Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Top data mining tools for the healthcare industry

Judith Santos-Pereira [a], Le Gruenwald [b], Jorge Bernardino [a,c,*]

[a] Polytechnic of Coimbra, ISEC, Rua Pedro Nunes, Quinta da Nora, 3030-190 Coimbra, Portugal
[b] University of Oklahoma, School of Computer Science, 110 W. Boyd St., Room 150 DEH, 73019 Norman, Oklahoma, USA
[c] Centre of Informatics and Systems, University of Coimbra, Pinhal de Marrocos, 3030-290 Coimbra, Portugal

ARTICLE INFO

Article history:
Received 3 February 2021
Revised 3 May 2021
Accepted 1 June 2021
Available online xxxx

Keywords:
Data mining
Healthcare
Open-source data mining tools

ABSTRACT

The healthcare industry has become increasingly challenging, requiring retrieval of knowledge from large amounts of complex data to find the best treatments. Several works have suggested the use of Data Mining tools to overcome the challenges; however, none of them has suggested the best tool to do so. To fill this gap, this paper presents a survey of popular open-source data mining tools in which data mining tool selection criteria based on healthcare application requirements is proposed and the best ones using the proposed selection criteria are identified. The following popular open-source data mining tools are assessed: KNIME, R, RapidMiner, Scikit-learn, and Spark. The study shows that KNIME and RapidMiner provide the largest coverage of healthcare data mining requirements.

Contents

* Corresponding author at: Polytechnic of Coimbra, ISEC, Rua Pedro Nunes, Quinta da Nora, 3030-190 Coimbra, Portugal.
E-mail addresses: santosj@hotmail.ca (J. Santos-Pereira), ggruenwald@ou.edu (L. Gruenwald), jorge@isec.pt (J. Bernardino).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2021.06.002
1319-1578/

Please cite this article as: J. Santos-Pereira, L. Gruenwald and J. Bernardino, Top data mining tools for the healthcare industry, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2021.06.002

1. Introduction

The healthcare industry daily generates large amounts of complex data from multiple data sources, such as electronic patient records, medical reports, hospital devices, and billing systems (Strang and Sun, 2020). These huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. In fact, several studies have suggested using an advanced data analysis technique called data mining to overcome these data challenges (Strang and Sun, 2020)(Gonzalez et al., 2016). Data mining is the process of discovering interesting patterns from massive amounts of data (Han et al., 2012), where standard statistical exploratory data analysis procedures (traditional statistics) could not discover useful insights (Hand et al., 2000). At present, in the healthcare industry, traditional statistical approaches are viewed as the primary data analysis technique and data mining as the secondary technique due to the limited exposure to the data mining area by medical researchers and healthcare practitioners (Reddy and Aggarwal, 2015). While the groundwork of both techniques is mathematics, data mining extends it with other subjects such as machine learning, database systems and visualization which brings important gains over the traditional statistics techniques (Tekieh and Raahemi, 2015).

Data mining tools are software packages that have the ability to analyze large amounts of data to discover meaningful patterns and predict outcomes (Tan et al., 2006). Some data mining tools have data cleaning features that automate the cleaning process of data, as well as the ability to extract valuable information from different data types such as numeric, text, document, image, graph, speech, audio, and video. This type of tools allows the extraction of valuable information from data, also known as Knowledge Discovery in Databases (KDD) (Almeida and Bernardino, 2016). Furthermore, data mining tools have the ability to perform inductive analysis. This ability is fundamental in the case where researchers are trying to understand a health condition that is unknown; due to not knowing very well a condition, researchers have the difficulty to create a hypothesis to prove or reject through data analysis. Moreover, these tools have the ability to consider the whole dataset for analysis which can bring new insights to research. In this paper, we choose to analyze open source data mining tools rather than proprietary ones due to their free acquisition cost, which is an important aspect for healthcare researchers who often work for non-profit organizations or projects with limited budgets.

Nowadays, there is a wide range of open-source data mining tools and usually their vendors do not clearly specify their application domains adequately, leaving users lost in choosing tools for their applications. Hence, the aim of this paper is to propose data mining tool selection criteria and present a survey on popular open-source data mining tools that have been suggested for the healthcare industry (Sharma et al., 2016)(Gui et al., 2016), but not assessed nor compared with their domain requirements as we will do in this work.

In this paper, we describe the popular open source data mining tools (Poll, 2019) (Gartner, 2019) – KNIME (KNIME, 2017), R (RProject, 2021); RapidMiner (RapidMiner, 2017); Scikit-learn (Scikit-Learn, 2017) and Spark (Spark, 2021) - comparing them using the proposed selection criteria to guide healthcare industry users. The data mining tools were selected based on their popularity from the KDnuggets annual software poll (Poll, 2019), as well as their appropriateness to the healthcare domain presented in the Gartnerś Magic Quadrant Report for Data Science and Machine Learning Platforms (Gartner, 2019).

To the best of our knowledge, this is the first work that has both conducted an open-source data mining tool survey for the healthcare industry and proposed data mining tool selection criteria for this domain.

The main contributions of this work are the following:

- Contributing to the healthcare industry that needs to know the data mining methods commonly applied in healthcare (Section 2) and its data domain requirements by identifying the data characteristics (Section 3).
- Guiding healthcare data analysts in choosing their open-source data mining tools by proposing data mining tool selection criteria based on the healthcare data domain requirements (Section 5) and comparing a set of open-source data mining tools using the proposed criteria (Section 7).
- Helping data miners be up-to-date on the trends of this challenging field by disclosing the most popular data mining tools on the market (Sections 6) as well as giving a summary of the related surveys (Section 8).

The remainder of this paper is structured as follows. Section 2 describes the most common data mining methods used in the healthcare industry. Section 3 describes the data characteristics of healthcare applications. Section 4 identifies the critical capabilities that a data mining tool must have to perform healthcare data analysis. Section 5 presents the proposed data mining tool selection criteria. Section 6 presents the selected open-source data mining tools. Section 7 compares the selected tools using the proposed selection criteria and suggests the best tool for healthcare applications. Section 8 presents the related work on data mining tools. Finally, Section 9 provides conclusions and future work.

2. Data mining methods commonly used in healthcare

In order to extract knowledge from big data, a healthcare system requires unconventional and mature data storage, management, analysis, and data mining tools (Pramanik et al., 2020).

Data mining provides the methodology and technology to transform massive amounts of data into useful information for decision making (Dash et al., 2019). Since data mining tools do not all support the same data mining methods, it is important to identify the ones that are most commonly used in the healthcare industry (Tekieh and Raahemi, 2015) to guide us in the selection of the most suitable data mining tools. Therefore, the identified data mining methods are also part of the domain requirements covered in our proposed data mining tool selection criteria presented in Section 5. In this section, we describe the identified methods and their applications to support their selection.

2.1. Classification

Classification is a data analysis method constructing models that predict categorical labels (target attributes) (Han et al., 2012). This method is used when the data is required to be classified into different groups based on a target attribute (Tekieh and Raahemi, 2015), and/or predict the probability of a target label outcome based on historical records. This method has been used in various healthcare applications: a classification method was applied to better identify if a patient has dementia based on his/her neuropsychological test in (Maroco et al., 2011). In (Elhoseny et al., 2018), the Support Vector Machine and Artificial Neural Network algorithms were used to find correlations between a specific intestinal microbiota and the presence or absence of diabetes in order to predict metabolic diseases like diabetes. In another work, the Support Vector Machine algorithm was also used along with a computational method that optimizes it, named Particle Swarm Optimization, to predict seminal quality (Sahoo and Kumar, 2014). In (Mirroshandel et al., 2016), the KStar algorithm was used to predict the outcome of individual sperm implantation on humans in order to increase the implantation rate for intracytoplasmic. In (Kourou et al., 2015), a survey of works on data mining applications in the cancer prognosis and prediction field was presented. It turned out that all the presented works had applied algorithms that are classifiers. Just to name a few, Decision tree algorithms were used to predict breast cancer survival (Delen et al., 2005) and Bayesian Network to predict the recurrence of oral cancer considering several data types (clinical imaging and genomic data from tissue and blood) (Exarchos et al., 2012).

2.2. Clustering

Clustering is the process of partitioning a set of data objects (or observations) into subsets (Han et al., 2012). This technique is used when we do not have much information about the different types of data objects involved in a population. As it is an unsupervised learning method, it tries to find the cluster of data objects that are similar to each other without considering any specific target label (Tekieh and Raahemi, 2015). Since clustering is a method specially used in the *descriptive* analysis stage, several works have applied clustering algorithms to categorize the handled data prior to classification. In (Sharma et al., 2016) a survey was presented on medical publications that have used *Classification* and *Clustering* methods where the following works were pinpointed for the *Clustering* method: the K-Means clustering algorithm was used to contribute to the diagnose of heart disease patients (Shouman et al., 2012) and to categorize colon tumors (Kumar and Wasan, 2010). Clustering methods were also used to categorize proteins into functional groups (Xu et al., 2012), to predict the likelihood of diseases (Paul and Hoque, 2010) and to detect disease-specific clusters within medical image data (Bruse et al., 2017).

2.3. Association

Association is the process of finding association rules between attributes. This method is used when the relationship of attributes in a dataset needs to be identified (Tekieh and Raahemi, 2015). We can apply this method to, for instance, see if there is an association between a high blood pressure condition and a salt eating habit, and if so, build an association rule from it. For example, the popular APRIORI association rule mining algorithm was used to find associations between clinical data from diabetic patients (Stilou et al., 2001). Other association rule mining algorithms were proposed to find associations between time, place and patienťs infections on public health surveillance data (Brossette et al., 1998); between clinical data and therapeutic treatments (Ting et al., 2010); between medical data and rhinitis conditions (Yang et al., 2016); and between patienťs data for coronary heart disease diagnosis (Orphanou et al., 2016).

2.4. Outlier detection

Outlier detection is the process of identifying attributes that are not normal or outcomes that are unusual. This method is often used to find discrepancies in data with the aim of cleaning the data or detecting abnormal values presented in medical databases, such as the works done in (Kumar et al., 2008) and (Bellaachia and Bari, 2012).

In this section, we were able to verify that all identified data mining methods have been used across various works in the healthcare industry which emphasize not only the applicability of data mining in the healthcare research field, but also the need to seek for a data mining tool that covers these data mining methods. Next, we will present the specific data characteristics of healthcare applications.

