OCTA

6-8 February, 2020

ALECSO, Tunis -Tunisia

Multi-Conference OCTA

OCTA International Multi-Conference on:

Organization of Knowledge and Advanced Technologies

https://multiconference-octa.loria.fr/

Unifying the scientific contributions of the following conferences:

SIIE

8th. International Conference on:

Information Systems and Economic Intelligence

ISKO-Maghreb

Digital Sciences: impacts and challenges on Knowledge Organization

CITED

2nd. International Symposium on:

Advanced Technologies, Renewable Energies and Economic Development

TBMS

1st. international symposium on:

Big-Data-Analytics Technologies for Strategic Management: innovation and competitiveness

ORGANIZERS

Université de Tunis

ISKO-Maghreb

UNIVERSITÉ DE LORRAINE

IEEE Tunisia Section www.ieee.tn

المنظمة العربية للتربية والثقافة والعلوم

SPONSORS

Tunisia China Friendship Association

ATLAS

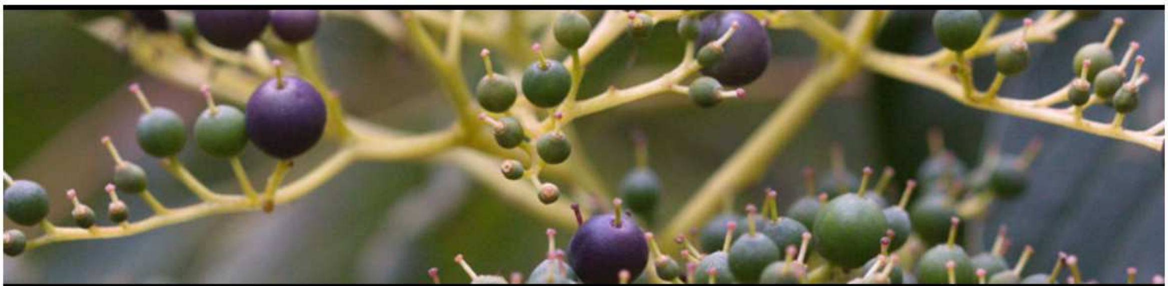
SPECIAL ISSUES JOURNALS & BOOKS

International Journal of Geo-Information

IEEE Xplore® Digital Library

TBMS '2019

Technologies Big-Data-Analytics pour le Management Stratégique : innovation et compétitivité. | Big-Data-Analytics Technologies for Strategic Management: innovation and competitiveness.



PROCEEDINGS

1st. international symposium TBMS’2019 on:

“Big-Data-Analytics Technologies for Strategic Management: innovation and competitiveness”.

November 21th. – 23th. , 2019 in Tunis (Tunisia)

https://tbms2019.loria.fr/

Organization in the event of the International Multi-Conference OCTA’2019 on: « Organization of Knowledge and Advanced Technologies » (https://multiconference-octa.loria.fr/)

Multi-Conference OCTA OCTA'2019 IEEE Tunisia Section www.ieee.tn UNIVERSITÉ DE LORRAINE ISKO-Maghreb

TBMS’2019 CFP (to download) & Special issues in Journals & Books

TBMS *“ Big-Data-Analytics Technologies for Strategic Management: innovation and competitiveness”:*

States like companies and organizations reorganize themselves in a serious way to the treatment of big masses of information: it is the advent of Big Data. As other economic, industrial and social actors radically reorganize their activities around the management of this data in Big Data. For example, « Chief data officer » in the United States, « Chief digital officer » responsible for the « Government Digital Service » in

Great Britain, « General Data Administrator » in France, etc. These public administrators are responsible for ensuring that the country values its data, presented as new sources of material and intangible growth unexplored or partially.

The same phenomenon of Big Data is at work in the academic world with more enthusiasm. In line with major research centers such as the Media Lab (1985) of the Massachusetts Institute of Technology (MIT), many universities have created centers dedicated to the analysis of Big Data in constant evolution with the « Data Science ». For the humanities, it is the birth of the « Big data and Society » which aims to study Big Data and their impact on societies « . These are the practical Big Data implications and how they reconfigure the relationships, the expertise, the methods, the concepts and the academic knowledge in the social, professional and business sectors.

If Big Data interest but also evoke multidisciplinary debate? The enthusiasm of the artisans of these techniques is not always shared. In this case, the repeated revelations of massive surveillance of populations, the research world as well of civil society highlight the dangers of these methods, which can quickly turn into powerful technologies of government: a conscious and instrumental will to serve economic, political-societal and ideological interests.

Today, we have more data than ever before in human history. Data volumes multiplied by 100 between 1987 and 2007, then doubled on average every year. An increase infinitely greater than that caused by the invention of printing (J. Gutenberg), which had resulted in a doubling of data over 50 years.

Thus, comprehensive data set analysis can change our view of the world. The contribution of Big Data remains a project that insinuates reason and rationality in our complex world. If the scientific method is based on the premise that one can derive from abstract theories concrete assumptions about reality. The assumptions themselves can be tested using the data collected for this purpose. In this reasoning, Big Data will profoundly alter these two rational foundations, not to challenge scientific rationality, but to move it to a higher stage: for a more complex, broader and more exact interpretation of reality.

In a transdisciplinary enthusiasm, Big Data allows us to no longer bend reality to categories a priori and now to let the data give us themselves the categories which contain while faithfully reflecting the reality.

TBMS'2019 International Symposium on « **Big-Data-Analytics Technologies for Strategic Management: Innovation and Competitiveness** » explores the practical implications of Big Data and how it reconfigures relationships, expertise, methods, concepts and academic knowledge in all sectors: social, professional and economic.

Committees

Steering Committee OCTA'2019 :

Honorary President
Pr. Habib SIDHOM, President of the University of Tunis (Tunisia)

Presidents
Sahbi SIDHOM (LORIA – Université de Lorraine, France)
Amira KADDOUR (ENSTAB – Université de Carthage, Tunisia)
Anass EL HADDADI (ENSA Al-Hoceima, Morocco)
Mohamed ADDAM (ENSAH Al Hoceima, Morocco).
Abdelkrim MEZIANE (Centre de Recherche CERIST, Algeria)
Davy MONTICOLO (ENSGSI – Université de Lorraine, France)
Xi LU (Tsinghu University, China & Harvard University, Cambridge MA, USA)
Saoussen KRICHEN (Vice-President Université de Tunis, Tunisia)
Khaled KCHIR (Vice-President Université de Tunis, Tunisia)

Program committee / Comité scientifique :

Presidents
Anass EL HADDADI (ENSAH Al Hoceima, Morocco)
Sahbi SIDHOM (Université de Lorraine & LORIA Lab., France)
Mohamed ADDAM (ENSAH Al Hoceima, Morocco)

Members
Ajith Abraham Machine Intelligence Research Labs (MIR Labs), Washington USA
Albert Zomaya School of Information Technologies, University of Sydney Australia
Anand J Kulkarni Symbiosis International, Deemed University India
Gupta B.B National Institute of Technology, Kurukshetra India
Bart Baesens KU Leuven Belgium
Bich Tran Centre for Health Research · University of Western Sydney Australia
Bonghee Hong Pusan National University Korea
BOUHORMA Mohammed FST, UAE, Tangier Morocco
BOUNCEUR Ahcene Bretagne University, Brest France
CHERBI Rachid Paris Sud Saclay University France
Douglas Craig C. University of Wyoming, USA
Eleni Karatza Aristotle University of Thessaloniki Greece
Faïcel Chamroukhi University of Caen France
Fazle Baki Odette School of Business, University of Windsor Canada
FENNAN Abdelhadi FST, UAE, Tangier Morocco
Francisco Herrera University of Granada Spain
Guo Q. P. Wuhan University of Technology China
Hamido Fujita Iwate Prefectural University Japan
KECHADI Mohand Tahar University College Dublin Irlande

Mengjie Zhang School of Engineering and Computer Science, Victoria University of Wellington New Zealand
MSIIRDI Nacer Aix Marseille University France
Pablo Jose Villacorta Iglesias Stratio, Madrid Spain
Péter Kacsuk MTA SZTAKI Hungary
QUAFAFOU Mohamed Aix Marseille University France
Roy Cecil IBM Portugal Portugal
Schahram Dustdar Vienna University of Technology Austria
TARI Abdelkamel Univeristé de Bejaia Algeria
Tsui Thomas Chinese University of Hong Kong China
Vincent Blondel Université catholique de Louvain Belgium
Xu W. Jiangnan University China
Yi-ke Guo Imperial College United Kingdom

FluSpider a new vision of digital influenza surveillance system: based on Big Data technologies and Massive Data Mining techniques

Kaouther Nouira
Université de Tunis
ISGT, BESTMOD
2000, Le Bardo, Tunisia
Email: kaouther.nouira@planet.tn

Nesrine Ben Njima
Uninersité de Tunis
ISGT, BESTMOD
2000, Le Bardo, Tunisia
Email : nesrinebennjima@gmail.com

***Abstract*—With the development of the number of Internet users, we have becoming able to quantify real world social phenomena up to date. One of the most important problems in healthcare is monitoring the current level of pandemic influenza. In this article, we propose FluSpider, a new vision of digital surveillance system that is based on tracking web pages followers, to estimate the current level of influenza across the world. Our method uses Big Data technologies and Massive Data Mining methods to provide accurate results. Our evaluation shows that FluSpider system can solve the Influenza Like Illness epidemic monitoring problem in near real time and ahead of two weeks of traditional centers.**

Keywords*—*Influenza; Influenza Like Illness; ILI; Information retrieval; Click monitoring; Big data technologies.

I. Introduction

Nowadays, the World Wide Web is the most powerful source of information. It is our everyday source for all kinds of information. The analysis of the person's behavior on the web gives us an idea about critical subjects. One of the most critical subjects in healthcare, is monitoring the current level of transmittable disease such as pandemic influenza.

According to the large amounts of data available on the web, we need Data Mining methods to get reliable information and Big Data technologies to collect, store and process data. In this paper, we introduce a near-real time influenza surveillance system. Our novel method, is based on tracking web pages related to influenza disease by estimating the number of people with seeking behavior. Health seekers have been defined as internet users who search online for information on health topics and more especially about influenza whether they are patients, doctors or caregivers.

In general, a high level of views in selected pages, covers a high level of population interest in this subject. It helps us to estimate the health state of the country through diseases interested by their population. These estimations will be verified by the results of traditional centers like the Center for Disease Control and prevention (CDC), World Health Organization (WHO) and the European Center for Disease Control and Prevention(ECDC).

The results of the system present the level and the trend of influenza epidemic for each country in a period of time. Our aim is to enhance the world health through providing a daily results, that helps decision maker to respond earlier and prevent the spreading of the disease.

Our implementation, can be presented in two main sub processes: Data Management and Data Analysis.

On the first hand, data management, consists in using methods that collect and store data to be analyzed. In this step, Nutch search engine, information retrieval concept and Click Monitoring tool are used to retrieve data that will be stored on databases.

On the other hand, data analysis, consists in using techniques to extract knowledge from data. In this step, we place measurements to explore the data such us aggregation and Geo-location measures.

The structure of this paper is as follows. Section 2 investigates FluSpider related literature. Section 3 focuses on preliminaries. Section 4 illustrates methods and materials used. Section 5 presents the evaluation of the proposed method. Section 6 contains the discussion and the future perspectives. Section 7 provides the conclusion.

II. Literature on digital disease detection

In this part, literature specifically concerned with digital disease detection is investigated. Several systems based on the internet were found to detect transmittable disease. Some of them are still existing and other failed in front of the high velocity of the web noisy data.

A couple months ago, Google flu trends (GFT) was existing and it breaks down this august. GFT was launched in 2008 by Google foundation. It provides estimates of influenza activity for more than 25 countries by aggregating Google search queries [1]. The first version of GFT, gives a good estimation of Influenza-Like-Illness (ILI) activity. It was very close to the CDC data. And this is the reason of it is failing, GFT was trained on a small size of the CDC data set, more lower than the data collection. So, the data were overfitting the testing set. Moreover, the search terms selected in the model by Google was not structurally related, and so do not predict the future. Over time, the search terms used by people were changed because people became more efficient in their searching task. Moreover, GFT missed the nonseasonal 2009 influenza A-H1N1 pandemic, also in 2012-2013 predicted the double of

CDC data. Many updates were achieved in GFT model, but it doesn't lead for accuracy results and it ends with his quiet death [2].
Recently, researchers team from Harvard published a new model named ARGO (AutoRegression with Google search data). It improves the GFT results. ARGO model correlates Google Flu Trends data with flu outbreak data collected by CDC to estimate the number of flu cases [3]. Many other attempts to create an online system to monitor influenza activity in their local country. We can note, a system that counts query submitted into the medical Sweden web site to predict the influenza activity in Sweden [4]. Counting the number of visits to US web site [5].Analyze the keyword clicks associated with influenza like illness in Canada [6]. Using Tweets to predict influenza activity in US [7]. Wikipedia usage to estimate influenza in US [8].
All these articles show reliable results compared to the CDC data. but we don't know the real time data, or who well they work at the current time.
Moreover, there are systems displayed online and give an overview of the flu spread in the country. We can mention, Flu tracking is an online system that detects epidemics of influenza in Australia. It sends a flu survey to voluntary participants weekly and displays the results on Australia Map[9]. Also, Flusurvey system has the same method, but collects the answers from British people and displays it in the England Map [10]. WebMD [11], Everyday Health Flu Map[12], Mayo Clinic [13] are US Maps. They predict the flu severity for each state. WebMD and Mayo Clinic displays reported data from sentinel. Everyday Health Flu Map estimates the current level of influenza, according to the social media, regional weather and CDC data.

III. PRELIMINARIES

In this section, web data extraction and information retrieval are delineated.

A. *Information Retrieval*

Information Retrieval defined by Salton in 1968, as the science concerned with structure, analysis, organization, storage, searching and dissemination of information[22].
IR is finding the best subset that responds to the user's information need in a collection of documents. It is about analyzing words in documents, organizing them by appearances number in a document or in a collection, storing them in memory or hard disk, then searching the best documents that satisfies the user query and return results to the user.
To retrieve documents with relevant information, there are three approaches, Keyword-based approaches, Knowledge based approaches and Learning based approach. These types are widely used in the healthcare field.
Keyword based approaches use the frequency of words in the document to rank data that match with the query. For instance, Corley et al counts the frequency of Flu Content Posts by using lexical that match with the flu (H5N1, H3N1, etc.) to predict influenza from blog posts.
Lee et al employ a list of diseases, symptoms, treatments manually prepared to count the frequency of the Co-occurrence of keywords observations on Twitter [24].
Greaves et al use the frequency of Sentiment keywords (joy, anger, sadness) in social media to detect poor clinical care [25].
In Knowledge based approaches, a knowledge base (ontology) is employed with a set of techniques where entities and relationships are more captured.
There are many researches based on knowledge approaches. For examples, using BioCaster Ontology and informal keywords bases to predict influenza from Twitter. These bases are used to filter data by BCO syndrome and influenza from the syndrome [27].
Also, detect from the forums the drug abuse behavior using Drug Abuse Onto and World Net. DAO lookup to extract semantic entities and relationships and Word Net to create synonyms [28].
Finally, Learning based approaches use the machine learning techniques to solve the ranking problem, where the training data consist of queries and sentences, where each query is associated with a number of sentences.
Many supervised algorithms are developed as the use of Support Vector Machine for ILI prediction [29][30] [31] and for depression prediction [32]. and also, we note the use of Naive Bayes for Suicide prediction [33] and the use of logistic regression for influenza prediction [34] and collaborative approach for emotion detection [35].

a) Inverted Index: Inverted Index is a data structure that creates for each term T a list of all documents that contain T. It identifies each document by DocID, a document serial number. The postings lists are stored on disk and sorted by DOCID later. The query browsing walks through lists of terms that are equal to user searching terms and selects the intersection between lists. This is the result of searching task.

b) Ranked Retrieval: Rather than a number of documents that match with query expression. Using models of rank retrieval, documents are weighted by their importance order in the collection compared to the submitted query.
The ranking algorithm is based on an adaptation of the classic IR vector space model [22]. The vector space model is based on boolean numbers. if a row contains "1", then the word appears in the document and if 0 then it does not. This score is assigned to each document, it represents how well the document matches the query.
Jaccard coefficient is a model of scoring based on basic vector space model. It takes the number of items in intersection and divides by the number of items in the union. It returns a score of the number of words that match with a query compared to the number of words in all the collection.
Scoring with Jaccard coefficient, classic vector space model, inverse index does not take in count the term frequency in a document. It means that, rare terms in a collection are more informative than frequent terms. Other measures put in place to consider the occurrences of the term in the document, such as Term frequency (TF) weighting, inverse document frequency (IDF) weighting and Term frequency-inverse document frequency (TF.IDF). TF.IDF is the combination between two rating retrieval models TF and IDF. It is the product of TF weight and IDF weight. The TF model represents a matrix where each document is a count vector of terms occurrences in the document. The vector representation does not consider the position of terms in a document. And IDF model, is an inverse measure where rare terms are highly score than frequent words.

B. Web Data Extraction

Web Data Extraction is an emerging field of Information Extraction.
Information Extraction is defined by Gaizauskas and Wilks as the activity of extracting a structured information source from an unstructured information source (mainly text) [20].
Within Web Data Extraction, the unstructured information source is the web. Where, it offers opportunities to access and collect a large volume of data continuously generated, shared and disseminated online with limited human interaction, by using a broad class of software applications [21].
These data could be extracted by element (title, authors, links, etc.), page content or the full text of the page itself. But to extract data from many pages of the website is by building a crawler. It is the best way to gather a lot of data from pages when a list of their URL's does not exist.
In most cases this task is based on Natural Language techniques and it is also related to IR. Because, Information Extraction does not understand the text as a whole, but aims to extract text from relevant elements. The analysis is performed locally, only parts of the text are considered and the type of information to be contracted is known in priory. In contrast to IR which returns information dealing with a subject expressed by a query whose type is not fixed a priory. And the analysis covers all the texts in the collection. Moreover, according to Gaizauskas and wilks, IR retrieves documents from collections where as IE extracts relevant information from documents [20].

IV. Material and Methods

In this section, the proposed method is illustrated. FluSpider is a digital system that uses web data to provide estimates about the influenza activity.
To implement FluSpider system, we proceed as follows:
First, we gather web pages with influenza subject from the web. To collect these data, we extract all terms related to influenza from the US national library of Medicine called "Mesh", which controls vocabulary thesaurus. We display a search query with the word Influenza on the search box and we get a list with all words related to this disease. The list includes the following terms:

- Influenza in Humans
- Influenza
- Human Influenza
- Human Flu
- Avian flu
- Grippe

These terms are translated into 29 languages, presented in the table below, using two translators ECTACO [14] an online multilingual dictionary and Babel [15]. Table I shows languages provided by ECTACO to translate influenza terms and table II shows languages provided by Babel. Words translation helps us to cover web pages with different languages and attend huge population from different places in the world. This sample gives us an idea about the trend of the flu in the world.

TABLE I: The translation languages of influenza related terms ECTATO

Arabic	Bulgarian	Czech	German	Greek
Spanish	Estonian	Persian (Farsi)	French	Hungarian
Italian	Yiddish	Korean	Dutch	Polish
Portuguese	Romanian	Russian	Serbian	Slovak
Albanian	Swedish			

TABLE II: The translation languages of influenza related terms using Babel

Ukrainian	Japanese	Hindi	Chinese	Danish

Second, we use the selected terms to search online for influenza related pages. The main page of search engine is the starting point for Internet surfers. which helps them to find documents that match with the submitted query. For different individuals the importance of the information is not the same. For this reason, we keep influenza general terms.
Third, we collect data related to influenza terms from web pages. We need a web crawler to extract data, which is spread out across many web pages. These data should be indexed to find pages related to ILI. So, a web indexer and a web searcher are used. The web indexer indexes documents through creating in inverted index, it is a presentation that helps the web searcher to evaluate the document and make search narrow. And, the web searcher deals with query processing, it counts the number of keywords in each document and returns a list of documents with higher score [16].
The web crawler, web indexer and web searcher are major components of a search engine. So, Nutch[17] is used. It is an open source search engine build on the top of Hadoop and provided by Apach foundation [18]. It includes all the search engines tools with offering the possibility to adapt them to our needs. It uses to collect data from the web and stores directly to HDFS, But instead of using Nutch's default database, we are using HBase to store crawled pages for real time access. HBase is the open source project of bigTable, which is developed by Google. It provides a downstream destination system accessible in real time and a large tables with extensible columns to store a lot of data [48].
The stored data are URLs and page contents. The URLs are tracked via a click monitoring tool. It is Java code based on Shell command, It monitors link clicks, an other words, web site access. It returns a list of user IP address and URLs of viewed web pages. The IP addresses are stored. These addresses are used to Geo locate the user, how are interested in influenza subject. The intention behind the Geo location task is to recognize epidemic level severity of each country, and in the end to have an overview of the world state.
IP is an identifier for a computer or device connected to internet network. And, Every particular region in the world has a list of IP address blocks allocated only for here [19].
So, to identify the location of these visitors, we download a database with IP address blocs by country from https://db-ip.com/db/. This database contains 150 countries from a total number of 192 countries on the world. Despite that, it is an accurate database, within updates are made monthly. The data presented on the database downloaded are the first and the last IP address allocated to each country.

To identify the location, we verify the user address, if it is between the two addresses, then this user is assigned to this country.
After recognizing the country of each user, visits can be added together per country, a day visit can include one or multiple web pages per internet user for this reason we use ip address as a constraint, that means once ip address is counted it is not re-added.
For the computation, we query the database. And, results export into a csv file.
and visualized through displaying charts(world map and lines). The map is colored by red shades, from the lighter to the darker. Each color represent a warning level. If the number of influenza web pages follower is higher than 75% of internet user number of that country, then it will be highlighted in dark red to mention the influenza spread danger and so.
The swine flu (H1N1) and Bird flu (H5N1) are subtypes of influenza diseases. Both are a communicable disease and have the same symptoms of regular human flu, but, the distribution of these viruses can vary from geographic area to another and between time of year. According to the similarity between influenza and their two subtypes. We take the data that we collected previously with influenza terms and we fetch H1N1 and H5N1 related words. These words are also extracted from Mesh dictionary. We translate these words into multiple languages to retrieve related information. The content of web pages related to influenza disease is already stored. As explained below, H1N1 and H5N1 are similar to ILI, but both of them have their own words. For this reason, we used influenza document collection to retrieve content related to these epidemics. That lead us to use Information retrieval technique and the rank measure TF.IDF to get relevant information from our collection of text.

Terms related to H1N1 disease are:

- Influenza A virus H1N1 subtypes
- H1N1 virus
- Influenza pandemic 1918-1919
- H1N1 pandemic 2009
- H1N1 pdm09
- Swine influenza
- Swine flu

Terms related to H5N1 disease are:

- Influenza A virus H5N1 subtype
- H5N1 virus
- Bird Influenza
- Bird flu
- Avian flu
- Avian influenza

V. Results

In this Section, the results are evaluated. Our analysis is based on 15 websites in each languages. It includes different websites (forums, blogs, medical websites, Wikipedia, etc). There are the most ranked websites by Google and Baidu for Japanese websites.
After running Nutch with terms related to influenza. It crawls the websites and extracts 64 620 web pages related to influenza subject. The data waves weighted almost 1,4 Terabyte. As we noted before, those terms are in different languages.
The translation is done using online software's. Terms are general and few. For that reason, we examine each term translation and we adjust the translation by adding an other words or change the translation or even the term structure. During this process, we find that "Human in Influenza"," Human influenza" or "Human Flu" have the same meaning and it could be represented in distinct words for better translation as "Influenza", "Human", "Flu". But, we need to construct an expression that will be used to extract web pages related to influenza. So, we add the preposition "in" to get the meaning for "Influenza" and "Flu" combined with "Human" term. For some languages Like Arabic, French, German, Persian (Farsi), Greek, Korean, Polish, Swedish, Turkish, Romanian ,Russian, The translation of these three words do not have any meaning. So we add the word "Body", to get the following expression "Influenza" + "in" + "Body" + "Human". Moreover, in some languages "influenza" and "Flu" terms have the same translation word, so we restrict one of them. We produce 189 different terms with 36 restriction and 12 combinations.
Moreover, for the sub-types of influenza (H1N1 and H5N1), the terms are divided into words and translated after. For example, " Influenza A Virus H1N1 subtypes " is the translation of each word by itself. The acronyms of influenza sub types like "H1N1" and "H5N1" are not translated, because these words have a technical use and do not change with languages. The translation of H1N1 and H5N1 related terms produce 120 different terms without any restriction.

