studied by Chatterji (2002) that extended the analysis of absence and sick pay by considering the impact on productivity of work attendance by sick workers, also called as presenteeism. The resulting contract embodies a richer complex of incentive compatibility considerations, which involves firms offering: a wage that is strictly higher than sickness pay offer and a sickness pay that is strictly positive. In addition, the results showed that the typical reaction to lower sickness pay in order to reduce absenteeism, may have another outcome, resulting in greater presenteeism instead.

2.2.4 Worktime schedule

According to the nature of their activities, services may adopt, with regard to workers on an appointment regime, the following patterns of working hours (Portuguesa, 2017b):

a) Flextime – are those enabling workers to manage their working hours, by choosing the beginning and end working hours, complying with the previously established fixed core morning and afternoon working hours;

b) Fixed working hours – are those which are divided by two daily periods, with beginning and end fixed hour, separated by a rest break;

c) Staggered working hours – are those enabling establish, service by service or by groups of workers, different fixed beginning and end working hours, while maintaining unchanged the daily work period;

d) Shift working hours- are those in which, for need of regular and normal operation of the service, there is the performance of work in, at least, two daily and successive periods, being each one of them of duration not lower than the daily average work duration;

e) Continuous journey working hours – are those in which are uninterrupted work, except for a rest period of no more than thirty minutes, which for all purposes is considered as working time;

f) Half journey working hours – are those which are known as part-time work and it consists in the provision of work in half of the normal full-time working schedule;

g) Specific working hours – are those which are envisaged to adjust to workers' needs or of his/her members of the family (i.e. student-worker).

According to the 109th article, the typical daily work period is interrupted by a rest break of duration not lower than one hour nor higher than two hours, except in duly grounded exceptional cases, so as workers do not perform more than five hours of consecutive work, save in case of continuous work (Portuguesa, 2017b).

In the employment contract for public functions a model of working hours was not set. Instead, it was left to public employer entities upon consultation of the workers' representative entities and works councils or in the default of these ones the inter-trade union commissions, trade union commissions or trade union representatives (Presidency, 2010).

Referring to the 105th article, the law maker set the limits of duration of the daily normal work period (7 hours) and weekly (35 hours), as well as the interruption for the rest break so as to ensure a performance of work not higher than 5 consecutive work hours (Portuguesa, 2017b).

These issues may the subject of change by collective labor regulation instrument, enabling, for example, reduction of the rest break and reduction or increase of the work period (Presidency, 2010).

Evidence suggests that flextime appears to lower the incidence of tardiness, although the reduction appears to be significant only for women employees (Ala-Mursula, Vahtera, Kivimaki, Kevin, & Pentti, 2002). Another benefit of flexible work scheduling in the form of flextime is decreased absenteeism (Casey & Grzywacz, 2008). Employers seem to benefit when employees have working time autonomy to the extent that more flexible working time arrangements reduce absenteeism if they facilitate the combination of paid work with other activities. In the longer term, they may further decrease absenteeism by also improving worker health, through reduced stress and increased job satisfaction (Possenriede, 2011).

2.2.5 Absences

A major source of cost savings for organizations is lower rates of absenteeism (Kelly et al., 2008; Kossek & Hammer, 2008). As Golden (2011) stated, rates of absenteeism are one indicator of employee commitment that affects organizational productivity. Effectively, in the United States there is little variation between industries, although the absence rate is highest in public sector (especially education and health) jobs and lower in construction.

De Paola et al. (2014) also studied absenteeism in public sector, specially the effects of an Italian policy intervention reducing sick leave compensation and increasing the monitoring of the health status of absent employees in the absenteeism. Interestingly, the reform started to change civil servants' behavior before its effective implementation. Effectively, they discovered that workers' probability of taking days off work for sick leave decreased strongly (the estimated effect was of about 53%) once the reform was effectively implemented. The authors also found that absence behavior is responsive both to wage reductions and to changes in monitoring intensity.

This study will explore different kinds of absences and, in order to understand them, the next section will explain the differences between licenses/leaves, unplanned absences and vacations.

2.2.5.1 Licenses and leaves

According to the 34th article, in the scope of protection to parenthood releases are also granted to cope with the need of absence of parents in following situations (Portuguesa, 2016):

- Prenatal consultations;
- Breastfeeding and nursing;
- Release for assessment for adoption;
- Release from performance of work in the night period;
- Release from performance of work by pregnant worker, worker who has recently given birth or breastfeeding worker, on grounds of protection of her safety and health;

It is important to remark these absences do not impact the workers' rights, namely their salary, unlike the unjustified leaves (Portuguesa, 2016).

Although, there are other leaves in which total loss of remuneration is verified, such as (Presidency, 2010):

- Leave up to 90 days;
- Leave for one year;
- Extended leave;
- Leave for accompanying spouse placed abroad;
- Leave for performing functions in international organizations;
- Leave for the performance of functions in trade union association;
- Leave for attendance of training courses;

- Non-specified leave.

Regarding releases, another pattern is provided in the PPA. It consists of granting workers release from appearing in the service, that, in a determined working day are bound to the duty of assiduity. It is not regarded as holiday, it does not suspend holidays and workers who are enjoying holidays are not entitled to an extra day of holiday for compensation. For example, at 12th of May of 2017, due to Pope's visit to Portugal, the PPA's workers could decide to not show up to work without repercussions in their salaries (Primeiro-Ministro, 2017).

Note that as Ziebarth's (2013) study suggested that cuts in statutory sick pay did not significantly affect the incidence of long-term absenteeism. Employees did not significantly adjust their long-term sick leave behavior on the decision to enter an episode of long-term sick leave. As for the effects on the decision to reduce the length of the long-term sick leave episodes, the models also produced mostly insignificant and small, but negative, reform effects. However, for two subsamples, the poorer half of the sample as well as middle-aged employees working full-time, the findings suggest that cutting long-term sick pay may have reduced the length of such absences significantly.

2.2.5.2 Unplanned Absences

According to the 133rd article, an employee is absent when not present in the work place during the normal worktime schedule. If the absence is inferior to the normal worktime schedule, then the nonappearance durations are add up, respectively, to determine the absence period (Portuguesa, 2017b).

In the same article, there is a mention to the difference between justified and unjustified absences. The justified absences are the following (Portuguesa, 2017b):

a) Those given, for 15 consecutive days, at the time of marriage;
b) Those motivated by the death of the spouse, relatives or alike;
c) Those motivated by the provision of evidence in an educational establishment;
d) Those motivated by the impossibility of providing work when is not attributable to the worker, namely the observance of medical prescription following the use of medically procreative technique assisted, illness, accident or fulfillment of legal obligation;

e) The one motivated by the provision of urgent and essential assistance to the child, grandchild or member of the household of the worker;
f) Those motivated by the transfer to educational institution responsible for the education of minor, by reason of the educational situation, for the strictly necessary time, up to four hours per quarter, for each minor;
g) Those of worker elected to the collective representation structure of the workers, in accordance with 316th article;
h) Those given by candidates for elections to public office, during the legal period of their election campaign, in the terms of the corresponding electoral law;
i) Those motivated by the need for outpatient treatment, medical consultations and complementary tests which cannot be carried out outside the normal working hours and only for the time strictly necessary;
j) Those motivated by prophylactic isolation;
k) Those given for blood donation and rescue;
l) Those motivated by the need to submit to selection methods in insolvency proceedings;
m) Those given on account of the period of leave;
n) Those that by law are considered as such.

Naturally, all the others not mention above are considered as unjustified absences (Portuguesa, 2017b).

Easton and Goodale (2005) studied the unplanned absences and strategies to recover from them. They found that the reduction in total profits due to absenteeism is strongly influenced by the staffing strategies and absence recovery policies that firms adopt to cope with absenteeism. Forty-hour work weeks and zero anticipated absenteeism with holdover absence recovery appears to be a fairly robust combination, on average reclaiming nearly 60% of the profit consumed by unchecked absenteeism. For firms unwilling or unable to implement active absence recovery policies, however, planned overtime staffing strategies with absence anticipation appear less vulnerable to absenteeism.

Another study suggests that, there is a strong positive relationship between the generosity of granting sick leave days and days of absenteeism. Furthermore, the results showed that there is a positive relationship between income differences of neighboring states and sick leave days. Which can be interpreted as an evidence that employees of lower income countries in the OECD, like Portugal, have an incentive to report in sick on their regular job and, instead, work in the unofficial market of the high-income neighbor state (Osterkamp & Rohn, 2007).

2.2.5.3 Vacations

Despite being an absence, vacations according to 237th article of Employment Contract, holidays constitute a right that cannot be waived by the worker and its enjoyment shall not be replaced by any economic compensation whatsoever (Portuguesa, 2016), which is why it will not be considered when creating the data model to forecast absenteeism. In the same article it is mentioned that the right to holidays is obtained on the 1st January of each calendar year and is related to, as a rule, the service performed in the preceding calendar year (Portuguesa, 2016).

Referring to the 238th article, each worker is entitled to a minimum holiday period of 22 remunerated working days (Portuguesa, 2016). The minimum holiday period may still be increased in the framework of performance rewarding systems, without prejudice to increases granted to each job attachment (Portuguesa, 2017b).

The worker is entitled to a holiday allowance, as mentioned the 152nd article, which value is the same as the salary. The aforementioned should be paid in June or together with the salary of the month just before the vacation's enjoyment month (Portuguesa, 2017b).

Although vacations will not be considered as an absence, it is important to note that, as Westman and Etzion (2001) stated in their study, vacations alleviated perceived job stress and thus also the experience of burnout. However, the most notable contribution of the study was the finding of a respite effect on a behavioral strain, the absenteeism, measured objectively. It was found that absenteeism for non-health reasons decreased after vacation, which implies that taking a vacation can be regarded as a stress management technique. As such, it corroborates findings concerning the beneficial effects of stress management intervention on burnout and absenteeism. Thus, in order to decrease absenteeism, the organization should try to regulate vacations according to stressful periods.

2.3 Human resources and data mining

As Bernik et al. (2007) stated, we are now in the knowledge era. Thus, the basic element of HR research is the acquirement of knowledge. The amount of information currently available is huge and the problem of the HR management is mainly the filtering and integration of information. Numerous scholars in the area of HR management that delved in this problem have reached the conclusion that it is not only the acquirement of information, but mostly its management, stemming from the fact that information is one of the basic resources of every organization. Like an organization manages material, human and other resources, it also manages the information that flow either within the company or outside. Because of that, information management has recently become one of the priority tasks of HR management (Bernik et al., 2007).

In this context, data mining techniques can be used to extract and discover the valuable and meaningful knowledge from a large amount of data. Additionally, among the major tasks in data mining are classification and prediction, concept description, rule association, cluster analysis, outlier analysis, trend and evaluation analysis, statistical analysis, and others. Classification and prediction tasks are among the popular tasks in data mining, widely used in many areas especially for trend analysis and future planning (Jantan, Hamdan, & Othman, 2010).

As matter of fact, there are many areas which adapted such approach to solve their problems such as in finance (Cowan, 2002), tourism (Moro, Rita, & Oliveira, 2018), medical and engineering (Kusiak, Kernstine, Kern, McLaughlin, & Tseng, 2000), marketing (Silva, Moro, Rita, & Cortez, 2018), telecommunication (Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012), customer relationship (Chen, Wu, & Chen, 2005), social media (Moro et al., 2016), among others. Nevertheless, data mining application has not attracted much attention from people in HR field (Chien & Chen, 2008). In fact, prediction applications are mainly developed in business and industrious fields and quite restricted studies involved human talent in an organization (Jantan, 2009). HR data can provide a rich resource for knowledge discovery and for decision support system development (Jantan et al., 2010).

Regardless of the location, the HR database mostly contains data on the employee and the job position that are sorted into these categories (McLeod & Sanctis, 1995):

- Basic employee data;

- Data on professional, work related and personal development;
- Data on work and job performance;
- HR social data;
- Work safety data, data on managerial and societal activities of the employees.

In fact, this study is based on a dataset where the information about the employee and its job is sorted just like McLeod and Sanctis (1995) mentioned.

2.4 Data mining and public sector

Finding useful patterns in data has been given, through the years, a lot of names, one of those being data mining. Typically used by statisticians, data analysts, and the management information systems communities, data mining gained its popularity in the database field (Fayyad & Uthurusamy, 1996).

For the past few years, the uncontrolled growth of database brought data mining to the forefront of new business technologies. Estimates show that double of the world's data is being stored every 20 months, which creates a big opportunity for the increase of data mining use (Witten, Frank, & Hall, 2016).

Nowadays, with the world's complexity growing day by day, data should be intelligently analyzed so it can lead to new insights, better decision making and competitive advantages (Witten et al., 2016).

Effectively, within the public sector the same problem resides, as the need to have tools in order to dig through huge collections of data, often referred as "official data", are needed to recognize trends or patterns (Antti, Syvajarvi, & Stenvall, 2010).

Mining public sector's data is crucial to evaluate public programs and investments, for example in social security, health and economic growth (Antti et al., 2010). Although, a major concern is always brought up, which is the balance between managing personal information, respecting its privacy, as well as data integrity and data security, and maximizing the available information for general policy purposes (Antti et al., 2010).

In order to performance's measurements requirements are met, policy makers see themselves studding a mix of different types of information and making decision based on that (Hamlin, 2007). Although, with insufficient high-quality data and available information as well as high-stakes pressures to demonstrate organization improvements, the data for these purposes is still more likely to be misused and manipulated.

It is certain that organization and government activities confront requirements like predicting and forecasting (Antti et al., 2010), so this study shall give them some conclusions to support their next decision, as other studies about that used data mining in the public sector tried to do (Cleary, 2011; Durairaj & Ranjani, 2013; Moro, Cortez, & Rita, 2014).

2.4.1 Support vector machines

The foundations of SVMs have been developed by Vladimir Vapnik (1995) and are gaining popularity due to many attractive features, and promising empirical performance. The formulation embodies the SRM principle. SVMs were developed to solve the classification problem, but recently they have been extended to the domain of regression problems (for prediction of continuous variables). SVMs can be applied to regression problems by the introduction of an alternative loss function that is modified to include a distance measure. The term SVM is referring to both classification and regression methods, and the terms Support Vector Classification (SVC) and Support Vector Regression (SVR) may be used for more precise specification (Kantardzic, 2011).

As Kantardzic (2011) mentions in his book, an SVM is a supervised learning algorithm creating learning functions from a set of labeled training data (Figure 2). It has a sound theoretical foundation and requires relatively small number of samples for training; experiments showed that it is insensitive to the number of samples' dimensions. Initially, the algorithm addresses the general problem of learning to discriminate between members of two classes represented as n-dimensional vectors. The function can be a classification function (the output is binary) or the function can be a general regression function.

