Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Descoberta do Conhecimento - 2017/2018

Predicting the Interval of Recovery Days For Breast Tumour Patients

Hugo Carvalho[a], João Nuno Almeida[b], Marcos Luís[c]

[a]*MIEI - A74219, Departamento de Informática, University of Minho*
[b]*MIEI - A75209, Departamento de Informática, University of Minho*
[c]*MIEI - A70676, Departamento de Informática, University of Minho*

Abstract

Treatments for breast tumours are delicate and complicated procedures. Each day a patient spends in post-surgery recovery, translates into an extra expense for the hospital in question. On top of that, the life quality of each patient also decreases. Given these premises, the goal behind this paper, is to predict and to understand which aspects influence the interval of recovery days in which each patient, who has been submitted to a specific kind of treatment, will fall into and remain in post-surgery care. This way we may hopefully, increase the life quality of each patient by decreasing the number of days they will have to spend in this stage, and we may improve the fund management of the medical centres giving these treatments. To reach this goal, we used previous patients medical records combined with Data Mining techniques following the CRISP-DM methodology and software tools like WEKA, Excel and RapidMiner to generate the DM models.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Data Mining; CRISP-DM; WEKA; Excel; Breast Tumour; Post-Surgery;

1. Introduction

Currently, organisations are relying more and more in decision making systems (DSS)[6]. These tools have been very useful for users in various areas of expertise. Data Mining is the process in which large amounts of data (compiled in datasets) is analysed, using numerous techniques and working methodologies. These techniques and methods, have a common goal, trying to find hidden patterns in the data, that may help in some way the people involved. There are two types of Data Mining techniques: descriptive and predictive. Descriptive techniques include clustering and k-Means algorithms, and it is useful when there are no obvious natural grouping within the dataset. Predictive techniques can be divided in two branches: classification techniques and regression techniques. In regression techniques, the goal is to find within the data, a mathematical way (usually in the form of an equation) that allows to predict a target value. In classification techniques such as the ones explored in this study, the main focus is to predict a certain type of nominal labels.

Data Mining is a complex process, not only when it comes to analysing the data, but also in the moment of finding a source of data. Numerous privacy laws, and privacy regulations stop engineers from being able to use the data stored for the purposes of a study.

In the health care business, the one we are focusing in this essay, the use of Data Mining models, as been used mainly to give a certain kind of power to the professionals of this area. Knowing beforehand, or having at least a

Peer-review under responsibility of the Conference Program Chairs.

foreknowledge, of what may happen to the patients in care, helps medical personnel in a major way by virtue of allowing them to be prepared ahead of time for the occurrence of a certain scenario.

2. Background and Related Work

2.1. Breast Tumours

A breast tumour, just like any other tumour, is a mass of abnormal tissue, created by an unnatural accelerated cell reproduction within the tissue in question. There are two kinds of breast tumours, benign tumours or malign tumours, most commonly known as cancer.

When not treated, benign breast tumours, may cause pain and discomfort, even though they are not as aggressive as malign tumours, benign tumours may occasionally grow and start affecting surrounding tissues and organs. Malign tumours or cancers, are extremely aggressive towards surrounding tissues, for their only goal is to attack and destroy them, creating metastasis, i.e., secondary tumours in other body parts.

The most common breast tumour symptom is a lump, that feels different from the rest of the breast tissue. Breast tumours are usually gender biased with 99% of the total number of cases being diagnosed in women. The most affected age group is woman with 65 years or more.

Like any other evasive surgery, there are risks associated with the treatment of breast tumours. Studies have concluded that age is not a risk factor for post-operative complications, neither is something that influences the days spent in post-operative care, and that the rate of POCs (post-operative complications) is currently 15.2%. However there is still no conclusion on what is the predicted interval of days needed to recover from such treatments, because no one can really put their finger on what factors affect that variable, even though the current average time is 7 days. Having the knowledge beforehand of what motivates the need to remain in the post-operative condition or not, would allow doctors and health professionals to, if possible, avoid certain types of treatments or approaches that they knew would increase the number of days admitted in the hospital in a certain variety of patients. This would improve the patients life quality by reducing the total time spent in a hospital environment. Having a prediction in advance of the interval of recovery days in which each patient would fall into, would also allow the hospital to manage their funds more accurately, distributing the remaining funds more accordingly.

