Data Mining for Healthcare Decision Making

Matheus Carvalho Peixoto
Pontifícia Universidade Católica de Goiás, PUC Goiás
Goiânia, Brasil

Kátia Kelvis Cassiano Lozano
Universidade Federal de Goiás, UFG, Goiás
Goiânia, Brasil

Talles Marcelo Gonçalves de Andrade Barbosa
Pontifícia Universidade Católica de Goiás, PUC Goiás
Goiânia, Brasi

ABSTRACT
This paper presents a case study with the application of machine learning techniques to generate knowledge about breast cancer in Goiás / Brazil regarding morbidity and access to health services. Data mining techniques were used, involving descriptive and predictive data analysis, revealing characteristic patterns of the Goiás municipalities that allowed their grouping by similarity. The results suggest that such models can be used to generate information that supports decision making processes in the definition and applicability of public health policies.

General Terms
Data Mining, Descriptive Exploratory Analysis, Predictive Analysis, Clustering

Keywords
Data Mining, Descriptive Exploratory Analysis, Predictive Analysis, Clustering

1. INTRODUCTION
Public policies refer to a set of programs, actions, and activities implemented by the State - at the federal, state and municipal levels - about a specific sector (health, education, for example) to guarantee the right of citizenship. These actions directly affect citizens, so it is important that they are assertive and not allocated to public resources.

Given the exponential growth of data and the significant evolution of computer systems, producing information to add value to decision making has become a major challenge. According to a survey conducted by the International Data Corporation (IDC), in the year 2025, it is estimated a generation of 160 zettabytes from heterogeneous sources: corporate data, health system records, socioeconomic data, social media content [1].

According to [2], data mining is described as a process that automates the extraction of characteristic patterns to generate knowledge from the interpretation of existing relationships and can be applied in various contexts, such as the development of decision support solutions.

In this context, health data analysis is an interdisciplinary area that combines data with a set of computer science and mathematics techniques and strategies such as artificial intelligence, machine learning, graphical interfaces, and statistical models to reveal patterns and relationships between data [3] [4].

One of the major challenges related to health data analysis concerns the selection of the database. In Brazil, the Department of Informatics of the Unified Health System (DATASUS) provides the organs of the Unified Health System (SUS) with information systems and computer support. It is also responsible for sustaining and preserving data collection SUS [5]. These data can be used in data analysis processes to foster the definition and implementation of public health policies in an assertive manner.

This paper presents the results of a case study developed through the application of descriptive and predictive data analysis techniques, to generate knowledge about breast cancer in the state of Goiás / Brazil and, thus, providing subsidies for evaluation and management of public health policies.

2. METHODOLOGY
The methodology applied in the present study is presented in Figure 1.

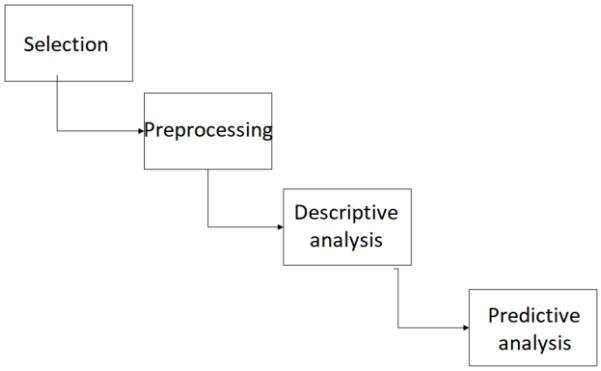

Fig 1. Data Analysis Process Diagram.

The data used in the present study were directly selected from the DATASUS open data repository, available at [6]. In this repository, data selection is performed in the context of filters and originally made available as reports, one for each year and age group.

A script in Python language was developed with the purpose of grouping the records of the municipalities of Goiás, inserting the year and age attributes of the period from 2010 to 2013. As a result of this data selection step, a file in Comma Separated Values (CSV) format was generated.