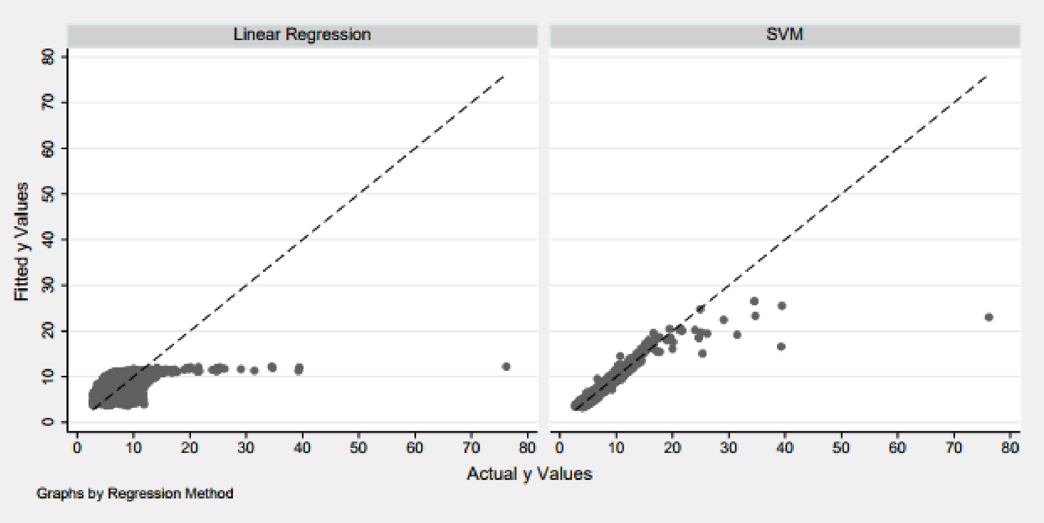


Figure 1 - Fitted versus actual values for both linear regression and SVM (Guenther & Schonlau, 2016).

An SVM includes a set of techniques which produces nonlinear boundaries by constructing a linear boundary in a large, transformed version of the feature space (Hastie, Tibshirani, & Friedman, 2009). Essentially, what occurs is that through the usage of a kernel, an approximation function, and addition of a loss function, the hinge loss function, the desired outcome is optimized. The jumbled data in the input space takes form into a separating hyper plane, the feature space, which will be easier to analyze because the data becomes structured and, therefore, further analysis is made possible until it develops into intelligible information (Silva, 2016).

2.4.2 Recency, frequency and monetary value (RFM)

As mentioned by Moro et al. (2015), one way of characterizing a database of customers is by computing their recency, frequency and monetary (RFM) characteristics. These allow to capture customer behavior in a very small number of features (Table 1). Still, the relative importance among RFM varies with the characteristics of the product and industry (Moro et al., 2015). As matter of fact, feature weight has been a subject of research in order to improve classification accuracy (Ahn, Kim, & Man, 2006). Actually, there are already some studies to determine the relevance of each of the three RFM features like the use of the K-means algorithm to build clusters by RFM attributes (Cheng & Chen, 2009), the analytic hierarchy process to determine the relative weights of RFM (Liu & Shih, 2005), and measuring loyalty through a user's RFM and a smart object that attributes weights to RFM features based on its goal (Kwon & Lee, 2011).

Table 1 - Adapted RFM telemarketing features analyzed
(Moro et al., 2015)

Factor	**Citation**	**Application to telemarketing**	**Application to absenteeism**
Recency	How recent is the last purchase? Time of most recent purchase Period since the last purchase The total days between the day of the latest purchase and analysis (days) The period since a customer last purchase	Months since the last purchase up to date	Days since the worker failed to show up to work? Was the motive the same?
Frequency	How often does customer buy a product? Number of prior purchases Number of purchases made within a certain period Consuming frequency (times) The number of purchases made within a certain period	Number of times the client subscribed the deposit previously	How many times, in the last year, was the employee absent? How many days was the worker absent since last time?
Monetary	The money spent during a certain period Amount of money of total consuming The amount of money that a customer spent during a certain period	Total amount of money the client subscribed in previous contacts	Total amount spent in temporary work to fill in the absent worker.

Based on the Moro's (2014) telemarketing RFM approach, in the present study it will be studied the historic variables which reflects the worker's behavior (contrasting with the consumer behavior used in telemarketing study), using recency as well as frequency, in order to get the most accurate model for the absenteeism in PPA.

If the model's results turn out to be positive, it may open the door to use this the RFM features in other areas than customer-oriented subjects like marketing or sales.

2.4.3 Regression performance metrics

According to Moro et al. (2016), one of the crucial steps in model building is assessing its adequacy in predicting what it is supposed to. As matter of fact, if one model fails to predict its output then it is inadequate (Diebold & Mariano, 2002). Which brings out the importance of forecast accuracy since the derived forecasts are used for guiding decisions-making (Moro et al., 2016).

Thus, there are some metrics to measure the accuracy of the forecasts, as Witten (2016) mentioned, such as the mean-squared error, which is one of the most commonly used measure, because it tends to be the easiest measure to manipulate mathematically. The Mean Absolute Error (MAE) is an alternative: just average the magnitude of the individual errors without taking account of their sign. Mean-squared error tends to exaggerate the effect of outliers—instances whose prediction error is larger than the others—but absolute error does not have this effect: all sizes of error are treated evenly according to their magnitude (Witten et al., 2016). Thus, in this study the MAE is chosen as one of those metrics to measure the model performance.

From the MAE value, it is possible to get another perspective of the deviation in predictive capacity of the model and that's by using the mean absolute percentage error (MAPE), the other metric that will be used to evaluate the model.

$$MAE = \frac{1}{n} \sum \left| y_{observed} - y_{predicted} \right|$$

Effectively, MAPE is the ratio of the MAE divided by the total of true values, e.g. it is the relative variation to those values and it can only be applicable if the actual value, of the predicted one, is greater than 0, otherwise the calculation is impossible (Moro et al., 2016).

$$MAPE = 100 \times \frac{1}{n} \sum \frac{\left| y_{observed} - y_{predicted} \right|}{y_{observed}}$$

2.4.4 Sensitivity analysis

It is often a challenge to extract knowledge in a way that is easy to understand, specially from a black box model, which lead to a new stream of research to tackle this problem. Therefore, methods like extracting rules from networks and the sensitivity analysis have emerged.

As Saltelli et al. (2002) mentioned, SA is the study of how the uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input. Therefore, SA is considered as the study of the relative importance of different input factors on the model output.

SA has been mostly used as a variable/feature selection method (e.g. to select the least relevant feature that is deleted in each iteration of a backward selection). However, SA can also be used to explain the model, as recognized in but more explored in and, thus, opening the black box (Cortez & Embrechts, 2011).

There are several types of algorithms to choose from within SA. The one chosen for this study is the data-based sensitivity analysis, mainly due to its capability of detecting input variable interactions. Effectively, DSA uses training samples and its' main goal is to harvest the possible interactions between inputs, while speeding up if a proportion of the training samples (randomly selected) are used instead of the whole training set (Cortez & Embrechts, 2013).

Although, it is important to recognize that the sensitivity of the parameter in the equation is what is being determined, not the sensitivity of the parameter in nature. If the model is wrong or if it is a poor representation of reality, determining the sensitivity of an individual parameter in the model is a meaningless pursuit (Pilkey & Pilkey-Jarvis, 2009).

3. Materials and methods

3.1 CRISP-DM

Over the last few years, the field of data mining became very important for different industries, co-corporations and businesses because of its ability to use huge amount of data that had previously no use and to analyze and predict trends and patterns (Shafique & Qaiser, 2014). Essentially, the risk of wasting the wealthy and valuable information contained by the big databases was arise and this requires the use of adequate techniques to get useful knowledge (Chen, Han, Yu, & Han, 1996) so that the field of data mining had been emerged in 1980's and is still making progress. With the emergence of this field different process models were introduced. These process models guide and carry the data mining tasks and its applications (Shafique & Qaiser, 2014). One of these models it will be used in this study and it's called as CRoss-Industry Standard Process for Data Mining (CRISP-DM).

Effectually, CRISP-DM is a popular methodology for increasing the success of DM projects (Chapman et al., 2000). The methodology defines a non-rigid sequence of six phases, which allow the building and implementation of a DM model to be used in a real environment, helping to support business decisions (Moro, Cortez, & Laureano, 2011) (Figure 2).

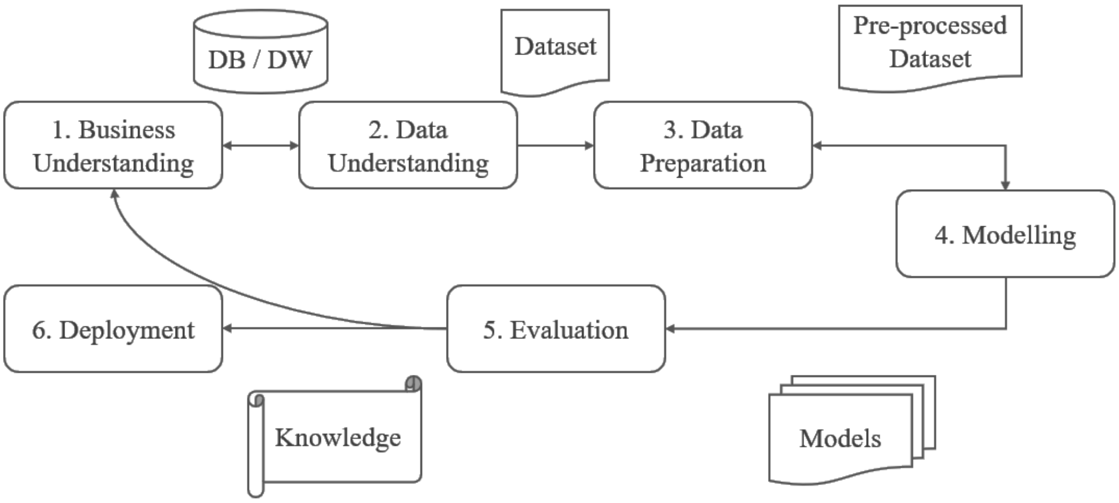


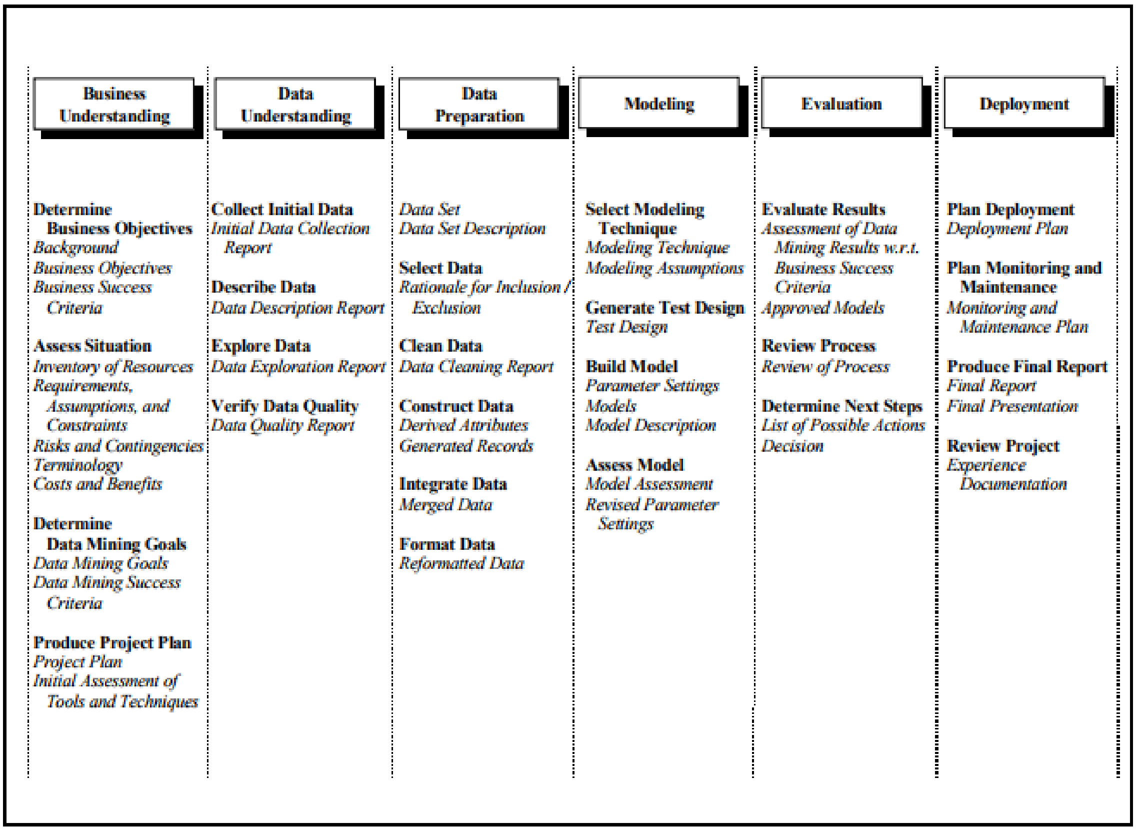
Figure 2 - The CRISP-DM process model, adapted from Moro et al. (2011)

According to Moro and Laureano (2011), CRISP-DM defines a project as a cyclic process, where several iterations can be used to allow final result more tuned towards the business goals. Effectively, this cycle goes through the following phases (Figure 3) (Wirth, 2000):

- Business Understanding - This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives.
- Data Understanding - The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

  There is a close link between Business Understanding and Data Understanding. In fact, the formulation of the data mining problem and the project plan require at least some understanding of the available data.
- Data Preparation - The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools. In this context, Witten described data as concepts (what needs to be learned), instances (independent records related to an occurrence) and attributes (which characterize a specific aspect of a given instance) (Witten et al., 2016);
- Modeling - In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques require specific data formats.
- Evaluation - At this stage in the project you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

- Deployment - Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.



Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

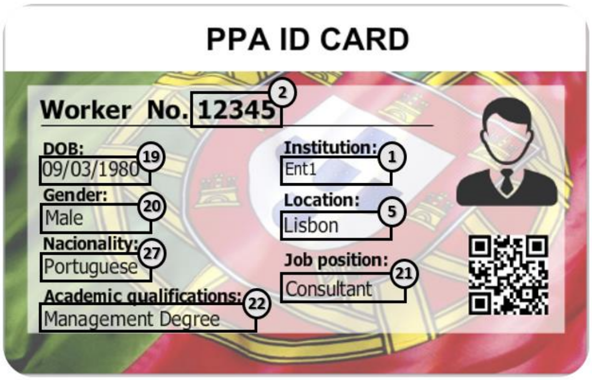
Figure 3 - Overview of the CRISP-DM tasks and their outputs
(Wirth, 2000)

3.2 Business and data understanding, and data preparation

The problem that lead to this report is, in fact, the lack of knowledge about the possible duration of an absence in the PPA, given a worker's profile and the reason why that is a need to get off work, as well as, other information that is relevant to get a solid prediction.

