THE APPLICATION OF DATA MINING BY CLASSIFICATION IN A DATABASE OF NOTIFIED COVID-19 CASES IN MANAUS-AM

Fábio Gomes Cantuário
fcantuario80@gmail.com
Graduate department of the university center - FAMETRO, Manaus - AM, Brazil

Luiz Eduardo Santos de Araújo
14eduard@gmail.com
Graduate department of the university center - FAMETRO, Manaus - AM, Brazil

Rilmar Pereira Gomes
rilmargomes@hotmail.com
Graduate department of the university center - FAMETRO, Manaus - AM, Brazil

David Barbosa de Alencar
david002870@hotmail.com
Graduate department of the university center - FAMETRO, Manaus - AM, Brazil

Abstract

This scientific article aims to present information on the cases of comorbidity that most aggravate the symptoms of SARS-CoV-2 (Covid 19) with data extracted from the database of the official website of the Ministry of Health, which defined a system to monitor the information detected in the diagnoses of each patient. Since the beginning of the pandemic, the city of Manaus has suffered great consequences in relation to the SARS-CoV-2 virus (Covid-19). predicting patients at higher risk of death. We describe the origin and spread of the virus and the use of the SGBD software MySql and MySql Workbench to improve data in the selection and pre-processing, with the resources of the weka tool for knowledge learning, ending with the objective achieved in the classification of comorbidities that further aggravate the clinical conditions.

Keywords: COVID-19; MySql; KDD; Weka; Workbench; Data Mining.

1. INTRODUCTION

Every day, society produces tons of virtual data through the most different means and forms, like a mountain of information, which, if not interpreted, only serves to occupy space in technological devices. Due to this huge amount of information it is challenging to extract knowledge from this data and the traditional techniques to analyze it end up becoming impractical.

The process of discovering hidden connections and predicting future trends dates back to the 1990s and encompasses the areas of statistics, artificial intelligence and machine learning, thus triggering the emergence of the data mining area.

The use of data mining can bring several benefits to companies and organizations, making good use of information, understanding its relevance, evaluating possible results and accelerating decision making. Data mining seeks to discover relationships and has applications in several areas, such as: assessing the reputation of companies through social networks; bank credit approval; discover market trends or classify a new species of plant or animal.

The interest in the development of artificial intelligence applications for the health area and the potential of such techniques in the prevention of pandemics has been growing in recent years. The pandemic for the new coronavirus (SARS-CoV-2), presents itself as one of the greatest challenges ever faced by modern society, which requires quick decision-making to face this international public health emergency.

The Ministry of Health of Brazil, through the Health Surveillance Secretariat, has been monitoring the Severe Acute Respiratory Syndrome (SRAG) in Brazil, since the Influenza A (H1N1) pandemic in 2009. Since then, this surveillance has been implemented in network of Influenza and other respiratory viruses, which first acted only with sentinel surveillance of Influenza Syndrome (SG).

In 2020, the surveillance of COVID-19 (Coronavirus Disease) was incorporated into this network. COVID-19 is the human infection caused by the new coronavirus, which is a family of viruses, where some infect humans. Another coronavirus had already caused an epidemic in 2002 in China, causing Severe Acute Respiratory Syndrome (known by the acronym in English Sars).

On March 11, 2020, Covid-19 was characterized by the World Health Organization (WHO) as a pandemic. In Brazil, the first case of Covid-19 was confirmed on February 26. Amazonas confirmed the first case of the disease on March 13, 2020.

Until February 8, 2021, the country registered 9,550,301 cases and 232,248 deaths. In the State of Amazonas, according to the State Health Surveillance Foundation, through data from the COVID-19 Daily Case Bulletin of February 8, 2021, it registered 691,241 reported cases, 283,658 confirmed cases and 9,116 deaths.

DATASUS is the IT department of the Unified Health System (SUS) and aims to provide the legacy of the epidemiological databases of SARS, the Influenza and other respiratory viruses surveillance network, from the beginning of its implementation in 2009 until today, with the incorporation of COVID-19 surveillance. Its objective is to collect, process and disseminate health information in the country, in addition to assisting in the support of information technology and systems necessary for the planning, operation and control of SUS agencies. The official system for the registration of cases and deaths of SARS is the Influenza Epidemiological Surveillance Information System (SIVEP-Gripe).

People with comorbidities are considered to be at risk for the disease caused by the Coronavirus, as are elderly people, smokers and those who are considered immunodepressed, such as those who are undergoing cancer treatment with chemotherapy or who have undergone bone marrow transplantation. When treated correctly, diseases considered comorbid to Coronavirus have their risks reduced, as is the case of former cancer patients without evidence of cancer or people with HIV with undetectable viral load.

Due to the large volume of data produced during the pandemic, among them the data from the SIVEP-Gripe system, a manual and traditional analysis would become unfeasible, thus, the main motivation of the present work lies in the fact of the efficiency of the use of mining of data in these large volumes of data generated by the pandemic in Brazil, more specifically in the city of Manaus, in the period from 01/01/2021 to 02/02/2021, and the application of the classification task for knowledge extraction. To achieve this motivation, a more appropriate algorithm to perform data mining will be selected, the reliability of the generated model will be analyzed, provide subsidies to determine the efficiency of the applicability of the generated model.

1. THEORETICAL FRAMEWORK

Before the application of the basic tools of this article is executed, it is important to conceptualize and understand the main subjects that support the proposed theme, of which they will be: Knowledge Discovery in Databases - KDD (Knowledge Discovery); Data Mining Method and its tasks, as well as the Naive Bayes algorithm, the Confusion Matrix and the Weka Tool, which is one of the most simple and widely used data mining tools in academia.

2.1 Knowledge Discovery in Databases - KDD (Knowledge Discovery)

The KDD according to Fayyad et al (2006) is a non-simple way to extract information from a large amount of data stored and to be analyzed. Previously unknown, but potentially useful.

