Unravelling Patterns from a Blood Donation Data Set using Machine Learning Approaches

Francisco J. Lopes

ABSTRACT

Blood management is a concerning problem for humans. Albeit the existence of technological progress in the field of substitutes for blood products, there will always be a need for whole blood from donors and the derived products that come from it, so an optimization of the available resources should be made. With that in mind, this thesis has the goal of discovering valuable information to optimize operations of IPST, making better use of the existing teams, resources, and daily inventory. For this reason, we start by presenting a background of the most relevant concepts, followed by a review of the state of the art in this domain. Afterwards, we perform an exploratory data analysis of the blood donation data set provided by IPST to have a better understanding of it. This analysis focuses on studying the time-geographic aspect of the donations, as well as the donors that performed them. We then discuss the machine learning approaches that we use to study and find patterns in the data set, and how they were applied. The results obtained highlight the characteristics of the most relevant donations made in 2017, 2018, and 2019, while providing interestingness metrics that are able to support those choices. We also present the results of a more geographically focused study that we performed, in which we discovered the characteristics of the most relevant patterns for each of the Portuguese districts in the same time frame. We end by discussing and presenting the main conclusions of our study.

KEYWORDS

Blood Donation; Data Mining; Network Science; Geographic Patterns; Sequential Patterns

1 INTRODUCTION

The Instituto Português do Sangue e da Transplantação (IPST)[1] is a public Portuguese institute of the Ministry of Health and the recognized authority for the collection and regulation of blood donation and transplantation at the national level. Its headquarters are located in Lisboa and its operations are divided into three regional centers (south, center, and north). Each regional center has the same characteristics, having its own finite resources (e.g. personnel).

IPST is dedicated to supporting human life through areas of intervention transversal to the whole medical and surgical activity, by ensuring the sustainability of health care, assuring the supply of blood and its components, as well as the cells, tissues, and organs for transplantation. It is divided into three main Centros de Sangue e da Transplantação (CSTs) (Blood and Transplant Centers), in Lisboa, Porto, and Coimbra. Its mission is to ensure and regulate transfusional medicine and transplantation activity on a national level, as well as to ensure the offer, collection, analysis, processing, preservation, storage, and distribution of human blood, blood components, organs, tissues, and cells of human nature. IPST's vision is translated to promoting the offering as a gesture that is transversal to the whole IPST's activity with the goal of contributing to human life on time and quality and to do so ensuring that good practices and innovation accompany the state of the art [18].

Blood donations can be made at hospitals, schools, companies, fire stations, etc. Upon collection, each blood unit is tagged with a unique number and associated bar code, according to the International Society of Blood Transfusion (ISBT) 128 norm, which is used to track information, from collection to transfusion (e.g. blood type, collection time, and location). IPST is able to keep track of this information at the national level with a dedicated software system called Aplicação de Sistema de Informação de Sangue (ASIS), which has operated since 2002 and contains historical data for blood activity in Portugal. Over the years, hospitals have been integrating their information systems with ASIS, to enable their communication of collections, storage, distribution, and transfusions. Currently, only 10-20% of national blood collections are not integrated with ASIS, since 5 hospitals have not yet implemented the proper automatic reporting procedures from their own information system into ASIS. Nevertheless, being the recognized authority, IPST still has access to periodic reports from these hospitals and is thus able to assess national collections and needs [11]. Accordingly, every year IPST creates: 1) An *activity plan*, which defines the adopted strategies, ranks options, and plans actions and mobilization of resources for that year. In the medium-term, this plan is made based on annual targets set at the national level, driven by the contracts with donors' associations and planned events (e.g. festivals, World Youth Day, etc.), and in the short-term, the collection dates and locations are fixed, and the logistics are determined; and 2) An *activity report*, that refers what decisions were taken, points out the deviations that were made from the initial plan, evaluates the results, and structures the information that might be relevant for the future.

In 2014, an emergency plan was presented, with the goal of minimizing the impact of an accident or catastrophe in the fulfillment of IPST's mission regarding blood, which influences the amount of blood components to be stored and readily available [17].

1.1 Problem

Blood and its respective components are needed for different kinds of situations, such as emergency or regular surgeries and other procedures, like the treatment of anemic patients, premature infants, cancer, liver diseases or burn injuries. However, even though there is an activity plan done by IPST, it is impossible to predict exactly the characteristics of blood donations that are collected or how many there will be. This may lead to a lack of blood units with specific requirements, due to which it is not possible to keep the health care system running as planned, or having too many blood units, which end up going to waste since they were not used.

According to the IPST's activity report from 2019 [19], the year was marked by the alignment to some measures and orientations that concern the efficient use of resources, reduction of waste, and

[1] http://ipst.pt/index.php/pt/

inefficiencies. In this context, and to continue improving the efficiency of blood supply operations in Portugal, a project called LAIfeBlood was created with the goal of providing IPST with new tools that are capable of helping with this improvement.

1.2 Goal

As part of the LAIfeBlood project, the goal of this thesis is to first focus on analyzing the historical data made available by IPST (a blood donation data set) and then use them to find patterns. These patterns would then be studied to retrieve information and understand how they can be helpful to IPST. From them, we intend to provide new hints to IPST on how to fulfill its activity plans while increasing the resources and daily inventory for that year by optimizing the operations of IPST teams. Examples of useful hints could be: 1) If two locations are close to each other, try to collect blood in both of them on the same day, or be somewhere in between them and ask people to go donate there instead; 2) If there are brigades with high attendance in a specific location, try to increase the frequency of blood collections performed by them there over the year, and decrease the amount of time spent on each of them; 3) If there are too many donations in a certain place with low attendance, reduce the number of collections there, while increasing the number of collections in a place where there is a low number of donations, but with potential for a high donation rate; 4) If the most relevant brigades for each district are identified, more resources can be provided to them, while reducing the resources given to the less relevant ones.

2 BACKGROUND

2.1 Network Science

Network science is a field that aims to study and analyze complex networks [25]. A network is made up of various elements represented by nodes N (or vertices) and the connections made between those elements, known as edges E (or links). A network can be undirected, if all the edges are bidirectional, or directed if its edges have a direction from one node to another. Some characteristics of a network are: 1) the Degree k; 2) the Degree Distribution P_k; 3) the Average Degree z; 4) the Average Path Length $\langle L \rangle$; and 5) the Clustering Coefficient $\langle C \rangle$. There are different types of networks depending on the application, and this thesis focuses on similarity networks, since the goal is to have the nodes (donors) connected according to their similarities.

2.1.1 Similarity Network. A similarity network is a type of network in which the edges represent the similarity between two nodes, i.e., if two nodes are similar to each other, then they are connected; if not, then they are not. Similarity can be measured using similarity metrics. However, similarity between two nodes is subjective since what is considered similar depends from author to author, depending on their own opinions and expertise.

