T.C.

ALTINBAS UNIVERSITY

Electrical and Computer Engineering

EFFECTIVE INTEGRATION OF DATA MINING TECHNIQUES WITH BUSINESS INTELLIGENCE USING WEB MINING

OMER MUNEAM MUSHREF

Master Thesis

Supervisor

Prof. Dr. Osman Nuri UÇAN

Istanbul, 2019

EFFECTIVE INTEGRATION OF DATA MINING TECHNIQUES WITH BUSINESS INTELLIGENCE USING WEB MINING

By

Omer Muneam Mushref

Electrical and Computer Engineering

Submitted to the Graduate School of Science and Engineering

In partial fulfillment of the requirements for the degree of

Master of Science

ALTINBAS UNIVERSITY

2019

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

____

Prof. Dr. Osman Nuri UÇAN
Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

Prof. Dr. Osman Nuri UÇAN	School of Engineering and Natural Sciences, Altinbaş University	____________
Prof. Dr. Oğuz BAYAT	School of Engineering and Natural Sciences, Altinbaş University	____________
Asst. Prof. Dr. Adil Deniz DURU	Physical Education and Sport, Marmara University	____________

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

____

Asst. Prof. Dr. Çağatay AYDIN
Head of Department

____

Prof. Dr. Oğuz BAYAT
Director

Approval Date of Graduate School of
Science and Engineering: ____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Omer Muneam Mushref

DEDICATION

I dedicate this thesis to my family, classmates and all my friends for the support and encouragement throughout my education and life. Special dedication goes to my supervisor Prof. Dr. Dr. Osman Nuri UÇAN, my sisters and my mother for their support and prayers during my research work.

ACKNOWLEDGEMENTS

I might want to offer my thanks to every one of the individuals who have upheld me all through the regularly extended periods of this voyage. I might want to thank my advisor, Prof. Dr. Osman Nuri UÇAN for being my compass notwithstanding when I believed I was lost and being in extraordinary part in charge of the zenith of this work. I might likewise want to thank my supervisors for their supportive exhortation, which incredibly enhanced the nature of this work. Finally, I thank this institution for hosting me during these years, securely earning its place as my home. I am very thankful to my mom and dad, whose values and education motivate me to keep asking questions; to my siblings and family for their infinitely appreciated love and to my country which, although inanimate, keeps me anchored and offers me an example of resiliency.

ABSTRACT

EFFECTIVE INTEGRATION OF DATA MINING TECHNIQUES WITH BUSINESS INTELLIGENCE USING WEB MINING

Mushref, Omer Muneam
M.S., Electrical and Computer Engineering, Altınbaş University
Supervisor: Prof. Dr. Osman Nuri UÇAN

Date: June 2019
Pages: 50

Web mining is transforming the World Wide Web and changing it into a more useful environment through which users can obtain information efficiently. The world web contains massive quantities of multimedia files, text documents, educational content, and images which is continuously increasing with every new user. Web mining is a type of data mining that entails the extraction of data available on the internet, especially discovering patterns from Web-related data sources such as hyperlinks, server logs, web content, and web documents. Web mining can be accomplished in three different ways, i.e., web content mining, web usage mining, and web structure mining, etc. as they contribute to analyze different elements of a webpage. The effective implementation of these three elements of web mining has made it quite easy to retrieve information. The web mining concept has grown over the years due to the fact that the current web users prefer to upload data compared to download. This alteration in trend has led to a surge in the internet data making it difficult and time-consuming to retrieve informative patterns and knowledge. In this research, we will describe the fundamentals of web mining, well applying three essential data mining techniques to illustrate that web mining is the next driver of business intelligence. The data mining technologies will be applied on a dataset of 1098 phrases extracted from two news agencies, i.e., Reuters and Newswire. The result illustrated that the application of the Naïve Bayes and KNN to extract patterns from the data and to use PageRank to give the phrase weights makes it easy to determine a user's requirement in terms of the content of a website. The comparative analysis has illustrated that the KNN technique performed better that the Naïve Bayes and PageRank with an accuracy of 96.7%, 89.3% and 91.1% respectively. The field of Business Intelligence was picked due to the field's long-term development of data analysis tools and their diversity. The top three tools of the various tools covered in our research were compared to an average one in order to enlighten the reader in regards to each tool's strengths, weaknesses, and general functionality and electivity. Following the research is an in-depth review and examination of the customized analytical tool being developed, as well as various tests, conducted on actual company data, in order to discover flaws, faults, and deficiencies early, so as to have sufficient time to fix each of them and assure the software remains effective in the future. The work will then be summarized along with evaluated results from field tests of the developed custom application.

Keywords: Web mining, KNN, Web structure mining, Naïve Bayes, PageRank, Web usage mining, web content mining.

TABLE OF CONTENT

Pages

LIST OF TABLE

Pages

LIST OF FIGURES

Pages

LIST OF ABBREVIATIONS

ANN	:	Artificial Neural Networks
API	:	Application Programming Interface
BI	:	Business Intelligence
CSV	:	Comma-separated values, file that stores tabular data
DLR	:	Dynamic Line Rating
DMS	:	Distribution Management System
ERP	:	Enterprise Resource Planning
EU	:	European Union
GIS	:	Geographic information
GPRS	:	General Packet radio Service
GPS	:	Global Positioning System
HDI	:	Human Development Indicator
NB	:	Naïve bayes
IP	:	Internet protocol
k-NN	:	K nearest neighbor
LAN	:	Local Area Network
MLT	:	Machine Learning Techniques
WM	:	Web Mining
SVM	:	Support Vector Machine

