



Chapter 4

Healthcare Support Using Data Mining: A Case Study on Stroke Prediction



Georgios Michailidis, Michail Vlachos-Giovanopoulos,
Paraskevas Koukaras , and Christos Tjortjis 

Abstract Data and information have become a valuable asset for small and big organizations in the past few decades. Data is the main ingredient for strategic decision-making, which could give businesses a significant advantage over their competitors, by providing customized services or overall experience to their customers and attracting new ones. For this purpose, data mining techniques are being utilized so that valuable information can be discovered and exploited. There is a vast amount of data generated in the field of healthcare that is not getting fully exploited by traditional methods, for reasons, such as their complexity, velocity, and volume. Therefore, there is a demand for the development of powerful automated data mining tools for the complete utilization of these data, and the uncovering of patterns and precious knowledge about patients, medical claims, treatment costs, hospitals, etc. This work focuses on exploiting the best-known data mining techniques: classification, clustering, and association rule mining, which are utilized extensively in the healthcare industry for incident prediction and general medical knowledge acquisition. The data mining process comprises several steps, such as data selection, pre-processing, transformation, interpretation, and evaluation. The section of the experimentation includes a stroke incidents dataset fetched from the Kaggle dataset provider. This chapter also provides a literature survey on data mining applications in the healthcare sector, while discussing the abovementioned machine learning concepts.

G. Michailidis · M. Vlachos-Giovanopoulos · P. Koukaras · C. Tjortjis (✉)
The Data Mining and Analytics Research Group, School of Science and Technology, International
Hellenic University, 570 01 Thermi, Greece
e-mail: c.tjortjis@ihu.edu.gr

G. Michailidis
e-mail: gmichailidis@ihu.edu.gr

M. Vlachos-Giovanopoulos
e-mail: mvlachos@ihu.edu.gr

P. Koukaras
e-mail: p.koukaras@ihu.edu.gr

Keywords Data mining · Healthcare support · Machine learning · Stroke prediction

Term Definition Table

Term	Definition
Bmi	Body Mass Index
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRM	Customer Relationship Management
Data mining algorithm	A set of heuristics and calculations that creates a model from data
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
Feature	An independent variable that can be used as input of a machine learning model
FP-Growth	Frequent Pattern Growth
GDPR	General Data Protection Regulation
LEADERS	Lightweight Epidemiological Advanced Detection Emergency Response System
Machine learning model	An expression of an algorithm that processes data to explore patterns or make predictions
OLAP	Online Analytical Processing
Operator	In RapidMiner, a group of functions that perform actions on input data through parameters and outputs the results of said actions
Parameter	A special kind of variable used in an algorithm to refer to one of the pieces of data provided as input to it
Process	In RapidMiner, it is a set of sequentially connected operators represented by a flow design, where each operator provides its output as input to the next one
ROC	Receiver Operating Characteristic
SEMMA	Sample, explore, modify, model, assess
SVM	Support Vector Machine

4.1 Introduction

Over the last few decades, data analysis has become a vital tool for both small and large enterprises. Any business with a presence on the various types of social media platforms [1] and the web is collecting a vast amount of data about customers, user behavior, web traffic, demographics and more. The trajectory, as well as the

success of any organization in the market, depends highly on the appropriate exploitation of these data. In other words, data is the main component for the strategic decision-making process, which could give the business a significant lead against their competitors by providing customized services or overall experience to their customers and attracting new ones [2].

The process of exploring and analyzing, by automatic or semi-automatic means, large quantities of data to summarize them into valuable information and discover meaningful patterns and extract useful knowledge, is called “data mining” (a.k.a. knowledge discovery) [2]. Consequently, data mining comprises functional elements that transform data stored into a data warehouse, manage them in a multidimensional database, make data access easier to experts, analyze data using tools and techniques, and present them in a way that valuable information can be discovered [2, 3].

4.1.1 Data Mining

The entire concept of data mining has been developed in the 90’s and emerged as a powerful tool that is capable of discovering previously unknown patterns and valuable information from massive datasets comprised of thousands of records [3–6]. Previous studies mentioned that data mining approaches enable data owners to analyze and find unexpected connections in their data, which turn out to be supportive for decision making [7]. The data mining process is split into two parts, Data Preprocessing and Data Mining. During the data preprocessing phase, different algorithms and processes perform data cleaning, integration, reduction, and transformation. On the other hand, data mining phase involves pattern mining, evaluation, and knowledge representation, as shown in Fig. 4.1.

Skills and knowledge are necessary ingredients for implementing data mining since the performance depends on the person who is performing the procedure due to non-appearance of a basic structure. The Cross-Industry Standard Process for Data Mining (CRISP-DM) suggests a principle for data mining tasks, and splits them into six stages: business understanding, data understanding and preparation, modeling, evaluation, and deployment [8]. The first stage is rather crucial because it recognizes the business objectives and, therefore, the success standards of mining tasks. Data is unquestionably an essential component since it is the raw material. Without data we cannot mine information. Thus, according to CRISP-DM, data understanding and preparation are considered as prerequisites for modeling. Subsequently, modeling is the definite data analysis.

Most data mining operating systems contain Online Analytical Processing (OLAP). That is, non-traditional statistical analysis, like neural networks, decision trees, link and association analysis, and traditional statistical methods, like clustering, discriminant and regression analysis. This wide range of techniques is expected, given that data mining has been considered a descendant of three different disciplines, specifically computer science (also counting machine learning and artificial intelligence), statistics and database management. The evaluation phase allows models

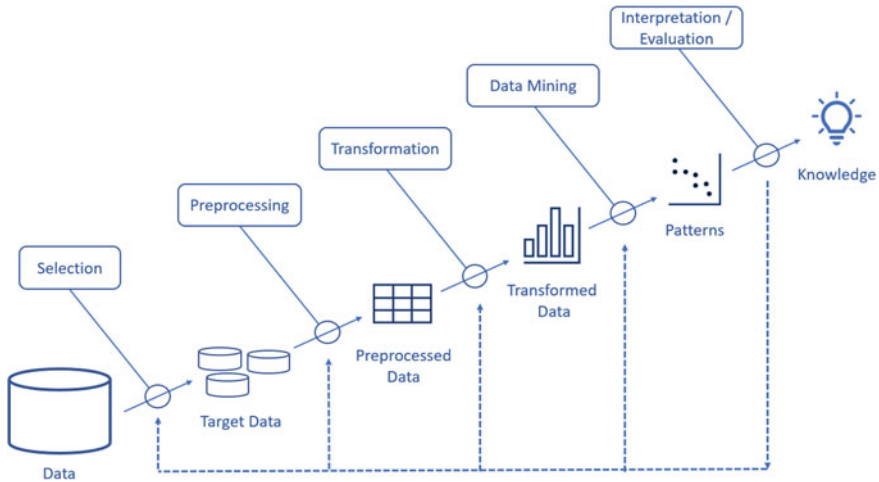


Fig. 4.1 Stages of data mining

and results to be compared based on a common criterion, such as lift charts, profit charts, or diagnostic classification charts. Finally, deployment is where data mining pays off. It has to do with the implementation and utilization of the models [4].