The best way to evaluate the ranked results, is by calculating two metrics the precision and the recall. To do that, we took a random sample of 10 URL's collected previously by Nutch and stored in the HBase. These URL's are in English and associated into different terms related to influenza subject. Then, we displayed these terms, "Influenza", "Flu", "Human influenza", " influenza in Human", "Influenza in body Human". Also, we calculated the recall and precision functions. The relevant pages that match with the query are those who have an informative nature. That means, that they include definition of influenza or their symptoms or protection measures, etc. According to the definition of relevant page. we check our database and we find 6 relevant pages. Our system is configured to download the content retrieved of the page and its URL in the HBase, so we change the path of Nutch storage into a text file.
Recall *P(retrieved/relevant)* is the fraction of relevant items in the collection that are retrieved.
Recall = Relevant items retrieved /Relevant items.
Precision *P(relevant/retrieved)* is retrieved items that are

relevant to the user query's.
Precision= Relevant items retrieved / Retrieved item.

TABLE III: Evaluation of the ranking models of influenza related terms

Influenza related terms	Recall	Precision
Influenza	6/6	6/10
Flu	1/6	1/7
Human influenza	5/6	5/8
influenza in Human	5/6	5/8
Influenza in body Human	4/6	4/7

Moreover, the recall number is close to "1" more our system is efficient in finding documents that match with the query. For "Influenza" term we get a recall value equal to 1 because all documents found are relevant to the query. But, the precision value is lower than the recall value due to the increase of the number of retrieved documents. Our system detect all documents with influenza subject, the 10 retrieved documents are the 10 selected for testing. In the flu terms, we note that the recall number is inferior than the precision number ($0.16 < 0.17$). this result is explained by, the flu related documents found are more important than the number of relevant documents presented in the collection ($6 < 7$). With the other terms, we note that the recall number is superior than the precision number. It is caused by, display the system an important part of relevant documents in the collection.

The H1N1 and H5N1 pages are retrieved in the influenza collection. To evaluated, we proceed as follow, the sample of 10 URL's previously chosen for evaluate the raking retrieval model of influenza is still used. Within this sample we will find documents that match with H1N1 and H5N1 terms. The H1N1 testing set are "H1N1 virus", "H1N1pdm09", "Swine Influenza" . And the H5N1 terms are "H5N1 virus", " Influenza A virus H5N1 subtype", "Bird flu".
The number of relevant content in the collection is 4 documents H1N1 and 3 documents H5N1 from 10 documents in the collection.

TABLE IV: Evaluation of the ranking models of H1N1 related terms

H1N1 related terms	Recall	Precision
H1N1 virus	3/4	3/7
H1N1pdm09	0/4	0/1
Swine Influenza	3/4	3/4

The Recall value is more important than the precision value ($3/4 > 3/7$). Also, the number of retrieved documents with "H1N1 virus" term is superior than relevant documents in the collection ($7 > 4$). That means, that documents retrieved by our system do not match perfectly with our need, but an important number of relevant pages contents are detected, where 3 relevant documents found from a total number of 4 documents in the collection. the four documents selected as relevant, they satisfy our need to detect H1N1 seeker behavior, tree pages from four pages include the term "H1N1 virus".
For "H1N1pdm09" , the system found only one document in the collection that respond to the query but it is not qualified as relevant. So, "H1N1pdm09" is a rare term in the collection, that it does not represent perfectly H1N1 subject.
For "swine influenza" term, the recall value is equal to the precision value ($3/4 = 3/4$). But, The number of relevant documents found is lower than the number of relevant documents in the collection ($3 < 4$).
We conclude that "H1N1 virus" and "swine flu" terms are pertinent query expressions. They respond to the user's need. In contrast to "H1N1 virus", "swine flu" is more accurate, where it has the highest precision value ($0.74 > 0.428$).

TABLE V: Evaluation of the ranking models of H5N1 related terms

H5N1 related terms	Recall	Precision
H5N1 virus	2/3	2/7
Influenza A virus H5N1 subtype	1/3	1/2
Bird flu	2/3	2/4

Within H5N1 related terms, we note that "H5N1 virus" term has a recall value more important than the precision value, as the case in "H1N1 virus". With a high number of retrieved documents. This can be explained by the presence of this term in 7 document of 10 documents in the collection. "H1N1 virus" and "H5N1 virus" are highly related to influenza subject (more than 70% of documents contain those words). Although "Influenza A virus H5N1 subtype" term includes the therm " virus H5N1", but the system found only 2 documents that match with the query less than the result returned by "H5N1 virus", and this could be resulted by the ranking model TF.IDF, where it ranks frequent words with low value than rare words. And because, "virus","H5N1" and "Influenza" are frequent words, there are assigned with a less important value than "subtype".
We note that all terms detect relevant documents. and this reflected the accuracy of those those expressions.

After evaluating the accuracy of pages retrieving. Now, we evaluate the performance of the system.
We visualize our map and we note that 63 countries on the world are keeping track of influenza pages. Some of them have an important number of follows than other countries. Based on the results of the last three weeks of 2015, their accuracy will be examined, down below, using influenza surveillance data collected by CDC centers and World Health Organization.

On the map, we can notice four groups based on colors: red, dark orange, orange, light orange and light yellow. Each color of them represents a class of risk. The red color means a severe state in those countries, which has more than 75 percent of internet users that are following influenza subject online. We notice that in these weeks, Russian Federation is in a severe state.
The highest class is represented by the dark orange color. This

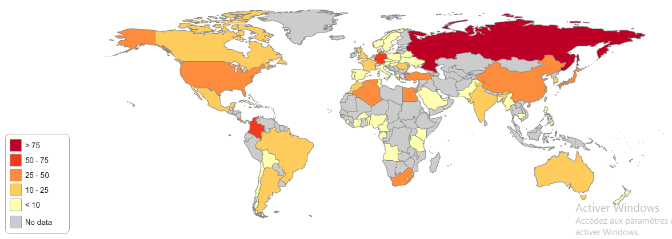


Fig. 1: Map visualization of the spread of influenza across countries during Week 50, 53, 52 and 53

class contains countries with a number of followers between 50 percent and 75 percent. This class includes Germany and Colombia.
The orange class is assigned to countries with moderate level. We qualify a country as moderate if the number of followers is superior than 25 percent and less than 50 percent. Here, we note United States, China, Bahrain, Hong Kong, Turkey, Algeria, Romania, Egypt, Japan, Costa Rica and South Africa. Canada, Morocco, Mexico, Brazil, Armenia, Argentina, India, Australia, Denmark, Republic of Korea, France, United Kingdom and Lebanon. These are countries with an intermediate level, where the number of person tracking influenza subject on the web is less than 25 percent and higher than 10 percent.
In the final class, we mention countries with an activity lower than 10 as Bolivia, Sierra Leona, Ivory Coast, Burkina Faso, Austria, Nigeria, Benin, Togo, Cameroon, Gabon, Angola, Tanzania, Kenya, Saudi Arabia, United Arab Emirates, Italy, Spain, Portugal, Belgium, Norway, Sweden, Poland, Czech Republic, Greece, Pakistan, Bangladesh, Myanmar (Burma), Cambodia, Tunisia, Slovakia, Slovenia, Belarus, Cuba, Chile, Philippines and Ukraine. These countries involve into lower class and there are colored by light yellow.
All countries are ordered by their risk level in Fig 2.

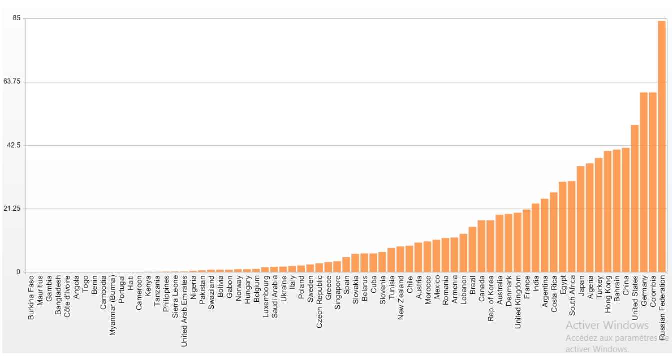


Fig. 2: The trend of influenza on the world during Week 50 and 53

Looking at the trend of influenza in the world, we notice that the Russian Federation has the highest ILI level. Where, the average number of influenza page's followers per week is estimated 84.2 percent of internet users. The WHO National Influenza Center in Russia mentioned an increase of influenza activity in Russia during the week 53 compared to the week 50 [36]. As we mentioned too, the level of influenza raised from 71.5 million to 99.5 million followers in the last week.
The Russian people are more interested about the H5N1 influenza sub type than the H1N1 sub type. Within 100 internet users, 22 persons are looking online for H5N1 related information per week. Although, only 5 persons from 100 are followed up H1N1 web pages per week.
Seeing both figures Fig. 5 and Fig. 7, we find that, Russian Federation is ordered in the second position after Egypt by their H5N1 followers number and in the 22 position by their H1N1 number. In spite of the fact that National Influenza Centers (NICs) reported cases of A(H1N1)pdm09 in these weeks, we notice that people still worried about avian flu that is highly spread in Russia (Fig. 8).
Germany and Colombia are in the second row have the same level of risk with an average of 60.2 followers.
The influenza activity in Germany is increasing. It grows up with an average of 0.03 every day on these weeks, except weekends, where the activity level has a low decrease. The H1N1 level in Germany is higher than H5N1 level, 19.8 percent of internet users follow H1N1 influenza related pages per week versus 8.2 for H5N1 pages.
In contrast, influenza activity in Colombia is decreasing. The H5N1 level is more important than H1N1 level. There are 12.5 percent of Colombian internet users are followed web pages with H5N1 as a subject against 3.9 percent of H1N1 pages.

The trend of influenza in United States is increasing slightly. The H5N1 related pages are more accessed than H1N1 pages. In America, there are an average of 20 million views per day. It includes 8 million views to H5N1 web pages and 7 million to H1N1 related pages .
CDC reported also a slightly increase of influenza activity in US, during those weeks. Moreover, it identifies 7 regions from 10 with elevated activity.
We note also an important activity in china. An average of 42 from 100 persons consult pages related to ILI per week. The H5N1 activity is very spread in it. Where, we notice an important interest about theirs related pages. More than half of influenza views are related to avian flu. The number of followers of H1N1 related pages is very low almost 1 percent per week despite to the reported results of Chinese National Influenza Center and WHO, which many cases are detected with H1N1 sub-type (Fig. 3).
Bahrain is also in sever situation, the ILI activity is decreasing. It passed from 36 thousand views in week 50 into 15 thousand views in week 53. The most pages viewed are pages related to H1N1 influenza disease. There are 14.8 percent of views related to H1N1 against 8.5 are related to H5N1. WHO reported also a decline in the influenza activity during those weeks [39] and a high proportion of H1N1 outbreaks (Fig. 3).
The activity level in Hong Kong is increasing slightly during two weeks. It is going from 40.6 percent in week 50 to 44.2 percent in week 53. With almost 5 million views in these weeks. We note an increase in avian flu and swine flu. The avian flu passed from 13.2 percent to 26.4 percent, and the swine flu rise from 11 percent to 22 percent during

those weeks. Center for Health and protection in Hong Kong reported an increase on the influenza activity, which rose by 38 outbreaks in the last weeks. Cases detected in those weeks have C, B, A(H3) and A(H1) influenza sub types [40][41], But in Fig. 8 we can notice a high level of H5N1 in china and Hong Kong.
Many outbreaks were recorded in Turkey in 2015. The ECDC reports 127 outbreaks from 188 positive influenza cases between Israel and Turkey. The level of influenza is increased during these weeks. FluSpider system detects also this increase, which expands from 22.9 percent of influenza followers in week 50 to 53.5 in week 50. The majority of reported cases by ECDC are sub typed A(H1N1). 40 percent of views detected by our system are also related to H1N1 sub type and 24 percent of views are related to H5N1 influenza sub type. The avian flu is highly spreadable in Turkey (Fig. 8) as the swine flu (Fig .3). At the end of the year, A(H1N1) gets more interest than A(H5N1).

The influenza activity in Algeria is increasing. 36.4 percent of internet users follow influenza pages per week. During 13, December to 9, January, it rose by 2.3 percent. The proportion of H5N1 pages followers is 14.4 percent per week. There is no great disparity between the proportion ratio of H1N1, which equal to 14.5 percent. Our information sources, WHO and CDC centers do not report the influenza activity in this period, but a number of deaths are reported by Algerian media [43][44]. Where, four deaths are recorded in December, 18.
In western Asia countries such as Japan, China, Australia, Cambodia and republic of Korea, the influenza activity is increasing. Like we explained before for China. The most viewed pages are related to H5N1 disease except for Republic of Korea and Australia, Which people are more interested in H1N1 related pages than H5N1. In contract, of the reported results by WHO. Where, they mentioned that the influenza activity in this area is increasing due to H1N1 and B viruses [39]. This is can be explained by the sensibility of people about H5N1 subject.
The influenza activity in South Africa is increasing slightly. Where, it increases by 300 thousand views during week 51 to week 53. The number of H5N1 consultations is higher than H5N1. The average rate of swine flu related pages views per week is equal to 8.5 percent. However, it is only 9.9 percent for avian flu.
The highest level of avian flu is in Egypt (Fig.6). There are almost 31 people among 100 follows pages related to H5N1 influenza sup type per week. Against, an average of 10 viewed to H1N1 pages in the same period. The influenza level is moderate in Egypt as the case in Costa Rica. There, the number of influenza tracker is almost stable. It saves the same trend. During the week the number of views increases, but it is reduced at the end of the week. We record the same behavior for Argentina, Mexico, Romania. WHO report also an important activity in Costa Rica contrariwise to central America, the area that it belong to. Where, the activity is estimated low [39].
In France, we notice a high interest about the influenza subject. This interest is increasing by time. There are 7.6 million people looking for influenza pages online during the first studied week and it raises to 15.9 million views at the end of the period. The most viewed pages are related to avian Flu, it is equal to 7.8 percent per week against 6.8 percent to H1N1 pages. The same trend is recorded in Italy, Belgium, Spain, Slovenia, Austria, Czech Republic, Poland, Denmark, Slovakia, Ukraine, UK, Belarus. Which, the activity is increasing but it still low. Even, WHO report a low activity in Europe during 2015[39]. The Sentinel surveillance of influenza-like reported a low intensity activity in this geographic area except for France, Italy and Norway, where the influenza spread is more important that other places[44][45][46].

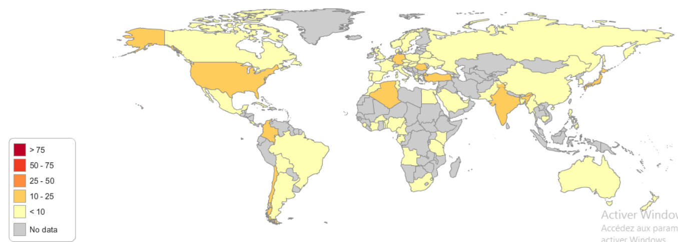


Fig. 3: Map visualization of the spread of H1N1 disease across countries during Week 50 and 53

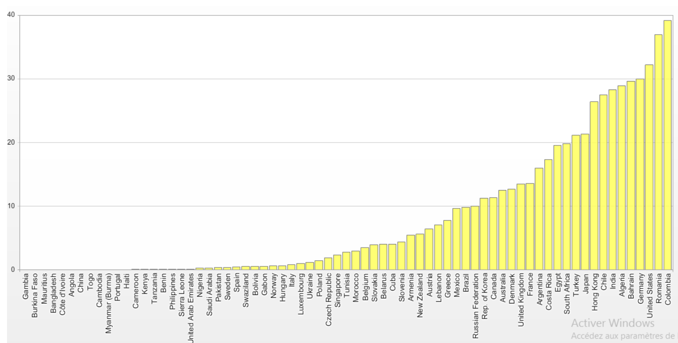


Fig. 4: The trend of H1N1 disease on the world during Week 50 to 53

We note an increase of H1N1 activity in the most of countries and specially at the end of the second week. Although there is not any country detected with influenza emergency have an important swine flu level. Some of them are in a normal situation as Colombia, Romania, United States, Bahrain and Germany, Turkey and Hong Kong. The others countries as Russia Federation and china have a low risk.
The list of normal risk includes other countries such as Algeria, Chile and India.
Countries with a low H1N1 risk are Canada, France, Egypt, Australia, South Africa, United Kingdom, Denmark, Japan, Republic of Korea, Argentina, Costa Rica, so on.

In three studied weeks, the trend of avian flu is decreasing in most countries except in France, Turkey and India, where it is increasing slightly. The rest countries in Northern and South West Europe, H5N1 trend remained nearly stable at a very low level. Some countries have an interest level higher than other countries as Egypt, Russian Federation, United States, China, Colombia, Germany, Chile, Romania, Algeria (Fig.5)(Fig.6). ECDC has already detect an avian spread in several countries

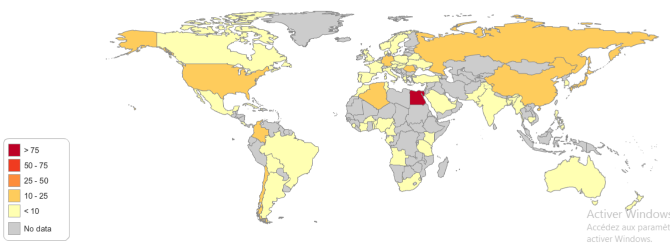


Fig. 5: Map visualization of the spread of H5N1 disease across countries during Week 50 and 53

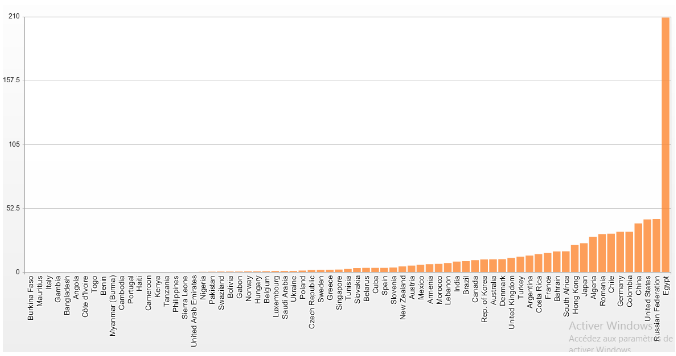


Fig. 6: The trend of H5N1 across countries within week 50-week 53: from December 6,2015 into December 26,2015

detected with normal activity as Egypt, United States, Russian Federation and China also in turkey, India and France that could explain the increase of followers numbers unlike other countries (Fig.7).

Fig. 7: World-map reported by ECDC representing countries with avian influenza A(H5N1) outbreaks in 2014-2015 [38]

VI. Discussion

In spite of the high spread of influenza in some countries of middle Africa, Western Africa and Asia, there are few records of web access activity to flu pages.
For instance, the visiting number of influenza pages in Benin does not exceed 7 visits during the whole studied period. Also in Nigeria, where the avian flu is widely transmitted. That could be explained by a poor internet network coverage where, a minority of people can uses internet. In Benin for example 5.3 percent of population have internet. In the other hand, some countries have an important activity spite of the low internet coverage. As, Algeria where 18.09 percent of population have internet network and have a moderate influenza level. So, there are other factors who influence the influenza activity as Human Development Index (HDI). This is a score attributed to each country. It represents the education level, life expectancy and Income per worker.
We notice that countries with Low human development are not interested to search about influenza subject and get deeper knowledge. In these countries, the influenza result does not present the current state of the country.
For high and moderate human development, our system gives a good estimation of the influenza activity and in two weeks ahead national influenza centers(NICs).
Some times, people are more interested about some subject than other. Like the case in China, Russian Republic and Germany. That could provide inaccurate information.

VII. Conclusion

In this paper, we propose FluSpider, a surveillance system that is based on internet data, to estimate the current level of influenza across the world. It helps to enhance the world health by providing a daily information. Previously, many methods have been developed in order to detect ILI disease. However, no one of them has handled several countries in the same platform and time. This is why our system come into prominence. Indeed, it is the first one that can cover the limitation mentioned above. FluSpider, cover 63 countries on the globe. We use terms extracted from medical dictionary and translated into different languages to access to multiple influenza pages spread on the web. Our method tracks these web pages and records the IP address of their followers. These addresses will be geo-located and displayed on the map after counting the total number of views per day in the country. The results are adjusted by the number of internet users. In the other hand, we use Information Retrieval (IR) to find pages related to influenza sub-types (Avian flu (H5N1) and Swine flu (H1N1)) in our corpus. Our corpus is a collection of web content and URLs extracted previously from the web using terms translated and collected by a crawler. To collect, handle and process this data we are using a Big Data solution, Nutch. Our evaluation shows that FluSpider system can solve the ILI epidemic monitoring problem in near real time and ahead of two weeks of traditional centers.
We test the results of three successive weeks, and we compared with the results of Influenza surveillance centers as World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), European Center for Disease Control and Prevention (ECDC), National centers and media for countries where data is not available at the studied period.
We find that our system provides an accurate results, 82 percent of countries have the same trend as data published by Influenza surveillance centers. Moreover, the information is presented daily and it covers many places that traditional centers does not report their activity as Japan, Algeria, South Africa, etc.

Otherwise, the proposed method saves not only human efforts, but also money and time and it can be used used for other problems as well.

Acknowledgment

The authors would like to acknowledge English technical help support provided by Mr. Durga Viswanath Gadi.

References

[1] Jeremy. Ginsberg and Matthew H. Mohebbi and Rajan S. Patel and Lynnette. Brammer and Mark S. Smolinski and Larry. Brilliant, *Detection influenza epidemics using search engine query data*. Nature Vol 457, 2009.

[2] Lazer. David and Ryan. Kennedy and Gary. King and Alessandro. Vespignan, *The Parable of Google Flu: Traps in Big Data Analysis*. Science Vol 343, 2014.

[3] A. Hulth and G. Rydevik *Web query-based surveillance in Sweden during the influenza A(H1N1)2009 pandemic, April 2009 to February 2010*. Euro Surveill , 2011.

[4] Philip M. Polgreen and Yiling Chen and David M. Pennock and Forrest D. Nelson, *Using Internet Searches for Inuenza Surveillance*. HEALTHCARE EPIDEMIOLOGY, 2008.

[5] Gunther.Eysenbach, *Infodemiology: tracking flu-related searches on the web for syndromic surveillance*. AMIA Annu Symp Proc, 2006.

[6] Alessio. Signorini, Alberto. Maria. Segre, Philip. M. Polgreen, *The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic*. PLoS ONE 6(5), 2011.

[7] David J. McIver and John S. Brownstein, *Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time*. PLoS Comput Biol 10(4), 2014.

[8] Shihao. Yanga and Mauricio. Santillana and S. C. Koua, *Accurate estimation of influenza epidemics using Google search data via ARGO*. pnas vol.112 no.47, September 30, 2015.

[9] Anonymous, *http://www.flutracking.net/*. accessed date Dec-21-2015.

[10] Anonymous, *https://flusurvey.org.uk/*. accessed date Dec-21-2015.

[11] Anonymous, *http://symptoms.webmd.com/cold-and-flu-map-tool/*. accessed date Dec-21-2015.

[12] Anonymous, *http://www.everydayhealth.com/flu/map/*. accessed date Dec-21-2015.

[13] Anonymous, *http://www.mayoclinic.org/diseases-conditions/flu/in-depth/flu-tracker/itt-20113960*. accessed date Dec-21-2015.

[14] Anonymous, *http://online.ectaco.com/*.

[15] Anonymous, *https://www.babelfish.com/*.

[16] Sergey. Brin and Lawrence. Page, *Reprint of: The anatomy of a large-scale hypertextual web search engine*. Computer Networks 56(38253833), 2012.

[17] Rohit. Khare and Doug. Cutting and Kragen. Sitaker and Adam. Rifkin , *Nutch: A Flexible and Scalable Open-Source Web Search Engine*. CommerceNet Labs Technical Report 04-04, 2004.

[18] R. Housley and J. Curran and G. Huston and D. Conrad , *The Internet Numbers Registry System*. Internet Engineering Task Force (IETF) , 2013.

[19] Robert. Gaizauskas and Yorick. Wilks, *Information extraction: beyond document retrieval*. Journal of Documentation 70-105 (54) , 1998.