2.2. Related Work

In this section we will presented some works that have applied Data Mining techniques, in an effort to support the existence of behavioural patterns in the breast cancer research area.

Delen et al. (2004)[3] used three Data Mining techniques to predicted the survivability of patients diagnosed with breast cancer. Real data from more than 200000 cases was compiled in a dataset, used, and applied in two of the most popular Data Mining algorithms: artificial neural networks and decision trees. Along with these, a statistical method (logistic regression) was used, to develop the prediction models. A 10-fold cross validation, was also used to measure the unbiased estimate, of the predictions obtained using the algorithms referenced above. The results achieved in terms of accuracy, with the use of the decision tree algorithm, were the best ever recorded in literature with a value of 93.6%. In second came the values achieved by the artificial neural network algorithm, with an accuracy level measured at 91.2%. Lastly came the results achieved by the logistic regression models with an accuracy of 89.2%.

Asri et al. (2004)[4] used machine learning algorithms to predict the risk of breast cancer and to diagnose it. The data used was real, and extract from the Wisconsin Breast Cancer datasets. For this study, four machine learning algorithms were considered: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbours (k-NN). In the end, the most accurate machine learning algorithm proved to be the SVM algorithm, with and accuracy measured at 97.13% and the lowest error rate of the four at 0.02%.

3. Methodology, Data and Data Mining

In this section we will explore the subjects of the methods used to reach conclusions, the data mining process itself and techniques used and the data sources available.

3.1. Dataset

The data sample used to represent the population of breast tumour patients, was extracted from medical records of 176 real anonymous patients, from ages 14 to 85. It comprises the months of January and February of the year 2017, and there is no information within the data sample that specifies from which hospital(s) were the records taken, therefore we can not make any conclusions about the nationalities of the people involved. Because the patients are anonymous we can not take any elation about their sex either. The data sample contains, for each patient, information in the form of 23 attributes, about their personal health history, the treatment they were submitted to, their behaviour in the post-surgery period and any complication that might have occurred.

3.1.1. Privacy

One of the biggest worries we had while conducting this study, was to always keep intact the privacy of the patients involved in the dataset. Delicate situations are being handled, and is not our intention at all to harm or violate a persons right to privacy.

3.2. CRISP-DM

We followed the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, during the Data Mining Process. Essentially it divides the Data Mining process into six phases: *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Discussion*. This *framework* was followed incrementally, by the order specified in order to increase the success of the Data Mining process as a whole and its final conclusions.

3.3. Data Mining

To generate the Data Mining models needed, three different techniques were used: *J48*[5], *Naive Bayes*[5] (NB) and *k-Nearest Neighbours* (kNN)[5]. We used Weka to implement these machine learning algorithms.

NB or Naive Bayes relies on the Bayes theorem of conditional probability that allows one to calculate the probability of a certain event based on prior knowledge of conditions that might be related to the event. The NB algorithm as been one of the most popular algorithms in Data Mining studies, because of its ability to perform very well when there is, in fact, a strong dependency between attributes.

kNN or k-Nearest Neighbours, finds k nearest group of objects in the dataset used, that are the closest to the test instance. It bases its predictions, on patterns and classification that may have been found in surrounding instances. It allows weights to be assigned to the nearest neighbours, so that these contribute more for the final results and predictions.

J48 is a JAVA implementation of the C4.5 algorithm. It generates a decision tree and at each node a decision is made. The attribute that most effectively splits the samples into enriched subsets of data is chosen and makes the decision. The criteria for this choice is based on which attribute as the highest normalisation gain.