Data mining can be considered as a process, rather than a set of tools, and there is a methodology that illustrates this process. The SEMMA (Sample, Explore, Modify, Model, Assess) method splits data mining into five phases: (i) sample, to design a statistically representative sample of the data, (ii) explore, to apply exploratory, statistical and visualization techniques, (iii) modify, to select, create and transform the most important predictive variables, (iv) model, to model the variables to perform predictions, and (v) assess, to validate the model's accuracy. SEMMA is an iterative method. The internal stages can be performed repetitively based on the needs. Figure 4.2 illustrates the five stages of the SEMMA methodology [9].

4.1.2 Data Mining in Healthcare

Data mining applications are becoming more and more essential in healthcare [10, 11]. Various reasons have stimulated the use of data mining applications in healthcare. The enormous amounts of data generated by the healthcare industry are being under-utilized by using traditional methods due to the complexity, velocity, and volume of them. So, there is a need for developing powerful automated data mining tools for appropriate exploitation of these data, since they hide valuable information about patients, medical claims, treatment costs, hospitals, etc. [8].

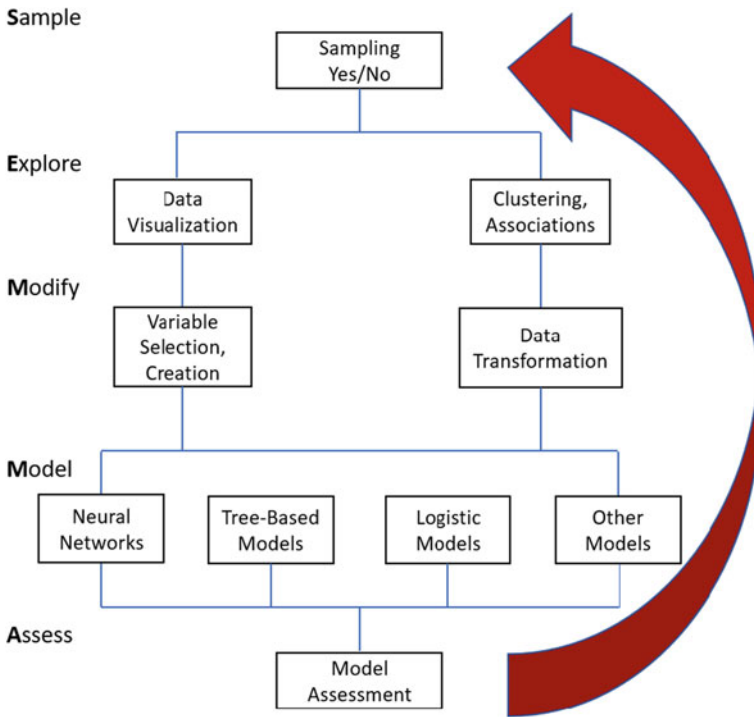


Fig. 4.2 SEMMA methodology

4.1.3 Applications of Data Mining in Healthcare

Various reasons have led to using knowledge discovery techniques in the healthcare sector. Data mining can upgrade decision-making by uncovering patterns and trends in huge amounts of complicated data. Financial burden has intensified the need for healthcare organizations to guide their management according to the analysis of clinical and financial data. Wisdom gain from knowledge discovery processes can motivate more sufficient and cost-effective operation and high-quality services.

Healthcare organizations adopting data mining techniques benefit and are better qualified for meeting future requirements. Another reason stimulating the utilization of data mining approaches in healthcare is the appreciation that they can produce knowledge that is supportive to every part of the healthcare system. For instance, data mining techniques can assist healthcare insurers to identify fraud and abuse. Medical centers, physicians and patients can benefit by finding better practices and treatments [4].

Such applications can be grouped as evaluation of treatment effectiveness, healthcare management, fraud, and abuse detection in health insurance, causes of diseases detection, and more [4]. The analysis of health data upgrades the healthcare system by boosting the effectiveness of tasks that have to do with the management of the

patients. Data mining techniques aim to support healthcare organizations on grouping patients with common health issues, so that they receive the appropriate treatment. They can also be used for predictions, such as predicting the length of hospitalization, predicting diagnosis, etc. Data mining technologies also provide benefits by analyzing various characteristics that cause a disease, like lifestyle, food, education, working environment etc. [8].

4.1.4 Chapter Overview

The following section, reviewing the literature, showcases various applications of data mining techniques in healthcare sector, as well as utilizes some of the most notable data mining methodologies. Such applications could be healthcare management, fraud and abuse detection, and treatment effectiveness, which are going to upgrade the overall healthcare quality. Moreover, machine learning concepts that are related with healthcare support are also discussed. With concepts like classification, regression, clustering and association rule mining, valuable predictions can be performed, such as to predict if a patient is likely to suffer from a particular disease.

Typically, various preparatory actions must be completed before reaching the stage where algorithms may be applied. First, data selection needs to be applied, in order to keep only the most relevant content. The steps that follow are data cleaning, integration, reduction, and transformation, which remove duplicate observations, outliers, noise, and deal with missing values, and data sampling to avoid overfitting. The next step is data mining, where the preprocessed data are fed into classification, regression, clustering, and association rule mining algorithms. It is a trial-and-error process, as there is no rule to choose the best algorithm for each data mining case. Finally, after running the data mining algorithms, some results are produced. From these results, the interpretation and evaluation step arise, when new useful knowledge is attained.

4.2 Literature Review

This section contains a description of several data mining applications in the healthcare sector, as well as machine learning concepts related with healthcare support. As already mentioned, information gained by utilizing data mining methodologies upgrade the overall healthcare quality, from both patient's and healthcare management's perspectives. In this chapter, the most noteworthy applications and machine learning concepts are reviewed.

