Hospital bed management support using regression data mining models

Sérgio Oliveira[1], Filipe Portela[1], Manuel F. Santos[1], José Machado[2], António Abelha[2]
[1]Algoritmi Centre, University of Minho, Guimarães, Portugal
sergiomdcoliveira@gmail.com; {cfp, mfs}@dsi.uminho.pt
[2] CCTC, University of Minho, Braga, Portugal
{jmac, abelha}@di.uminho.pt

Abstract. The lmitations found in hospital management are directly related to the lack of information and to an inadequate resource management. These aspects are crucial for the management of any organizational entity. This work proposes a Data Mining (DM) approach in order to identify relevant data about patients' management to provide decision makers with important information to fundament their decisions. During this study it was developed 48 DM models. These models were able to make predictions in the hospital environment about beds tournover/patients discharges. The development of predictive models was conducted in a real environment with real data. In order to follow a guideline, the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was adopted. The techniques used were the Regression Tree (RT) and Support Vector Machine (SVM) in order to perform regression tasks. Regression models were able to predict patient's discharges with Relative Absolute Error (RAE) lower than 100% -]38.26; 96.89[. Significant results were achieved when evaluated the Mean Absolute Error (MAE) -]0.619; 4.030[and Mean Squared Error (MSE) -]0.989; 34.432[.The use of these models can contribute to improve the hospital bed management because forecasting patient discharges makes possible to determine the number of beds available for the subsquent weeks.

Keywords: Hospital Management, Patients Management, Beds Management and Data Mining.

1 Introduction

Nowadays organizations can acquire high datasets from multiple business processes. Based on the idea that the organizations can construct large volumes of data for extracting knowledge, this knowledge cannot be extracted without the use of appropriate tools (e.g., data mining) [1]. Hospitals operate on this principle since their databases may contain important hidden knowledge for example, the patient clinical status or the respective prognosis. When are applied Data Mining (DM) techniques in the data acquired it is expected a significant contribution to knowledge discovery from the database [2].

The application of DM processes in the health sector can be an asset. DM can help health insurers to detect fraud or abusive situations in the decision making process influencing the relationship with its clients. Doctors can identify more effective treatments as well as best practices and patients can receive better health care. DM provides not only the methodology but also the technology to transform enormous volumes of data into useful information for the decision makers [3].

Based on these promises a goal was defined: to determine whether it is possible or not to construct DM models capable of supporting the hospital bed management process based on the number of patient discharges.

The data used to accomplish the process of knowledge discovery were obtained using variables in a quantitative descriptive format. This aspect strengthened the need of using the regression approach. Although, also be possible to follow the classification approach through the creation of classes, at a first stage the models were induced using the regression approach. The study was developed in a real environment by using real data, i.e., the numbers of patient discharges of some hospital services. The data used are from the Centre Hospital of Porto (CHP) – Hospital Santo António.

This document is structured in six chapters. After a first introduction of the problem, the conceptual framework is designed and the techniques and metrics for evaluating regression models are presented in the second chapter. In the third chapter, the development process is presented driven by the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology.

In the fourth chapter, results are analysed and the most relevant aspects of the work are discussed. The fifth chapter concludes the paper presenting relevant aspects of the work. Finally, possible future directions are presented.

2 Background

Generally a hospital is by nature an extremely complex organization. This happens because the hospital is composed by highly differentiated services where it is required an intensive work and differentiated workforce [4].

The World Health Organization (WHO) [5] describes the hospital as being part of a medical and social organization whose function is to provide a comprehensive health service and care for the population, both curative and preventive, and whose outpatient services must reach families in the home environment. The hospital is also a centre for training of health workers and biosocial research. To manage a hospital it is necessary to have a deep understanding of the institution where the decision maker works. He has to know the rules and routines of the services that the hospital provides, he also has to be able to identify the strengths and find which aspects need improvements. It is from these aspects that the manager should outline a clear and organized plan to provide an efficient and effective hospital management [6]. Management combines a set of variables, technical tools and technologies in order to ensure organizational' success [4]. The core business of an Hospital is to optimize patients admission, minimizing the length of stay and maximizing the treatment quality [7]. In this sense, one of the most important features is the beds distribution by the various services. Their management reflects the efficiency and quality of hospital management [8]. The introduction of new technologies in hospitals provides opportunity to improve the work, as example, the tasks can be performed faster, more consistently and with less cost [9].