To better address this issue of what can be considered similar, some authors have proposed similarity functions depending on the type of data that is being used. For heterogeneous data sets, the similarity functions need to be different from the ones used for homogeneous ones, like the Euclidean Distance or the Cosine Similarity. Some similarity functions that can be used for heterogeneous data sets are: HEOM, GOW, ER, GEM, and HVDM [22].

We focused on the Heterogeneous Euclidean-Overlap Map (HEOM) [22] since it is able to deal with the type of data considered in this thesis and is simple to implement and to make changes to. It consists of a Euclidean distance that treats a variable i differently depending on whether it is nominal or quantitative.

2.1.2 Community Finding. In network science, a *Community* is a group of nodes that have a high likelihood of being connected to each other rather than to the nodes of other communities. So, community finding algorithms focus on grouping communities according to different approaches, such as modularity optimization, which is the one that will be focused on in this thesis, due to its relevance for finding coherent communities.

One of the most relevant metrics in community finding is the *Modularity M*, which measures the strength of the division of a network into modules/communities. The higher the modularity value, the more similar the nodes in each community are. Some examples of *Modularity Maximization* algorithms are the `CNM`, `PL`, `WT` and the `Louvain Method` [3]. The latter is a heuristic method for community finding proposed by Blondel et al. [6]. The `Louvain Method` has lower computational complexity and shorter running time than any other of the studied algorithms, while providing good results in terms of accuracy, and higher modularity when compared to other modularity maximization algorithms [6]. For this reason, this was the algorithm we chose to use in this thesis.

2.2 Data Mining

Data mining can be divided into two broad areas: descriptive data mining, using mainly unsupervised learning methods, and predictive data mining using supervised approaches. The area of focus in our thesis was *Unsupervised learning*, which has the goal of discovering information (in the form of groups, patterns, etc.) from input data that is not labeled and consists of *Clustering* (including *Biclustering*) and *Association* algorithms. We focused on Association approach in this thesis, more specifically pattern mining.

2.2.1 Pattern Mining. Frequent Pattern Mining helps to discover *frequent patterns* that conceptually represent relations among discrete entities (or items). Depending on the complexity of these relations, different types of patterns are discovered. [9] There are several types of frequent patterns, for example, frequent itemsets, frequent subsequences (or sequential patterns), and frequent substructures. The most common kind of frequent patterns are *itemsets*, where the relation is the co-occurrence of items that appear together in a transactional data set. There are also sequential patterns, which require a temporal or geographic ordering between items. Pattern mining algorithms are divided into four main categories, these being: 1) Transactional Mining, or Itemset Mining; 2) Sequential Mining; 3) Tree Mining; and 4) Graph Mining. The ones that will be focused on are the *Transactional* and *Sequential* types.

There are also the concepts of *maximal frequent itemsets* and *closed frequent itemsets*. According to [14], an itemset X is maximal if it is not a sub-itemset of any other frequent itemset and closed if it does not have a super-itemset with the same support. X is closed in a data set D if there exists no proper super-itemset Y such that Y has the same support count as X in D. If X is both closed and frequent in D, it is a closed frequent itemset in D. X is a maximal frequent itemset in a data set D if X is frequent, and there exists no super-itemset Y such that $X \subset Y$ and Y is frequent in D. These itemsets can have a single or multiple items.

Transactional Mining. Focuses on finding frequently co-occurring itemsets in transactions. For this type of algorithms, the sequence of transactions and the structure of data are not taken into account (unstructured). It is the first step of Association Rule Mining, which then creates rules based on the frequent patterns found, according to a minimum relative support and confidence thresholds. These two are considered interestingness measures and can be applied for the evaluation of association rules. Other examples of interestingness measures are the lift, conviction, and leverage. The most well-known and the basis of many other algorithms are `Apriori` and `FP-Growth`. These two, together with `FP-Max`, were the ones used/tested in this thesis.

Sequential Mining. Has the goal of discovering sequential patterns given a data set of sequences, where each sequence is an ordered list of transactions and each transaction is a set of items. A sequential pattern is also composed of a list of sets of items. The sequential pattern mining algorithms considered in this thesis were `PrefixSpan` and `CM-SPADE`, which take into account item occurrence order but not transaction time intervals, meaning that two sequences containing the same itemset at separate times are deemed to have the same pattern. We considered the `Fournier08` algorithm, which takes transaction timestamps into account, allowing for the retrieval of closed sequential patterns with defined time intervals between transactions. `CM-SPADE` is an improved version of the `Spade` algorithm, while `Fournier08` took inspiration from `Hirate-Yamana` and `BIDE+`.

2.3 Blood Management

According to Beliën and Forcé [4], there are eight classification fields that a person should take into account when referring to this topic: 1) Type of blood products; 2) Solution Method; 3) Hierarchical level; 4) Type of problem; 5) Type of approach; 6) Type of algorithm; 7) Performance measures; and 8) Type of study. Regarding the collection and distribution of blood products, Blake and Hardy [5] explain how it works in Canada, for red cell units, where they discuss how a generic modeling framework is useful for regional blood supply chains.

3 LITERATURE REVIEW

3.1 Network Science

With the emergence of the patient similarity network paradigm, the concept of similarity networks has been brought to the spotlight again [26]. With the goal of adapting this concept and applying it in this dissertation, we did a review of papers regarding similarity functions and networks, as well as community finding.

3.1.1 *Similarity Network.*

To study the creation of a function that can calculate the similarity between two nodes, Lumijarvi et al. [22] refer and compare five different Heterogeneous Proximity Functions: 1) HEOM; 2) GOW; 3) ER; 4) GEM; and 5) HVDM. Harikumar [15], proposed another distance metric in the form of a triplet. Before applying the similarity function, the numeric attributes are normalized. The distance is calculated depending on the type of the attributes. Finally, the distance between two objects is the sum of all the previous distances calculated. Klenk et al. [20] proposed the use of a similarity function with a weighting term to calculate the similarity for patient data. This term was assigned to each dimension and corresponds to its influence on the similarity.

3.1.2 *Community Finding.*

In the paper presented by A. Carreiro et al. [8], the similarity network was used as an input to Gephi and a community finding algorithm from that software was used to find and analyze the communities. Finally, information regarding the communities was retrieved, such as the features that characterize each one and their importance.

3.2 Data Mining

The literature regarding data mining on issues related to blood donation is considerably small, but it has been growing in the past few years. We follow an overview of the current state of the art that is relevant to this project.

