Understanding Surgery and Post-Surgery Complications in Breast Tumor Patients

Carlos Campos[2], Diana Costa[2], Hugo Peixoto[1], José Machado[1], José Oliveira[2], and Vitor Castro[2]

[1] Algoritmi Research Center, University of Minho, Campus Gualtar, Braga 4710, Portugal

[2] University of Minho, Campus Gualtar, Braga 4710, Portugal

Abstract. Female breast cancer incidence rates have been increasing in Portugal for years, and it is most common cancer today[3]. Every year, around 6,000 cases of breast cancer are diagnosed and around 1,000 women die from the disease. Cancer treatments can be very stressful and the post-surgery complications are common. This work aims to predict when complications may occur, helping medical staff to prevent them, reducing hospital costs and improving patient recovery. After analyzing a dataset with 176 entries from a public portuguese hospital, the best results were found with Support Vector Machine algorithm, using Evolutionary optimization, obtaining a accuracy of 96.88%, a recall of negative class of 96.38% and a recall of positive class of 100%. The precision of each class is 100% for a false predict, and 82.76% for a true predict. With such high values for the several metrics, including a tendency to predict a false positive complication, the goal was accomplished with success.

Keywords: Breast Cancer, Classification, Complication, CRISP-DM, Data Mining, Morbimortality

1 Introduction

Day by day, companies are turning to technology to achieve and empower the objectives they propose[14]. In health care, the aim is to improve the response times and decrease the costs, making a more sustainable and comfortable service.

The increasing capabilities of modern computers allow institutions to apply Data Mining techniques in the daily work-flow. Data Mining is the process in which large amounts of data is analyzed, using numerous mathematical and statistic techniques and working methodologies. The goal, when applying this techniques, is to find a hidden pattern in the collected data, which may help in the treatment process of the involved patients. There are two types of Data Mining techniques: descriptive and predictive. The second one is the useful one for the purpose of this study, that promotes the anticipation of health complications after the cancer treatment. Predictive techniques can be divided in two branches: classification techniques and regression techniques. In classification problems, as is the one in study, the main focus is to predict a certain type of nominal labels.

Due to the ethical and law restritions, Data Mining can be a complex process. In health care businesses, like the one being focused in this essay, this restrictions lead to a close control of the data the group is given to work with. Despite of this problems, the benefits for patients are enormous and so the group was given the opportunity to work on the empower of medical staff. Knowing beforehand 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 is a mass of abnormal tissue, created by a deficient and unnormal accelerated cell reproduction. Tumours can be benign or malignant, being the second one mostly known as cancer. Benign tumours usually present no harm for the person who has it, but may be discomfortable and can start to mess with the surrounding tissue when growing. Malignant tumours, on the other side, are agressive with the surrounding tissues, usually using them to grow in size and danger, as their malfunction causes the cells to attack and destroy the good cells and use the all energy available to grow, even creating metastasis that spread over the body.

In the case of breast tumours, the most common symptom is a lump that feels harder than the surrounding tissue. The majority of breast cancer happens in women[3] and the most common cure is surgery associated with chemotherapy and radiotherapy. Surgeries are a very invasive method so it's pretty common, especially in situations like when the body is already debilitated, to have complication, albeit all the efforts from the medical teams. These complications, associated with the fragile medical situation of the patients, can grow in importance and lead to problems for the rest of the patient's life. To prevent the appearance of complications it would be good to know beforehand if any is expected. Having this in consideration, the objective of the work is to find if, in certain conditions, patients can develop some complication, which would help medical staff to observe closely this cases and reduce the costs of medical treatments and future medical problems.

2.2 Related Work

It's important to be aware of the existing studies in the area. In this section will be 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.

Asri et al. (2004) 4 [12] 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%.

Delen et al. (2004) 3 [11] used three Data Mining techniques to predicted the survivability of patients diagnosed with3 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 at the time, with a value of 93.6%. The values achieved by the artificial neural network algorithm, with an accuracy level measured at 91.2%, achieved the second higher accuracy. The worse results were achieved by the logistic regression models with an accuracy of 89.2%.