The traditional method of transforming data into knowledge depends on the interpretation and manual analysis of one or more analysts, serving as an interface between data and users. When data volumes grow too large, this type of analysis becomes impractical, slow, expensive and very subjective.

These data, in most cases, are in binary format, are typically bulky and difficult to understand. Thus, KDD aims to transform this data, which can be in the form of: short reports, a descriptive model of the process that generated the data, or a predictive model to estimate future cases.

KDD refers to the entire discovery process, while Data Mining or Data Mining refers to a step in the process that consists of applying data analysis using algorithms that respect the limitations of computational capacity thus producing patterns (CAMILO and SILVA, 2009).

The KDD process consists of a sequence of steps as shown in figure 1 and that must be performed sequentially, because at the end of each step, the result obtained serves as an aid to the next step, being able to repeat previous steps whenever necessary and that will be discussed below.

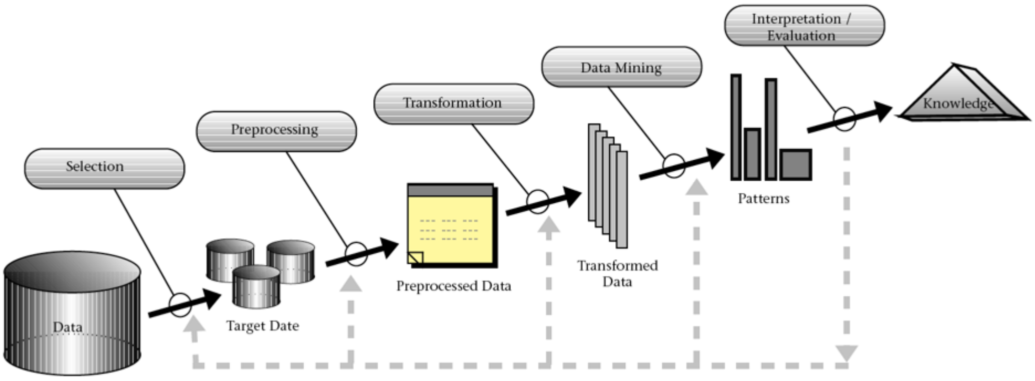


Figure 1: The KDD steps

Source: danielteofilo.wordpress.com/2015/02/16/kdd-knowlegde-discovery-in-database /

2.1.1 Data

The data is all untreated and unprocessed records in which it has a large volume containing all characteristics or attributes. Data are codes that constitute the raw material of the information, that is, it is the untreated information that still has no relevance. They represent one or more meanings of a system that alone cannot transmit a message or represent some knowledge (SILVA, 2007).

2.1.2 Selection

It is a set of data or sample of data with which the discovery process will be carried out. For Zanardi (2007), in this phase, a set of data is selected or a subset of attributes or data instances is focused on, in which the discovery must be made.

2.1.3 Pre-Processing

These data may go through a pre-processing stage where problems such as noise and incomplete data will be dealt with. Data Cleanup methods can be Value Correction, Measurement Error and Data Collection, Noise, External, Inconsistent Values and Duplicate Data.

In the pre-processing phase, data cleaning is performed using operations such as noise elimination and incorrect or incomplete data is neglected. Zanardi (2007) says that it collects the necessary information for modeling and correcting noise and for strategies for manipulating lost data fields.

2.1.4. Transformation

In the data transformation phase, projection and reduction operations may be used, reducing the number of variables under consideration or finding representations of the data that do not vary.

Mendes (2011), says that data transformation consists of converting data to a format interpretable by Data Mining tools, associated with Data Integration, which can come from multiple sources, meet the specific needs of Data Mining.

2.1.5. Data Mining

Data mining is one of the phases of KDD, a step where they are applied to mining techniques, often repetitively, which look for patterns and rules hidden in the data according to the desired objective.

As the main step in the KDD process, Data Mining consists of the application of specific algorithms on the data, in order to abstract knowledge. According to Nangiyalil (2007), these algorithms use inductive learning techniques on databases and are able to extract knowledge through examples, applying interactive methods repeatedly.

Santos (2009, p. 1) defines data mining as the name given to a set of techniques and procedures that tries to extract information of a higher semantic level from raw data, in other words, allowing the analysis of large volumes of data. data for knowledge extraction. (GALVÃO and MARIN, 2009).

2.1.6. Assessment

Phase that interprets the mined and evaluated standards, often in the form of graphs or reports, selecting the useful knowledge of this entire process. (OLIVEIRA; CARVALHO, (2008) FREITAS; LAVINGTON, (1998). It may be necessary to go back to some previous step in an iterative process.

2.2 Data Mining Method

The methods are existing technologies, regardless of the data mining context, since, applied in KDD, they produce good results in the health area, transforming data into useful knowledge.

In addition to the models, Data Mining can be characterized by the choice and application of the algorithms to be applied. Among the main existing methods, we will work with the Classification Rules.

There are several methods that can be used in Data Mining, and their application depends on the type of input data and the type of information desired as an output. The methods are divided between verification and discovery methods. Among those of discovery, they are divided between methods of prediction and description. Predictive tasks seek to forecast a target attribute value. The descriptors try to find patterns, correlations, trends, groups, anomalies.

According to Fayyad et al (1996), the objectives of prediction and description can be achieved using specific data mining methods that are classified according to the models: Predictive and Descriptive. However, for this research only the Predictive format will be used, namely:

- Predictive: Able to predict future or unknown values through some variables. That is, anticipating future behavior or value. Methods that generally use this model are those of Classification, Regression and Detection of deviations.

2.2.1 Classification

The predictive function aims to map a data item from several predefined items. It is one of the most common tasks of data mining and seeks to identify which class a particular record belongs to (SILVA, 2014). In this task, the model analyzes the set of records provided, with each record already containing the indication to which class it belongs, in order to "learn" how to classify a new record (supervised learning). Mendes (2011) explains that this approach allows the construction of classification models, that is, being able to recognize the function that describes a certain class to which an item belongs, from a set of input data.