1. INTRODUCTION

Websites have become a crucial conveyance channel for most businesses and individuals trying to relay diverse information. Thus, it is to find new ways of making this web information more useful, two businesses and individuals [1]. The website is made of a collection of closely related web pages that may contain different types of information, images, digital assets, and videos [1]. This has made the World Wide Web an enormous resource of information stored or presented in different formats that may be useful to business progress [2]. The growth in the usage of technology in businesses has introduced a cut-throat competition that needs businesses to be able to extricate data from web content and applying it to companies' marketing strategy [3]. It is necessary for companies to manage this massive information and always provide the user with information that is related to their queries [3] [4]. The analysis of these vast amounts of data, especially fetching extracting and processing, cannot be effected manually and therefore, automated extraction tools are needed. The primary means by which users find data from the internet is via the use of different search engines like Google, Bing, Yahoo, MSN, Yandex, etc. [1]. Data mining is an emerging technique used in analyzing useful information and extracting data from colossal data warehouses using various patterns, algorithms, tools, and intelligent methods. This process can assist existing and emerging businesses in analyzing data with regards to user conduct and predicting future trends [4]. The data mining process has four fundamental schemes for effective data extrication. First, the data source is traditionally a huge dataset that is located in large databases known as data warehouses, which may have the problem of defining it [2] [4]. Secondly, data exploration which entails the investigation of relevant information contained in the bulk and familiar data. Third, various models and design will be implemented to evaluate this unfamiliar data, four trends, and patterns. The tested and successful models will be deployed as part of the last strategy of data mining changing raw data into convenient information. This data may assist companies in improving their marketing strategies to increase profits and gain a competitive edge.

Web mining becoming a popular technology with many businesses as it's providing a tool to learn two vital and active areas in operation in the World Wide Web. Initially, companies have been utilizing powerful analytical devices to sift through volumes of supermarket scanner data and analyzing them to obtain vital market research information [5]. The recent increase in computing power, storage capacity, and availability of statistical software have led to a drastic increase in the accuracy of the analysis while reducing the cost [6] [2]. Although the study of web data is becoming vital for most companies, it has paused a lot of challenges to researchers; thus, the reason development entails application and adoption of data mining approaches for web mining. Therefore, web mining can be described as the application of data mining technologies and tools automate the process of discovering and extracting significant information from websites [2] [7]. The internet is a resource full of massive quantities of information obtained from the linked web pages. The repository contains data inform of multimedia objects and text. Web text mining approach is the first technique that can be used for web mining, which utilizes a key-based method [3]. Alternatively, other technologies apply a phrase for a text representation to search through a set of documents. The phrase based approach has been deemed better than the keyword-based technique as a phrase provides more information compared to a single term. Studies have shown that the phrasal indexing language is not powerful compared to the word based, although the phrase carries less equivocal and more succinct meaning [1] [3]. The phrase based approach reduces performance due to the following three facts; a phrase has an inferior statistical property two words, phrases have increased the redundancy and noise, and Afraid has a low frequency of occurrence [4]. Researches have been focusing on creating a solution to this problem, such as the use of sequential pattern mining leading to the emergence of structure mining.

Our Contribution:

This study will evaluate the fundamental concepts of web mining discussing the available knowledge on three techniques used for data analysis and extraction. Further, three data mining techniques, i.e., Naïve Bayes, KNN and PageRank, will be applied to a set of 1098 news-phrases to demonstrate the robustness of web mining and how it can be incorporated into businesses. The final step will be to utilize the results from the MATLAB simulation of the scenario to recommend the vital areas of application of web mining in Business Intelligence (BI).

Web content mining describes a procedure in which resourceful functional data is extricated from websites [1] [3] [8]. These content may include text documents, videos, audios, images, structured record, and hyperlinks. Over the last few decades, there has been a considerable increase in the number of web pages to billions, and the growth is continuing. Searching these billions of web pages is a daunting task and time consuming [1] [3]. The content mining technique retrieve queried data by carrying out various mining techniques that narrow the search to ease the process of discovering relevant data.

The content mining technique can be accomplished procedurally as illustrated in figure 2 below.

1.1 WEB CONTENT MINING

Web content mining describes a procedure in which resourceful functional data is extricated from websites [1] [3] [8]. These content may include text documents, videos, audios, images, structured record, and hyperlinks. Over the last few decades, there has been a considerable increase in the number of web pages to billions, and the growth is continuing. Searching these billions of web pages is a daunting task and time consuming [1] [3]. The content mining technique retrieve queried data by carrying out various mining techniques that narrow the search to ease the process of discovering relevant data. The content mining technique can be accomplished procedurally as illustrated in figure 1.1 below.

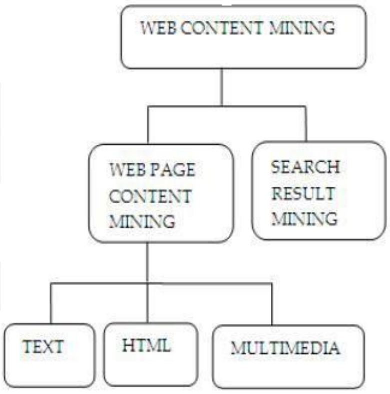

Figure 1.1: The web content mining procedural technique

Web content can be defined as unstructured data; therefore, extraction there is the need to use text and data mining technologies [3]. Information in text documents can be extracted using text mining, natural language, or machine learning. Primarily text mining is applied for the retraction of information from the content source; hence, different techniques are applied to provide these unknown data [8] [7]. Some of these text mining technologies include:

- Summarization
- Information visualization
- Categorization
- Clustering
- Information extraction
- Topic tracking

Furthermore, structured techniques can be used to mine structured data. In most cases, the structured data is a representative of the host page on the web [1][7]. the meaning of structured information is conventionally easy compared to the unstructured cases in content mining; therefore, the following techniques are applicable for structured content mining.

- Page contact mining
- Web crawler
- Wrapper generation

The semi-structured data describes a type of partially structured data in the text that is grammatical [9]. The content is structured in a hierarchy but is not predefined; thus, this data is in the form of tags, e.g., XML, HTML [9]. Data mining techniques used for extraction of semi-structured data include Top-Down extraction, OEM-Object Exchange model, and Web Data Extraction Language. To perform any data analysis the data quality is essential also the preparation of this data into the right format is the key to start the analysis sought. Hence, it is important to know beforehand the purpose of the analysis by having one or other purpose the procedure to go into the data processing may change completely. A good data preparation will also report better and more accurate results. Conventionally, apart from text other computational data in web mining maybe inform of multimedia data such as images, audios, and videos. Techniques used for multimedia mining include Color Histogram Matching, SKITCAT, Multimedia Miner, and Shot Boundary Detection [8].