In addition to those previously mentioned, it's important to refer other studies done in the healthcare area involving data mining techniques [2] and classification problems [1].

3 Methodology, Materials and Methods

3.1 Methodology

This is a study whose data mining process follows the CRISP-DM method[7], which describes approaches commonly used by specialists to tackle problems of this nature[13]. The biggest advantage in using this method is that it is independent of industry, tools and data. The phases of the methodology include Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

3.2 Materials

The dataset given for this investigation was provided by a Portuguese hospital, and has 176 anonymous personal records about mammary or axillary surgeries performed on women with different types of tumors, as well as some information about the background of the patient and treatments. The data involves women of all ages (from adolescence to menopause), and there are reports of complications during surgery or during post-surgery.

3.3 Methods

The classification algorithms this article will approach are k-Nearest Neighbors (kNN), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF)

and Support Vector Machines (SVM). These are usually considered the best algorithms, and some have the best precision and robustness, in average, between classification problems. They will be explained in order to better understand the reasons why the algorithms were chosen[4].

Therefore, kNN is a classifier algorithm where the learning is based on "how similar" is some data from other things exist in close proximity. So, it can do distance weighting, and is rather a simple and fast algorithm. LR is a simple, rudimental and useful statistical machine learning algorithm that classifies the data by considering outcome variables on extreme ends and tries makes a logarithmic line that distinguishes between them. It is the go-to method for binary classification problems, which is the case. The NB makes an assumption that all the variables in the dataset are "Naive", that is, not correlated to each other. It is based on the so-called Bayesian theorem (probability of an event occurring given the probability of another event that has already occurred), and despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. RF, or Random Forest, builds multiple decision trees and merges them together to get a more accurate and stable prediction. Also, Random decision Forests correct for decision trees' habit of overfitting to their training set, which is desirable. Finally, the SVM[5] combines linear modelling with learning based on instances: it finds a hyperplane in an N-dimensional space(with N being the number of features) that distinctly classifies the data points. That is, this algorithm chooses a limited number of samples from each group and constructs a linear function building separated boundaries between datasets. When no linear separation is feasible, the kernel approach will be used to automatically add the training samples into a higher dimensional space and to learn a separator in that zone.

4 Data mining process

4.1 Business Understanding

The goal of the work presented in this paper is to predict if a certain patient will have complications or not, after the cancer treatment surgery. For the prediction to be possible, the group worked with a dataset that described the age and living habits of the patient, the treatment that has been done and the type and progression of the tumour. Predicting if a certain patient will get a complication translates in a better management of efforts, leading teams to focus on the right patients and decreasing the costs with future treatments for the individual. It is expected a reduction of costs by not having to deal with post-surgery complications, while increasing the satisfaction and health of the patient that can leave the hospital earlier. A better life quality is expected to be achieved, since patients often get diseases and virus the more they stay in the hospital environment.

Lately, hospitals have been collecting data of the patients at a high rate, taking advantage of the powerful computer systems available at such a low cost. More individual data being collected at each time of the process of treatment

allows to better understand and profile the patient. Even though Data Mining techniques are not commonly used in this field, due to the sensible data being used, it is a great opportunity for the use of those techniques. Those effort may, indeed, contribute for a better overall care system and a more sustainable economy.

4.2 Data Understanding

The dataset consists of 176 entries, reduced to 150 entries after the removal of some registers considered irrelevant or unusable. The Portuguese hospital where the records came from is unknown, due to privacy concerns. Each one of the entries is described by a total of 16 attributes, which were reduced from a total of 23. The 16 available are divided in 3 groups, concerning to Personal History, Surgery and Post-Operative conditions.