In the topic of blood donations or donors, clustering algorithms such `K-Means` and `Two-Step` were used and discussed by Ashoori and Taheri in [2] and by Venkateswarlu and Raju in [30]. `Two-Step` was also applied by Testik et al. in [29]. In the same paper, they also used `CART`, which is a decision trees algorithm. This algorithm was also used by Ashoori et al. in [1], and compared to other algorithms of the same type, such as `CART` or `C4.5`, such as `C5.0`, `CHAID`, and `QUEST`. `CART` was also applied by Santhanam and Sundaram in [28]. `C4.5`, another decision trees algorithm was used by Ramachandran et al. in [27], similarly to Boonyanusith and Jittamai in [7], who then compared the results with the ones obtained from using a multilayer perceptron (MLP), which is an Artificial Neural Network (ANN) algorithm. An MLP was also used by Mostafa in [24], who compared the respective results with the ones obtained with a Probabilistic Neural Network. ANN and Support Vector Machines approaches were used and compared by Darwiche et al. in [10].

3.2.1 *Pattern Mining.*

Even though, to the best of our knowledge, only a single paper applies a pattern mining algorithm ([21]), this type of algorithms seemed to be an interesting approach to use on the data set that was used in this thesis. For this reason, the rest of the literature regarding pattern mining focuses on topics other than blood donation, and that served as a basis what type of approaches was used, such as papers where the algorithms were proposed, can be represented or were used in another area of expertise. To better understand this topic, we started by considering the book by Han et al. [14] and the paper by Chaoji et al. which presented Transactional Pattern Mining algorithms such as `Apriori`, and `FP-Growth` [9]. The representation of the results obtained with Association Rule Mining algorithms is often difficult, since it is mostly based on the values of the support, confidence and lift of each rule and there is a large number of found rules, making it difficult to interpret those results. With this in mind, Michael Hahsler and Radoslaw Karpienko [13] present arulesViz, that provides the most popular visualization techniques for association rules. Regarding Sequential Pattern Mining algorithms, T. Li et al. [21] presents a Fuzzy Sequential Pattern Mining. P. Fournier-Viger et al. conducted a survey regarding recent sequential pattern mining algorithms and their utility in [12]. And A. S. Martins et al. [23] used transactional mining and sequential pattern mining in a data set of amyotrophic lateral sclerosis patients.

4 EXPLORATORY DATA ANALYSIS

4.1 IPST Data Set

The data set made available by IPST is composed of heterogeneous data gathered from 1995 until August 2020, and it consists of 5 787 731 rows and 55 columns (or features). Each row of the data set represents an instance of a blood donation by a donor (from a total of 1 055 831 different donors) that was collected by a specific blood collection brigade. For this reason, blood donation and blood collection have the same meaning in this thesis. Each column has information about the donor, the donation, the association, and the brigade responsible for the blood collection. This information can vary from the blood type of the donor, to the identifier of the brigade. The data can be divided into five core concepts of the data set that each feature can be attached to. A feature can belong to one or more of these core concepts. This allows us to group the features by them and facilitate how the features are defined and their utility. The core concepts are: 1) the donor; 2) the donors' association; 3) the blood collection brigade; 4) the blood collection session; and 5) the blood collection, or blood donation. Each of these concepts is identified and characterized by a set of features.

4.2 Data Distribution

To do a general statistical analysis of the data set, we used the Pandas library[2] to study and analyze our data set, and the Matplotlib library[3] to plot the graphs that will be presented below. Firstly, we start by making a *time-geographic* focused study. In the stacked bar chart from Figure 1, we can see the total amount of donations for each year together with the distribution of the contributions to them by the three different CST (SP, SL, SC as defined in the identification of a donors' association) present in the feature that identifies the donors' associations. The null bar represents the donations with a missing value in this feature, so they did not have a CST assigned to them in the data set. This division was made since the data set can be organized or divided by CST, which is a logical division of the data to better explore it, since they are independent of each other. In this plot, we can see that there was an increase in the number of donations registered by IPST up until 2009, when it began to decrease again until 2020 (please bear in mind that the data for 2020 only goes up until August of that year). This figure also highlights the increase in donations for the Porto CST in comparison to the other two, even though the number of donations country-wide has been decreasing over the last 10 years. Another important aspect to note in this chart is the fact that from the years 2017 until 2019, there has not been much variance in the number of donations.

Studying the distribution of the donations collected by brigades originated in the Portuguese districts from 1995 to 2019 (we did not count 2020 since the data does not represent the complete year), we can see that overall each of these districts has had a decrease in the number of donations. The districts composing the Porto CST (Porto and Braga) have had a lower decreasing rate as it was expected in comparison to districts that are part of other CSTs. The districts with the lowest number of donations are the ones that compose the southern part of Portugal (Beja, Évora, Faro and Portalegre), and there have not been any donations from Beja and Portalegre since 2014 and 2015, respectively.

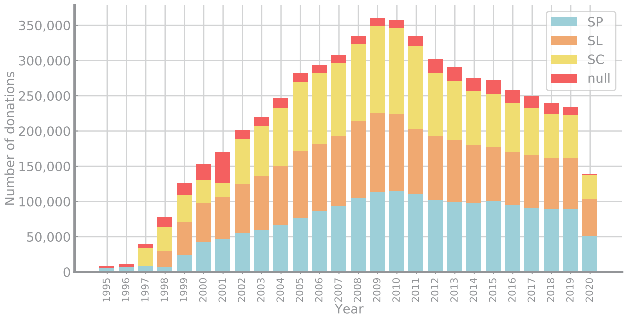

Figure 1: Distribution of the number of donations from 1995 to 2020, with the contribution of each CST to the total number of donations for each of those years. Each color represents a different CST. SP is the Porto CST, SL is Lisboa's, and SC is Coimbra's.

The distribution of the donations over the months for each year shows that the month with the largest tendency for a higher number of donations are March, May, October and November. We can also see that the months that correspond to the Summer, like June, July, and August, together with the months that correspond to the Winter, mainly January and February, are the ones with the lowest donation rate. The years of 1995, 1996, 1997 and 2020 are outliers.

Moving on to a *donor* focused study, where we look at the distributions of all the donations according to the different characteristics of the donors that performed them. We start by comparing the district of origin of the brigades and of birth of the donors. Porto is the district where most of the donors come from, however this result is not exactly expected considering that the brigades originated in Lisboa are the ones that receive a higher amount of donations. This presents an interesting result that the citizens born in the Porto district are the ones with the highest donation ratio, however they do not all donate blood to brigades original from the Porto district, showing a large gap between those values. The most frequent blood types over the entirety of the donations are the expected ones, A+ and O+. There is also small amount of AB- and B-, which are the rarest types. The distribution of the donations, according to the gender of the donors, shows that overall there are more males donating blood than females. However, in the 2017-2019 time frame, there have been slightly more females than males donating. Regarding the distribution of the total number of donations, taking into account the current age of the donors. The minimum age requirement to donate blood is 18 years old, and the maximum is 65 years old. This way, we only take into consideration the ages from 18 to 90 years old, since donors that were 65 in 1995 would now be 91 years old. From it, we observe that the age bracket with the highest number of donations is the one where the donors are at the ages between 40 and 50 years old. There has been an increase in the number of donations by younger donors, especially in the 20 to 25 year bracket. Ultimately, there are 3 001 different values registered

[2] https://pandas.pydata.org/
[3] https://matplotlib.org/

in the donor's job feature. We can see that overall the most common value for this feature is the one defined as "unknown", however for the 2017-2019 time frame the most common value changes to "factory employee".