4.2.1 Data Mining Applications in Healthcare

4.2.1.1 Healthcare Management

Data mining approaches can be used to detect and track the chronic disease states and high-risk patients, and make good use of hospital resources, which is an important task in healthcare [4, 8]. The main goal is to identify patients who need additional care, enhanced health quality, and cost-effective services. Other factors, such as physical condition and demographic details of patients, play a significant role in the appropriate utilization of the available hospital resources. These resources, both physical and human ones, are managed through an automated data mining procedure.

For instance, Group Health Cooperative, through data mining approaches, offer several cost-effective healthcare services. A non-profit organization enables patients to access their medical information through an online platform, where they can also fill the prescription form and securely communicate with the healthcare providers. Seton Medical Centre used knowledge discovery techniques to make information on the patient's health available, decrease admitted duration of patients in the medical centers, and thus improve healthcare quality. Blue Cross, through data mining, constructed a model which contributes to high-performance and cost-conscious management of diseases. With the help of data mining, Sierra Health Centre instructs appropriate management of the treatment and its cost and recognizes the prospects of upgrading health quality [8].

Data mining approaches can also be used to process big amounts of data and detect patterns related to bioterrorist attacks. For instance, the Lightweight Epidemiological Advanced Detection Emergency Response System (LEADERS) used knowledge discovery methods to reveal various epidemics [4].

4.2.1.2 Customer Relationship Management (CRM)

CRM is a central approach in managing interactions between commercial organizations and their customers, also in healthcare. When talking about customer interactions, we refer to interactions through call centers, billing departments, physicians' offices, inpatient settings, and ambulatory care settings [4]. Data mining can be used to identify preferences, usage patterns and current needs of individuals, to improve the relation with them [4, 8]. These approaches can also be used to predict future needs of individuals, and other products that a patient may purchase, whether a patient is likely to comply with prescribed treatment or whether preventive care is likely to decrease future use.

Customer Potential Management Corp. has established a Consumer Healthcare Utilization Index signifying how likely it is for a person to make use of healthcare services, defined by 25 major diagnostic categories, selected diagnostic related groups or medical service areas. With the help of healthcare transactions of millions

of patients, this index can recognize the ones who can derive advantage from health-care services, support the ones who most need access to specific care, and constantly improve the means used to reach groups of people for enhanced health and long-term patient relationships. OSF Saint Joseph Medical Centre has used the index to deliver the appropriate messages and services to the appropriate patients at the right time. The outcome of this action were more efficient communications and higher earnings [4].

It was also suggested that knowledge discovery of patient survey data can define logical expectations about waiting times and uncover information about the kind of services patients anticipate from the providers, and how to improve these services [4].

On the other hand, pharmaceutical companies can derive advantage from health-care CRM and data mining, since by keeping track of physicians prescribing drugs along with the reason, they are able to recommend the most efficient and effective treatment plans, match physicians to specific groups of patients, and track the trajectory of an epidemic to better handle the situation. Moreover, such companies can also benefit from analyzing big amounts of genomic data to detect a link between a patient's genetic makeup and their response to a drug therapy [4].

4.2.1.3 Fraud and Abuse Detection

Healthcare insurers can use data mining to identify illegitimate prescriptions or referrals, fraudulent insurance and abnormal or fake patterns in medical claims by laboratories, clinics, physicians, or patients [8]. Many incidents were recorded when health insurance companies used these techniques and reported millions of dollars of annual savings. For instance, Texas Medicaid Fraud and Abuse Detection System which detected 1400 cases and saved \$2.2 million in 1998, Australian Health Insurance Commission saved millions of dollars in 1997, and ReliaStar Financial Corp. has reported a 20% increase in annual savings [4].

4.2.1.4 Treatment Effectiveness

Data mining techniques may be used to compare causes, symptoms, and treatment plans to assess the efficacy of medical therapies [4, 12]. Physicians and patients are given the ability to analyze different treatments and their effectiveness to decide which technique is a better and cheaper choice [8]. For instance, the results of different groups of patients who received different medicines for the same health issue can be compared to decide which medicines perform better and are of lower cost. United HealthCare has analyzed its treatment data to find out how to offer more efficient medicine. It also created clinical profiles through which physicians can gain knowledge and compare their practice patterns with those of other physicians. Consequently, knowledge discovery techniques can uncover information about the appropriate therapy for particular diseases [4].

4.2.1.5 Improved Patient Care

Huge amounts of data are collected in the form of electronic health records. Digitized data regarding patients contribute to the enhancement of the healthcare system's quality. Using data mining, a predictive model was created in order to process and analyze these data and discover valuable information that lead to making appropriate decisions regarding the upgrade of the healthcare quality. Healthcare providers are able, through knowledge discovery processes, to recognize not only present, but also future needs and preferences of patients and offer more qualitative services. Data mining techniques can also equip patients with knowledge they need to have about diseases and their prevention, and they can also group patients with common features [8].

4.2.1.6 Hospital Infection Control

Data mining can be used to analyze and detect previously unknown or abnormal patterns in infection control data. Then, through association rules, unforeseen and useful information can be mined from the public surveillance and hospital control data. Finally, experts analyze this information and propose actions to control the infection in the hospitals [8].

4.2.1.7 Hospital Ranking

Through various data mining tools, several hospital details are examined and analyzed, in order to define their ranks. The ranking system considers the ability of taking care of high-risk patients. Higher ranked medical centers deal with high-risk patients on top priority. On the contrary, lower ranked hospitals are not able to handle such situations [8]. In addition, risk factors need to be considered from the healthcare providers' side when reporting information. This means that a hospital's rank will be lower because they reported a greater difference between predicted and actual death rate, no matter if their success rates are equal to those of other healthcare providers. So, it is important that the reports contain consistent and reliable information [9].

4.2.2 Machine Learning Concepts Related with Healthcare Support

4.2.2.1 Classification

Classification is one of the most popular techniques of data mining in healthcare. It splits data samples into training and test sets and predicts the target class of each observation. One good example of classification in the healthcare sector is the association of a risk factor to patients after analyzing their disease patterns. Classification belongs to the supervised learning methods (there are also semi-supervised approaches such as [13]), which means that the class categories are known.