[20] Robert. Baumgartner, Wolfgang. Gatterbauer, Georg. Gottlob, *Web Data Extraction System*. ENCYCLOPEDIA OF DATABASE SYSTEMS , 2009.

[21] G. Salton and M. J. Mcgill, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[22] Corley. C. D and Cook. D. J and Mikler. A. R and Singh. K.P*Using Web and Social Media for Influenza Surveillance*. Advances in computational intelligence. , 2010.

[23] Lee. K and Agrawal. A and Choudhary. A *eal Time Disease Surveillance using Twitter Data: Demonstration on Flu and Cancer*. KDD13,Chicago Illinois, USA, 2013

[24] Greaves. F and Ramirez-Cano. D and Millett. C and Darzi. A and Donaldson. L *Harnessing the cloud of patient experience: using social media to detect poor quality healthcare*. BMJ Quality and safe Online First, 2013.

[25] Doan. S and Ohno-Machado. L and Collier. N *Enhancing Twitter Data analysis with simple semantic filtering: Example in tracking Influenza like illness*. University of California, San Diego, USA, 2012.

[26] Sokolova. M and Jafer. Y and Matwin. S and Schramm. D *How Joe and Jane tweet about their health: Mining for Personal Health information on Twitter*. In Proceeding of recent advances in Natural Language Processing. Hissar, Bulgaria. pp. 626-632, 2013.

[27] Cameron. D and Smith. GA and Daniulaityte. R and al*PREDOSE: A Semantic Web Platform for Drug Abuse Epidemiology using Social Media*. Journal of biomedical informatics, 2013.

[28] Achrekar. H and Gandhe. A and Lazarus. R and Yu. S and Liu. B*Twitter improves seasonal in Influenza prediction*. Department of Computer Science. University of Massachusetts Lowell, Massachusetts, USA, 2012.

[29] Aramaki. E and Maskwa. S and Morita. M*Twitter Catches the flu: detecting influenza epidemics using Twitter*. In Proceedings of 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK. pp 1568-1576, 2011.

[30] Sadilek. A and Kautz. H and Silenzio. V*Predicting Disease Transmission from Geo Tagged Micro Blog Data*. Proceedings of the Twenty-Sixth AAAI Conference on Artficial Intelligence. pp. 136-142, 2012.

[31] De Choudhury. M and Gamon. M and Counts. S and Horvitz. E*Predicting Depression via Social Media*. Association for the advancement of Artificial Intelligence, 2013.

[32] Abboute. A and Boudjeriou. Y and Entringer. G and Az. J and Bringay. S and Poncelet. P*Mining Twitter for Suicide*. University of Montpelier, 2013.

[33] Culotta. A*Toward Detecting In Influenza Epidemics by Analyzing Twitter Messages*. In first workshop on Social Media Analysis (SOMA10), Washington. USA, 2010.

[34] S. Tuarob and W. Chu and D.Chen and C.S.Tucker*TwittDict: Extracting Social Oriented Keyphrase Semantics from Twitter*. Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction, pages 2531,Beijing, China, July 30, 2015

[35] S. Tuarob and W. Chu and D.Chen and C.S.Tucker*TwittDict: Extracting Social Oriented Keyphrase Semantics from Twitter*. Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction, pages 2531,Beijing, China, July 30, 2015

[36] *INTEGRATED DATA OF INFLUENZA MORBIDITY AND DIAGNOSIS Week 53*. WHONIC at Research Institute of Influenza and D.I. Ivanovsky Institute of Virology, December 20, 2015

[37] European Centre for Disease Prevention and Control *Rapid risk assessment Situation overview:highly pathogenic avian influenza virus A of H5 type*. December, 4 2015. Stockholm: ECDC; 2015.

[38] World Health Organization *Influenza Update N 253 WHO Report* . December, 28 2015: WHO; 2015.

[39] Center for Health and protection in Hong Kong *FLU EXPRESSS*. VOLUME 12, NUMBER 51 (PUBLISHED ON Dec 24 2015) .

[40] Center for Health and protection in Hong Kong *FLU EXPRESSS*. VOLUME 12, NUMBER 51 (PUBLISHED ON Dec 31 2015) .

[41] World Health Organization *Influenza transmission zones*. WHO 2011, Last updated: 2011-03-15 .

[42] FluTrackers.com *https://flutrackers.com/forum/forum/middle-east/h5n1-tracking-ac/algeria-ad/742977-algeria-media-report-regarding-a-suspect-bird-flu-h5n1-death-in-tebessa-moh-seasonal-flu-pandemic-h1n1-09* . accessed date: January, 17, 2016 .

[43] elkhabar newspaper *http://www.elkhabar.com/press/article/96997/*. accessed date: January, 17, 2016.

[44] European Centre for Disease Prevention and Control (ECDC) *CDTR Week 51, 13-19 December 2015*. .

[45] European Centre for Disease Prevention and Control (ECDC) *CDTR Week 52, 20-26 December 2015*. .

[46] European Centre for Disease Prevention and Control (ECDC) *CDTR Week 53, 27 December 2015-2 January 2016*. .

[47] White, T *Hadoop : The denitive guide [Book]*. OReill, 2015

Analysis of breast cancer data: a comparative study on different feature selection techniques

Linda Ouerfelli[a,*], Kaouther Nouira[a], Ahmad Ferchichi[a,1]

[a] *Université de Tunis, Institut Supérieur de Gestion de Tunis*
LR99ES04 BESTMOD, 2000, Le Bardo, Tunisia

Abstract

Choosing the relevant features is important to provide a better understanding of the data and improve the prediction performance. In this paper, we present a comparative study of various feature selection methods applied on a breast cancer dataset. In addition, this work investigates the stability of these techniques when perturbation on the dataset is added. Artificial Neural Network and Random Forest are used for classification. The results are compared when using all the features and when using only the top ranked. The classification performance are comparable in either cases.

Keywords: Feature selection, Stability analysis, Classification, Breast cancer

Introduction

Breast cancer is the most common cancer affecting women worldwide, with nearly 1.7 million cases diagnosed in 2012 [1]. However, breast cancer has a low fatality rate. Survival rate have been increasing mainly because of medical advances allowing for an early diagnosis and a personalized treatment. The disease can be cured if detected before symptom's development. When detected at an early stage, the five years survival rate is almost perfect [2].

Because of the significant incidence of this cancer, extensive research continues to be conducted worldwide to identify causes and solutions. Although, no specific factor is associated with breast cancer development, some risk factors are common in prediction models namely age, family history of breast cancer and age at first birth. More recent models include other identified factors namely breast density, use of hormone therapy and body mass index (BMI). Most of the studies employ clinical, genetic and histological or any combinations of these parameters as input to the prognostic procedure. Identifying the risk factors of breast cancer will help in the prevention by targeting high-risk individual. For example, to reduce the risk minimizing alcohol consumption, increasing physical activities and maintaining a

*Corresponding author
Email address: linda.ouerfelli@gmail.com (Linda Ouerfelli)

healthy weight are advised [3]. Adding these modifiable factors might help to know the extent to which they affect the risk of breast cancer. There are three methods to detect cancer namely: mammography, Fine Needle Aspiration (FNA) and surgical biopsy. The accuracy of each is depicted in table 1. For an early diagnosis, mammography is often used but not all cancers are detected and there is the risk of a poor-diagnosis. Although, the screening may result in false-positive, over-treatment or over-diagnosis, the early detection saves lives.

Table 1: Accuracy of methods to detect breast cancer [4]

Methods	Accuracy
Mammography	$68-79\%$
FNA	$65-98\%$
Surgical Biopsy	100%

Scientists have developed strategies for cancer prediction and prognosis. Plus, large amounts of data have been collected for research. However, the accurate prediction of disease susceptibility, recurrence or survival is still problematic. An obvious trend in the proposed models includes the use of machine learning methods. As these techniques permit the discovery of patterns and relationships in complex datasets. According to [5], the application of machine learning improved the accuracy of cancer prediction by 15-20 % in the last years. Indeed, various feature selection and machine learning techniques have been used for breast cancer prognosis and prediction.

The paper is structured as follows. Section 1 provides some background on breast cancer risk models. Sections 2 and 3 introduce the concepts of feature selection and feature stability respectively. Next, the feature selection methods used in the testing along with the stability metrics and the classification algorithms are presented. In section 5, an analysis of the dataset used for testing is given and the experimental results are illustrated in section 6.

1. Risk Models for Breast Cancer

A risk prediction model is a statistical tool used to estimate that within a time period an individual with certain factors will develop breast cancer. The factors can be any combination of genetic, environmental or clinical. There are two main statistical models used in the literature namely logistic regression and COX portional hazards regression. They are used to identify and score a variety of different risk factors.

Scoping the literature, there are many reviews of risk prediction models for breast cancer. A study published in 2009 by Jacobi et al. [6], listed seven models two of which included modifiable risk factors namely the Gail model [7] and the Tyrer-Cuzick model [8]. Meads et al. [9] provided a review of models including at least one modifiable factor. Of the tested models, many require further validation as they have not been validated in more than one dataset [10].

The most well-known is the Gail model [7] which includes non-genetic risk factors with limited information on family history. These risk factors are age, age at menarche, age at first live birth and number of previous breast biopsies. While studies indicate that the Gail model is well calibrated, more recent study [11] suggests that it may actually under-predict the risk as it limits the family history of breast cancer to first degree relatives for example.

Current models can produce a more accurate risk estimation [12]. However, when applied to differing populations, these models can under- or overestimate the risk. Thus, it might be useful to know if adding a certain factor will lead to improved performance such as single-nucleotide polymorphisms (SNPs) [13], hormones and growth factors [14] and mammography density [15]. For example, previous attempts have been made to introduce mammographic density to prediction models.

The model proposed by Chen and colleagues [16] is an extension of the Gail model. It is constructed on the same population as the initial model but included weight and mammograms. However, information was not available for all subjects thus the data was much smaller than the one used in the initial Gail model. In this study, percentage density was added to the Gail model which increased the C-statistic from 0.602 to 0.664 but this continuous measure of breast density is not often available in practice (requires digital scanning and specialized software). Similarly, Tice et al. [17] proposed a model based only on breast density adjusted for age and ethnicity which performed as well as the Gail model. In 2006, Barlow et al. [18] proposed two prediction models using logistic regression, one for pre-menopausal and one for post-menopausal women. The models were constructed with 4 or 10 factors depending on the menopausal status. The risk models gain the use of hormone therapy and BI-RADS density compared to the Gail model. The reported C-statistic increased from 0.605 (95% CI 0.60 to 0.61) to 0.62 (65% CI 0.62 to 0.63). Tice and colleagues [19] proposed a model developed on a different subset of the BCSC dataset that Barlow and colleagues used. Their reported C-statistic yields an increase from 0.61 to 0.66.

The purpose of this paper is to determine the factors that increase the risk of breast cancer.

2. Feature Selection

The focus of feature selection process is to detect the relevant features and discard the irrelevant ones in such way that the selected subset correctly describes the input data while providing good prediction performance and reducing the noise and irrelevant variables effect [20]. The selection of the best feature subset may lead to an improvement of the learner performance (in terms of speed, generalization and simplicity), a reduced cost and gain some insight of the domain [21]. Hence, its importance in data preprocessing.

As it is shown in [22], feature selection is an NP-Hard problem. Applying and testing the effectiveness of feature selection methods on real data is challenging especially without knowing the relevant attributes. Besides, the quality of the employed dataset may influence the results such as binary or multiple class output, noisy and missing data, number of redundant or irrelevant features. The difficulty resides also in the size of the input variables and the size of the selected subset. The amount of studies on feature selection proves this. New methods are constantly proposed and various comparative studies are conducted. But, there is no Best-method agreed upon. Researchers mainly focus on a specific problem, a feature selection technique category or the learner performance.

Feature selection methods can be categorized according to various criteria. One popular category is from the classifier perspective that is if the feature selection method is conducted independently from the classification process, it is referred to as filter, if the selection is done using the classifier to evaluate the subset it is referred to as wrapper and embedded are methods that perform the selection process along with the training. Another perspective consists in dividing feature selection methods into those evaluating subsets and those evaluating individual attributes.

3. Feature Stability

Traditionally, finding a subset of features that best improves the classification accuracy is the main goal of feature selection. The selected features need to be highly relevant and nonredundant. Thus, a majority of algorithms concentrate on this issue. But, besides high accuracy another important point of concern is feature stability. This problem is significant for real-applications where the selection is used as knowledge discovery tool. For instance, in analyzing cancer biomarkers a stable subset of markers is crucial in discovering links between two cancer labels [23] and a strong indicator of reproducible results [24].

Stability of a feature selection algorithm refers to its sensitivity to variations in the training set [25]. For example, two algorithms might have the same performance in terms of classification but act differently stability wise. A possible solution is to choose the best algorithm based on both accuracy and stability.

Perturbations in the training set can occur by simply adding a new sample. According to [26], The number of selected features k and the underlying characteristics of datasets can affect the stability. Such factors are the dimensionality m, the sample size n and the variation of the distribution of the dataset. Another important issue to consider is the importance of the chosen feature selection method as different algorithms work differently on a given dataset.

Assessing the stability of a feature selection method consists in defining a function that takes as input the resulting feature subsets and returns as output a stability value. However, the function depends on the type of feature selection procedure. As introduced by [25], there are three different types of outputs namely a weighting or scoring on the features, ranking and subset selection. There exist a relation between these types as weighting can be translated into a ranking and ranks can be mapped into a feature subset by adding a threshold. Hence, there is stability measures for each type of output. Only the metrics we are using in the experiment section are presented. For further informations on stability measures refer to [24] and [27].

4. Methodologies

In the following subsections, the feature selections methods used are presented. The methods are Chi-Square, T-score, F-score and Gini Index. Next, the stability metrics used in this study are described. The metrics are respectively Jaccard Index, Consistency Index and Spearman Correlation measure. Finally, the two classification algorithms used for testing namely Random Forest and Artificial Neural Network.

4.1. Feature Selection Methods

The rest of this section is devoted to a brief description of the tested feature selection techniques. Since we are interested in evaluating the attributes in this database, we focused mainly on filter techniques. There are several methods that analyze features individually and produce a ranking. Univariate filter based methods rely on some statistical measure to evaluate the relevance of features. In this work, we focus on these representative methods, namely: Chi-square, T-score, F-score and Gini-index. A brief description of each is presented here after.

Chi-Square. The Chi-Square is a univariate relevance criterion designed for discrete variables [28]. It is a supervised approach that performs a discretization technique. It measures the association between features and classes. Initially, each observed value of a numeric feature is placed into its own interval. The Chi-Square χ^2 metric is used in the next step to determine the relative frequencies of the classes in adjacent intervals. If they are similar, the intervals are merged with respect to a predetermined threshold. The Chi-Square score can be computed using the following formula 7, for a given feature f_i with r different values [29].

$$\chi^2(f_i) = \sum_{j=1}^{r}\sum_{s=1}^{c}\frac{(n_{js}-\mu_{js})^2}{\mu_{js}} \tag{1}$$

where c is the number of classes, n_{js} is the number of instances with the j^{th} feature value in class s, $\mu_{js} = \frac{n_{j*}n_{*s}}{n}$ with n is the total number of instances, n_{j*} is the number of instances with j^{th} feature value and n_{*s} is the number of instances in class s.

T-Score. The T-Score is limited to binary classification problems. The intuition behind this score for features selection is to assess whether a given feature can differentiate between the two classes c_1 and c_2. The T-score for the feature f_i is given as follows [29]:

$$t - score(f_i) = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \tag{2}$$

where μ_1 and μ_2 are respectively the mean value for the instances from the first and second class, σ_1 and σ_2 are the corresponding standard deviation values and n_1 and n_2 are the number of instances from each class.

F-Score. The F-Score is a univariate feature selection method. It measures the discrimination between two sets. The higher the score is, the more likely the feature is more discriminative. The features are ranked accordingly. Attributes may be selected depending on a predefined threshold value obtained by computing the average value of F-Scores [30]. The F-score of the i^{th} feature is defined as follows:

$$f - score(f_i) = \frac{(\bar{X}_i^{+} - \bar{X}_i)^2 + (\bar{X}_i^{-} - \bar{X}_i)^2}{\frac{1}{n_+ - 1}\sum_{j=1}^{n+}(x_{j,i}^{+} - \bar{X}_i^{+})^2 + \frac{1}{n_- - 1}\sum_{j=1}^{n-}(x_{j,i}^{-} - \bar{X}_i^{-})^2} \tag{3}$$

where $\bar{X}_i$, $\bar{X}_i^+$ and $\bar{X}_i^-$ are respectively the averages of all the classes, positive and negative. The discrimination between the positive and negative sets are indicated by the numerator. The denominator, on the other hand, indicates the discrimination within each of the two sets.

Gini Index. Gini Index is a statistical measure used to quantify the impurity of attributes. It quantifies if a feature f_i is able to separate the instances of different classes. The Gini Index was originally developed for splitting attributes in decision trees. The main idea of this index is that an attribute is selected as the divisive criterion if it can provide the minimum Gini Index after it covers all possible partitions [31].

$$Gini(f_i) = \sum_i P(f_i|c_i)^2 P(c_i|f_i)^2 \tag{4}$$

where $P(f_i|c_i)$ is the probability that feature f_i appears in every class c_i and $P(c_i|f_i)$ is the conditional probability that the feature f_i belongs to the class c_i when the feature f_i occurs.

4.2. Feature Stability Metrics

To compute the stability, a pairwise comparison of the selection results is performed and the assessment is averaged with respect to the number of comparisons. Let $X = x_1, \ldots, x_n$ be the original set of features. Suppose we carry out l runs of a feature selection algorithm F and record all the output sequences $S = (S_1, \ldots, S_l)^T$ (see equation 6) where $s_i \subseteq X$. Then the stability $\varphi(S)$ is defined as:

$$\varphi(S) = \frac{2}{l(l-1)} \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} \varphi(S_i, S_j) \tag{5}$$

Where $\varphi(S_i, S_j)$ is a predefined stability measure. In the following, we will present the different metrics used in this paper.

$$S = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_l \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{l1} & x_{l2} & \ldots & x_{ln} \end{pmatrix} \tag{6}$$

Jaccard Index. Also, referred to as the Tanimoto Distance ([24]). It measures the intersection between two sets S_i and S_j divided by the cardinality of the union of the sets (as presented in equation 3). This index ranges between 0 and 1 with 0 indicating that there is no overlap between the sets and 1 means that the two sets are identical. r_{ij} indicates the intersection of S_i and S_j.

$$\begin{aligned}\varphi_{Jaccard}(S_i, S_j) &= \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \\ &= 1 - \varphi_{Tanimoto}(S_i, S_j) \\ &= \frac{r_{ij}}{k_i + k_j - r_{ij}}\end{aligned} \tag{7}$$

Consistency Index. Based on the observation that many measures are biased by the cardinality of the feature sets, the author in [32] proposed a similarity measure $\varphi_{Consistency}$ between two sets S_i and S_j as presented in equation 4. The two sets must have identical cardinality that is $|S_i| = |S_j| = k$. The $\varphi_{Consistency}$ ranges between -1 and 1.

$$\varphi_{Kuncheva}(S_i, S_j) = \frac{nr_{ij} - k^2}{k(n-k)} \tag{8}$$

Spearman Correlation Measure. The Spearmans rank correlation coefficient presented in [25] and [33], evaluates the stability of ranked feature sets as presented in the equation 9. It takes values in $[-1, 1]$ with 0 meaning no correlation and 1 or -1 indicating a positive or negative correlation respectively.

$$\varphi_{Spearman}(S_i, S_j) = 1 - 6\sum_{r=1}^{n} \frac{(x_{ir} - x_{jr})^2}{n(n^2-1)} \tag{9}$$

4.3. Classification algorithms

In the following subsections, we describe the classification algorithms used for testing the selected features namely Random Forest and Artificial Neural Network.

Random Forest. Random forest is a classification algorithm that uses an ensemble of trees [34] It constructs each tree using a different bootstrap data sample. In contrast to standard trees where nodes are split according to the best split among variables, in random forest the nodes are split using the best among a random subset of variables. To obtain low-bias, the generated trees are un-pruned. Also, bootstrapping and random selection result in low correlation of the individual trees. Random forest can handle both two class and multi-class problems, noise and redundant/irrelevant variables.

Artificial Neural Network. Artificial Neural Network try to mimic the brain neurons. They are computational tools composed of highly inter-connected neurons. They consist of three types of neuron layers: input, hidden and output. The neuron store the information of the network as weights of its connections. The training process is repeated until a level of accuracy is achieved. In this process, the model makes predictions and is corrected when the predictions are wrong. Artificial neural network can handle large quantities of data. The classifier is known for its high classification efficiency. Most classification problem can be solved using one hidden layer.

5. Data: Materials and Analysis

BCSC dataset is comprised of seven mammography registries ([35]) located in Denver, New Hampshire, New Mexico, North Carolina, San Francisco, Seattle and Vermont. The screening occurred in the period from January 1, 1996 through December 31, 2002 (except for one registry, period through December 31, 2001). The data is available in the following link: http://breastscreening.cancer.gov/rfdataset/.

This study includes 2392998 mammograms. The age group combines women aged between 35 to 84 years. The women with previous breast cancer and or with breast augmentation were excluded. The women included had a prior mammogram within the last five years with a negative result. A total of 11638 women were diagnosed with breast cancer within one year of the screening [18]. The primary objective of BCSC is the study of breast carcinoma detection and outcome. Patient information was recorded using divers questionnaires. The information include birth date, race, education, personal and family history of breast cancer, prior breast procedures. There are 14 available factors in this study as described in table 2.

Table 2: The factors available in BCSC

This table provides the factors and possible values in BCSC

Factor	**Value**
Age	1 - 10
Menopause	0 = pre - menopausal; 1 = post - menopausal; 9 = unknown
Race	1 = white; 2 = Asian, PacificIslander; 3 = black; 4 = NativeAmerican; 5 = other=mixed; 9 = unknown
Hispanic	0 = no; 1 = yes; 9 = unknown
Density	BI-RADS breast density codes : 1; 2; 3; 4; 9
BMI	1 = 10-24.99; 2 = 25-29.99; 3 = 30-34.99; 4 = 35 or more; 9 = unknown
Age at first birth	0 = Age < 30; 1 = Age >=30 ; 2 = Nulliparous; 9 = unknown
Number of first degree relatives	0 = zero; 1 = one; 2 = 2 or more; 9 = unknown
Previous breast procedure	0 = no; 1 = yes; 9 = unknown
Last mammogram	0 = negative; 1 = false positive; 9 = unknown
Surgical Menopause	0 = natural; 1 = surgical; 9 = unknown or not menopausal
Hormone Therapy	0 = no; 1 = yes; 9 = unknown or not menopausal
Invasive	0 = no; 1 = yes
Cancer	0 = no; 1 = yes

Due to the non-standardization of the questionnaires between the different registries and the fact that some risk factors were not added at the same time, there is a high rate of missing values as observed in table 3. For example, the age at menarche was not reported for all women. Although it was confirmed as a risk factor for breast cancer. The BMI ($weight\ in\ kilograms = height\ in\ meters^2$) was not recorded from the start.

Table 3: The percentage of missing values per factor

This table provides the factors and possible values in BCSC

Factor	Missing Value
Age	0
Menopause	181959(7.60%)
Race	379804(15.87%)
Hispanic	486054(20.31%)
Density	652386(27.26%)
BMI	1336105(55.83%)
Age at first birth	1328294(55.50%)
Number of first degree	363319(15.18%)
Previous breast procedure	250312(10 : 46%)
Last mammogram	559018(23.36%)
Surgical Menopause	1247700(52.13%)
Hormone Therapy	980452(40.97%)
Invasive	0
Cancer	0

The age factor contains ten intervals of five years ranging from 35 to 84. The menopausal status was determined by age and self reported cessation of periods. Women aged between 35 and 54 years who stated that they still have their periods are considered as pre-menopausal. Women are categorized as post-menopausal if aged 55 or older plus women in the age group 45-55 who stated that their periods had stopped permanently or due to hormone therapy. Family history refers to the number of first degree relatives with breast cancer, values range is none, one and two or more. If there is a positive history of breast cancer but the exact number is unknown, the code 1 is attributed. Breast density recorded includes four categories according to BI-RADS(stands for Breast Imaging Reporting and Data System) coding system. These categories are: almost entirely fat, scattered fibroglandular densities, heterogeneously dense and extremely dense. The density can be recorded as unknown if a different coding system was used or it was not coded. The result of the previous mammogram was also reported. Women were excluded if they had a true positive or false-negative screening result as it indicates a prior breast cancer. The interpretation recorded are categorized as false-positive corresponding to BI-RADS categories 0, 4, 5 or 3 with immediate follow-up or true-negative corresponding to the categories 1, 2 or 3 without immediate follow-up. The complete BI-RADS categories are explained in details in table 4. The distribution of number of breast cancer diagnosed per demographic factor is presented in table 5.