In the following tables will be shown some statistical information that may help to comprehend how the data distributes. All this data is unchanged and only serves the purpose of understand the data available to perform the study. Table 1 and Table 2 shows those statistical distributions along the Personal History Group. Table 3 does it accordingly to the Surgery group and Table 4 and Table 5 does the same for the Post-Operative Group.

Name	Minimum	Maximum	Mean	Standard Deviation
Age	14	85	53,023	15,560

Fig. 1: Numerical attributes in group: Personal History

Name	Range	Percentage (%)
Tobacco	Yes	15.9
	No	84.1
Diabetes	Yes	8.0
	No	92.0
Immunosuppresants	Yes	3.4
	No	96.6
Hypocoagulation	Yes	2.3
	No	97.7
QTx NA	Yes	8.8
	No	91.2

Fig. 2: Nominal attributes in group: Personal History

Name	Range	Percentage (%)
Date Surgery	02/01/2017 to 27/02/2017	100.00
Surgery/Ambulatory	Surgery	58.0
	Ambulatory	42.0
Benign/Malignant	Benign	19.9
	Malignant	80.1
Diagnosis	Invasive Carcinoma NST	51.7
	Fibroadenoma	9.7
	DCIS High Grade	4.5
	ILC	4.5
	Others	29.6
Laterality	R	49.7
	L	46.9
	B	3.4
Breast Intervention	Enlarged Lumpectomy	38.9
	Excisional biopsy	17.7
	MRT	8.8
	MRM	7.1
	Others	27.5
Armpit Intervention	SG	81.6
	AE	17.1
	Exertion of ganglion	1.3

Fig. 3: Nominal attributes in group: Surgery

Name	Minimum	Maximum	Mean	Standard Deviation
Days	0	26	1.710	2.972

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

Name	Range	Percentage (%)
Antibiotics	Yes	10.2
	No	89.8
Hypocoagulation	Yes	54.3
	No	45.7

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

For the target attribute, is important to consider that the majority of cases don't have any kind of complication. The distribution has 80% of cases of no complication, while the other 20% present several different complications. Later, this several types of complications will be grouped as one group, representing the possibility of having any kind of complication.

4.3 Data Preparation

At this phase, it was necessary to select and prepare the data to be used by the DM models.

Firstly, to ensure that there was no incomplete or inconsistent information, all the data with null or noise values were removed. More specifically, from the set of attributes related with surgery, "Date Cx" was removed because the information it held was not relevant to the present study and it was not useful to predict if a patient would have complications. Still in the surgery attributes, "Other Interventions" was also deleted because it was an observation like attribute, and it contained a lot of noise and scattered data. Although "Other Interventions" was deleted, the data it contained helped on the standardization of other attributes, as explained ahead. There were also a set of attributes related to the complications, with information like the date when the complication happened, the number of days a patient stayed in the post-surgery and the procedure used in order to deal with the complication. These attributes were all deleted because they were mostly compose of nulls, leading only to further noise in the models. After this attributes were excluded from the dataset, all the records underwent through a null data deletion process, leaving 150 records to be used by the data mining models.

The next step was to normalize and standardize the data. All the values that had only two possibilities, as "true" and "false", were normalized, and all the interventions made in the surgery, diagnosis and laterality were attributed to acronyms, including, for scattered information, the acronym "OI" (for "Other Interventions"). At this time, the group decided to analyze the field "Other Interventions", which allowed to complete some empty fields on "Breast Intervention" and "Armpit Intervention" with a new value created for that purpose. Besides this, and still analyzing the intervention attributes, if only one of them had a null value, the acronym for "Without Intervention" was used. In fact, in certain surgeries, doctors decide to operate via the patient arm only, while in other cases they do it through the breast, meaning that a blank register here is not a typical null.

Lastly, to evaluate the effect of oversampling the data on the performance of the DM models, 12 additional instances were created with cases of patients that had complications in the treatment process, allowing a better accuracy, recall and precision values as an outcome. This was necessary because of the low count of complications observed.