According to the number of classes, it is characterized as binary classification, for two possible classes, such as “high” or “low” risk patient, and multiclass classification, for more than two possible classes, such as “high”, “medium” and “low” risk patient. The initial dataset gets split into training and test dataset. The algorithm uses the training set and tries to explore the relationship between the attributes in the training set to predict the result. Then, the algorithm utilizes the test set, where the class attribute is not known yet, contrary to the training set. By analyzing the input, the algorithm performs the prediction. Finally, as for the evaluation of the performance, with the help of the accuracy is it defined how “good” predictions the algorithm achieved [14].

Some of the most popular classification algorithms are K-Nearest Neighbors, Decision Tree, Support Vector Machine (SVM), Neural Network, and Bayesian Methods [14, 15].

Das et al. tried to diagnose heart diseases with the help of data mining. For this purpose, they mainly used a neural networks ensemble model, which achieved almost 90% accuracy by utilizing data from Cleveland heart disease database. So, they managed to create an intelligent medical decision support system to help healthcare assistants [16].

Curiac et al. conducted an experiment analyzing psychiatric patient data to examine the most important factors and their correlations with some diseases. They used the Bayesian Networks technique and they found that it is a very useful tool which can be used to support physicians in the process of prediction and diagnosis of psychiatric diseases [17].

4.2.2.2 Regression

Regression is another popular method of knowledge discovery, which is used to discover the significant variables in terms of correlation [8, 14]. It is a mathematical model which, similarly to classification, uses the training set to be constructed. There are two types of variables that are used in statistical modeling, that are called dependent (usually represented using “Y”) and independent (usually represented

using “X”) variables respectively. Every model can have one or more independent variables, while the dependent variable is only one.

According to the number of independent variables, regression can be characterized as Linear or Non-linear. Linear regression refers to explaining the relation between a dependent variable and one or more independent ones and works only with numerical data. When the data are of categorical type, logistic regression can be used, which is a type of non-linear regression, and is divided into Binomial and Multinomial types. Binomial regression’s role is to predict the outcome for a dependent variable when there are two possible outcomes, while multinomial deals with three or more possible outcomes [14]. Other common regression methods are Support Vector Regression, Decision Tree Regression, and Multilayer Perceptron [14].

Tomar and Agarwal analyzed real time body information through sensors and predicted the activity to provide continuous monitoring and better healthcare services. Based on their experiment, they proved that Least Square Support Vector Regression achieved more accurate results than other Support Vector Regression techniques examined [18].

Alapont et al. presented a research project related to hospital management, where they examined various machine learning techniques, such as Linear Regression, Least Med Squared Linear Regression, SVM for Regression, Multilayer Perceptron, K Star, Locally Weighted Learning, Tree Decision Stump, Tree M5P, and IBK. After executing the algorithms, they found that Linear Regression and Tree M5P produced the best results in terms of hospital management, such as human and physical resources, ward management, emergencies, etc. [19].

4.2.2.3 Clustering

Another popular data mining method is Clustering, which is an unsupervised learning, meaning that there are not predefined classes [8, 20]. It is used mainly in cases where there is not adequate knowledge about the objects of large datasets, trying to group them based on similar characteristics [5]. The groups that will be formed are called clusters, and data points within the same cluster are characterized by higher similarity, while those assigned to different clusters have lower similarity [8]. The most common clustering techniques used in healthcare are Partitional, Hierarchical, and Density-Based Clustering [14].

Bertsimas, et al. attempted to predict healthcare costs by utilizing medical and cost data over three years for 800,000 insured people. For their experiment they first set a baseline method by using the healthcare cost of the last 12 months of the period of examination as the forecast of the overall healthcare cost in the result period. Then, they used classification trees and clustering. The clusters that were formed contained groups of people with similar cost characteristics and often similar medical characteristics. They concluded that knowledge discovery methods significantly improved performance compared to the baseline method, with the classification tree algorithm performing better on the lowest-cost buckets, while clustering doing better on the

higher-cost buckets. Moreover, clustering achieved better results when utilizing both cost and medical data [21].

Peng et al. applied clustering to understand the data and detect suspicious healthcare frauds from large amounts of real-life health insurance data. To reach their goal, they used two clustering techniques, SAS EM and CLUTO. After conducting their experiment, they concluded that the second method needs less time, while the clusters formed by the first method are more valuable for their purpose [22].

4.2.2.4 Association Rules

Association rules is one of the most important knowledge discovery techniques used to detect valuable links and recognize frequent patterns among data objects [23]. It plays a significant role in the healthcare sector as it helps in detecting relationships among diseases, health state and symptoms. Health insurance companies also use association approach to analyze data and detect fraud and abuse. The most common association algorithm is Apriori [14].

Patil et al. introduced a new approach to classify whether a patient is likely to suffer from diabetes or not [24]. For this purpose, they first used equal interval binning based on medical expert opinion, and then they used association to produce rules and understand the relationship among their data.

Kai et al. used association rule mining to develop a support system for clinical decisions. The goal of their project was to support healthcare assistants identify and offer appropriate lifestyle guidance [25].

4.3 Methodology and Results

In the previous literature review section, we specified some interesting case studies of data mining applications in the field of healthcare support. In this experimentation section, we present various data mining tools and algorithms, either predictive ones, like classification, or descriptive ones, such as clustering and association rule mining. That was intended to extract valuable knowledge, information and patterns with the aim to predict stroke incidents from medical records.

4.3.1 Methodology Outline

We used a dataset from Kaggle that contains stroke incident records to perform several experiments. Our main objective was to use data mining tools to extract useful knowledge from the dataset and develop a predictive tool that will predict whether a person will suffer from stroke.

For the experimentation, we used classification, clustering, and association rule mining. Decision Tree, Random Forest and Naïve Bayes were used for classification, k-means, k-medoids and DBSCAN for clustering and FP-Growth for association rule mining. During these experiments, it was necessary to perform some pre-processing to filter out invalid records or fix wrong values.

4.3.2 Experiments

4.3.2.1 Dataset

We utilized a dataset named “Stroke Prediction Dataset” which contains medical records about stroke incidents, retrieved from the open-source dataset provider Kaggle. In total, the dataset consists of 5110 records and 12 attributes of all attribute types, such as categorical, real, integer or binary. As mentioned earlier, the general purpose was to use data mining to develop a predictive tool to predict whether a person will suffer from stroke. Therefore, the class attribute, which is also provided by the dataset, must be of binary type and take two values, “suffered from stroke” and “not suffered from stroke”. As for the other attributes, integer ones are “id” and “age”, real ones are “average glucose level” and “body mass index (bmi)”, categorical ones are “gender”, “work type”, “residence type” and “smoking status” and binary ones are “hypertension”, “heart disease” and “ever married”. All experimentations with the different data mining techniques were performed with RapidMiner.