Table 4: BI-RADS Categories

Category	Explanation
0	The assessment is incomplete (not enough information)
1	Negative, no suspicious abnormalities
2	Negative, benign findings
3	Probably benign, follow-up in 6 months
4	Suspicious, recommended biopsy
5	Highly suggestive of malignancy, immediate biopsy
6	Known cancer

6. Results and Discussion

In this section, we will examine the selection results along the following points:

- The frequency of a specific feature appearance at a given rank
- The stability performance of a feature selection method giving the dataset perturbation
- The classification performance

To compare the selection sets, a testing procedure is established. The experimental design is described in the next section, followed by a discussion on the frequency of appearance of certain features in a given rank. Then, a comparison of the stability performances is given and the classification performance given a feature subset.

6.1. Experimental Settings

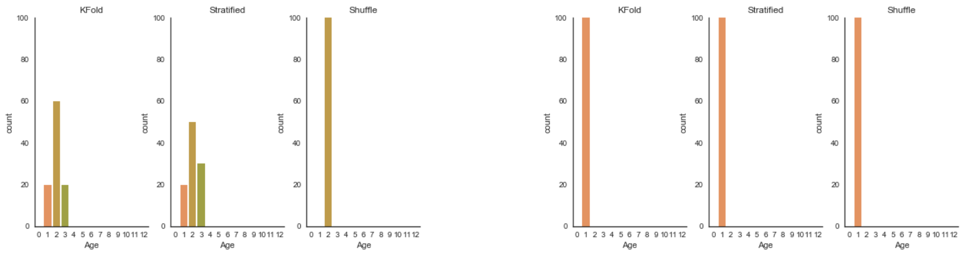
Data perturbation is used to test the stability of feature selection methods. Cross validation is performed with three different schemes. First, KFold cross validation is performed where the data is split into ten folds. Second, shuffle cross validation is applied where random splits are used to generate the training and testing sets. Last, a stratified cross validation is used to split the data where the percentage of samples for each class is preserved.

6.2. Discussion on Features Frequencies

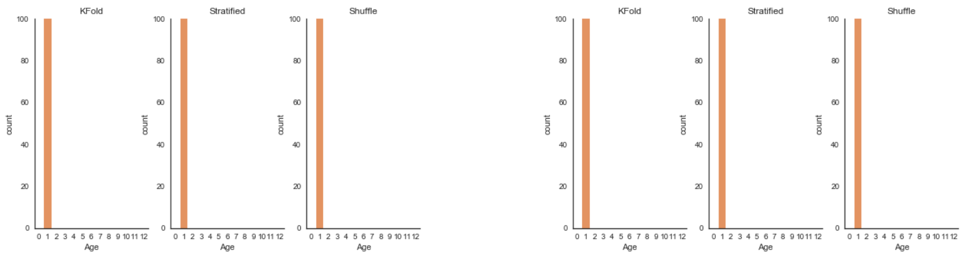
Figures 1, 2 and 3 depict examples of the frequency that a specific feature appears at a rank giving a perturbation process and a feature selection method. In the case of age, the results vary when KFold is applied as presented in figure 1. The rank ranges between second and fourth with the percentage of appearance at the third position is slightly higher and it is ranked third when the shuffling split is applied. Whereas, for the other perturbations schemes age is ranked second.

The methods Chi-Square, F-Score and T-Score place the factor breast density mainly at the fifth and sixth position although it is an important factor according to experts. The frequencies are presented in figure 2.

The positions of features: Age first birth, BMI, Hispanic, last mammogram, number of first degree relatives, previous breast procedures and race vary greatly across the different

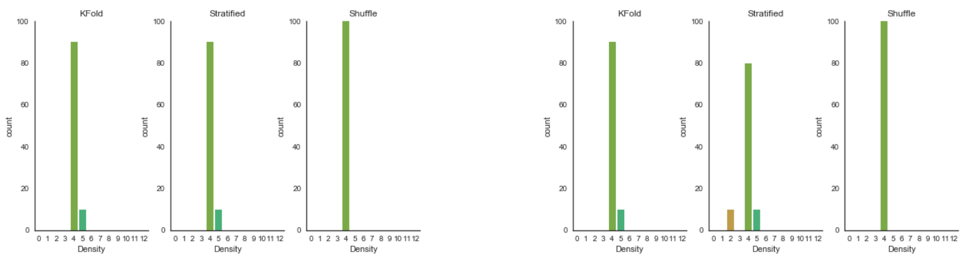


(a) Age frequencies generated by Chi-Square and F-Score

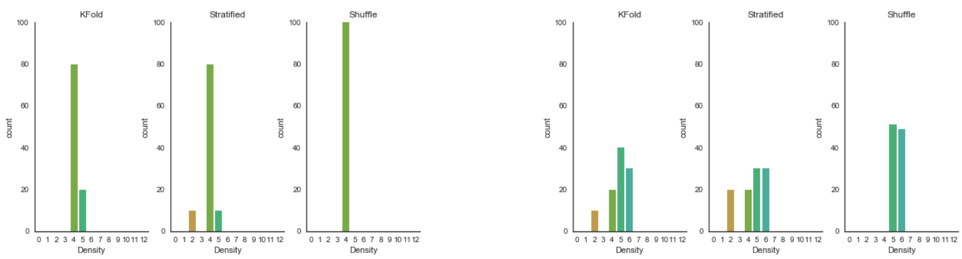


(b) Age frequencies generated by T-Score and Gini-Index

Figure 1: Example of features frequencies



(a) Density frequencies generated by Chi-Square and F-Score



(b) Density frequencies generated by T-Score and Gini-Index

Figure 2: Example of features frequencies

Table 5: Breast cancer per demographic factors

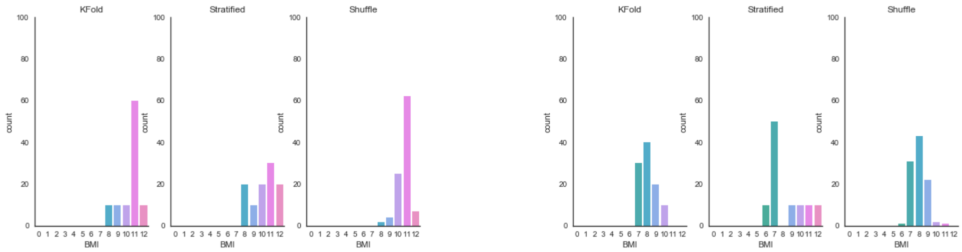
Factors	Values	Cancer 0	%	1	%	All
Age	1	42670	0.997942	88	0.002058	42758
	2	286633	0.997744	648	0.002256	287281
	3	385965	0.996692	1281	0.003308	387246
	4	426608	0.996022	1704	0.003978	428312
	5	332337	0.994628	1795	0.005372	334132
	6	261945	0.994019	1576	0.005981	263521
	7	230437	0.993674	1467	0.006326	231904
	8	201686	0.993009	1420	0.006991	203106
	9	144013	0.992495	1089	0.007505	145102
	10	69066	0.991815	570	0.008185	69636
	All	2381360	0.995137	11638	0.004863	2392998
Race	1	1729317	0.994995	8698	0.005005	1738015
	2	102573	0.995874	425	0.004126	102998
	3	120958	0.995261	576	0.004739	121534
	4	28277	0.997109	82	0.002891	28359
	5	22186	0.995424	102	0.004576	22288
	9	378049	0.995379	1755	0.004621	379804
	All	2381360	0.995137	11638	0.004863	2392998
Hispanic	0	1740832	0.994986	8772	0.005014	1749604
	1	156756	0.996288	584	0.003712	157340
	9	483772	0.995305	2282	0.004695	486054
	All	2381360	0.995137	11638	0.004863	2392998

selection methods and permutation schemes. An example of these variations is presented in figure 3 for the BMI factor.

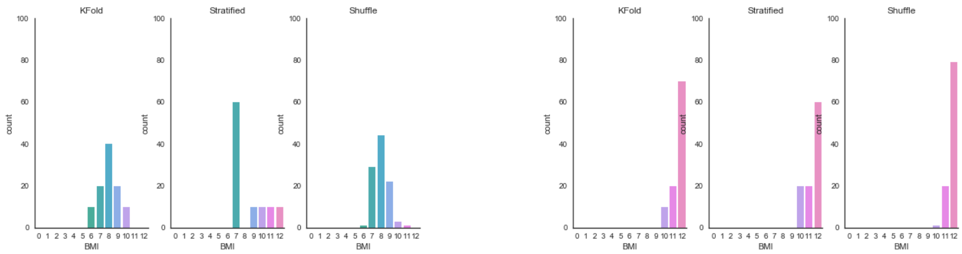
The top six ranked features produced by F-Score and T-Score are the same except for some variations in the frequencies as depicted in table 7. Gini Index puts Menopause third while it appears in sixth position for all other methods and permutations. It ranks third in 10% of the runs for both F-Score and T-Score when KFold and stratified cross validation are performed. Also, the factor previous breast procedures ranks sixth when the Gini Index is calculated. Whereas the density factor appears in the top six ranked features for all the other methods as it ranks fifth.

6.3. Discussion on Stability of Feature Selection Algorithms

In figure 4, the evolution of the stability results for $\varphi_{Jaccard}$, $\varphi_{Consistency}$ and $\varphi_{Spearman}$ for the different methods considered are presented. The results reflect the sensitivity of the methods to perturbation. Jaccard scores are high for the top ranked features especially if the subset size is small, i.e. number of features considered is low. Similarly, Spearman scores can be low for the top ranks as differences are more heavily accounted for. On the other



(a) BMI frequencies generated by Chi-Square and F-Score



(b) BMI frequencies generated by T-Score and Gini-Index

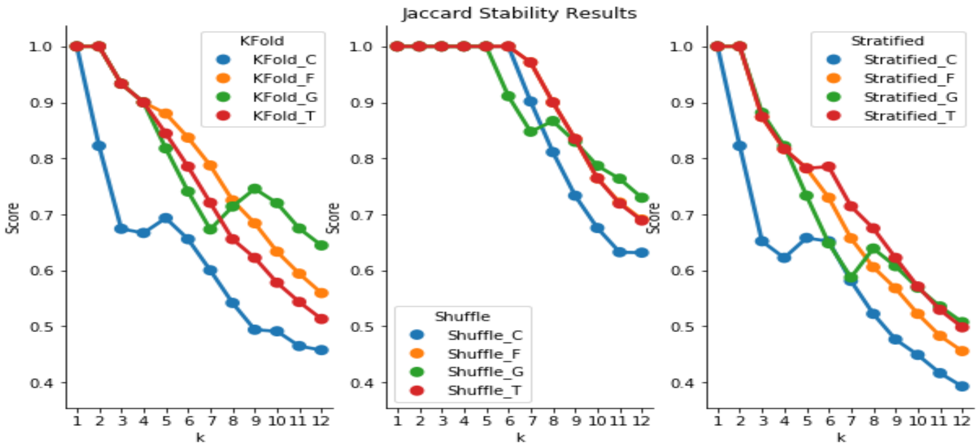
Figure 3: Example of features frequencies

hand, the consistency index decreases as k (number of features) increases. This score has a high value if the consistency exceeds the correction of chance term.

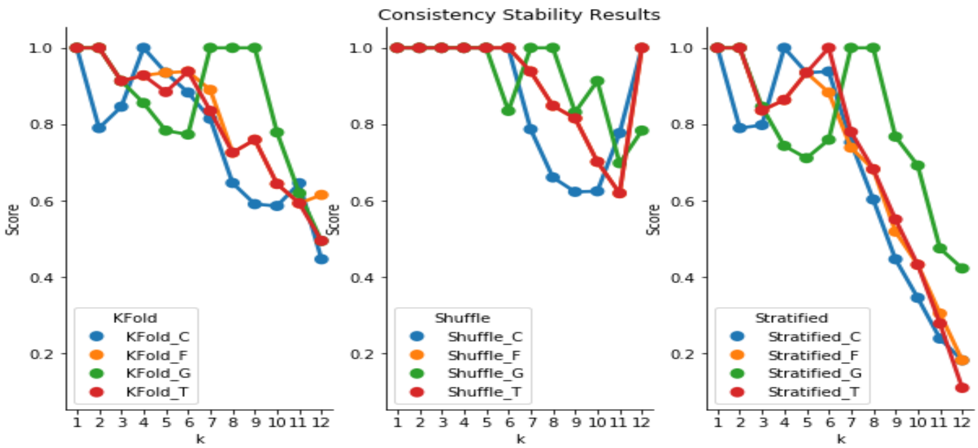
In the stratified cross validation case, the consistency index decreases noticeably in comparison with the other perturbation schemes. It reaches 0.1093 for T-Score selection when 12 factors are selected and 0.1815 for F-Score and Chi-Square. As expected, the Spearman Index increases as more features are added to the subset of selected factors. It reaches its maximum of 0.9835 for the Gini index and shuffle cross validation. The Jaccard index can be extremely stable for the top ranked features as depicted for the top six factors when T-Score and shuffle cross validation are applied. The variations in the Jaccard index scores can be attributed to the fact that certain features are selected less frequently in this subset. For example, in the case of the Gini Index and stratified cross validation the frequencies vary from 40% for the factor Previous Breast Procedure to a 100% for Invasive. Consequently, the Jaccard index decreases from 1.0 to a 0.5074 score. Also, in case of stratified cross validation and Chi-Square there is a dip at $k = 3$. It corresponds to a decrease of the age frequencies as it only equals 50% of the runs.

6.4. Classification Performance

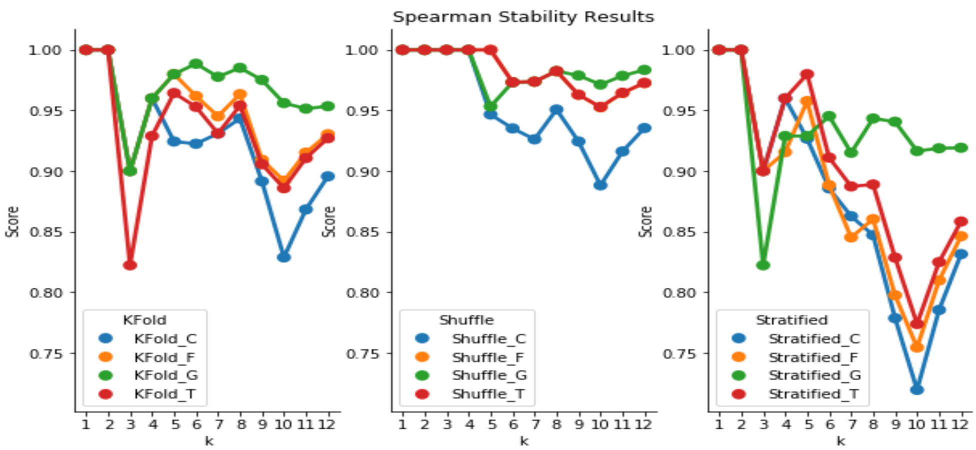
The performance is evaluated using the 10-fold cross-validation. Using the response of the classifier, confusion matrices are generated and true positive rate (TPR), false positive rate (FPR), Recall and Precision are calculated. Furthermore, composite measures are generated using these metrics, i.e. F-Measure and Accuracy. F-Measure is the harmonic mean of Precision and Recall. The accuracy is the percentage of correctly identified classes.



(a) Evolution of the Jaccard Index per k



(b) Evolution of the Consistency Index per k



(c) Evolution of the Spearman Index per k

Figure 4: Evolution of the stability indexes

These metrics are used to compare the performance of Random Forest (RF) classifier and Artificial Neural Network (ANN) classifier when all features are used and when the top six features are used. The classification results are presented in table **??**. It is worth noting that the results are the average scores. The results of RF classifier are generated using 10 trees. The results of ANN classifier are generated with one input layer consisting of one node for each feature, one hidden layer composed of half the input nodes and one output layer consisting of one node.

Tables 8, 9 and 10 present the performance results of the two classifiers. The accuracy rates are high for all the tests performed except for the case of stratified cross-validation and random forests. It equals 0.627 when all features are available and increases to 0.866 with the top six features ranked using the Gini index. Overall, the average accuracy for the tested cases equals 0.99.

The decrease in the number of features used did not affect the performance's rate and in some cases the accuracy and F-Measure rates slightly increased. For example, the accuracy increased from 0.9988 to 0.9990 for ANN and KFold cross-validation. The F-Measure, in the case of RF and shuffle cross-validation, evolved from 0.9994 to 0.9995.

7. Conclusion

This paper has explored risk factors for predicting breast cancer. To achieve this, feature selection techniques were used to rank the factors and stability metrics were used to test the sensitivity of these algorithms to variations in the data. A ranking subset was obtained after computing the factors frequencies of appearance at a certain rank. The data was analyzed using two classification techniques: random forest and artificial neural network and the results were compared. The classification performance using half the factors set achieves the same scores as when the full set of factors is used. It is worth mentioning that the factor Invasive has a high correlation with the target class (score of 0.9). This can explain the fact that it is always considered relevant by all the selection methods. From the experiments, the Jaccard Index scores indicate the sensitivity of the methods to variations in the training set and in the number of features considered as the scores vary between 1.0 and 0.4. Future research effort should be allocated to test the output of wrapper methods and other classification algorithms. Also, other breast cancer datasets should be considered to find relevant features.

8. Bibliography

[1] B. Stewart, W. C.P, World cancer report 2014., Tech. rep., Technical report, International Agency for Research on Cancer (2014).

[2] M. Hayat, N. Howlader, M. Reichman, B. Edwards, Cancer statistics, trends, and multiple primary cancer analyses from the surveillance, epidemiology, and end results (seer) program, The Oncologist 12 (1) (2007) 20–37.

[3] L. Kushi, C. Doyle, M. McCullough, C. Rock, W. Demark-Wahnefried, E. Bandera, S. Gapstur, A. Patel, K. Andrews, T. Gansler, T. A. C. S. . Nutrition, P. A. G. A. Committee, American cancer society guidelines on nutrition and physical activity for cancer prevention, CA: A Cancer Journal for Clinicians 62 (1) (2012) 30–67.

[4] X. Xiong, Y. Kim, Y. Baek, D. Rhee, S. Kim, Analysis of breast cancer using data mining statistical techniques, in: Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Network, 2005, pp. 82–87.

[5] J. Cruz, D. Wishart, Applications of machine learning in cancer prediction and prognosis, Cancer Informatics 2 (2006) 59 – 77.

[6] C. Jacobi, G. de Bock, B. Siegerink, C. van Asperen, Differences and similarities in breast cancer risk assessment models in clinical practice: which model to choose?, Breast Cancer Research and Treatment 115 (2) (2009) 381–390.

[7] M. Gail, L. Brinton, D. Byar, D. Corle, S. Green, C. Schairer, J. Mulvihill, Projecting individualized probabilities of developing breast cancer for white females who are being examined annually 81 (1990) 1879–86.

[8] J. Tyrer, S. Duffy, J. Cuzick, A breast cancer prediction model incorporating familial and personal risk factors 23 (2004) 1111–30.

[9] C. Meads, I. Ahmed, R. Riley, A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance, Breast Cancer Research and Treatment 132 (2) (2012) 365–377.

[10] J. Cintolo-Gonzalez, D. Braun, A. Blackford, E. Mazzola, A. Acar, J. Plichta, M. Griffin, K. Hughes, Breast cancer risk models: a comprehensive overview of existing models, validation, and clinical applications, Breast Cancer Research and Treatment 164 (2) (2017) 263–284.

[11] E. Amir, O. Freedman, B. Seruga, D. Evans, Assessing women at high risk of breast cancer: A review of risk assessment models, Journal of the National Cancer Institute 102 (10) (2010) 680–691.

[12] A. Howell, A. Anderson, R. Clarke, S. Duffy, D. Evans, M. Garcia-Closas, A. Gescher, T. Key, J. Saxton, M. Harvie, Risk determination and prevention of breast cancer, Breast Cancer Research 16 (5) (2014) 446.

[13] K. Michailidou, P. G.-N. A. Hall, et al., Large-scale genotyping identifies 41 new loci associated with breast cancer risk., Nature genetics. 45 (4) (2013) 353–361.

[14] H. Endogenous, B. C. C. Group, Circulating sex hormones and breast cancer risk factors in postmenopausal women: reanalysis of 13 studies., British Journal of Cancer. 105 (5) (2011) 709–722.

[15] C. W. Huo, G. L. Chew, K. L. Britt, W. V. Ingman, M. A. Henderson, J. L. Hopper, E. W. Thompson, Mammographic density a review on the current understanding of its association with breast cancer, Breast Cancer Research and Treatment 144 (3) (2014) 479–502.

[16] J. Chen, D. Pee, R. Ayyagari, B. Graubard, C. Schairer, C. Byrne, J. Benichou, M. Gail, Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density., JNCI: Journal of the National Cancer Institute. 98 (17) (2006) 12151226.

[17] J. A. Tice, S. R. Cummings, E. Ziv, K. Kerlikowske, Mammographic breast density and the gail model for breast cancer risk prediction in a screening population, Breast Cancer Research and Treatment 94 (2) (2005) 115–122.

[18] W. Barlow, E. White, R. Ballard-Barbash, P. Vacek, L. Titus-Ernstoff, P. Carney, J. Tice, D. Buist, B. Geller, R. Rosenberg, B. Yankaskas, K. Kerlikowske, Prospective breast cancer risk prediction model for women undergoing screening mammography., JNCI: Journal of the National Cancer Institute. 98 (17) (2006) 1204–1214.

[19] J. Tice, S. Cummings, R. Smith-Bindman, L. Ichikawa, W. Barlow, K. Kerlikowske, Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model., Ann Intern Med. 148 (2008) 337347.

[20] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[21] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, Knowledge and Information Systems 34 (3) (2013) 483–519.

[22] I. Kononenko, I. Bratko, Information-based evaluation criterion for classifier's performance, Machine Learning 6 (1) (1991) 67–80.

[23] Y. Li, J. Si, G. Zhou, S. Huang, S. Chen, Frel: A stable feature selection algorithm, IEEE Transactions on Neural Networks and Learning Systems 26 (7) (2015) 1388–1402.
[24] S. Nogueira, G. Brown, Measuring the Stability of Feature Selection, 2016, pp. 442–457.
[25] A. Kalousis, J. Prados, M. Hilario, Stability of feature selection algorithms: a study on high-dimensional spaces, Knowledge and Information Systems 12 (1) (2007) 95–116.
[26] S. Alelyani, H. Liu, L. Wang, The effect of the characteristics of the dataset on the selection stability, in: 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, 2011, pp. 970–977.
[27] T. Khoshgoftaar, A. Fazelpour, H. Wang, R. Wald, A survey of stability analysis of feature subset selection techniques, in: 2013 IEEE 14th International Conference on Information Reuse Integration (IRI), 2013, pp. 424–431.
[28] H. Liu, R. Setiono, Chi2: feature selection and discretization of numeric attributes, in: Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, 1995, pp. 388–391.
[29] J. Li, K. Cheng, S. Wang, F. Morstatter, T. R.P., J. Tang, H. Liu, Feature selection: A data perspective, CoRR.
[30] K. Polat, S. Gne, A new feature selection method on classification of medical datasets: Kernel f-score feature selection, Expert Systems with Applications 36 (7) (2009) 10367–10373.
[31] W. Zhu, Y. Lin, Using gini-index for feature weighting in text categorization 9 (2013) 5819–5826.
[32] L. Kuncheva, A stability index for feature selection, in: Proceedings of the 25th Conference on Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications, 2007, pp. 390–395.
[33] S. Nogueira, K. Sechidis, G. Brown, On the Use of Spearman's Rho to Measure the Stability of Feature Rankings, Springer International Publishing, 2017, pp. 381–391.
[34] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.
[35] R. Ballard-Barbash, S. Taplin, B. Yankaskas, V. Ernster, R. Rosenberg, P. Carney, W. Barlow, B. Geller, K. Kerlikowske, B. Edwards, C. Lynch, N. Urban, C. Chrvala, C. Key, S. Poplack, J. Worden, L. Kessler, Breast cancer surveillance consortium: a national mammography screening and outcomes database., American Journal of Roentgenology 169 (4) (1997) 1001–1008.

