Case of Study - Gastric Cancer

Francisco Matos, Francisco Oliveira, Gil Cunha, Luís Costa

University of Minho - Gualtar Campus - Department of Informatics
{a77688,a78416,a77249,a74819}@alunos.uminho.pt
http://www.di.uminho.pt/

Abstract. Gastric cancer, also known as stomach cancer, is a cancer that develops from the lining of the stomach, being the third leading cause of death by cancer.
The aim of this paper is to create the best prediction model possible for the morbidity state of a patient, in a period of 30 days, given some health conditions. This also allows to analyse the efficiency of certain treatments on patients who suffer of gastric cancer. The model is developed through Data Mining techniques and following CRISP-DM methodology.
The algorithms used were Naive Bayes, Bayes Net and J48, explored with data mining tools like *Weka, RapidMiner and ANNs*.
All the objectives have been fulfilled, as one can see in the conclusion of this paper. Several classifiers/models have been developed and compared by its performance, which Naives Bayes proved to produce the best model classifier in this study with an accuracy of 99.47% and a RMSE of 0.07.

Keywords: Prediction Model, Morbidity, Gastric Cancer, Health conditions, Data Mining, CRISP-DM, Weka, RapidMiner, ANNs, algorithms, accuracy, RMSE

1 Introduction

The objective of this study is to predict, using Data Mining, the probability of cure from gastric cancer. Data mining is the process of analysing hidden patterns of data according to different perspectives for categorisation into useful information to cut costs and increase revenue.[19] The data referred is presented in a data set, and the programs that will be used for the process are Weka and RapidMiner as well as artificial neural networks (ANNs). Using a data set provided by our professors, we intend, during this project, to use it to find the best treatment and the best odds of total cure from this disease, using the attribute "Morbilidade 30 dias" (morbidity in 30 days). The data set to be used has 154 entries and 66 attributes.

2 Background and related work

Gastric cancer, also known as stomach cancer, is a cancer that develops from the lining of the stomach. Early symptoms may include heartburn, upper abdominal pain, nausea and loss of appetite. Later signs and symptoms may include weight loss, yellowing of the skin and whites of the eyes, vomiting, difficulty swallowing and blood in the stool among others. The cancer may spread from the stomach to other parts of the body, particularly the liver, lungs, bones, lining of the abdomen and lymph nodes.

The following projects are similar to ours regarding the intention to predict morbidity within 30 days after a surgical operation:

- ***Good overall morbidity prediction with the POSSUM scoring system in patients having a total hip or knee replacement***

  The Physiological and Operation Severity Score for the enumeration of Mortality and Morbidity (POSSUM) and P(Portsmouth)-POSSUM predict the risks of complications and mortality within 30 days after surgery. The purpose of this study was to evaluate the POSSUM and P-POSSUM scoring systems in patients who underwent surgery for a total hip or knee replacement. A total of 227 patients with an elective primary total hip or knee replacement were included. The predicted postoperative morbidity was analysed in these patients and compared with the observed value 30 days after surgery. Logistic regression analysis was used to assess the correlation of variables and outcome.

- ***Mortality and cardiovascular morbidity within 30 days of discharge following acute coronary syndrome in a contemporary European cohort of patients***

  Given the increasing focus on early mortality and readmission rates among patients with acute coronary syndrome (ACS), the study was designed to evaluate the accuracy of the GRACE risk score for identifying patients at high risk of 30-day post-discharge mortality and cardiovascular readmission.

3 Methodology

This project is structured following CRISP-DM method. CRISP-DM stands for cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project and it breaks the process of data mining into six major phases. [11]

The sequence of the phases is not strict and moving back and forth between different phases as it is always required. A data mining process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions, and subsequent data mining processes will benefit from the experiences of previous ones.

3.1 Business Understanding

The main business goal of the project is to be able to predict the morbidity within 30 days after the treatment of a patient (this value indicates whether the patient is still suffering from gastric cancer or not). This prediction is relevant to know if it is necessary to do more and/or different treatments from the regular ones. It is also sensitive, so the prediction needs to be highly accurate since it can be decisive for the treatment. Applying the results of this paper in real situations can reduce the time of the patient in the health facilities, which reduces the resources needed to treat him, and increase the efficiency of the processing. It is also good for the patient due to the fact that he spends less time being treated and hospitalised.

3.2 Data Understanding

On the topic of Data Understanding, the data used here is about Gastric Cancer, provided to us by our mentors, and all attributes description and values can be seen at the Appendix A.

Attribute Selection

Regarding the selection of attributes, many were discarded due to high percentages of missing values, and others because they were not relevant for our objective.

With this in mind, the selected attributes will be presented next and the correspondent reason for the choice, keeping in mind that the attribute *Morbilidade 30 dias* ("morbility in 30 days") will be used as analysis of success or insuccess of each case (***output***).

- ***Sexo (Gender)***[3] : By studying several reports regarding the importance of gender in the recovery of gastric cancer patients, we were able to understand that there was a connection between the recovery and gender, more specifically the fact that women have a bigger survival chance. Therefore we added this attribute to our model.

- ***Motivo (Motive):*** The techniques applied in different appointments vary from one another. Thus, they may influence the prediction of gastric cancer differently, since different results are produced, some of which are more useful for

prediction than others. A prediction of early gastric cancer increases the likelihood of cure, as it is more effective to fight cancer when it is at an early stage.

- ***Estadio (Stage):*** [4] The stage at which cancer is found is critical in predicting patient survival. Detection of cancer at a less advanced stage ensures a complete recovery of 95% of patients over a 5-year period. Thus, identifying the patient's stage at the time of admission is critical to predicting patient recovery.

- ***Complicacao (Complications):*** The existence of complications during the treatment process may compromise the results of the treatment, making it more or less efficient, depending on the complications. Since the efficiency of the treatment is one of the factors that influence the value of morbidity, that is, if the disease is successfully removed from the patient, it is obvious that this attribute is important in the present case study.

- ***Primary Last:*** [5] The relapse of a patient is a good indicator, since the reappearance of the disease, after applying a certain treatment strategy, may indicate that it is not as effective when applied to the present patient. The lower the efficacy of treatment, the greater the likelihood of morbidity.

- ***Idade Grupos Etários (Age Group)***[3]: The human body, with age, loses resistance abilities against adverse factors, so it is expected that the greater the age, the lower the hypothesis of total recovery of the patient. This fact is supported by the article, which indicates that in each age group a significant decrease in the chance of recovery of the patient was detected. Therefore, it was considered relevant to account this factor for analysis.

- ***Via accesso (Access way)*** [6]: There are two different surgery techniques, laparoscopy and laparotomy. In order to understand if one of these has a better chance of fully recovering the patient we considered that the technique used should also be used in the model.

- ***Clavien Dindo***[7]: The Clavien Dindo indicator is important because it allows understanding how many patients have developed complications and in what state they are, requiring intervention or not.