In the preprocessing step, the data was treated to eliminate redundancies and inconsistencies, such as missing data in some attributes or records. Through resource engineering, attributes were derived from concatenating characteristic attributes.

Using Power BI Desktop software, a descriptive exploratory analysis was performed using interactive dashboards that relate the characteristic attributes related to morbidity and mammography. Such analysis is important for the discovery of patterns, allowing an understanding of the problem as a function of the variables of interest.

From the application of machine learning techniques, a predictive analysis was performed. An unsupervised learning

model for grouping data by pattern similarity has been developed in Python programming language. The clustering technique used was K-Means and, for its implementation was used in the *scikit-learn* library. The *onehotencoder* and *standardscaler* functions from the *scikt-learn* library were used to adapt the data to a machine-readable format.

3. DATA ANALYSIS

3.1 Descriptive Analysis

Descriptive analysis aims to summarize and explore the behavior of the data. It allows the combination of characteristic attributes for information generation. It provides inputs for pattern generation.

Descriptive analysis was performed using Power BI Desktop, commonly used for data intelligence applications. It Allows interactivity between graphics and pattern identification.

3.2 Predictive Analysis

Clustering techniques aim to segment the data set according to the similarity of the characteristic patterns. The K-Means algorithm is presented in Figure 2.

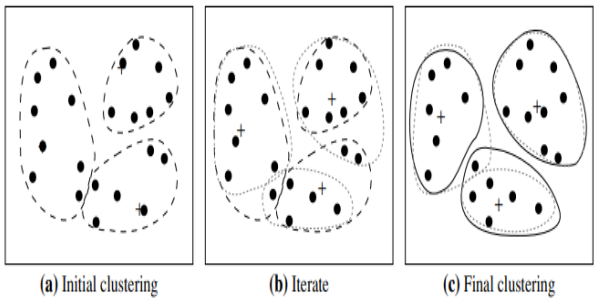

Fig 2. Cluster analysis [7].

The K-means algorithm works as follows. Initially, the cluster centers *k* are randomly selected. Each attribute is assigned to the cluster to which it is most similar, based on the Euclidean distance between the object and the cluster average. The algorithm iteratively improves variation within the cluster. For each cluster, it calculates the new average using the objects assigned to the cluster in the previous iteration. All objects are reassigned using the updated media as the new cluster centers. Iterations continue until the assignment is stable. In other words, until the clusters formed in the current iteration are the same as those formed in the previous iteration [8].

A major challenge in the K-means algorithm is choosing the optimal value of *k*, as it can affect model performance. There are several methods for determining the optimal value of *k*, depending on the type of learning (supervised or unsupervised) and the intrinsic parameters of the data. In this work, was opted for the Silhouette Score Method.

The silhouette coefficient combines the idea of cohesion and group separation. Cohesion measures how observation is similar to the assigned cluster and, mathematically, results from the sum of the squares of the observation distances from the cluster's centroid. Separation measures how different is an observation from close groups is; in other words, measures how well separated one cluster is from the others.

Figures 3 and 4 show the Silhouette Score Method on mammography and morbidity databases, respectively. The optimal value of k for the mammography base, as shown in Figure 3, was 14. The optimal value of k for the test of morbidity base, according to Figure 4, is 2.

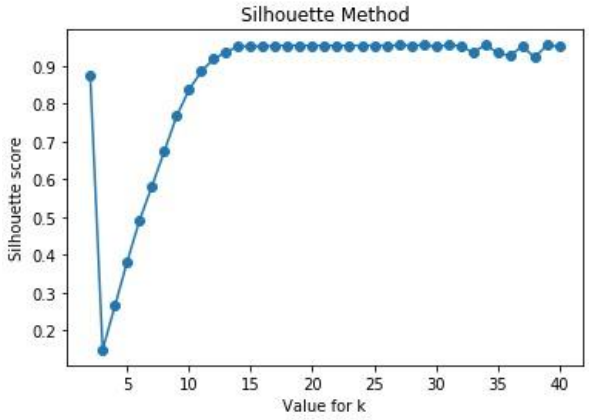

Fig 3. Silhouette coefficient for mammography base.