3. Data characteristics of healthcare applications

Data characteristics were obtained through our analysis of several works carried out in the healthcare domain that depicts its challenges (Tekieh and Raahemi, 2015) and points out its data characteristics (Strang and Sun, 2020)(Raghupathi and Raghupathi, 2014)(Tortorella et al., 2021) (Smys, 2019)(Saeed et al., 2018). Patient health data can be securely captured using health monitoring systems. A variety of sensors and complex algorithms are used to analyze the data and then share it through Internet of Things (IoT) solutions. The medical professionals can then make appropriate health recommendations also remotely. The critical challenges of healthcare services are big patients' data, big resources, and big applications by retrieving and storing those processes in the shortest possible time (Elhoseny et al., 2018). In the next subsections, we describe the data characteristics that we will consider as part of the healthcare domain requirements (Wang et al., 2018) in our proposed data mining tool selection criteria presented in Section 5.

3.1. Large amounts of data

Advances in data generation and collection technologies led to an enormous data growth in healthcare databases. Patienťs management software, medical equipment, clinical analyses and medical imaging software are only a few examples of these technologies. With all these different healthcare software products daily functioning, researchers are facing with an unmanageable scale of data (volume). Therefore, having a data mining tool that can handle large amounts of data is critical to data analysis.

3.2. Cloud data

The integration of Cloud Computing and IoT provides a new storage, processing, scalability and networking capabilities which are so far limited in the IoT due to its characteristics in the area of healthcare (Sun et al., 2017)(Li et al., 2017)(Muhammad et al., 2015)(Muhammad et al., 2016)(Ray, 2018). Data can be stored in logical pools for, for instance, real time access. In the healthcare industry, gathering data in the cloud is not very common currently. However, some works are suggesting it to centralize patient's data (Newhouse, 2016) - one of the healthcare challenges. Therefore, we believe that this data characteristic can be a must to seek for in a data mining tool.

3.3. Streaming data

Stream data flows in and out of a computer system continuously and with varying update rates. They are often generated by real-time surveillance systems, remote sensors or other dynamic environments (Makhabel, 2014). The specificity of this type of data is that it needs to be processed and analyzed in real-time. Biosensor data streaming and analytics is a key component of smart e-healthcare. However, existing IoT ecosystem is unable to materialize the real-time bio-sensor data streaming and analytics within resource constrained environments (Pratim Ray et al., 2020). In the healthcare industry, various applications are being proposed with this streaming data characteristic: systems for patient's blood pressure and temperature tracking (Aziz et al., 2016), wearable diagnostic devices to combat childreńs pneumonia (Mala et al., 2016), prevent drug abuse (Wang et al., 2017) and assess cognitive impairments (Alam et al., 2016) are some examples. Therefore, a data mining tool with streaming data analysis abilities is fundamental.

3.4. Multiple data sources

Modern healthcare systems are still struggling to provide patient-centered healthcare instead of clinical-centered healthcare as it is essential to implement major aspects of modern healthcare such as continuity of care, evidence-based treatment, and more importantly, preventing medical errors (Song, 2016). Therefore, one of the most common situations in the healthcare industry is to have clinical data being handled from several software products, scattered in different places (i.e. distributed data storage) and owned by different healthcare personnel such as physicians and clinical staff (Wan, 2016). This data characteristic raises the requirement that a data mining tool import and integrate data from various types of data sources.

3.5. Different data types

With the widespread use of medical information systems, information acquisition now expands to different data types such as the following (Wang et al., 2018)(Primova et al., 2020)(Kaur and Rani, 2015):

- **Numeric:** data that contains only numbers (e.g. age, weight).
- **Text:** alphanumeric content (e.g. car plate number)
- **Document:** unstructured free-text gathered in files such as Microsoft Word, Acrobat PDF documents or even simple text files. Document mining, so called text mining, has been used in the healthcare industry to, for example, extract information of protein–protein interactions within several documents (Zhou et al., 2006).
- **Image:** In medical procedures, imaging is increasingly employed as a preferred diagnostic tool. It can be a SPECT scan, an MRI scan or even a collection of ECG signals. For example, these medical images have already been used for tumor classification in digital mammography (Antonie et al., 2001).
- **Graph:** In a healthcare example application, a graph may represent a chemical compound where the nodes correspond to atoms and the links correspond to bonds between atoms (Aridhi and Mephu Nguifo, 2016). The aim of this type of study may be to mine the bonds between the atoms to better understand a chemical structure. Therefore, some authors also refer to this type of data as links.
- **Audio:** In medical procedures, audio data can be any kind of audio signals (e.g. cardiac beating).
- **Speech:** Spoken recorded words (e.g. patient´s speech).
- **Video:** Audio-visual content that can come from, for example, a patienťs surgery (Wang et al., 2018).

The need to use large, and at the same time still constantly growing, amounts of information in solving diagnostic, therapeutic, statistical, managerial and other tasks, determines today the creation of information systems in medical institutions (Xu et al., 2012). Since healthcare data can be more than just numbers and alphanumeric contents, the need for data mining tools that can perform knowledge discovery in their domain-related data types is crucial (Elhoseny et al., 2018).

3.6. Dirty data

In the healthcare industry, data is usually collected through Electronic Medical Records (EMR). The data collected via these systems are mainly gathered for analytical purposes and contain many issues – incorrectness, missing data, miscoding, incongruences, and incompleteness (Tekieh and Raahemi, 2015). The main reasons for these data characteristics are that most data registered in EMR are observational and not experimental. Therefore, they might not represent all cases involved in, for instance, a patient disease and then cause misleading data registrations (Tekieh and Raahemi, 2015).

Dirty data generation is also caused by distributed data storage. For instance, each healthcare service can have different names or coding for the same attribute, and at data integration time, we would be confronted with an incongruent data set.

We consider the incompleteness of data as another data characteristic under the umbrella of the dirty data characteristic because, as with the Incorrectness, Missing data, Miscoding and Incongruence characteristics, they all can be in part automatically corrected with data preprocessing features or already implemented functions (built-in functions) that data mining tools have.

Having a data mining tool that can automatically clean the data (i.e. prepare the data to mine) is a huge benefit for healthcare researchers; otherwise, they would have a very time-consuming task, prone to errors, by doing it manually. Therefore, having a data mining tool that can clean and transform the data is important, especially if it has data preprocessing features to ease the task.

3.7. Complex data

Researchers may handle data that they do not fully understand (i.e. complex data) from the scientific point of view. This makes their analysis task harder or impossible when it comes to apply hypothetical-deductive analysis to unravel health conditions as it is done with traditional statistics. Thus, having a tool that performs inductive analysis like data mining tools is useful.

Another aspect that contributes to data complexity is that data is gathered by different providers and in different ways, which makes it difficult to understand how and why the data is being gathered in order to analyze it correctly (Wan, 2016). Some data mining tools have data exploration and visualization capabilities that can support the process of better understanding the data to mine. Thus, having a data mining tool with such capacity is essential to mine complex data.

By considering the data mining gains over the traditional statistical techniques together with the healthcare data characteristics described above, we can see why data mining is being used in the healthcare domain. As the amount of collected health data is growing significantly every day, it is believed that a strong analysis tool that is capable of handling and analyzing large health data is essential.

Analyzing the health datasets gathered by electronic health record (EHR) systems, insurance claims, health surveys, and other sources, using data mining techniques is very complex and is faced with very specific challenges, including data quality and privacy issues (Tekieh and Raahemi, 2015). Therefore, the next section presents our investigation on critical capabilities that a data mining tool must have for healthcare applications.

4. Critical capabilities of data mining tools for healthcare

Generally, software selection lies in the selection of features for the application's needs. Hence, we will describe the critical features/capabilities that a data mining tool must have to mine data in the healthcare domain. The identified critical features/capabilities are based on the ones used by the IT consultant Gartner for its annual report on advanced analytics platforms (Linden et al., 2016) and data science platforms (Linden et al., 2017) as well as its definitions. These capabilities are selected taking also in consideration data mining methods commonly used in healthcare (Section 2) and data characteristics of healthcare applications (Section 3). At the end of each capability, we identify its related domain requirements.

4.1. Performance and scalability

Good performance and scalability reduce the time taken to load the data, as well as to create, validate and deploy the models. As data volume and complexity grow and the demand for faster insights rises, these capacities become important, especially in the healthcare domain where the following domain requirements have to be handled: large amounts of data, cloud data and streaming data (described in Sections 3.1–3.3, respectively).

4.2. Data access

This critical capability addresses the ability of a tool to access and integrate data from various sources and of different types (numeric, text, image etc.). Hence, the data mining tool will have to be able to handle the following domain requirements: multiple data sources and different data types (described in Sections 3.4 and 3.5, respectively).

4.3. Data preparation

The data preparation capability provides the ability to clean, transform and filter data in order to prepare it for modelling. This feature also enables the tool to perform basic descriptive statistics and pattern detection with descriptive data mining methods to support the data preparation. Since healthcare applications contain a lot of dirty data (one of the domain requirements described in Section 3.6), the selected data mining tool will have to cover this ability.

4.4. Data exploration and visualization

This capability allows a range of exploratory steps, including interactive visualizations, to support data mining methodology. The main domain requirement that makes this critical is the complex data (described in Section 3.7) that the healthcare sector generates.

4.5. Advanced modelling

This capability provides the ability to create data mining models that anticipate future behavior, estimate unknown outcomes or study behaviors. These models are created with a set of data mining methods. Therefore, the related domain requirements that make this capability critical are the data mining methods mostly applied in the healthcare industry which are: Classification, Clustering, Association and Outlier (described in Section 2). Since data modelling requires the selection of the best built model, the automation of this process is also considered as a domain requirement since it will ease the search for the best model from a set of built candidates.

4.6. User experience

This capability is provided through the ease of use of a tool, the type of interface that a tool has (e.g. graphical (GUI), a console (CLI) or a programmeŕs interface (IDE)), the skill level required to use it (e.g. programming languages), as well as the support provided by documentation and guidance of community support.

Some data mining tools have the Visual Composition Framework (VCF) feature that enables the construction of advanced analytic models without coding. Since there are healthcare researchers who do not know how to program, this feature is then essential for their applications and thus, promote a good user experience. Therefore, VCF is placed as a domain requirement under the good user experience.