- ***Local_T_P:*** The location of the cancer is important because the risk of spread is different for each site.

- ***Objectivo_cirurgia_realizada (goal of surgery performed):*** The purpose of the surgery performed is important because there are treatments that focus on the quality of life and not the extent of it.

Attribute Analysis

In this section, it is described all the details of each attribute, regarding its role, type and values.

Input:

→ Gender - Sexo

- **Type:** Numeric
- **Values:**
  - **1:** 103 instances (66.9%)
  - **2:** 51 instances (33.1%)

→ Motive - Motivo

- **Values:**
  - **0:** 39 instances (25.3%)
  - **1:** 79 instances (51.3%)
  - **2:** 2 instances (1.3%)
  - **3:** 2 instances (1.3%)
  - **4:** 2 instances (1.3%)
  - **5:** 0 instances (0.0%)
  - **6:** 30 instances (19.5%)
  - **7:** 0 instances (0.0%)

→ Stage - Estadio

- **Type:** Numeric
- **Mode:** Stage IV (37%)
- **Min:** Stage 0 (5.8%) - Very early stage of cancer development
- **Max:** Stage IV (37%) - Very advanced stage of cancer development
- **Null:** 11.7%

→ Complication - Complicacao

- **Type:** Numeric
- **Null:** 0.6%

→ Primary Last

- **Type:** Numeric
- **Null:** 4.5%

→ **Group Age - Idade Grupos Etários**

- **Type:** Numeric
- **Values:**
  - **31-40:** 7 instances (4.5%)
  - **41-50:** 24 instances (15.6%)
  - **51-60:** 29 instances (18.8%)
  - **61-70:** 44 instances (28.8%)
  - **71-80:** 35 instances (22.7%)
  - **81-90:** 14 instances (9.1%)
  - **91-150:** 1 instance (0.6%)

→ **Access way - Via acesso**

- **Type:** Numeric
- **Values:**
  - **0:** Laparoscopia, 64 instances (41.6%)
  - **1:** Laparotomia, 59 instances (38.3%)
  - **Missing:** 31 instances (20.1%)

→ **Clavien Dindo**

- **Type:** Numeric
- **Values:**
  - **0:** Without morbimortality (Sem morbimortabilidade), 117 instances (76.0%)
  - **1:** I, 4 instances (2.6%)
  - **2:** II, 5 instances (3.2%)
  - **3:** IIIa, 8 instances (5.2%)
  - **4:** IIIb, 3 instances (1.9%)
  - **5:** IVa, 1 instances (0.6%)
  - **6:** IVb, 6 instances (3.9%)
  - **7:** V, 10 instances (6.5%)

→ **Local_T_P**

- **Type:** Numeric
- **Mode:** Antro (43,5%)
- **Nulls:** 0%

→ **Goal of surgery performed - Objectivo_cirurgia_realizada:**

- **Type:** Numeric
- **Mode:** Curative (46,8%)
- **Nulls:** 20,1%

Output:

→ **30 days Morbidity - Morbilidade 30 dias**

- **Type:** Numeric
- **Values:**
  - **0:** 120 (77.9%)
  - **1:** 33 (21.4%)
  - **Missing:** 1 (0.6%)

3.3 Data Preparation

Before start the data mining process, it's very important to prepare the data set, transforming its data to the right format. The majority of algorithms work best with numeric values instead of strings. Considering all attributes selected before, we can see that we are dealing with numeric real values, so we make sure that all values are in this same format. Also, the output (Morbilidade 30 dias) is binary and this helps not only in algorithm calculations, but also in getting better results. With this in mind, we can now proceed the data preparation process, which can be divided in 3 steps: **removing null values, discretization and normalization**.

Removing null values

Some attributes had null values, which can impair the final results of the prediction models. Thus, handle with this situation is essential to the success of the models. During this step, the following attributes were considered:

- **Primary Last:** replacement of 7 null values by value 1, since it is the value of the mode (most abundant value of the attribute, about 93.5 %);

- **Complicacao:** replacement of 1 null value, by value 0, since it is the value of the mode (about 75.3 %);

- **Via acesso & Objetivo cirurgia realizada:** After analyzing the data set, it was verified that all entries that had null value in the attribute "Via Accesso" also had, correspondingly, a null value in the attribute "Objetivo cirurgia realizada". Thus, it would be assumed that if the access way (via acesso) is unknown, the goal of the surgery performed (objetivo cirurgia realizada) is also unknown. Therefore, it was decided to delete these entries (31 entries), since they had null values in both attributes, in order to reduce noise and inconsistency in the data.

- **Morbilidade_30_dias:** mostly the output must not have null values, because it influences the predictions' results more than any other. However,

this attribute only has one null value which led to its entry removal (patient 152).

Discretization

The algorithms used to build the prediction models work best with less scattered data. The more dispersion there is (the more distant the values of a certain attribute are between them), the worse the learning process is.

The data set presents attributes with many variations on the values, therefore *Weka* was used to generate graphs of the distributions of each field/attribute and thus to choose those that have more variation, in order to reduce the error hypothesis and to make the training model more efficient later.

The graphical distribution of each attribute obtained through the program can be found in the figure B, on the appendices this document. With this information, we can select those that present a greater distribution and categorize its values, reducing the dispersion on the data.

As we can observe, the attributes **Clavien Dindo, Idade Grupos Etários, Estadio and Local_T_P** have a wide range of possible values, which confers the possibility of dispersion in the data. However Clavien Dindo and Local_T_P have low dispersion - most of the values in Clavien Dindo are 0 and in the Local_T_P are 4, so it does not have too much influence. On the other hand, Estadio and Idade Grupos Etários can have a significant influence.

As it can be seen, Estadio and Idade Grupos Etários attributes have many dispersed values, so the next step is to reduce this dispersion by limiting the range of possible values of each one.

In Estadio attribute it's possible to have 10 different values, so we reduce it to 6 categories:

- **Value 0** : 0;
- **Value 1** : I (Ia + Ib);
- **Value 2** : II (Ia + Ib);
- **Value 3**: III (IIIa + IIIb + IIIc) ;
- **Value 4** : IV;
- **Value 5** : nao estadiados .