4. Data Mining Process

4.1. Business Understanding

The goal of the work this paper is presenting, is to predict the interval of days a breast tumour patient will remain in post-surgery care. The aspects to be considered consist of, the personal health history of each patient, the treatment he was submitted to and the complications that may have occurred in the meantime, given the way each patient deals with the disease and the treatment. Knowing beforehand, what patterns may affect the interval of recovery days in which a patient will fall into, in post-surgery care, not only increases the life quality of future breast tumour patients by avoiding, if possible, certain treatments that may extend that period, but it also allows hospitals and medical centres to predict and manage the costs involved in keeping the patients admitted with a lot more care and knowledge ahead of time.

4.2. Data Understanding

The dataset used consists of 176 entries. Each entry represents, the records and information of an anonymous real former breast tumour patient. Unfortunately we can not specify from which hospital(s) were the records taken. Each entry is described by a set of 23 attributes, divided in 4 groups: Personal History, Surgery, Post-Operative, Complications.

Table 1 and Table 2 show the statistical distribution of the attributes in Personal History, and presents the attributes themselves to the reader. Table 3 shows the statistical distribution of the attributes in Surgery, and presents the attributes themselves to the reader. Tables 4 and 5 show the statistical distribution of the attributes in Post-Operative, and presents the attributes themselves to the reader. Tables 6 and 7 shows the statistical distribution of the attributes in Complication, and presents the attributes themselves to the reader.

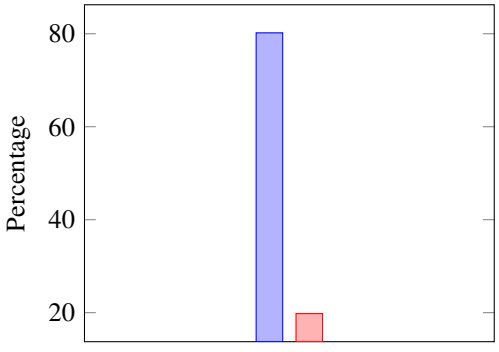
Table 1: Numerical attributes in group: Personal History.

Name	Minimum	Maximum	Mean	Standard Deviation
Idade	14	85	53.023	15.560

Table 2: Nominal attributes in group: Personal History.

Name	Range	Percentage (%)
Tabaco	Sim	15.9
	Não	84.1
Diabetes	Sim	8.0
	Não	92.0
Imunosupressores	Sim	3.4
	Não	96.6
Hipocoagulação	Sim	2.3
	Não	97.7
QTx NA	Sim	8.8
	Não	91.2

Within the dataset there is an unnamed target attribute. This unnamed attribute, draws a conclusion on whether there was a post-treatment complication or not, for each patient. The attribute specifies which complication occurred if that is the case. Because that is not the issue of this work, we will normalise the information in a binary way: yes or no, for the emergence of a complication. As we can see in the following graph, a great percentage of the patients, *80.2%*, did not show signs of any complication that required them to be admitted for longer than predicted, nevertheless, there is a percentage of patients, *19.8%*, that suffered from these complications. However, even though this is the only explicit target attribute, this is not the target attribute we will use considering our goals with this work. Our goal is to predict the interval of recovery days for breast tumour patients, and therefore our target attribute will be the attribute *Dias*.



Graph 1. Target Attribute distribution

Table 3: Nominal attributes in group: Surgery.

Name	Range	Percentage (%)
Data Cx	02/01/2017 a 27/02/2017	100.0
Cx/Ambulatório	Cirurgia	58.0
	Ambulatório	42.0
Benígno/Malígno	Benígno	19.9
	Malígno	80.1
Diagnóstico	CI NST	51.7
	Fibroadenoma	9.7
	CDIS Alto Grau	4.5
	CL Invasor	4.5
	Others (37)	29.6
Lateralidade	D	49.7
	E	46.9
	B	3.4
Intervenção Mama	TA	38.9
	BE	17.7
	MRT	8.8
	MRM	7.1
	Others (15)	27.5
Intervenção Axila	GS	81.6
	EA	17.1
	Exérese de gânglio	1.3
Outras Intervenções	CVC TI	54.5
	Remoção CVC TI	12.5
	Simetrização	5.7
	RMI	2.3
	Others (21)	25.0

Table 4: Numerical attributes in group: Post-Surgery.