Having all the data cleaned and all the fields not null, the polynomial values were transformed to numerical ones. This allows a better and faster optimization of the models. The graphic 6 shows the distribution of **Complication** class.

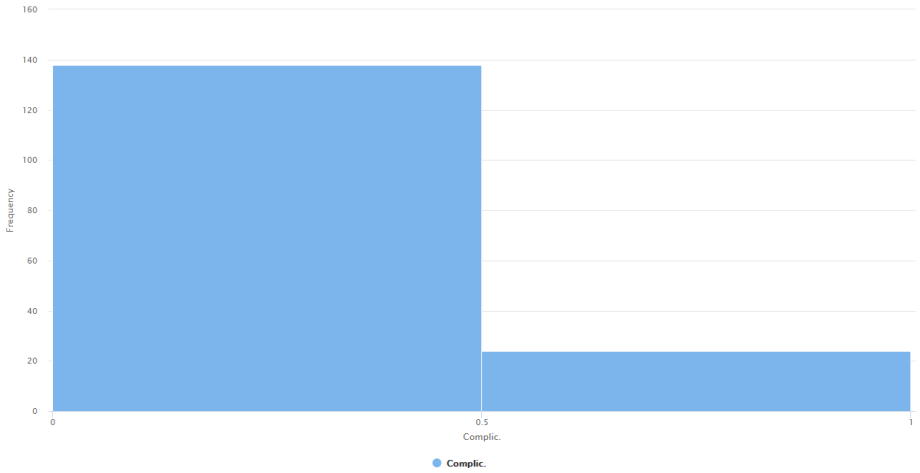

Fig. 6: figure
Distribution of the target class Complication

4.4 Modeling

With the data transformed and processed, this section now explores the different Data Mining Models (DMM) used to reach the final results. So, the DMM has the following formula:

$$DMM_n = \{A_f, S_i, DMT_y, SM_c, DA_b, TG_t\} \tag{1}$$

or, in current language,

DMM = {Approach, Scenarios, Data Mining Techniques, Sampling Models, Data Aproaches, Target}

For the approach, it was defined that this was a classification problem, and that five data mining techniques would be used for this type of theme, among them, kNN, LR, NB, RF, SVM and some other variations of this algorithms. These algorithms were executed with the default values in RapidMiner. Two data approaches were used, with and without oversampling, both tested with cross validation with 10 folds, because it is a technique that fights overfitting and it's considered of the best test methods. The target variable was defined as whether or not there are complications, and the chosen scenarios are defined below. The scenarios allow to identify which factors have the most impact in predicting complications.

- S1: {All attributes}
- S2: {Age, Tobacco, Diabetes, Immunosuppressants, Hypocoagulants, QTx NA}
- S3: {Cx/Ambulatory, Benign/Malignant, Diagnosis, Laterality, Breast Intervention, Armpit Intervention}

- S4: {Days, Antibiotic, Hypocoagulants}
- S5: {Age, Tobacco, Diabetes, Immunosuppressants, Hypocoagulants, QTx NA, Benign/Malignant}
- S6: {Cx/Ambulatory, Benign/Malignant, Diagnosis, Laterality, Breast Intervention, Armpit Intervention, Days, Antibiotic, Hypocoagulants}

The first scenario includes all the attributes, and S2 is helpful to check if the patient's background has influence on having complications. S3 is the scenario responsible to check if the patient disease has influence on having complications, and S4 means if the post-surgery details have influence on having complications. S5 tells if the patient's background, combined with the type of tumor, has influence on having complications and, finally, S6 checks if the patient disease, combined with post-surgery details, has influence on having complication.

Thus, briefly,

$$A_f = \{Classification\},\ S_i = \{S1, ..., S6\},\ DMT_y = \{kNN, LR, NB, RF, SV\},$$
$$SM_c = \{CrossValidation - 10folds\},$$
$$DA_b = \{Withoversampling, Withoutoversampling\},\ TG_t = \{Complications\}.$$

Therefore, the data mining model will be:

$$\text{DMM} = \{1 \text{ Approach, } 6 \text{ Scenarios, } 5 \text{ Techniques, } 1 \text{ Sampling Method, } 2 \text{ Data Approaches, } 1 \text{ Target}\}$$

with a total of 60 models induced.