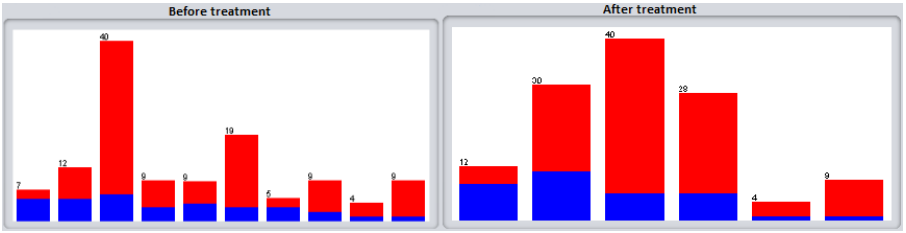

Fig. 1. Estagio - Stage: Discretization Graph

In Idade Grupos Etários it's possible to have 7 different values, so we reduce to 4 categories:

- **Value 1** : 31-50;
- **Value 2** : 51-70;
- **Value 3** : 71-90 ;
- **Value 4** : 91-150 (but there are no entries of this type because the only existing entry was deleted in previous treatments ;

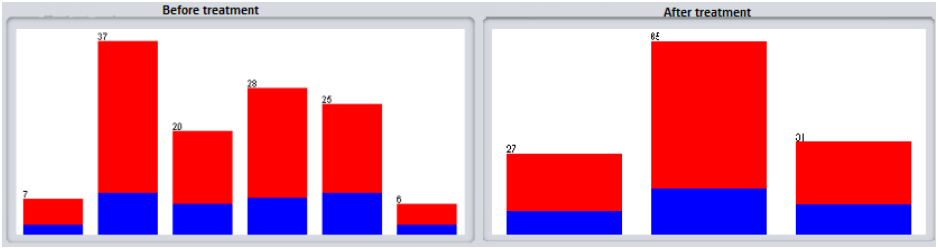

Fig. 2. Idade Grupo Etários - Group Age: Discretization Graph

Normalization

The process of *normalization* refers to the transformation of data into a standard distribution. Normalizing the data set allows to compare the effects of different factors in the business without regard to scale [8].

Several tests were performed, with and without certain types of discretization and normalization, and it was determined that the best option would be to standardize values taking into account the following formula:

$$(Z_i^k)_N = \frac{Z_i^k - Z_{min}^k}{Z_{max}^k - Z_{min}^k}$$

This is a **Feature-Scaling formula** ***(min-max normalization)*** [**9**]. It normalizes everything within a scale between 0 and 1, e.g.: Attribute Sexo (sex) varies between 0 (male) and 1 (female), instead of 1 (male) and 2 (female). This method was applied to all the fields from the data set.

Oversampling

Lastly, the oversampling technique took place, since the data set is unbalanced: it has more 0's values (90 values) than 1's (33 values) in the output. This technique replicates the cases in which morbidity was verified (value 1) so that the number of their occurrence is similar to cases where it did not occur (value 0), improving the outcomes of this study. It was replicated two times, which

leaves the data set with 90 cases with 0 value and 99 cases with 1 in the output (Morbilidade 30 dias).

4 Modeling

In this section it is explored the different Data Mining Models (DMMs) used in this project.

The final objective is to predict the morbidity 30 days after the treatment of a patient, which means that the appropriate approach is **classification** and the target is to understand if the patient has gastric cancer or not.

It was used three different data mining techniques, explored in two data mining tools - *Weka*[15] and *Rapid Miner*[16]:

- **Naive Bayes** is a simple technique for constructing classifiers that uses the Bayes theorem and assume the independence of the attributes. The label is predicted by the most probable class.

- **Bayes Net** is a type of probabilistic graphical model that uses Bayesian inference for probability computations. Bayesian networks aim to model conditional dependence, and therefore causation, by representing conditional dependence by edges in a directed graph.

- **J48** builds decision trees from a set of training data using the concept of information entropy. The splitting criterion is the normalised information gain (difference in entropy). The attribute with the highest normalised information gain is chosen to make the decision.

There were also used some meta-algorithms such as *Bagging* and *AdaBoost*, in an attempt to reduce error and increase accuracy in decision trees and other algorithms, respectively.

- **Bagging**, also called *Bootstrap aggregating*, is designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps avoiding *overfitting*.

- **Adaboost**, short for *Adaptive Boosting*, is sensitive to noisy data and outliers. In some problems it can be less susceptible to the *overfitting* problem than other learning algorithms.

The development of **Artificial Neural Networks (ANNs)** to predict the output, was studied aswell. Therefore, the data set was analysed through these soft computing and machine learning systems, using **RStudio**[17] with the package `neuralnet`[18].

The aim of ANNs is to process all the information that is provided to them, learning from it so that in the future it will be able to make decisions. Thus, artificial neural networks are a way of storing knowledge through learning and experiences. This learning is possible with the time variation of the synapse value, which determines the weight (importance) of the signal to enter the neuron (excitatory, inhibitory or null).

5 Evaluation

Now that we have the data set normalized we can proceed to create our models to properly data mine and evaluate our results, in order to achieve some knowledge.

5.1 Results

Correlations and Association Rules

By studying the correlations between attributes and extracting association rules from the data set, we can reach a more deep understanding of those attributes and maybe obtain some *hidden knowledge* (information) from the data.

First of all, it was calculated the correlations between each attribute from input with the one of output. From the correlations obtained, the only ones that deserve being mentioned are:

- **Correlation *[Morbilidade 30 dias] - [Clavien Dindo]:*** Index 0.923, positive;
- **Correlation *[Morbilidade 30 dias] - [Complicacao]:*** Index 0.710, positive;

Both are logical and very strong correlations and the fact of being positive means that the more complications and the greater the Clavien Dindo stage, the greater the probability of morbidity in 30 days.

Secondly, it was also verified with the association rules that whenever there are cases in which there are no complications, the value of the output (Morbilidade 30 dias) is 0. Thus we may conclude that the data set is unbalanced and it has few records, insufficient to get good results and make an efficient prediction model. In an attempt to cover this problem, that is why it's used the oversampling technique.

Weka

To assess the performance of each model, built with different algorithms with the data mining tool *Weka*, tables were created with some metrics, such as *Precision, Sensitivity, Specificity, Accuracy, ROC Area and Root Mean Squared Error*. This values were calculated through the confusion matrix, which presents the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). With these results, it is possible to calculate sensitivity, specificity and accuracy in order to evaluate the algorithms performance. Thus, if some models present similar results we can verify which one is the best by the value of these metrics. It should be noted that the greater the value of ROC Area, the better is the model to predict.

All the algorithms were applied using *Weka's* default settings. About **J48** algorithm, it was used with a 0.25 of *confidence factor* and 2 of *minNumObj*. It was also used with *Subtree Raising and ReducedErrorPruning*, to check if there's any difference between them:

• **Subtree raising**: Applying the subtree raising method, trying to eliminate an intermediate node of the tree, which implies a redistribution of the training instances.

• **ReducedErrorPruning**: Applying the reduced error pruning method, which takes the decision to "prune" taking into account the error estimated.

However, it was not used the **unpruned** option to avoid the affect of **overfitting**.