4.3.2.2 Classification

With regards to classification, our goal was to learn a classifier that correctly predicts cases of people who might suffer from stroke. Before proceeding to performing experiments, it was necessary to pre-process the data. First, the class attribute as well as the attributes “heart_disease” and “hypertension” were converted from integer to binominal, because according to the dataset description, they take values 0 or 1, meaning false or true.

Then, we set the label role to the class attribute and id role to the “id” attribute and handled the missing values either by changing any “N/A” and “Unknown” text to blank value to turn them into actual missing values and then with the “Replace Missing Values” operator, the missing values were replaced with the average of the attribute or filtered out in case the average could not be calculated. With regards to “bmi” attribute, it was of polynomial type, so in order to be converted to real, we applied the “Parse Numbers” operator. Finally, to balance the dataset, the Sample operator was used to filter the dataset up to 280 records per class, almost as much as the smallest set, to avoid learning a biased classifier. The pre-processing task can be seen in RapidMiner application in Fig. 4.3.

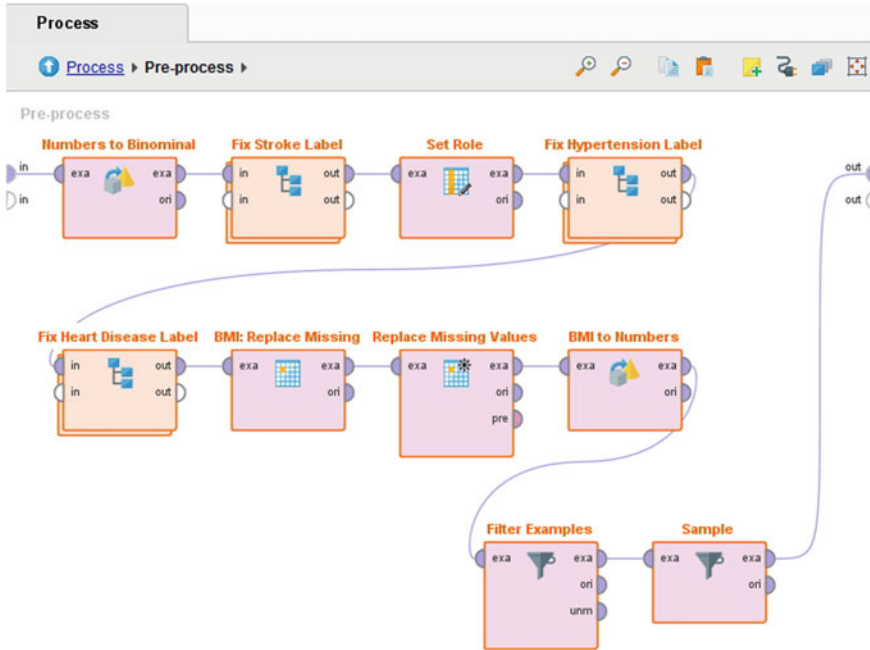


Fig. 4.3 Pre-processing process

Table 4.1 Decision Tree accuracy per criterion

Criterion	Accuracy (%)
Gain ratio	69.69
Information gain	70.75
Gini index	66.19

Then, the dataset was ready to be used for classification to train different classifiers and measure their prediction accuracy. The classification algorithms that were used were Decision Tree, Random Forest, and Naïve Bayes. For each one, it was important to perform a tenfold cross validation, with the use of “Cross Validation” operator, using accuracy as the main criterion of performance evaluation.

Decision Tree

Inside the “Cross Validation” operator, the “Decision Tree” was used with the default parameters except the criterion. We performed the classification for gain ratio, information gain and gini index criterion values. Table 4.1 presents the accuracy results for each criterion. As it seems, with the information gain criterion the learned classifier has the highest accuracy.

Table 4.2 Random Forest accuracy per criterion

Criterion	Accuracy (%)
Gain ratio	74.48
Information gain	74.26
Gini Index	75.10

Table 4.3 Classifier accuracy comparison

Classifier	Accuracy (with best criterion)
Decision Tree	70.75% (with information gain)
Random Forest	75.10% (with gini index)
Naïve Bayes	77.59%

Random Forest

Similarly, with the previous experiment, by swapping the “Decision Tree” operator with the “Random Forest” one, inside “Cross Validation”, we performed classification with random forest and used the same criterion values, to test and find which one produces the best results. This time the gini index criterion produces the classifier with the highest prediction accuracy, from what can be seen in Table 4.2.

Naïve Bayes

Finally, we swapped “Random Forest” with “Naïve Bayes” operator. This methodology does not have any parameters, like the criterion from the previous ones, to modify, test and get the best results. So, it always produces a classifier with accuracy of 77.59%.

Classifiers Comparison

The results of our experimentations showed that the Naïve Bayes produces the best classifier with overall accuracy of 77.59%, as shown in Table 4.3.

However, the accuracy by itself may not be enough to prove that a classifier is the best. For the classifier comparison to be more prestigious and since we are dealing with matters in the medical domain, we used the Receiver Operating Characteristic (ROC) curve, to find the most accurate classifier [26]. The ROC curves of the three classifiers shown in Fig. 4.4, validate that the Naïve Bayes classifier is indeed the most accurate.