The classification task consists of a set of records or objects, in which each record is represented by two types of attributes (x, y), in which X represents a set of attributes, of the discrete or continuous type; and Y represents a special attribute of the discrete type, usually referred to as a class.

Thus, Mendes (2011) teaches that each of the classifying methods uses an algorithm called the Learning Algorithm that aims to build models with good generalization capacity and / or that accurately predict the labels of previously unknown record classes. These algorithms work with the following elements:

- Training set: A set of records with labels of known classes is provided;
- Classification model: It is built to relate the set of attributes to the class labels.
- Test set: These are records with labels of unknown classes.

The performance of the classification model is evaluated after the application of the selected algorithm. This assessment is based on counting the test records provided by the model, checking which ones are correct or not. Thus, the performance of the algorithm is calculated by the metric of precision and error rate. It is worth mentioning that most classification algorithms seek models that, when applied to the set of tests, have greater precision or the lowest error rate. (MENDES, 2011)

2.2.2 Naive Bayes Algorithm

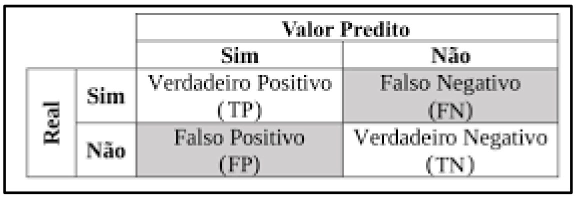
Silva (2014) defines the Naive Bayes algorithm as a method that uses conditional probability based on Thomas Bayes' theorem. Camilo and Silva (2009) say that according to Bayes' theorem, it is possible to find the probability that a certain event will occur, given the probability of another event that has already occurred: Probability (B given A) = Probability (A and B) / Probability (A).

This method is easy to build, does not require any complicated iterative parameters and estimation schemes. It can be easily applied to large data sets, easy to interpret, generally performs the classification task well and is robust enough. For the application of the Naive Bayes algorithm, the data must pass through the pre-processing transformation stage where they will be discretized.

2.2.3 Confusion Matrix

In the Classification task, a matrix of predicted results is used in comparison with the original classes observed, known as a confusion matrix.

This matrix seeks to understand the relationship between successes and errors that the model presents and presents itself as follows in Figure 2.



		Valor Predito	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Figure 2: The confusion matrix

Source: diegonogare.net/2020/04/performance-de-machine-learning-matriz-de-confusao /

The results can be summarized in four initial values, being:

- True Positive (TP): class originally predicted and observed is part of the positive class. (*True Positive* – TP) originally predicted and observed class is part of the positive class.

- ● False Positive (FP): predicted class returned positive but the original observed was negative;
- ● True Negative - TN: the predicted and observed values are part of the negative category;
- ● False Negative (False Negative - FN): it represents that the predicted value resulted in the negative class but the original observed was of the positive class.

2.2 Weka Tool (Waikato Environment for Knowledge Analysis).

Weka, according to Witten, Frank and Hall (2011), is free software developed in Java, widely used for academic data mining, its strong point is the classification task, but it can work with the association rules and data clusters .

Weka can access data from the database through the JDBC connector and its own data files with the extension ARFF (Attribute-Relation File Format). It was developed to be a collection of Machine Learning algorithms, where many of them incorporate concepts of artificial intelligence.

2.3 MYSQL

MySql is a relational DBMS (Database Manager System), which uses SQL (Structured Query Language, or translating, Structured Query Language). MySQL is also multiuser and multitasking. (WELLING, THOMSON, 2003).

2.4.1 *Workbench*

Workbench is a unified visual tool for database architects, developers and DBAs. MySQL Workbench provides data modeling, SQL development and comprehensive administration tools for server configuration, user administration, backup and more. MySQL Workbench is available on Windows, Linux and Mac OS X. (SURHONE, TIMPLEDON, MARSEKEN, 2010).

2. MATERIALS AND METHODS

The methodology applied in this research can be defended as a case study, Ellram (1996) also adds the possibility of using case studies to analyze past events in similar cases and make predictions.

The collection of data on the subject to be presented, regardless of the techniques used, through statistical sources. It also refers to applied, descriptive and qualitative research using an inductive method. (BARDIN, 2009; GERHARDT and SILVEIRA, 2009).

There were also consultations with studies that have already elaborated works on the subject of data mining, KDD, mining algorithms, SGBD MySQL and the Weka Tool. Thus, the research path involved the practical application of the Data Classification technique, which is one of the phases of the KDD process, to allow the generation of a data model from the application of the Naive Bayes algorithm to extract data patterns .

In order to carry out this analysis, it was necessary to capture data from the system made available by the Ministry of Health's website, DataSus, which provides information that serves to support objective analyzes of the health situation, evidence-based decision making and the elaboration of health action programs..

Therefore, initially, these data were exported to the MySQL database management system, to carry out the treatment and pre-processing of the data. Then, these were treated and analyzed in the data mining tool Weka, where the classification algorithms were applied to extract a test model. For this, the Naive Bayes algorithm was used, generating models for tests and the confusion matrix for comparing the results.

Checking the spreadsheet and studying the basis, an attempt was made to analyze whether the sex, age and the comorbidities recorded are related to the severity of the patient's clinical evolution, which may be death or cure. Among the listed data are the sociodemographic, related to sex, age and other indicators related to the individual and the epidemiological ones, which include clinical and laboratory information, such as results of diagnostic tests, hospitalizations and symptoms of diseases.