With that in mind, it was extracted from HCM ERP system, which manages the human resources of seven public entities, data that will be tested in order to get a stable combination to predict the duration of the next absence, by using the error measures referred before. The two main sources of information were an absenteeism map and a query to get the most important details about the worker.

With the intention of understanding where the information was extracted, it was illustrated in the Figures 4 to 7, where it's shown the workers' profile and work schedule, as well as, the absence details, which resulted in the features used to create the model used to predict the number of days a worker will miss work.



PPA ID CARD

Worker No. 12345

DOB: 09/03/1980

Gender: Male

Nacionality: Portuguese

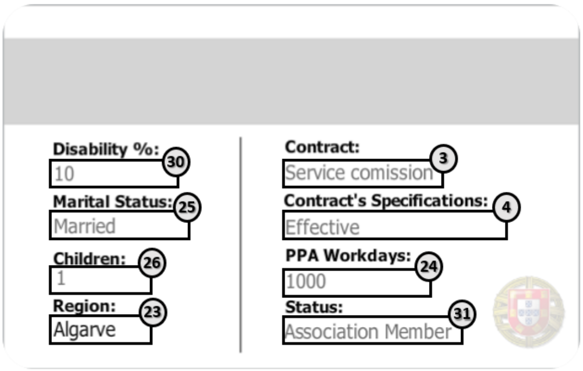
Academic qualifications: Management Degree

Institution: Ent1

Location: Lisbon

Job position: Consultant

Figure 4 - PPA ID Card Front



Disability %: 10

Marital Status: Married

Children: 1

Region: Algarve

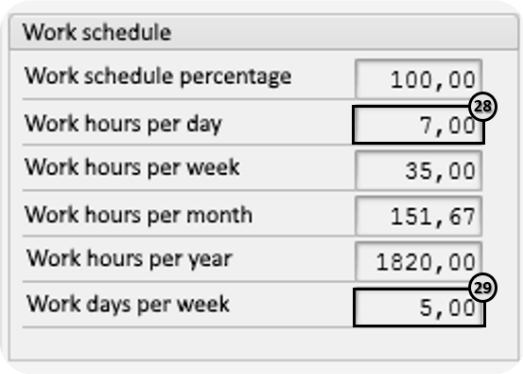
Contract: Service comission

Contract's Specifications: Effective

PPA Workdays: 1000

Status: Association Member

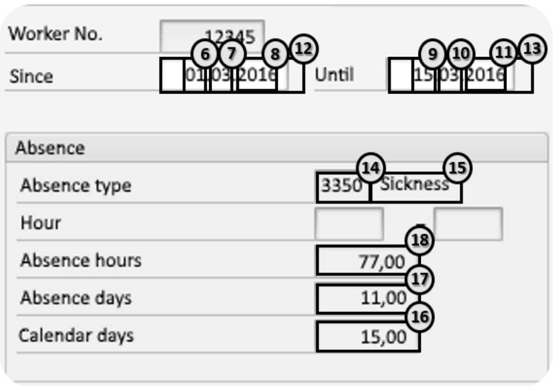
Figure 5 - PPA ID Card Back



Work schedule

Work schedule percentage	100,00
Work hours per day	7,00
Work hours per week	35,00
Work hours per month	151,67
Work hours per year	1820,00
Work days per week	5,00

Figure 6 - Work schedule's details



Worker No. 12345

Since 01 03 2016 Until 15 03 2016

Absence

Absence type	3350 Sickness
Hour	
Absence hours	77,00
Absence days	11,00
Calendar days	15,00

Figure 7 – Absence's details

Table 2 lists all the features gathered after the data collection process, as matter of fact, the first two columns ("N.º" and "Feature Name") identify the features, giving each one a number that corresponds to the locations from where they were collected, as identified in Figures 4 to 7. In order to understand dataset's attributes, it was added columns to Table 2 as "Origin", defines the source of the information retrieved, "Source Type", groups features by the nature of the information, "Data Type", corresponds to the R tool data types, "Description", gives a short description of the feature, and "Status", acknowledge features that were included or excluded of the model created.

Table 2 - Features list

N.º	Feature Name	Origin	Source Type	Data Type	Description	Status
1	Entity	Extracted	Entity	Character	Worker's entity which is responsible for paying salaries and account the absences	Included
2	WorkerNumb	Extracted	Entity	Character	Number which identifies the worker	Excluded
3	Contract	Extracted	Entity	Character	Contract type between the worker and the institution.	Included
4	ContractSpecs	Extracted	Entity	Character	Contract's specifications	Included
5	Workplace_ Location	Extracted	Entity	Character	Where the worker does its activities	Excluded

6	DayBegin	Extracted	Absence	Integer	First day of absence	Excluded
7	MonthBegin	Extracted	Absence	Integer	Absence start month	Excluded
8	YearBegin	Extracted	Absence	Integer	Absence star year	Excluded
9	DayEnd	Extracted	Absence	Integer	Last day of absence	Excluded
10	MonthEnd	Extracted	Absence	Integer	Absence end month	Excluded
11	YearEnd	Extracted	Absence	Integer	Absence end year	Excluded
12	DateBegin	Extracted	Absence	Date	Absence begin date	Excluded
13	DateEnd	Extracted	Absence	Date	Absence end date	Excluded
14	AbsenceCode	Extracted	Absence	Integer	Code which identifies the absence	Excluded
15	AbsenceDesc	Extracted	Absence	Character	Absence description	Included
16	CalendarDays	Extracted	Absence	Numerical	Absence calendar days (including weekends)	Excluded
17	AbsenceDays	Extracted	Absence	Numerical	Duration of the absence in working days	Included
18	AbsenceHours	Extracted	Absence	Numerical	Duration of the absence in working hours	Excluded
19	BirthDate	Extracted	User	Date	Worker's birthdate	Transformed
20	Gender	Extracted	User	Character	Worker's gender	Included
21	JobPosition	Extracted	Entity	Character	Worker's job position	Included
22	AcademicQual	Extracted	User	Character	Worker's academic's qualifications	Included
23	LivingDistrict	Extracted	User	Character	Workers' living district	Included
24	PPAWorkDays	Extracted	User	Integer	Worker's days in PPA (antiquity)	Included
25	MaritalStatus	Extracted	User	Character	Worker's marital status	Included
26	ChildrenNumb	Extracted	User	Integer	Worker's number of children	Transformed
27	Nacionality	Extracted	User	Character	Worker's nationality	Included
28	WorkHours_Day	Extracted	Entity	Numerical	Worker's working hours per day	Included
29	WorkWeekdays	Extracted	Entity	Numerical	Worker's working days per week	Excluded
30	DisabilityPercent	Extracted	User	Integer	Worker's disability percentage	Included
31	Specificities	Extracted	User	Character	Worker's specificities	Included
32	DaysNo_Absences	Computed	User	Integer	How many days with no absences?	Included
33	TimesAbsent_LastYear	Computed	User	Numerical	How many times the worker got absent during the year?	Included
34	DaysAbsent_SinceLast	Computed	User	Numerical	Sum of the worker's absences until the present absence	Included
35	Age	Computed	User	Numerical	Worker's age	Included
36	TimesAbsent_SameMotive	Computed	User	Numerical	How many times the worker got absent for the same motive/reason?	Included
37	HaveChildren	Computed	User	Character	Does the worker have kids?	Included
38	AbsentAfter_Vacation	Computed	User	Character	Is the absence after vacation?	Included
39	DaysAfter_Vacation	Computed	User	Integer	How many days the worker got absent after vacations?	Included
40	VacationDays	Computed	User	Numerical	How many vacations days did the worker have?	Included

After the assembly of the dataset, Table 2 was populated, which resulted in 31 extracted features and 9 computed ones. Likewise, the features collected were separated in groups by source type, so features like "Entity", "WorkerNumber", "Contract", "ContractSpecs", "JobPosition", "WorkHoursDay" and "WorkWeekdays", which define the contractual link between the worker and the employer, they were classified as "Entity" data. Similarly, features as "DateBegin", "DateEnd", "AbsenceCode", "AbsenceDesc", "CalendarDays", "AbsenceDays" and "AbsenceHours", that describe the absence, they were grouped as "Absence" data, followed by the "User" data that has features such as "Gender", "AcademicQual", "LivingDistrict", "PPAWorkDays", "MaritalStatus", "ChildrenNumb", "Nacionality", "DisabilityPercent" and "Specificitiesin" in order to characterize the worker's bio. Finally, it was needed to compute some features to complete the worker's profile, thus "User" data, as "DaysNoAbsences", "TimesAbsentLastYear", "DaysAbsentSinceLast", "Age", "TimesAbsentSameMotive", "HaveChildren", "AbsentAfterVacation", "DaysAfterVacation" and "VacationDays".

Although, as Han et al. (2012) noted, real-world data tend to be incomplete, noisy, and inconsistent, so, it's required to select the suitable data and eliminate all the other that did not meet the requirements.

Initially, the dataset was composed by 40 different attributes including the output variable, contemplating 59,163 observations, which represent the workers' absences, with different lengths and causes. Though, after analyzing the outliers (Figure 8) and some others incongruencies, the number of observations dropped to 36,499, as well as the features used, that were set in 25, classified as "Included" in the "Status" column. Though, the high amount of records with 1 day of absence sets some concern, because of its possible influence in the conclusions.

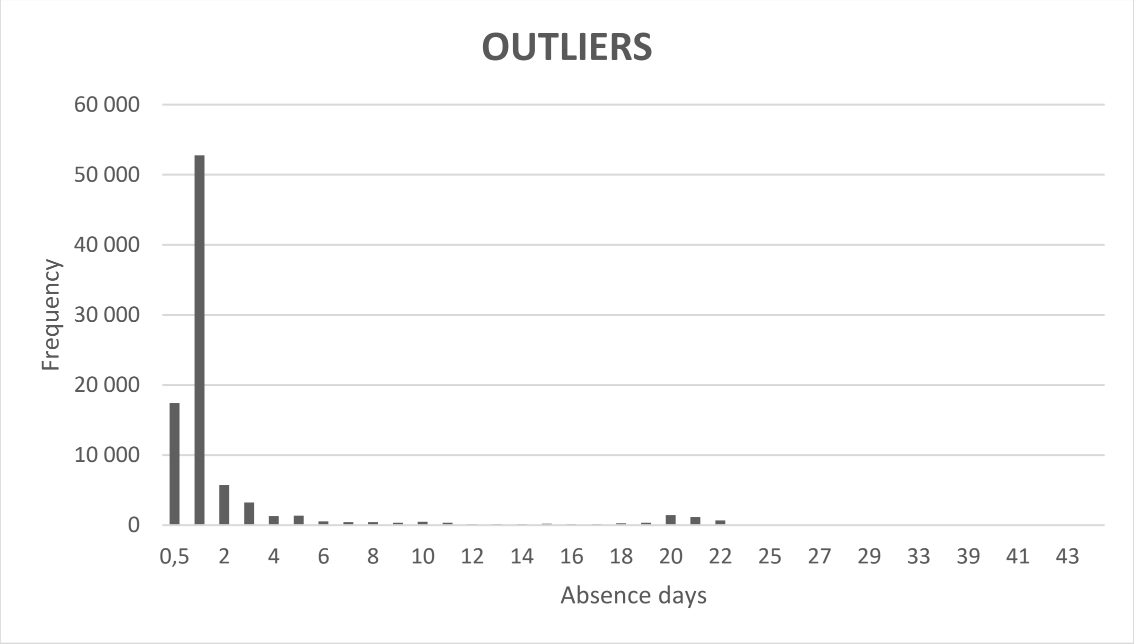


Figure 8 - Outliers

3.3 Modeling and evaluation

After collecting all the data and getting the dataset clean and ready for extracting knowledge, it is time to use the most suitable methods, to get the best results. Figure 9 pictures these two stages.

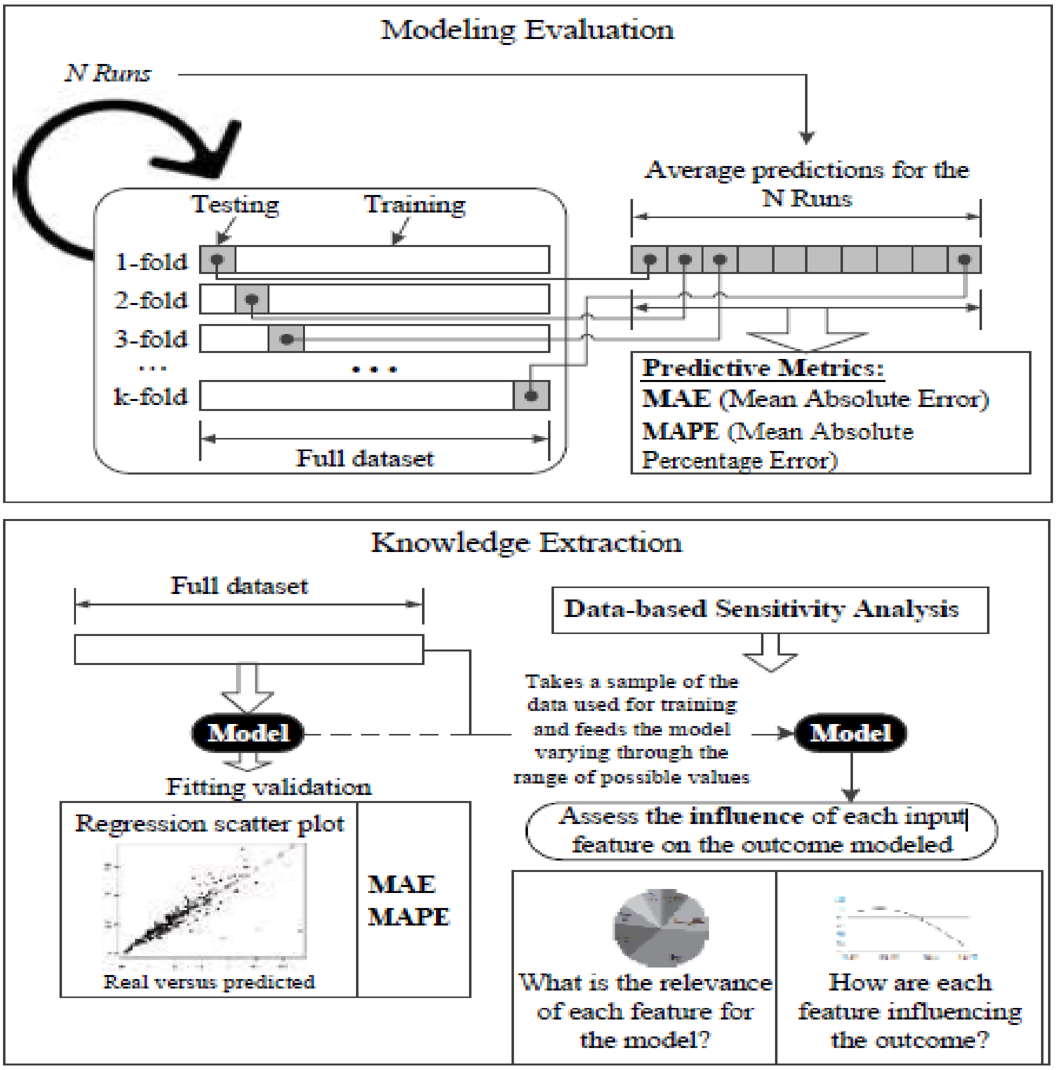

Modeling Evaluation
N Runs
Average predictions for the N Runs
Testing
Training
1-fold
2-fold
3-fold
...
k-fold
Full dataset
Predictive Metrics:
MAE (Mean Absolute Error)
MAPE (Mean Absolute Percentage Error)
Knowledge Extraction
Full dataset
Data-based Sensitivity Analysis
Takes a sample of the data used for training and feeds the model varying through the range of possible values
Model
Model
Fitting validation
Assess the influence of each input feature on the outcome modeled
Regression scatter plot
Real versus predicted
MAE
MAPE
What is the relevance of each feature for the model?
How are each feature influencing the outcome?