Table 6: Breast cancer per personal factors

Factors		Cancer				
	Values	0	%	1	%	All
Menopause	0	566489	0.996962	1726	0.003038	568215
	1	1633524	0.994339	9300	0.005661	1642824
	9	181347	0.996637	612	0.003363	181959
	All	2381360	0.995137	11638	0.004863	2392998
Density	1	147881	0.997787	328	0.002213	148209
	2	779023	0.995704	3361	0.004296	782384
	3	670282	0.994472	3726	0.005528	674008
	4	135294	0.994728	717	0.005272	136011
	9	648880	0.994626	3506	0.005374	652386
	All	2381360	0.995137	11638	0.004863	2392998
BMI	1	506423	0.995139	2474	0.004861	508897
	2	323649	0.994766	1703	0.005234	325352
	3	144069	0.994794	754	0.005206	144823
	4	77421	0.994860	400	0.005140	77821
	9	1329798	0.995280	6307	0.004720	1336105
	All	2381360	0.995137	11638	0.004863	2392998
Age at First Birth	0	718588	0.995006	3607	0.004994	722195
	1	140563	0.994876	724	0.005124	141287
	2	200141	0.994628	1081	0.005372	201222
	9	1322068	0.995313	6226	0.004687	1328294
	All	2381360	0.995137	11638	0.004863	2392998
Number of First Degree	0	1710426	0.995383	7934	0.004617	1718360
	1	293855	0.993532	1913	0.006468	295768
	2	15408	0.990804	143	0.009196	15551
	9	361671	0.995464	1648	0.004536	363319
	All	2381360	0.995137	11638	0.004863	2392998
Previous Breast Procedure	0	1714642	0.995579	7614	0.004421	1722256
	1	417631	0.993343	2799	0.006657	420430
	9	249087	0.995106	1225	0.004894	250312
	All	2381360	0.995137	11638	0.004863	2392998
Last Mammogram	0	1791368	0.995241	8566	0.004759	1799934
	1	33779	0.992158	267	0.007842	34046
	9	556213	0.994982	2805	0.005018	559018
	All	2381360	0.995137	11638	0.004863	2392998
Surgical Menopause	0	713650	0.993989	4316	0.006011	717966
	1	425152	0.994899	2180	0.005101	427332
	9	1242558	0.995879	5142	0.004121	1247700
	All	2381360	0.995137	11638	0.004863	2392998
Hormone Therapy	0	725211	0.994535	3985	0.005465	729196
	1	679400	0.994220	3950	0.005780	683350
	9	976749	0.996223	3703	0.003777	980452
	All	2381360	0.995137	11638	0.004863	2392998
Invasive	0	2381360	0.999034	2303	0.000966	2383663
	1	0	0.000000	9335	1.000000	9335
	All	2381360	0.995137	11638	0.004863	2392998

Table 7: Top six ranked features
per feature selection methods

Methods		Factors					
Chi Square	KFold	Invasive 100%	Hormone Therapy 80%	Age 60%	Surgical Menopause 80%	Density 90%	Menopause 70%
	Shuffle	Invasive 100%	Hormone Therapy 100%	Age 100%	Surgical Menopause 100%	Density 100%	Menopause 100%
	Stratified	Invasive 100%	Hormone Therapy 80%	Age 50%	Surgical Menopause 70%	Density 90%	Menopause 80%
F-Score	KFold	Invasive 100%	Age 100%	Hormone Therapy 90%	Surgical Menopause 90%	Density 90%	Menopause 80%
	Shuffle	Invasive 100%	Age 100%	Hormone Therapy 100%	Surgical Menopause 100%	Density 100%	Menopause 100%
	Stratified	Invasive 100%	Age 100%	Hormone Therapy 80%	Surgical Menopause 80%	Density 80%	Menopause 70%
T-Score	KFold	Invasive 100%	Age 100%	Hormone Therapy 90%	Surgical Menopause 90%	Density 80%	Menopause 70%
	Shuffle	Invasive 100%	Age 100%	Hormone Therapy 100%	Surgical Menopause 100%	Density 100%	Menopause 100%
	Stratified	Invasive 100%	Age 100%	Hormone Therapy 80%	Surgical Menopause 80%	Density 80%	Menopause 90%
Gini Index	KFold	Invasive 100%	Age 100%	Menopause 90%	Hormone Therapy 90%	Surgical Menopause 70%	Previous Breast Procedure 50%
	Shuffle	Invasive 100%	Age 100%	Menopause 100%	Hormone Therapy 100%	Surgical Menopause 100%	Previous Breast Procedure 60%
	Stratified	Invasive 100%	Age 100%	Menopause 80%	Hormone Therapy 80%	Surgical Menopause 60%	Previous Breast Procedure 40%

Table 8: Comparison of RF and ANN performance for the Stratified Cross-validation

	Random Forest			ANN		
	All	Top 6 (Others)	Top 6 (Gini)	All	Top 6 (Others)	Top 6 (Gini)
TPR	0.99849	0.997894	0.998904	0.999034	0.998262	0.998629
FPR	0.989554	0.993084	0.971368	0	0	0
Precision	0.62822	0.427892	0.866437	1	1	1
Recall	0.998227	0.995317	0.998537	0.999034	0.998264	0.998630
Accuracy	0.627694	0.42977	0.866141	0.999038	0.998267	0.998634
F-Measure	0.743465	0.529237	0.895625	0.999517	0.999131	0.999314

Table 9: Comparison of RF and ANN performance for the Shuffle Cross-validation

	Random Forest			ANN		
	All	Top 6 (Others)	Top 6 (Gini)	All	Top 6 (Others)	Top 6 (Gini)
TPR	0.999033	0.999033	0.999033	0.999033	0.998631	0.997867
FPR	0.009127	0	0	0	0	0
Precision	1	1	1	1	1	1
Recall	0.999033	0.999033	0.999033	0.999033	0.998633	0.99787
Accuracy	0.999001	0.999036	0.999036	0.999036	0.998636	0.997873
F-Measure	0.999498	0.999516	0.999516	0.999516	0.999316	0.998933

Table 10: Comparison of RF and ANN performance for the KFold Cross-validation

	Random Forest			ANN		
	All	Top 6 (Others)	Top 6 (Gini)	All	Top 6 (Others)	Top 6 (Gini)
TPR	0.999034	0.999034	0.999034	0.998839	0.999034	0.998716
FPR	0.00299	0	0	0	0	0
Precision	0.999988	1	1	1	1	1
Recall	0.999034	0.999034	0.999034	0.998839	0.999034	0.998717
Accuracy	0.999026	0.999038	0.999038	0.998843	0.999038	0.998721
F-Measure	0.999511	0.999517	0.999517	0.999419	0.999517	0.999358

Gabor Filter Algorithm for medical image processing: evolution in Big Data context.

N. Bourkache*, S. Sidhom**, and M. Laghrouche*.

Abstract—In the health field, several thousand images are generated every day in medical imaging establishments. On the one hand, the volume of information involved is still far from being fully controlled. On the other hand, the development of machine learning tools today opens the way to a new generation of image analysis in this context of "BigData". Moreover, our approach is a part towards this dynamic research using Gabor Filter Algorithm in image processing, analysis and diagnosis. In order to test the robustness of our algorithm and its degree of application in the context of BigData, we tested, in a first analysis phase, our algorithm on an image-database containing 320 mammograms. The precision obtained is estimated at 75% for a recall of 33%. In a second analysis phase, we performed the test on an image data-base containing 1000 medical images. The precision obtained is estimated at nearest 70% for a recall of 33%. Although the precision obtained in this first step is far from perfect, our processing algorithm remains promising and shows a good adaptation in contest of "Big Data”. The purpose of this research work is to contribute to machine learning. Therefore, the analysis process based on Gabor Filter can distinguish, more and concisely, tumor mammograms from healthy mammograms in Big Data context.

Key words— medical image (mammogram), image processing, image retrieval; Gabor Filter algorithm; medical databases; Big Data environment (extended study context), tumor diagnosis.

1 Introduction

CURRENTLY, the number of medical imaging exams is estimated to be close to 4 billion per year worldwide (Medical imagery - May 2016: Cour des comptes - www.ccomptes.fr - @Courdescomptes). Radiology establishments, every day, perform thousand radiologic examinations. The huge volume of data produced, thus feeding databases which size continues to grow considerably each year. If we take for example the last five years, in 2015, the volume of data already represents a few hundred petabytes (10^{15} bytes). In 2020, ten or so zettabytes (10^{21} bytes, i.e. 1 billion terabytes) are expected!

If we take the case of cancer, we observe in medical protocols that imaging is closely linked in diagnosis, even essential, in all phases: from diagnosis, treatment and follow-up after treatment in monitoring healing. In this area, the interest of imagery is paramount. Because, early detection (by imaging) may well reveal anomalies that predict imminent cancer! In such cases, and in order for appropriate medical measures to be taken early, it is desirable that the conclusion of the radiologist is as precise as possible. In this context, Computer Aided Diagnosis (CAD) systems are essential and then help in decision-making. In this article, we will present our medical imagery analysis tool based on Gabor Wavelets (or Filters).

- *Noureddine Bourkache & Mourad Laghrouch: University of Mouloud Mammeri (Tizi-Ouzou) & LAMPA Lab., Algeria. E-mail: nourbourkache@yahoo.fr ; larouche_67@yahoo.fr*
- *Sahbi Sidhom: University of Lorraine (Nancy) & LORIA Lab., France. E-mail: sahbi.sidhom@loria.fr*

2 Related works

Since the 1980s, image analysis tools have been constantly evolving. Particularly, the CBIRS (Content-Based Image Retrieval System) takes an important place and involves the interest of the scientific community since the 90s. In this type of systems, the image analysis is based on the extraction of the morphological features of the image namely: Color, shape, or texture. In the medical field (especially in cancer screening), recent works are mainly oriented towards “diagnostic assistance” and “decision-making”. We find in research works [1], [2], [3] a set of Classification Algorithms of Lung Nodules into Benign or Malignant (tumors) represented in CBIR Systems. In CT (Computed Tomography) *scans,* several methods in images classification and analysis are proposed. For example, [4] propose a classification and analysis of pulmonary nodules in CT images using random forest. In prostate cancer diagnosis, [5] offers a cancer classification based in genetic algorithm. In breast cancer, [6] presents a new method based on the expert annotation and automatic selection of cell types by their transcriptome profiles. In recent work, [7] offers a similarity measure method for mammogram retrieval. For large data volumes and in the order of Big Data, several approaches are proposed [8], [9], [10].

3 Research work

In this paper, we present our image analysis algorithm based on the extraction of morphological features (ie. image digital signature) from medical images. The objective of this work is to provide a learning tool and diagnosis aid in breast cancer.

3.1 Gabor Filter Algorithm in image processing: Texture representation

In our approach, we had chosen to study the texture parameter to be able to construct the digital component of images. Parameters are collected in the form of a complex vector, called: "texture vector" or "image digital signature". This one should be not sensitive to image transformations: particularly for translation and rotation of the image. For this important reason, we have chosen in our study the coding by Gabor wavelets (or Filters).

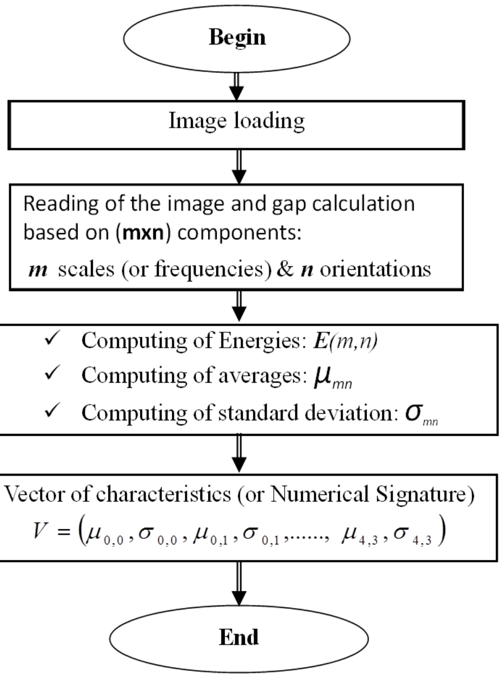


Fig. 1. Flowchart of the main algorithm for image indexing applying the Gabor "Filters" Model.

As illustration, cf. Fig. 1, the main algorithm for image indexing applying the Gabor filter Model.
This architecture represents the principal steps of features image representation.

In a context of applied research, the Gabor wavelet proves to be an interesting tool for texture analysis applied to image and it is largely adopted in performance measures.

3.2 Gabor Filter analysis

For an image I(x, y) having dimensions MxN, its conversion into discrete Gabor Wavelet is given by the following convolution formula:

$$G_{mn}(x,y) = \sum_s \sum_t I(x-s, y-t)\psi^*_{mn}(s,t) \quad (1)$$

Ψ^* is the combined of $\Psi(x,y)$ such as the formula:

$$\psi(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + j2\pi\cdot f\cdot x\right] \quad (2)$$

After applying the Gabor filter to the image with various orientations and levels of filtering, we get two computed formula of $\sigma_{m,n}$ (average) and $\mu_{m,n}$ (standard deviation):

$$\mu_{m,n} = \frac{E(m,n)}{MxN} \quad (3)$$

with: $$E(m,n) = \sum_x \sum_y \left|G_{m,n}(x,y)\right| \quad (4)$$

$$\sigma_{m,n} = \frac{\sqrt{\sum_x \sum_y \left(\left|G_{m,n}(x,y)\right| - \mu_{m,n}\right)^2}}{MxN} \quad (5)$$

The values of $\sigma_{m,n}$ and $\mu_{m,n}$ represent the components of the characteristics vector (V). Thus, for four orientations and five scales, this vector V had the following formula:

$$V = (\mu_{0,0}, \sigma_{0,0}, \mu_{0,1}, \sigma_{0,1}, \ldots\ldots, \mu_{4,3}, \sigma_{4,3}) \quad (6)$$

At this level, images are represented by the characteristic vectors in the space of numeric attributes.

3.3 Searching step

The research phase, the similarity measure between images is defined by a set of distances in the same defined space.

The similarities are computed with the image-query *Q* and the images-targets *T* (ie. Stored in the image-database) using for each vector values the distance *D(Q,T)* using the formula:

$$D(Q,T) = \sum_m \sum_n d_{mn}(Q,T) \quad (7)$$

where:

$$d_{mn} = \sqrt{(\mu^Q_{mn} - \mu^T_{mn})^2 + (\sigma^Q_{mn} - \sigma^T_{mn})^2} \quad (8)$$

The fig. 2, illustrates the various steps of the search process. The extraction of image characteristics is always carried out by the Gabor Filter Model applied to the query image.

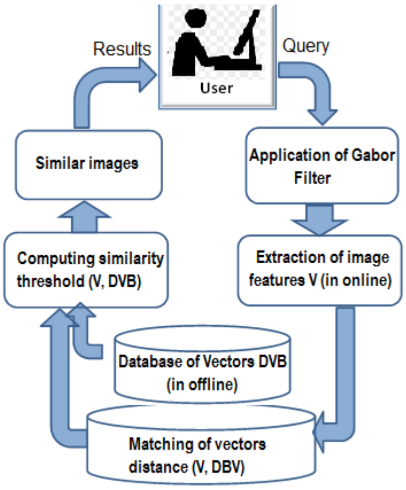


Fig. 2. Principal steps of the Information Retreival (IR) process by image content: application of the Gabor Filter Model.

4 CORPUS OF ANALYSIS AND RÉSULTS.

Our analysis corpus includes a mammographic image database (320 mammograms) and an enlarged image database (1000 medical images): cf. Fig.3.

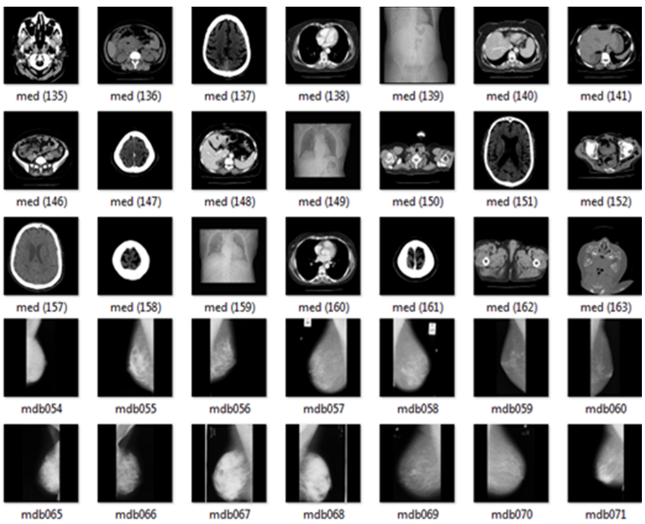


Fig. 3. Corpus of images (overview)

4.1 Test based on mammographic images

In our first test phase, we performed an analysis on 320 mammograms. The performance study based on the estimation of precisions according to the recall. The results obtained are illustrated in table 1.

In practice, we have carried out several tests on our image database. We estimated the details of the different recall values (from 5% to 95%). As shown in fig. 4.

For readability, we have simply presented in table.1 three reference values, namely: the details corresponding respectively to the reminders of 33% (Third of the relevant

TABLE 1
AVERAGE REFERENCE VALUES OBTAINED IN THE 1ST. TEST PHASE.

Main precision and recall values obtained	
Recall	Precision
33 %	75 %
50 %	64 %
66 %	53 %

images from the entire database have been found), 50% (half relevant images were found) 66% two-thirds of relevant images were found). In this first test phase, we see that the accuracy is around 64% for a recall of 50%: i.e. almost two images out of three, among the images found, are relevant when half the images of the entire database has been returned. While the accuracy rises to 75% when a third of the relevant images of the database is found against an accuracy of 53% for a recall of 66%.

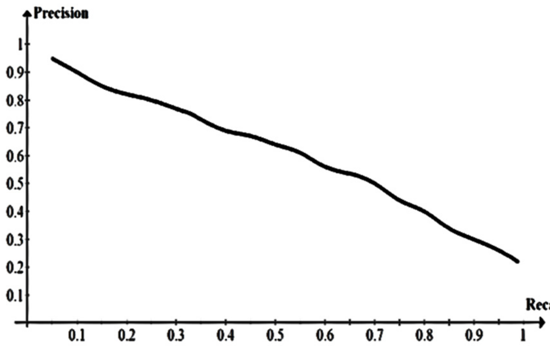


Fig. 4. Representation of the average values of the precisions according to the recall of the 1st. test phase

4.2 Test based on extended database

In the second test phase, we performed an analysis on a corpus of 1000 medical images. The results obtained are illustrated in Table. 2 and Fig 5.

TABLE 2
AVERAGE REFERENCE VALUES OBTAINED IN THE 2ND. TEST PHASE.

Main precision and recall values obtained	
Recall	Precision
33 %	70 %
50 %	59 %
66 %	42 %

In the second test phase, we obtained an accuracy of 70% for a recall of 33%: seven images out of ten found are relevant when a third of the relevant images from the database are returned. However, when the system finds two-thirds (66% recall) of the relevant images of the database, the accuracy is around 42%. This relative decrease in precision, in this second phase (1000 medical images) is essentially, due to the approximation from the texture point of view, of the images constituting our test database.

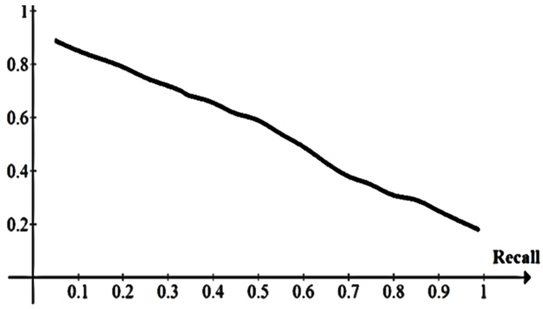


Fig. 5. Representation of the average values of the precisions according to the recall of the 2nd. test phase.

From results we note that the transition from our 1st corpus of images (320 mammograms) to the 2nd corpus (1000 medical images) has relatively affected the average values of the precision obtained. If we take for example the three reference values given in table 2, we see that the values of the average precisions have dropped a little compared to those obtained in table 1.

However, we also find that tripling the corpus of images (going from 320 images to 1000 images) has not weakened the performance of our system too much. This encourages us to process a larger corpus or even the transition to processing Big Data.

5 CONCLUSION AND PERSPECTIVES

In this work, we were able to apply the Gabor filter to different corpora of images (on the number side as well as the homogeneity side) the performances obtained are not perfect at this level. On the other hand, the observations made allow us to take a correct path in the treatment of Big-data. Moreover, since the low-level analysis of an image by the Gabor filter is carried out pixel by pixel, the processing time is relatively considerable for large image corpora. Even if major processing is done offline, the transition to BigData can, a priori, impose a more demanding computing time! to overcome this requirement, we are thinking of carrying out the processing on supercomputers (clusters) where each node of the cluster is responsible for processing, in parallel with the other nodes, a well-defined part of the global corpus. We then envisage carrying out this task in other work to come.

REFERENCES

[1] Bhavanishankar K., Sudhamani M.V. (2020) Classification of Lung Nodules into Benign or Malignant and Development of a CBIR System for Lung CT Scans. In: Smys S., Tavares J., Balas V., Iliyasu A. (eds) Computational Vision and Bio-Inspired Computing. ICCVBIC 2019. Advances in Intelligent Systems and Computing, vol 1108. Springer, Cham. https://doi.org/10.1007/978-3-030-37218-7_64. Janv. 2020

[2] Xie, Y., Xia, Y., Zhang, J., Song, Y., Feng, D., Fulham, M., Cai, W.: Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT. IEEE Trans. Med. Imaging **38**, 991–1004 (2018)

[3] Pham, T. D., Watanabe, Y., Higuchi, M., &Suzuki, H. (2017). Texture Analysis and Synthesis of Malignant and Benign Mediastinal Lymph Nodes in Patients with Lung Cancer on Computed Tomography. Scientific Reports, 7, 1-10

[4] Baboo, S.S., Iyyapparaj E.: A classification and analysis of pulmonary nodules in CT images using random forest. In: 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 1226–1232. IEEE (2018)

[5] Lu, X., & Chen, D. (2017). Cancer Classification through Filtering Progressive Transductive Support Vector Machine Based on Gene Expression Data. AIP Conference Proceedings, 1864 (1), 1-5.

[6] Identification and transfer of spatial transcriptomics signatures for cancer diagnosis (2020) 22: 6. https://doi.org/10.1186/s13058-019-1242-9. Janv. 2020.

[7] A similarity measure method fusing deep feature for mammogram retrieval (2019). Journal of X-Ray Science and Technology, vol. Pre-press, no. Pre-press, pp. 1-17, 2019. DOI: 10.3233/XST-190575

[8] Bertrand Thirion, «Introduction aux approches BigData pour l'imagerie médicale», Collège de France, 2 mai 2018: https://www.college-de-france.fr/site/gerard-berry/symposium-2018-05-02-09h35.htm

[9] Gérard Berry, «Imagerie médicale et apprentissage automatique: vers une intelligence artificielle?»: https://www.college-de-france.fr/site/gerard-berry/symposium-2017-2018.htm

[10] Isabelle Thomassin-Naggara et Olivier Clatz, «Technologies innovantes au chevet du dépistage du cancer du sein : quel futur pour les femmes ?», Collège de France, 2 mai 2018: https://www.college-de-france.fr/site/gerard-berry/symposium-2018-05-02-15h15.htm

Authors :
First N. BOURKACHE is a Teaching researcher at the Mouloud Mammeri Tizi-Ouzou University & LAMPA Lab., Algeria.

Second S. Sidhom Assistant Professor Lorraine University & LORIA Lab. (Kiwi Team)
https://members.loria.fr/SSidhom/
https://www.researchgate.net/profile/Sahbi_Sidhom

Third M. LAGHROUCHE Profesor at the Mouloud Mammeri Tizi-Ouzou University & LAMPA Lab., Algeria.