The KDD phases carried out in practice will be addressed below:

3.1 Dados

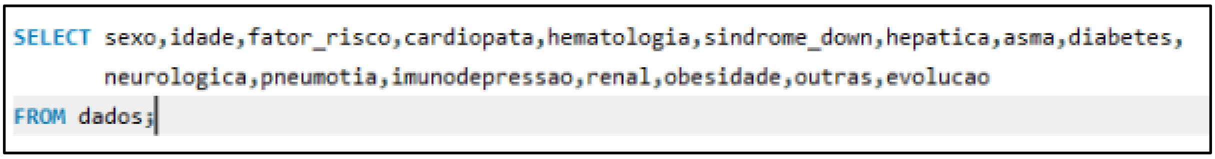
The base where the work was carried out was taken from the official records of the Ministry of Health website, DATASUS, in which the file INFLUD21-08-02-2021 was extracted in .csv format, which refers to the data registered in the SRAG system 2021 - Severe Acute Respiratory Syndrome Database.

This program manages the rates of hospitalizations and diagnoses of patients seen, with 98,975 records, containing data from COVID-19 from all over Brazil in the period from January 1, 2021 to February 8, 2021 (39 days). This short period of time is justified by the large volume of data in the system, the low computational processing power and the intention of the work being only to demonstrate the applicability of the concepts.

The INFLUD21-08-02-2021 file contains data from notification records from all over Brazil, with various information recorded in the COVID-19 case notification. Among these data are sociodemographic data, related to sex, age and other indicators related to the individual; and epidemiological ones, which include clinical and laboratory information, such as diagnostic test results, hospitalizations and disease symptoms.

3.2 Selection

Using the MySQL Workbench tools in conjunction with MySQL DBMS, the records contained in the CSV file were imported into a database called SIVEP and the data was added to the CADASTRO nomenclature table, containing the total record extracted from the site that was on total of 98975 records. The selection phase is the learning of the application domain, where the data to be analyzed and grouped with which it was intended to work were selected, as shown in the figure 3.



```
SELECT sexo,idade,fator_risco,cardiopata,hematologia,sindrome_down,hepatica,asma,diabetes,
       neurologica,pneumotia,imunodepressao,renal,obesidade,outras,evolucao
FROM dados;
```

Figure 3: Sql Selection Script

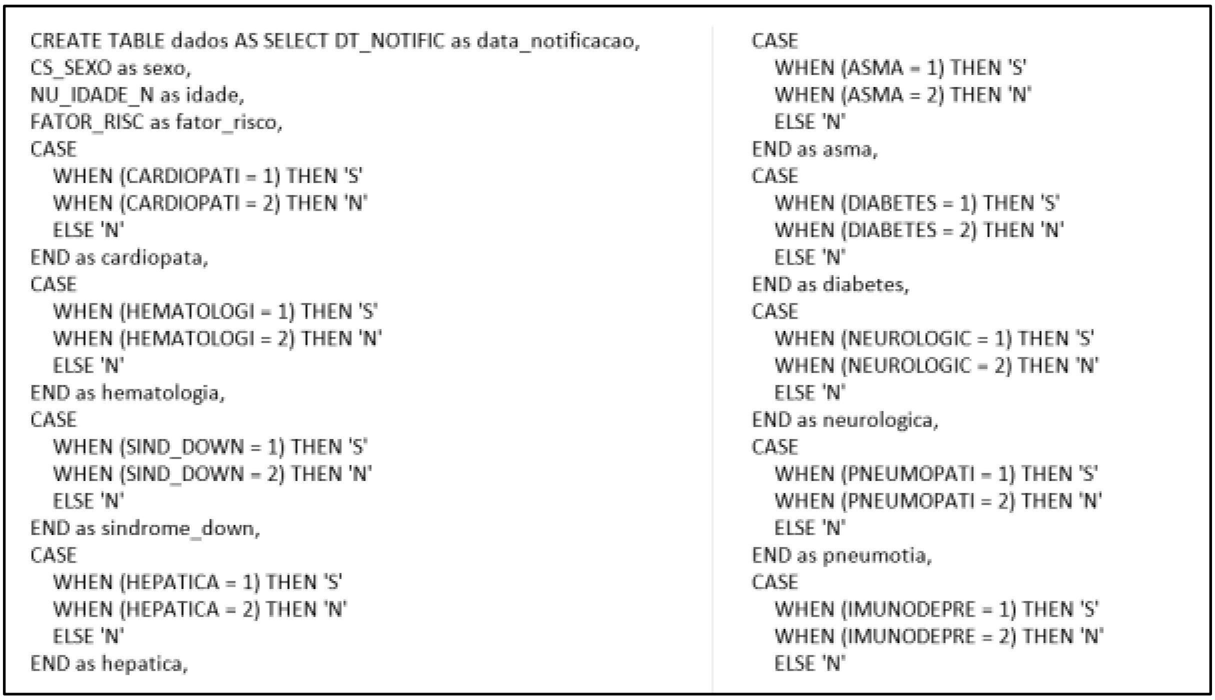
Source: The Author.

3.3 Pre-Processing and Data Transformation

In this phase, data cleaning and basic noise removal operations took place. In this case, there was a need to generate a new table called DATA and in this table there was the pre-processing of the necessary records to be used in the next phase. The values of the fields were defined as 1-Yes, 2-No and 9-Ignored for the comorbidity fields, and the name of the fields was changed for a better understanding of the data. Some fields had blank values. Blank fields with the value 'N' were processed

The objective of this phase was the search for useful attributes in the data, taking into account the defined objectives and the use of transformation methods in order to reduce the effective number of variables under consideration.

In the selection and pre-processing phases, a script was created to create a new DATA table, with the converted records, eliminating noise, as shown in figure 4 below.