The data set was approached in several ways: with and without oversampling, to deal with the lack of consistency and unbalance in the data set, and with different sampling methods: *holdout sampling and cross-validation.*

Data Approach: Without Oversampling

– Sampling Method: Cross-Validation

It was used a 10-Fold Cross-Validation:

Classifier	*Precision*	*Sensitivity*	*Specificity*	*Accuracy*	*ROC Area*	*Root Mean Squared Error*
Naive Bayes	0.992	0.992	0.997	99.187 %	0.998	0.1088
Bayes Net	0.992	0.992	0.997	99.187 %	0.998	0.1058
J48	0,971	0.967	0.988	96.748 %	0.961	0.1761
J48 w/ reducedErrorPruning	0.978	0.976	0.991	97.561 %	0.978	0.1675
Naive Bayes w/ AdaBoostM1	0.976	0.976	0.953	97.561 %	0.998	0.1517
Bayes Net w/ AdaBoostM1	0.984	0.984	0.975	98.374 %	1	0.1214
J48 w/ Bagging	0.971	0.967	0.988	96.748 %	0.967	0.1753

– Sampling Method: Holdout Sampling

With holdout sampling, the data set was splitted into to: 66% of the original data set is used as a training set and the other 34% as a test set.

After obtaining the results, it was verified that all models present the same accurary value: 95.2381 %. However other metrics differ, but not significantly. Overall, despite all models have similar performance, **BayesNet** (with and without AdaBoostM1)is the best, due to presenting the lowest Root Mean Squared Error (0.1647) and greater ROC Area (0.994).

Data Approach: With Oversampling

– Sampling Method: Cross-Validation

Classifier	*Precision*	*Sensitivity*	*Specificity*	*Accuracy*	*ROC Area*	*Root Mean Squared Error*
Naive Bayes	0.990	0.990	0.988	98.9418 %	0.998	0.0866
Bayes Net	0.995	0.995	0.994	99.4709 %	0.998	0.0811
J48	0.980	0.979	0.977	97.8836 %	0.961	0.1444
J48 w/ reducedErrorPruning	0.985	0.984	0.983	98.4127 %	0.983	0.1271
Naive Bayes w/ AdaBoostM1	0.995	0.995	0.994	99.4709 %	1	0.0711
Bayes Net w/ AdaBoostM1	0.995	0.995	0.994	99.4709 %	1	0.0715
J48 w/ Bagging	0.980	0.979	0.977	97.8836 %	0.965	0.1443

– Sampling Method: Holdout Sampling

After obtaining the results, it was verified that all models present the same accurary value: 98.4375 %. However other metrics differ, but not significantly. Overall, despite all models have similar performance, **J48** (with and without Bagging)is the best, due to presenting the lowest Root Mean Squared Error [RMSE] (0.1236) and greater ROC Area (0.984).

Without the attributes Complicacao and Clavien Dindo, the accuracy of the models drops between 60% $\sim$ 70%, depending on the algorithm. This shows, once again, that these attributes are important to predict the morbidity state of the patient, in 30 days.

RapidMiner

The results, using RapidMiner and the data set that was provided were as follow:

Classifier	*Accuracy*	*Root Mean Squared Error*
Naive Bayes	95.24%	0.218
Bayes Net	97.62%	0.156
J48	97.62%	0.154
J48 w/Pruning	97.62%	0.154

As using the oversampled data set we were able to get the following results:

Classifier	*Accuracy*	*Root Mean Squared Error*
Naive Bayes	97.56%	0.153
Bayes Net	99.12%	0.097
J48	98.68%	0.115
J48 w/Pruning	94.74%	0.195

Artificial Neural Networks

In order for an artificial neural network process the data set (used with oversampling), there was a need to form a training set (to train the network) and a test set (to test its predictions). So the data set was shuffled to break any tendentious patterns and splitted in 66% as training and 34% as test.

In a previous phase, the processed data set was analysed to evaluate the attributes' level of significance. These are the results:

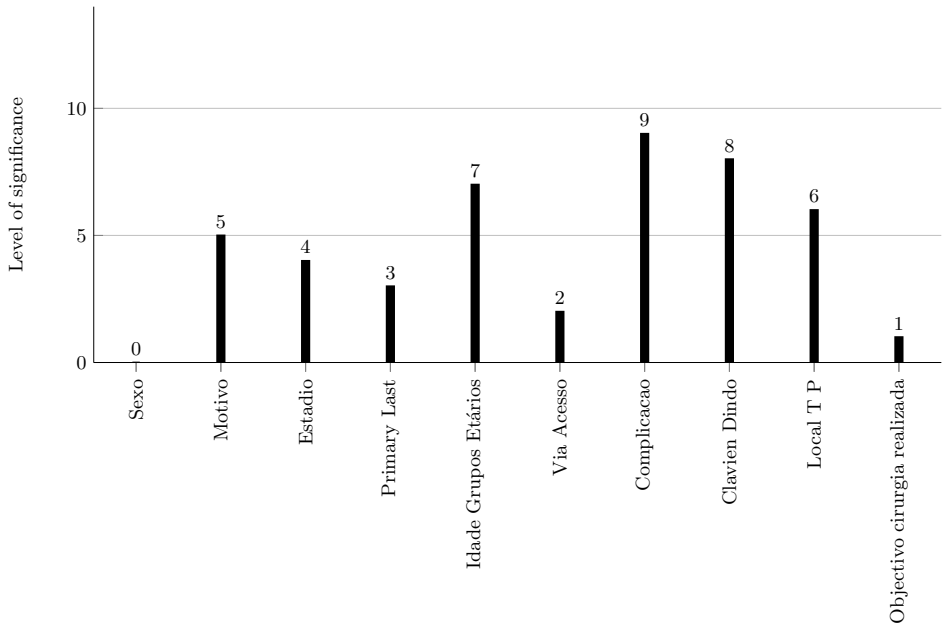

Fig. 3. Levels of significance of the attributes in the *data set* considered

As it can be observed in the graph above, Complicacao and Clavien Dindo are the ones with most significance, corroborating previous conclusions.

The next step was to find out the best architecture for the artificial neural network, i.e. the number of intermediate layers, number of nodes per layer, initial weights and which algorithm to apply. As for the number of layers, a brief survey was conducted with the objective of deciding how many layers should be used to obtain the best performance. It was consider one to two intermediate layers, apart from input and output base layers. As for the number of nodes per layer, it ranged from 1 to 10. The initial weights were established after, ranging from -5 to 5.

The algorithms applied were: ***"backprop"***- back propagation, ***"rprop+"*** and ***"rprop-"*** which refer to a flexible back propagation with or without backsliding, respectively.

The following table exposes the obtained results:

	1° Layer	**2° Layer**	**RMSE**	**Accuracy**
	1	0	0.7806	43.750 %
	1	9	0.6374	59.375 %
	2	10	0.1250	96.875 %
backprop	3	9	0.1768	98.438 %
	1	0	0.7500	43.750 %
	1	9	0.6374	0.5938 %
	2	10	0.1768	96.875 %
rprop+	3	9	0.1250	98.438 %
	1	0	0.7500	43.750 %
	1	9	0.6374	0.5938 %
	2	10	0.1768	96.875 %
rprop-	3	9	0.1250	98.438 %

5.2 Discussion

Firstly, we shall analyse the results obtained with *Weka*. Using an holdout sampling the results were equal across the classifiers, with an accuracy of 95.2381%, with the only difference being other metrics, meaning that *BayesNet* was the best since it had the smallest RMSE . Using cross-validation, we were able to deduce that *Bayes Net* was once again the best classifier, since it had the least value of RMSE and biggest ROC Area. Using oversampling, the results were a little different, and the best classifier was in fact *NaiveBayes* when using cross-validation sampling. However, with an holdout sampling every classifier had the

same accuracy, pointing out *J48*, which had the least Root Mean Squared Error and highest ROC Area with 0.984.