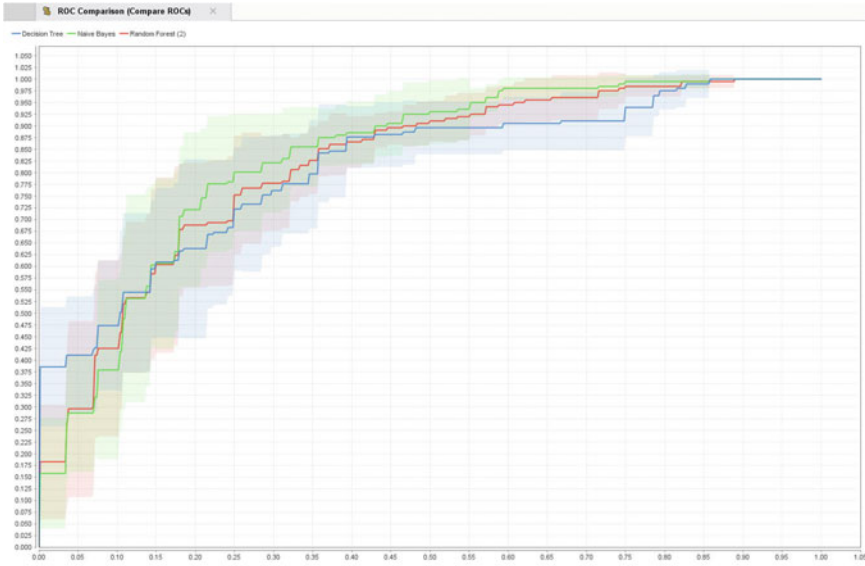


Fig. 4.4 ROC curve comparison

4.3.2.3 Clustering

Clustering is a data mining methodology which partitions data in different groups or clusters based on similarity criteria. That is, data in the same cluster bear some conceptual resemblance and those from different clusters are dissimilar [27]. For the clustering experiments, we used the same stroke incident dataset and the k-means, k-medoids and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms. Similarly with the classification experimentation, some pre-processing was also applied to the dataset before running the clustering algorithm, to select the appropriate attributes, clear missing values or outliers and transform the data. The attributes that were selected for all experiments, were “bmi”, “average glucose level” and age, since those are the only numerical ones, hence being interesting for clustering purposes.

k-Means

After pre-processing the data, the k-means algorithm was used for the initial experiment. Given that the dataset has two classes, “had stroke” and “not had stroke”, we set the parameter $k = 2$ for the cluster number, meaning that 2 clusters would be produced. After clustering, the first cluster had 4355 records, while the second one had 755, as shown in Table 4.4.

Table 4.4 K-means clustering

Clusters	Records
Cluster 0	4355
Cluster 1	755

Table 4.5 k-medoids clustering

Clusters	Euclidean distance	Manhattan distance
	Records	
Cluster 0	973	983
Cluster 1	4137	4127

Table 4.6 DBSCAN clustering

Epsilon parameters	Number of clusters	Number of noise records
1	0	5110 (all records)
2	2	5108
2.5	22	5017
3	38	4607
3.5	17	3533
4	9	2702
4.5	8	2104
5	5	1705
5.5	8	1368
6	4	1117

k-Medoids

The following experiment utilized k-medoids, which is similar to the k-means. For this experiment, we used the same cluster size as before, meaning that we set the parameter $k = 2$, but this time for the “Measure Types” parameter, we used two different types, Euclidean and Manhattan distance. The results of clustering can be seen in Table 4.5.

DBSCAN

For the third and final clustering experiment, we used the DBSCAN. In contrast with k-means and k-medoids, for this algorithm the initial cluster size parameter k cannot be set, and it accepts only one parameter, the epsilon parameter. We tried different values for the epsilon parameter, and it produced interesting results, where a lot of this values produced clusters with only a few records while considering most of the records as noise. The results of this experiment can be viewed in Table 4.6.

4.3.2.4 Association Rule Mining

As a final step, we used association rule mining [28]. With this technique, one can extract useful patterns, relations, and associations between records of large datasets [29]. For this purpose, we used Frequent Pattern (FP)-Growth and the “Create Association Rules” operator in RapidMiner, with various values for support and confidence, after applying once more some pre-processing to the dataset. Figure 4.5 shows the RapidMiner process that was developed for this experiment.

The results of an experimentation example can be seen here, with support level being at 60% and confidence at 70%. A total of 50 different rules were produced and the more interesting one are presented in Table 4.7.

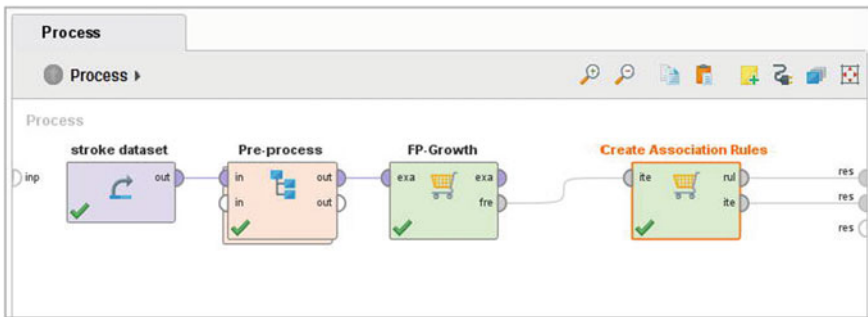


Fig. 4.5 Association rule mining process

Table 4.7 Association rules

Premises	Conclusion	Support (%)	Confidence (%)
No hypertension	No heart disease	83	95
Married	No heart disease	70	93
No heart disease	No hypertension	83	88
Married	No hypertension	65	85
No heart disease	Married	70	75
No hypertension	Married	74	88
No heart disease, No hypertension	Married	73	88

4.4 Conclusion

4.4.1 Discussion

Several experiments were carried out using a dataset that included stroke incidence records. The main goal was to employ data mining methods to acquire meaningful knowledge and predict if a patient would have a stroke. To that end, classification, clustering, and association rule mining were utilized. It was also necessitated to undertake some data pre-processing steps to filter out invalid or rectify incorrect values. Therefore, for classification, Decision Tree, Random Forest, and Nave Bayes were employed achieving a stroke prediction accuracy of 70.75%, 75.10% and 77.59%, respectively. For clustering we incorporated k-means, k-medoids and DBSCAN showcasing controversial results with the best ones forming 2–4 clusters of patients with strokes or not. Finally, for mining association rules we used FP-Growth retrieving a few strong rules with the top three associating “No hypertension” with “No heart disease” (Support 83% and Confidence 95%), “Married” with “No heart disease” (Support 70% and Confidence 93%) and “No heart disease” with “No hypertension” (Support 83% and Confidence 88%).

Even though data mining is relatively new, it has become a useful asset for any organization because of its ability to extract new knowledge from enormous databases fast. This is very important in healthcare, since we are dealing with records from patients. But also, there is the potential to extract useful knowledge and information by carefully evaluating and mining such data.