Figure 9 - Modelling evaluation and Knowledge extraction scheme (Silva et al., 2018)

The first step was to run a SVM model to predict they days of absence that a worker would miss, based on the other features, and evaluate the results through a cross-validation scheme with 10-folds. In order to evaluate the model's prediction accuracy, two metrics were chosen (MAE and MAPE) and the deviations of the predictions from the real results were represented as a regression scatterplot.

The second step was to extract, in fact, knowledge from the model just built, using the DSA, which expose the features that influence the days of absence the most, in percentage, as well as, how each feature manages to influence the modeled variable.

4. Results and discussion

4.1. Model evaluation

As described in Section 3.3 and illustrated in Figure 9, modeling performance was first measured using an evaluation scheme including a realistic 10-fold cross-validation procedure to test the model with unforeseen data.

In order to see the model in action, Table 3 combined three random absences, as well as, the worker's profile, that were used as input, with the predictions that came out of the model.

Table 3 - Prediction results for three absences

Feature	#1	#2	#3
Entity	Ent5	Ent5	Ent3
Contract	CEPF	Service commission	CEPF
ContractSpecs	Effective	Effective	Effective
WorkplaceLocation	Lisbon	Lisbon	Middle
AbsenceDesc	Sickness	Treatments	Day Off
Gender	Female	Female	Male
JobPosition	Technician	Director	Technician
AcademicQual	Degree	Degree	Highschool
LivingDistrict	North	Middle	Middle
PPAWorkDays	5198	12511	4043
MaritalStatus	Married	Married	Married
Nacionality	Portuguese	Portuguese	Portuguese
WorkHoursDay	8	7	6.5
Specificities	No Specificities	No Specificities	Associations Member
DisabilityPercent	0	0	0
DaysNoAbsences	2	26	69
TimesAbsentLastYear	12.2	2	21
DaysAbsentSinceLast	7.49	4.07	248
Age	39	59	34
TimesAbsentSameMotive	0	2	1
HaveChildren	No	No	Yes
AbsentAfterVacation	No	No	Yes
DaysAfterVacation	0	0	248
VacationDays	0	0	6
AbsenceDays	3	0.5	1
Days Predicted	2.06	0.56	1.05
Absolute deviation	0.94	0.01	0.05
% deviation	31.33	2.00	5.00

As stated in the previous section, MAE and MAPE are used to evaluate model's prediction accuracy (Table 4). Effectively, the SVM achieved an average absolute deviation of 0.26, which translates into a predicted value very close to the real days of absence. More so, from the range of 0.5 to 4.0 days, it predicted with an error of less than half a day (about 6 hours). Similarly, the MAPE metric resulted in 19.26%, which represents a low average predicted score deviation and allows to proceed to the knowledge extraction. Additionally, there have been other studies that extracted valid insightful knowledge from a model with a MAPE of around 27% (e.g., Moro et al., 2016).

Table 4 - Performance metrics

Metric	Result
MAE	0.26
MAPE	19.26%

Figure 10 displays the result of the 10-fold cross-validation method to test the model with unforeseen data. Since the data is chronologically ordered, with the passing of time and folds, the RFM's variables are getting richer every iteration, which is demonstrated by a substantial number of correct predictions at the last fold. Together with MAE and MAPE's scores, these results show how robust the model is and how good the conclusions can be.

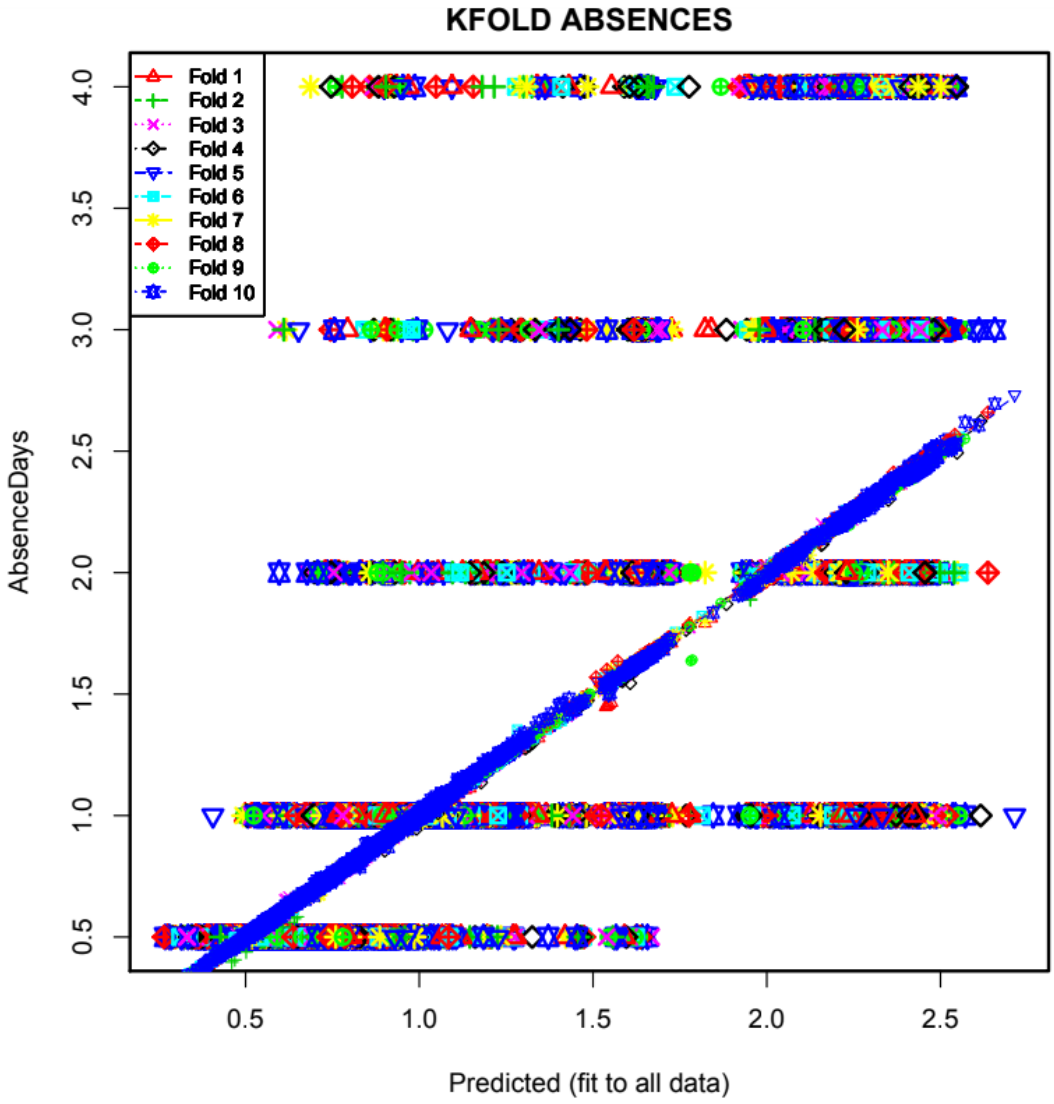


Figure 10 - K-fold model on the variable Absence Days

4.2. Knowledge extraction

This is stage is the one most important in the study, mostly because of its major role in contributing to understand how the cause and the worker's profile influences the days of absence. With that in mind, one can manage its resources to fill/predict the gap of a job position whose owner is absent. Therefore, understanding how long a worker will be absent can lead to leverage managerial decision support in the human resources department across the PPA.

DSA allowed to perceive to which extent the features that fed the SVM algorithm described the output variable. Figure 11 displays a graph of the most significant features' relevance, while Table 5 shows the percentage values for all the 25 features rounded to the hundredth.

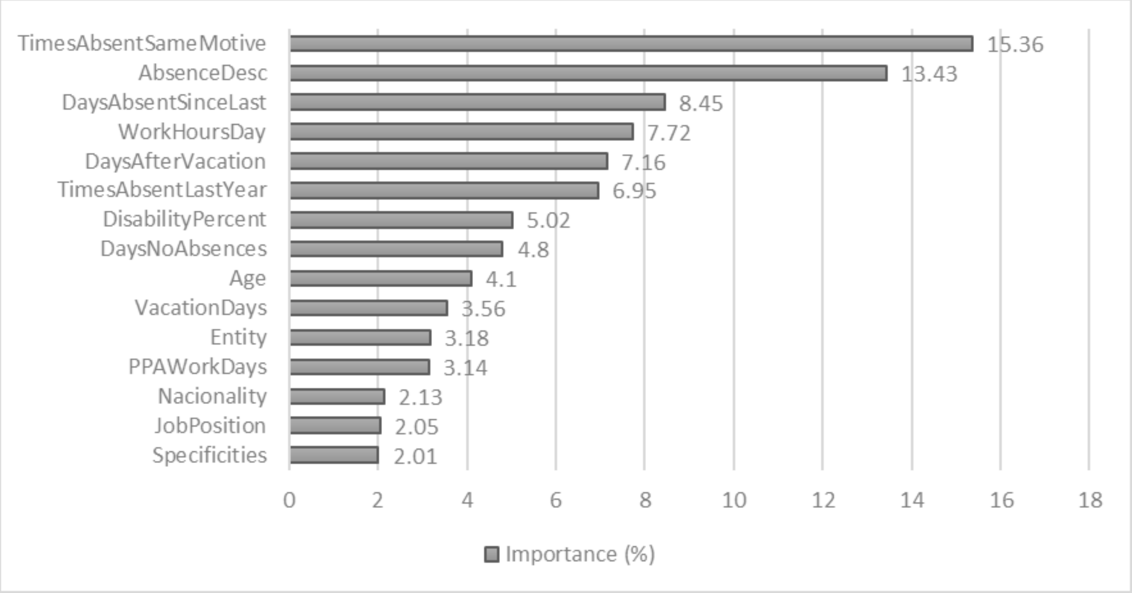


Figure 11 - Feature's relevance

Table 5 - Features' relevance

Feature	Relevance (%)
TimesAbsentSameMotive	15.36
AbsenceDesc	13.43
DaysAbsentSinceLast	8.45
WorkHoursDay	7.72
DaysAfterVacation	7.16
TimesAbsentLastYear	6.95
DisabilityPercent	5.02
DaysNoAbsences	4.80
Age	4.10
VacationDays	3.56
Entity	3.18
PPAWorkDays	3.14
Nacionality	2.13
JobPosition	2.05
Specificities	2.01
AcademicQual	1.85
MaritalStatus	1.83
LivingDistrict	1.76
Contract	1.63
ContractSpecs	1.39
Gender	1.21
HaveChildren	0.99
AbsentAfterVacation	0.29

After observing both Figure 11 and Table 6, it is visible that subject among the six most relevant features, collecting about 59% of the importance, are linked with the absenteeism profile of the worker, its motive and its recurrence, being the work hours per day the only exception. After those, there're no pattern for the relevance features, as there are a mix of the worker's characteristics, as well as, contract specifications and absenteeism records.

Effectively, the most relevant feature "TimesAbsentSameMotive", with 15.36% of the importance, was a feature computed based on the RFM methodology, as Miglautsch (2000) referred to it as a simple framework for quantifying customer behavior, which creates the possibility to apply RFM to other subjects, employees specifically, in order to understand their behavior and predict their actions.

The second feature in the list of most relevant is "AbsenceDesc", representing 13.43%, that indicates the motive of the absence, which was expected to have a big impact in the prediction of the duration of the absence (Bakker, Demerouti, de Boer, & Schaufeli, 2003).

Under the 10% relevant features, the following four variables came up with close results. Three of them were also computed based on the RFM procedure, "DaysAbsentSinceLast", "DaysAfterVacation" and "TimesAbsentLastYear", reinforcing the idea of using RFM concepts in topics like the absenteeism. The only one not fitting this pattern is the "WorkHoursDay", holding 7.72%, which goes along with the finding of the correlation between long work hours and reduced sickness absence, by Bernstrøm & Houkes (2017).

The next step after analyzing the relevance of features on consecutive absence days is to dive deeper into each of the most relevant ones, with relevance above 4.0%, in order to understand how they affect the absenteeism.

Although, it is important to mention that, due to high amount of records with one day of absence (Figure 8), collected in the dataset, most of the following graphs, with the impact of the most relevant features in the absenteeism, will not be possible to see past 2 days of consecutive absence.

Starting with the most important feature, it is noticeable the influence of being absent more than one time for the same reason (Figure 12). If it is the first time the worker is absent then it is expectable that its' next absence will be for more than one day. Although, if the worker keeps skipping work due to the same motive, the duration of the following absences shall decrease to less than a day. Although, these findings go against some other studies that stated that the odds of a new episode of sickness absence were 1.95 times higher for employees who already had such an episode compared with employees without sickness absence (Duijts, 2006; Laaksonen, He, & Pitkäniemi, 2013).