5 UNRAVELLING PATTERNS

5.1 Initial Data Pre-Processing

We started by transforming some of the features present in the data set and removing the wrong values present in them, as follows:

- The feature with the donor's birthdate was transformed into the donor's age to represent the age of the donor as an integer, this way being easier to compare than a date. Then we kept the rows where the age was above or equal to 18 years old and below or equal to 91, as aforementioned;
- The feature with the blood collection date was divided into four different features with the year, month, week, and day of the donation, respectively, again because comparing integers or strings is easier than comparing a date;
- Correcting the features with the donor's blood type ground and RhD, as suggested by IPST.

None of the approaches was able to process the whole data set (with 5 369 765 rows and 57 features) with acceptable computational time and resources. To address this issue, the data set was divided into different data sets of donations for 2017, 2018, and 2019 (with 239 414, 229 922, and 223 545 rows, respectively). In this way, the criteria for the sampling done was the year, considering that in these three years there is a similar distribution of the data, and we can see that there was a stabilization of the data, so most of the characteristics and distributions of the features from the previous years are maintained in this time frame. From this point on, whenever we are talking about further changes to the data sets, these three are the ones we are referring to.

5.2 Community Finding

5.2.1 Data Pre-Processing. We began by removing the features where the data was unbalanced, since to make a similarity network these would often have the same value between two different nodes, which could bias the connection between both nodes. We decided to choose only one of/collapse the features that gave the same information as others. This was mostly because different columns were referring to the location in which the donation was made but with different granularities. With this, the similarity network would have a bias towards two nodes that share the same location, since it is represented by more than a single feature in the data set. Then we removed the features that would not add any new information to the network that was going to be created, since they gave only information about the donation that would not be relevant to the similarity between two nodes, or because they had a high percentage of missing values (more than 20% threshold of missing values). Finally, the number of features selected had to be small, since the algorithm scaled with the number of features. So, after performing the data pre-processing, we were able to select the seven most relevant features: 1) a new feature with the donation number; 2) the donor's age; 3) the donor's gender; 4) the donor's blood type group; 5) the donor's blood type RhD; 6) the brigade's district of origin; and 7) the month of the blood collection.

5.2.2 Similarity Network Creation. The similarity network consisted of the relationship between two donations. To create this network, we needed to have a metric that reflected this relationship. The data set was composed of heterogeneous data. So, we needed to use different distance metrics to represent how close the relationship between two nodes was, while having to deal with any missing values. Since each node was represented in the form of tabular data, which has the values of its features: 1) If the data was categorical, we checked if the value present in a column X was the same for both nodes; and 2) If the data was quantitative (the donor's age being the only case), it was compared by intervals of age ranges $-x < Age < x$, to check if the other donor's age could be included in that age range.

Then, to combine the data, a global distance was calculated by using an approach to the HEOM [22] for each type of data. This distance function was able to deal with nominal and quantitative data types in a data set. However, how quantitative values were treated was different from the original function, since we were dealing with age ranges. It was adapted to:

$$h_i(a,b) = \begin{cases} 1 & \text{if } a \text{ or } b \text{ is a missing value} \\ I_D(a,b) & \text{if the } i\text{th variable is nominal} \\ J_D(a,b) & \text{if the } i\text{th variable is quantitative} \end{cases}, \quad (1)$$

where I_D and J_D are overlap functions:

$$I_D(a,b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

$$J_D(a,b) = \begin{cases} 0 & \text{if } abs(a,b) < x \\ 1 & \text{otherwise} \end{cases} \quad (3)$$

Weights could also be added to each feature, but more information and knowledge about the problem would be needed.

Towards building the similarity network, due to the size of the data sets, the computational time and RAM needed to run the tools and techniques chosen was too high, even after transforming the three data sets into arrays. This was seen in both the creation of the similarity network (when comparing each row to every other row, to find out if they were similar) and the community finding algorithm (when creating the network from the edge list file given). After decreasing the number of rows several times, we discovered that in order to be able to run this approach, we needed a random sample of $n = 10\,000$ rows (donations) from each of the three data sets.

After computing the distances with the tweaked HEOM, the network can be built. However, for network analysis, the most common approach is to use similarities instead of distances. The similarity is calculated by $S(i,j) = 1 - D(i,j)$, where S and D are the similarity and distance between two nodes, respectively. Then, a threshold was defined to determine which links are to be created and added to the edge list file. The edge list file was composed of two columns, one with the starting and the other with the ending node of the link. The threshold was defined at 75%, which would consider two donations to be similar if and only if they had four or more features in common out of all the features, except the one with the donation number. This threshold value was chosen since it reflected our idea of what it meant to be similar. The edge list file was then used as input for NetworkX[4], a Python package for the creation and study of complex networks. From the point of view of the resulting network, the nodes represent the donations and the links represent that the two donations connected are similar.

5.3 Results

After using the Louvain method implementation available in NetworkX to find communities in each of the three similarity networks, their modularity and giant connected components were calculated respectively. The giant connected component, which is the largest connected group of nodes, is useful since it allows one to filter the nodes of a network by keeping that group while removing the disconnected groups of insignificant size.

Table 1: Networks' and Giant Connected Components' (GCC) characteristics given 10 000 donations of each year.

	Network		GCC		
Year	#Nodes	#Links	Modularity	#Nodes	#Links
2017	9 598	1 543 913	0.588	9 598	1 543 913
2018	9 563	1 527 430	0.632	9 562	1 527 429
2019	9 594	1 541 782	0.646	9 593	1 541 781

Table 1 shows the modularity value and the giant connected component size for the donation similarity networks of 2017, 2018, and 2019 given 10 000 nodes. The values obtained for the modularity are considerably good and the GCCs cover almost all the nodes of each network. After studying the group of nodes that were left for each network, none was significant enough and ended being filtered out. With NetworkX and its spring layout, plots were obtained (example in Figure 2) by coloring each of the communities with a different color to visualize the results better. They showed 6 different communities created for 2017, 7 for 2018, and 6 for 2019.

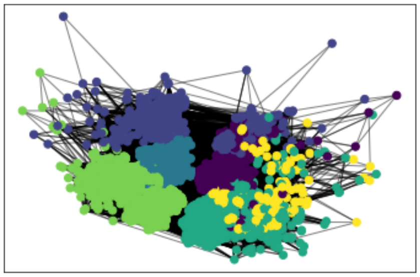

Figure 2: Representation of the communities found with NetworkX for the donation similarity network of 2019. Different Colors represent different communities.