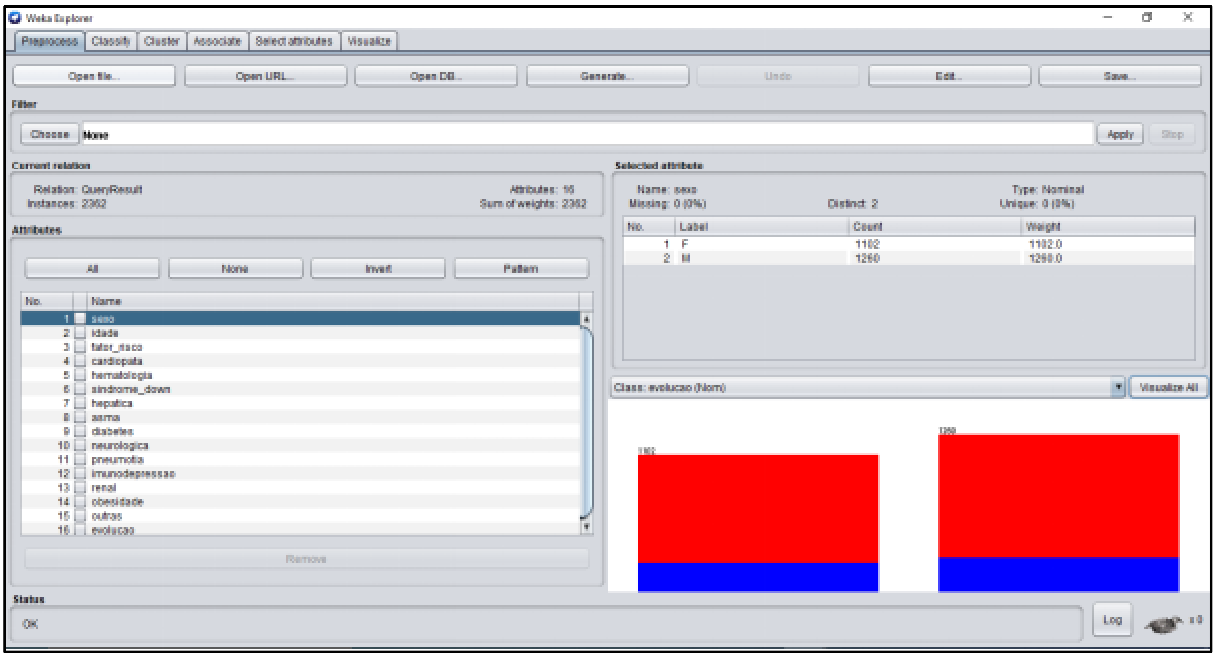
```
CREATE TABLE dados AS SELECT DT_NOTIFIC as data_notificacao,
CS_SEXO as sexo,
NU_IDADE_N as idade,
FATOR_RISC as fator_risco,
CASE
   WHEN (CARDIOPATI = 1) THEN 'S'
   WHEN (CARDIOPATI = 2) THEN 'N'
   ELSE 'N'
END as cardiopata,
CASE
   WHEN (HEMATOLOGI = 1) THEN 'S'
   WHEN (HEMATOLOGI = 2) THEN 'N'
   ELSE 'N'
END as hematologia,
CASE
   WHEN (SIND_DOWN = 1) THEN 'S'
   WHEN (SIND_DOWN = 2) THEN 'N'
   ELSE 'N'
END as sindrome_down,
CASE
   WHEN (HEPATICA = 1) THEN 'S'
   WHEN (HEPATICA = 2) THEN 'N'
   ELSE 'N'
END as hepatica,
CASE
   WHEN (ASMA = 1) THEN 'S'
   WHEN (ASMA = 2) THEN 'N'
   ELSE 'N'
END as asma,
CASE
   WHEN (DIABETES = 1) THEN 'S'
   WHEN (DIABETES = 2) THEN 'N'
   ELSE 'N'
END as diabetes,
CASE
   WHEN (NEUROLOGIC = 1) THEN 'S'
   WHEN (NEUROLOGIC = 2) THEN 'N'
   ELSE 'N'
END as neurologica,
CASE
   WHEN (PNEUMOPATI = 1) THEN 'S'
   WHEN (PNEUMOPATI = 2) THEN 'N'
   ELSE 'N'
END as pneumotia,
CASE
   WHEN (IMUNODEPRE = 1) THEN 'S'
   WHEN (IMUNODEPRE = 2) THEN 'N'
   ELSE 'N'
```

Figure 4: Pre-Processing and Transformation Script

Source: The Author

The transformation phase is necessary to present the values that are recognized by the tool. In this case, we used the WEKA tool, which has 4 ways to load data to work. The Open DB option was selected, which corresponds to opening a database connection and performing the select operation of the fields in the studied table as shown in the sql script in figure 3.

When entering this script code, the screen as shown in figure 5 is shown, where the graph corresponds to the Procedure table, in which the blue color refers to the HEALING frequency in the table, and the red color refers to the DEATH value in relation to the table..



Filter: None

Relation: QueryResult
Instances: 2362
Attributes: 16
Sum of weights: 2362

No.	Name
1	sexo
2	idade
3	fator_risco
4	cardiopata
5	hematologia
6	sindrome_down
7	hepatica
8	asma
9	diabetes
10	neurologica
11	pneumotia
12	imunodepressao
13	renal
14	obesidade
15	outras
16	evolucao

Name: sexo
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	F	1102	1102.0
2	M	1260	1260.0

Class: evolucao (Nom)