DM as well as some classic statistic methods have been implemented in hospital databases since 1990 [10], [11], [12]. Ever since, DM is becoming increasingly popular if not increasingly essential in healthcare [3].

2.1 Hospital Bed Management

Hospital beds are one of the scarcest resources of hospitals. In most cases the beds are arranged according to hospital specialties in order to provide a better service for the patients. The beds management reflects not only the services efficiency but also the quality of hospital management. In order to improve the hospital management with a focus on bed management there are some studies about hospital patient admission using DM techniques. These studies presented as main goals decrease the length of stay of patients and performing a good beds management.

A project conducted at hospital Chiba University Hospital, studied which tasks awarded for medical care which were directly correlated with patient length of stay. This study revealed a strong correlation between some variables, presenting coefficients between 0.837 and 0.867, in a range of [-1, 1]. The study concluded that the obtained results have shown a strong correlation between patient discharges and hospital management quality [11].

Another study was conducted at the National University Hospital of Singapore. This study had as main objective identify which was the key variable associated to mismatch allocation of beds by department. The obtained models had acuity values of 74.1% and 76.5% and allowed to find that the key variable was the medical speciality. Through this study was possible to determine strategies to beds allocation in the respective hospital [13].

2.2 Data Mining

Technological advances have provided new ways to create and store data. Organizations accumulate data related with their processes (billing, business transactions and accounting) based on the idea that large volumes of data can be a source of knowledge [1].

From a technical standpoint, DM is a process that uses artificial intelligence techniques, statistics and mathematics to extract useful information and knowledge (or patterns) from large volumes of data. These patterns can be in the form of business rules, affinities, correlations, or in terms of forecasting models [14]. The goal of predictive methods is to automatically build a behavioural model, obtaining new samples, in order to be able to predict values of one or more variables related to the sample. The design of patterns enables knowledge discovery and can be easily used as a decision base [15].

For this work the implementation of the DM was achieved through the use of a statistical environment R. R presents itself as a programming language and an environment for statistical development [16]. The library e1071 [17] was used to implement the DM techniques Support Vector Machine (SVM) and Regression Tree (RT), and to resolve the regression forecasting problem. To evaluate the DM models it was used the rminer library [18].

3 CRISP-DM

The DM process is complex but when it is driven by a methodological context becomes easier to understand, implement and develop. As such, CRISP-DM methodology was followed to carry out the present study.

3.1 Business Understanding

The first phase of the CRISP-DM is focused on understanding the project objectives and requirements from the business point of view. For this work, it was necessary to conduct some research focusing on the CHP, with the intention to prove the necessity of develop this study.

First of all it was identified that, until this date, CHP has no mechanism to predict the flow of medical discharges. This evidence was used as foundation for the work process and became an objective. The main objective of this work was to predict the number of patients discharged weekly in order to improve beds management.

The data used to perform this study were provided by the hospital above mentioned. The DM models designed require only two types of variables as input: the target variable (the number of patient discharged distributed by multiple services) and the respective date.

3.2 Data Understanding

The sample provided comprises a period between 1/1/2009 and 31/12/2012. For the sample timeline, the years 2009, 2010 and 2011 have 365 days and the year 2012 has 366 days (leap year). The referred timeline is composed by 62302 registers. These registers correspond to discharges from ninety one hospital services. Each record contains three diferent fields:

- Date: corresponds to the day, month and year in which (s) the patient (s) was/were discharged;
- Service: hospital service that discharged the patient;
- Number of discharges: contains the number of patients who were discharged. This field is directly related to the date and the hospital service. This means that the number was grouped by date and service.

To predict hospital discharges by week, the first task was to group the number of patients by service and by the respective week of the year. Although they were received data from ninety one services, only four of these services presented daily records for all days from 2009 to 2012. In this context only those services were selected to perform this study: Orthopedics, Obstetrics, Childbirth and Nursery.

The following Table 1 presents the statistical analysis for each one of the services.

Table 1 – Data characterization

	Orthopedics	Obstetrics	Childbirth	Nursery
Maximum	78	91	8	92
Minimum	19	40	0	34
Average	≈48.6	≈62.9	≈2.4	≈57.3
Coefficient of Variation	≈22.634%	≈17.011%	≈70.833%	≈19.197%
Standard Deviation	11	10.7	1.7	11
Total Discharges	10108	13080	495	11924

3.3 Data Preparation

Since this study is intended to use regression approach it was necessary to ensure that the data used would enable the application of the respective DM approach. As already mentioned, this work proposes to make predictions of hospital discharges by week. So it was necessary to trasnform the records, grouping by weeks, following the standard that a week begins on Sunday and ends the following Saturday. Through this principle were created groups of 52 weeks by each one of the years: 2009, 2010, 2011 and 2012.