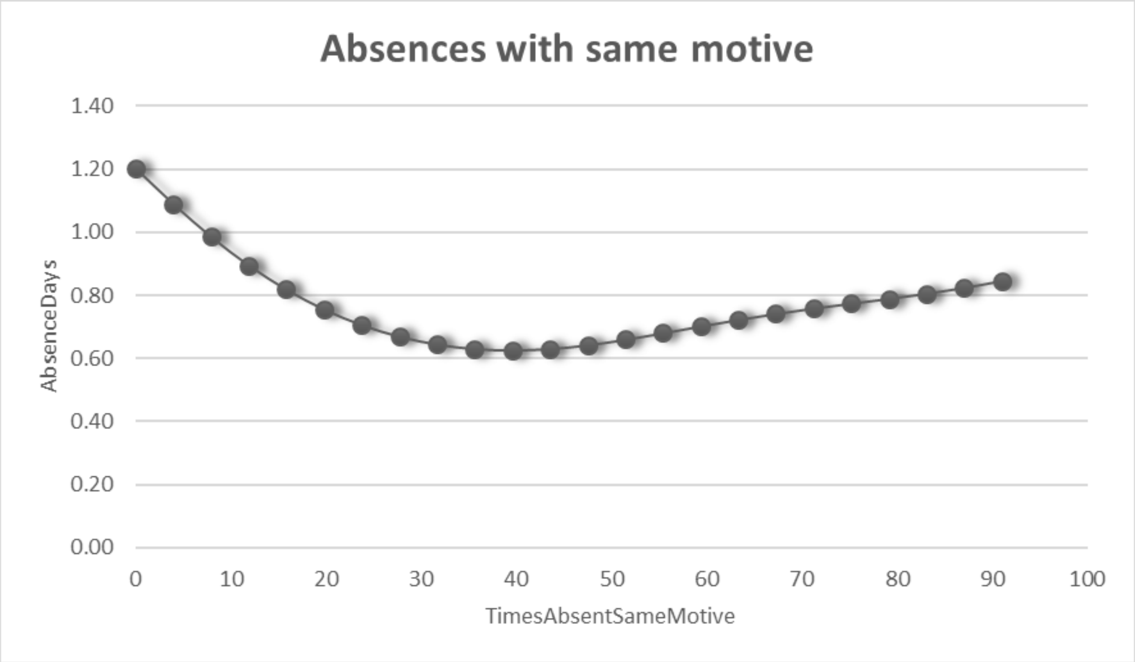


Figure 12 - Effect of motive recurrence in absenteeism

Figure 13 shows the different absences' motives and how many consecutive days the employee it is expected to miss work with that justification. Effectively, bereavement leave and sickness are motives that should concern the human capital department, because these indicate absences of two days or more, which can lead to loss of productivity. Sickness is also pointed by Joseph (2015) as one of the main reasons to the absenteeism in public sector. Family assistance should also be noticed, as it can get close to two consecutive days of absence. On the other hand, ambulatory care should lead to absences with a duration between half a day to one full day of work.

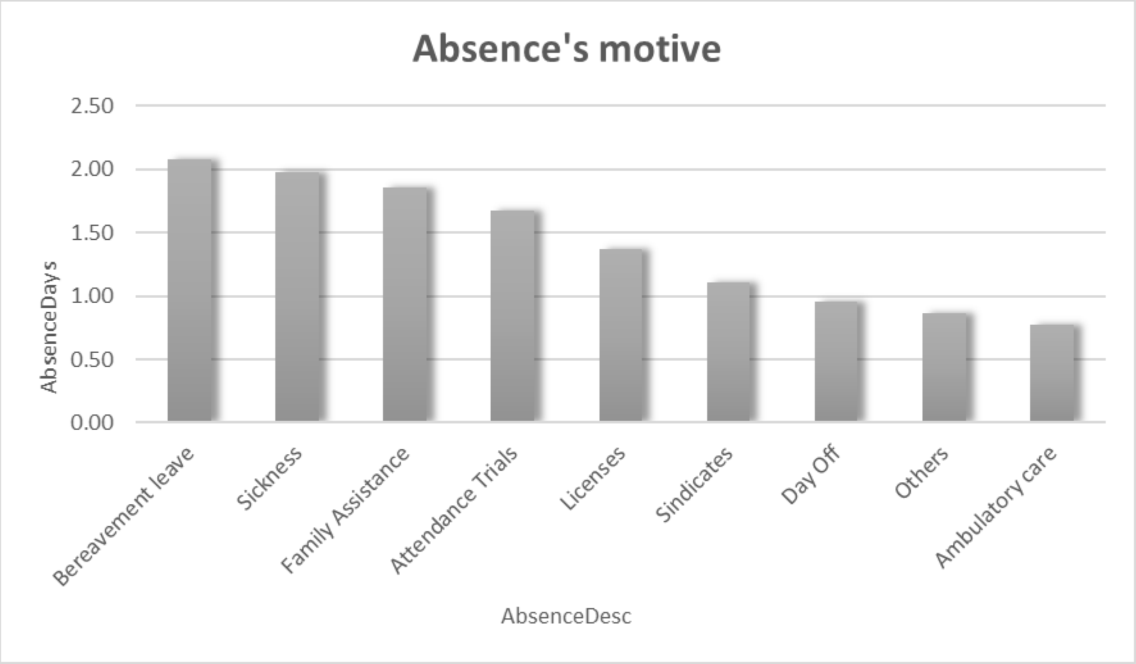


Figure 13 - Absenteeism's motives and their impact

The following feature in the list is "DaysAbsentSinceLast" that represents the amount of cumulative days of absence that the worker has been accumulating until the present absence and how does this number influence the duration of it. In Figure 14, it is clear that with the increasing number of days that the employee has been missing work, the duration of its absence is also longer. Same result was achieved by Roelen et al. (2011) in their study, which exhibited that employees with a history of prolonged sickness absence in past years were more likely to have above average sickness absence.

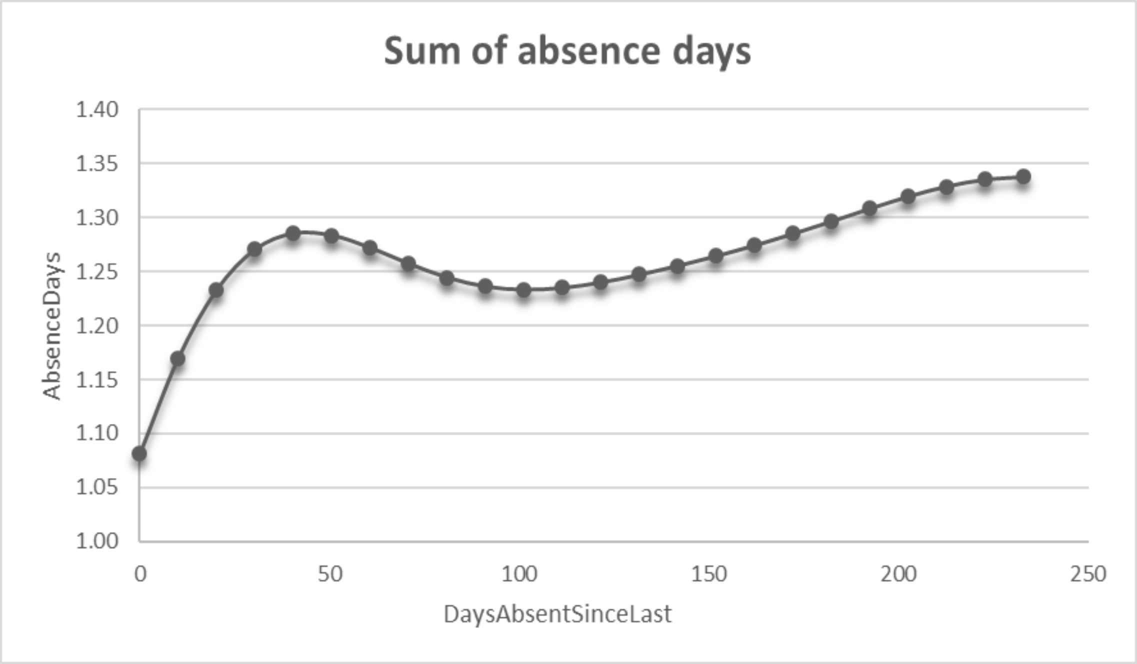


Figure 14 - Cumulative days of absence influence

Working time is also one of the most relevant features and as Figure 15 shows, working hours influence the duration of the absence. As matter of fact, it is possible to see the employees who works 5 hours per day tend to be absent for a longer time than the others. After analyzing the dataset, it is noticeable that female assistants, with a not effective service commission contract, are the ones with around 5 hours working time and so might be the ones that will be absent for an extended period of time. Markussen et al. (2011) stated that absenteeism generally rises with work hours, which is partially demonstrated by this study, i.e. until the 5 hours period the studies are in an agreement, but after that, this study shows the opposite, the decreasing of the absenteeism's duration.

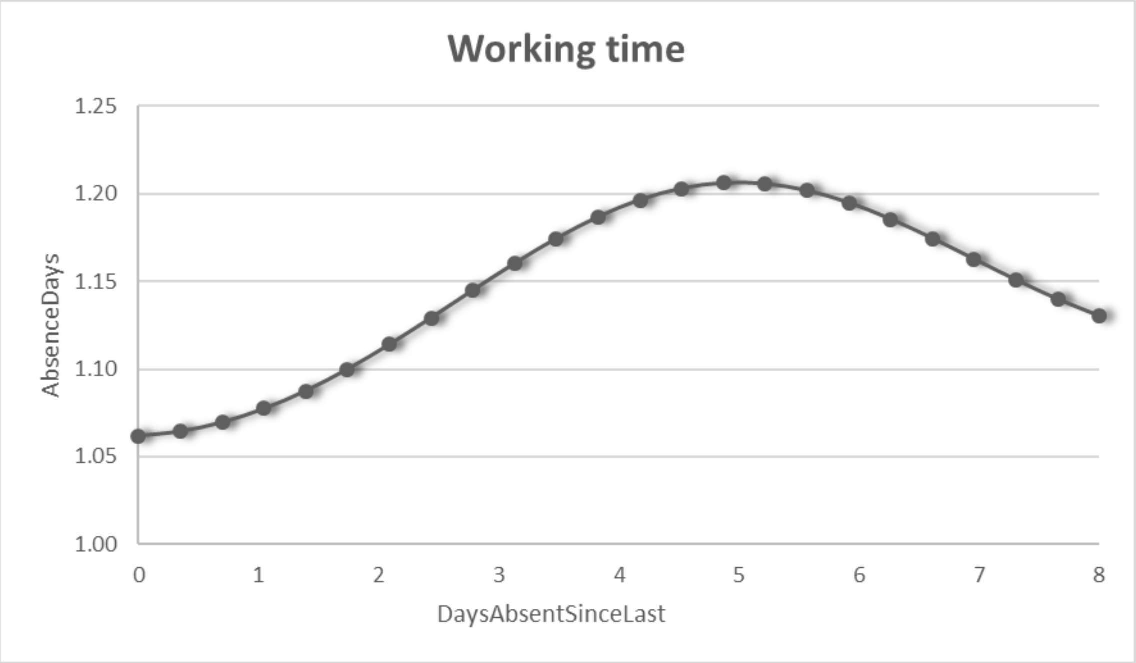


Figure 15 - Working time impact

Other variable computed by RFM methodology was "DaysAfterVacation", which allow the understanding if workers who do not go on vacations tend to be absent for a longer time that the ones that rather go on vacation now and then. Effectively, Figure 16 shows the effect that vacations does in the absenteeism and that is, employees who do not go on vacations for an extended period or a lot of times are leaning to be absent less time than the others. These results do not match with other studies (Westman & Etzion, 2001), but as De Bloom, Geurts and Kompier (2012) stated, the relation between vacation duration and the strength and endurance of vacation effects on employee health and well-being is still unclear, so it is also uncertain the correlation between vacations and absenteeism.

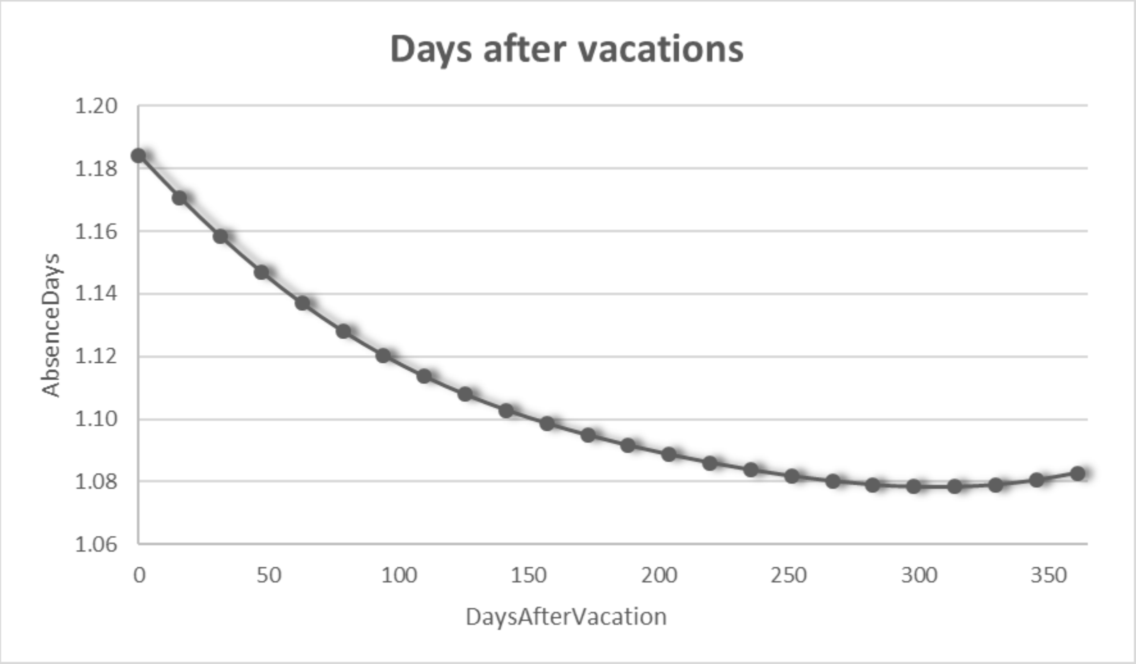


Figure 16 - Absence after vacations

Besides understanding the impact of the cumulative absence days ("DaysAbsentSinceLast"), it is also important to understand if those absences were for a long period time and short repetitions or short absences but multiple times through the year and that is when comes in hand the Figure 17. Actually, Figure 17 displays the influence of the number of times a worker has been absent in the duration of the next absence and it is visible that the employees who miss work a lot of times tend to do it for an extended period, stabilizing after the 30th time. Furthermore, this feature is connected to the "DaysAbsentSinceLast", since both contributes to longer periods of absence, which is also backed up by Roelen et al. (2011) study.