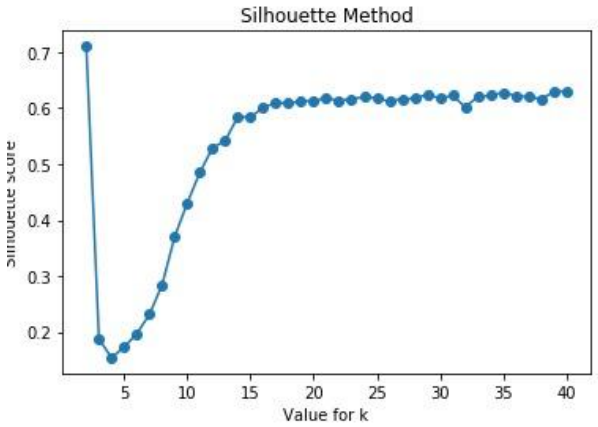

Fig 4. Silhouette coefficient for morbidity base.

4. RESULTS AND DISCUSSION

4.1 Results for Descriptive Analysis

The initial objective was to explore aspects of the incidence of breast cancer in Goiás, but the unavailability of data did not allow this approach to be developed.

The results of the descriptive analysis are presented as interactive dashboards. The visualizations obtained from the tool illustrate the behavior of the data by municipality and age group.

The visualization shown in Figure 5 shows the ten municipalities of the state of Goiás with the highest average morbidity. The red line represents the average of the average morbidity rate.

Figure 6 illustrates the municipalities with the highest average of mammography exams performed in which the red line represents the average of the examination. The Yellow represents the median of the average of the examination.

The ten municipalities with the highest average morbidity rate are in the interior of the state of Goiás, while the highest average of examinations performed is in Goiânia. The outlier value presented in Goiânia suggests an aspect of patient migration to access to public health.

Figure 7 shows the annual evolution of the average morbidity rate by age group for the ten municipalities with the highest

average morbidity rate in which the blue line represents the range of people over the age of 50. The red line represents the range of people under 50 years old. The average morbidity for the age group older than 50 years was higher when compared to the age group younger than 50 years from 2010 to 2013, not distinguishing it from the context of the other municipalities.

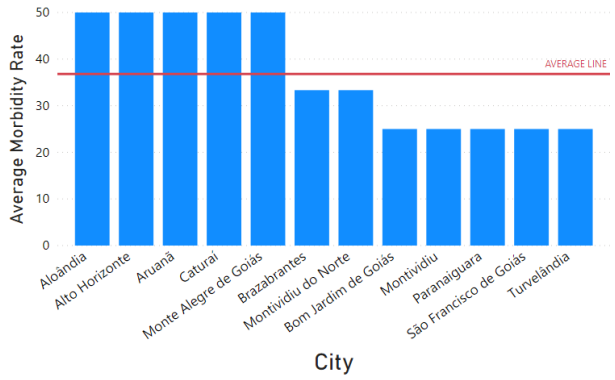

Fig 5. Morbidity rate by the municipality.

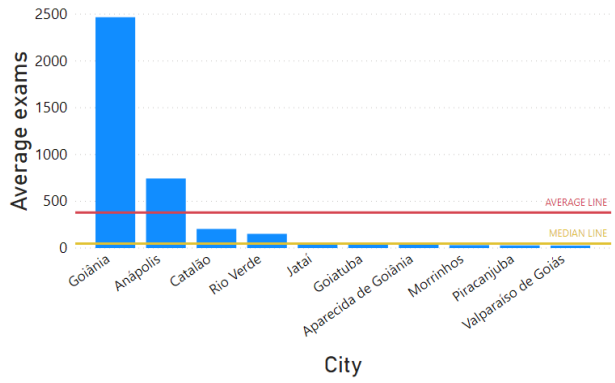

Fig 6. Average mammography by the municipality.