Name	Minimum	Maximum	Mean	Standard Deviation
Dias	0	26	1.710	2.972

Table 5: Nominal attributes in group: Post-Surgery.

Name	Range	Percentage (%)
Antibióticos	Sim	10.2
	Não	89.8
Hipocoagulação	Sim	54.3
	Não	45.7

Table 6: Numerical attributes in group: Complications.

Name	Minimum	Maximum	Mean	Standard Deviation
Dias Pós	1	85	20.714	24.724
Dias Tx	—	—	—	—

4.3. Data Preparation

In order to create accurate DM models, the data had to be prepared. The first step in the data preparation process was to eliminate all the null entries of the dataset, that were causing nothing but inconsistency. First of all, we started

Table 7: Nominal attributes in group: Complications.

Name	Range	Percentage (%)
Data Dx	02/01/2017 a 27/02/2017	2.3
Complicação	Hematoma	14.8
	Seroma	11.1
	Atraso Cicatrização	7.4
	Deiscência	7.4
	Others (15)	59.2
Tratamento	Conservador	45.5
	Médico	22.7
	Drenagem Aspirativa	18.2
	Drenagem Cirúrgica	4.5
	Penso	4.5
	Pico	4.5

by eliminating the attributes ”Nome”, ”Data Cx” and ”Dias Tx”, because they weren't in any way relevant considering the goals of this study. In total, at the end of this phase, we took out 169 entries of the original 176. We handled these missing values, by deleting every entry that corresponded to a missing value of less than 5% in a certain attribute. By replacing the missing values of a certain attribute, by its mean or mode, depending on whether it was a numerical or nominal type, if the percentage of missing values was greater than 5% and less than 80%. And by deleting the attribute from the dataset in its whole if the percentage of missing values was bigger than 80%. We also felt the need to change the instances of some attributes. For nominal attributes where the range of values consisted in two possibilities, for example, ”Yes” or ”No”, we normalised the data in order to have a range of values of 0 and 1. We transformed the attribute ”Idade” in a categorical attribute by creating three age intervals: Young (0 to 45 years old), Middle (46 to 62 years old) and Elder (62 years old or more). In cases where the range of values was more extensive, for example in the attribute ”Intervenção Mama”, we used the 4 most common occurrences and flagged the rest with the value ”Outras”. For our target attribute, because the goal is to predict the interval of recovery days in which each patient will fit into, given his medical situation, and therefore we are talking about a classification problem, we decided to normalise the values in the attribute Days, by creating two intervals: NormalStay (0 to 1 days in recovery), and LongStay(2 or more days in recovery), based on what Weka considers as the best balance, given the records and entries of the dataset. Next we handled the outliers within the remaining data to prevent them from influencing our final conclusions. We excluded from the dataset every entry that had a numerical attribute, out of the range [Mean-2SD, Mean+2SD]. Ultimately the dataset we were left to work with, had 156 entries, each one with 16 attributes. Table 8 presents the decisions taken for each attribute.

Table 8: Data Preparation decisions.

Name	Missing Values	Decision	Name	Missing Values	Decision
Nome	0%	Remove	Idade	0%	Normalise
Diabetes	0%	Normalise	Imunosupressores	0%	Normalise
QTx NA	3.41% (Remove)	Normalise	Cx/Ambulatório	0%	Normalise
Diagnóstico	0%	Normalise	Lateralidade	0.57% (Remove)	Normalise
Intervenção Axila	56.2% (Mode)	Keep	Outras Intervenções	50.0% (Mode)	Normalise
Name	Missing Values	Decision	Name	Missing Values	Decision
Tabaco	0%	Normalise	Hipocoagulação	0%	Normalise
Benígno/Malígno	0%	Normalise	Intervenção Mama	35.8% (Mode)	Normalise
Dias	0%	Normalise	Dias Pós	92.1%	Remove
Antibióticos	0%	Normalise	Hipocoagulação	0.57% (Remove)	Normalise
Complicação	84.1%	Remove	Tratamento	87.5%	Remove

4.4. Modelling

This phase consisted of inducing the Data Mining Models (DMM) in Weka using the prepared data. As the described approach corresponds to a classification problem, we used 3 different DM techniques: J48, Naive Bayes and k-Nearest Neighbours (kNN). These algorithms were used with the default settings in Weka.