OK

x 0

Figure 5: Pre-Processing and Transformation Script

Source: The Author

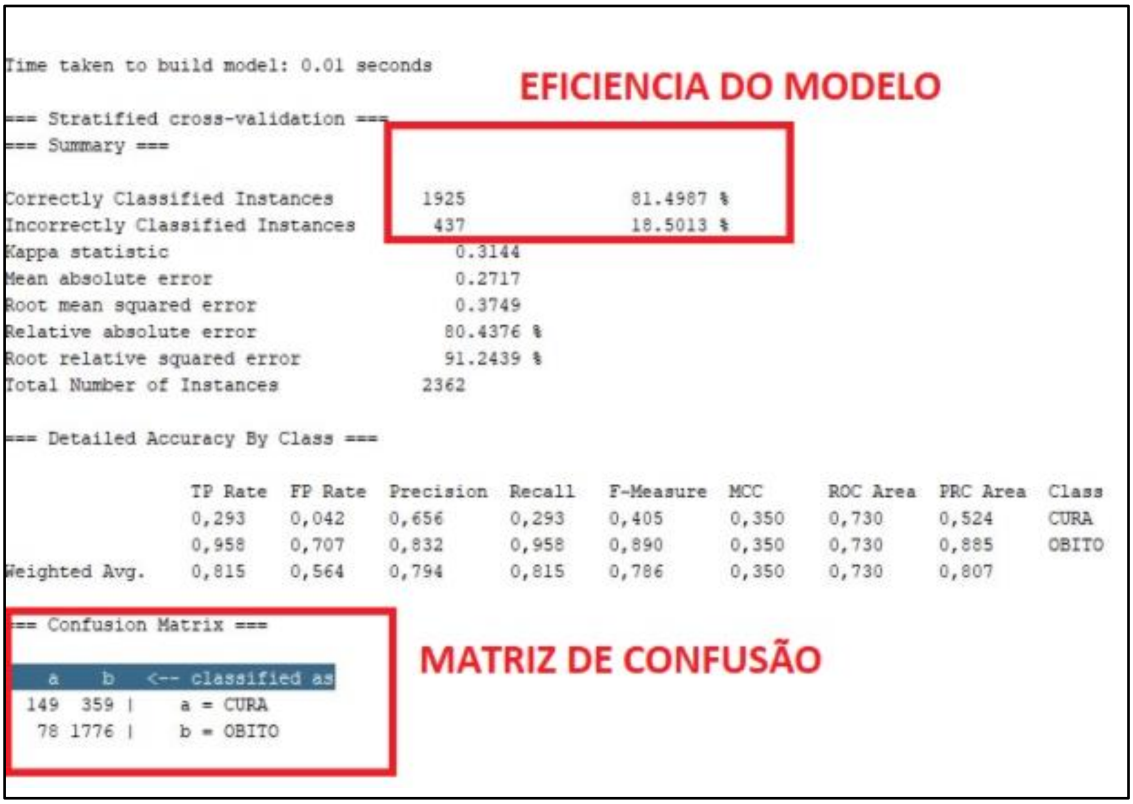
3.4 Data Mining

The task chosen for data mining of the selected base was that of classification, presented in sequence using the Naive Bayes algorithm. The validation method used in the execution of the algorithms was Cross-validation of k parts. As mentioned by Tan et al (2009), this approach segments the data into k equal partitions, where during each execution one of these partitions is chosen for testing, while the others are used for training. This procedure is repeated k times, so that each partition is used for testing exactly once. The total error is found by adding the error of all k executions. (MENDES, 2011).

3. RESULTS AND DISCUSSIONS

After applying the Classification method to the database with the Naive Bayes algorithm, the Weka tool provided a sequence of data with the results obtained.

2362 instances from the 16 attributes selected in the KDD selection phase were analyzed using cross-validation for model creation and data testing with the model



```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1925               81.4987 %
Incorrectly Classified Instances       437               18.5013 %
Kappa statistic                          0.3144
Mean absolute error                      0.2717
Root mean squared error                  0.3749
Relative absolute error                 80.4376 %
Root relative squared error             91.2439 %
Total Number of Instances             2362

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,293    0,042    0,656      0,293    0,405      0,350    0,730     0,524     CURA
                 0,958    0,707    0,832      0,958    0,890      0,350    0,730     0,885     OBITO
Weighted Avg.    0,815    0,564    0,794      0,815    0,786      0,350    0,730     0,807

=== Confusion Matrix ===

    a    b   <-- classified as
  149  359 |    a = CURA
   78 1776 |    b = OBITO
```

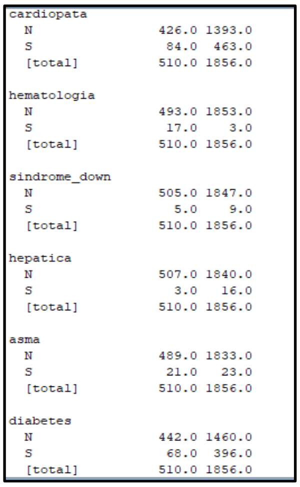
Figure 6: Results after applying the algorithm

Source: The Author

Figure 6 shows that the efficiency of the model was 81%, correctly classifying 1925 instances and 437 incorrect classifications. The Confusion Matrix presented the results of the classifications, detailing where the number of instances were considered correct and incorrect according to the class considered, which was the evolution attribute.

- True Positive: 149Falso Positivo : 78.
- True negative: 1776.
- ● False Negative: 359.

It was observed that the cases that most express the values of risk of death in relation to comorbidities are: Other Morbidities, Diabetes and Cardiopaths, as shown in figure 6.