Furthermore, some data mining tools can also facilitate all kinds of collaborations across all modelling steps with team members that are at different locations. Since some healthcare data mining projects require collaboration due to their complexity, this feature is then important to be considered in a tool since using a third-party application to do such service would be difficult for the tool to be used. Therefore, collaboration is also placed as a domain requirement under the user experience.

There are several types of data mining tools: data mining tools that can on their own perform all data mining methodology steps, so called end-to-end analytic tools; tools that are libraries that perform data mining tasks (machine learning packages); tools for statistics and computing; and tools that preprocess and mine large-scale data. Since each tool type is adequate to a specific project context (e.g. project collaborators with advanced technical skills can go for machine learning packages), Tool Type is also considered as a domain requirement under the user experience criterion.

In order to select the most suitable open-source data mining tools for healthcare applications, we propose the data mining tool selection criteria based on the above identified critical capabilities and compare the popular open-source data mining tools using these criteria. In the next section, we present our proposed data mining tool selection criteria.

5. Data mining tool selection criteria

In (Mikut and Reischl, 2011), the authors stated that researchers are mainly interested in data mining tools with well-proven domain related data mining methods, a graphical user interface (GUI) and interfaces to domain-related data formats or databases (Mikut and Reischl, 2011). Therefore, to build the proposed data mining tool selection criteria presented in Table 1, we had to identify the healthcare data requirements (the most common data mining methods (Section 2) and data characteristics (Section 3)) and relate them with the critical capabilities (Section 4) that a data mining tool should have.

The identified critical capabilities not only cover the good user experience criteria suggested by Ralf Mikut et al. (Alam et al., 2016) through their GUI requirement, but also include all the capabilities needed to perform data mining for healthcare applications.

Table 1
Data Mining Tool Selection Criteria.

Critical Capabilities	Domain Requirements
Performance and Scalability	*Large Amounts of Data*
	Cloud Data
	Streaming Data
Data Access	*Multiple Data Sources*
	Different Data Types
Data Preparation	*Dirty Data*
Data Exploration and Visualization	*Complex Data*
Advanced Modelling	*Classification*
	Clustering
	Association
	Outlier
	Automation
User Experience	*Programming Language*
	Operating System
	Interface
	Visual Composition Framework
	Collaboration
	Ease of Use
	Tool Type
	Community Support

The identified critical capabilities were based on the ones suggested by the IT Consultant Gartner (Linden et al., 2016).

The proposed selection criteria presented in Table 1 can be seen as a check list of principal features (critical capabilities presented in Section 4) that must be looked for in any data mining tool at the selection time in order to select the most suitable tool to mine healthcare data. Furthermore, the proposed selection criteria can also be used for tool comparisons since they will help identify which data mining tool covers most of the healthcare domain requirements as presented in Section 7.

6. Open-Source data mining tools

In this section, we describe the popular open-source data mining tools: KNIME, (2017), R (RProject, 2021); RapidMiner (RapidMiner, 2017); Scikit-learn (Scikit-Learn, 2017), and Spark (Spark, 2021). This selection is based on softwarés popularity disclosed in the KDnuggets annual software poll (Poll, 2019), the softwarés ability to execute presented in the Gartnerś 2019 Magic Quadrant for Data Science and Machine Learning Platforms report (Gartner, 2019), as well as the software's suitability for the healthcare domain suggested in the survey on Medical data mining applications (Sharma et al., 2016). While R is a statistical programming language, we also include it as a part of the open-source data mining tools for our study as it has many packages that support implementation of data mining tasks.

6.1. Knime

KNIME(2017) is a Java-based end-to-end analytic tool that integrates, transforms, analyzes and deploys data. It was developed by a team of developers from a Silicon Valley software company specialized in pharmaceutical applications and has been used in pharmaceutical research, among other fields. Some authors state that it is a tool with powerful capabilities in pre-processing, cleansing, modeling, analysis, and mining tasks for KDD (Almeida et al., 2016), and IT consultants such as Gartner consider KNIME as one of the leading solutions (Linden et al., 2017).

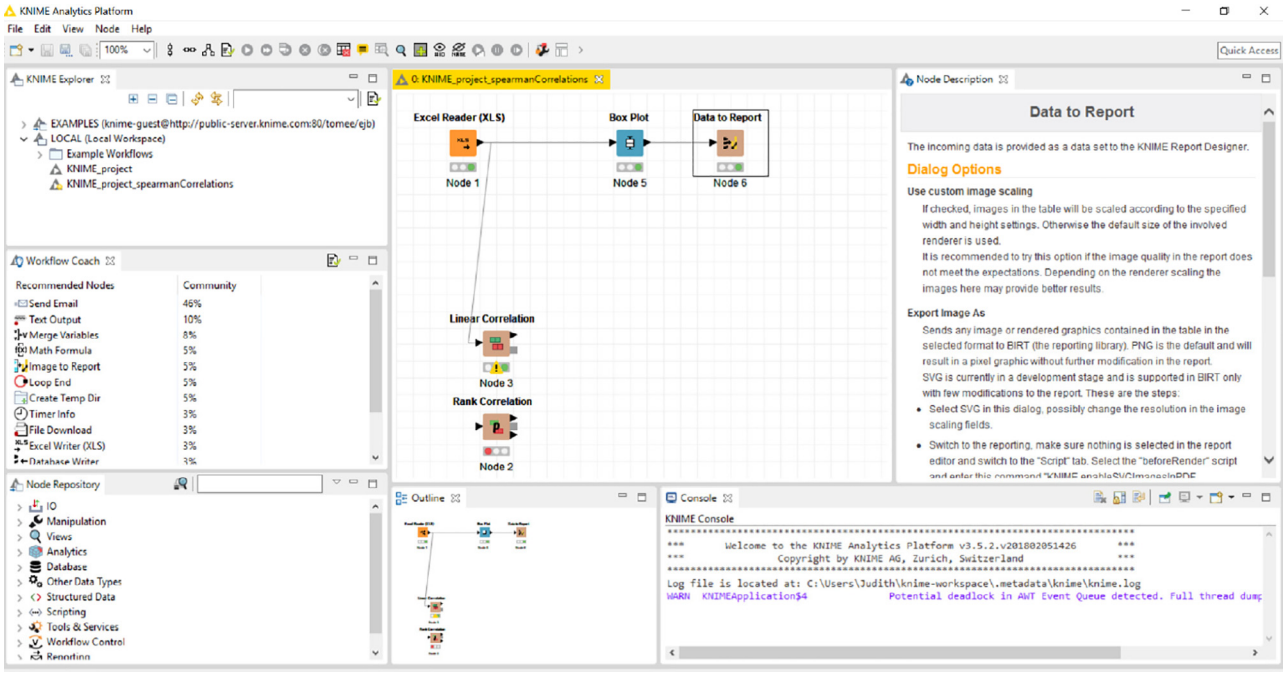

Fig. 1. KNIME Interface.

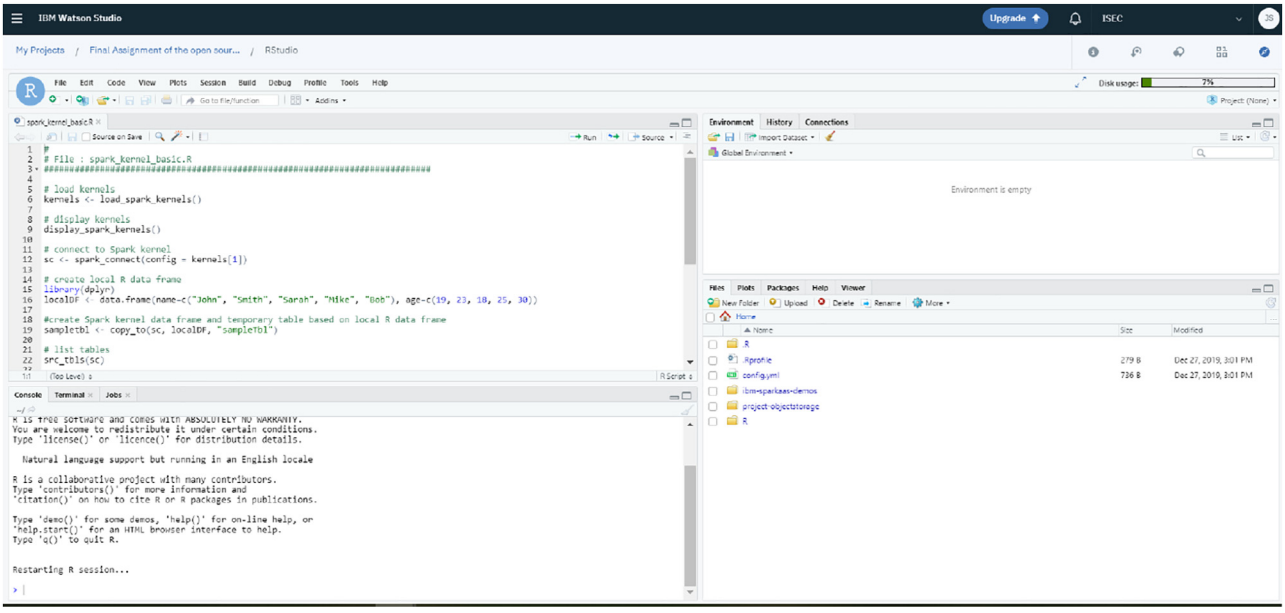

Fig. 2. RStudio Interface.

A print screen of a KNIME interface is shown in Fig. 1. In the top center of the print screen, a data workflow made with the KNIME VCF feature is shown - the depicted VCF computes Spearman correlations. On the right, the description of a selected component is seen, and on the left, the tool's workspace.

The KNIME development team claims that their concerns are to develop a tool that could process and integrate huge amounts of diverse data that would be robust, modular and highly scalable encompassing various data loading, transformation, analysis, and visual exploration. Their aim is to help users perform KDD in a faster and easier way. In fact, the KNIME tool has a Graphical User Interface (GUI) based on the popular Eclipse IDE; adheres to the visual programming paradigm with its VCF feature; has several already built examples to reduce the end users learning curve; can integrate and blend several types of data in a data mining task, such as, databases, simple text files, documents, images (BioSolveI, 2011), graphs and Hadoop-based data; and can integrate with other data mining tools like Weka and R, which gives access to several or personalized algorithms over its groups of already coded data mining algorithms. These already coded algorithms include Bayes, Clustering, Rule Induction, Association Rules, Neural Network, Decision Tree, Miscellaneous Classifiers (such as the K-Nearest Neighbor), Ensemble Learning, Multi-Dimensional Scaling (MDS), Principal Component Analysis (PCA), Predictive Model Markup Language (PMML), Support Vector Machine (SVM), and Feature Selection. Most of these algorithms have been already applied in the healthcare industry.