The data tables were developed according to the number of weeks and the respective years, i.e., the rows correspond to the number of weeks (52) and the columns correspond to years from 2009 to 2012. This data representation is called conventional. It was also used the data sliding window representation (with window N = 4), however this data representation did not provide the expected results, thus becoming obsolete.

Table 2 presents an example of a conventional data table for the first five weeks by each one of the four years.

Table 2 – Data characterization

Week Number	Year 2009	Year 2010	Year 2011	Year 2012
1	56	73	66	45
2	62	73	73	54
3	47	75	54	70
4	42	70	62	70
5	41	76	72	67

3.4 Modelling

To induce the regression models were considered two techniques: SVM and RT. The data mining techniques were selected considering three distinct criteria: easy understanding of the techniques, engine efficient and the suitability for training the models with a small dataset. Based in these three principles, SVM achieved the second and third goals and presents as being efficiently using a reduced number of data in the training phase. The RTs has in common the fact of being easy to understand and efficient.

In order to test the models, a mechanism was implemented and two sampling methods were addressed: 10-folds Cross Validation (CV 10-folds) and Take-One-Out Cross Validation (LOOCV). The 10-folds CV was implemented due to the good results

demonstrated on multidisciplinary databases [19]. LOOCV was implemented because it is a suitable method for database when it is used just a few dozen records [20], because the data tables only have 52 or 205 registers.

SVM and RT techniques were submitted to the tune function. This function included in library e1071 tries to search for the hyperparameters intervals previously supplied and identifies the best model and respective hyperparameters.

In the SVM technique it was used two distinct kernels, Radial-Basic Function (RBF) and linear. When two different kernels are used it is necessary to perform different parameterizations because the hyperparameters were different between the kernels. For both kernels it was determined a range of values and the cost parameter C. Its range was defined in terms of the values obtained by the power $2^{(1,..,4)} = [2, \dots, 16]$ where $C > 0$.

The cost parameter C introduced flexibility in the separation of categories in order to control the trade-off between the training errors and the rigidity of the margins. The Gamma hyperparameterγ was defined in the same way as C, the range was determined according to the values obtained by power, $2^{(-1,0,1)} = [0.5, 1, 2]$. This parameterization was used only for the RBF kernel. The value of γ determines the curvature of the decision boundary [21].

The function ε − insensitive loss - was introduced in SVMs. Using this function the models ignored the errors, i.e., the pipe created around the waste and the errors were ignored. This value was determined by default (0.1). The RTs were developed using the CART algorithm through the method of feature selection or splitting rule, Gini Index (GI).

The goal of the GI is to calculate the value for each attribute using the attribute for the node with the lowest impurity proportion [1]. The GI index determines the impurity of D from a data partition or a training set of attributes $Gini(D) = 1 - \sum_{i=1}^{m} pi^2$, where pi corresponds to the probability of an attribute of D belonging to class Ci, this value is estimated by $|Ci \cap D|/|D|$. The sum is calculated as a function of m classes [22].

The developed models can be represented by the following expression.

$$M_n = A_f + S_i + D_x + MRD_z + TDM_y + MA_k.$$

The model M_n belongs to the approach (A) regression and is composed by a service (S), a type of variable (TV) a method of data representation (MDR), a DM technique (TDM) and a sampling method (SM):

$$A_f = \{Regression_1\}$$
$$S_i = \{Orthopedics_1,\ Obstetrics_2, Parturition_3, Nursery_4\}$$
$$TV_x = \{discrete\ quantitative\ variavles\}$$
$$MDR_z = \{Conventional_1, Sliding\ Window_2\}$$
$$TDM_y = \{SVML_1, SVMRBF_2, DTGI_3, DTIG_5, NB_5\}$$
$$SM_k = \{10 - folds\ CV_1, LOOCV_2\}$$

Using this notation for representing DM models is possible to present a particular model implemented. For instance, the model (M_1) follows the regression approach using the service data from Obstetrics and as the conventional data representation method, the

technique SVM with RBF kernel and the sampling method 10-folds CV and it is expressed by:

$$M_1 = A_1 + S_2 + TV_1 + MDR_1 + TDM_2 + SM_1.$$

3.5 Evaluation

After applying the DM models it was necessary to carry out the models evaluation. To evaluate the results presented by DM models approach several regression metrics were adopted: Mean Squared Error (MSE), Mean Absolute Error (MAE) and Relative Absolute Error (RAE).