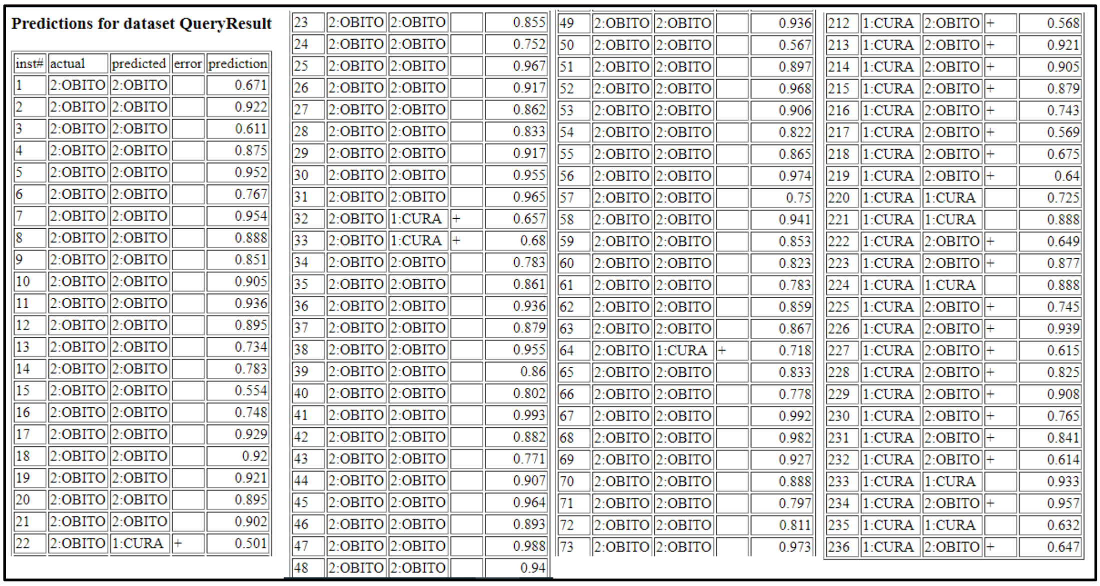


```
cardiopata
  N              426.0 1393.0
  S               84.0  463.0
  [total]        510.0 1856.0

hematologia
  N              493.0 1853.0
  S               17.0    3.0
  [total]        510.0 1856.0

sindrome_down
  N              505.0 1847.0
  S                5.0    9.0
  [total]        510.0 1856.0

hepatica
  N              507.0 1840.0
  S                3.0   16.0
  [total]        510.0 1856.0

asma
  N              489.0 1833.0
  S               21.0   23.0
  [total]        510.0 1856.0

diabetes
  N              442.0 1460.0
  S               68.0  396.0
  [total]        510.0 1856.0
```

Figure 6: Results after applying the algorithm

Source: The Author

The Weka tool makes it possible to export to other result formats such as HTML, in which the generated model and test can be analyzed side by side, as shown in figure 8 below:



Predictions for dataset QueryResult

inst#	actual	predicted	error	prediction
1	2:OBITO	2:OBITO		0.671
2	2:OBITO	2:OBITO		0.922
3	2:OBITO	2:OBITO		0.611
4	2:OBITO	2:OBITO		0.875
5	2:OBITO	2:OBITO		0.952
6	2:OBITO	2:OBITO		0.767
7	2:OBITO	2:OBITO		0.954
8	2:OBITO	2:OBITO		0.888
9	2:OBITO	2:OBITO		0.851
10	2:OBITO	2:OBITO		0.905
11	2:OBITO	2:OBITO		0.936
12	2:OBITO	2:OBITO		0.895
13	2:OBITO	2:OBITO		0.734
14	2:OBITO	2:OBITO		0.783
15	2:OBITO	2:OBITO		0.554
16	2:OBITO	2:OBITO		0.748
17	2:OBITO	2:OBITO		0.929
18	2:OBITO	2:OBITO		0.92
19	2:OBITO	2:OBITO		0.921
20	2:OBITO	2:OBITO		0.895
21	2:OBITO	2:OBITO		0.902
22	2:OBITO	1:CURA	+	0.501
23	2:OBITO	2:OBITO		0.855
24	2:OBITO	2:OBITO		0.752
25	2:OBITO	2:OBITO		0.967
26	2:OBITO	2:OBITO		0.917
27	2:OBITO	2:OBITO		0.862
28	2:OBITO	2:OBITO		0.833
29	2:OBITO	2:OBITO		0.917
30	2:OBITO	2:OBITO		0.955
31	2:OBITO	2:OBITO		0.965
32	2:OBITO	1:CURA	+	0.657
33	2:OBITO	1:CURA	+	0.68
34	2:OBITO	2:OBITO		0.783
35	2:OBITO	2:OBITO		0.861
36	2:OBITO	2:OBITO		0.936
37	2:OBITO	2:OBITO		0.879
38	2:OBITO	2:OBITO		0.955
39	2:OBITO	2:OBITO		0.86
40	2:OBITO	2:OBITO		0.802
41	2:OBITO	2:OBITO		0.993
42	2:OBITO	2:OBITO		0.882
43	2:OBITO	2:OBITO		0.771
44	2:OBITO	2:OBITO		0.907
45	2:OBITO	2:OBITO		0.964
46	2:OBITO	2:OBITO		0.893
47	2:OBITO	2:OBITO		0.988
48	2:OBITO	2:OBITO		0.94
49	2:OBITO	2:OBITO		0.936
50	2:OBITO	2:OBITO		0.567
51	2:OBITO	2:OBITO		0.897
52	2:OBITO	2:OBITO		0.968
53	2:OBITO	2:OBITO		0.906
54	2:OBITO	2:OBITO		0.822
55	2:OBITO	2:OBITO		0.865
56	2:OBITO	2:OBITO		0.974
57	2:OBITO	2:OBITO		0.75
58	2:OBITO	2:OBITO		0.941
59	2:OBITO	2:OBITO		0.853
60	2:OBITO	2:OBITO		0.823
61	2:OBITO	2:OBITO		0.783
62	2:OBITO	2:OBITO		0.859
63	2:OBITO	2:OBITO		0.867
64	2:OBITO	1:CURA	+	0.718
65	2:OBITO	2:OBITO		0.833
66	2:OBITO	2:OBITO		0.778
67	2:OBITO	2:OBITO		0.992
68	2:OBITO	2:OBITO		0.982
69	2:OBITO	2:OBITO		0.927
70	2:OBITO	2:OBITO		0.888
71	2:OBITO	2:OBITO		0.797
72	2:OBITO	2:OBITO		0.811
73	2:OBITO	2:OBITO		0.973
212	1:CURA	2:OBITO	+	0.568
213	1:CURA	2:OBITO	+	0.921
214	1:CURA	2:OBITO	+	0.905
215	1:CURA	2:OBITO	+	0.879
216	1:CURA	2:OBITO	+	0.743
217	1:CURA	2:OBITO	+	0.569
218	1:CURA	2:OBITO	+	0.675
219	1:CURA	2:OBITO	+	0.64
220	1:CURA	1:CURA		0.725
221	1:CURA	1:CURA		0.888
222	1:CURA	2:OBITO	+	0.649
223	1:CURA	2:OBITO	+	0.877
224	1:CURA	1:CURA		0.888
225	1:CURA	2:OBITO	+	0.745
226	1:CURA	2:OBITO	+	0.939
227	1:CURA	2:OBITO	+	0.615
228	1:CURA	2:OBITO	+	0.825
229	1:CURA	2:OBITO	+	0.908
230	1:CURA	2:OBITO	+	0.765
231	1:CURA	2:OBITO	+	0.841
232	1:CURA	2:OBITO	+	0.614
233	1:CURA	1:CURA		0.933
234	1:CURA	2:OBITO	+	0.957
235	1:CURA	1:CURA		0.632
236	1:CURA	2:OBITO	+	0.647