4.5 Evaluation

In order to measure the performance of each data mining model, some values from the confusion matrix were used for model evaluation, including *accuracy* which represents the percentage of predictions that are correct, *recall* meaning the percentage of a given class that is correctly identified and *precision* which is the percentage that is correct for a given predicted class. All of the metrics were automatically calculated in *Rapid Miner*.

Tables 2, 3, 4, 5, 6 and 7 in section Attachments, represent the metrics values for the models that performed better in each scenario proposed, with and without oversampling the input data. At this point it was necessary to understand and combine all the metrics considered, in order to select the more capable model.

Table 1 presents the best model and technique obtained for each scenario. All the results are discussed right after the referred table.

Models with the best performances							
Scenario	**Model**	**OVS**	**Accuracy**	**Specificity**	**Sensitivity**	**Precision 0**	**Precision 1**
S1	SVM (evo)	Yes	96.88%	96.38%	100.00%	100.00%	82.76%
S2	NB (kernel)	Yes	75.85%	76,81%	70,83%	93,81%	34,69%
S3	NB (kernel)	Yes	71.54%	73.19%	62.50%	91.82%	28.85%
S4	NB	No	87.33%	91.30%	41.67%	94.74%	29.41%
S5	SVM (pso)	Yes	81.40%	88.41%	41.67%	89.71%	38.46%
S6	RF	Yes	86.40%	94.20%	41.67%	90.28%	55.56%

Table 1: Best model obtained for each scenario

It's very important to look at recall and precision values, as they allow to understand the real behaviour of the models created. For instance, a 90% accuracy wouldn't be real if 90% of the cases would be of one class and the model only predicted that class. However, looking only to these parameters wouldn't confirm good models, since it's very important for the general accuracy to be good. For example, it would be great if all the cases when a patient is probably getting a complication were detected, but not at a cost of always predicting that possibility.

Several tests where conducted, as seen in the previous section, and the best ones where selected for each of the possible scenarios. All the models values where calculated using cross-validation with 10 folds, which gives enough confidence to decide and choose which are the real best. It's important to know that the models were also run with 5 folds on cross-validation but the results were close to the same, only without as much confidence as the previous. For that reason, 5 folds cross-validation cases are not present in this document. In Table 1, where the best model of each scenario are described, it's possible to see several values for recall and precision. Looking for the best accuracy leads us to the scenario S1, the one which uses all the attributes to predict the possibility of complication. To confirm this accuracy, and having in mind the importance of detecting possible complications, recall should be analyzed. Recall can be interpreted as the amount of positive test samples that were actually classified as positive. With a recall of 100% for the positive cases, where a patient would have a complication, this model matches exactly with the needs of reducing post-surgery complications by predicting them earlier. The recall for the no-complication class is also very good, with a value of 96,38%. Adding to these positive values, precision , which can be intuitively understood as the classifier's ability to only predict really positive samples as positive, shows also very good results. It's possible to understand that the models tends to behave in favour of the false detection of complication, where it always classifies well the no-complication cases. No other model could come close to this results, so **SVM (evolutionary), with oversampling, with all the data columns**, proves to be the best model available.

The Support Vector Machine (Evolutionary) uses an Evolutionary Strategy for optimization. This operator is a SVM implementation using an evolutionary algorithm to solve the dual optimization problem of an SVM. Formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high-

or infinite- dimensional space, which can be used for classification, regression, or other tasks. The majority of cases with the best results were obtained with oversampling, meaning that there are many cases clustered in a single class. Oversampling also creates bias to select more samples from one class than from another, to compensate for an imbalance that is either already present in the data, or likely to develop if a purely random sample were taken. In this study, there were very few cases of complications and that needed to be addressed. Also, the bias to positively predict a complication protects patients from further problems.