1.1.1 Content Mining Techniques And Algorithms

As earlier noted the multiple methods that can be utilized for content mining to extricate data from enormous sets of data similarly various types of algorithms can be used to retrieve information from webpage [1] [3]. The decision tree is an example of an algorithm used for data extraction. It is classification and structure like a tree made up of root nodes, leaf nodes, and branches. This approach is hierarchical, where the root node split into branches that contain the lymph nodes which house the class labels [3]. The Naive Bayes algorithm (Native Bayes Classifier) is another easy and simple but Powerful classification that can be used for data extraction [9]. This algorithm utilizes Bayes' Theorem in calculating probabilities for each class in predefined data sets via counting the combination on values. The solution is the class with the highest probability as Illustrated by the equation below;

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

Where

P is the probability of a class

$A\ and\ B$ the classes in the data

Additionally, the Support Vendor Machine (SVM) and Neural Network (NN) are other algorithms that can be used for classification and extraction of data. The SVM can be utilized for both non-linear and linear data sets [2] [3]. The SVM employs an optimal separating hyperplane to segregate between classes using distinct classification characteristics [9]. The NN applies back-propagation with multiple layers that feed each subsequent layer to compute the solution. Once the data motivation is clear, so once having the specific idea to what we want to extract from the data, we can start looking at how the raw data is presented. It is not likely to receive the data with the precise format we would like to have in order to start the analysis, usually raw data is hard to visualize and therefore analyze.

Processing the raw data is needed in order to transform the data into the appropriate format, this data ready for the analysis can be called tidy data or processed data. To achieve it computational data processing is used, by writing and running a piece of code called processing script that will allow getting the data ready. When the data is processed, the next step before starting the data analysis is the cleaning the data. What cleaning data does is to prevent and correct the errors or the incompleteness of the dataset, for instance with duplications, missing data, NAs…

1.2 STRUCTURE MINING

The structure mining is a fundamental elements of web mining, which is concerned with the structure of hyperlinks [8]. The structure mining technique is a description of the structure of a website. Structure mining defines the relationship of the linked web pages in any website; therefore, the continued growth in the number of websites increases the challenge of extracting informative data [1] [3] [5]. The structure mining analyses hyperlinks on any website to collect data and summarize it based on similarity and relationships. The intra-page describes a type of mining that is executed at the document level where is the inter-page mining define data extraction at the hyperlink level. Figure 1.2 illustrates the steps followed in web structure mining and analysis.

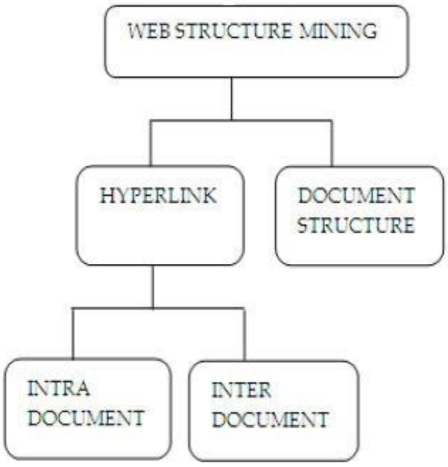

Figure 1.2: The web content mining procedural technique

1.2.1 Web Structure Mining Tools And Algorithms

The most common web structure mining algorithms are PageRank and HITS algorithms. These algorithms focus on the importance of web pages while evaluating their link structure. The PageRank algorithm was curated in 1998 by L. Page [1][3]. This algorithm forms the underlying core concept for Google search engine and has been used to determine the importance over page based on the number of pages linking to it. The importance of a page affects the backlink increasing its weight compared to the link from non-important pages [7][8]. The PageRank algorithm can be described using the relationship below.

$$\text{PR(A)} = \text{(1-d)} + \frac{d(PR(T1))}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)(1)} \qquad (1.2)$$

Where

PR(T1) is the rank of a page

d is the damping factor 0 to 1

Ti is the link created between pages

C(Ti) is the number of the outbound links.

The HITS algorithms are utilized for web structure analysis. Jon Kleinberg developed the technique for page ranking using hubs and authorities [3] [8]. The HITS algorithm is made of two technologies, i.e., sampling for the related pages and iterative step for the hubs and the authorities. The aim is to organize and visualize all the data using different representations according to the specific goal of the visualization. It is interesting to provide a tool that allows the visual representation for each individual customer and satisfies all the purposes sought.

1.3 USAGE MINING

This method is sometimes referred to as log mining 112 the process of recording user access data to a website which is collected in the form of logs. The data collected in logs is in the form of residual user-data that is left after visiting a website such as the IP address, pages visited, visiting time, etc. [8]. The information can assist businesses to comprehend user behavior so that they can improve the website structure and the information accessible to the users. Web usage mining is an automatic approach that archives the access patterns of the users, which is located on the web server [3] [2]. The information in the web service recorded in access logs where significant information such as URL, IP addresses, visiting time, etc. that may assist organizations to serve their customers based on their behavior. Web usage mining analyses data in log files to understand the behavior of the user at the time of interacting with the website [7] [6]. Web usage entails tracking of patterns and information which can either be customized or general trucking. In general, tracking the data is collected based on the webpage history whereas in customize ranking the information is collected for a specific user as illustrated in figure 1.3 below.