Its leading solution is the open source KNIME Analytics Platform that can be extended with the KNIME Commercial Software. KNIME is supported by various operating systems including Windows, Mac OS and Linux.

The KNIME's strongest points are the following (Almeida et al., 2016) (Ramesh et al., 2020):

- It has a short learning curve: it has a familiar GUI to a lot of programmers due to its Eclipse based IDE and because data mining algorithms are already coded.
- Data mining tasks can be performed on several types of data.
- It is a complete solution, which incorporates over 100 processing nodes for data I/O, preprocessing and cleansing, modeling, analysis and data mining as well as various interactive views, such as scatter plots, parallel coordinates and others.
- It can easily integrate with other data mining tools offering access to a vast library of statistical routines through the R solution, for example.
- The aspect that truly sets it apart from other data mining tools is its ability to interface with programs that allow the visualization and analysis of molecular data.

Its weakest points are the following (Ramesh et al., 2020):

- It has limited error measurement methods.
- It has no wrapper methods for descriptor selection.

6.2

R (RProject, 2021) is a programming language as well as an environment for statistical computing and graphics that is actually the first choice by statisticians. Its use by statisticians is old due to R being the successor of the statistical language S, originally developed by Bell Labs in 1970s. The R source code is written in C++, Fortran and in R itself (Jović et al., 2014). Hence, this tool has the ability to easily integrate with a code made in any of these languages, as well as in C and Python, which makes it a powerful tool that can perform any kind of data mining tasks.

The integrated development environment (IDE) of R is named RStudio (RStudio, 2016). This IDE not only supports direct code execution built with conditional, loops, input and output commands, but also, includes a console, as well as tools for plotting, history, debugging and workspace management. Hence, this tool allows the full cycle of data mining process to be performed including manipulation, calculation and display, as well as effective methods for data storing, handling and intermediate data analysis through graphical facilities (Almeida and Bernardino, 2016). A print screen of the RStudio interface, open in the IBM Watson Studio environment, is shown in Fig. 2 where some of its features can

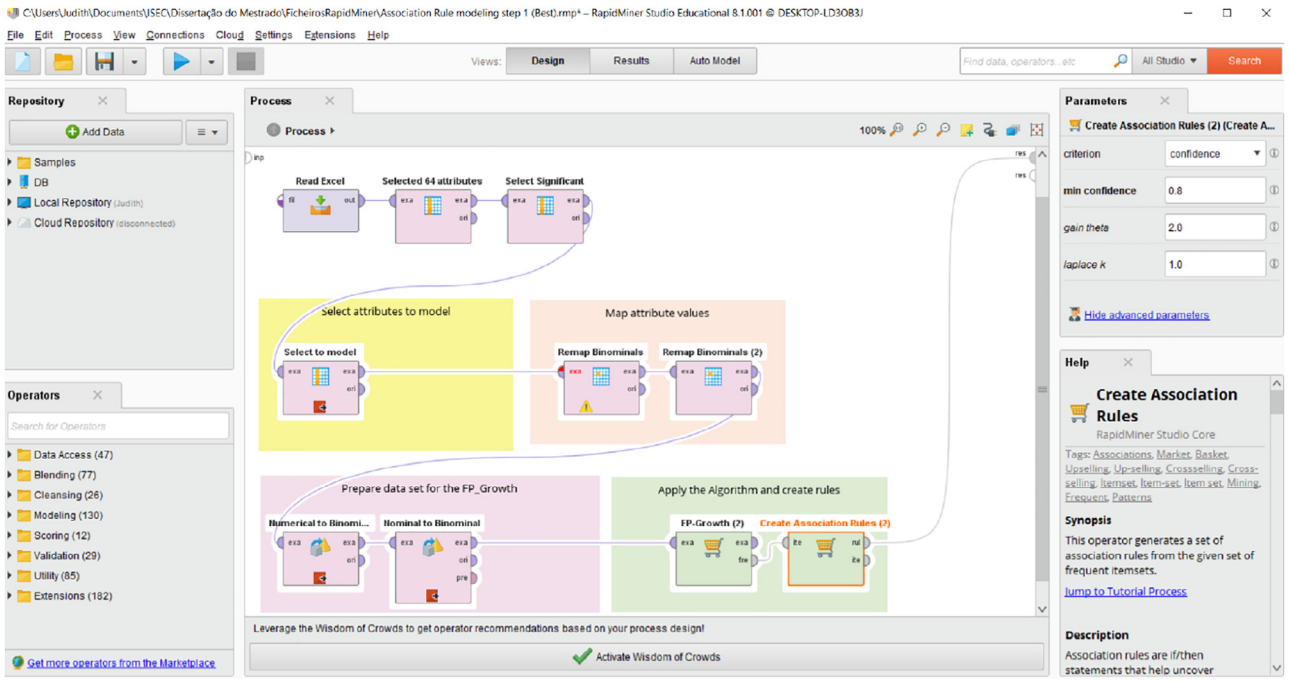

Fig. 3. RapidMiner Interface.

be seen: the source code at left; the console at the bottom left; and the workspace at right.

As shown in (RStudio, 2016)(RStudio, 2017), there are several R packages that can be added to the RStudio IDE to have a more interesting solution. For example, the Markdown package can turn analyses into reports, presentations and dashboards to faster present data results; the Shiny package can build an interactive web application to share the data analysis results; the ggplot2 package eases the visualization of data with multi-layered graphics; the Haven package allows loading foreign data formats (SAS, SPSS, and Stata); and the tidyr package incorporates other packages to merge, visualize and model the data.

The RStudio IDE is available in open source or commercial license and runs on Windows, Mac OS and Linux desktops, as well as in a browser connected to RStudio Server (the open-source license version) or RStudio Server Pro (the commercial license version). This last connection type can be very useful for projects where data mining tasks are made as team tasks, or there is a need to remotely access the mined results. Prior to the RStudio installation, the R solution will have to be installed.

The major advantages of this tool are (Almeida and Bernardino, 2016):

- Users can build algorithms that suite their tasks and domains.
- It is easy to integrate R with other coding languages.
- It provides extension capabilities through the use of packages.
- It centralizes access and computation through its Server version.
- Its-Haven package permits loading foreign data formats like SAS, SPSS and Stata.

Its major weaknesses are:

- Users need to have advanced technical knowledge.
- Users have to know how to program in the R language.

6.3. RapidMiner

RapidMiner (RapidMiner, 2017) is a mature Java-based end-to-end analytic tool for data mining, text mining, predictive analytics and business analytics (Almeida et al., 2016) developed by the company of the same name. This solution has been used in several areas and it is actually the most popular stand-alone and open-source solution on the market (Poll, 2019), as well as a market leader in its field (Gartner, 2019). The RapidMiner solution has five tools/editions: i) *RapidMiner Studio*, a client application with a graphical user interface (GUI) that supports the implementation of a complete predictive analytic workflow with a VCF feature that can cover the main data mining tasks such as data integration, cleaning, transformation, exploration, modeling, and validation; ii) *RapidMiner Server*, a server for collaborative team work, running automated and scheduled jobs, deployment and integration with other systems, and the creation of web based application; iii) *RapidMiner Radoop*, a set of capabilities to perform data mining in Hadoop to accelerate the analysis on large amounts of data and overcome the complexity that the Spark Hadoop has for non-technical users; iv) *RapidMiner Extensions*, additional capabilities provided by the community such as Text Processing, Web Mining, WeKa Extension, Text Analysis by AYLIEN, and Series Extension; and v) *RapidMiner Cloud*, capabilities that enable data mining jobs to be processed in the cloud.

From the RapidMiner website, this tool can access more than 40 file types including SAS, ARFF, Stata and via URL; access to text documents, web pages, PDF, HTML and XML, as well as to the NoSQL databases, MongoDB and Cassandra; model with a greater set of modeling capabilities and algorithms like similarity calculation, clustering, market basket analysis, decision trees, rule induction, Bayesian modeling, regression, neural networks, support vector machine, memory-based reasoning, model ensembles and estimate model performance with several validation techniques and performance criteria; and analyze large amounts of data through several other tools like Hadoop, Spark, Hive, MapReduce, Pig, and Mahout.

Fig. 3 shows a print screen of the RapidMiner interface where in the center of the picture a workflow that applies FP-Growth association mining algorithm on a clinical data set.

RapidMiner is an open-source and commercial solution that is priced by the amount of data (above 10,000 rows for the RapidMiner Studio) and memory used (above 2 GB of RAM for

the RapidMiner Server). For community or educational purposes, RapidMiner Radoop and RapidMiner Cloud are free.

RapidMiner Studio tool is supported by Windows, Mac OS and Linux.

Its major advantages are the following (Van Poucke et al., 2016) (Almeida and Bernardino, 2016):

- It supports all computer environments.
- All its methods can run in-memory, in-database or in clusters with Hadoop - useful to analyze large amounts of data.
- It has a wide range of data visualization output such as 3D graphs, scattered matrices, and maps.
- It provides a visual interface (GUI) that abstracts the user from implementation details.
- It has an API that provides extension capabilities, versatility of configuration and connection to other tools like R and Spark which eases the use of these more complex tools.

Its major weaknesses are (Linden et al., 2017):

- Users have reported issues with its documentation.
- The price of the commercial solution can be unpredictable since it is based on the amount of the data that is managed.
- Its free editions are limited (i.e. Rapid Miner Studio Free Edition) can only manage 10,000 rows of data unless Radoop tool is used with the community support.

6.4. Scikit-learn

Scikit-learn (Scikit-Learn, 2017) is a machine learning library for the Python programming language that performs data mining and data analysis tasks. Its development started in a Google Summer Code project by David Cournapeau and has been used in several areas including the healthcare domain (Almansa and, Macedo, 2021)(Culotta and Aron, 2014) for applications, such as neuroimaging (Michaud, 2014).