To evaluate the models, they were submitted to a mutually exclusive division of the subsets, through the implementation of two procedures: 10-folds CV and LOOCV. In the implementation of the respective procedures ten runs were performed for each one of them. Around 100 experiments were performed for each test configuration with models that use the 10-folds CV procedure. The number of experiments performed using the LOOCV procedure was 520.

Table 3 shows the values obtained through the implementation of various metrics. This table also marks the Best Models (BM), following the representation of M_n.

Table 3 – Evaluation of regression models

Best Model	Service	TDM	SM	MAE	MSE	RAE
BM_1	S_1	TDM_2	$SM_{1,2}$	≈4,030	≈34,432	≈74,075%
BM_2	S_2	TDM_2	$SM_{1,2}$	≈1,863	≈10,925	≈38,260%
BM_3	S_3	TDM_2	$SM_{1,2}$	≈0,619	≈0,989	≈96,894%
BM_4	S_4	TDM_2	$SM_{1,2}$	≈2,433	≈20,150	≈53,818%

The technique which provided better results was the SVM with the RBF kernel. In the Regression Error Characteristic (REC) plots it is possible to identify the superiority of accuracy of SVMs (as the error tolerance increases) relative to the top RTs developed models, these results are represented in Figure 1.

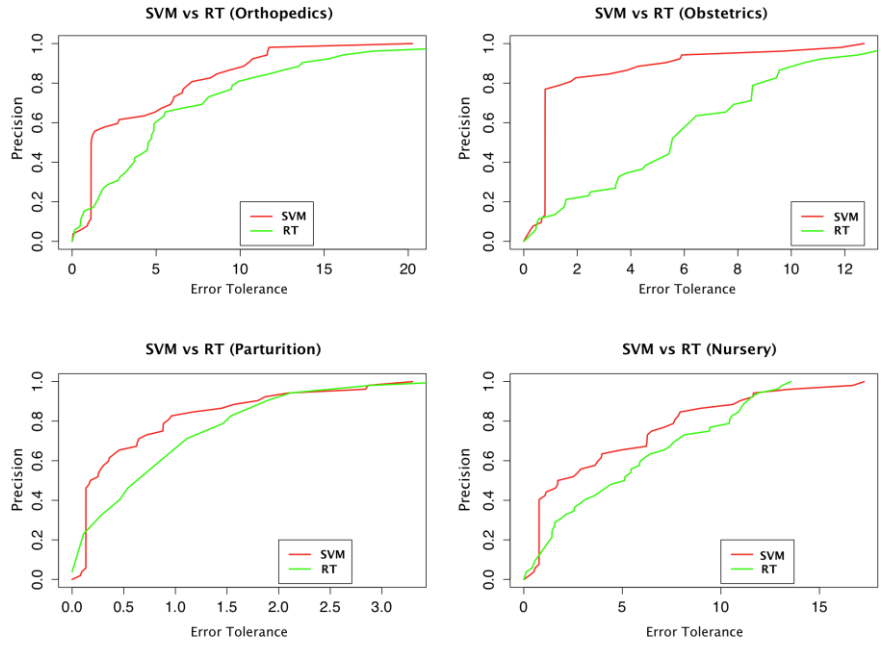

Figure 1 – REC curves for prediction models

The best results were also obtained by using the two sampling methods, 10-folds CV and LOOCV. Regarding to the representation of the data, it was found that the conventional method provided optimum results. Through the values expressed by the metric MAE it was observed that the average of absolute errors are significantly lower for BM_2 and BM_4 models.

In order to identify the predictions made with dissimilar values of the real values, i.e., identify the existence of outliers, it was used the MSE metric. The values presented by the metric shown some inadequate predictions for some services in a few weeks. This observation was relevant since it aims to identify the least possible discrepant values. Yet in order to perform evaluation and validation of the models generated it was used the RAE metric. This metric helped to identify the best models. In order to demonstrate the quality of the results, some graphical representations were designed in order to study the various services. The results can be observed in Figure 2.