After finding the communities, it was still difficult to understand which characteristics each partition had in particular and what distinguished each group from each other. To analyze this, a study of which features better explain each community was done. To do so, the features were plotted for each community of each year's network. From their distributions, it was possible to see that the three features that better represent each community would be Blood Type, Gender, and District. For the feature: 1) **Blood Type**, there was an almost clear separation between two sets of communities, such that over the three years most of the communities can be divided into either having blood type A+ or O+; 2) **Gender**, the partitions were considerably well divided into Feminine or Masculine; and 3) **District**, the partitions were decently divided into districts from the northern part of the country (Porto, Braga, and Aveiro) or districts from the central part of the country (Lisboa and Leiria);

From this information, an important thing to note is the fact that the communities that were found follow a pattern throughout the years, since most communities appear connected by two features. The patterns found were: **1)** Gender=Male, Blood Type=A+ (2017-2019); **2)** Gender=Male, Blood Type=O+ (2017-2019); **3)** Gender=Females, Blood Type=A+ (2017-2019); **4)** Gender=Females, Blood Type=O+ (2017-2019); **5)** Region=North, Blood Type=A+ (2017-2019); **6)** Gender=Males, Region=Center (2017-2019); **7)** Gender=Females, Region=North (2018-2019); and **8)** Blood Type=O+, Region=Center (2018-2019). We also found a pattern connected by three features: Gender=Male, Blood Type=O+, Region=Center (2018-2019). So, using these features, it was possible to extract some characteristics that can distinguish each of the communities that were found.

5.4 Transactional Mining & Association Rule Learning

5.4.1 Pre-Processing. We started by reducing the number of columns in each data set, since the algorithms used do not scale well horizontally, as they are focused on performing well over transaction data sets with a high number of rows (transactions) and a small number of features (items). To perform this reduction, we started by removing the features that would not be relevant to the experiment. This way, the first features to be removed were the ones with a high percentage of missing values, which we considered as having more than a 20% threshold of missing values. We then tested the algorithms with different sets of features in smaller data sets. After some discussion, we arrived at the conclusion that we should remove some features that were not relevant to our experiment, since they did not present any sort of new information that could be used for our goal. This was because it was already included in other features, or because the values themselves were not important to our goal. To reduce the number of features even more, we decided to collapse some of the remaining ones into a single one. These features, which could be new or not, were: 1) a feature with the brigade's info. It was composed of the several features that together characterized and identified a blood collection brigade. Seeing that a brigade keeps its characteristics throughout the time, by grouping them into one, we would have a feature that has the information about the brigade that does not change over time; 2) a single feature with the blood type group and RhD of a donor; and 3) the feature with the week of the blood collection in a year, which was picked over the ones with the day, month, or year. Lastly, we were left with four features with integer values. These features, when represented in the form of patterns and association rules, cannot be

[4] https://networkx.org/

distinguished between them, so we do not know what the value we are seeing represents. For this reason, we decided to tag each of these features with the initials of each word present in it. As an exception, the feature with the number of expected donors for a blood collection was kept as an integer.

To use the algorithms, we utilized the Mlxtend library[5], which is a Python library with multiple data science tools. The implementations of the algorithms in the Mlxtend use binary data sets. So, in order to use our data set, we had to first transform it into a binary one. In it, each line i is a donation, each column j is a feature's value, and in position ij the value 1 represents that the feature value j appears at donation i.

5.4.2 Applying the Algorithms. With the binary data set, we were then able to input it into the Mlxtend library and perform transactional mining using the `Apriori` and `FP-Growth` algorithms. After considering different minimum support thresholds, we were able to determine that for our data, the `FP-Growth` algorithm had better performance than the `Apriori` algorithm. We needed a considerably small support threshold to obtain a meaningful number of frequent patterns, this way containing more than just the obvious ones. After obtaining the frequent patterns using a minimum support of 0.01, we could perform the next step in association rule mining, which was to generate the association rules based on them. To do so, we chose the confidence as a metric of interest and then determined a threshold for it. Afterwards, we used the `FP-Max` algorithm to obtain the maximal frequent patterns while keeping the minimum support threshold. This way, the total number of patterns obtained was smaller, due to the removal of the sub-patterns of larger frequent patterns. One downside of this algorithm, which had to be kept in mind, was the fact that it focuses only on the maximal support and limits the capability of generating rules since the antecedent and the consequent supports were not computed beforehand. So, to complete the data set, we merged it with the frequent patterns data set, according to the patterns that were present in the former. To better visualize and interpret the association rules we obtained, we used the ArulesViz library in R. The two types of plot we decided to use from ArulesViz were the scatterplot and the grouped matrix plot.

5.4.3 Results. We started by studying the patterns obtained with the `FP-Growth` and `FP-Max` algorithms for the whole data sets of the years 2017, 2018, and 2019, using a minimum support threshold of 0.01. Over the course of the three years, the top patterns did not change substantially. From the patterns obtained, we can see an increase in the number of donors that are single and donate blood to brigades from the district of Lisboa, especially in 2019. Having obtained the frequent and maximum patterns, we generated the association rules based on them. After comparing the results with different minimum support and confidence thresholds, we chose to keep the same minimum support as before (i.e. 0.01) and a level of confidence above 50% (minimum confidence threshold of 0.5) to perform the rest of the experiment. The district of Coimbra is the one being represented the most in these rules. They highlight the most relevant brigade for the district, and inform us that it is highly likely to receive blood donations from males that reside in Coimbra, when the number of expected donations is 12.

We then decided to study the differences between the districts. So, we divided each of the data sets according to the home districts of the brigades that performed the blood collections. For this study, we decided to keep the same minimum support and confidence thresholds for all the new data sets. Due to size constraints, we will only present the most interesting results we obtained for the Lisbon district. For results regarding other districts, please check the full thesis document. As an example for this district, we can see Table 2 showing 3 out of the top 10 association rules for Lisboa in 2017.

Table 2: 3 association rules out of the top 10 obtained for the Lisboa's 2017 data set using the `FP-Max` algorithm with minimum support and confidence thresholds of 0.01 and 0.05, respectively. Sorted by lift.