This tool has only a command-line interface (CLI) solution without a GUI. Therefore, a skilled programmer in Python (Jović et al., 2014) is required, even if most of the data mining algorithm groups - clustering, classification, regression and dimensionality reduction - as well as features to support model selection and data preprocessing methods are already implemented. Model results are visually seen, and its main advantages is its performance and its capacity to mine all data types (numeric, text, document, image, graph, speech, and audio).

Since Scikit-learn is built on top of NumPy (to handle large arrays of data) and SciPy (the fundamental library for scientific computing), these packages will need to be firstly installed. Scikit-learn can be installed in Windows, Mac OS, and Linux operating systems.

The main advantages of Scikit-learn are (Scikit-Learn, 2017) (BenLorica, 2015):

- Its commitment to documentation and usability - several coded examples are shared in its website.
- Its models are implemented by a dedicated team of experts.
- It covers most of the data mining algorithms.
- It can mine all data types.
- It has a simple flow chart that guides the user for its model selection.
- It delivers good performance.
- It supports a programming language that is preferred by many data miners – Python.

Its major weaknesses are:

- It is not an end-to-end analytic platform; thus, the integration with other solutions might be necessary.
- It has a high learning curve and modeling is made through coding without VCF features.
- It does not handle large amounts of data in a straightforward way.

6.5. Spark

Spark (Spark, 2021) is a fast and general engine for large-scale data processing that can execute data mining methods across nodes when the data is too large to fit and manage in one computer. It was originally developed in the Scala programming language at the University of California in Berkeleýs AMPLab (Wikipedia, 2017), but it is now maintained and under active development in the Apache Software Foundation with other similar tools like the Hadoop MapReduce. Spark has been used in areas that especially require large-scale data, and thus have been proposed (Gui et al., 2016)(Dhoka and Kudale, 2016) and used (Pita, 2015) in the healthcare industry.

According to Sparḱs official web site (Spark, 2021), Spark runs programs up to 100 times faster in memory, or 10 times faster on disk than its opponent, Hadoop MapReduce. This performance statement was supported by other sources (Noyes, 2015)(Kedia et al., 2016)(Gu and Li, 2013) since Spark was firstly designed to overcome Hadoop̀s shortages in iterative operations (Gu and Li, 2013) which data mining methods have. Thus, Spark has been gaining popularity in the data mining field due to its performance enhancement in mining large amounts of data with its optimized engine and libraries. According to KDnuggets, the usage of tools to process large-scale data grew 39%, driven mainly by the big growth seen in the usage of the Spark Machine Learning Library (a Spark library for data mining) named MLlib (Poll, 2019). Since Spark is a trend-setter to manage large amounts of data and a foundation for many data science advances (Linden et al., 2017), all the selected popular tools discussed in this paper have now the ability to connect to it, and therefore, have the capacity to mine large data sets, and handle streaming data as well as different data types.

Since Spark has more than just the MLlib Library in its rich ecosystem (Meng et al., 2016) that can support the data mining methodology, we will describe all Spark Libraries that are included as modules at Sparḱs installation. However, we will focus more on the MLlib Library due to its strongest relation with this papeŕs goal. These Libraries are described based on the information presented in the Sparḱs official website (Spark, 2021) and related application (Databricks, 2016):

- **Spark SQL** – Enables the query of several structured data types (Numeric, String, Binary, Boolean, Datetime data types as well as complex data types such as, Array, Maps and Structured types (Spark, 2017) with SQL queries inside Spark programs. Thus, this Library can be useful to, for instance, visualize in different ways the data to mine in the Sparḱs API. Furthermore, Spark SQL provides a common way to access and join a variety of data sources, including Hive, Avro, Parquet, ORC, JSON and other databases that can be connected with JDBC - a database connection standard.
- **MLlib** – Allows the use of the built-in data mining methods in the Sparḱs APIs such as *Classification*, *Clustering*, and *Association*. However, the *Outlier* method is not yet developed. The MLlib

Table 2
Comparison of Data Mining tools.

Critical Capabilities	Domain Requirements	KNIME (KNIME, 2017)	R (RProject, 2021)	RapidMiner (RapidMiner, 2017)	Scikit-Learn (Scikit-Learn, 2017)	SPARK (Spark, 2021)
Performance and Scalability	***Large Amounts of Data***	+ (KNIME Sparkexecutor)	Add-on (sparklyr package)	+ (RapidMiner Radoop)	Add-on (spark-sklearn package)	+
	Cloud data	+	+	+	± (IPython shell)	+
	Streaming data	+	Add-on (stream)	+	±	+ (Spark Streaming Library)
Data Access	***Multiple data sources***	Add-on (Imports from most data sources)	Add-on (Imports from most data sources)	+ (Imports from most data sources)	Add-on (Imports from most data sources)	+ (Imports from most data sources)
	Different data types	Add-on (Supports numeric, text, document, image, graph, speech, audio and video data)	Add-on (Supports numeric, text, document, image, graph, speech, audio and video data)	Add-on (Supports numeric, text, document, image, graph, speech, audio and video data)	Add-on (Supports numeric, text, document, image, graph, speech, audio and video data)	+ (Supports numeric, text, document, image, graph, speech, audio and video data)
Data Preparation	***Dirty Data***	+	Add-on (Has built-in functions for data preprocessing)	+	± (Has built-in functions for data preprocessing)	+ (Has built-in functions for data preprocessing)
Data Exploration and Visualization	***Complex data***	+	Add-on (Has built-in functions for data visualization)	+	Add-on (Has built-in functions for data visualization)	Add-on (Has built – in functions for data visualization)
Advanced Modelling	***Classification***	+	± (Algorithms have to be coded)	+	+	+
	Clustering	+	± (Algorithms have to be coded)	+	+	+
	Association	+	± (Algorithms have to be coded)	+	+	+
	Outlier	+	± (Algorithms have to be coded)	+	+	± (Algorithms have to be coded)
	Automation	+	± (with Add-on packages)	+	± (with coding)	± (with coding)
Good User Experience	***Programming Language***	Code Free	R, C, Fortran	Code Free	Python + NumPy + SciPy + matplotLib	Java, Scala, Python or R
	Operating System	Window, Linux, Mac OS X	Window, Linux, Mac OS X	Window, Linux, Mac OS X	Window, Linux, Mac OS X	Window, Linux, Mac OS X
	Interfaces	GUI	CLI, IDE	GUI	CLI, IDE (with IPython console)	CLI, API
	Visual Compo. Framework	+	–	+	–	–
	Collaboration	+	Add-on (Git & GitHub)	+	Add-on (with GitHub)	Add-on (with GitHub)
	Ease of use	+	± (Has the RStudio IDE)	+	– (Has a CLI and IDE interface)	– (Has CLI and API interfaces)
	Tool Type	End-to-end Analytic Tool	Statistics and Computing	End-to-end Analytic Tool	Machine Learning Package	Large-scale data & process mining
	Community Support	Moderate (∼15 K users)	Very large (∼2M users)	Large (∼2K users)	Moderate	Large (344 331 users)

interoperates with NumPy in Python (as of Spark 0.9) and R libraries (as of Spark 1.5). Furthermore, this Library also includes utilities to perform feature transformations such as standardization and normalization, as well as to build summary statistics, which is useful for *Data Preprocessing*. In terms of supported data sources, MLlib allows any Hadoop data source (e.g. HDFS, HBase, Hive, and local files), making it easy to plug into Hadoop workflows, as well as Cassandra, the Amazon simple Storage Service S3 and Tachyon (for in-memory storage).

- **Spark Streaming** – Enables the writing of streaming jobs which can be useful to mine streaming data along with the MLlib Library previously disclosed.
- **GraphX** – Unifies the exploratory analysis with the iterative graph computation within a single system which enables the view of the same data as both graphs and collections. It also enables the transformation and join of both data types, as well as the writing of custom iterative graph algorithms with the Pregel API.

Spark runs on Windows, Linux, and Mac OS. It can run locally (in one machine with the Java installation), in several machines or in the Cloud (Spark, 2021). At Spark installation, a Pre-built Hadoop package and other related programming languagés software need to be installed. Project collaboration can be performed through the development platform named GitHub.

The major advantages of Spark are:

- It can be easily used with Hadoop tools.
- It has an API that supports several languages – Scala, Java, Python, and R.
- It exhibits excellent performance and scalability (Meng et al., 2016) - data mining methods can be run through several computers.
- It is fitted to mine large data sets.
- It is a solution that is constantly improving along with the MLlib Library.

Its major weaknesses are:

- It does not have a GUI. However, one of its suggested APIs can be used for coding.
- It is not a code free solution, that is to say that it does not have the VCF feature.
- Users need to be tech savvy – know Java, Scala, Python or R, must install a data mining environment, which implies the use of several software packages.

7. Data mining tool comparison

In this section, we present a comparative analysis of the data mining tools described in the previous section and discuss how to select the most suitable tool for the healthcare industry. The analysis is based on the proposed data mining tool selection criteria presented in Section 5. Therefore, the selected tools in Table 2 are compared based on:

1) *Critical Capabilities* – Performance and Scalability, Data Access, Data Preparation, Data Exploration and Visualization Advanced Modelling, and User Experience;
2) *Data Characteristics* - Large Amounts of Data, Cloud Data, Streaming Data, Multiple Data Sources, Different Data Types, Dirty Data, Complex Data; and
3) *Data Mining Methods* applied in the healthcare industry - Classification, Clustering, Association and Outlier – with the Automation model requirement

Tools are also compared against the User Experience capability they have based on the papers (Jović et al., 2014)(Almeida and Bernardino, 2016) which are: Programming Language, Supported Operating System, Interfaces, Visual Composition Framework, Collaboration, Ease of Use, Tool Type and Community Support.

The Tool comparison is made by specifying, inspired by Alan Jović's work (Jović et al., 2014):

(**+**) if the tool either handles the feature on its own;

(**±**) if it needs an external add-on or tool to perform the task (Add-on) and shows more or less the ability to handle it;

(−) if it does not handle it at all.

The data mining toolś information presented in Table 2 was also gathered through each tooĺs official website where we have considered the statements related with the open-source tooĺs versions due to the aim of our investigation.