Improving Customer Relationship Management using Machine Learning techniques: A Tunisian Case Study

Afef Ben Brahim, Wafa Raboudi

Université de Tunis, Tunis Business School, LARODEC, BP 65, 2059 BIR El Kassaa, Tunisia
afef.benbrahim@tunis-business-school.tn, wafaraboudi@yahoo.fr

Abstract. In the CRM domain, predicting interesting customer profiles relies on studying their different characteristics and behavior. Data mining tools provide functions and algorithms that allow such knowledge extraction. We investigate a research study of data mining concepts and an application of its different functions for a Tunisian hotel case study. The proposed data mining application results in successful knowledge extraction that will be very precious to enhance the loyalty of the hotel customers.
Keywords: Data Mining, Hospitality, CRM, Machine learning

Introduction

For the hotel industry, the most important thing is to know the customers it is hosting. It is highly relevant for a company to identify its target before launching a marketing campaign. This allows focusing on the customer segment that better maximize the profits and appreciate the services provided. In communication, quality is more important than quantity especially for hoteliers.
Hotels can gather tremendous amounts of data; those data are valuable only if they are well treated. To help hoteliers extract relevant information and detect hidden patterns from the available data, a process exists that associates the analytical techniques with the database system; which is the "data mining" process. It uses statistical techniques to find relevant patterns and relationships among data that can be used in the decision making process.
Many success stories in the Hospitality domain are the outcome of the use of data mining techniques. But, what we noticed is that Tunisian hotels do not take the advantage of this powerful tool yet despite the poor customer relationship management (CRM) that exists in our hotels.
"The Flamingo Beach" Hotel in Djerba, Tunisia, has accepted our suggestion to use its customers' data and try to detect the hidden patterns and extract useful knowledge.
The objective of this study is to help responsibles retain the maximum number of guests and predict the most profitable profiles based on previous behaviors and the relation that exists between customer's information.
Those objectives will be met after the implementation of the different functions provided by data mining.
The section of this paper involves a study about data mining data mining concept and its evolution over the past years defining the different functions it provides.
Then, in Section 2, we introduce how data mining has been used in the Hospitality domain to achieve different goals and improve its CRM. These achievements made us realize how much Tunisian hotels need to take advantage of this concept and ameliorate its services. This is illustrated in the third Section where we implement the different functions of data mining on a sample of data collected from "The Flamingo Beach" hotel using an open source tool. We continue with a Section dedicated to experimental study then conclude this paper.

1 Data Mining and its application

This section aims to give the main ideas and concepts behind data mining which is being increasingly known.

1.1 Data mining concept

Nowadays, our world is often being referred to as the “information age” or “Computer age” era which is characterized by the Digital Revolution. In the information age, huge amounts of information are being gathered using different information tools and technologies which are changing and developing from regularly.

At the very beginning and before being a valuable information, data are collected from everywhere and stored in an unstructured way counting on the power of computer to help sort and go through this mixture of data. Unfortunately, these collections are rapidly being overwhelming and the extraction of data is becoming more and more complex as well as time-consuming.

To solve this big issue, structured database and database management systems have been created providing various functions that facilitate the entry and storage of the large amount of information and allow an efficient and effective retrieval of certain information whenever needed.

Unquestionably, those technologies have played a key role and contributed a lot in enhancing data acquisition and retrieval methods. But with the increasing use of information today in different domains: from business transactions and scientific data to satellite pictures and military intelligence, the previously discovered tools are insufficient and no more efficient for decision making process. The users’ requirements exceeded the ability and capability of a database management system due to the massive amount of data.

Thus, the need of a more developed and useful techniques have become imperative and obvious to help people analyze large amount of data and discover hidden patterns and knowledge. Without those powerful tools, the size of data would not have any importance.

So, in order to help experts extract relevant information and detect hidden patterns from enormous data, a new process to integrate between the analytical technique and database system was created which is called “Data mining”.

According to the Principles of data mining written by David J. Hand, Heikki Mannila and Padhraic Smyth, data mining is “the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data analyst”

In other words, data mining is the process of discovering and finding useful trends in a huge quantity of data and to make sense of it. It is one step among different steps that define the Knowledge Discovery in Databases (KDD) which refers to the extraction of significant, once unknown and potentially convenient information. However, KDD and data mining are often being used interchangeably.

In another definition of Rajaraman ullman, Data mining is the set of different models that contributed in the manipulation of data in a better way. We have mainly four models: the first one is for the data explanation, the second is to predict the future data samples, the third model is used to summarize the data and finally the last model is to extract important characteristics of the data.

Data mining is a repetitive process that includes several steps to transform data into useful information which will lead itself to powerful and successful decisions. It involves the following phases : Data mining functions:

In general, the use of Data mining is not specified for only one type of data or information. Any set of information can be treated using different functions that are provided with data mining. But we have to put in mind that, algorithms and techniques may vary with the type of data (Osmar R. Zaiane, 1999).

Furthermore, Data mining functions are used depending on the different needs of the user and on the specific sort of pattern to be found. They can be used independently, iteratively, or in combination. Actually, data mining tasks are divided into two different types. First, Descriptive Data mining tasks, in which we describe the general characteristics of the existing data. A retailer trying to associate products that can be purchased together could be considered as descriptive data mining task. Second, Predictive Data mining tasks, at this phase, we conduct conclusions based on the current data to end up making predictions about a certain situation depending on the case. This could be illustrated for example in the medical industry when a doctor tries to predict the illness of his patient based on the result of the medical analysis.

Data mining includes many functions, the mostly known and used one and the multiple types of patterns they can find are defined in the list below:

- **The association mining function:**

Association is a data mining function that discovers the probability of the co-occurrence of items in a collection (Oracle). The relationships between co-occurring items are expressed as association rules which allow finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in databases (Agrawal et al, 1993).

Association analysis is being used especially for commodity management, advertising and direct marketing.

- **The classification mining function**

Classification is a data mining technique that assigns categories to a collection of data in order to help in more accurate predictions and analysis. It is one of several methods intended to make the analysis of very large data sets effective (Mike Chapple, 2016).

Many classification techniques such as K-nearest neighbor algorithm, decision tree based method, naïve Bayes, etc. are used to extract models from the existing dataset to define the class of an instance based on its attributes values. Then, these models will be used to predict the class of new instances. Classification function is very useful when it comes to marketing since it leads to efficient results with lower costs.

- **The clustering mining function**

Clustering is the process of partitioning a set of data objects or observations into subsets. Each subset is a cluster, such that objects in a cluster are similar and related to one another, yet dissimilar and unrelated to objects in other clusters. These similarities and dissimilarities are based on the attributes values that describe the objects and often involve distance measures such as Euclidean distance, squared Euclidean distance, city-block distance, etc.

Before starting the clustering, the problem should be formulated, that is variables, that describe the similarity between objects on which the clusters are based on, have to be selected carefully because irrelevant variables may damage the results.

The clustering process is composed of several steps as illustrated in the following figure:

1.2 Data mining Goals and real-life applications:

Data mining is a new technology and an innovative way to acquire precious, valuable and new business concepts by studying and examining the information located in the database. This concept will enable the discovery of the business' strengths and hence helps moving from reactive to proactive decision making processed since data will be used in a more meaningful way.

These days, with the propagation of data mining concept and the growth of business owners' needs that can no longer be handled with traditional tools, it is the more and more obvious that data mining is the most suitable solution to be used in all sort of activities and domains.

Data Mining is primarily used today by companies with a strong consumer focus: retail, financial, communication, and marketing organizations, to go through their transactional data and determine pricing, customer preferences, product positioning, and impact on sales, customer satisfaction and corporate profits.

Moreover, the number of success stories of businesses using data mining is increasing continuously. It is important to notice that behind the success of a data mining project, a good data treatment and exploration exists. It is not about the huge amount of data you are able to collect, but how well you can use these data and transform it into useful information and knowledge (Wilson, 2001, p.26).

In the following paragraph, we will represent some real life application of data mining from business to scientific domain (John Wiley & Sons, 2011):

- **Financial Data Analysis:**

  Banks and most of the financial organizations are offering various services to their clients.

On one hand, in the banking industry, data mining is being strongly employed in many areas such as modeling, predicting credit fraud, measuring the risk, analyzing profit and when dealing with advertising campaigns.
On the other hand, in the financial industry, this new technology has supported the financial applications such as forecasting, portfolio management, commodity price prediction, etc.
Actually, these studies are achieved with the presence of high quality and reliable information existing in the industry database which will conduct, therefore, in enhancing the powerful presence of the industry in the market.

One example is "Mellon bank" in the USA which has used the information extracted from the credit-cards of its customers to define and predict their behaviors. Mellon bank employed IBM intelligent miner to develop a credit card-attrition to predict which customers will stop using its card in the coming months. Based on these results, it adopted new marketing strategies to retain its customers' loyalty.
Another example is at the financial organization "American express" which has developed several applications. The most useful one was "loan application screening". It employed different statistical methods to divide loan applications into 3 categories: those that should be accepted, those that should be rejected, and those which required a human expert to judge.

- **Retail industry:**

Retailers were the first to use data warehousing before it was propagated into other industries. They rapidly noticed its benefits in the decision making process, as managing the inventory and forecasting the sales have become more efficient. This early employment of data warehousing has given the advantage to retailers to benefit from data mining since they are collecting huge amounts of data about sales, service records, clients shopping history, etc.
With data mining techniques especially clustering, association and classification, and based on the collected information, retailers nowadays can easily classify customers into different segments based on their previous purchase and behavior. As a result, this will improve the services provided to each segment which will lead to customer satisfaction and retention. Finally, data mining helps retailers reduce the costs and increase the profit while adopting appropriate direct-marketing applications.

A real-world example of retail industry using data mining is with "Safeway", a grocery chains in the UK. Safeway generates over than $10 billion of sales and it uses IBM intelligent Miner to extract information from its products-transactions. One of the valuable results was that 25% of its loyal customers who spend the most are purchasing some products ranked below 200 in sales. Other groceries, without the data mining results, would stop selling these products which will disappoint its customers. But thanks to data mining techniques, Safeway continued in selling these products despite their low ranking and the most important thing, it retained its customers.

- **Customer Relationship Management:**

Certainly, there are plenty of other areas where the use of data mining is being essential today. Among all the needs behind using this technology, we notice that almost for all domains, building honest customer relationships is the main concern to guarantee a profitable business. This is why; customers should be at the center of an organization; because in some industries, the clients do not take into account the price of service as much as they do with the quality provided.

Data mining technology has made customer relationship management (CRM) a new area where firms can gain a competitive advantage since they are starting to realize that surviving a competitive marketplace requires closer relations with clients.
In fact, CRM is an enterprise customer-centric approach that companies use to manage and analyze customer interactions and data throughout the customer lifecycle, with the goal of improving business relationships with clients, assisting in customer retention and driving sales growth.
Normally, CRM is a framework of four elements: know, target, sell and service. Firstly, the company tries to know and discover the market and get detailed information about customers in order to select the customers that are worth targeting. Then, choose which product to offer to which customers. Finally, companies try to retain customers through different services. (Cap Gemini n.d.)
CRM in the Hotel industry

Nowadays, we find that CRM is highly being used in the hotel industry since the domain implicates to have a direct relationship with clients. Initially, information technology was viewed by hotel industry as a back office function that supports mainly the accounting and finance area. But, in the past decade, things have changed and hotel industry has progressed.
Hotel industry leaders have though a lot about the role of technology. Among the reached conclusions were “Going forward, technology will be the most competitive weapon for any hospitality company. If hospitality organizations want to compete successfully, they must do so by using technology to drive value to both the customer and to the firm”. (Kotlen et al. 1999)
In this domain, knowing well the customers; where they are from, how many nights they spend, etc. will help the company to plan marketing strategies and to generate profits. With the technological development, hotels are collecting large amounts of customers data which are organized in databases that can be used in the decision making process.
Yet, finding significant variables and relationships in these consumer-information systems can be a hard job if it is done manually, that’s why some hotel industries tried to use other technique such as data mining process technology to exploit the collected data.
A major task for data mining application for CRM in the hospitality domain is building appropriate segmentation and predictive models. In Tunisia, CRM concept is being slowly integrated into different industry and especially in the hotel industry where customer satisfaction matters the most. In this context, we have chosen to apply our skills in data mining to improve the services of a Tunisian hotel; The "Flamingo Beach Hotel Djerba". The objective is to extract knowledge about clients that will enhance the relationship with them, identify the most interesting profiles, conquer new customers, etc.
After exploring the data that are organized in the database, we will employ data mining functions with the purpose of meeting the following objectives:

- Retain the maximum number of clients and maximize the number of next visits.
- Maximize the number of extra services for each client which will increase itself the profits of the Hotel.
- Predict consumer-behavior and trends and the services that will be offered to future clients based on previous purchases.
- Create targeted seasonal promotions and plan personalized advertisements.

2 Data mining in the hospitality domain

To remain powerful and strong in a competitive market, hotels should develop strategies to maintain their loyal clients and target new customers. Thus, Hotels owners need to find out the most profitable ways to preserve the loyal customer relationship. Since the propagation of data mining, its functions and techniques have been one of the best tools helping hotels to understand their customers. There are many success stories nowadays that emphasize the benefits behind using data mining. Hotels reveal that they are using data mining in their business, yet they remain silent toward the algorithms and tools they use. Listed below, are three examples of hotels that have used data mining to improve their services:

- **Caesars Entertainment Corporation in Las Vegas:**

It is previously known as Harrah's hotels entertainment. It owns and operates over 50 casinos and hotels which had made of it a leading corporation in the hotel and gaming industry. It achieved this success when it had adopted customer services oriented strategy on data mining techniques. They launched a loyalty-card program “Total reward” and followed their customers’ purchases and activities (Loveman, 2003). Based on the collected information, Caesars proposed to its customers the most-effective incentives and offers.
Using data mining techniques, they developed quantitative models to predict lifetime value of its customer and used them to center marketing and service delivery programs in increasing customer loyalty (Bligh & Turk, 2004; Freedman, 2003).

In addition to that, Caesars found out that 26 percent of its customers represented 82 percent of the company revenue (Magnini, Honeycutt, & Hodge, 2003). As a result, the company continued in mining the data, modeling, and executing powerful marketing programs.

- **Island Shangri-La in Hong Kong:**

Island Shangri-La Hotel conducted a study to improve its targeting marketing strategies using its database which is consisted of different files containing multiple records about customers. The fields' records of the database describe the behavior and attitude of particular customers (Law, 1998).

Below (Figure 1) is a sample of the hotel database design containing three files: Room file, customer file and payment file. It is how the hotel stores its data:

Room File

Room Type	Name
SK	Smoking king
SQ	Smoking queen
HS	Hospitality suite
PS	Penthouse suite
NK	Non-smoking king
NQ	Non-smoking queen

Customer File

Name	Persons	Origin	Number of Nights Stayed	Room Type
Rob Jones	1	Permanent folios	2	HS
Kevin Adams	2	Weekend rate	2	NQ
Norm Smith	1	Corporate	3	NK
Rob Jones	3	Permanent folios	3	PS

Payment File

Name	Room Type	Payment Method	Rate Reason
Rob Jones	HS	AE Gold Card	Package
Kevin Adams	NQ	Cash	Special rate
Norm Smith	NK	Visa Classic	Special rate
Rob Jones	PS	AE Gold Card	Package

Figure 1 Hotel Database Sample

The data collected in these files is used to conduct special packages offerings to each type of customer. For example, if Island Shangri-La would like to offer a special package to visitor for three nights for a Penthouse suite, it will target customers with the same characteristics as Rob Jones rather than Kevin Adams or Norm Smith. This offer is made under the hypothesis that those people would be visiting Hong Kong when the hotel would be offering its special packages. Based on this logic, the hotel sends promotional brochures to different clients.

By mining in the database, the hotel can extract different patterns and build various models which will help in organizing specific marketing strategies. Because the selection criteria are now easier, the new method will reduce not only marketing costs but also time which can be in a result allocated to other beneficial activities.

- **South Korea's Hotels**:

A group of eleven different hotels in South Korea have adopted data mining technique approach to develop the profiles of hotel customers. This idea came when the revenue generated from tourism has doubled and as a result the hotel industry decided to expand the capacity and build new hotels to support the number of guests (Korean Tourism Organization, 2000).

Following this expansion, the competition between the hotels has increased gradually and the need to adopt powerful strategies to survive and to remain in the market is becoming more and more necessary. Hotels have to adapt their services to the changes. This cannot be achieved until the hotel managers develop detailed profiles of hotel "valued customers".

The hotel industry had made a study to figure out the profiles of customers returning to the same hotel, which customers generate more profit to a hotel than another, which segment of customers fit the service quality provided and which customers represent risk to move to another hotel.

The study is made on a group of clients who had stayed at 11 different luxury hotels sharing the same pricing characteristics, location and service amenities in Seoul, South Korea. The industry conducted a survey where the participants answered questions associated with their demographic profile, the frequency, the reason and the length of their visit to the hotel and whether they are satisfied or not with the service quality.

After the data collection, the industry applied data mining techniques to meet the objective and classify the customers.

It mainly chose decision trees among the different methods because of their visual appeal and simplicity in setting up useful "if-then" rules which will be used to give a clear customer segmentation to conduct targeted marketing and promotional strategies (Menon and Sharda, 1999).

Listed below is the collection of the most important "if-then" rules derived from the study:

The first result (Figure 2) summarizes the behavioral pattern of the customer's hotel choices which depends on their level of satisfaction toward the courtesy of the hotel employee.

Rule 1

If a customer is somewhat satisfied with the courtesy of hotel employees,
and the customer uses a credit card to pay for the hotel room charge,
and the customer is somewhat or very satisfied with the efficiency of a business center,
and the customer is somewhat satisfied with the variety and quality of leisure and/sport facilities of the hotel,
and the customer believes the some degree of importance in the promptness of check-in and -out to hotel service quality,
then the customer is likely to choose either the Hilton or Plaza hotel for his/her stay.

Figure 4 If-then rules for hotel choice

Rule 8

If a customer stays at the hotel for a convention,
then the customer is likely to be a manager.

Rule 9(a)

If a customer stays at the hotel for travel/pleasure,
and the customer's gender is male,
and the customer stays at either the Amiga or the Plaza hotel,
then the customer is likely to be an entrepreneur.

Figure 2 If then rules for the occupations of guests

The second rules (Figure 3) are about the occupation of the guests and how it affects the purpose behind the trip.

Other results show that nationality plays a crucial role on the behavior with regard to room choice and travel purpose and others focused on the age, gender of the clients and the length and frequency of visits.

Those rules contributed to the discovery of new patterns and knowledge in addition to the multiple correlations between the different attributes that constitute the profile of the customer. Therefore, the construction of a successful customer retention strategy is more accurate.

Data mining has proven its efficiency in the hospitality industry through different examples and the success stories behind it are too numerous to be mention. Despite these results, Tunisia has not taken advantages of these tools yet and this is probably illustrated with the poor services often provided by Hotels especially after the revolution. This how we have came with the idea to employ data mining in a Tunisian hotel and take advantage of its functions to improve its services.

3 Proposed application

3.1 Organization

The “Flamingo Beach” Hotel is a charming hotel opened in 2005 located on the northeastern coast of the island of Djerba. We propose to exploit our knowledge about Data mining techniques in order to improve the services and enhance the relationship with clients.

3.2 Data description

Since the opening of the hotel, data and clients’ information have been used to be registered manually in traditional ways. The only details entered into the database, in an excel file format, are about staying days and booking details. Figure 4 demonstrates a sample, of a specific month (July 2010), of how data is stored and organized.

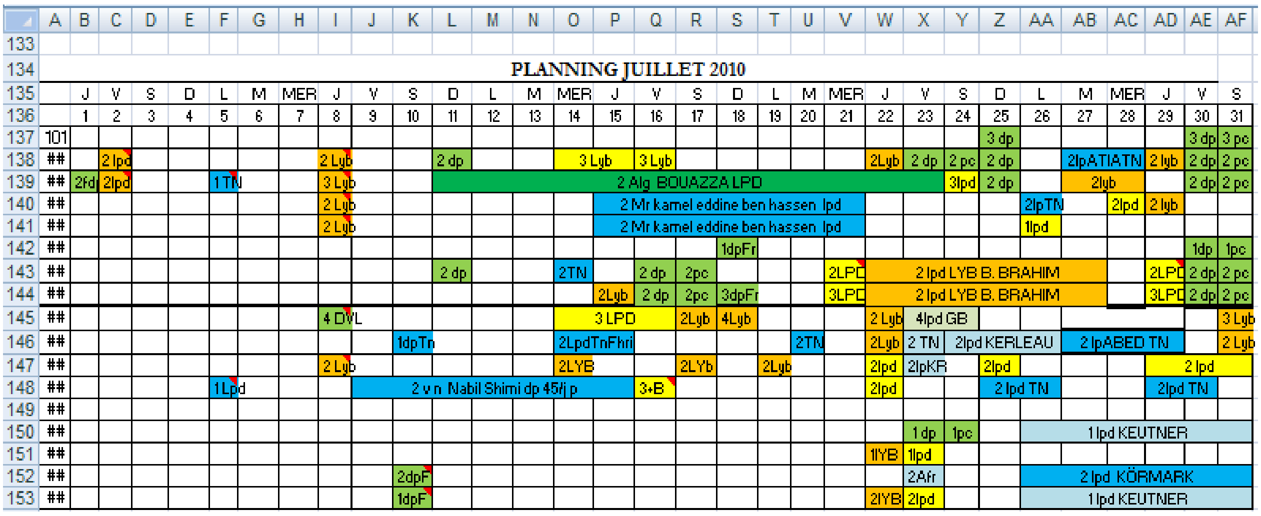


Figure 3 Booking details of the hotel

From this file, we can only extract the period and the length of the stay and the room type of a specific client. But the information related to each customer that will help us in our study is not registered here. In Fact, at the time of the registration, each customer is asked to fill in a form with information related to their demographic profile. After that, these forms are saved in the hotel archive and are rarely used.

In order to get valuable knowledge and interesting pattern which will help us to achieve our goals, we made a sample of 158 clients. We combined information from the excel files and the form and came out with a new file containing both the personal information of a customer and details about his visit. As a result, the new excel file contains 158 instances with 11 attributes represented in Figure 5.