Antecedents	Consequents	Support	Confidence	Lift
HOSPHES_POSTO AVANCADO_OUTRA_LISBOA	5	0.0123	0.9833	77.6769
SOLTEIRO, 120.0	A32PST_BRIGADA_FACULDADES_LISBOA	0.0114	0.9479	74.9894
10.0, ESTUDANTE	HOSPHSM_POSTO AVANCADO_OUTRA_, SOLTEIRO	0.0110	0.7975	25.5754

The Lisboa district's association rules with the highest lift over the three years highlighted donations that were made by donors that were university students. Following this, we can also see that Lisboa has a tendency for donations done in universities ("FACULDADES"), which was to be expected looking at the donors' jobs highlighted for the district. We can also see a high tendency for donations in other establishments ("OUTRA"). We believe these refer to mostly hospitals, since they are often related to brigade identifiers like "HOSPHSM", which in this case refers to Hospital Santa Maria, and the fact that there is a large number of hospitals in Lisboa validates this idea. The rules mostly highlight donors that are single, which once again corroborates the idea that the main donor population in Lisboa is composed of students or people that finished university recently and are still establishing their lives. Regarding the gender of the donor, Lisboa is overall balanced in terms of the rules that are highlighted, but we can see an increase in the lift of rules containing females over the three-year time frame. Lisboa's rules highlight the blood type of the donor, focusing on blood types A+ and O+. This information was to be expected, since these two are the most common blood types both in the data sets and in the world. The weeks in which the donations are made are not highlighted in the top 20 rules. This may be explained by the fact that Lisboa has a high number of donations done in hospitals and most of them are open over the entirety of each year to accept blood donations, meaning that there is an equal distribution of donations over the year. Similarly, the number of total donations made by a donor is not highlighted, meaning that the values do not present a tendency. Another aspect that might justify the fact that Lisboa has a high number of donations done in hospitals, is that the number of donations expected by brigade is mostly 10 and sometimes 120. The latter was expected, since Lisboa is the capital of Portugal, so

[5] http://rasbt.github.io/mlxtend/

it would be normal to expect a high number of donors to show up to a blood collection done by a brigade, but 10 does not follow this thought process. However, if we look at the fact that hospitals are open throughout the year, this value can be explained as the number of expected donors for a single day in one of them.

5.5 Sequential Pattern Mining

5.5.1 Pre-Processing. The initial part of the pre-processing done for this approach was similar to the one done for the transactional mining and association rule learning one, so we will not cover it again. We began by using the same three data sets and performed the same pre-processing thought process up until the point before creating tags for each feature with integer values. To use the algorithms for this approach, we utilized the implementations available in the SPMF library[6]. SPMF is a Java open-source software and data mining library, created by Philippe Fournier-Viger, and contains implementations of different pattern mining algorithms. The input for the algorithms we planned on using was a sequence data set or a time-extended sequence data set. The difference between a sequence data set and a time-extended one is that, besides having sequences where each sequence is a list of itemsets, each itemset is annotated with a timestamp.

The input file format for the SPMF library's implementation of the algorithms we used had to respect the following characteristics: 1) It is a text file where each line represents a sequence (or a time-extended sequence) from a sequence data set; 2) Each of the items in the itemsets is represented by a positive integer, and items from the same itemset within a sequence are separated by single spaces; 3) The end of an itemset is indicated by a "-1", and after all the itemsets, the end of a sequence is indicated by a "-2"; and 4) If it is a time-extended sequence each itemset is first represented by its timestamp, which is a positive integer between the "<" and ">" symbols. The implementations of these algorithms assume that the items are sorted according to a total order in each itemset, and that no item appears more than once in the same itemset. So in order to use our data, we created the input file with the respective format, while indexing each of the different values present in our data sets to unique integer values. Each sequence represented a blood collection brigade, and each itemset in the sequence represented a donation that was collected by that brigade.

5.6 Applying the Algorithms

We started by using the `PrefixSpan` algorithm, since it is one of the most popular sequential pattern mining algorithms. However, due to the large number of different values present in some of the features, the number of different indexed values was too large for the algorithm to be able to run with acceptable computational time. So we started reducing the cardinality of the features by: 1) Keeping only the 10 most common values from the donor's job feature and referring to the other ones as a new value "other"; 2) Replacing the donor's local of residence according to the zip code with the donor's district of birth, reducing the value cardinality from 4 555 to 29 and joined the different values corresponding to all the Portuguese islands into a single new value "ILHAS" reducing the value cardinality from 29 to 19; 3) Binning the values in the feature with the total number of donations. The bin dimension increases since the number of donors with a higher donation count is not as common as donors with a smaller one; 4) Binning the values in the donor's age feature. The first bin starts at 18, since that is the minimum legal age to be able to donate blood; 5) Binning the values in the feature with the expected number of donors for a blood collection; 6) Binning the values in the feature with the donation's month so that a year would be divided into trimesters. However, this change was not enough to decrease the computational time. So, we looked at and tested the other sequential pattern mining algorithms available in the SPMF library, and we concluded that the fastest algorithm was `CM-Spade`, as it was mentioned in the survey conducted by P. Fournier-Viger et al. [12]. With the algorithm exponentially scaling with more than 3 itemsets per sequence, it was not able to run our data sets with acceptable computational time either. For example, the data set for Lisboa in 2019, had 52 649 donations collected by 384 different brigades, meaning that on average each sequence had 137 itemsets. To solve this issue, we compressed the data sets as much as possible, by reducing the total amount of donations per brigade to a maximum of 4 (one per trimester). To do this, we calculated the mode of each of the features composing all the donations done in a brigade in each trimester and collapsed them into the format of a single donation. Following the example given above regarding the 2019 Lisboa data set, it had now been reduced to an average of 1 itemset per sequence.

To increase the amount of information obtained from the data sets, we started to gradually remove some of the changes that had been made to the data sets until, after several attempts, we reached a point where we decided to stop. After such attempts, we ended up completely removing the changes made in 3) and 5). In 1) and 6), all the different jobs and months were taken into account, respectively. The removal of the changes made in 6) also meant that the total amount of donations per brigade was increased to 12 instead of 4, so now in the 2019 Lisboa data set we had an average of 2 itemsets per sequence (793 donations collected by 384 different brigades). The changes done in 2) and 4) were kept. Seeing that the readability of the patterns was difficult, we also decided to tag each of the features with the respective feature name followed by an "=" symbol.

Finally, with our data sets transformed into the proper input file format, we used them as input to the SPMF library's implementation of the `CM-Spade` and `Fournier08` algorithms. Since the `Fournier08` algorithm needed a time-extended file, we used the month m to represent the timestamps $< m >$.

5.6.1 Results. For the results shown, please note that the end of an itemset is represented by a |. These algorithms allow us to check the support of a sequence of donations that were done by donors with particular characteristics in a certain brigade over a year. Taking into account that we applied the mode for each month in order to reduce the data set, please keep in mind that each itemset characterizes the most common type of information for each feature regarding a donor in each month. As the dimension of the tables containing the patterns with the greatest lengths was too large, in this paper we decided to only present a table with the results for Lisbon in 2019 using the `CM-Spade` algorithm. For more results, please refer to the full thesis document.