From the graphs drawn in figure 2 for the various services it is possible identify the difference between the forecast value and the value verified in the reality. For example the predictions made for the Nursery presented a lower error than Orthopedics service. The difference between the range of values (real and forecasted) verified in the Nursery service it is less than the variation obtained in Orthopedics service.

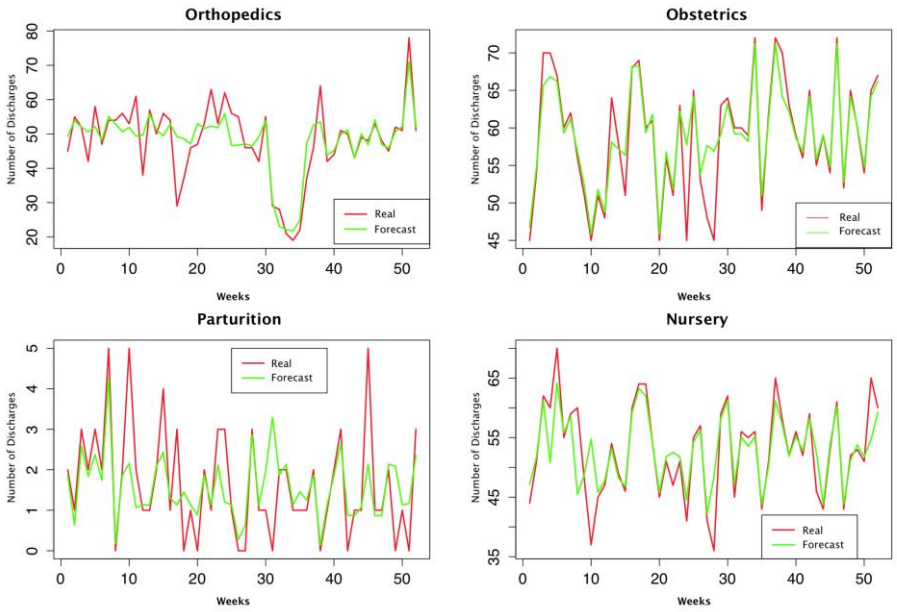

Figure 2 – Forecasts for the four services

4 Discussion of Results

Through the application of the RAE metric it was verified that BM_1 and BM_3 models had a performance near the naïve average predictor, most concretely for Childbirth (BM_3) service the return value from RAE is closest to 100%. Evaluating all the services, BM_2 and BM_4 presented the best results. It is important to point the result obtained by the Nursery service, because it is significantly lower than the results presented by other services with a value of RAE ≈38.260%.

The magnitude of the errors using MAE metric is substantially lower for the four services in particular to the services of Obstetrics and Nursery with errors of 1,863 and 2,433 respectively. These errors correspond to a very small value when compared with the average discharge value of these two services.

The average of the annual values (2012) for the respective services (SAVG) are 58,81 and 52,87 patients. The results presented by MAE for the services (SMAE) when compared with the average values demonstrates that the errors obtained for these two services are small. The percentage of error obtained for the Obstetrics service is 3,17% of the average value and for the Nursery service is 4.6%. These values (X) were obtained through the formula:

$$x = (SMAE * 100)/SAVG$$

The application of the MSE metric demonstrated some misfits to the values predicted when compared with the real values. Through MSE it was possible to observe the

existence of outliers in the predictions made. It should be noted that the obtained results for the Obstetrics and Nursery services had relatively low values when compared with their respective average values. The application of this metric it was important because it allow analyse the mismatch between the predictions made and real results.

5 Conclusion

This study aimed to carry out effective predictions of hospital discharges for the four services above mentioned based on the Data Mining (DM) process. The extraction of knowledge was achieved through the application of two DM techniques (SMV and RT). Its application to the data sources resulted in the discovery of knowledge contained in CHP database.

This study demonstrated that regression models are very promising to solve problems that fall in the studied environment. Moreover, is also possible to conclude that the conventional method of representing data combined with sampling 10-folds CV and SVM technique demonstrated to be the most appropriate tools for extracting knowledge from the studied data source.

Concluding, this study can prove that it is possible to predict the number of weekly discharges of patients by using the number of patients discharged by day registered in the past.