By considering the toolś User Experience presented in Table 2, we can say that the most suitable tools are the RapidMiner and the KNIME. They are easy to use and code-free solutions, supported by the mainstream operating system Windows and with a GUI to ease its use. Furthermore, these tools have enough users (community support) to help new ones on their daily questions and features/tools to support and ease their work - the VCF feature for data modelling and an integrated tool to collaborate with other team members. By considering the other critical capabilities, we can say that they all can be used since all domain requirements are fulfilled. However, we cannot say the same thing for the other compared tools: R needs programming knowledge to write data mining methods, build models and select the better ones, as Scikit-Learn and Spark do; and Scikit-Learn has difficulties to mine streaming and cloud data. Since these tools are not end-to-end analytic tools with GUI interfaces, their data visualization capabilities are lower than RapidMiner and KNIME. Therefore, managing Complex and/or Dirty data can be harder for analysts who do not have advanced technical knowledge.

In the last years, RapidMiner has been scored similarly to KNIME in the Gartneŕs annual evaluations where they have been both standing as vendors leaders on the data mining toolś market, alongside commercial solutions. KNIME has highly scored in the 2016 evaluation on its Data Access, VCF, Automation and Collaboration capability (Almansa and Macedo, 2021) but has been slightly surpassed in 2017 by its opponent RapidMiner (Linden et al., 2017). Regarding the other described tools, they have not been covered in these evaluations because Gartner only evaluates end-to-end analytic tools that handle specific features such as the VCF requirement.

According to (Mikut and Reischl, 2011), researchers are mainly interested in tools with well-proven domain related data mining methods, a graphical user interface (GUI) and interfaces to domain-related data formats or databases. Thus, KNIME and RapidMiner go with this suggestion. Specifically KNIME was born in the healthcare field which provides features that other tools do not - like the ability to interface with programs that allow the visualization and analysis of molecular data (Ramesh et al., 2020).

Due to their similarities, KNIME and RapidMiner have been evaluated in several works (Almeida et al., 2016)(Al-odan and Saud, 2015)(Singh et al., 2016). In a work carried out in (Singh et al., 2016), KNIME was seen as a better solution than RapidMiner in terms of Data Import and Export Support, but with equivalent capacities on Data Manipulation and Transformation. In terms of good user experience, a work held in (Al-odan and Saud, 2015) had qualitatively evaluated the following data mining tools: RStudio; RapidMiner, Weka, KNIME and Orange. These tools were assessed by 17 participants from fresh graduates of IT related fields to individuals with more than 15 years of work experience and the results show that KNIME and RapidMiner achieve far better results than the rest of the evaluated tools. However, they have both their strengths and weaknesses. In terms of intuitiveness of use and software consistency, KNIME scored better than RapidMiner. Nevertheless, the RapidMiner tool presented a better navigation, usability, installation manual, configuration guide, troubleshooting guide and user tutorials than KNIME. In this work, the RStudio tool had the lowest scoring. In (Al-odan and Saud, 2015), the weak RStudio qualitative scoring was due to the nature of RStudio tool which is an IDE for R. Thus, a great part of the functionality and performance of RStudio depends greatly on R itself. By extrapolating this explanation to the Spark and Scikit-learn tools, we believe that the qualitative scoring would be even lower in this evaluation for being a set of Libraries accessed through a Command Line Interface (CLI) with APIs. Thus, in Table 2, this thought is reflected in the ease of use requirement where we have put a minus sign (-) under Spark and Scikit-learn, and a (±) sign under RStudio. In spite of Sparḱs use difficulty worsened by the non-existence of a VCF feature, it is a good tool. Gartner even states that Spark is a trendsetter and a foundation for many data science advances (Linden et al., 2017). However, its use difficulty can be managed by accessing it through the RapidMiner or KNIME tool that can connect to Spark, and thereby creating an abstraction layer on top of it to ease its use.

To the best of our knowledge, an evaluation of the performance of Spark vs. Scikit-learn and R does not exist which is in part understandable since Spark Libraries, especially the MLlib Library, was originally built to mine large amounts of data in contrast to the other two tools. Thus, Spark is usually compared with similar tools such as Hadoop MapReduce (Gu and Li, 2013), that we have seen in Sparḱs description.

Several works have been conducted on the performance evaluation of RapidMiner and KNIME. For example, the work in (Singh et al., 2016) concluded that RapidMiner uses less memory in its data mining jobs than its rival, KMINE, in spite of its similar execution time, and that it is a better solution than its rival in terms of "Node IO Limit", "Scripting Language Support", "Visualization Support" and "Interface Usability"; and the work in (Almeida et al.,

2016) showed that RapidMiner always give better performance than the KNIME. All these works have not specified which tool versions they have evaluated which makes it hard to identify if a tool feature has overcome its capabilities on the latest version. Furthermore, Sven Van Poucke, a data scientist in the healthcare domain, has reviewed RapidMiner and KNIME and has stated that, their visual environments and free coded methods make them suitable for the medical community that usually does not have programming skills (Van Poucke et al., 2016).

By considering all the evaluations performed by other authors and our own analysis, we can say that KNIME is a good solution, but RapidMiner should be first considered if it fulfills the data requirements of the application under investigation due to its good evaluation results seen in the related works.

8. Related work

Considering the existing related works, we can say that they all provide information about the popular open-source data mining tools; but to the best of our knowledge, we have not found a survey on data mining tools for the healthcare industry that investigates the domain requirements to elect the most suitable tool. Nevertheless, due to the contribution that the related works on data mining tools have brought to our work, in this section we describe the essence of these works.

In (Mikut and Reischl, 2011), an overview of existing data mining tools was presented. Its main contribution is the presentation of the tools' categorization criteria based on different user groups, data structures, data mining tasks and methods, visualization and interaction styles, import and export options for data and models, platforms, and license policies. The authors used these tools' categorization criteria to classify the data mining tools into nine tool types (e.g. Data mining suites, Business intelligence packages, Mathematical packages etc.). The benefit of the authorś approach is that they listed most data mining tools by specifying their tool types. Moreover, they also identified which types of tools are suitable for identified user groups, which are business applications, applied research, algorithm development, and education. Since the applied research group is the target of our investigation, we have considered the suggestion of the authors to select a tool that has well-proven domain related data mining methods, a GUI, and an interface to domain-related data formats or databases. The drawback of this work is that they did not describe the identified data mining tools in depth and did not compare them. However, this work contributed to our proposed data mining tool selection criteria.

In (Begum, 2013), the authors discussed the KDD process and various open source tools (R, Weka, Orange, RapidMiner, and Tanagara). Its highlights are the identification of data mining current and future trends in most domains – cloud and distributed computing were identified as well as the heterogeneous and complex data characteristic. Another useful input is the identification of data mining methods to tackle the trends that are basically challenges of the KDD process. However, the discussed data mining tools were not compared or selected by any criteria.

Due to the increase popularity of data mining tools, a number of data mining tool surveys were conducted. In (Ramesh et al., 2020) a theoretical analysis of six open source data mining tools, Weka, Keel, R, Knime, RapidMiner, and Orange, was given. The strongest points of this paper it is that each data mining tool was described through its technical specifications, features and specializations, as well as its advantages and limitations. Unfortunately, the authors did not say why they have considered these tools in their survey, neither specified the areas to which they are suited. Furthermore, for healthcare researchers who are beginners in the data mining field and want to identify the most suitable tools for their needs, all the technical aspects shown can be overwhelming. Therefore, we believe that an elaborating survey that targets one specific application field like ours will likely be more helpful to choosing a data mining tool to adopt.

In (Jović et al., 2014), a data mining tool survey was conducted that specifies the reason of its tools of choice – mostly on the results of the KDnuggets poll. Their work emphasizes the quality of RapidMiner, R, Weka, and KNIME platforms identified in the KDNuggets poll, but also acknowledges the significant advancements made in other tools like Orange and Scikit-Learn. The strongest aspect of this paper it its technical aspects: the selected tools were compared based on their general characteristics (i.e. programming language, license, etc.), applicability (i.e. Big Data, Text Mining, etc.) and data mining algorithms and procedures (Data visualization, Decision tree algorithms, etc.). Therefore, this work is related to our work and is helpful for the technical descriptions and comparisons of the commonly selected tools.

The performance of the classification algorithms (Naïve Bayes, Random Forest, Random Tree, and Bagging) were evaluated through the use of three data mining tools (Weka, RapidMiner and Support Vector Machine) in (Mishra and Thakur, 2014). The major contribution of this work is the identification of the best algorithm and tool for spam/junk mail classification. One of the benefits of this work is the performance evaluation of the tools in a specified field. In spite of the tools being compared by its performance, they were not described.

A similar work was carried out in (Singh et al., 2016) that presents an evaluation of the data mining tools, RapidMiner and KNIME, on a customized predictive and descriptive model built with the VCF feature for telecom monitoring data. The strongest points of this paper are that it shows how it is possible to build a model in the KNIME and RapidMiner with the VCF feature, as well as its performance evaluation through benchmarking. In spite of the tools being analyzed qualitatively and quantitively, the authors did not present a qualitative comparative table. We believe that this type of comparison would more likely help in selecting a tool to adopt.

In (Al-odan and Saud, 2015), a comparative study of data mining tools suited for small to medium enterprises (SMEs) was presented. The reviewed data mining tools were KNIME, Rapid Miner, Weka, RStudio, and Orange. These tools were evaluated by 17 participants with more than 15 work years' experience in the IT field to assess their user experience through their intuitiveness, consistency, navigation, usability, installation manual, configuration guide, troubleshooting guide, and user tutorials. Since the ease of use is one of our proposed criteria for data mining tool selection, under the user experience critical capability, this work contributed to our tool's comparative analysis. The weakest point of this paper is that the data mining tools are not clearly described in detail in spite of their comparison.

In (Aalam and Siddiqui, 2016) seven data mining tools - Weka, ELKI, Orange, R, KNIME, Scikit-learn, and Rapid Miner – were compared for clustering. The positive aspect of this paper is that it describes and compares qualitatively (programming language, interface type, covered clustering algorithm, etc.) the data mining tools. However, its focus is only on clustering.