Nevertheless, it's important to comprehend why the other models didn't have as good outcome as the first one. From scenario S1 to scenario S6, all data about the patient (age, smokes, diabetes, immunosuppressants, hypocoagulant and QTx NA) was taken out of consideration. From the removal of that information the model lost 10% of accuracy and, more important than that, 59% of sensivity. This shows clear signs of high importance of those features when predicting or having the possibility of having further complications. To prove this, from scenario S1 to S2, where only data about the patient was used, the accuracy decreased by about 20% but the sensitivity saw a decrease of only 30%. It becomes even more clear the importance of the patient habits and lifestyle in the post-surgery recovery process. From scenario S1 to S3, being S3 the one where only surgery and diagnosis data was used, it is possible to observe the biggest decrease in accuracy, from 97% to 72%, with sensitivity dropping almost 38%. This shows clear sign of the less important nature of this features, which is normal because all the diagnosis and treatments should be similar (patients should be treated always the same way, when having the exact same problem). However, the "Benign/Malignant" tumour indication felt like an important attribute to the group. To prove this, a comparison between scenario S2 and S5 was made, where a accuracy increase of 6% was seen. However, looking at sensitivity, there was a decrease of 29%, showing clear signs that this attribute has littler interference with post-surgery recovery. Finally, a comparison between S1 and S4 scenarios was made, where scenario S4 has only post-surgery data. The decreases in accuracy and sensitivity showed less importance of this features when compared to S2, although being justified by the low number of features present. It's evident that personal lifestyle, habits and chronic problems have way bigger importance in post-surgery recovery than other features. Regardless of that, all features are important do predict complications, as seen in scenario S1.

5 Conclusion and Future Work

Female breast cancer incidence rates have been increasing in Portugal for years, and it is most common cancer today[3]. Every year, around 6,000 cases of breast cancer are diagnosed and around 1,000 women die from the disease. Cancer treatments can be very stressful and the post-surgery complications are common. This essay has the objective to illustrate all of the methodology behind CRISP-DM, guiding all the steps that helped us to achieve a good result in this problem. Which consists of predicting if a patient will have difficulties in his/her recovery during a breast cancer treatment, helping medical staff to prevent them and with that we may be capable of reducing hospital costs, and improving patient recovery. With this study, we're also able to prove the success of Data Mining models in attain a goal as this classification.

The best results were found with Support Vector Machine algorithm, with Evolutionary optimization, obtaining a accuracy of 96.88%, a specificity of 96.38% and a sensitivity of 100%. The same model posses a negative predict value of 100% and a precision of 82.76%. With such high values for the several metrics the goal was accomplished with success. Despite of this outcomes, it wasn't possible to reach conclusion reasoning towards the most relevant features, since the best result was reached in scenario 1, and this is the complete dataset, and as such we can not define which features influence more the recovery of the patient.

One problem to bear in mind is the existence of false positives, since this problem is in the health area, and could be indicated as future work, but since 100% accuracy was achieved in predicting positive cases, there's no need to concern. Anyway, since this result was obtained with a oversampled dataset, it's a imperative task get more records of breast cancer treatments. An important factor in the development of breast cancer in an individual, is the history of his/her family, an interesting idea would be to explore if this same factor also influences the fact that the patient will or will not have complications in his/her recovery. For such, a next step would be to include data relevant to this condition in the dataset itself.