[6] https://www.philippe-fournier-viger.com/spmf/

The top three patterns (sorted by length) obtained using the CM-Spade algorithm for the 2019 data set, with a minimum support threshold of 0.05, are shown in Table 3. As an example, in a certain month, if we have donations that were done by a majority of donors that were married and born in Lisboa, there is 0.0521 support of in a following month the donations being mostly done by donors that were female, between 45 and 55 years old, married and born in Lisboa, which was then followed by a month with mostly donations where the donor was female, married and born in Lisboa.

Table 3: Top 3 patterns obtained for the 2019 data set of Lisboa using the CM-Spade algorithm with minimum support threshold of 0.05. Sorted by length.

Pattern	Support
dador_estado_civil = CASADO dador_naturalidade_distrito = LISBOA | dador_sexo = feminino dador_idade = [45,55[dador_estado_civil = CASADO dador_naturalidade_distrito = LISBOA | dador_sexo = feminino dador_estado_civil = CASADO dador_naturalidade_distrito = LISBOA |	0.0521
dador_tipo_sangue = A+ dador_estado_civil = CASADO dador_naturalidade_distrito = LISBOA | dador_tipo_sangue = A+ dador_estado_civil = CASADO dador_naturalidade_distrito = LISBOA | dador_tipo_sangue = A+ dador_estado_civil = CASADO dador_naturalidade_distrito = LISBOA |	0.0521
dador_idade = [45,55[dador_estado_civil = CASADO dador_naturalidade_distrito = LISBOA | dador_sexo = feminino dador_estado_civil = CASADO dador_naturalidade_distrito = LISBOA | dador_sexo = feminino dador_estado_civil = CASADO dador_naturalidade_distrito = LISBOA |	0.0521

6 CONCLUSIONS

In this thesis, we have applied three different novel approaches to a blood donation data set from IPST. We started by transforming the data set into a similarity network to be used as input for the Louvain Method, which is a community finding algorithm. Then we applied the Apriori, FP-Growth, and FP-Max algorithms to the data set to perform a transactional mining and association rule learning approach. Finally, we also applied sequential pattern mining methods, namely PrefixSpan, CM-Spade, and Fournier08 to the data set. We discuss below the results and performance of the different approaches from the point of view of computational time and resources needed to run them, and the results they generated.

Regarding performance, out of the three approaches, the one best suited for the IPST data was the transactional mining and association rule learning approach, seeing that it was not only able to deal with the highest amount of data, but was also the fastest one to generate results. However, and even if it was the best out of these three approaches, it is important to keep in mind that it was still similarly affected by the fact that our data was composed of a large data set, both vertically and horizontally. These approaches are mostly used to deal with transactional data, so they were optimized to deal with data sets with a high number of rows, but a low number of columns, which was not the case in our data set and was reflected in the performance of the used approaches. To deal with this, we had to reduce the data set's size and divide it into smaller subsets according to specific years (2017, 2018, and 2019) and/or districts and perform other changes as a form of pre-processing, which are described in the results' section in more detail.

As for the quality of the results themselves, we can consider them to be relevant or interesting if they satisfy at least one of the following criteria: 1) can be easily interpreted by humans; 2) are valid with a certain degree of certainty; 3) can be useful in some way; and 4) present new information. Taking these criteria into account, out of the three, the one that presented the worst results was the community finding approach. This choice was made since it was the one that lacked a means of explanation that could be easily interpreted by humans, considering that the modularity focuses on measuring the strength of the division of a network into modules. So, it does not present any real readability for a human, and it does not present any degree of certainty, especially since we were the ones defining the similarity threshold in order to create the network, generating a bias. Focusing on the other two approaches, we can see that there are some differences in the results obtained between them in the top 5 districts in terms of length. This might be due to the fact that, in order to perform sequential pattern mining, we had to join the information of a brigade for each month in a single row by performing the mode of the different features. In the sequential pattern mining approach, it is also important to keep in mind that the results obtained by CM-Spade and Fournier08 when sorted by length are different, presenting two types of information. In the former, as long as the pattern follows a sequence, the time frame does not need to be taken into account, while in the latter, we also take into account the exact time frame, so the results present more accuracy when defining the time sequence.

To conclude, we were able to find patterns that contained information to be provided to IPST in the form of hints to satisfy its activity plans while optimizing the operations of the teams they organize, which was the goal of this thesis. Some examples of hints that could be given according to the pattern we uncovered include: 1) The most relevant brigades of each Portuguese district were identified, so more resources could be dispatched to them, while reducing the amount of resources spent on the less relevant ones. This could be done, for example, by improving the frequency of blood collections in a place where the most relevant brigades have been collecting blood donations, while decreasing the frequency of less relevant ones; 2) The jobs with the highest number of donations were also identified for each district, so IPST could create new brigades or dispatch older ones to places near their places of work, such as universities in Lisboa to collect blood from higher education students, or factories in Aveiro to collect blood from factory employees; 3) The months with the highest donation rate were also identified, which allows us to tell IPST hints on when they could scatter their brigades throughout the year for each district; 4) We identified patterns that showed that there is an increase in the number of donors that are in the 20 to 25 year bracket, or female donors over the past three years. So, IPST could advertise more to donors with these characteristics in order to bring a higher number of them to donate or keep donating, since it has been proven to be successful. Another option would be to advertise for the other types of donors that are still part of a minority to increase the number of donations made by them.

We would also like to bring up the fact that even though some of the information could be gathered without using this sort of algorithms, when using pattern mining algorithms we are able to retain the probability of some aspect happening in the real world, for example, in the form of the support metric. This can be considered relevant information when we take into account that these algorithms might justify some of the ideas that people that work in this field may have, or even contradict them.

6.1 Limitations and Future Work

Besides the amount of data present in the data set proving to decrease the performance of the approaches, it together with the quality of the data (for example, there were cases where the cardinality for some of the features was too high, which hindered the obtainment of more concrete results, or features where there was a high number of wrong or missing values) were two issues that were referred throughout this document as having an impact on the approaches chosen. However, we believe that if we had any other approach, the same issues would have come up. With this in mind, we believe that in the future, the data needs to be treated better as a whole, removing the incongruities as well as taking care of the unknown or other values present in the data set.

When this reiteration of the data is done, we could reapply our approaches to see the difference in the results, as well as new ones. One example of a new approach would be a biclustering one. For it, we could use algorithms like Qubic2 [31] or BicPams [16] to study a biclustering approach and the modules discovered by it when applied to the IPST data. To have a professional point of view on the topic and to know the utility and novelty of our results, we have also communicated the most relevant results to IPST.

7 ACKNOWLEDGEMENTS

The database complies with current legislation for the protection of data, in this case, from benevolent blood donors - their identity being unknown. Data were used for academic purposes only.

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference DSAIPA/AI/0033/2019 (project LAIfeBlood) and UIDB/50021/2020 (INESC-ID multi-annual funding).