Next, using *Rapid Miner* and using Accuracy and Root Mean Squared Error as metrics to evaluate which classifier had the best result, we can see that, when using the given data set, the best classifier was *J48* with or without pruning. Yet, once again, using an oversampled data set we verify that the best classifier this time was actually *Bayes Net* with the best accuracy and smallest RMSE.

Finally, using ANNs, we were able to also obtain very accurate results using 3 nodes in the first layer in all algorithms used, getting an accuracy of 98.438% and a RMSE of 0.1250. Despite this very high value, it still was not higher then the data mining results we observed above.

Looking at all results, we can compare the classifiers that were used in both Weka and RapidMiner. The classifiers were, Naive Bayes, Bayes Net, J48 and J48 with Pruning. With weka we can see better values of accuracy in every classifier, however this is not a concluding point on Weka being more accurate then Rapid Miner.

The positive impact of the oversampling on the results proves how unbalanced the data set is. Looking back at the data set it becomes obvious some of its flaws, for example, there aren't enough entries, it is based regarding negative morbidity cases, there are a lot of missing values present and access way is directly linked to surgery goal, which means that if one of them is null, the other also is. However, the biggest flaw in the data set is the fact that it answers itself using the attributes *complications* and *clavien dindo*. It is possible to almost always correctly predict the result of the treatment making it a very simple model.

The following graphics show that whenever the clavien dindo is 0 or the complications are 0 (equivalent to no complications) the patient fully recoveries from the disease. Blue represents no morbidity within 30 days, while red means the patient is still not cured.

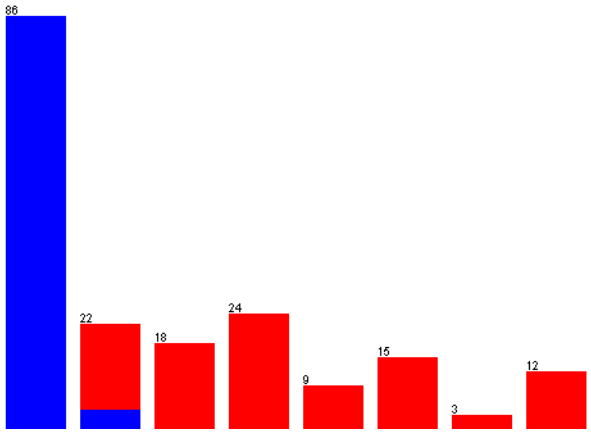

Fig. 4. Clavien Dindo Relative to Morbidity

The column on the left represents a value of 0 in Clavien Dindo.

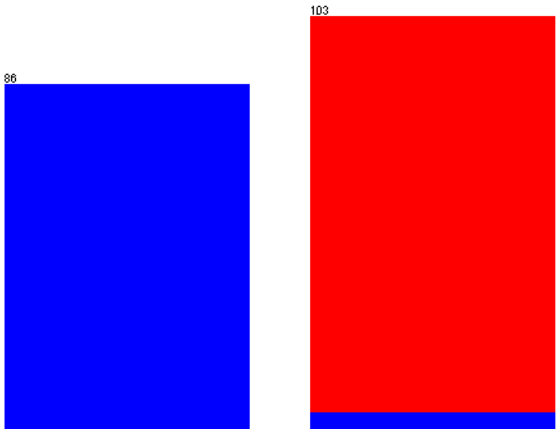

Fig. 5. Complications Relative to Morbidity

The column on the left represents no complications.

The next image shows the decision tree generated by the algorithm J48 in Weka. Here we can see the weight of complications in the probability of full recovery from the patient. This shows how unbalanced the data set is, and how little info we can extract from it.

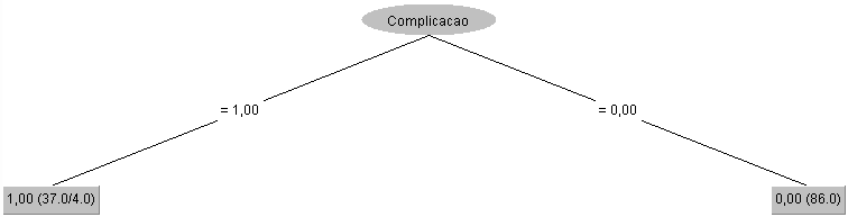

Fig. 6. Weka's J48 - Decision Tree

Overall, all results considered and comparing the values of the tables in **section 5.1**, we may conclude that the best model/classifer created, with this particular data set, is **Naive Bayes with AdaBoostM1**, giving the best combination of accuracy (99.4709%), ROC Area (1) and RMSE (0.0711).

6 Conclusion and Future Work

This study could have been improved if the data set was larger and more balanced allowing us to have a bigger knowledge pool and deliver better results and analysis.

Thus we may conclude that the data set is unbalanced and it has few records, insufficient to get good results, make an efficient prediction model and gather any definite conclusions, however, it still shows some of the power in data mining, since it was able to correctly predict results despite smaller data sets ratio for training.

Looking at the main focus of this study, predict whether a patient will recover after a treatment or not, we can say we achieved good results, despite the situation, since the accuracy values were very high. This was achieved best when using Naive Bayes with AdaBoostM1 that gave us a 99.4709% accuracy, a ROC Area of 1 and a value of root mean squared error of 0.0711. With the algorithms we were able to successfully identify that complications and clavien dindo as the key attributes to properly predict the morbidity in 30 days.

As future work, an attempt to collect more data is in view, so to create a larger and more versatile data set, retrieving more information and, consequently, more knowledge from more trust worthy analyses and results. New strategies for data mining could also be implemented in order to have a wider range of results to evaluate.

Despite all the adversities we still reached some conclusions, and understand the use of Data Mining, and its powerful capabilities in medicine to detect patterns and allow early treatment of the patient.