Guest Name	Room Type	Gender	Nationality	AGE	Revenue	Indiv/TO	Season	Nb of visits	Duration	Formula	Extra services
THIERRY HENRION	DOUBLE	MALE	FR	54	ACCEPTABLE	INDIV	MIDDLE SEASON	2	2 NIGHTS	FULL BOARD	NOTHING
MONIQUE MICHEL	DOUBLE	FEMALE	BELGIUM	54	ACCEPTABLE	INDIV	MIDDLE SEASON	2	2 NIGHTS	FULL BOARD	NOTHING
AISSA BAKKA	FAMILY ROOM	MALE	LIBIYA	53	VERY HIGH	INDIV	HIGH SEASON	1	1 NIGHT	FULL BOARD	NOTHING
SAMIA BAKKA	FAMILY ROOM	FEMALE	LIBIYA	45	VERY HIGH	INDIV	HIGH SEASON	1	1 NIGHT	FULL BOARD	NOTHING
ABIR BAKKA	FAMILY ROOM	FEMALE	LIBIYA	20	VERY HIGH	INDIV	HIGH SEASON	1	1 NIGHT	FULL BOARD	NOTHING
TAHA BAKKA	FAMILY ROOM	MALE	LIBIYA	22	VERY HIGH	INDIV	HIGH SEASON	1	1 NIGHT	FULL BOARD	NOTHING
DANIELLE GOULEY	SINGLE	MALE	FR	64	HIGH	HP	LOW SEASON	1	1 NIGHT	HALF BOARD	DESERT TOUR
ANTONIO ANTONNO	DOUBLE	MALE	SPAIN	68	HIGH	HP	LOW SEASON	1	1 NIGHT	FULL BOARD	DESERT TOUR
ISABEL COSTA	DOUBLE	FEMALE	SPAIN	55	HIGH	HP	LOW SEASON	1	1 NIGHT	FULL BOARD	DESERT TOUR
MAGAN MLUSIA	DOUBLE	FEMALE	SPAIN	56	HIGH	HP	LOW SEASON	1	1 NIGHT	FULL BOARD	DESERT TOUR
FERNANDO ORDUVIA	DOUBLE	MALE	SPAIN	73	HIGH	HP	LOW SEASON	1	1 NIGHT	FULL BOARD	DESERT TOUR
CLAUDIA MARIA	SINGLE	FEMALE	GERMANY	44	VERY HIGH	HP	HIGH SEASON	5	16 NIGHTS	HALF BOARD	ISLAND TOUR
FAISAL BEN HASSEN	DOUBLE	MALE	TN	48	VERY HIGH	INDIV	HIGH SEASON	6	7 NIGHTS	B&B	BOAT TRIP
JANE KLEIN	DOUBLE	FEMALE	USA	36	VERY HIGH	INDIV	HIGH SEASON	6	7 NIGHTS	B&B	BOAT TRIP
SALEM ABDENNABI	DOUBLE	MALE	LIBIYA	55	VERY HIGH	INDIV	MIDDLE SEASON	2	3 NIGHTS	FULL BOARD	DESERT TOUR
KHADIJA ABDENNABI	DOUBLE	FEMALE	LIBIYA	45	VERY HIGH	INDIV	MIDDLE SEASON	2	3 NIGHTS	FULL BOARD	DESERT TOUR
JOANNA MEYLAN	SINGLE	FEMALE	SWITZERLAND	54	HIGH	HP	HIGH SEASON	4	7 NIGHTS	FULL BOARD	BOAT TRIP
SABINA SAUWAI	SINGLE	FEMALE	CHINESE	63	HIGH	HP	HIGH SEASON	4	7 NIGHTS	FULL BOARD	BOAT TRIP
LAI WING WENDY	SINGLE	MALE	CHINESE	53	HIGH	HP	HIGH SEASON	4	7 NIGHTS	FULL BOARD	BOAT TRIP
MOEZ BEN YAHMED	DOUBLE	MALE	TN	40	HIGH	FRAM	MIDDLE SEASON	1	3 NIGHTS	HALF BOARD	DESERT TOUR
SAFA BEN YAHMED	DOUBLE	FEMALE	TN	30	HIGH	FRAM	MIDDLE SEASON	1	3 NIGHTS	HALF BOARD	DESERT TOUR
CHRISTIAN LE BECQ	DOUBLE	MALE	BELGIUM	74	HIGH	FRAM	HIGH SEASON	3	6 NIGHTS	FULL BOARD	BOAT TRIP
DANIELLE LE BECQ	DOUBLE	MALE	BELGIUM	72	HIGH	FRAM	HIGH SEASON	3	6 NIGHTS	FULL BOARD	BOAT TRIP
KATY RANNOU	DOUBLE	FEMALE	FR	53	VERY HIGH	FRAM	MIDDLE SEASON	4	7 NIGHTS	FULL BOARD	ISLAND TOUR
ADEL MOSRATI	DOUBLE	MALE	TN	54	VERY HIGH	FRAM	MIDDLE SEASON	4	7 NIGHTS	FULL BOARD	ISLAND TOUR
ATIG BARNOUS	SINGLE	MALE	LIBIYA	44	HIGH	INDIV	LOW SEASON	1	1 NIGHT	B&B	DESERT TOUR
ROGER DESMARCHELIER	SINGLE	MALE	FR	40	HIGH	INDIV	HIGH SEASON	5	14 NIGHTS	HALF BOARD	BOAT TRIP
FREDERIC DUGUET	DOUBLE	MALE	FR	33	HIGH	FRAM	HIGH SEASON	4	7 NIGHTS	HALF BOARD	BOAT TRIP
HUGUETTE DUGUET	DOUBLE	FEMALE	FR	30	HIGH	FRAM	HIGH SEASON	4	7 NIGHTS	HALF BOARD	BOAT TRIP
HEINRICH KLEISLI	DOUBLE	MALE	SWITZERLAND	55	HIGH	HP	HIGH SEASON	5	21 NIGHTS	FULL BOARD	ISLAND TOUR
NELLY KLEISLI	DOUBLE	FEMALE	SWITZERLAND	52	HIGH	HP	HIGH SEASON	5	21 NIGHTS	FULL BOARD	ISLAND TOUR
MARTIN AUDREY	DOUBLE		FR	50	ACCEPTABLE	FRAM	LOW SEASON	1	7 NIGHTS	FULL BOARD	NOTHING

Figure 4 Database sample

These attributes are divided into demographic details about the client (gender, age, occupation, and nationality), details about their staying period (room type, duration, season, formula) and other extra information. At first sight, we found out that the clients represent many different nationalities and demographic sector as shown in Table 1.

Demographic features	Frequency	Percentage
Gender:		
(1)Male	83	52,53
(2)Female	75	47,47
Age		
(1)under 20	5	3,16
(2) 20-29	8	5,06
(3) 30-44	37	23,42
(4) 45-59	68	43,04
(5) Over 60	40	25,32
Revenue (%occupation)		
(1) Acceptable	16	10,13
(2) High	96	60,76
(3) Very high	46	29,11
Nationality		
(1) TN	21	13,29
(2) FR	74	46,84
(3)SWITZERLAND	7	4,43
(4) LIBYA	17	10,76
(5)OTHERS	39	24,68

Table 1 Demographic features of the Hotel customers

Unfortunately, these data do not provide any useful information without a deep study. That's why; we will take the advantage of applying data mining functions and techniques on our data to discover hidden patterns and build new models. The description of the selected attributes is shown in table 2:

Attribute	Description	Type
Room type	The type of the booked room. It could be a single, double or family room.	Nominal
Gender	The Gender of the client.	Nominal
Nationality	Then nationality of the client.	Nominal
Age	Client's age obtained from the formula	Numeric
Revenue	The revenue the clients generates defined as acceptable, high and very high according to their occupation	Nominal
Indiv/TO	This attributes indicates whether the clients booked in the hotel from a tour operator (TO) or by himself.	Nominal
Season	The season is divided into 3 categories: low, middle and high.	Nominal
Formula	The hotel provides 3 sorts of formula: Half board, full board and breakfast&bed(B&B)	Nominal
Extra services	The client can purchase different types of services which are: Desert tour, Island tour or Boat trip.	Nominal
Next visit	This attributes specifies whether the client visited again the hotel or not.	Nominal

Table 1 Attributes' description

Details about the used functions and algorithms of data mining are described and explained in the next section.

3.3 Proposed data mining functions

In our work, we choose to experiment the main data mining functions used in the hospitality domain, which are: clustering, classification and association rules. In this section, we explain in details the functions and algorithms implemented in our study.

Our dataset consists of an ensemble of 158 instances X={x_1, x_2… x_{158}} characterized by 11 attributes A= {a_1=room type, a_2=gender, a_3=nationality, a_4=age, a_5=revenue, a_6=indiv/TO, a_7=season, a_8=duration, a_9=formula, a_{10}=extra services, a_{11}=nb of visits}.

3.3.1 Clustering: K-means algorithm

Segmentation is very important for hoteliers; it divides customers according to their purchasing behavior or their demographic characteristics into different groups which help them understand and know who their customers are. Each group can be targeted through a specified marketing campaign that matches the customer's needs and budget level. The segmentation process can be ensured using clustering.

Various algorithms are provided in the literature but we will employ the K-means algorithm because of it is simple to use and provides relevant knowledge (Jain).

The algorithm clusters the **n** observations into **K** clusters, where **K< n** with K the number of clusters desired entered by the programmer. Once the K parameter is selected, the algorithm then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again.

First, it randomly selects K points as cluster centroid which is the mean values of the variables for all the cases or objects in a particular cluster. Second, each object in the dataset is assigned to the closest cluster center based on the Euclidean distance (or other distance measures) between the two points which is the frequently used measure to calculate distance between instances and clusters. The Euclidean distance is calculated as follows:

$$Euclidean\ distance = d(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

Third, the cluster center of all objects in the cluster is recalculated. Finally, the previous steps are repeated until the clusters converge and the same objects are assigned to each cluster in consecutive rounds (Mike Chapple, 2016).

3.3.2 Classification: Decision trees

As we mentioned before, our dataset consists of 158 instances characterized by 11 attributes. In supervised learning, labeled data is used where every instance consists of an ensemble of attribute values and a required output value. With a specific supervised learning algorithm, the training data will generate a function which will be applied on new instances. These algorithms will permit to correctly define the class labels for new instances.

Classification is a supervised learning function which is very beneficial for Hotels. It organizes customers into predefined segments that allow the monitoring of the market groups' structure. Classification uses the information contained in the dataset such as the demographic and lifestyle data in order to build predictive models that can classify activities.

There are various algorithms that ensure classification function in the literature.

In our study, we are mainly going to exploit J48 algorithms for building decision trees.

A decision tree is a tree-like graph decision support tool that lists all possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm (Agrawal et al. 1993).

It is known among the most simple and easiest algorithms, according to Rohit Arora and Suman from Hindi College of Engineering, and a set of rules can be extracted easily. It is as accurate as efficient when it comes to classification which has made it the mostly used between all algorithms.

A decision tree classifies tuples of attributes by sorting them from root to leaf node also known as classes. The components of a decision tree are illustrated in the following figure:

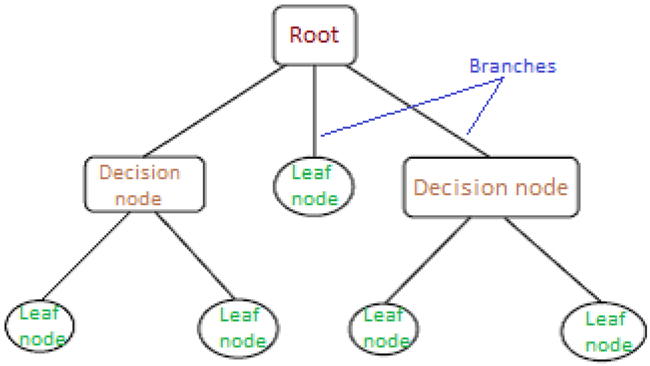


Figure 6 Decision Tree Components

The first thing to do before building the tree is to choose the first attribute or the root that will split the data. This step is ensured using a statistical property called the information gain. Information gain is the difference between the entropy, which is measure of the uncertainty contained in a piece of information, before and after a decision (Giura, 2016).

After that, the algorithm continues to build the decision tree, by evaluating the remaining attributes under the initial branches.

Data mining tools allow building the decision tree from which the "if-then" rules could be derived and provide statistical output as well that gives deeper insight on the classified data.

3.3.3 Association rules: Apriori algorithm

In the hospitality domain, association rules include the detection of connections and relations between the instances. They detect what attributes enhance the length of the stay and determine why specific profiles have higher probability to return than other profiles.

We will apply the Apriori algorithm which is the most popular algorithm to extract the association rules from the databases. Existing data mining tools enable us to sort the association rules according to different metrics in order to select the most interesting ones.

The most important and used metrics in the Apriori algorithm are:

- Support: it determines how often a rule is applicable to a given dataset
- Confidence: Given the occurrence of the RHS, each LHS has an associated confidence representing its probability to occur.
- Lift: Lift indicates the strength of a rule over the random co-occurrence of the antecedent and the consequent, given their individual support. It provides information about the improvement, the increase in probability of the consequent given the antecedent

Below is a summary of the rules calculation:

$$Rule:\ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Detailed explanation of each function can be found in Section 1.

4. Implementation and experiments

For our experiments, we use Weka tool for data mining. WEKA recognizes the attributes and evaluates each one; it displays the frequency of each value if the attribute is categorical and for continuous values, it calculates the mean, min, max and the standard deviation. These statistics will help get an overall idea about the attributes value of the dataset as shown in the example given in Figure 7.

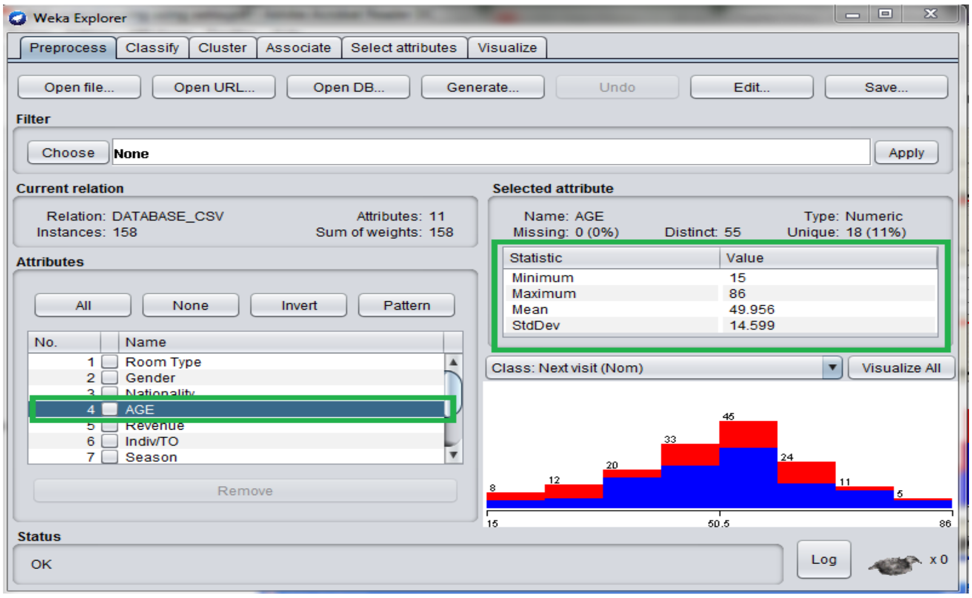


Figure 7 Attribute statistics

In some cases and with certain functions, it would be necessary to conduct some modification to the attributes to make the algorithm implemented. These modifications could be ensured with the use of the set of filters provided with WEKA.

If we deal with a classification task, we can visualize all the attributes and see how their values are assigned to each class label. In the example of Figure 8, the class to predict is "Next visit" and:

- The class “YES” is represented with the blue color.
- The class “NO” is represented with the red color.

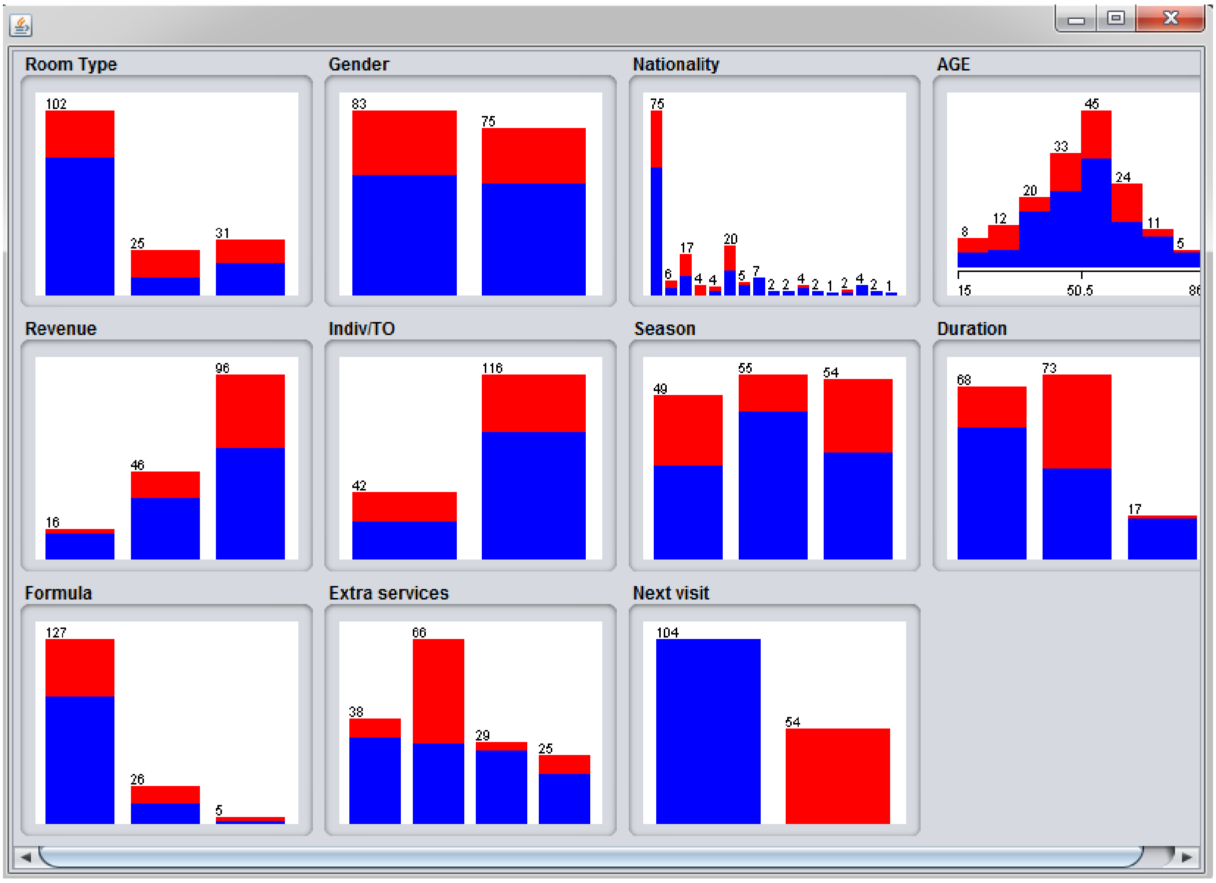


Figure 8 Attributes’ visualization

In some cases and with certain functions, it would be necessary to conduct some modification to the attributes to make the algorithm be implemented. These modifications could be ensured with the use of the set of filters provided with WEKA.
Shown below the results we found after we applied the different data mining functions on our dataset and the most valuable information.

4.1 Clustering

4.1.1 Experimental Setting

To perform clustering, we go to the cluster tab in the WEKA explorer and then choose the appropriate algorithm which will be in our case the SimpleKMeans.
SimpleKMeans algorithm in WEKA deals with both the numeric and nominal attributes, so there is no need to do any modification to the dataset. When computing the distance measures, the algorithm normalizes the numeric attributes automatically. After choosing the algorithm, we can modify the parameter in conducting the clustering. By default, the number of clusters is set to 2, the distance measure is the Euclidean distance and the number of seed is 10.

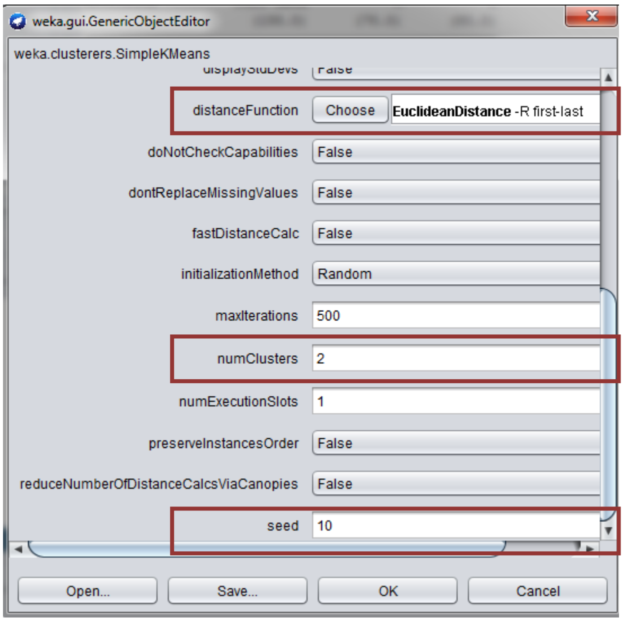


Figure 9 Clustering parameters

The seed value is used in generating a random number which is, in turn, used for making the initial assignment of instances to clusters (Mobasher, 2010).
Also, the clustering is based on the cluster mode chosen as shown in the Figure 10:

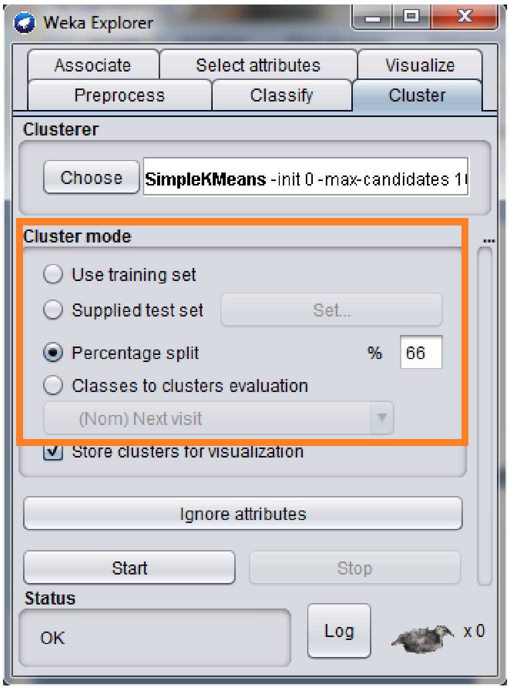


Figure 10 Cluster mode

- **Use training set** (default): WEKA classifies the training instances into clusters according to the cluster representation and computes the percentage of instances falling in each cluster.
- **Supplied test set** or **Percentage split:** WEKA can evaluate clustering on separate test data.
- **Classes to clusters evaluation:** At first, the class attribute is ignored and WEKA generates the clustering. Then during the test phase it assigns classes to the clusters, based on the majority value of the class attribute within each cluster.

In the experiments, all the 158 instances with the 11 attributes were used and the percentage split was used as validation protocol. Now that our parameters are specified we can start our clustering.

4.1.2 *Results analysis*

✓ *Experiment 1:*

For the first experiment, we set the number of clusters to 10 and left the other parameters with their default values.

After performing the algorithm we got the results of the 10 clusters shown in the Figure 11. The result window shows the number of iterations done in order to get the final clusters and the within cluster sum of squared errors (SSE). SSE is the sum of the squared differences between each observation and its group's mean. It can be used as a measure of variation within a cluster. If all cases within a cluster are identical the SSE would then be equal to 0. In our case the SSE is equal to 371.96. In addition to that, it displays the centroid of each cluster with the percentage of instances assigned to each one.

Then, the results show the initial randomly picked clusters from which the process begins. Based on the Euclidean distance, each instance will be assigned to the closest cluster and the cluster centers of the objects are recalculated. These steps are repeated until all instances are assigned to a cluster and their assignments no longer change. We got our 10 clusters as illustrated in Figure 11.

Final cluster centroids:

		Cluster#									
Attribute	Full Data	0	1	2	3	4	5	6	7	8	9
	(158.0)	(4.0)	(27.0)	(7.0)	(9.0)	(14.0)	(20.0)	(16.0)	(8.0)	(20.0)	(33.0)
Room Type	DOUBLE	FAMILY ROOM	DOUBLE	DOUBLE	FAMILY ROOM	SINGLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE
Gender	MALE	MALE	MALE	MALE	FEMALE	MALE	MALE	MALE	FEMALE	FEMALE	FEMALE
Nationality	FR	LIBIYA	FR	BELGIUM	FR	FR	TN	FR	FR	FR	FR
AGE	49.9557	35	55.1852	66.8571	42.1111	57.2857	49.15	48.4375	63	40.7	46.6061
Revenue	HIGH	VERY HIGH	HIGH	HIGH	HIGH	HIGH	VERY HIGH	HIGH	HIGH	HIGH	HIGH
Indiv/TO	TO	INDIV	TO	TO	TO	TO	INDIV	TO	TO	TO	TO
Season	HIGH SEASON	HIGH SEASON	HIGH SEASON	MIDDLE SEASON	LOW SEASON	LOW SEASON	MIDDLE SEASON	LOW SEASON	LOW SEASON	MIDDLE SEASON	HIGH SEASON
Duration	Short	Short	Medium	Medium	Short	Short	Short	Medium	Long	Short	Medium
Formula	FULL BOARD	FULL BOARD	FULL BOARD	FULL BOARD	FULL BOARD	FULL BOARD	HALF BOARD	FULL BOARD	FULL BOARD	FULL BOARD	FULL BOARD
Extra services	DESERT TOUR	NOTHING	NOTHING	BOAT TRIP	DESERT TOUR	DESERT TOUR	DESERT TOUR	DESERT TOUR	DESERT TOUR	DESERT TOUR	BOAT TRIP
Next visit	YES	NO	YES	YES	YES	NO	YES	YES	YES	NO	YES

=== Model and evaluation on training set ===

Clustered Instances

0	4	(3%)
1	27	(17%)
2	7	(4%)
3	9	(6%)
4	14	(9%)
5	20	(13%)
6	16	(10%)
7	8	(5%)
8	20	(13%)
9	33	(21%)

Figure 11 Cluster centroids with k=10

At first sight, we notice that the customers belong to a certain class, they all generate high revenues. They are also in the same age interval. This gives an idea about the market segments will be targeting. Almost for all clusters, the clients (foreign clients) come to the Hotel with a tour operator who is dealing with the "Flamingo Beach Hotel" except for ***cluster 0*** and ***5*** which represents segments of Tunisian and Libyan clients and come to the hotel on their own. We can note also that those segments represented in ***cluster 0, 4*** and **8** do not visit the hotel again. This is explained by the fact that they come for a short period which represents one night. We can predict that they visit Djerba for a business purpose or it is just a station toward their final destination and chose “Flamingo Beach Hotel” to spend the night.