6 Future Works

After concluding this study it is essential to define a set of directions that the further work can take. In the future special attention will be taken in the following aspects:

- Introduction of new variables in the prediction models, such as the number of patients admissions, length of bed occupancy and bed occupancy ratio;
- Estend the experiments to other hospital services;
- Implement a prototype of a DSS in a real environment and identify whether the system is in accordance with reality.

Acknowledgements

This work has been supported by FCT – Funda. ão para a Ciência e Tecnologia in the scope of the project: PEst-OE/EEI/UI0319/2014.

The authors would like to thank FCT (Foundation of Science and Technology, Portugal) for the financial support through the contract PTDC/EIA/72819/ 2006 (INTCare) and PTDC/EEI-SII/1302/2012 (INTCare II).

References

[1] M. Santos and C. Azevedo, *Data Mining Descoberta do conhecimento em base de dados*. FCA - Editora de Informática, Lda, 2005.

[2] M. Santos, M. Boa, F. Portela, Á. Silva, and F. Rua, "Real-time prediction of organ failure and outcome in intensive medicine," in *2010 5th Iberian Conference on Information Systems and Technologies (CISTI)*, 2010, pp. 1 –6.
[3] H. Koh and G. Tan, "Data mining applications in healthcare," *J Healthc Inf Manag*, vol. 19, no. 2, pp. 64–72, 2005.
[4] J. Proença, A. Vaz, A. Escoval, F. Candoso, D. Ferro, C. Carapeto, R. Costa, and V. Roeslin, *O Hospital Português*. Vida Económica-Conferforum, 2000.
[5] WHO, "Expert Committee on Health Statistics," 261, 1963.
[6] I. Santos and J. Arruda, "Análise do Perfil Profissional dos Gestores dos Hospitais Particulares da Cidade de Aracaju- SE," *Revista Eletronica da Faculdade José Augusto Vieira*, vol. Nª -7, 2012.
[7] M. Neves, "Os Médicos vão ter de ser os motores da reforma do sistema," *Revista Portuguesa de Gestão & Saúde*, no. 5, 2011.
[8] G. Yang, L. Sun, and X. Lin, "Six-stage Hospital Beds Arrangement Management System," presented at the Management and Service Science, 2010.
[9] A. Dwivedi, R. Bali, and R. Naguib, "Building New Healthcare Management Paradigms: A Case for Healthcare Knowledge Management," presented at the Healthcare Knowledge Management Issues, Advances, and Successes, 2006.
[10] R. Bose, "Knowledge management-enabled health care management systems: capabilities, infrastructure, and decision-support," *Expert Systems with Applications*, vol. 24, no. 1, pp. 59–71, 2003.
[11] S. Tsumoto and S. Hirano, "Data mining in hospital information system for hospital management," presented at the ICME International Conference on Complex Medical Engineering, 2009. CME, 2009, pp. 1 –5.
[12] S. Tsumoto and S. Hirano, "Towards Data-Oriented Hospital Services: Data Mining-based Hospital Mangement," presented at the The 10th IEEE International Conference on Data Mining Workshops, Sydney, Australia, 2011.
[13] K. Teow, E. Darzi, E. Foo, X. Jin, and J. Sim, "Intelligent Analysis of Acute Bed Overflow in a Tertiary Hospital in Singapore," *Springer US*, 2012.
[14] E. Turban, R. Sharda, and D. Delen, *Decision Support and Business Intelligence Systems*, 9ª Edição. Prentice Hall, 2011.
[15] O. Maimon and L. Rokach, "Introduction to Knowledge Discovery and Data Mining," in *Data Mining and Knowledge Discovery Handbook*, 2ª Edição., Springer, 2010.
[16] L. Torgo, *Data Mining with R: Learning with Case Studies*. CRC Press - Taylor & Francis Group, 2011.
[17] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, "Misc Functions of the Department of Statistics (e1071)." 2012.
[18] P. Cortez, "Simpler use of data mining methods (e.g. NN and SVM) in classification and regression." 2013.
[19] I. Witten, E. Frank, and M. Hall, *Data Mining Pratical Machine Learning Tools and Techniques*, 3ª Edição. Morgan Kaufmann, 2011.
[20] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," *Encyclopedia of Database Systems*, 5 vols. Springer, 2009.
[21] A. Ben-Hur and J. Weston, "A User's Guide to Support Vector Machines," in *Data Mining Techniques for the Life Sciences*, O. Carugo and F. Eisenhaber, Eds. Humana Press, 2010.
[22] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3ª Edição. Morgan Kaufmann, 2012.