In (Almeida and Bernardino, 2016) a survey on seven open source data mining tools for SMEs - KEEL, KNIME, Orange, RapidMiner, RProject, Tanagra and Weka - were qualitatively compared (using programming language, interface existence, data types supported, etc.). The strongest points of this paper are the good descriptions of the selected tools and the identification of the Cloud Services and Big Data (large amounts of data) support of the tools. Since those aspects are important to assess data mining tools for the healthcare domain as well as due to its good tool

description, we have adopted some of its criteria. The weakest point of this paper is that it does not evaluate the performance of the selected tools. In a subsequent paper (Alam et al., 2016), the authors have addressed the interest of data mining for business and analyzed three popular open-source data mining tools – KNIME, Orange, and RapidMiner. The strongest point of this paper is that its tool evaluation besides comparing the execution time of the tested algorithms, has also compared the results on seven other performance metrics – Precision, Recall, F-Measure, ROC, Accuracy, Specificity and Sensitivity. Even though this work is related to the business domain, our work gained from it through its tool analysis and evaluation.

In (Sharma et al., 2016) a survey on the application of the Classification and Clustering data mining methods to heart and cancer diseases was provided. The paper briefly suggested the following data mining tools: Rapid Miner, Weka, R-Programming, Orange, KNIME, and NLTK. The strongest point of this work is its survey on its data mining applications in its projecťs aim. However, the way that it has suggested the data mining tools is its weakest point - they are briefly described, neither compared nor suggested based on healthcare requirements. In spite of that, this work provides the knowledge on the support of classification and clustering applications in the tools we investigated, as well as our data mining tool selection, for the healthcare industry. It is important to say that Weka, Orange and NLTK were not covered in our work due to its unpopularity seen in the KDNuggets poll (Poll, 2019) since our aim is to analyze popular tools to suggest a solution that is up-to-date and with well-proven applications.

In (Gui et al., 2016), a data management architecture for the personal health problem detection and real-time vital sign monitoring was proposed. The strongest point of this work is its data architecture suggestion that has considered the large amount of data characteristic with their specificities. However, the domain requirements are not deeply investigated, leaving behind other requirements such as the user experience of the proposed data management architecture. Its prototyped system was constructed with the Hadoop database named HBase, the Hadoop Data Warehouse Hive and the data mining Libraries MLlib and Spark Streaming. The proposed toolś combination makes sense since the Spark Libraries work perfectly with the Hadoop tools but it is not an architecture that is easy to work with for beginners in the data mining field. However, this work is related to ours since we have compared the Spark tool based on the identified requirements to assess its adequacy to the healthcare industry.

As we have seen in this section, a descriptive and comparative analysis among the popular open-source data mining tools for the healthcare domain has not been provided in the existing literature. Therefore, in this paper, we aim to fill this gap with the purpose to help researchers and other personnel of the healthcare industry to choose an appropriate open-source data mining tool for their applications.

9. Conclusions and future work

Computer scientists are not usually trained in domain specific medical concepts, whereas medical practitioners and researchers also often have limited exposure to the data mining area. To fill this gap, we have firstly identified the requirements of the healthcare industry in terms of its data characteristics and mostly used data mining methods to propose a set of data mining tool selection criteria based on its critical capabilities. Afterwards, the popular open-source data mining tools - KNIME, R, RapidMiner, Scikit-learn, and Spark - were described and compared using the proposed data mining tool selection criteria. Finally, the comparison of the tools was discussed to suggest the most suitable tools for the healthcare industry.

Through our tool analysis and comparison, we have concluded that RapidMiner and KNIME are the best solutions due to their wider coverage of the identified healthcare requirements compared with the other tools. Nevertheless, through our related work investigation made on the evaluation of these two selected solutions, we have concluded that RapidMiner is a better solution than KNIME if RapidMiner fulfills the data requirements of the specific healthcare application under investigation.

As future work, we aim to apply RapidMiner in a real use case in the healthcare field to contribute to the research on infertility field.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aalam, P., Siddiqui, T., 2016. Comparative study of data mining tools used for clustering. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 3971–3975.

Alam, M. A. U., Roy, N., Holmes, S., Gangopadhyay, A., Galik, E., 2016, "Automated Functional and Behavioral Health Assessment of Older Adults with Dementia," in: 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), pp. 140–149.

Almansa, L. F., Macedo, A. A., 2021, "Sistema de Informação para Perguntas e Respostas em Doenças Crônicas," in: XXXVI Congresso da Sociedade Brasileira de Computação, pp. 2587–2596.

Almeida, P., Bernardino, J., 2016. A survey on open source data mining tools for SMEs. Adv. Intell. Syst. Comput. 444, 253–262.

Almeida, P., Gruenwald, L., Bernardino, J., 2016. Evaluating open source data mining tools for business. In: Proc. 5th Int. Conf. Data Manag. Technol. Appl. no. Data, pp. 87–94.

Al-odan, H.A., Saud, A.A.A.K., 2015. Open Source Data Mining Tools. In: 1st International Conference on Electrical and Information Technologies ICEIT'2015 Open, pp. 369–374.

Antonie, M.-L., Zaane, O.R., Coman, A., 2001. Application of data mining techniques for medical image classification. In: Proceedings of the Second International Workshop on Multimedia Data Mining in conjunction with ACM SIGKDD conference, pp. 94–101.

Aridhi, S., Mephu Nguifo, E., 2016. Big graph mining: frameworks and techniques. Big Data Res. 6, 1–10.

Aziz, K., Tarapiah, S., Ismail, S.H., Atalla, S., 2016. Smart real-time healthcare monitoring and tracking system using GSM/GPS technologies in 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC), pp. 1-7.

Begum, H., 2013. Data mining tools and trends – an overview. Int. J. Emerg. Res. Manag. Technol. ISSN, 2278–9359.

Bellaachia, A., Bari, A., 2012. A flocking based data mining algorithm for detecting outliers in cancer gene expression microarray data. In: 2012 International Conference on Information Retrieval & Knowledge Management, pp. 305–311.

BenLorica, "Six reasons why I recommend scikit-learn - O'Reilly Media," 2015. [Online]. Available: https://www.oreilly.com/ideas/six-reasons-why-i-recommend-scikit-learn. [Accessed: 16-Jan-2017].

BioSolveIT GmbH, "newsletter #20 - Q2/2011," 2011. [Online]. Available: https://www.biosolveit.de/newsletter/archive/issue20.html. [Accessed: 23-Jun-2017].

Brossette, S.E., Sprague, A.P., Hardin, J.M., Waites, K.B., Jones, W.T., Moser, S.A., 1998. Association rules and data mining in hospital infection control and public health surveillance. J. Am. Med. Informatics Assoc. 5 (4), 373–381.

Bruse, J.L. et al., 2017. Detecting Clinically Meaningful Shape Clusters in Medical Image Data: Metrics Analysis for Hierarchical Clustering applied to Healthy and Pathological Aortic Arches IEEE Trans. Biomed. Eng. pp. 1-1.

A. Culotta, Aron, Culotta, and Aron, "Estimating county health statistics with twitter," in: Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14, 2014, pp. 1335–1344.

Dash, S., Shakyawar, S.K., Sharma, M., Kaushik, S., 2019, "Big data in healthcare: management, analysis and future prospects," J. Big Data 6, vol. 54.

Databricks, "A Tale of Three Apache Spark APIs: RDDs, DataFrames, and Datasets - The Databricks Blog," 2016. [Online]. Available: https://databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-dataframes-and-datasets.html. [Accessed: 28-Apr-2017].

Delen, D., Walker, G., Kadam, A., 2005. Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med. 34 (2), 113–127.

Dhoka, S., Kudale, R.A., 2016. Use of big data in healthcare with spark. Int. J. Sci. Res. 5 (11), 401–403.

Elhoseny, M., Abdelaziz, A., Salama, A.S., Riad, A.M., Muhammad, K., Sangaiah, A.K., 2018. A hybrid model of Internet of Things and cloud computing to manage big data in health services applications. Future Generation Comput. Syst. 86, 1383–1394. https://doi.org/10.1016/j.future.2018.03.005.

Exarchos, K.P., Goletsis, Y., Fotiadis, D.I., 2012. Multiparametric decision support system for the prediction of oral cancer reoccurrence. IEEE Trans. Inf. Technol. Biomed. 16 (6), 1127–1134.

Gartner, "Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms", 2019.

Gonzalez, G.H., Tahsin, T., Goodale, B.C., Greene, A.C., Greene, C.S., 2016. Recent advances and emerging applications in text and data mining for biomedical discovery. Brief. Bioinform. 17 (1), 33–42.

Gu, L., Li, H., 2013, "Memory or time: Performance evaluation for iterative operation on hadoop and spark," in: Proceedings - 2013 IEEE International Conference on High Performance Computing and Communications, HPCC 2013 and 2013 IEEE International Conference on Embedded and Ubiquitous Computing, EUC 2013, 2014, no. November 2013, pp. 721–727.

Gui, H., Zheng, R., Ma, C., Fan, H., Xu, L., 2016. An Architecture for Healthcare Big Data Management and Analysis. Springer, Cham, pp. 154–160.

Han, J., Kamber, M., Pei, J., 2012, Data mining: concepts and techniques. Elsevier/Morgan Kaufmann, 2012.

Hand, D., Blunt, G., Kelly, M., Adams, N., 2000. Data Mining for Fun and Profit. Stat. Sci. 15 (2), 111–131.

Jović, A., Brkić, K., Bogunović, N., 2014. An overview of free software tools for general data mining. In: 37th Int Conv. MIPRO ... no. May, pp. 26–30.

Kaur, K., Rani, R., 2015, "Managing Data in Healthcare Information Systems : Many Models, One Solution," Computer (Long. Beach. Calif)., 48(3), p. 52-59.

Kedia, S., Wang, S., Ching, A., 2016, "Apache Spark @Scale: A 60 TB+ production use case," Facebook code, 2016. [Online]. Available: https://code.facebook.com/posts/1671373793181703/apache-spark-scale-a-60-tb-production-use-case/. [Accessed: 28-Apr-2017].

KNIME, "KNIME | KNIME Analytics Platform," 2017.

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. CSBJ 13, 8–17.

Kumar, V., Kumar, D., Singh, R. K., Bhoj, M. P., 2008, "Outlier Mining in Medical Databases: An Application of Data Mining in Health Care Management to Detect Abnormal Values Presented In Medical Databases," IJCSNS Int. J. Comput. Sci. Netw. Secur., 8(8).

Kumar, P., Wasan, S.K., 2010. Analysis of X-means and global k-means USING TUMOR classification. In: 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE), pp. 832–835.