✓ *Experiment 2:*

After this first experiment, we conduct another one with K=5 as shown in Figure 12.

For this clustering, the number or iterations executed to get the final result is 4 and the value of SSE decreased to 293.76 which is better.

For example the centroid for ***cluster 0*** shows a segment of cases representing French male guests with an average age of 49.6. They booked a double room with full board formula for a short duration during the middle season. These guests are generating high revenue and took the desert tour as an extra service. They have chosen the hotel with a tour operator and they did not come a second time.

Approximately, the clusters’ centroids have the same features. They are all between the age of 44 and 53 and they generate high revenue. They mainly differ in the season and the extra services they chose which will affect the probability of the next visit. In addition to that, the clusters show that "French" is the nationality present in all clusters' centroids, and also Tour operator "TO" as travel mode.

```
Number of iterations: 4
Within cluster sum of squared errors: 293.7629395696884

Initial starting points (random):

Cluster 0: SINGLE,MALE,FR,55,'VERY HIGH',TO,'LOW SEASON',Short,'FULL BOARD','DESERT TOUR',NO
Cluster 1: 'FAMILY ROOM',FEMALE,FR,50,HIGH,TO,'HIGH SEASON',Medium,'FULL BOARD','BOAT TRIP',NO
Cluster 2: DOUBLE,MALE,UK,40,HIGH,TO,'LOW SEASON',Long,'FULL BOARD','DESERT TOUR',YES
Cluster 3: DOUBLE,FEMALE,FR,29,'VERY HIGH',TO,'MIDDLE SEASON',Short,'FULL BOARD',NOTHING,YES
Cluster 4: DOUBLE,MALE,TN,30,HIGH,INDIV,'MIDDLE SEASON',Short,'FULL BOARD','DESERT TOUR',NO

Missing values globally replaced with mean/mode

Final cluster centroids:
```

Attribute	Full Data	Cluster# 0	1	2	3	4
	(158.0)	(30.0)	(45	4.0)	(31.0)	(18.0)
Room Type	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE
Gender	MALE	MALE	MALE	FEMALE	FEMALE	FEMALE
Nationality	FR	FR	FR	FR	FR	FR
AGE	49.9557	49.6	52.9778	53.2647	44.9677	45.3333
Revenue	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH
Indiv/TO	TO	TO	TO	TO	TO	TO
Season	HIGH SEASON	MIDDLE SEASON	HIGH SEASON	MIDDLE SEASON	LOW SEASON	MIDDLE SEASON
Duration	Short	Short	Medium	Medium	Short	Short
Formula	FULL BOARD	FULL BOARD	FULL BOARD	FULL BOARD	FULL BOARD	FULL BOARD
Extra services	DESERT TOUR	DESERT TOUR	NOTHING	BOAT TRIP	DESERT TOUR	DESERT TOUR
Next visit	YES	NO	YES	YES	YES	NO

Cluster 0

Cluster 0

```
Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       30 ( 19%)
1       45 ( 28%)
2       34 ( 22%)
3       31 ( 20%)
4       18 ( 11%)
```

Figure 12 Clustering Output

The marketing strategies for these clusters will be based on the season and the extra services.

4.1.3 Discussion on clustering results

The main objective toward these segments would be certainly the enhancement of the pobability of next visit. This will be ensured only if the clients spend longer period in the Hotel and discover the personnalised provided services. Thus, the strategy would be to send seasonal promotion, to either the customer himself or the tour operator he deals with, with details of the activities that are appropriate to that season.

For the rest of the clusters, they represent segments of loyal customers. The strategy will differ because they are loyal customers that the Hotel wants to retain.

We will plan to personalize advertisement for each segment based on the criteria and services they had purchased before. To maintain a good relationship with customers, we will present customized offers for special occasion to each customer of a targeted cluster; marriage anniversary for married guest and birthday for single customers. Because, from previous actions, we notice that these behaviors please the customers.

Targeting only interesting cluster profiles will reduce advertisement and personal services costs, and will maximize the perentage of satisfaction and loyalty of these customers in return.

4.2 Classification

For our case study, we are interested by predicting if a given client profile will have a next visit to the hotel or not. This will allow targeting these interesting profiles with personalized services. A classification algorithm is used for this purpose.

4.2.1 Experimental setting

In our study, we used the J48 algorithm. There is no need to do any change to the attributes values because the algorithm deals with both the numeric and nominal attributes contrary to the ID3 algorithm which only handles nominal attributes and we use Cross validation protocol in which the dataset is split into K non-overlapping subsets and K models are built. Each model is built on all subsets except the k^{th} one which is used to evaluate performance of model k. The average of the K performances is given as result.

Finally, we must define some evaluation metrics that will be used in the analysis. The J48 algorithm displays the tree in an ASCII version in addition to the statistical results obtained using the following evaluation metrics:

- ✓ **TP Rate** (true positives rate): the rate of the correctly classified instances.
- ✓ **FP Rate** (false positives rate): the rate of the incorrectly classified instances.
- ✓ **Precision:** proportion of instances that are truly of a class divided by the total instances classified as that class.
- ✓ **Recall:** proportion of instances classified as a given class divided by the actual total in that class. It gives the same results as the TP rate.

- ✓ **F-Measure**: A combined measure for precision and recall calculated as:

$$F - Measure = \frac{\mathbf{2 * Precision * Recall}}{\mathbf{Precision + Recall}}$$

- ✓ **Confusion Matrix:** A table that describe the classification model with the TP, TN, FP, FN rates

After we defined and specified the needed parameters we click on the “start” button to generate the model.

4.2.2 Results analysis

The experiment was performed to predict whether a customer would visit again the Hotel or not. The model took 0.17 seconds to be built; it is of size 31 with 21 leaves. The classification results are illustrated in the following figure:

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

Correctly Classified Instances         136               86.0759 %
Incorrectly Classified Instances        22               13.9241 %
Kappa statistic                          0.6849
Mean absolute error                      0.1922
Root mean squared error                  0.351
Relative absolute error                 42.6296 %
Root relative squared error             73.9647 %
Total Number of Instances              158

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,913	0,241	0,880	0,913	0,896	0,686	0,864	0,900	YES
	0,759	0,087	0,820	0,759	0,788	0,686	0,864	0,757	NO
Weighted Avg.	0,861	0,188	0,859	0,861	0,859	0,686	0,864	0,851	

Figure 13 Statistical measures of J48

The results display that 136 instances were correctly classified with 86.07% while 22 instances were incorrectly classified with 13.92%. The model represents a good result with high accuracy rate. To efficiently evaluate the performance of the new model for each class we should take a look at other measures. The model presents a TP value of 0.913 which is also a good performance and has a tendency to identify false appearance represented with the FP rate of 0.241. Another method to calculate the the accuracy rate is from the confusion matrix:

$$accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

$$= \frac{95 + 41}{94 + 41 + 9 + 13}$$

= 0.86075

Predicted

Actual		YES	NO
	YES	95	9
	NO	13	41

Table 3 Confusion matrix of the model

Although those measures give a good result, they are sometimes not reliable because they are sensitive to class distribution. It will be better to look at another measure, the ROC area value. The Closer the value to 1, the better the classification is. The ROC area value in our case is equal to 0.864 which is close to 1.

Another value to analyze is the F-measure which is in this case equal to 0.896. As defined earlier, the F-measure combines both the precision and recall value and since the value is good we can conclude that the dataset is *balanced*.

With J48 algorithm, we can visualize the decision tree (Figure 14) where we can see clearly the predicted classes.

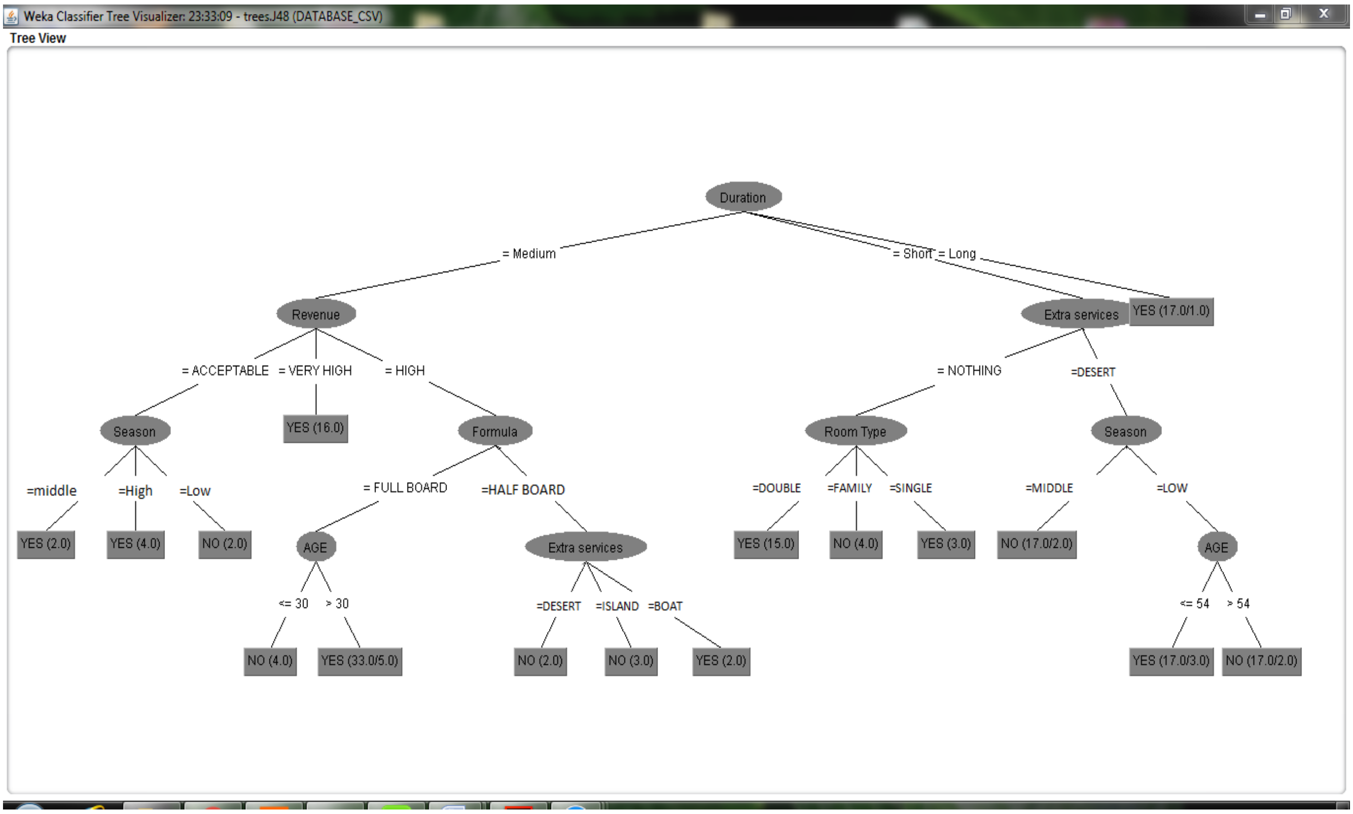


Figure 14 Decision tree of the model

From the decision tree, we can have more meaningful results to analyze. We can notice that not all the attributes were used in the implementation of the decision tree and that the root attribute that has the highest information gain value is the DURATION.
We had made an analysis based on the decision tree and on the experience of an expert in the hospitality domain. We will start the analysis from the root:

✓ *Medium Duration (2-7 nights)*

If the customers stayed for a *medium* duration and their revenue is *very high*, it's evident that they will come another time since they like the lovely room view and the service.
It their *revenue* is *acceptable* and we are focusing on the *middle* or the *high season*, they will also come next time, but those who are coming during the *low season*, they will probably choose another destination for next vacations.
Now, if the *revenue* is *high*, we will see the *formula* and the *extra services choice*. If the category of the visitors are salt and pepper hair (age>30), they will appreciate the hotel and visit it another time. They will become loyal guests of *full board* booking since they like the charm, the location, the food and the personalized service that is no more considered a luxury any more for them. But those who are *less than 30*, they will maybe try another hotel category next time where they might find night clubs and sport activities exactly like those who took *half board* formula and went on an *island* or *desert tour*. Also, visitors interested by the *boat trip extra service* will appreciate this adventure and come another time.

✓ *Long Duration (more than 7 nights)*

From the other side, customers who come for *long* vacations will also visit this charm hotel next time. They are generally retired customers of *more than 60 years old* who are evading from stress and had become loyal customers visiting the Hotel every season.

✓ *Short Duration (1 night)*

This part depends on the Extra service. If there are no extra services and the customer booked a *double* or a *single* room, he will probably come again either alone or in couple. But for customers

who booked a family room, they will change their destination next time either because they came for a business trip or they generate low revenue.
For those who come for *desert tour*, they return if they visit Djerba and the Flamingo Beach Hotel only during their winter vacations (low season) since our desert is very warm during the other seasons and this explains, in the latter case, why they do not return again.

4.2.3 Discussion of classification result

From the decision tree and without being an expert in the Hospitality domain, we can now understand the different profiles of customers visiting the Flamingo Beach. At their coming and based on their demographic information and especially on the *Duration* of the stay, we can suggest to the customer the services that he will probably appreciate based on the similarity between his profile and the old profiles from which we built the model. For the customers who may match the profile of those that will not visit the hotel again, this strategy may have an impact on them and makes them visit the hotel again, not because of the services provided but of the special attention to be given to them. The experience of the hotel's owner allowed him to notice that what the customers valued the most was the way they were treated. With the use of the classification function, we now have an internal strategy to be implemented in addition to the external one. We no longer need an expert to deal with the customers, from the built model and with some training we can provide the clients with the best suited services for them.

4.3 Association rules

The Apriori algorithm for association rule function can only be implemented on categorical or nominal attributes. Thus, before starting the implementation we should convert the numeric attribute "age" in our dataset to nominal type. This can be ensured by the "NumericToNominal" unsupervised filter for attributes.

4.3.1 Experimental setting

As we did with the other algorithms, we should specify the parameter to be used first.

The default metric set by WEKA is confidence with a minimum value of 0.9. To determine whether a rule is valid or not, we should consider both the support and confidence values (defined in Section 2). But they are not enough to examine the goodness of the rule. That's why we will choose the lift, also known as the improvement, as an evaluation metric to sort the rules. It is evident that if the LHS and the RHS occurred together; this is not random but because of a relationship existing between them. The lift value should be greater than 1. The higher the value, the greater the rule.
In the drop-down list, we select lift to be the metric type and enter 1.5 as the minimum metric value. We can also change the number of rules which is set by default to 10. The upper bound for minimum support is set to 1.0 and the lower bound to 0.1. Apriori algorithm in WEKA starts with the upper bound support and incrementally decreases support by delta increments which by default are set to 0.05. The algorithm stops when either the specified number of rules is generated, or the lower bound for min support is reached. The significance testing option is only applicable in the case of confidence and is by default not used (-1.0).
After converting the attributes value to nominal and choosing the appropriate parameters we can run the program.

4.3.2 Results Analysis

In the window "associator output" the program displays the run information and the discovered rules. The minimum support for the rules is 0.25 and the number of cycles performed is 15.
For each rule, we notice that it includes the support count for both RHR and LHS, as well as the confidence, lift, leverage and conviction values. Let's remind that we had explained the first two values in Section 2. The leverage has the same measurement idea as lift, except that it measures the difference between the probability of co-occurrence of the antecedent and consequence of the rule as the independent probabilities of each side. The value of leverage should be higher than 0 in order

to be acceptable. The conviction is also similar to lift, but it measures the effect of the antecedent if it is false.
To measure the efficiency of the rules, it is enough to concentrate on the combination of support and lift values. The association rules are illustrated in the following figure:

```
Best rules found:

 1. Season=LOW SEASON 54 ==> Extra services=DESERT TOUR 45    conf:(0.83) < lift:(1.99)> lev:(0.14) [22] conv:(3.14)
 2. Extra services=DESERT TOUR 66 ==> Season=LOW SEASON 45    conf:(0.68) < lift:(1.99)> lev:(0.14) [22] conv:(1.97)
 3. Room Type=DOUBLE Indiv/TO=TO Formula=FULL BOARD 70 ==> Nationality=FR Next visit=YES 39    conf:(0.56) < lift:(1.69)> lev:(0.1) [15] conv:(1.47)
 4. Nationality=FR Next visit=YES 52 ==> Room Type=DOUBLE Indiv/TO=TO Formula=FULL BOARD 39    conf:(0.75) < lift:(1.69)> lev:(0.1) [15] conv:(2.07)
 5. Room Type=DOUBLE Nationality=FR Formula=FULL BOARD 46 ==> Indiv/TO=TO Next visit=YES 39    conf:(0.85) < lift:(1.67)> lev:(0.1) [15] conv:(2.84)
 6. Indiv/TO=TO Next visit=YES 80 ==> Room Type=DOUBLE Nationality=FR Formula=FULL BOARD 39    conf:(0.49) < lift:(1.67)> lev:(0.1) [15] conv:(1.35)
 7. Duration=Short 73 ==> Extra services=DESERT TOUR 51    conf:(0.7) < lift:(1.67)> lev:(0.13) [20] conv:(1.85)
 8. Extra services=DESERT TOUR 66 ==> Duration=Short 51    conf:(0.77) < lift:(1.67)> lev:(0.13) [20] conv:(2.22)
 9. Room Type=DOUBLE Nationality=FR 50 ==> Indiv/TO=TO Formula=FULL BOARD Next visit=YES 39    conf:(0.78) < lift:(1.64)> lev:(0.1) [15] conv:(2.19)
10. Indiv/TO=TO Formula=FULL BOARD Next visit=YES 75 ==> Room Type=DOUBLE Nationality=FR 39    conf:(0.52) < lift:(1.64)> lev:(0.1) [15] conv:(1.39)
```

Figure 15 Association rules of the Apriori algorithm

The best 10 rules were sorted according to the specified parameters and the threshold values for support and lift. We will analyze these rules and highlight the most important patterns.

- ✓ ***Rule 1 & 2:*** the first 2 rules have a lift value of 1.99. They illustrate the relationship between the attributes Season and Extra services. Among 54 customers who visited the "Flamingo Beach" during the low season 45 of them go on a Desert tour ***(rule1)*** and among 66 customers who visit the desert 45 of them come during the low season ***(rule2).***
- ✓ ***Rule 3, 4, 5, 6, 9 & 10:*** Those rules depict the relationship that exists between the room type of value "double", the "full board" formula, whether the client comes alone or with a tour operator, the French nationality and the next visit attribute's value "YES". The related association rules have a lift value that lies between 1.69 and 1.64…
- ✓ ***Rule 7 & 8:*** These 2 rules describe the connection between the duration attribute with value "short" and the "Desert Tour" as an extra service. The lift value of these rules is equal to 1.67. Most of the customers who came for a short period visited the Desert and vice versa.

4.3.3 Discussion of association rules results:

From the depicted rules, we have now an overall idea about the associations that exist between the different attributes. We can assume that two main reasons exist behind the clients' visit. There are those who come during the low season, go on a Desert tour and come to our hotel to only spend the night. So, the number of booking increases during this season and it should be taken into consideration to prepare enough rooms to host the customers. The other category of clients is especially retired French people. They come through a tour operator the hotel is conventionned with. All they search for is the clam and the friendly atmosphere to spend their vacation.

4.4 Discussion

Building a strong relationship with the customers does not mean collecting huge information about them. In fact, the size of data is not as important as the way it is exploited and the knowledge you get from it. To get valuable results and try to achieve our objective our goals in predicting the customer's behavior, we implemented the different mining techniques.
First, we conducted a clustering analysis using the SimpleKMeans algorithm to group the clients into different segmentation based on their past personal and behavioral characteristics. We discovered interesting information about the different segments. But despite these results, we noticed that the model displayed a little bit high within cluster sum of squared errors value. That is, instances in the

same cluster may have dissimilarities. This is why we decided to use another data mining function, the classification. We used the J48 algorithm to build the decision tree and the model built was successful. The accuracy rate was satisfying and high and all the other measures including the F measure and ROC area emphasized the goodness of the model and the balanced dataset. Finally, to have deeper information, we implemented a third function which is Association rules using the Apriori algorithm. The algorithm resulted in a set of association rules describing the existing relationships and connections between the different attributes which describe the reason of the visit. From the previous experiments, we managed to extract the most interesting and attractive profiles visiting the hotel. Since the “Flamingo Beach” is a charming hotel, it attracts mainly retired and middle age customers who are looking for the calm and the personalized services provided in the hotel. Those customers will get personalized services in special occasion with a list of activities that are appropriate to the season. Furthermore, we notice that the low season is the season when customers prefer to visit the desert. That’s why; we will plan seasonal promotions. In additional to the marketing strategy planned, internal actions could be done with the customers without the presence of an expert of the domain. This is thanks to classification model and the association rules result.

5 Conclusion

Given that CRM is very important in retaining and maintaining relationships with customers, this study addressed the issue of applying data mining techniques to extract useful knowledge of a hotel' customers. The Dataset used in our study was gathered from "The Flamingo Beach Hotel" customers' information and booking's details. After collecting the data, it was prepared in appropriate way to be used. Then we performed appropriate data mining functions, namely the data clustering, classification and finally the search of association rules.

From this study, we could extract the most segments of customers to be targeted and according to which criteria which will lower the advertising costs and increase the revenue. We were also able to build a model from which we can predict the customer loyalty and finally we discovered the association between the attributes that characterize the profile. Data mining is the key to improve any business and help get a competitive edge exactly like what we presented and what we did and implemented on this study.

Although we had come with great results, they would have been better if we could integrate all customers' information. But, due to the limited time we could only retrieve data of 158 customers. Our future plan will be to integrate all the information in the dataset, because the more you collect from customers, the more value you can deliver to them, and the more revenue you can generate as a consequence.

References

Agrawal, R.; Imieliński, T.; Swami, A. (1993). "*Mining association rules between sets of items in large databases*". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207-216

Chapple, M. (2016, November 25). *Classification in Data mining.* Retrieved from Thoughtco: https://www.thoughtco.com/classification-1019653

Chapple, M. (2017, march 30). *K-means clustering.* Retrieved from thoughtco: https://www.thoughtco.com/k-means-clustering-1019648

Data mining concept. (n.d.). Retrieved from Oracle: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/algo_apriori.htm#DMCON061

Data mining functions. (n.d.). Retrieved from IBM: https://www.ibm.com/support/knowledgecenter/en/SSEPGG_9.5.0/com.ibm.im.easy.doc/c_dm_functions.html

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.

Gemini, I. &. *our elements of customer relationship management.* CapGemini White Paper.

Giura, L. (2016, January 31). *Decision trees using WEKA.* Retrieved from Technobium: http://technobium.com/decision-trees-explained-using-weka/

Jain, A. K. *Data Clustering: 50 Years Beyond K-Means.* Michigan, Michigan State University.

John Wiley & Sons, I. (2011). *Data mining: Models, Methods, and Algorithms, Second Edition. Mehmed Kantardzic.* Institute of Electrical and Electronics Engineer.

Korean Tourism Organization. (2000). Retrieved from http://www.knto.or.kr/cai-bin/TSS/tss_main.cgi?_year2000&i_type=13

Kotlen, P., Bowen, J., & Makens, J. (1999). *Marketing for Hospitality and Tourism.* Upper Saddle River: Prentice-Hall.

Law, R. (1998). *Hospitality data mining myths.* FIU Hospitality Review.

Menon,S. and Sharda, R. (1999), "Data mining update: new models to pursue old objectives", OR/MS Today, Vol.26 No. 3, pp.26-9

Mobasher, B. (2010). *Data mining with Weka*. Retrieved from http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/index.html

Organization, K. T. (2000). Retrieved from http://www.knto.or.kr/cai-bin/TSS/tss_main.cgi?_year2000&i_type=13

PETRE, R.-Ş. (2012). Data mining in Cloud computing. *Database Systems , III* .

R.Zaiane, O. (1999). *Introduction to data mining.* Retrieved from webdocs: https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/

Rouse, M. (2011, February). *Association Rules.* Retrieved from TechTarget: http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining

The data mining process. (n.d.). Retrieved from https://www.ibm.com/support/knowledgecenter/en/SSEPGG_9.5.0/com.ibm.im.easy.doc/c_dm_process.html?view=embed