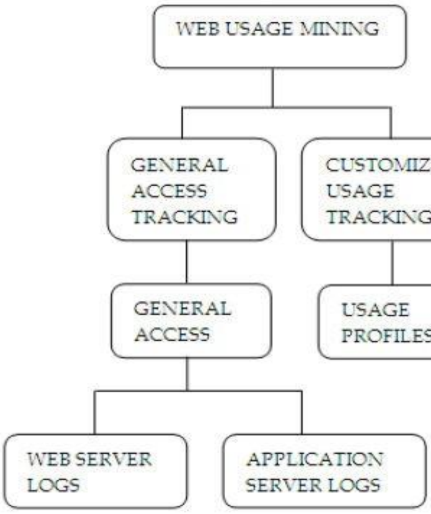

Figure 1.3: Different types of tracking technologies in Web Usage mining

1.3.1 Web Usage Tool And Algorithms

The several algorithm things used for web usage mining. The Apriori algorithm is a significant supervised technique that utilizes association rule to determine the frequency of a set of data, for instance, items in a transaction [5] [6]. The algorithm observed a database identifying the largest dataset which forms the base model for the identification of other datasets. The algorithm contains a predefined support level which is used to compare with the data. In a comparative analysis, when the support level is greater than the minimum component of the data set then it will be considered as a frequent dataset in case the support level is below the item number then it will be considered small [8] [7]. FP growth is another algorithm that is efficient for web usage mining, and similar to the Apriori algorithm, it utilizes the association rule to discover the frequent sets of data. This technique uses a tree which is a data structure that is made up of a single root node “null” with the subtree nodes forming children [8] [2]. The fuzzy mean technique utilizes clustering for usage mining. This technique is an unsupervised algorithm that can be employed to a different degree of connected data. The FCM algorithm is used to group objects into several clusters.

2. BACKGROUND

For the last twenty years, companies in almost every industry invested time and money towards improving their ability to collect data about their customers and to exploit the collected data to gain a competitive advantage against each other. As a result of significant progress in the field of computer science and the growth in computational power, the volume of collected data surpassed the capability of teams of statisticians employed to extract information and knowledge from large datasets by manual analysis. As the problem grew beyond the reach of human manual labor, it inspired the development of algorithms to scan and interact with multiple databases, to enable deeper and more thorough analysis than previously possible [31]. These algorithms formed the basis of several computer programs which we now call Business Intelligence tools (or "BI tools" for short).

The concept of extracting useful information from massive datasets, often called Data Mining, attracted many companies such as Oracle or Microsoft to invest their time, money and human resources to develop their own BI tools, driven by the promise of high rewards from renting out these tools. In the first part of bachelor thesis, we will discuss four tools called Tableau, Power BI, Micro-Strategy. According to the Gartner diagram depicted in figure 2.1, Tableau, Power BI is currently the market leaders in BI tools in 2017 which provokes questions about what makes these tools unique. To assure a proper comparison a basis in the form of Micro-Strategy tool is introduced. Comparison of their performance is made in six different categories:

- Intuitiveness of control - are these programs user-friendly? Ability to work with large datasets - are these programs fast? Built-in functions - are these programs robust with many tools?
- Writing own functions - is it easy to write scripts in these programs?
- Availability of information materials and tutorials - is it easy answer questions which rise during the development process?
- Ability to achieve the set goal - does the result meets the predefined standards?

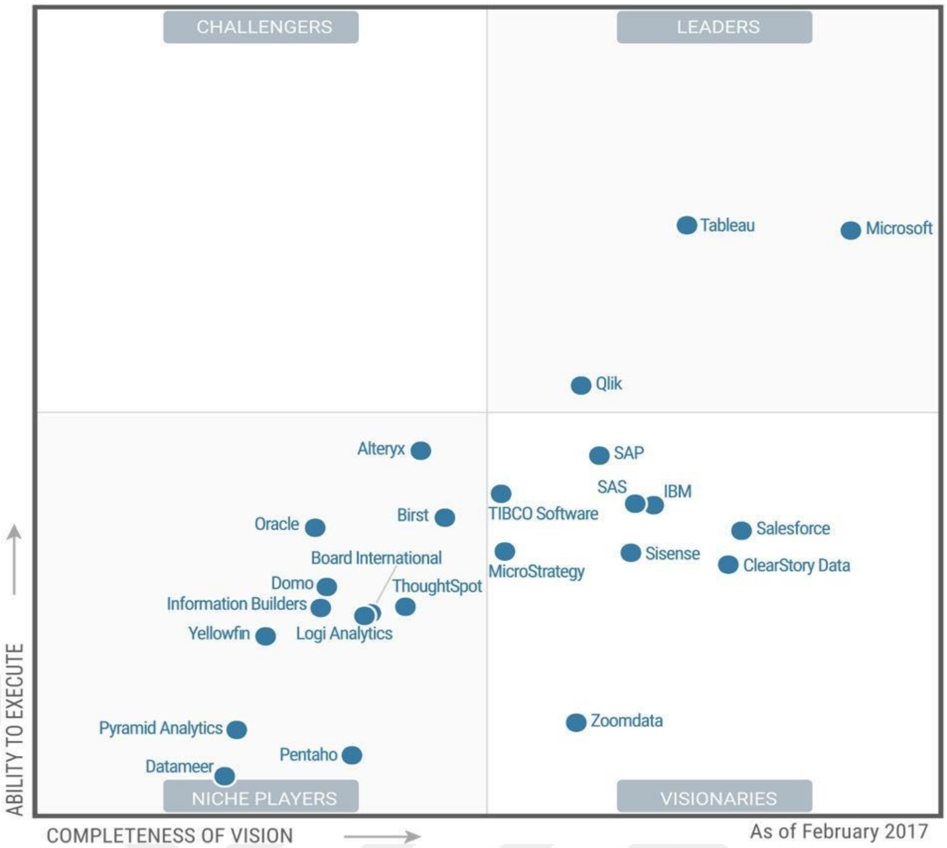

Figure 2.1: Gartner magic quadrant for Business intelligence 2017 [17]

The business intelligence mentioned above illustrates a real situation in which we average the power used by companies at CTU building and present the results for easier analysis. Thanks to the intuitiveness and simplicity of the aforementioned structure, it is easy to gain an extensive overview of the entirety of the examined period, and in most cases detect and prevent the various malfunctions and issues which may rise from an extended use of the heat pump or from wrong initial settings [28].

The business intelligence structure is also capable of simplifying massive amounts of data, for example summarizing years’ worth of data and categorizing it according to the days of the week or month in which it was measured. For time intervals which are longer than one month, the graph becomes unintelligible and devoid of useful information. It is therefore recommended, for maximal clarity, to use the structure to present data in one of the preset formats of hours, days, weeks, or month. As we can see the heating system is set well because the largest values of the power are found in days and hours during which people are present at school.