Li, C., Darema, F., Chang, V., 2017. Distributed behavior model orchestration in cognitive internet of things solution. Enterp. Inf. Syst., 1–21

Linden, A., Kart, L., Hare, J., Herschel, G., 2016, "Critical Capabilities for Advanced Analytics Platforms,".

Linden, A., Krensky, P., Hare, J., Idoine, C. J., Sicular, S., Vashisth, S., 2017, "Magic Quadrant for Data Science Platforms,".

Makhabel, B., 2014. Mining stream, time-series, and sequence data. In: Learning data mining with R. Packt Publishing, p. 314.

Mala, K., Kumar, B.M., Vignesh, R., Kumar, K.M., 2016. A wearable diagnostic device to combat children's pneumonia. In: 2016 IEEE Global Humanitarian Technology Conference (GHTC), pp. 654–659.

Maroco, J. et al., 2011. 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 4 (1), 299.

Meng, X. et al., 2016. MLlib: machine learning in apache spark. J. Mach. Learn. Res. 17, 1–7.

Michaud, P., 2014, "Scikit-Learn donne de l'intelligence à nos systèmes," Inria, 2014. [Online]. Available: https://www.inria.fr/centre/saclay/actualites/scikit-learn-donne-de-l-intelligence-a-nos-systemes. [Accessed: 17-Jan-2017].

Mikut, R., Reischl, M., 2011. Data mining tools. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1 (5), 431–443.

Mirroshandel, S.A., Ghasemian, F., Monji-Azad, S., 2016. Applying data mining techniques for increasing implantation rate by selecting best sperms for intra-cytoplasmic sperm injection treatment. Comput. Methods Programs Biomed.

Mishra, R., Thakur, R.S., 2014. An efficient approach for supervised learning algorithms using different data mining tools for spam categorization. In: 2014 Fourth International Conference on Communication Systems and Network Technologies, pp. 472–477.

Muhammad, K., Sajjad, M., Mehmood, I., Rho, S., Baik, S.W., 2015. A novel magic LSB substitution method (M-LSB-SM) using multi-level encryption and achromatic component of an image. Multimed. Tools Appl. 75 (22), 14867–14893.

Muhammad, K., Sajjad, M., Baik, S., 2016. Dual-level security based cyclic18 steganographic method and its application for secure transmission of keyframes during wireless capsule endoscopy. J. Med. Syst. 114 (40), 1–16.

Newhouse, S. J., 2016, "HPCS 2016 KEYNOTES TUESDAY KEYNOTE Big Data Analysis in European Clouds : The Challenges for Life Science," no. Hpcs.

Noyes, K., 2015, "Five things you need to know about Hadoop v. Apache Spark," InfoWorld, 2015. [Online]. Available: http://www.infoworld.com/article/3014440/big-data/five-things-you-need-to-know-about-hadoop-v-apache-spark.html. [Accessed: 28-Apr-2017].

Orphanou, K., Dagliati, A., Sacchi, L., Stassopoulou, A., Keravnou, E., Bellazzi, R., 2016. Combining Naive Bayes classifiers with temporal association rules for coronary heart disease diagnosis. In: 2016 IEEE International Conference on Healthcare Informatics (ICHI), pp. 81–92.

Paul, R., Hoque, A.S.M.L., 2010. Clustering medical data to predict the likelihood of diseases. In: 2010 Fifth International Conference on Digital Information Management (ICDIM), pp. 44–49.

Pita, R. D. da R., 2015, "Correlação probabilística implementada em spark para big data em saúde," Instituto de Matemática. Departamento de Ciência da Computação.

Poll, KDnuggets, 2019, "What software you used for Analytics, Data Mining, Data Science, Machine Learning projects in the past 12 months?"

Pramanik, M., Lau, R.Y.K., Azad, M.A.K., Hossain, M.S., Chowdhury, M.K.H., Karmaker, B.K., 2020. Healthcare informatics and analytics in big data. Expert Syst. Appl. 152.

Pratim Ray, P., Dash, D., Moustafa, N., 2020. Streaming service provisioning in IoT-based healthcare: an integrated edge-cloud perspective. Trans. Emerg. Tel. Tech. 31.

Primova, H.A., Sakiyev, T.R., Nabiyeva, S.S., 2020. Development of medical information systems. J. Phys., Conf. Ser. 1441 (1).

Raghupathi, W., Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. Heal. Inf. Sci. Syst. 2, 3.

Ramesh, G.S., Rajini Kanth, T.V., Vasumathi, D., 2020, "A Comparative Study of Data Mining Tools and Techniques for Business Intelligence". in: Pant, M., Sharma, T., Basterrech, S., Banerjee, C. (eds.) Performance Management of Integrated Systems and its Applications in Software Engineering. Asset Analytics. Springer, Singapore.2020 DOI:10.1007/978-981-13-8253-6_15.

RapidMiner, 2017, "Data Science Platform | Machine Learning | RapidMiner,".

Ray, P., 2018. A survey on Internet of Things architectures. J. King Saud Univ. – Comput. Inf. Sci. 30 (3), 291–319.

Reddy, C., Aggarwal, C., 2015. Healthcare Data Analytics. CRC Press.

RProject, "R: What is R?".

RStudio, "RStudio – RStudio," 2016. [Online]. Available: https://www.rstudio.com/products/rstudio/. [Accessed: 27-Jan-2017].

RStudio, "R Packages – RStudio," 2017. [Online]. Available: https://www.rstudio.com/products/rpackages/. [Accessed: 27-Jan-2017].

Saeed, S., Shaikh, A., Memon, M. A., Naqvi, S. M. R., 2018, "Impact of Data Mining Techniques to Analyze Health Care Data," J. Med. Imaging Heal. Informatics, pp. 8(4), 682–690.

Sahoo, A., Kumar, Y., 2014. Seminal quality prediction using data mining methods. Technol. Heal. Care 22 (4), 531–545.

Scikit-Learn, "Classifier comparison — scikit-learn 0.18.1 documentation," 2017. [Online]. Available: http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html. [Accessed: 17-Jan-2017].

Scikit-Learn, 2017, "scikit-learn: machine learning in Python — scikit-learn 0.18.1 documentation,"

Sharma, R., Singh, S., Khatri, S., 2016. Medical data mining using different classification and clustering techniques: a critical survey. Second International Conference on Computational Intelligence & Communication Technology.

Shouman, M., Turner, T., Stocker, R., 2012, "Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients," Sch. Eng. Inf. Technol. Univ. New South Wales Aust. Def. Force Acad. Northcott Drive, Canberra ACT 2600, no. August 2014, pp. 1–7.

Singh, S., Liu, Y., Ding, W., Li, Z., 2016. Evaluation of data mining tools for telecommunication monitoring data using design of experiment. In: 2016 IEEE International Congress on Big Data Evaluation, pp. 283–290.

Smys, S., 2019, "Survey on Accuracy of Predictive Big Data Analytics in Healthcare," J. Inf. Technol., p. no.02, pp.77-86.

Song, Y.-T., 2016. Toward connected personal healthcare: Keynote address. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1–2.

Spark, "Spark SQL and DataFrames - Spark 2.1.0 Documentation." [Online]. Available: http://spark.apache.org/docs/latest/sql-programming-guide.html#data-types. [Accessed: 29-Apr-2017].

[17] Spark, "Apache SparkTM - Lightning-Fast Cluster Computing."

Stilou, S., Bamidis, P.D., Maglaveras, N., Pappas, C., 2001. Mining Association Rules from Clinical Databases: An Intelligent Diagnostic Process in Healthcare. MEDINFO.

Strang, K.D., Sun, Z., 2020. Hidden big data analytics issues in the healthcare industry Health Informatics Journal 26(2) 981–998 doi:10.1177/1460458219854603.

Sun, G. et al., 2017. Efficient Location Privacy Algorithm for Internet of Things (IoT) Services and Applications. J. Netw. Comput. Appl., Elsevier 89, 3–13.

Tan, Pang-Ning, Steinbach, Michael, Kumar, Vipin, 2006. "Introduction to data mining." Library of congress. Vol. 74.

Tekieh, M. H., Raahemi, B., 2015, "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 2015 -ASONAM, pp. 1057–1062.

Tekieh, M. H., Raahemi, B., 2015, "Importance of Data Mining in Healthcare," in: Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. 2015 - ASONAM '15, pp. 1057–1062.

Ting, S.L., Wang, W.M., Kwok, S.K., Tsang, A.H.C., Lee, W.B., 2010. RACER: Rule-Associated CasE-based Reasoning for supporting General Practitioners in prescription making. Expert Syst. Appl. 37 (12), 8079–8089.

Tortorella, G.L., Saurin, T.A., Fogliatto, F.S., Rosa, V.M., Tonetto, L.M., Magrabi, F., 2021. Impacts of Healthcare 4.0 digital technologies on the resilience of hospitals. Technol. Forecasting Social Change 166. https://doi.org/10.1016/j.techfore.2021.120666.

Van Poucke, S. et al., 2016. Scalable predictive analysis in critically Ill patients using a visual open data analysis platform. PLoS One 11 (1).

Wan, K., 2016, "Characteristics and Classification of Big Data in Health Care Sector," pp. 1439–1446.

Wang, J., Fang, H., Carreiro, S., Wang, Honggang, Boyer, E., 2017, "A new mining method to detect real time substance use events from wearable biosensor data stream," in: 2017 International Conference on Computing, Networking and Communications (ICNC), pp. 465–470.

Wang, Yichuan, Kung, LeeAnn, Byrd, Terry Anthony, 2018. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol. Forecasting Social Change 126, 3–13. https://doi.org/10.1016/j.techfore.2015.12.019.

Wikipedia, "Apache_Spark," 2017. [Online]. Available: https://en.wikipedia.org/wiki/Apache_Spark. [Accessed: 28-Apr-2017].

Xu, W. et al., 2012. Proteomic characteristics of spermatozoa in normozoospermic patients with infertility. J. Proteomics 75 (17), 5426–5436.

Yang, L., Li, Z., Luo, G., 2016. MH-ARM: a multi-mode and high-value association rule mining technique for healthcare data analysis. In: 2016 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 122–127.

Zhou, D., He, Y., Kwoh, C. K., 2016, "Validating Text Mining Results on Protein-Protein Interactions Using Gene Expression Profiles," no. February 2016, pp. 580–585.