References

1. Peixoto, C.; Peixoto, H.; Machado, J.; Abelha, A.; Santos, M.F. 2018. *Iron value classification in patients undergoing continuous ambulatory peritoneal dialysis using data mining.* 4th International Conference on Information and Communication Technologies for Ageing Well and e-Health, vol. 2018-March, pp. 285-290
2. Prata, M., Peixoto, H., MacHado, J., Abelha, A. 2018. *Data mining in urgency department: Medical specialty discharge prediction.* 16th International Industrial Simulation Conference 2018, pp. 28-35
3. Global Cancer Observatory, `http://gco.iarc.fr/today/data/factsheets/populations/620-portugal-fact-sheets.pdf`
4. Towards Data Science - Machine Learning Algorithms, `https://towardsdatascience.co`
5. Alteryx - Why use SVM, `https://community.alteryx.com/t5/Data-Science-Blog/Why-use-SVM/ba-p/138440`
6. Chioka - Class Imbalance Problem, `http://www.chioka.in/class-imbalance-problem/`
7. SmartVision - What is the CRISP-DM methodology?, `https://www.sv-europe.com/crisp-dm-methodology/`
8. Machine Learning Mastery - A Gentle Introduction to k-fold Cross-Validation, `https://machinelearningmastery.com/k-fold-cross-validation/`
9. Simon's Blog - Why are precision, recall and F1 score equal when using micro averaging in a multi-class problem?, `https://simonhessner.de/why-are-precision-recall-and-f1-score-equal-when-using-micro-averaging-in-a-multi-class-problem/`
10. Vijay Kotu and Bala Deshpande, 2015, *Predictive Analytics and Data Mining*, Morgan Kaufmann, Waltham USA
11. Delen D., Walker G., Kadam A. (2004).Predicting breast cancer survivability: a comparison of three data mining methods. Artmed 2004.07.002
12. 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
13. Hian Chye Koh and Gerald Tan, 2009, *Data Mining Applications in Healthcare* ,Journal of Healthcare Information Management — Vol. 19, No. 2
14. Z. Bosnjak ; O. Grljevic ; S. Bosnjak, 2009, *CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data*, 2009 5th International Symposium on Applied Computational Intelligence and Informatics.

6 Attachments

Scenario 1					
Model	Accuracy	Recall 0	Recall 1	Precision 0	Precision 1
Naive Bayes	47.33%	50.00%	16.67%	87.34%	2.82%
Naive Bayes (kernel)*	91.36%	91.30%	91.67%	98.44%	64.71%
LR (svm)	90.00%	97.10%	8.33%	92.41%	20.00%
LR*	85.18%	94.93%	29.17%	88.51%	50.00%
RF	91.33%	99.28%	0.00%	91.95%	0.00%
RF*	90.18%	97.10%	50.00%	91.78%	75.00%
SVM (pso)	60.00%	63.04%	25.00%	90.62%	5.56%
SVM (evolutionary)*	96.88%	96.38%	100.00%	100.00%	82.76%
KNN	92.00%	100.00%	0.00%	92.00%	0.00%
KNN*	83.42%	96.38%	8.33%	85.81%	28.57%
models marked with * were built with oversampled input data					

Table 2: Metrics for models evaluation regarding scenario 1

Scenario 2					
Model	Accuracy	Recall 0	Recall 1	Precision 0	Precision 1
Naive Bayes	26.00%	23.19%	58.33%	86.49%	6.19%
Naive Bayes (kernel) *	75.85%	76.81%	70.83%	93.81%	34.69%
LR	91.33%	99.28%	0.00%	91.95%	0.00%
LR (evolutionary) *	54.74%	51.45%	75.00%	99.21%	21.18%
RF	91.33%	99.28%	0.00%	91.95%	0.00%
RF*	82.68%	95.65%	8.33%	85.71%	25.00%
SVM (libsvm)	92.00%	100.00%	0.00%	92.00%	0.00%
SVM (pso) *	77.76%	88.41%	16.67%	85.92%	20.00%
KNN	92.00%	100.00%	0.00%	92.00%	0.00%
KNN *	80.88%	94.93%	0.00%	84.52%	0.00%
models marked with * were built with oversampled input data					