The second data structure is named "Scatter plot", also known as a scatter diagram. Similarly to the aforementioned Business intelligence, the Scatter plot uses two variables to separate data samples - in the case of this work variables such as average outdoor temperatures, Instantaneous power, etc. In case that the samples are meant to be colored differently, a third variable will be included. In contrast to business intelligence each axis can use a different aggregation function. An example of such a scatter plot is introduced below in figure 2.2.

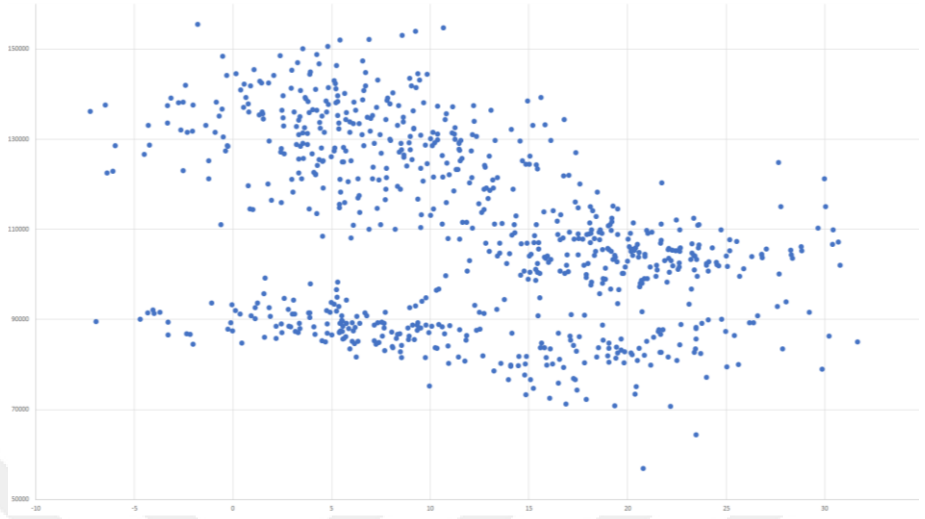

Figure 2.2: Example of scatter plot for intelligent business [21].

After introducing these two key components necessary for the analysis of the field of Energy intelligence, we will proceed to survey the Business intelligence tools mentioned at the beginning of this chapter and compare their performance. It is important to note that many of the currently available tools are programmed and optimized for different types of analysis, so there is no simple way to compare and contrast them. All tools together are examined in very specific field of analysis [24].

At the end of each review of the separate Business tools, a Business intelligence using web mining and a Scatter plot are shown to illustrate the capabilities of that particular tool and provide a screenshot of the whole program environment. As said before, depending on the analysis' goal one parameter or another would be used. So, although the possible approaches to the consumption data an endless, below different representation ways are shown that would be useful to know and better understand the consumer. This section uses various consumers for the different representations, in order to see variety of load profiles and also test the validity of the illustrations. Regarding the separation of load profiles between weekdays and weekends, it is seen that the businesses differ for weekdays and weekends. There are no significant differences among the businesses in weekday and in a weekend day. For the sake of the current study, this figure gives a hint for later discussions whether both weekdays and weekends' profiles should be considered or only the Weekdays profiles; when mining the consumers load patterns.

2.1 FRIENDLY BUSINESSES

The Power BI tool is pretty similar visually to other software developed by Microsoft such as Word, Excel or PowerPoint, allowing an effortless transition from other software and a comfortable learning experience. Also, tool placement in the program window is familiar and convenient, which results in an intuitive experience without any unnecessary searching. When it comes to visual adaptation to 4k screens, Power BI is the most compatible BI tool included in this research. There were no scaling issues, and everything was perfectly readable. Even letters in dialogue windows were complete without any trimming. However, the developers would be wise to consider improving the scaling of graphs or charts to fit into the program window in next version of the program [12-15]. For example, when analyzing large tables, there is the possibility of missing a part of the graph due to it being hidden behind the side menu. Such issues can be solved manually by changing letter sizes, but even after these changes, the graph may still be covered and not fully available. The aim of the data analysis is to group a set of observations based on one or various properties that make them similar to each other, hence the objects inside the same group will be more similar to any other object placed in other groups as shown in figure 2.3. The properties to group or data the observations are set by the analyst according to the aim of the grouping; this means that the same set of observations can be mined in many different ways; depending on what is considered similar, it is a task of the analyst to define the appropriate similarity/distance metric.

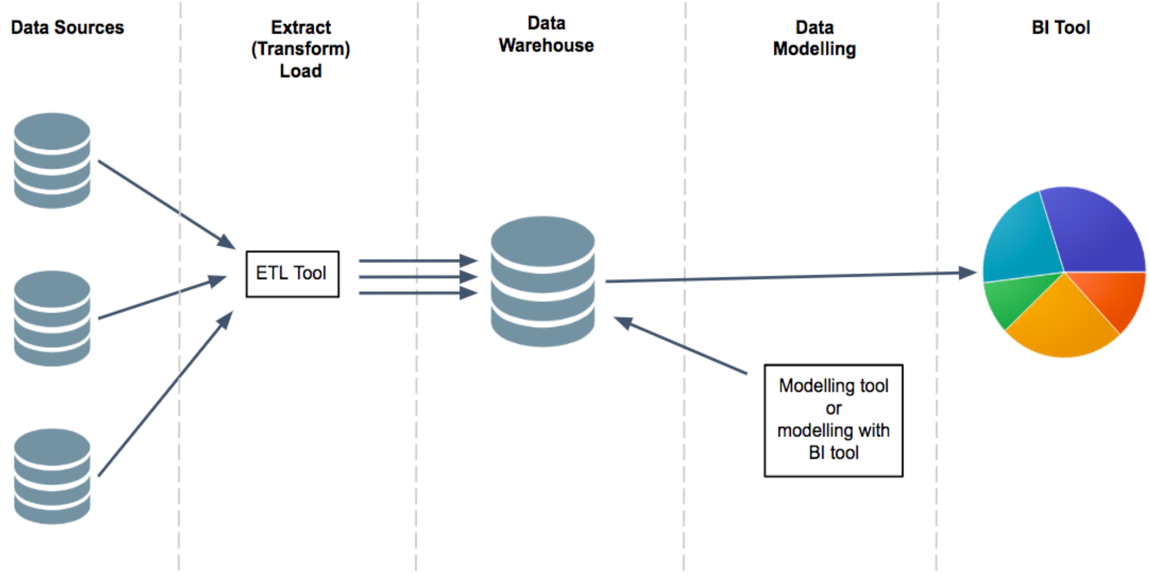

Figure 2.3: Web mining has introduced the use of numerous tools and algorithms to extract data for intelligent businesses [27].