References

1. Portal Regional da BVS, *Câncer Gástrico - Gastric Cancer*
   http://docs.bvsalud.org/biblioref/2018/05/883263/ca-gastrico-finalb_rev.pdf
   Accessed in 26/04/2019

2. AEOP - Associação de Enfermagem Oncológica Portuguesa, *Linhas de Consenso - Consensos & Estratégias —2011 - CANCRO GÁSTRICO*
   https://www.aeop.pt/ficheiros/13b396a4dbd836a1dd8a254922ac7247.pdf
   Accessed in 26/04/2019

3. Yang D, Hendifar A, Lenz C, Togawa K, Lenz F, Lurje G, *Survival of metastatic gastric cancer: Significance of age, sex and race/ethnicity*
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3397601/
   Accessed in 10/05/2019

4. Arnis Kirshners, Serge Parshutin and Marcis Leja, *Research in application of data mining methods to diagnosing gastric cancer*
   https://link.springer.com/chapter/10.1007/978-3-642-31488-9_3
   Accessed in 10/05/2019

5. Yasuo Ohashi and Mitsuru Sasako, *Survival after recurrence in patients with gastric cancer who receive S-1 adjuvant chemotherapy: exploratory analysis of the ACTS-GC trial*
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5910584/
   Accessed in 10/05/2019

6. FA FANG,FENG HAN,YIN-LU DING and HAI-JIANG WANG, *Comparison of laparoscopy-assisted surgery and laparotomy for treating locally advanced distal gastric antral cancer*
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786921/
   Accessed in 10/05/2019

7. Kyung-Goo Lee, Hyuk-Joon Lee, Jun-Young Yang, ... ,
   Risk Factors Associated with Complication Following Gastrectomy for Gastric Cancer: Retrospective Analysis of Prospectively Collected Data Based on the Clavien–Dindo System
   https://link.springer.com/article/10.1007/s11605-014-2525-1
   Accessed in 10/05/2019

8. Stephanie - Statistics How To, *About Normalized Data*
   https://www.statisticshowto.datasciencecentral.com/normalized/
   Accessed in 21/05/2019

9. Wikipedia, *Feature scaling*
   https://en.wikipedia.org/wiki/Feature_scaling
   Accessed in 21/05/2019

10. Wikipedia, *Cross-industry standard process for data mining*
    https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining#cite_note-Harper06
    Accessed in 12/06/2019

11. sv-europe, *What is the CRISP-DM methodology?*
https://www.sv-europe.com/crisp-dm-methodology/
Accessed in 12/06/2019

12. Wikipedia, *Gastric Cancer*
https://en.wikipedia.org/wiki/Stomach_cancer
Accessed in 12/06/2019

13. NCBI, PubMed Center - morbidity prediction with the POSSUM scoring system
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6145357/
Acessed in 12/06/2019

14. NCBI, PubMed Center - Mortality and cardiovascular morbidity within 30 days
http://www.revportcardiol.org/pt-pdf-S0870255115001250
Acessed in 12/06/2019

15. Weka 3
https://www.cs.waikato.ac.nz/ml/weka/
Acessed in 13/06/2019

16. RapidMiner
https://rapidminer.com/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6145357/
Acessed in 13/06/2019

17. RStudio
https://www.rstudio.com/
Acessed in 13/06/2019

18. cran.r-project, *Neural Net*
https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf
Acessed in 13/06/2019

19. Devmedia, *Data Mining*
https://www.devmedia.com.br/conceitos-e-tecnicas-sobre-data-mining/19342
Acessed in 13/06/2019

Appendices

Appendix A

Data Understanding - Attributes' Description

1. **ID:** Value that identifies the patient.

2. **Sexo:** Value that identifies the patient's gender.

3. **Proveniencia:** Value that identifies where the patient comes from. This attribute has the following values:
   0 - Unknown;
   1 - Extern consult;
   2 - Emergency services;
   3 - Hospitalization Surgery;
   4 - Internment of another specialty;
   5 - Transferred from another hospital;

4. **Motivo:**Value that identifies the reason for the patient's arrival. This attribute has the following values:
   0 - HDB (High digestive bleeding);
   1 - Endoscopic exams;
   2 - Other imaging tests;
   3 - Follow-up neoplasia operated - relapse;
   4 - Cholestasis;
   5 - Analytical results + Clinical;
   6 - Clinical (without HDB);
   7 - Followed by Neoplasia consultation in palliation;

5. **Data_nascimento:** Date of birth of the patient.

6. **HPreOp:** Study of the biological tissues of the patient before the operation. This attribute has the following values:
   0 - Not available;
   1 - Papillary Adenocarcinoma;
   2 - Tubular Adenocarcinoma;
   3 - Other Adenocarcinoma;
   4 - Signet ring cells adenocarcinoma;
   5 - Adenosquamous adenocarcinoma;
   6 - Squamous cell carcinoma;
   7 - Small Cell Carcinoma;
   8 - Undifferentiated Carcinoma;

9 - Others;
10 - Adenocarcinoma not elsewhere classified;
11 - Nonspecific/inflammatory changes;
12 - MALT lymphoma;
13 - High-grade dysplasia;
14 - Diffuse large B-cell lymphoma;

7. **HPreOp:** Study of the biological tissues of the patient after the operation. This attribute has the following values:
0 - Not available;
1 - Papillary Adenocarcinoma;
2 - Tubular Adenocarcinoma;
3 - Other Adenocarcinoma;
4 - Signet ring cells adenocarcinoma;
5 - Squamous cell carcinoma;
6 - Squamous cell carcinoma;
7 - Small Cell Carcinoma;
8 - Undifferentiated Carcinoma;
9 - Others;
10 - Adenocarcinoma not elsewhere classified;
11 - Nonspecific/inflammatory changes;
12 - MALT lymphoma;
13 - High-grade dysplasia;
14 - Diffuse large B-cell lymphoma;

8. **Ganglios_ressecados:** Number of resected lymph nodes;

9. **Margens:** Value that identifies the type of surgical margin. This attribute has the following values:
0 - R0;
1 - R1;
2 - R2;

10. **Grau_diferenciacao:** Value that identifies the degree of differentiation. This attribute has the following values:
0 - Little differentiated;
1 - Moderately differentiated;
2 - Well differentiated;
99 - Unknown;

11. **Ganglios_invadidos** Number of invaded ganglia.

12. **T:** Type of primary tumor. This attribute has the following values:
0 - Tx;
1 - T0;

2 - Tis;
3 - T1;
4 - T1a;
5 - T1b;
6 - T2;
7 - T3;
8 - T4;
9 - T4a;
10 - T4b;

13. **N:** Value indicating whether the cancer has spread to the lymph nodes. This attribute has the following values:
    0 - Nx;
    1 - N0;
    2 - N1;
    3 - N2;
    4 - N3;

14. **M:** Value that indicates whether the cancer has spread to distant organs. This attribute has the following values:
    0 - Mx;
    1 - M0;
    2 - M1;

15. **Estadio:** Value that indicates the stage of cancer of the patient. This attribute has the following values:
    0 - 0;
    1 - Ia;
    2 - Ib;
    3 - IIa;
    4 - IIIa;
    5 - IIIb;
    6 - IV;
    7 - IIb;
    8 - IIIc;
    9 - Not staged;