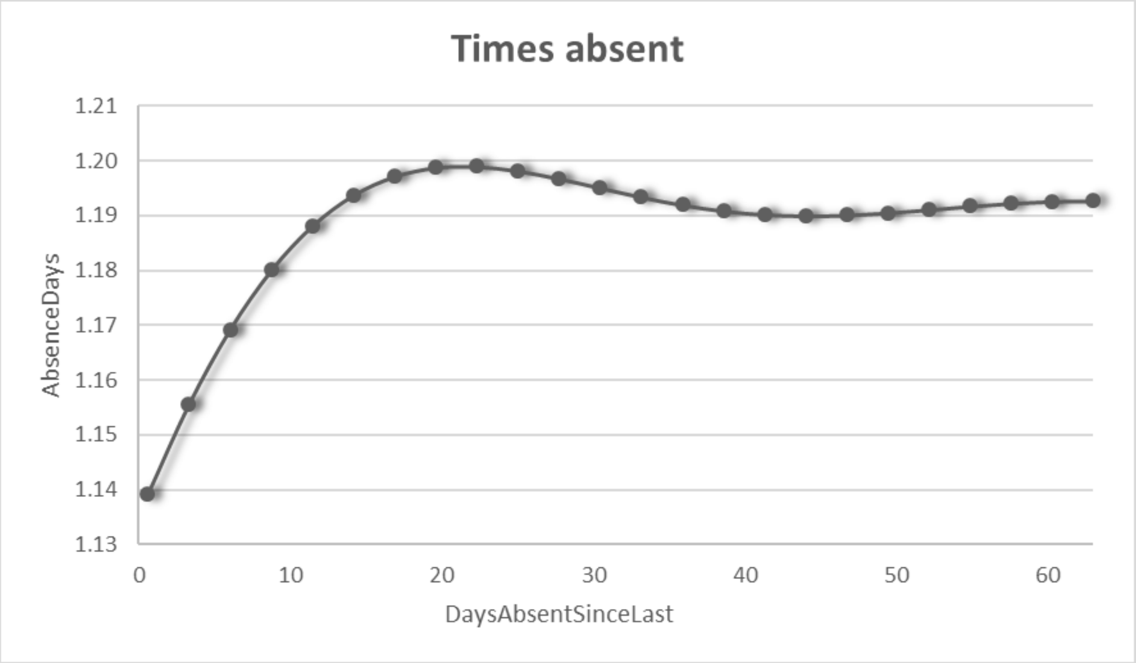


Figure 17 - Times absent

The most important feature based on the worker's profile is the disability percentage. Figure 18 shows its influence on the duration of the absence, until being considered as 30% disabled the duration of the absence is increasing getting to around 28.5 hours, but then it starts falling and when it comes to 100% of disability that number is around 27 hours, which is better than a worker with no disability.

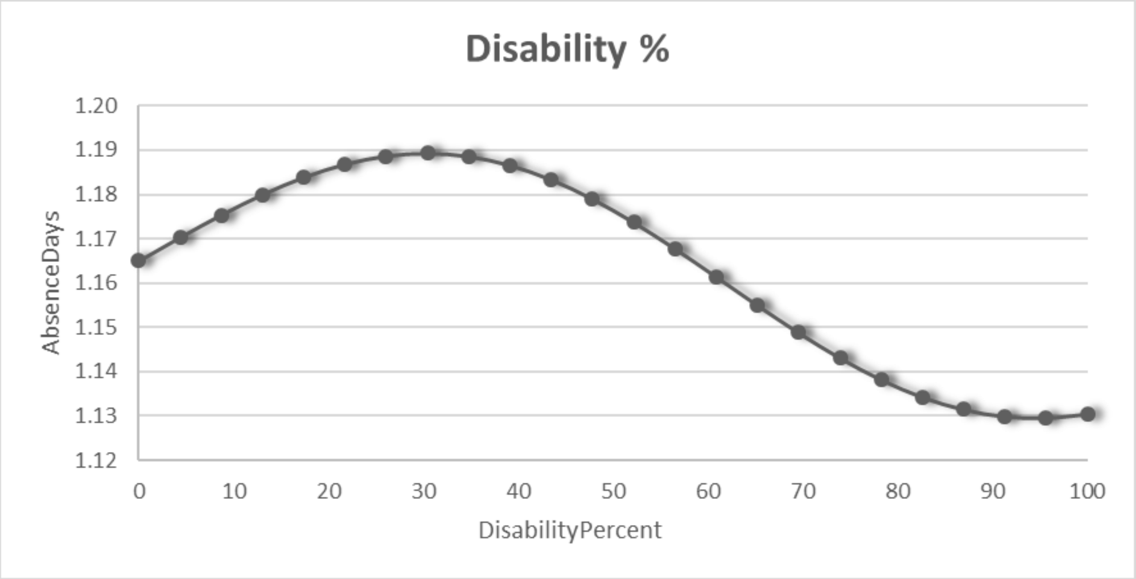


Figure 18 - Disability percentage

Following the disability percentage comes the number of days with no absences, other variable originated by the application of the RFM's methodology, and how it influences the duration of the worker's next absence. It is important to mention that the dataset has a calendar days base (365 days) and when counting the number of absences days, it might include weekends, because of the different types of schedules the workers can have, for example, shifts or additional work. After examining Figure 19, it is possible to understand that workers that do not get absent for a period of 140 days, or close to 5 months, are more likely to be absent for a longer duration, getting to an expected 1.22 days or around 30 hours of absence. Subsequently, the duration of the absence tends to get lower until the 360 days mark, which is the maximum length of no absences in the dataset, there the duration should stabilize and it is predictable that the next absence should have a length of 28 hours.

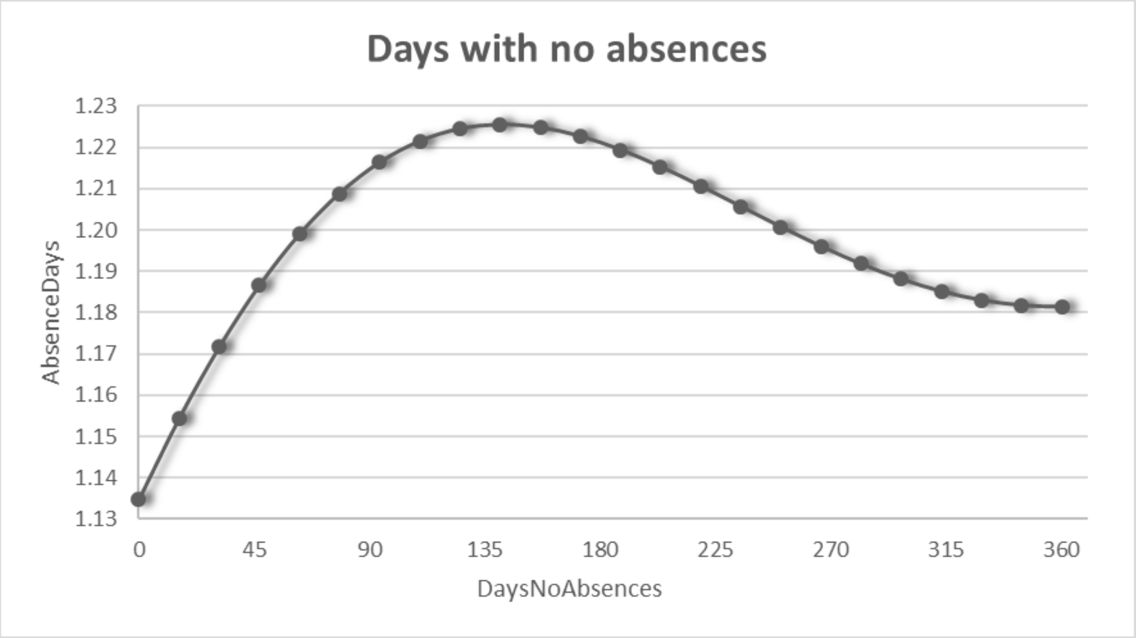


Figure 19 - Days with no absences

Finally, the last variable to comprehend its behavior is the age of the worker. it is pretty clear that the age has a direct impact in the absence duration and it is obvious that an older worker tends to be absent for a longer duration than a younger one, as shown in Figure 20. Effectively, there is a difference between a 25 years old worker and a 70-year-old one about their absence duration, which is 1.5 hours longer for the older one. These findings go along with the Markussen's et al. (2011) study, which shows that major disease absence rises monotonously and significantly with age.

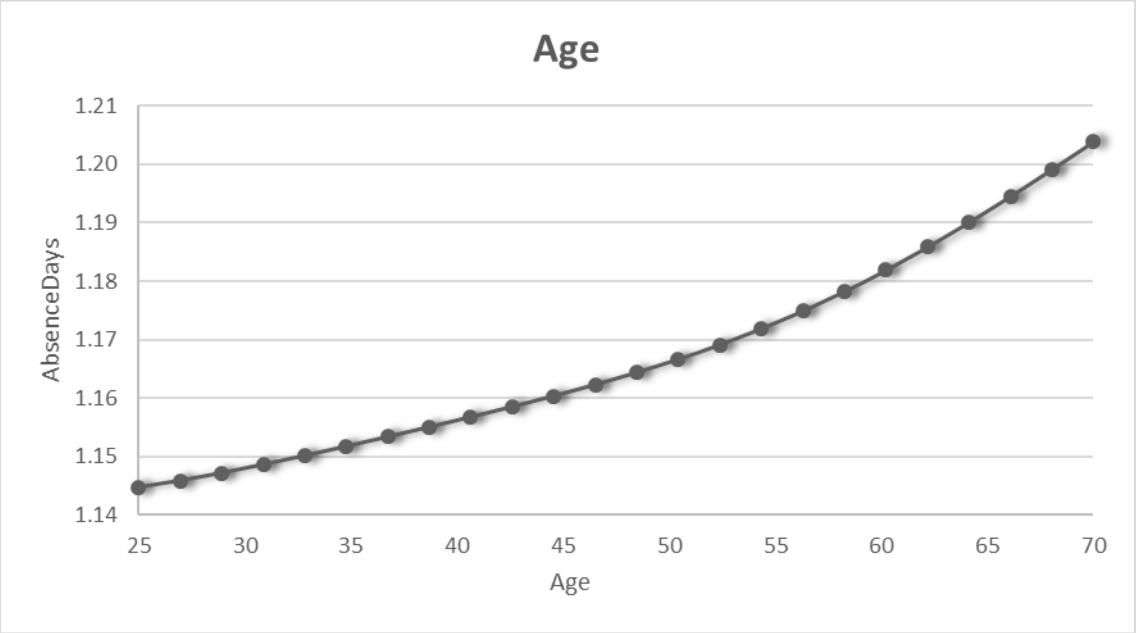


Figure 20 – Age

5. Conclusions

The lack of productivity should always be in the mind of the companies, so absenteeism must be understood and controlled in order to prevent flaws in the processes and, ultimately, profit loss. PPA is the most numerous organization in Portugal, with over 350 thousand workers, which, in 2017, held a 10.47% absenteeism rate (Estatística, 2018).

Absenteeism is influenced by innumerous factors but in this report the focus was on the motive of the absence, the worker's absenteeism history, its profile and its job specificities.

Effectively, the purpose of this study is to use those factors the next time the worker informs its superiors about needing to miss work in order to predict for how long the worker will be absent, which was already demonstrated as possible (Roelen et al., 2011) but not deepened enough for the PPA scenario, nor unfolded into the studied features. Providing that precious information to the HCM's department will help them make a better decision regarding the allocation of their workers and productivity management.

Furthermore, this study stands out by using the RFM methodology, mostly used in marketing studies, to compute new variables in order to evaluate the worker's absenteeism history and how it can influence its next absence. As Domingos (2012) pointed, feature engineering is a key task in any data-driven model.

Surprisingly, the RFM methodology had a big influence in this study, managing to get all its computed variables in the 25th most important features, especially considering that five out of six of them were in top 10 most important features, concealing around 43% of the total relevance, which opens the door to use this methodology in other field of study besides marketing.

After cleaning the dataset and removing the outliers, despite the high amount of records with one day of absence, it was run an SVM model to predict they days of absence that a worker would miss, based on the other features, which came out with very good results, getting a MAPE of 19.26% and a successful test using a 10-fold cross-validation scheme, meaning that the model can predict the duration of the next absence, up to 4 days, with an error inferior of a 5-hour period.

Achieving those result allowed to proceed with the knowledge extraction stage. Using the data-based sensitivity analysis, it was possible to unveil that the six most relevant features, gathering about 59% of the importance, which are linked, with the exception of work hours per day, by the absenteeism profile of the worker, i.e., absence's motive and its recurrence. After those, there's no visible pattern for the relevance features, as there are a mix of the worker's characteristics, as well as, contract specifications and absenteeism records.

The discovery that the features related to profile of the worker is less relevant than absence related features is quite interesting, mostly because it opens the door for a generic model that can be created without too much knowledge about the worker and, subsequently, can model can be generalized to all HR departments or all companies. This finding goes along with other studies' authors, that were already mentioned in this study, who used the past absences to predict future ones (Laaksonen et al., 2013; Reis et al., 2011; Roelen et al., 2011).

With so many features and results, it was felt the need to group up the information before discussing it. Table 6 shows the major findings of the study, relating the indicator (scenario of a feature) with the expected result in absence days (and working hours, using the average 6 hours per day).

Table 6 - Major findings

Indicator	**Expected** **(working hours = average 6 hours per day)**
Absence with same motive	Up to 1 day (5.1 hours)
Bereavement leave and sickness	2 days of absence (12 hours)
Ambulatory care	Less than 1 day (4.6 hours)
High number of absences days (233 days)	1.3 days of absence (8 hours)
5 working hours per day	1.2 days of absence (7.2 hours)
No vacations for a long time (361 days)	About 1 day (6.5 hours)
Absent regularly (22 times)	About 1.2 days (7.2 hours)
30% disability	1.19 days of absence (7.1 hours)
Higher disability % (over 30%)	1.13 days of absence (6.8 hours)

No absences for a long period (140 days)	1.22 days of absence (7.3 hours)
Older worker (70 years old)	1.2 days of absence (7.2 hours)

By further taking advantage of the sensitivity analysis, it was possible to perceive how each of the most relevant features affected the number of consecutive days of absence and some conclusions were drawn.