Data mining could be applied to many different healthcare domains, such as measuring the treatment effectiveness provided from medical stuff, helping with decision making with regards to healthcare management, detecting fraudulent cases in the healthcare insurance domain, finding out about the preferences of patients to provide better services, discovering new patterns for better infection control, improving patient care and ranking hospitals based on their ability to handle high-risk patients.

4.4.2 Issues and Challenges of Data Mining in Stroke Prediction and Healthcare

Performing experiments on patient stroke records to produce a classifier that predicts stroke incidents was challenging. The need for knowledge in healthcare, especially in the stroke domain, slows down experimentation, as it is required to study material and perform a targeted data mining analysis. Data mining needs specific attributes to be selected, based on relations between them and how they co-relate with the stroke incident.

Another challenge with stroke prediction process was the data format. The dataset has a lot of missing values, and those missing values are not just blank entries, but

they are represented by terms like “N/A” or “Unknown”. This can cause some analysis barriers in the application environment of RapidMiner. RapidMiner’s “Replace Missing Values” operator only works on actual blank values and when a record is filled with “N/A” or any kind of text, it is not considered blank. The solution is to first use the “Replace” operator and replace the “N/A”, “Unknown” and every word that implies missing values with the actual empty value. Then the “Replace Missing Values” operator can be used correctly.

In conventional data mining applications, data scientists are interested in discovering useful patterns in datasets and describing those patterns in simple terms. On the other hand, in the healthcare domain, data scientists may aim to identify irregular cases present in a small number of records, which do not follow the usual patterns. In contrast with conventional data mining, in most healthcare applications it is of critical importance to detail every piece of new-found knowledge as thoroughly as possible. Even minor mistakes could have fatal results on peoples’ lives [5].

The General Data Protection Regulation (GDPR) is another challenge. Health related data are sensitive and private and must not be accessed by anyone without sufficient authorization [30]. Additionally, for data mining to be able to produce accurate and useful results, a vast amount of different, real data from many different patients is required. Those two obstacles combined hinder data mining related applications and can be critical in any healthcare related data analysis [5].

One more obstacle after knowledge extraction is that healthcare specialists may not modify their working routines because of some computer-assisted acquired knowledge. Even if data mining results provide major improvements on healthcare professionals’ routines, they may be reluctant to change their practices [4].

Moreover, the actual quality of data poses a great challenge. Since data mining requires many records, such data are gathered from many different sources, organizations, or data warehouses. All these sources do not use the same schemas. When collecting non-uniform data, difficulties may arise, as the same type of data are stored differently [14]. It is also possible for a data scientist to deal with missing values or noisy data that do not provide useful knowledge results [31]. The pre-processing required to deal with all these issues slows down the entire process [5].

In addition to the previous challenges, there is also the question of trustworthiness of data mining models. These models are predictive, meaning that they produce an estimated conclusion of some input data, and the conclusion is by no means to be taken for granted. They should be used by medical staff as an extra tool that provides a second opinion. Therefore, predictive models should be used wisely for critical decisions on patients’ health, as the estimated decision produced by the model could potentially endanger lives [5].

4.4.3 Future Directions and Insights

As mentioned above, a critical challenge of data mining is the quality of the data in healthcare. That is the reason why it is of importance for healthcare to find a more

efficient way to handle its data. With regards to (i) how those data are collected, (ii) stored in databases or data warehouses, (iii) pre-processed and finally (iv) what mining techniques can be used on them. These points may be considered as future work of this study.

To achieve that, a solution would be the standardization on how to handle healthcare data, as well as the implementation of a secure and efficient infrastructure that supports sharing of these private, personal data.

It is also important to mention that healthcare data could be either quantitative or qualitative. Quantitative are data that can be counted, measured, or expressed in a numerical format, while qualitative are more conceptual and descriptive. Normally, qualitative data can be considered notes, documents, any kind of text in general and even media files, like audio, video, and images. In the healthcare domain, qualitative data can be a doctor's notes from his/her patients, some clinical records or even diagnostic or x-ray images. Those data should also be included in a data mining process along with quantitative ones, in order to achieve even better results.

Finally, in data mining, it is important to perform some pre-processing and filter the available attributes. Outliers and noise should be filtered out, as they negatively impact on both accuracy and speed. Data mining provides statistical techniques that can detect such attributes and there are feature selection algorithms that can identify the most suitable attributes.

Acknowledgements We would like to express our gratitude to the anonymous reviewers who provided critical feedback during the preparation of this manuscript. Their remarks and recommendations significantly improved the quality of this work.

References

1. Koukaras, P., Tjortjis, C., Rousidis, D.: Social Media Types: introducing a data driven taxonomy. *Computing* **102**(1), 295–340 (2020). <https://doi.org/10.1007/s00607-019-00739-y>
2. Baitharu, T.R., Pani, S.K.: Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset. *Procedia Computer Science* **85**, 862–870 (2016). <https://doi.org/10.1016/j.procs.2016.05.276>
3. Tjortjis, C., Saraee, M., Theodoulidis, B., Keane, J.A.: Using T3, an Improved Decision Tree Classifier, for Mining Stroke-related Medical Data. *Methods Inf. Med.* **46**(05), 523–529 (2007). <https://doi.org/10.1160/ME0317>
4. Koh HC, Tan G. "Data mining applications in healthcare", *J Healthc Inf Manag*, 2005 Spring;19(2):64–72. PMID: 15869215.
5. M. H. Tekieh and B. Raahemi, "Importance of data mining in healthcare: A survey," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM 2015, Aug. 2015, pp. 1057–1062. doi: <https://doi.org/10.1145/2808797.2809367>.
6. Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan, I., Keane, J.: Comparing Data Mining Methods with Logistic Regression in Childhood Obesity Prediction. *Inf. Syst. Front.* **11**(4), 449–460 (2009). <https://doi.org/10.1007/s10796-009-9157-0>