16. **Local_T_P**: Location of the primary tumor. This attribute has the following values, indicating the location of the tumor:
    0 - Cardia;
    1 - Fund;
    2 - Small Curvature;
    3 - Great Bend;
    4 - Antro;
    5 - Small + Large curvature;

6 - Incisura angularis;
7 - Gastric stump - Great curvature;
8 - Gastric stump - Small curvature;
9 - Body;
10 - Gastric stump - unspecified;
11 - Gastric stump - Cardia;
12 - Gastric stump - Bottom;
13 - It is not known - they are metastases;
14 - Diffuse - Plastic Linite;
15 - Small curvature + Den;
16 - Gastric stump - Body;

17. **Data_Cx**: Date of surgery. Follow the DD / MMM / AA structure (example: January 14, 1999)

18. **Cx_Prop**: Type of surgery proposed. This attribute has the following values, indicating the proposed surgery to the wearer:
    0 - Laparoscopic Total Gastrectomy;
    1 - Laparoscopic Subtotal Gastrectomy;
    2 - Laparoscopic Atopic Gastrectomy;
    3 - Laparoscopic Gastrojejunostomy;
    4 - Total Gastrectomy + Laparoscopic Distal Esophagectomy;
    5 - Subtotal Gastrectomy + Laparoscopic Distal Esophagectomy;
    6 - Total Gastrectomy;
    7 - Subtotal Gastrectomy;
    8 - Atopic Gastrectomy;
    9 - Gastrejejunostomy Laparotomy;
    10 - Total Gastrectomy + Distal esophagectomy;
    12 - Subtotal Gastrectomy + Distal esophagectomy;
    13 - Exploratory laparoscopy;
    14 - Totalization of Gastrectomy;
    15 - Laparoscopic Gastrectomy Totalization;
    16 - Exploratory laparotomy;
    17 - Laparoscopy feeding jejunostomy;
    18 - Gastrojejunostomy laparoscopy;
    19 - Jejunostomy of Laparoscopic Feeding;
    20 - Laparoscopic gastrostomy;

19. **Cx_realizado**: Type of surgery performed. This attribute has the following values, indicating the surgery performed to the user:
    0 - Laparoscopic Total Gastrectomy;
    1 - Laparoscopic Subtotal Gastrectomy;
    2 - Laparoscopic Atopic Gastrectomy;
    3 - Laparoscopic Gastrojejunostomy;
    4 - Total Gastrectomy + Laparoscopic Distal Esophagectomy;

5 - Subtotal Gastrectomy + Laparoscopic Distal Esophagectomy;
6 - Total Gastrectomy;
7 - Subtotal Gastrectomy;
8 - Atopic Gastrectomy;
9 - Gastrejejunostomy Laparotomy;
10 - Total Gastrectomy + Distal esophagectomy;
12 - Subtotal Gastrectomy + Distal esophagectomy;
13 - Exploratory laparoscopy;
14 - Totalization of Gastrectomy;
15 - Laparoscopic Gastrectomy Totalization;
16 - Exploratory laparotomy;
17 - Laparoscopy feeding jejunostomy;
18 - Gastrojejunostomy laparoscopy;
19 - Jejunostomy of Laparoscopic Feeding;
20 - Laparoscopic gastrostomy;

20. **Reconstruçao**: type of traffic reconstruction. This attribute has the following values, indicating the type of Reconstruction performed to the user:
    0 - Billroth I;
    1 - Billroth II;
    2 - Y-en-Roux;
    3 - N/A;

21. **Alta_Bloco**: Reason for Discharge. This attribute has the following values, indicating the reason for hospital discharge:
    0 - Deceased;
    1 - Nursing High;
    2 - ICU discharge;

22. **Recessao_Linfatica** Lymphadenectomy. This attribute has the following values, indicating the type of Lymphadenectomy:
    0 - D0;
    1 - D1;
    2 - D2;
    3 - D2 + Distal pancreatectomy + Splenectomy;
    4 - D1 + 1;
    5 - D1 + 2 + Splenectomy;
    6 - D2 + Splenectomy;
    7 - D1 + 1 + Splenectomy;

23. **Objetivo_cirurgia_proposta:** Objective of surgery performed. This attribute has the following values, indicating the purpose of the surgery performed:
    0 - Healing;
    1 - Palliative;

2 - Staging;
3 - Emerging;

24. **Data_Internamento**: Date of hospitalization

25. **Intercorrencias_posOP:** Intercurrences / postoperative complications. It is the first of the series of postoperative complications. This attribute has the following values, indicating the type of complication:
0 - without intercurrences;
1 - occlusion;
2 - anastomosis dehiscence;
3 - hemorrhage;
4 - splenic infarction;
5 - abdominal abscess;
6 - pleural effusion;
7 - pneumotorax;
8 - urinary retention;
9 - evisceration;
10 - acute pancreatitis;
11 - esophago-pleural fistula;
12 - entero-cutaneous fistula;
13 - pelvic abscess;
14 - ischemia / venous insufficiency of loops;
15 - intraoperative pneumothorax;
16 - pneumonia.

26. **Dieta_Oral**: Oral diet start day

27. **Data_Alta**: Date of hospital discharge

28. **Resultado_Internamento**: Result of hospitalization
0 - Deceased;
1 - Improved;
2 - Same state;

29. **Destina_Alta**: Destination of discharge
0 - N/A;
1 - External consultation;
2 - Another institution;
3 - Lost for follow-up;

30. **Tratamento_complicacoes:** Treatment of postoperative complications. Refers to the treatment of the first occurrence of said sequence. This attribute has the following values, indicating the type of treatment:
0 - doctor;

1 - surgical;
2 - non-surgical drainage;
3 - untreated;
4 - NA.

31. **Recidiva (Relapse):**Reappearance of the symptoms of a disease that had already been cured in the same individual. This attribute has the following values, indicating whether or not there was a relapse:
    0 - no;
    1 - yes.

32. **Data_Recidiva (Relapse Date) :** Date of relapse;

33. **Complicacoes_pos_alta (Post-discharge Complications) :** Complications after discharge. It's the sequence of complications after discharge. This attribute has the following values, indicating the type of complication:
    0 - without intercurrences;
    1 - anastomosis dehiscence;
    2 - anastomosis stenosis;
    3 - occlusion / subocclusion;
    4 - surgical wound infection;
    5 - HDA;
    6 - feverish syndrome;
    7 - hemoperitoneum;
    8 - tumor stenosis;
    9 - incisional hernia;
    10 - splenic infarction;
    11 - hepatic infarction;
    12 - acute pancreatitis;
    13 - evisceration;
    14 - enterocutaneous fistula;
    15 - pelvic abscess;
    16 - probe occlusion.

34. **Faleceu (passed away) :** Passed away by gastric cancer. This attribute has the following values, indicating the situation of the death of the patient:
    0 - died for other reasons;
    1 - yes;
    2 - was already palliative;
    3 - unknown/no computer record;
    4 - did not passed away.