Firstly, if it is the first time an employee is absent for one of the motives, then it is expectable that it will miss work for more than one day per year. However, if the worker keeps skipping work due to the same motive, the duration of the absence shall decrease, which goes against some studies that demonstrate that having suffered a previous sick leave episode implies a significant increase in the risk of experiencing a subsequent one (Reis et al., 2011; Roelen et al., 2010). Although, it might be explained by the feeling of motive recurrence and the suspicion about the veracity of the justification, which is a limitation that has been mentioned in previous studies (M. Beil-Hildebrand, 1996), or because the employee feels like that being absent for a long time after missing work for the same motive might influence its "image" as employee (Mishali & Weiler, 2017).

Also, the motives bereavement leave and sickness tend to lead to a longer period of absence in PPA, followed by family assistance, which can get to almost two consecutive days of absence. Sickness is also mentioned as one of the main reasons to public sector's absenteeism (Jean & Guédé, 2015), where stress is highlighted as one of the main factors (George & K.A., 2015). Bereavement leave and family assistance are tied to the family aspect (Spetch, Howland, & Lowman, 2011). Among all health problems, ambulatory care shows itself as the least worrying motive with a duration between half a day to one full day of work for the following absence.

Effectively, the higher is the number of days that the employee missed work, longer will be the duration of its next absence, which goes along with Roelen et al. (2011) study about prolonged sickness absence in past years and its' impact in future absences.

Likewise, employees who work 5 hours per day will be absent for a longer time than the others, which, after analyzing the dataset, it is associated to female assistants, with a not effective service commission contract. This conclusion is on the same page as Markussen's et al. (2011) study until the 5 hour mark, but after that, this study shows the opposite, the decreasing of the absenteeism's duration with the rising of work hours.

Surprisingly, employees who do not go on vacations for an extended period have a tendency to be absent less time than the others. Taking a closer look to the dataset, it is possible to see that mainly assistants and technicians, who are effectives or in an equivalent regime, do not go on vacations that often and so are the ones who should be absent for shorter time. A bit of a controversial conclusion, as this discovery goes against some of other studies' results about the role of vacations in the absenteeism (Westman & Etzion, 2001), but it should not be discarded as the relation between vacation and the worker well-being is still unclear (De Bloom et al., 2012).

Not that surprising but workers who miss work a lot of times tend to do it for an extended period, although it tends to stabilize after the 30th absence. Furthermore, this finding is tied to the cumulative worker's absence days conclusion, since both contributes to longer periods of absence, which is also backed up by Roelen et al. (2011) study.

Additionally, PA employees with a 30% disability are the ones with a longer absence duration. Interestingly, the duration of an over 60% disabled PA worker's next absence is about 1.16 days (28 hours), which is shorter than the duration of a non-disabled worker's absence, as Kaye et al. (Kaye, Jans, & Jones, 2011) exposed in their study on why do not employers hire and retain workers with disabilities, there are clear stereotypes on people with disabilities about their poor performance and regular absenteeism, which, from the perspective of absenteeism, reveals to be inaccurate.

Moreover, workers that do not get absent for a period of 140 days, or close to 5 months, are more likely to be absent for a longer duration, a new finding enabled by the use of the RFM methodology (Vlasveld et al., 2012).

Lastly, an older worker tends to be absent for a longer duration than a younger one, which stands with the Markussen's et al. (2011) study, that shows age comes with a monotonous and significant rise of major disease absences.

Table 7 meets the study's findings with the literature review, summarizing authors points of view on the conclusions unveiled, i.e. identifies the studies whose conclusions are aligned, or not, with this study.

Table 7 - Conclusions review

Conclusion	Literature review	
	Approved	Objected
Absence with same motive, absence duration shall decrease	Beil-Hildebrand, 1996 Mishali & Weiler, 2017	Reis et al., 2011 Roelen et al., 2010
Most worrying motives: bereavement leave, sickness and family responsibilities	Jean & Guédé, 2015 George & Zakkariya, 2015 Spetch, Howland, & Lowman, 2011	Not applicable
High number of absences days, longer absence duration	Roelen et al., 2011	Not applicable
5 working hours per day, workers will be absent for a longer time	Markussen et al., 2011 (Partially)	Not applicable
No vacations for a long time, will be absent for a less time	De Bloom, Geurts, & Kompier, 2012	Westman & Etzion, 2001
Absent regularly, absences with extended periods	Roelen et al., 2011	Not applicable
Higher disability workers tend to be absent for a shorter time than a non-disabled	Kaye et al., 2011	Not applicable
No absences for a long period, extended duration of the next absence	Vlasveld et al., 2012	Not applicable
Older workers tend to be absent for a longer time	Markussen et al., 2011	Not applicable

Interestingly, the most concerning employee profile is an older, 30% disabled, worker with a schedule of 5 work hours per day, who just came out of vacation, that mentions that will miss work because it's feeling sick for the first time, but already missed work a lot of times and for an extended period of time, over 140 days ago.

To prevent absenteeism itself there are a lot of studies with solutions or that proven that a policy/reform had a positive impact in it (De Paola et al., 2014; Gosselin, Lemyre, & Corneil, 2013; Kocakülâh et al., 2016).

With these thoughts in mind, this study gives the public sector red flags for workers and their absences justifications, more so the HR department can now have a better perception for how long their employees will be missing work and act accordingly, including allocate other workers or subcontract to fill the gaps left by those absences.

Even though the results of the study were quite impressive, there are some limitations that should be mentioned as an opportunity to future researches. On one hand, the veracity of the absences' motives could not be proven, so if in future research there is a solution for this problem, the robustness of this study could be improved. On the other hand, the amount of records of one absence day were disproportional to the other records. So, it would be better to get a dataset with more variety of absence days records.

For a future research, it would be interesting to add continuity to this study, i.e. add more years to the dataset in order to update the data and refine the model. It would also be very valuable to cross absenteeism in the private sector with the results obtained in this study, so that it would be possible to know if there is a common model to be applied to both sectors or if each one needs its own.

References

Ahn, H., Kim, K. J., & Man, I. (2006). Global optimization of feature weights and the number of neighbors that combine in a case-based reasoning system. *Expert Systems*, *23*(5), 290–301. https://doi.org/10.1111/j.1468-0394.2006.00410.x

AICEP. (2018). *Portugal Principais Indicadores Económicos*. Retrieved from http://www.portugalglobal.pt/PT/Biblioteca/Economianet/PortugalIndicadoresEconomicos.pdf

Ala-Mursula, L., Vahtera, J., Kivimaki, M., Kevin, M., & Pentti, J. (2002). Employee control over working times: Associations with subjective health and sickness absences. *Journal of Epidemiology and Community Health*, *56*(4), 272–278. https://doi.org/10.1136/jech.56.4.272

AlAzzam, M., AbuAlRub, R. F., & Nazzal, A. H. (2017). The Relationship Between Work–Family Conflict and Job Satisfaction Among Hospital Nurses. *Nursing Forum*, *52*(4), 278–288. https://doi.org/10.1111/nuf.12199

Anafarta, N. (2011). The Relationship between Work-Family Conflict and Job Satisfaction: A Structural Equation Modeling (SEM) Approach. *International Journal of Business and Management*, *6*(4), 168–177.

Armstrong, M. author. (2014). *Armstrong's handbook of human resource management practice* (13th ed.). New York: Kogan Page.

Bakker, A. B., Demerouti, E., de Boer, E., & Schaufeli, W. B. (2003). Job demand and job resources as predictors of absence duration and frequency. *Journal of Vocational Behavior*. https://doi.org/10.1016/S0001-8791(02)00030-1

Beil-Hildebrand, M. (1996). Nurse absence—the causes and the consequences. *Journal of Nursing Management*, *4*(1), 11–17. https://doi.org/10.1111/j.1365-2834.1996.tb00022.x

Beil-Hildebrand, M. (1996, January). Nurse absence-the causes and the consequences. *Journal of Nursing Management*. https://doi.org/10.1111/j.1365-2834.1996.tb00022.x

Bernik et al., M. (2007). Using information technology for human resource management decisions. *8th WSEAS International*, 130–133.

Bernstrøm, V. H., & Houkes, I. (2017). A systematic literature review of the relationship between work hours and sickness absence. *Work & Stress*, *32*(1), 1–21. https://doi.org/10.1080/02678373.2017.1394926

Boswell, W., Boudreau, J. W., & Tichy, J. (2005). The relationship between employee job change and job satisfaction: The honeymoon-hangover effect. *Journal of Applied Psychology*, *90*(5), 882–892. https://doi.org/10.1037/0021-9010.90.5.882

Calvano, L. (2013). Tug of War: Caring for Our Elders While Remaining Productive at Work. *Academy of Management Perspectives*, *27*(3), 204–218. https://doi.org/10.5465/amp.2012.0095

Carvalho, T., & Bruckmann, S. (2014). Reforming Portuguese Public Sector: A Route from Health to Higher Education. In *Reforming Higher Education Public Policy Design and Implementation* (Vol. 41, pp. 83–102). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-7028-7

Casey, P., & Grzywacz, J. (2008). Employee Health and Well-Being: The Role of Flexibility and Work–Family Balance. *The Psychologist-Manager Journal*, *11*(1), 31–47. https://doi.org/10.1080/10887150801963885

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76. https://doi.org/10.1109/ICETET.2008.239

Chatterji, M. (2002). Sickness, absenteeism, presenteeism, and sick pay. *Oxford Economic Papers*, *54*(4), 669–687. https://doi.org/10.1093/oep/54.4.669

Chen, Han, J., Yu, P., & Han, J. (1996). Data Mining: An Overview from Database Perspective. *IEEE Transactions on Knowledge and Data Engineering*, *Vol. 8, No,* 866–883.

Chen, R.-S., Wu, R.-C., & Chen, J. Y. (2005). Data mining application in customer relationship management of credit card business. *29th Annual International Computer Software and Applications Conference (COMPSAC'05)*, *2*, 39–40. https://doi.org/10.1109/compsac.2005.67

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, *36*(3 PART 1), 4176–4184. https://doi.org/10.1016/j.eswa.2008.04.003

Chien, C. F., & Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications*, *34*(1), 280–290. https://doi.org/10.1016/j.eswa.2006.09.003

Cleary, D. (2011). Predictive Analytics in the Public Sector: Using Data Mining to Assist Better Target Selection for Audit. *Electronic Journal of E-Government*, *9*(2), 132–140.

Cortez, P., & Embrechts, M. J. (2011). Opening black box Data Mining models using Sensitivity Analysis. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (pp. 341–348). https://doi.org/10.1109/CIDM.2011.5949423

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*, 1–17. https://doi.org/10.1016/j.ins.2012.10.039

Cowan, A. M. (2002). Data Mining in Finance: Advances in Relational and Hybrid Methods. *International Journal of Forecasting*, *18*(1), 155–156. https://doi.org/10.1016/S0169-2070(01)00128-5

De Bloom, J., Geurts, S. A. E., & Kompier, M. A. J. (2012). Effects of short vacations, vacation activities and experiences on employee health and well-being. *Stress and Health*, *28*(4), 305–318. https://doi.org/10.1002/smi.1434

De Paola, M., Scoppa, V., & Pupo, V. (2014). Absenteeism in the Italian Public Sector: The Effects of Changes in Sick Leave Policy. *Journal of Labor Economics*, *32*(2), 337–360. https://doi.org/10.1086/674986

DGAEP - Direção-Geral da Administração e do Emprego Público. (2013). Análise da evolução das estruturas da administração pública central portuguesa decorrente do PRACE e do PREMAC, 1–123.

DGAEP - Direção-Geral da Administração e do Emprego Público. (2017a). *Síntese estatística do emprego público*.

DGAEP - Direção-Geral da Administração e do Emprego Público. (2017b). Síntese Estatística do Emprego Público (SIEP). Retrieved December 17, 2017, from https://www.dgaep.gov.pt/index.cfm?OBJID=ECA5D4CB-42B8-4692-A96C-8AAD63010A54

DGAEP - Direção-Geral da Administração e do Emprego Público. (2018). Síntese Estatística do Emprego Público (SIEP). Retrieved July 1, 2018, from https://www.dgaep.gov.pt/upload/DEEP/SIEP1T2018/DGAEP-DEEP_SIEP_2018T1_15052018.pdf

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal Of Business & Economic Statistics*, *20*(1), 134–144. https://doi.org/Doi 10.1198/073500102753410444

Domingos, P. (2012). A few useful things to know about machine learning.

Communications of the ACM, *55*(10), 78. https://doi.org/10.1145/2347736.2347755

Duijts, S. F. A. (2006). Prediction of sickness absence: development of a screening instrument. *Occupational and Environmental Medicine*, *63*(8), 564–569. https://doi.org/10.1136/oem.2005.024521

Durairaj, M., & Ranjani. (2013). Data Mining Applications in Healthcare: A Study. *International Journal of Scientific & Technology Research*, *2*(10), 29–35. Retrieved from www.ijstr.org

Dybå, T., & Dingsøyr, T. (2008). Empirical studies of agile software development: A systematic review. *Information and Software Technology*, *50*(9–10), 833–859. https://doi.org/10.1016/j.infsof.2008.01.006

Eastont, F. F., & Goodale, J. C. (2005). Schedule recovery: Unplanned absences in service operations. *Decision Sciences*, *36*(3), 459–488. https://doi.org/10.1111/j.1540-5414.2005.00080.x

Estatística, I. N. de. (2018). Balanço Social 2017. Retrieved from https://www.dgaep.gov.pt/upload//Instrumentos_Gestao/2017/BS_2017.pdf

George, E., & K.A., Z. (2015). Job related stress and job satisfaction: a comparative study among bank employees. *Journal of Management Development*, *34*(3), 316–329. https://doi.org/10.1108/JMD-07-2013-0097

Golden, L. (2011). The effects of working time on productivity and firm performance : a research synthesis paper. *Conditions of Work and Employment Series*, (33), 1–34.

Gosselin, E., Lemyre, L., & Corneil, W. (2013). Presenteeism and absenteeism: Differentiated understanding of related phenomena. *Journal of Occupational Health Psychology*, *18*(1), 75–86. https://doi.org/10.1037/a0030932

Guenther, N., & Schonlau, M. (2016). Support vector machines. *Stata Journal*, *16.4*, 917–937.

Han, J.Kamber, M.Pei, J. (2012). *Data mining: concepts and techniques*. (Morgan Kaufmann Publisher, Ed.) (3rd ed.). Boston: Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed., Vol. 27). New York: Springer-Verlag. https://doi.org/10.1007/b94608

Haswell, M. (2003). Dealing with Employee Absenteeism. *Management Services*, *47*(12), 16.