In addition to that, there are so many algorithms and techniques to group observations, each one may also lead to a different grouping result. The selection of the appropriate technique is essential to perform the desired data analysis for businesses; also the typology of the input data to be mined is the main factor that will make choose one technique or another, taking into account that not every algorithm is valid for a specific problem.

2.2 BUSINESS DATASETS

During research, no special algorithms were used except three stated were found. Regardless, data was processed quickly even in the case of large datasets. When compared to other well established businesses, it achieved almost the same data processing time. On the other hand, when compared to Micro-Strategy, the difference in time required to process data was quite noticeable.

2.3 LITERATURE REVIEW

Web mining is an area in data mining that is concerned with data available on the internet true websites. The process entails the extraction of informative data that is located on web pages via the internet [7]. Currently uses apply various search engines to collect the needed information from the web [2][3][4]. Web mining has introduced the use of numerous tools and algorithms to extract data from web pages [3]. Web mining is increasing insignificance due to the quantity of text documents and the web that will require determination of relevant patterns, information, and knowledge. Web mining introduces techniques that automate the process making it rather easy and time-effective compared to when done manually [1] [5]. The collected information includes usage (data use, visited pages), structure (hyperlinks), and content (pages, text document). Web mining can be effected using the following three processes. First, information retrieval, which is a process of recovering pertinent and functional data from web pages [6] [4] [5]. The approach focuses on selected relevant data and the discovering of new information and knowledge from the massive quantity of data based on a user's query. Information retrieval is a procedural approach that includes searching, filtering, and matching data responses with queries [8]. Second, information extraction is a vital process that entails the extraction of the analyzed data. However, for the current study case it is not available such amount of detailed data as in [37]; the features’ data is much more limited in terms of samples and properties. The data referred to the household and householders is described in this work, this data is not fully complete and also needs to be further treated in order to eliminate bad data and duplicated features. Once refined, the data is able to be analyzed, in this case a graphical analysis using histograms is considered to represent the results and be the base to extract conclusions.

The ideal goal pursued is to discover if exists any pattern or relation when crossing the business profiles groups with the additional information related to data mining and web mining features. So, find the responsible features that could be the cause that define a specific load profile or another. In order to achieve this objective a significant number of samples is needed to consider the output results as a valid indicator and to be representative to extrapolate and use them for other purposes; for instance to determine the business profile of new users by knowing their characteristics.

Nonetheless, the actual features' dataset presents some limitations (small dataset, missing some relevant features, etc…) also the members in each mined data is so small, for instance the larger class has only few members using the machine learning algorithms. These facts, led to change the initial analysis' purpose to a one more adequate according to the dataset available, moving to a less ambitious analysis limited to find the most common household's and householder's features within each mined data. Hence, the output sought is to have a representative table stating the most common features for each business profile class.

The key focus of information extraction is obtaining a relevant fact from the analyzed data. Machine learning provides the third supportive element for web mining. Algorithms in machine learning and artificial intelligence improve the performance of web searches by learning a user's behavior and interest [2] [6]. Machine learning provides a more efficient way for information retrieval analysis and characterization of user behavior while enhancing the performance of a specific task. In this review will evaluate the core components of web mining and the various technologies, techniques and tools employed for these approaches as demonstrated in figure 2.4 below:

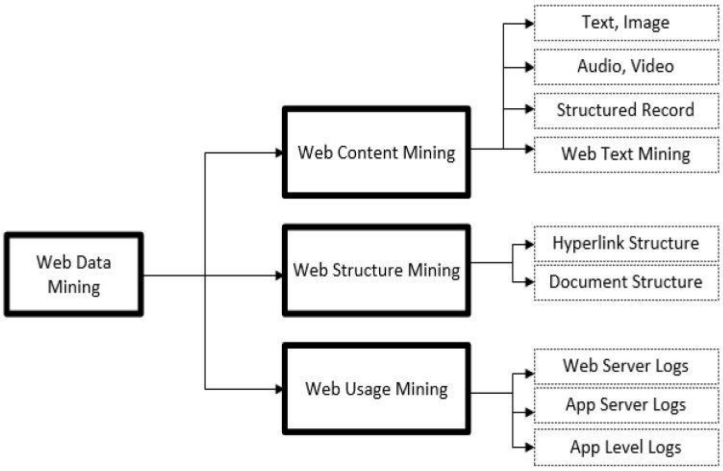

Figure 2.4: The taxonomy of the web mining process.

The web mining that allows reducing high dimensional data to a 2-dimensional space relying on group the vectors into regions of a map. The Hierarchical is an agglomerative bottom-up approach that allows to graphically represent its results; fact that permits to visualize the appropriate distance and linkage criteria when mining the data. Finally, the Naïve Bayes was chosen, it is the most common clustering method of vector quantization and it is needed to previously define the number of data instances. A business behavior change is needed in order to make people understand the importance of switching off; when is no needed. The base business refers to the minimum data used that the business constantly has, this is related to the night hours and the hours where there is no activity at home, for instance when the businesses are working. Minimizing this base business usage is key to reduce the data analysis when there is no activity and no need to use web mining. Therefore, this base businesses period should present a linear and regular profile, and as lower the better based on these techniques of web mining.

2.4 BUISNESS INTELLIGENCE IN AN ORGANIZATION

The term business intelligence arisen this last 2 decade, since renewable businesses were affordable for any business and the business price has continued growing. Depending on the legislation of each country net metering is able to sell this surplus or simple discount the businesses produced from the consumed from the intelligent system in an organization as shown in figure 2.5.