35. **Data_faleceu (Death date) :** Date of death of the patient;

36. **Recidiva_Nova_Abordagem (Relapse New Approach)**: New approach of healing, after relapse. This attribute has the following values, indicating the type of new approach:
    0 - palliative;
    1 - curative.

37. **Faleceu_Complicacao (Death by complication)**:Death due to postoperative complication. This attribute has the following values, indicating if the patient died due to a postoperative complication:
    0 - no;
    1 - yes.

38. **Tratamento_compl_PA (Complications Treatment):** Treatment of complications after discharge. Refers to the treatment of complications sequence, after discharge. This attribute has the following values, indicating the type of treatment:
    0 - medical;
    1 - surgical;
    2 - non-surgical drainage;
    3 - untreated;
    4 - NA.

39. **Complicacoes_30dias (Complications 30 days) :** complications 30 days after the operation. This attribute has the following values, indicating the type of complication:
    0 - without intercurrences;
    1 - occlusion;
    2 - anastomosis dehiscence;
    3 - hemorrhage;
    4 - splenic infarction;
    5 - abdominal abscess;
    6 - pleural effusion;
    7 - pneumotorax;
    8 - urinary retention;
    9 - evisceration;
    10 - acute pancreatitis;
    11 - esophago-pleural fistula;
    12 - entero-cutaneous fistula;
    13 - pelvic abscess;
    14 - ischemia / venous insufficiency of loops;
    15 - intraoperative pneumothorax;
    16 - pneumonia.

40. **Trat_complicacoes_30dias (Complications Treatment 30 days):** Treatment of complications that occurred within 30 days after the operation. This attribute has the following values, indicating the type of treatment:

0 - medical;
1 - surgical;
2 - non-surgical drainage;
3 - untreated;
4 - NA.

41. **Primary Last :** Value that identifies if the cancer has reocurred in the patient.

42. **Idade Internamento :** Value that identifies at what age the patient was admitted in the hospital.

43. **Tempo Internamento :** Value that identifies how much time the patient spent in the hospital.

44. **Tempo morte :** Value that indicates the number of days that have passed until the patient dies after hospitalisation

45. **Grupo gg ressecados :** Groups of ganglia that underwent surgery.
    0 - 0 - 15
    1 - 16 - 29
    2 - 30+

46. **Idade grupos etários :** Value representing the age group to which the patient belongs.
    1 - 31-40
    2 - 41-50
    3 - 51-60
    4 - 61-70
    5 - 71-80
    6 - 81-90
    7 - 91-150

47. **Via Acesso :** Value indicating the type of surgery performed
    0 - laparoscopia
    1 - laparotomia

48. **Complicação :** Value indicating whether health complications occurred during treatment.

49. **Idade doentes :** Value indicating age of patients.

50. **Morbilidade 30 dias :** Value indicating whether the patient is still suffering from the disease up to 30 days after treatment

51. **Mortalidade 30 dias :** Value indicating whether the patient died within 30 days after treatment

52. **Morbimortalidade :** Value that indicates if the patient has the disease and died up to 30 days after treatment

53. **Clavien Dindo :** Value indicating the patient's complications after the operation and the type of treatment needed
    0 - 0 - Does not carry the disease
    1 - l - No medical intervention required
    2 - ll - Medication Prescription
    3 - llla - Intervention without anesthesia
    4 - lllb - Intervention with anesthesia
    5 - lVa - Defective organ
    6 - lVb - Defective organs
    7 - V - Death of the patient

54. **filter Type :** Value indicating which patients were operated on and used as a laparotomy approach.

55. **ASA :** Value that indicates the degree of health of the patient, from healthy to extremely sick.

56. **Operado :** Value indicating whether the patient was operated

57. **Test :** Value indicating the existence of another pathology in the patient

Appendix B

Discretization - Attributes' Distribution Graphs

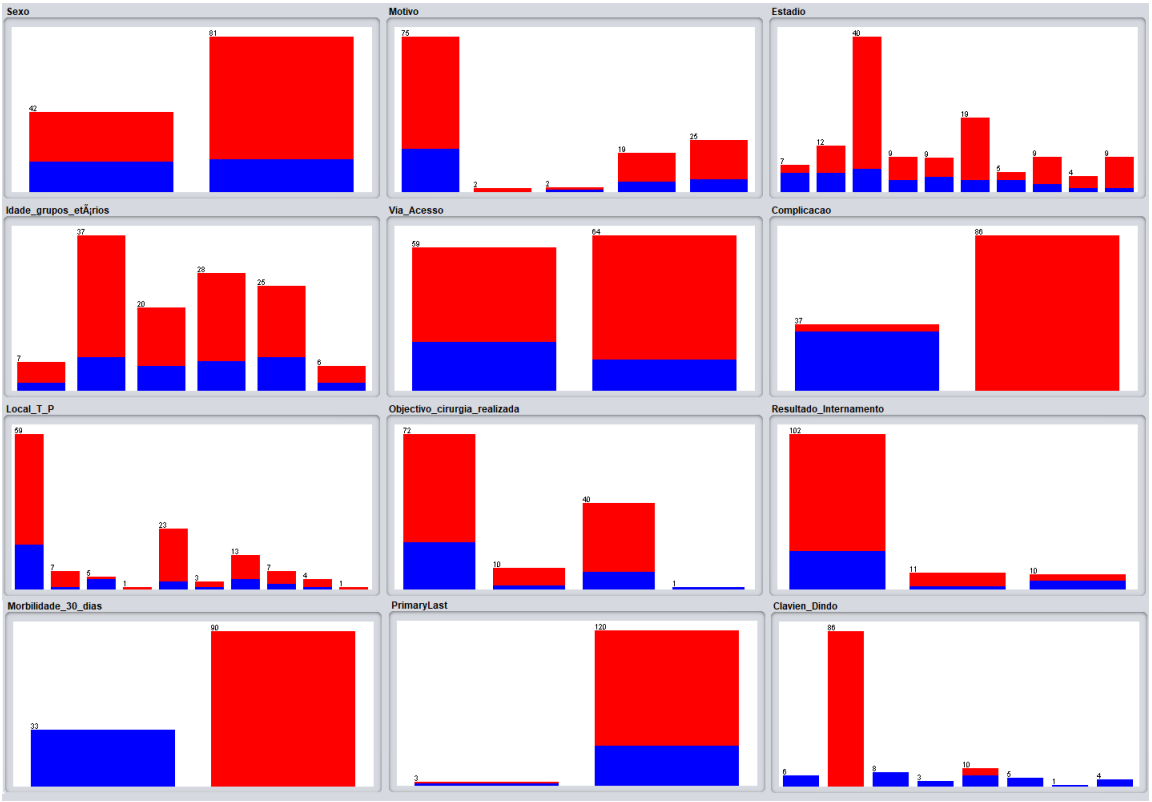

Fig. 7. Attributes' Distribution Graphs