Two different data approaches were made, one of them using oversampling and the other without it, both testing on 1/3 of the data (Holdout Sampling) and on all the data (Cross Validation with 10 folds). The different scenarios presented below were also created combining different variables in order to identify which factors have more impact on predict the number days of hospitalisation for a patient:

S1: All variables

S2: Idade, Diabetes, Imunossupressores, Hipocoagulação, QTx NA

S3: Cx/Ambulatório,Ben./Malig., Diagnostico, Lateralidade, Intervencão Mama, Intervencão axila, Outras intervencões

S4: Antibióticos, Hipocoagulação

S5: Diabetes, Hipocoagulação, Intervenção Mama, Intervenção Axila, Outras Intervenções, Antibióticos, Hipocoagulação

Each DMM can be described as belonging to an approach (A), being composed by a scenario (S), a data mining technique (DMT), a sampling method (SM), a data approach (DA) and a target (T):

$$DMM_s = \{A_f, S_i, DMT_y, SM_c, c, TG_t\}$$

A_f = Classification
S_i = S1,S2,S3,S4,S5
DMT_y = J48, NB, kNN
SM_c = Holdout Sampling, Cross Validation
DA_b = Without Oversampling, With Oversampling
TG_t = Days

4.5. Evaluation

To measure the performance of each DM model, its confusion matrix (CMX) was used, for it gives us four parameters that allows us to measure three different statistics. Using the values for True Negatives (TN), True Positives (TP), False Negatives (FN) and False Positives (FP) we are to derive the following formulas:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Tables 9, 10 and 11 show the best sensitivity, specificity and accuracy results respectively, for each DM Model and each technique.

Table 9: DM Models sensitivity results by DM Technique

DM Technique	Scenario	Sampling Method	Data Approach	Sensitivity
J48	S4	Holdout Sampling	With/Without Oversampling	1
NB	S4	Holdout Sampling	With/Without Oversampling	1
kNN	S4	Holdout Sampling	With/Without Oversampling	1

By analysing the content of each table, we see that the DM model with the best results for sensitivity with 100%, was achieved using the all three technique. The DM model with the best results overall for specificity with 100%, was

Table 10: DM Models specificity results by DM Technique

DM Technique	Scenario	Sampling Method	Data Approach	Specificity
J48	S2	Cross Validation	Without Oversampling	0.961
NB	S3	Holdout Sampling	Without Oversampling	0.921
kNN	S1	Holdout Sampling	Without Oversampling	1

Table 11: DM Models accuracy results by DM Technique

DM Technique	Scenario	Sampling Method	Data Approach	Accuracy
J48	S5	Holdout Sampling	Without Oversampling	0.827
NB	S3	Holdout Sampling	With Oversampling	0.871
kNN	S1	Holdout Sampling	Without Oversampling	0.846

achieved using the kNN technique. The DM model with the best results for accuracy with 87.1%, was achieved using the NB technique.

In order to filter the DM models with the best results, we defined a threshold value of 85% for sensitivity, 80% for specificity and 80% for accuracy. Table 12 shows the best DM models that reached the thresholds values defined for each measure.