7. Glover, S., Rivers, P.A., Asoh, D.A., Piper, C.N., Murph, K.: Data mining for health executive decision support: An imperative with a daunting future! *Health Serv. Manage. Res.* **23**(1), 42–46 (2010). <https://doi.org/10.1258/hsmr.2009.009029>
8. Tomar, D., Agarwal, S.: A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology* **5**(5), 241–266 (2013). <https://doi.org/10.14257/ijbsbt.2013.5.5.25>
9. Obenshain, M.K.: Application of Data Mining Techniques to Healthcare Data. *Infect. Control Hosp. Epidemiol.* **25**(8), 690–695 (2004). <https://doi.org/10.1086/502460>
10. T. Chatzinikolaou, E. Vogiatzi, A. Kousis, and C. Tjortjis, “Smart Healthcare Support Using Data Mining and Machine Learning,” in *EAI/Springer Innovations in Communication and Computing Book: “IoT and WSN based SmartCities: A Machine Learning Perspective,”* 2022.
11. P. Koukaras, D. Rousidis and C. Tjortjis, “Forecasting and Prevention Mechanisms Using Social Media in Health Care”, in Maglogiannis I., Brahnma S., Jain L. (eds) *Advanced Computational Intelligence in Healthcare-7. Studies in Computational Intelligence*, vol 891, March 2020, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-61114-2_8.
12. S. El-Sappagh, S. El-Masri, M. Elmogy, S. H. El-Sappagh, and A. M. Riad, “Data Mining and Knowledge Discovery: Applications, Techniques, Challenges and Process Models in Healthcare,” *International Journal of Engineering Research and Applications (IJERA)*, vol. 3, no. 3, pp. 900–906, May 2013, [Online]. Available: <https://www.researchgate.net/publication/250612388>.
13. P. Koukaras, C. Berberidis and C. Tjortjis, “A Semi-supervised Learning Approach for Complex Information Networks”, in Hemanth J., Bestak R., Chen J.IZ. (eds) *Intelligent Data Communication Technologies and Internet of Things. Lecture Notes on Data Engineering and Communications Technologies*, vol 57, February 2021, Springer, Singapore. https://doi.org/10.1007/978-981-15-9509-7_1.
14. Ahmad, P., Qamar, S., Qasim, S., Rizvi, A.: Techniques of Data Mining In Healthcare: A Review. *International Journal of Computer Applications* **120**(15), 38–50 (2015). <https://doi.org/10.5120/21307-4126>
15. Tzirakis, P., Tjortjis, C.: T3C: improving a decision tree classification algorithm’s interval splits on continuous attributes. *Adv. Data Anal. Classif.* **11**(2), 353–370 (2017). <https://doi.org/10.1007/s11634-016-0246-x>
16. Das, R., Turkoglu, I., Sengur, A.: Effective diagnosis of heart disease through neural networks ensembles. *Expert Syst. Appl.* **36**(4), 7675–7680 (2009). <https://doi.org/10.1016/j.eswa.2008.09.013>
17. D. I. Curiac, G. Vasile, O. Baniias, C. Volosencu, and A. Albu, “Bayesian network model for diagnosis of psychiatric diseases,” in *Proceedings of the International Conference on Information Technology Interfaces, ITI, 2009*, pp. 61–66. doi: <https://doi.org/10.1109/ITI.2009.5196055>.
18. Divya, D., Agarwal, S.: Weighted support vector regression approach for remote healthcare monitoring. *International Conference on Recent Trends in Information Technology, ICRTIT 2011*, 969–974 (2011). <https://doi.org/10.1109/ICRTIT.2011.5972437>
19. J. Alapont, A. Bella-Sanjuán, C. Ferri, J. Hernández-Orallo, J. D. Llopis-Llopis, and M. J. Ramírez-Quintana, “Specialised Tools for Automating Data Mining for Hospital Management,” in *Proc. First East European Conference on Health Care Modelling and Computation*, Aug. 2005, pp. 7–19.
20. Kanellopoulos, Y., Antonellis, P., Tjortjis, C., Makris, C., Tsirakis, N.: k-Attractors: A Partitional Clustering Algorithm for Numeric Data Analysis. *Appl. Artif. Intell.* **25**(2), 97–115 (2011). <https://doi.org/10.1080/08839514.2011.534590>
21. Bertsimas, D., et al.: Algorithmic prediction of health-care costs. *Oper. Res.* **56**(6), 1382–1392 (2008). <https://doi.org/10.1287/opre.1080.0619>
22. Y. Peng, G. Kou, A. Sabatka, Z. Chen, D. Khazanchi, and Y. Shi, “Application of Clustering Methods to Health Insurance Fraud Detection,” in *2006 International Conference on Service Systems and Service Management*, Oct. 2006, pp. 116–120. doi: <https://doi.org/10.1109/ICSSM.2006.320598>.

23. S. M. Ghafari and C. Tjortjis, "A survey on association rules mining using heuristics," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, Jul. 2019, doi: <https://doi.org/10.1002/widm.1307>.
24. B. M. Patil, R. C. Joshi, and D. Toshniwal, "Association rule for classification of type -2 diabetic patients," in *ICMLC 2010 - The 2nd International Conference on Machine Learning and Computing*, 2010, pp. 330–334. doi: <https://doi.org/10.1109/ICMLC.2010.67>.
25. E. Kai et al., "Empowering the Healthcare Worker Using the Portable Health Clinic," 2014 IEEE 28th International Conference on Advanced Information Networking and Applications, 2014, pp. 759–764, doi: <https://doi.org/10.1109/AINA.2014.108>.
26. Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., de Mendonça, A.: Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Res Notes*. **17**(4), 299 (2011). <https://doi.org/10.1186/1756-0500-4-299>. PMID:21849043;PMCID:PMC3180705
27. P. Berkhin, "A Survey of Clustering Data Mining Techniques," in *Grouping Multidimensional Data*, Berlin/Heidelberg: Springer-Verlag, pp. 25–71. doi: https://doi.org/10.1007/3-540-28349-8_2.
28. Kotsiantis, S., Kanellopoulos, D.: Association Rules Mining: A Recent Overview. *GESTS International Transactions on Computer Science and Engineering* **32**(1), 71–82 (2006)
29. Y. Liu, Institute of Electrical and Electronics Engineers, and IEEE Circuits and Systems Society, ICNC-FSKD 2017: 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery : Guilin, Guangxi, China, 29–31 July, 2017.
30. Abouelmehdi, K., Beni-Hessane, A., Khaloufi, H.: Big healthcare data: preserving security and privacy. *Journal of Big Data* **5**(1), 1 (2018). <https://doi.org/10.1186/s40537-017-0110-7>
31. B. Milovic, "Prediction and decision making in Health Care using Data Mining," *International Journal of Public Health Science (IJPHS)*, vol. 1, no. 2, Dec. 2012, doi: <https://doi.org/10.11591/ijphs.v1i2.1380>.