Hauge, L. J., Skogstad, A., & Einarsen, S. (2010). The relative impact of workplace bullying as a social stressor at work. *Scandinavian Journal of Psychology*, *51*(5), 426–433. https://doi.org/10.1111/j.1467-9450.2010.00813.x

Hughes, J., & Bozionelos, N. (2007). Work-life balance as source of job dissatisfaction and withdrawal attitudes. *Personnel Review*, *36*(1), 145–154. https://doi.org/10.1108/00483480710716768

INA - Direção-Geral da Qualificação dos Trabalhadores em Funções Públicas. (2017a). Fim do regime de requalificação. Retrieved December 17, 2017, from https://www.ina.pt/index.php/mobilidade-noticias/1648-fim-do-regime-da-requalificacao

INA - Direção-Geral da Qualificação dos Trabalhadores em Funções Públicas. (2017b). O que é a Mobilidade? Retrieved December 17, 2017, from https://www.ina.pt/index.php/mobilidade/2012-11-30-11-56-06/o-que-e-a-mobilidade-geral

INA - Direção-Geral da Qualificação dos Trabalhadores em Funções Públicas. (2017c). O que é a Valorização? Retrieved December 17, 2017, from https://www.ina.pt/index.php/mobilidade/valorizacao-profissional/o-que-e-a-valorizacao-professional

Jantan, H. (2009). Knowledge discovery techniques for talent forecasting in human resource application. *World Academy of Science, Engineering and Technology, 3*(2), 178–186. https://doi.org/10.4018/jtd.2010100103

Jantan, H., Hamdan, A., & Othman, Z. (2010). Human Talent Forecasting using Data Mining Classification Techniques. *International Journal of Technology Diffusion, 1*(4), 29–41. https://doi.org/10.4018/jtd.2010100103

Jean, U., & Guédé, L. (2015). Gauging the Issue of Absenteeism in the Workplace : Evidence from the Public. *International Journal of Business and Social Science, 6*(2), 65–71.

Jordan, J., Gurr, E., Tinline, G., Giga, S. I., Faragher, B., & Cooper, C. L. (2003). Beacons of Excellence in Stress Prevention: Research Report 133, 194. Retrieved from http://bradscholars.brad.ac.uk:8080/handle/10454/4056

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms* (2nd ed.). Totowa: Wiley-IEEE Press. https://doi.org/10.1002/9781118029145

Katz, R., Lowenstein, A., Prilutzky, D., & Halperin, D. (2011). Employers' knowledge and attitudes regarding organizational policy toward workers caring for aging family members. *Journal of Aging and Social Policy, 23*(2), 159–181. https://doi.org/10.1080/08959420.2011.554120

Kaye, H. S., Jans, L. H., & Jones, E. C. (2011). Why don't employers hire and retain workers with disabilities? *Journal of Occupational Rehabilitation, 21*(4), 526–536. https://doi.org/10.1007/s10926-011-9302-8

Kelly, E., Kossek, E., Hammer, L., Durham, M., Bray, J., Chermack, K., … Kaskubar, D. (2008). 7 Getting There from Here: Research on the Effects of Work–Family Initiatives on Work–Family Conflict and Business Outcomes. *The Academy of Management Annals*, *2*(1), 305–349. https://doi.org/10.1080/19416520802211610

Kim, J., & Garman, E. T. (2003). Financial Stress and Absenteeism: An Empirically Derived Model. *Journal of Financial Counseling and Planning*, *14.1: 31.*

Kocakülâh, M., Kelley, A., Mitchell, K., & Ruggieri, M. (2016). Absenteeism Problems And Costs: Causes, Effects And Cures. *International Business & Economics Research Journal (IBER)*, *15*(3), 89. https://doi.org/10.19030/iber.v15i3.9673

Kossek, E., & Hammer, L. (2008). Supervisor work/life training gets results. *Harvard Business Review, 86*(11), 36.

Kotsadam, A. (2011). Does Informal Eldercare Impede Women's Employment? The Case of European Welfare States. *Feminist Economics*, *17*(2), 121–144. https://doi.org/10.1080/13545701.2010.543384

Kusiak, A., Kernstine, K. H., Kern, J. A., McLaughlin, K. a, & Tseng, T. L. (2000). Data mining: medical and engineering case studies. *Industrial Engineering Research Conference*, 1–7.

Kwon, O., & Lee, N. (2011). A relationship-aware methodology for context-aware service selection. *Expert Systems*, *28*(4), 375–390. https://doi.org/10.1111/j.1468-0394.2010.00548.x

Laaksonen, M., He, L., & Pitkäniemi, J. (2013). The durations of past sickness absences predict future absence episodes. *Journal of Occupational and Environmental Medicine, 55*(1), 87–92. https://doi.org/10.1097/JOM.0b013e318270d724

Liu, D.-R., & Shih, Y.-Y. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information & Management, 42*(3), 387–400. https://doi.org/10.1016/j.im.2004.01.008

Madureira, C. (2015). A reforma da Administração Pública Central no Portugal democrático: do período pós-revolucionário à intervenção da troika. *Revista de Administração Pública*, *49*(3), 547–562. https://doi.org/10.1590/0034-7612129503

Markussen, S., Røed, K., Røgeberg, O. J., & Gaure, S. (2011). The anatomy of absenteeism. *Journal of Health Economics*, *30*(2), 277–292. https://doi.org/10.1016/j.jhealeco.2010.12.003

McHugh, M. (2001). Employee absence: an impediment to organisational health in local government. *International Journal of Public Sector Management*, *14*(1), 43–58. https://doi.org/10.1108/09513550110387066

McLeod, R., & Sanctis, G. De. (1995). A Resource-Flow Model of the Human Resource Information System. *Journal of Information Technology Management*, (No. 3), 1–15.

Miglautsch, J. R. (2000). Thoughts on RFM scoring. *Journal of Database Marketing & Customer Strategy Management*, *8*(1), 67–72. https://doi.org/10.1057/palgrave.jdm.3240019

Mishali, M., & Weiler, D. (2017). Psychological factors causing nonadherence to safety regulations in Israel's stone and marble fabrication industry: Unveiling the source of worker noncompliance. *Cogent Business and Management*. https://doi.org/10.1080/23311975.2017.1404717

Moro, S., Cortez, P., & Laureano, R. M. S. (2011). Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM methodology. *European Simulation and Modelling Conference*, (Figure 1), 117–121.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

Moro, S., Cortez, P., & Rita, P. (2015). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. *Neural Computing and Applications*, *26*(1), 131–139. https://doi.org/10.1007/s00521-014-1703-0

Moro, S., Rita, P., & Oliveira, C. (2018). Factors Influencing Hotels' Online Prices. *Journal of Hospitality Marketing and Management*, *27*(4), 443–464. https://doi.org/10.1080/19368623.2018.1395379

Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, *69*(9), 3341–3351. https://doi.org/10.1016/j.jbusres.2016.02.010

OECD. (2017). *Goverment at Glance 2017*. *OECD Publishing*. https://doi.org/10.1787/gov_glance-2017-en

Osterkamp, R., & Rohn, O. (2007). Being on Sick Leave: Possible Explanations for Differences of Sick-leave Days Across Countries. *CESifo Economic Studies*, *53*(1), 97–114. https://doi.org/10.1093/cesifo/ifm005

Painter, M. (2010). *Tradition and public administration*. (B. Peters, Ed.) (1st ed.). London: Palgrave Macmillan. https://doi.org/10.1057/9780230289635

Payne, S., Cook, A., & Diaz, I. (2012). Understanding childcare satisfaction and its effect on workplace outcomes: The convenience factor and the mediating role of work-family conflict. *Journal of Occupational and Organizational Psychology*, *85*(2), 225–244. https://doi.org/10.1111/j.2044-8325.2011.02026.x

Perry, J. L., Hondeghem, A., & Wise, L. (2010). Revisiting the Motivational Bases of Public Service Motivation: Twenty Years of Research and an Agenda for the Future. *Public Administration Review*, *70*(5), 681–690. https://doi.org/10.1111/j.1540-6210.2010.02196.x

Peters, B., Pierre, J., & Randma-Liiv, T. (2010). Global Financial Crisis, Public Administration and Governance: Do New Problems Require New Solutions? *Public Organization Review*, *11*(1), 13–27. https://doi.org/10.1007/s11115-010-0148-x

Pilkey, O. H., & Pilkey-Jarvis, L. (2009). *Useless arithmetics – why environmental scientists can't predict the future*. Columbia University Press.

PORDATA. (2017a). Administrações Públicas: despesas, receitas e défice/excedente (base=2011). Retrieved December 16, 2017, from https://www.pordata.pt/MicroPage.aspx?DatabaseName=Portugal&MicroName=Administrações+Públicas+despesas++receitas+e+défice+excedente+(base+2011)&MicroURL=2784&

PORDATA. (2017b). População residente: total e por sexo. Retrieved December 16, 2017, from https://www.pordata.pt/Portugal/População+residente+total+e+por+sexo-6

PORDATA. (2017c). Receitas de impostos do Estado em % do PIB. Retrieved December 16, 2017, from https://www.pordata.pt/MicroPage.aspx?DatabaseName=Portugal&MicroName=Receitas+de+impostos+do+Estado+em+percentagem+do+PIB&MicroURL=2773&

PORDATA. (2017d). Taxa de desemprego: total e por sexo (%). Retrieved December 16, 2017, from https://www.pordata.pt/Portugal/Taxa+de+desemprego+total+e+por+sexo+(percentagem)-550

Portuguesa, A. da R. (2005). Constituição da república portuguesa. *Assembleia Da República Portuguesa*, 91. https://doi.org/10.1017/CBO9781107415324.004

Portuguesa, A. da R. (2016). *Código do Trabalho*. (S. A. Porto Editora, Ed.) (13th ed.).

Portuguesa, A. da R. (2017a). Lei 35.

Portuguesa, A. da R. (2017b). *Lei geral do trabalho em funções públicas*. (S. A. Porto Editora, Ed.) (06-2014th ed.).

Possenriede, D. (2011). The effects of flexible working time arrangemets on absenteeism – the Dutch case. *Utrecht, Netherlands: Utrecht University School of Economics*, (February).

Presidency, M. of the. (2010). *Public Employment in European Union Member States*. (I. G. Valera, L. García-Manzano, P. Guerra, G. Carlos, E. Matas, A. Gómez Bada, & E. Grávalos, Eds.) (1st ed.). Madrid: Ministry of the Presidency. Technical Secretariat-General.

Primeiro-Ministro, G. do. (2017). Despacho n.º 3772/2017. *Diário Da República*.

Quinley, K. M. (2003). EAPs: A benefit that can trim your disbility and absenteeism costs. *Compensation & Benefits Report*, *17*(2), 6–7.

Reis, R., Utzet, M., La Rocca, P., Nedel, F., Martín, M., & Navarro, A. (2011). Previous sick leaves as predictor of subsequent ones. *International Archives of Occupational and Environmental Health*, *84*(5), 491–499. https://doi.org/10.1007/s00420-011-0620-0

Roelen et al., C. (2010). Recurrence of medically certified sickness absence according to diagnosis: A sickness absence register study. *Journal of Occupational Rehabilitation*, *20*(1), 113–121. https://doi.org/10.1007/s10926-009-9226-8

Roelen et al., C. (2011). The history of registered sickness absence predicts future sickness absence. *Occupational Medicine*, *61*(2), 96–101. https://doi.org/10.1093/occmed/kqq181

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2002). Global Sensitivity Analysis for Importance Assessment. In *Sensitivity Analysis in Practice* (pp. 31–61). https://doi.org/10.1002/0470870958.ch2

Schaufeli, W. B., Bakker, A. B., & van Rhenen, W. (2009). How changes in job demands and resources predict burnout, work engagement, and sickness absenteeism. *Journal of Organizational Behavior*, *30*(7), 893–917. https://doi.org/10.1002/job.595

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, *12*(1), 217–222.

Silva, A. (2016). Unveiling the Features of Successful Ebay Sellers of Smartphones – A Data Mining Sales Predictive Model.

Silva, A., Moro, S., Rita, P., & Cortez, P. (2018). Unveiling the features of successful eBay smartphone sellers. *Journal of Retailing and Consumer Services*, *43*, 311–324. https://doi.org/10.1016/j.jretconser.2018.05.001

Spetch, A., Howland, A., & Lowman, R. L. (2011). EAP Utilization patterns and employee absenteeism: Results of an empirical, 3-year longitudinal study in a national canadian retail corporation. *Consulting Psychology Journal*, *63*(2), 110–128. https://doi.org/10.1037/a0024690

Stansfeld, S., & Candy, B. (2006). Psychosocial work environment and mental health - A meta-analytic review. *Scandinavian Journal of Work, Environment and Health*, *32*(6), 443–462. https://doi.org/10.5271/sjweh.1050

Troika. (2011). *Portugal - Memorandum Of Understanding*.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. *Springer*. https://doi.org/10.1109/TNN.1997.641482

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031

Vignoli, M., Guglielmi, D., Bonfiglioli, R., & Violante, F. (2016). How job demands affect absenteeism? The mediating role of work–family conflict and exhaustion. *International Archives of Occupational and Environmental Health*, *89*(1), 23–31. https://doi.org/10.1007/s00420-015-1048-8

Vlasveld, M. C., Van Der Feltz-Cornelis, C. M., Bültmann, U., Beekman, A. T. F., Van Mechelen, W., Hoedeman, R., & Anema, J. R. (2012). Predicting return to work in workers with all-cause sickness absence greater than 4 weeks: A prospective cohort study. *Journal of Occupational Rehabilitation*, *22*(1), 118–126. https://doi.org/10.1007/s10926-011-9326-0

Westman, M., & Etzion, D. (2001). The impact of vacation and job stress on burnout and absenteeism. *Psychology & Health*, *16*(5), 595–606. https://doi.org/10.1080/08870440108405529

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959), 29–39. https://doi.org/10.1.1.198.5133

Witten, I. H., Frank, E., & Hall, M. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Elsevier.

Ybema, J. F., Smulders, P. G. W., & Bongers, P. M. (2010). Antecedents and consequences of employee absenteeism: A longitudinal perspective on the role of job satisfaction and burnout. *European Journal of Work and Organizational Psychology*, *19*(1), 102–124. https://doi.org/10.1080/13594320902793691

Yun, Y., Sim, J., Park, E.-G., Park, J. D., & Noh, D.-Y. (2016). Employee Health Behaviors, Self-Reported Health Status, and Association With Absenteeism. *Journal of Occupational and Environmental Medicine*, *58*(9), 932–939.

Ziebarth, N. (2013). Long-term absenteeism and moral hazard-Evidence from a natural experiment. *Labour Economics*, *24*, 277–292. https://doi.org/10.1016/j.labeco.2013.09.004