Table 12: Best DM models by sensitivity, specificity and accuracy

DM Technique	Scenario	Sampling Method	Data Approach	Sensitivity	Specificity	Accuracy
NB	S1	Holdout Sampling	Without Sampling	0.914	0.842	0.827
NB	S3	Cross Validation	With Sampling	0.876	0.825	0.854
NB	S3	Holdout Sampling	With Sampling	0.871	0.844	0.871
J48	S3	Holdout Sampling	Without Sampling	0.872	0.895	0.827
NB	S3	Cross Validation	Without Sampling	0.883	0.883	0.844
NB	S5	Cross Validation	Without Sampling	0.861	0.903	0.838
NB	S5	Holdout Sampling	Without Sampling	0.872	0.895	0.827

5. Discussion

The parameter that we will take into account for the analysis of the best models will be the sensitivity due to being the parameter that measures the proportion of actual positives that are correctly identified as such. If we look only at the sensitivity we can note that the results are 100% relative to scenario 4 (Antibióticos, Hipocoagulação), using different techniques such as J48 , NB and kNN. In this case we are only looking at the best cases using as an evaluator only this parameter which makes it unrealistic in practical terms, especially in real situations.

In terms of general classification and ordering by sensitivity value we can see that although the best case is in a scenario with all the attributes, the following models present attributes only regarding the intervention the indication of the type of cancer. The Naive Bayes algorithm as we can see occurred in 6 of the top 7 DM models because is based on conditional probabilities. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred.

The majority of the cases with the best results are without sampling, in other words this means that there are not many cases clustered in a single class.

6. Conclusion and Future Work

With this work, we were able to use real data to determine which interval of recovery days each patients will fit into, we also proved the success of DM models in predicting such goals. Given the results achieved by our DM models, it is safe to conclude that we have succeeded. Some of them were capable of achieving sensitivity results as high as 91%, specificity results as high as 96% and accuracy values higher than 87%. Considering the known standards for studies of this kind we consider our results satisfactory. The model that achieved the best specificity results while also achieving the established threshold for accuracy and sensitivity, used the J48 technique, the third scenario for attributes, Holdout Sampling as the sampling method and no oversampling which means the number of instances on the dataset was balanced when it comes to patients in each of the intervals of recovery days we have defined. The model that achieved the best sensitivity results while also achieving the established threshold for accuracy and specificity, used the NB technique, the first scenario for attributes, Holdout Sampling as the sampling method and no oversampling which means the number of instances on the dataset was balanced when it comes to patients in each of the intervals of recovery days we have defined. The model that achieved the best accuracy results while also achieving the established threshold for sensitivity and specificity, used the NB technique, the third scenario for attributes, Holdout Sampling as the sampling method and oversampling which means the number of instances on the dataset was not balanced when it comes to patients in each of the intervals of recovery days we have defined and for better accuracy in the future, more instances must be added. The best overall model achieved sensitivity results of 91.4%, accuracy results of 82.7% and specificity results of 84.2%. It used the NB technique, the first scenario for attributes, Holdout Sampling as the sampling method and no oversampling which means the number of instances on the dataset was balanced when it comes to patients in each of the intervals of recovery days we have defined. Nevertheless, we must conclude that the data used to induce these models must be improved by adding more instances, since we consider that 154 translates into a very small sample of entries, when compared with, for example, with Delen who used a Dataset with 200000 entries. These models are therefore, not suited for a decision-support system, but give the users a good notion of what to expect in terms of which is the interval of recovery days each breast tumor patient will fit into.

7. Acknowledgements

We thank Professor José Manuel Ferreira Machado, and Professor Hugo Peixoto for sharing with us their knowledge and experience in Data Mining studies, allowing us to conduct our own research.

References

1. Fonseca F., Peixoto H., Miranda F., Abelha A., Machado J. (2017). Step Towards Prediction of Perineal Tear. *Procedia Computer Science 00 (2017) 000–000.*
2. Morais A., Peixoto H., Coimbra C., Abelha A., Machado J. (2017). Predicting the need of Neonatal Resuscitation using Data Mining. *Procedia Computer Science 00 (2017) 000–000.*
3. Delen D., Walker G., Kadam A. (2004).Predicting breast cancer survivability: a comparison of three data mining methods. *Artmed 2004.07.002*
4. Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procs 2016.04.224*
5. https://newonlinecourses.science.psu.edu/stat857/node/161/
6. Power D. J. (2002). Decision support systems: concepts and resources for managers. Greenwood *Publishing Group*