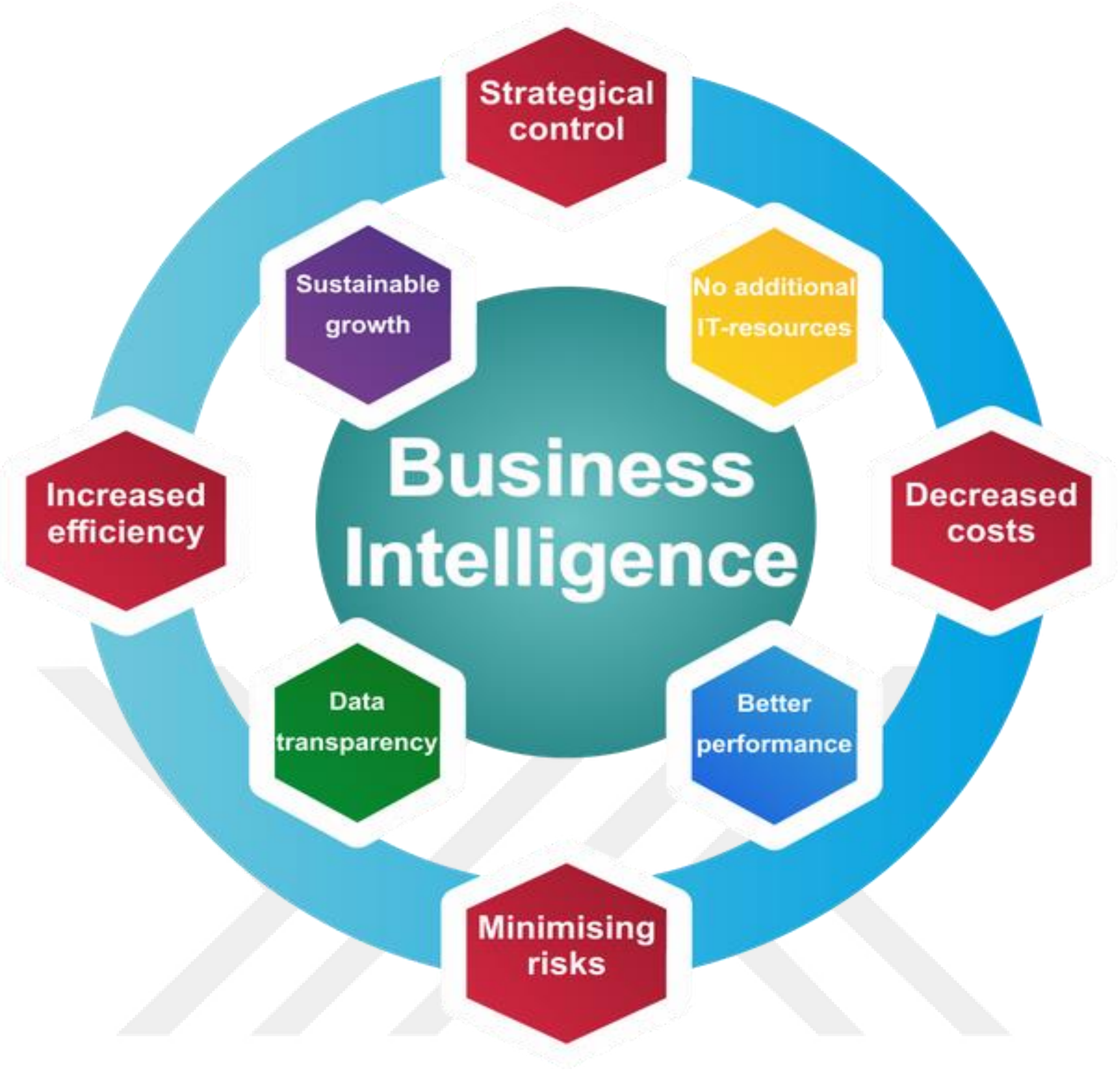

Figure 2.5: The business intelligence in an organization with different aspects [27].

This fact have a double benefit, gives to the customer the information to make smarter choices on when to consume the business, and for the utility allow to prevent blackouts and offer a better customer service. And communicate this information to the utility and also to the consumer, due to its connectivity as shown in figure 2.6. However, smart businesses are not exclusive of intelligence, also there are many factors available.

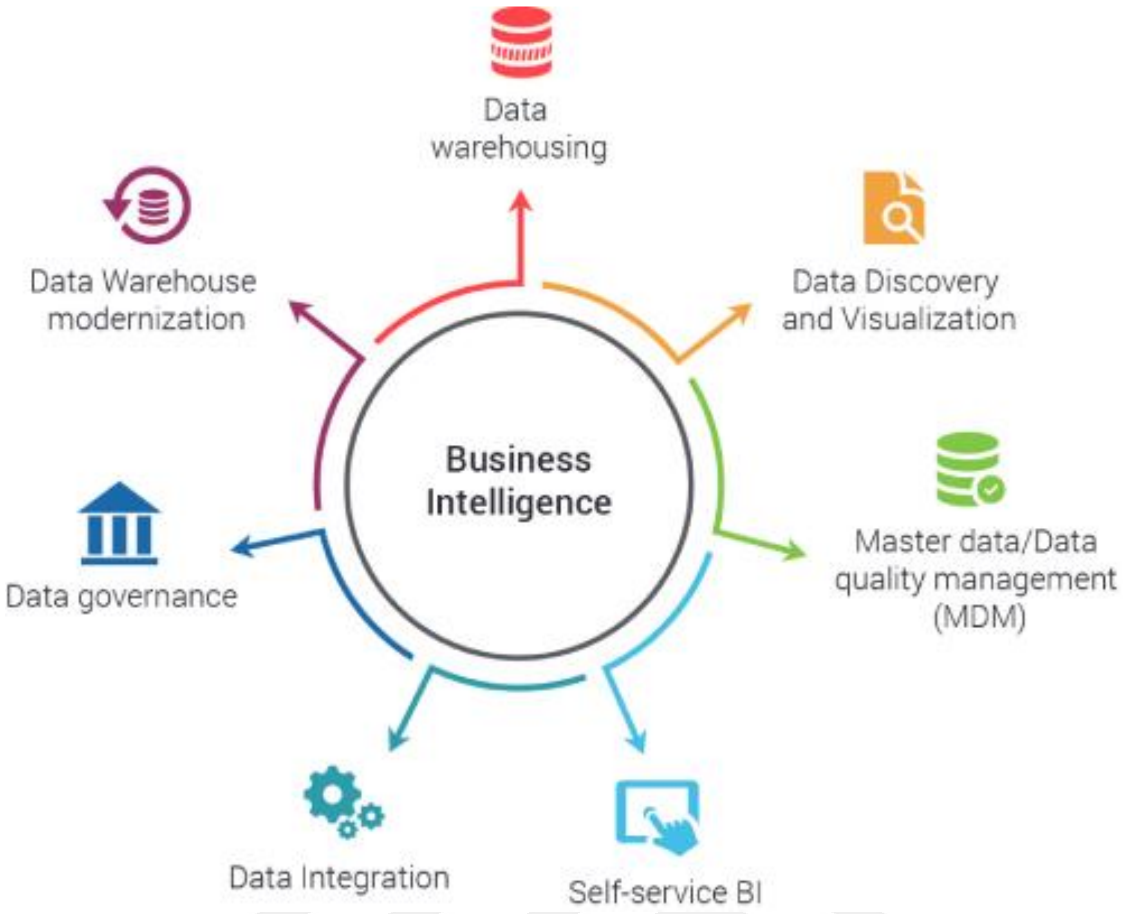
Data warehousing
Data Discovery and Visualization
Master data/Data quality management (MDM)
Self-service BI
Data Integration
Data governance
Data Warehouse modernization
Business Intelligence