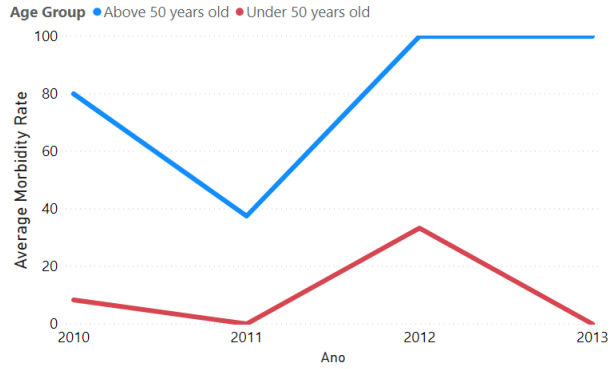

Fig 7. Morbidity rate by age group.

The average length of stay by age group for the municipalities with the highest average morbidity rate is shown in Figure 8. The average length of stay for both ranges is higher in these municipalities when compared to the general context.

In the general context, the average number of exams performed by the age group below 50 years is higher when compared to the age group above 50 years. Notwithstanding this, the average amount of examinations performed by these bands in the municipalities with the highest average morbidity rate is closer, as shown in Figure 9.

Overall, the average hospitalization costs for the age groups were close, with higher average costs for the age group over 50 years. For municipalities with the highest average morbidity rate, hospitalization costs are higher for the age group below 50 years in 2010 and 2013, as shown in Figure 10.

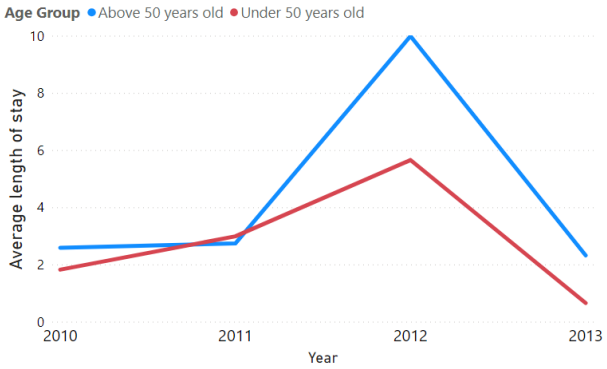

Fig 8. The average length of stay by age group.

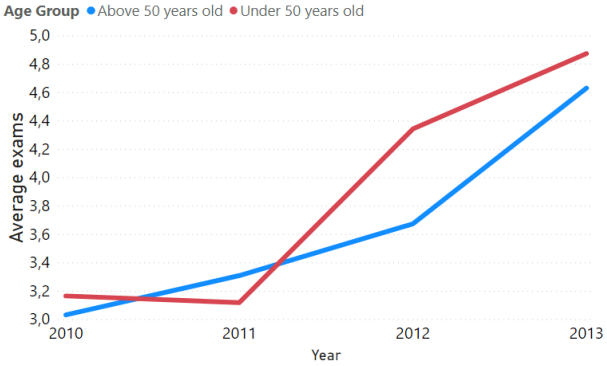

Fig 9. Average mammography by age group.

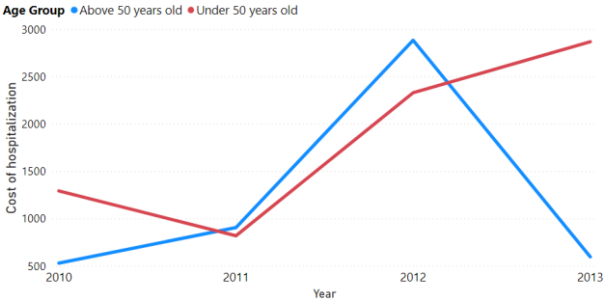

Fig 10. The average total cost of hospitalization.

4.2 Results for Predictive Analysis

Two models of group analysis are presented. One for morbidity base and one for mammography exam base.