REFERENCES

[1] Maryam Ashoori, Shahriar Mohammadi, and Hoda Eivary. 2016. Exploring Blood Donors' Status Through Clustering: A Method to Improve the Quality of Services in Blood Transfusion Centers. *Journal of Knowledge & Health* 11 (12 2016), 73–82. https://doi.org/10.1234/jkh.v11i4.1525

[2] Maryam Ashoori and Zahra Taheri. 2013. Using Clustering Methods for Identifying Blood Donors Behavior. In *5th Iranian Conference on Electrical and Electronics Engineering (ICEEE)*. 4055–4057.

[3] Albert-László Barabási et al. 2016. *Network science* (1st ed.). Cambridge university press. http://networksciencebook.com/chapter/9#testing

[4] Jeroen Beliën and Hein Forcé. 2012. Supply chain management of blood products: A literature review. *European Journal of Operational Research* 217, 1 (2012), 1–16. https://doi.org/10.1016/j.ejor.2011.05.026

[5] John T. Blake and Matthew Hardy. 2014. A generic modelling framework to evaluate network blood management policies: The Canadian Blood Services experience. *Operations Research for Health Care* 3, 3 (2014), 116–128. https://doi.org/10.1016/j.orhc.2014.05.002

[6] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics Theory and Experiment* 2008, 10 (04 2008), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

[7] Wijai Boonyanusith and Phongchai Jittamai. 2012. Blood Donor Classification Using Neural Network and Decision Tree Techniques. In *World Congress on Engineering and Computer Science*, Vol. 1. 499–503.

[8] André Carreiro, Sara C . Madeira, and Alexandre Francisco. 2013. Unravelling communities of ALS patients using network mining. In *ACM SIGKDD Workshop on Data Mining in Healtcare.*

[9] Vineet Chaoji, Mohammad Hasan, Saeed Salem, and Mohammed Zaki. 2008. An integrated, generic approach to pattern mining: Data mining template library. *Data Min. Knowl. Discov.* 17 (12 2008), 457–495. https://doi.org/10.1007/s10618-008-0098-x

[10] Mohamad Darwiche, Mathieu Feuilloy, Ghazi Bousaleh, and Daniel Schang. 2010. Prediction of blood transfusion donation. In *Proceedings of the Fourth IEEE International Conference on Research Challenges in Information Science.* 51–56. https://doi.org/10.1109/RCIS.2010.5507363

[11] FCT. 2019. LAIfeBlood Project. https://www.fct.pt/

[12] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, Yun Sing Koh, and Rincy Thomas. 2017. A survey of sequential pattern mining. *Data Science and Pattern Recognition* 1, 1 (2017), 54–77.

[13] Michael Hahsler and Sudheer Chelluboina. 2011. Visualizing Association Rules in Hierarchical Groups. *42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms* (01 2011).

[14] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier, USA.

[15] Sandhya Harikumar and Surya Pv. 2015. K-Medoid Clustering for Heterogeneous DataSets. *Procedia Computer Science* 70 (12 2015), 226–237. https://doi.org/10.1016/j.procs.2015.10.077

[16] Rui Henriques, Francisco L. Ferreira, and Sara C. Madeira. 2017. BicPAMS: software for biological data analysis with pattern-based biclustering. *BMC Bioinformatics* 18 (02 2017). https://doi.org/10.1186/s12859-017-1493-3

[17] IPST. 2014. Plano Nacional de Emergência para Eventos com Potencial Impacto na Missão do IPST,IP.

[18] IPST. 2020. Plano de Atividades do IPST,IP 2020-2022 | Homologado pela Ministra da Saúde.

[19] IPST 2020. Relatório de Atividades do IPST,IP - 2019 | Homologado pela Ministra da Saúde.

[20] Sebastian Klenk, Juergen Dippon, Peter Fritz, and Gunther Heidemann. 2010. Determining Patient Similarity in Medical Social Networks. In *MEDEX 2010 Proceedings*, Vol. 572. 6–14.

[21] T. Li, Y. Chen, Xiangwei Mu, and Ming Yang. 2010. An improved fuzzy k-means clustering with k-center initialization. In *Third International Workshop on Advanced Computational Intelligence.* IEEE, 157–161. https://doi.org/10.1109/IWACI.2010.5585234

[22] Janne Lumijärvi, Jorma Laurikkala, and Martti Juhola. 2004. A comparison of different heterogeneous proximity functions and Euclidean distance. *Studies in health technology and informatics* 107, Pt 2 (2004), 1362–1366.

[23] A. S. Martins, M. Gromicho, S. Pinto, M. Carvalho, and S. C. Madeira. 2021. Learning Prognostic Models using Disease Progression Patterns: Predicting the Need for Non-Invasive Ventilation in Amyotrophic Lateral Sclerosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 01 (may 2021), 1–1. https://doi.org/10.1109/TCBB.2021.3078362

[24] Mohamed M. Mostafa. 2009. Profiling blood donors in Egypt: A neural network analysis. *Expert Systems with Applications* 36 (04 2009), 5031–5038. https://doi.org/10.1016/j.eswa.2008.06.048

[25] National Research Council et al. 2006. *Network science.* The National Academies Press, Washington, DC. https://doi.org/10.17226/11516

[26] Shraddha Pai and Gary D. Bader. 2018. Patient Similarity Networks for Precision Medicine. *Journal of Molecular Biology* 430, 18, Part A (2018), 2924–2938. https://doi.org/10.1016/j.jmb.2018.05.037

[27] P. Ramachandran et al. 2011. Classifying Blood Donors Using Data Mining Techniques. *International Journal of Computer Science & Engineering Technology* 1 (02 2011), 10–13.

[28] T. Santhanam and Shyam Sundaram. 2010. Application of CART Algorithm in Blood Donors Classification. *Journal of Computer Science* 6 (06 2010), 548. https://doi.org/10.3844/jcssp.2010.548.552

[29] Murat Testik, Banu Yuksel-Ozkaya, Salih Aksu, and Osman Ilhan. 2010. Discovering Blood Donor Arrival Patterns Using Data Mining: A Method to Investigate Service Quality at Blood Centers. *Journal of medical systems* 36 (05 2010), 579–94. https://doi.org/10.1007/s10916-010-9519-7

[30] Bondu Venkateswarlu and G. V. S. Raju. 2013. Mine Blood Donors Information through Improved K-Means Clustering. *ArXiv* abs/1309.2597 (2013).

[31] Juan Xie et al. 2019. QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data. *Bioinformatics* 36, 4 (09 2019), 1143–1149. https://doi.org/10.1093/bioinformatics/btz692 arXiv:https://academic.oup.com/bioinformatics/article-pdf/36/4/1143/38712518/btz692.pdf