Figure 8: Comparison of the Generated Model with the Test

Source: The Author

The hit rate refers to the number of instances correctly classified divided by the total number of instances, which is necessary when it comes to predictive models, the higher the hit rate, the greater the efficiency of the database in the algorithm. A value considered ideal for a hit rate should be between 70% to 100% (BORGES, 2006). Therefore, it is necessary to analyze the efficiency of each technique in the scenario in which it will be applied, to verify if there is precision and reliability in the results with confidence (NASCIMENTO, 2017).

4. FINAL CONSIDERATIONS

This research had as its starting point the importance of the concepts of KDD, which were approached and applied in this work. This model requires attention to each of the phases that integrate it, because the success of a phase depends on the good development of the previous steps. Thus, the KDD was useful in replacing the traditional forms of analysis.

The stage that involved the registration of cases in the official system of the Ministry of Health, was necessary to develop planning strategies, because by analyzing the information on contamination records that occurred in Manaus, indicators can be generated for new strategies based on the health conditions of the people diagnosed with comorbidities.

In this sense, the efficiency of the applied model reached 81%, since the results presented with the extraction of the consistent standards generated useful and valuable information for coping with COVID. This is because, based on the experience presented here, it was possible to produce studies and reports that are really useful for treating patients more effectively.

However, after the application of the algorithm in this database, it was concluded that the Naive Bayes algorithm was efficient in the final generation of the model.

Therefore, this research presents itself as a contribution to future or more complex studies on data mining in the SIVEP-Gripe database..

REFERÊNCIAS

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1996.

CAMILO, C.; SILVA, J. Mineração de Dados: Conceitos, tarefas, métodos e ferramentas. Universidade Federal de Goiás (UFC), p. 29, 2009. ISSN 16113349.

OLIVEIRA, R. R; CARVALHO, C. L. Algoritmos de agrupamento e suas aplicações. Technical report, Universidade Federal de Goiás, 2008.

FREITAS, A. A; LAVINGTON, S. H. Mining Very Large Databases with Parallel Processing. The Kluwer international series on advances in database systems. Kluwer Academic Publishers, Boston, 1998.

LUKE WELLING, LAURA THOMSON. Tutorial Mysql. Ciência Moderna, 2004

Lambert M. Surhone, Miriam T. Timpledon, Susan F. Marseken. MySQL Workbench. Betascript Publishing, 2010

H. Witten, E. Frank, M. A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edition, Morgan Kaufmann, 2011.

R. Santos. “Weka na Munheca: um Guia para Uso do Weka em Scripts e Integração com Aplicações Java”. Instituto Nacional de Pesquisas Espaciais (INPE), 2005.

AGRAWAL, R. & SRIKANT, R. Fast algorithms for mining association rules. Proc. of the 20th Int'l Conference on Very Large Databases. Santiago, Chile, set. 1994. Disponível na Internet. http://www.almaden.ibm.com/u/ragrawal/pubs.html. 3 junho 1999.

Zanardi, Lucas Adão. Data mining: estudo e aplicação de algoritmos de data mining; orientador: Prof. Dr. Evandro de Araújo Jardini. Fernandópolis, 2007. 100 f.

MINISTÉRIO DA SAÚDE. Coordenação-Geral do Programa Nacional de Imunizações. Disponível em: https://opendatasus.saude.gov.br/dataset/bd-srag-2021 Acesso em: 14 fev. 2021. Autor: Datasus – Brasília: Ministério da Saúde, 2021.

Galvão,Noemi Dreyer ; Marin, Heimar de Fátima. ARTIGO DE REVISÃO.Técnica de mineração de dados: uma revisão da literatura. Acta Paulista de Enfermagem. versão impressa ISSN 0103-2100. Acta paul. enferm. vol.22 no.5 São Paulo set./out. 2009.

MENDES, Luciana. TCC Data Mining – Estudo de Técnicas e Aplicações na Área Bancária.São Paulo-2011.

NANGIYALIL, Sajeev G. Estudo de Ferramenta de KDD & Mineração de Dados. In: Trabalho Monográfico. São Paulo. UNICID-SP, 2007.

Tan P, Steinbach M, Kumar V (2009) Introdução ao Data Mining - Mineração de Dados. Editora Ciência Moderna Ltda

Ellram, L (1996) The use of the case study method in logistics research. Journal of Business Logistics. Oakbrook, Ill, v. 17, n. 2.

BORGES, H. B. Redução de dimensionalidade em bases de dados de expressão gênica. 2006. 123 f. Dissertação (Mestrado em Informática) - Pontifícia Universidade Católica do Paraná. Curitiba, 2006.

NASCIMENTO, F. R. S. Um Estudo Comparativo entre Algoritmos de Proteção da Privacidade e Segurança de Dados Aplicado à Bases de Dados na Área de Saúde. 2017. 52 f. Trabalho de Conclusão de Curso (Graduação em Sistema de Informação) - Universidade Federal do Rio Grande do Norte. Caicó – RN. 2017