Group analysis for mammography data referenced 14 groups. Each group was distributed according to the standards identified by the k-means algorithm in relation to the number of exams performed and diagnoses. The data are sparse, meaning there is a lot of variability in patterns by age group. Figure 11 illustrates the number of records assigned to each group.

For the morbidity basis, two groups were found. The descriptive analysis emphasized that in general, the characteristic pattern of morbidity of Goiânia differs well from the other municipalities concerning the evaluated

attributes. Few similarities were identified in 2010 for some age groups. Figure 12 shows the two groups identified.

The results of the analysis of the groups for morbidity corroborate the standards revealed in the descriptive analysis regarding the dissimilarity of Goiânia to the other municipalities.

The dissimilarity of Goiânia to the other municipalities suggests that there is a concentration of treatment in private networks. In the interior of the state, the public health system is more used, because in many municipalities is the only one available. The capital also receives people from all over the state for the treatment, which can corroborate this dissimilarity.

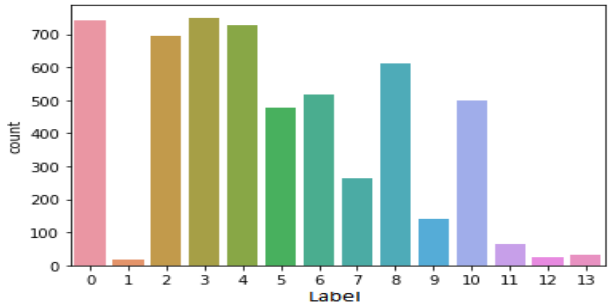

Fig 11. Mammography Base Groups.

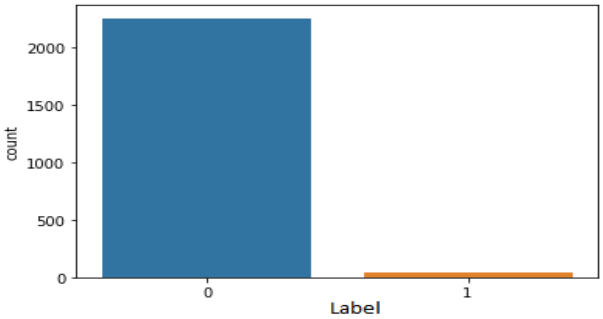

Fig 12. Morbidity Base Groups.

5. FINAL CONSIDERATIONS

The solution presented in this study is the potential for developing decision support systems and defining strategies for the implementation of public policies, as it generates information that allows greater assertiveness in the definition of these policies and greater use of public resources. It is a problem of social order.

The next step is to compare performance between data grouping models. Also, there is the possibility to label the database with the cluster results and use it for classification.

6. REFERENCES

[1] Reinsel, D., Gantz, J., Rydning, J. Data Age 2025: the evolution of data to life-critical. Framingham, MA, USA: IDC White Paper, 2017.

[2] Castro, L. N., Ferrari, D. G. Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

[3] J. Han, M. Kamber and J. Pei, “Introduction” in Data Mining Concepts and Techniques. Waltham: Morgan Kaufmann Publishers, 2012.

[4] Data analytics in medical data : A review

[5] DATASUS: Histórico / Apresentação. Disponível em: <http://datasus.saude.gov.br/datasus>. Acesso em: 10 de outubro de 2019

[6] SISMAMA: Informações estatísticas. Disponível em: <http://w3.datasus.gov.br/siscam/index.php?area=0402>. Acesso em 01 de agosto de 2019.

[7] J. Han, M. Kamber and J. Pei, “Cluster Analysis: Basic Concepts and Methods” in Data Mining Concepts and Techniques. Waltham: Morgan Kaufmann Publishers, 2012, pp. 453

[8] J. Han, M. Kamber and J. Pei, “Cluster Analysis: Basic Concepts and Methods” in Data Mining Concepts and Techniques. Waltham: Morgan Kaufmann Publishers, 2012, pp. 451