Table 3: Metrics for models evaluation regarding scenario 2

Scenario 3					
Model	Accuracy	Recall 0	Recall 1	Precision 0	Precision 1
Naive Bayes (kernel) *	76.67%	82.61%	8.33%	91.20%	4.00%
Naive Bayes (kernel) *	71.54%	73.19%	62.50%	91.82%	28.85%
LR	91.33%	99.28%	0.00%	91.95%	0.00%
LR (evolutionary) *	59.34%	63.04%	37.50%	85.29%	15.00%
RF	91.33%	99.28%	0.00%	91.95%	0.00%
RF*	85.15%	97.10%	16.67%	87.01%	50.00%
SVM (evolutionary)	80.00%	85.51%	16.67%	92.19%	9.09%
SVM (evolutionary) *	77.68%	84.78%	37.50%	88.64%	30.00%
KNN	92.00%	100.00%	0.00%	92.00%	0.00%
KNN *	83.97%	98.55%	0.00%	85.00%	0.00%
models marked with * were built with oversampled input data					

Table 4: Metrics for models evaluation regarding scenario 3

Scenario 4					
Model	Accuracy	Recall 0	Recall 1	Precision 0	Precision 1
Naive Bayes	87.33%	91.30%	41.67%	94.74%	29.41%
Naive Bayes (kernel)*	86.51%	94.93%	37.50%	89.73%	56.25%
LR (evolutionary)	82.67%	86.96%	33.33%	78.95%	6.11%
LR*	86.47%	99.28%	12.50%	86.71%	75.00%
RF	90.67%	97.83%	8.33%	92.41%	20.00%
RF*	85.85%	97.83%	16.67%	87.10%	57.14%
SVM (evolutionary)	82.67%	86.96%	33.00%	94.49%	21.74%
SVM*	87.10%	100.00%	12.50%	86.79%	100.00%
KNN	92.00%	100.00%	0.00%	92.00%	0.00%
KNN*	86.47%	100.00%	8.33%	86.25%	100.00%
models marked with * were built with oversampled input data					

Table 5: Metrics for models evaluation regarding scenario 4

Scenario 5					
Model	Accuracy	Recall 0	Recall 1	Precision 0	Precision 1
Naive Bayes	26.00%	23.91%	50.00%	84.62%	5.41%
Naive Bayes (kernel)*	75.26%	76.81%	66.67%	92.98%	33.33%
LR	90.67%	98.55%	0.00%	91.89%	0.00%
LR (evolutionary)*	56.07%	50.72%	87.50%	95.89%	23.60%
RF	90.67%	98.55%	0.00%	91.89%	0.00%
RF*	84.60%	97.10%	12.50%	86.45%	42.86%
SVM (libsvm)	92.00%	100.00%	0.00%	92.00%	0.00%
SVM (pso)*	81.40%	88.41%	41.67%	89.71%	38.46%
KNN	92.00%	100.00%	0.00%	92.00%	0.00%
KNN*	83.97%	98.55%	0.00%	85.00%	0.00%
models marked with * were built with oversampled input data					

Table 6: Metrics for models evaluation regarding scenario 5

Scenario 6					
Model	Accuracy	Recall 0	Recall 1	Precision 0	Precision 1
Naive Bayes	86.00%	89.86%	41.67%	94.66%	26.32%
Naive Bayes (kernel)*	82.76%	89.86%	41.67%	89.86%	41.67%
LR (svm)	92.00%	99.28%	8.33%	92.57%	50.00%
LR*	84.04%	96.38%	12.50%	86.36%	37.50%
RF	89.33%	96.38%	8.33%	92.36%	16.67%
RF*	86.40%	94.20%	41.67%	90.28%	55.56%
SVM (pso)	82.00%	86.96%	25.00%	93.02%	14.29%
SVM (libsvm)*	87.68%	100.00%	16.67%	87.34%	100.00%
KNN	92.00%	100.00%	0.00%	92.00%	0.00%
KNN*	85.22%	97.83%	12.50%	86.54%	50.00%
models marked with * were built with oversampled input data					

Table 7: Metrics for models evaluation regarding scenario 6