Figure 2.6. The Hierarchical is an agglomerative bottom-up approach that allows to graphically represent business with intelligence [27].

3. METHODOLOGY

The steps followed to conduct this work and obtain the intelligent business segmentation are described below for the data mining techniques based on their usage for classification with three described algorithms:

Based on our review of the various BI tools currently available in the market, which we have covered in the previous chapter, we may now proceed and attempt to implement our own application. Such software will be later converted into an API in order for intelligent business, to make full use of it, based on the experience and knowledge gathered during our aforementioned research and analysis.

The main inspiration for building a new tool from scratch was to provide better- tailored service for intelligent business, and equip the company with the necessary tools to better implement their existing methodologies to the newly developing field of Energy intelligence. Since Energy intelligence is a relatively new and yet-to-be fully explored field, companies intending to make full use of its capabilities in this field would need to develop its own tools and software to provide satisfactory service based on a specific analysis. In order to fit said company needs and provide optimal, practical service in real-world situations, such an application would need to meet three main requirements:

- The capability of extracting information from a Sample Timestamp (Day of the week, Hour), where "Sample" may be defined as a pair consisting of a timestamp and its respective value.
- The possibility of obtaining data from a wide range of data sources -such as, for example, CSV, Excel, or database files. segmentation of variety of mathematical operations such as Average,
- The ability to process user-defined mathematical operations, such as thresholding or data filtering.

The software which was developed for this work was designed with these four main requirements in mind, and if at some point during the development process some of these requirements were violated or not optimally met, the development process was stopped and reformed in order to provide a better solution that would meet the aforementioned requirements and satisfy them. Another consideration made while writing the application, which is not mentioned in the four requirements and yet is common for all commercial-collaborative applications, was to maintain the code structure, transparency, and edit ability. Since the tool written for this article was supposed to be used by an actual company on a variety of real-world cases, it was of paramount concern, even if not officially stated.

In the following section, we will discuss how the tool was built and implemented. Furthermore, we will review the structure of the program itself, and compare its capabilities with those of businesses, which was chosen thanks to its various capabilities as an inspirational for the development of the tool.

3.1 CALCULATION PROCESS

This section will introduce the main principles and logic behind the operational process which manages the computational nodes, which are elaborated in the next section. As the basic process of all calculations is similar, it can be explained in general with no need for a specific example from each of the processes as shown in figure 3.1.

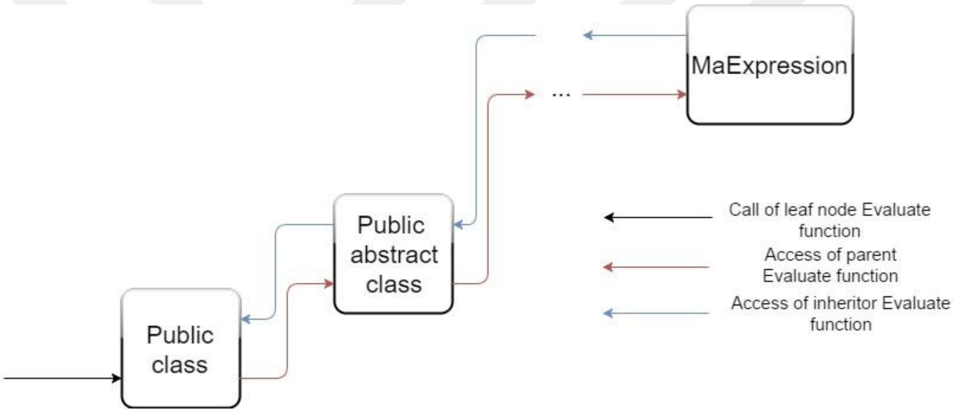
MaExpression
...
Public abstract class
Public class
Call of leaf node Evaluate function
Access of parent Evaluate function
Access of inheritor Evaluate function

Figure 3.1: Process calculation principles flowchart

Each and every calculation process starts by calling the function "Evaluate" of the leaf node in a tree. This function call, depicted in the figure above, is shown as a black arrow connecting the "Public class" block, which can express any leaf node in our structured tree. When we recall the structure of the whole node tree depicted in figure 3.1, we may see that each of the leaf nodes implements a specific function which is a combination of conditions and parameters imposed by the previously activated parent classes.

Once the "Evaluate" function call is made, the same function call is made in each of the parent classes as well, leading to a chain of "Evaluate" calls all the way up to the closest available node to the root of the tree called "MaExpression". In this section, this chain of events is expressed by a series of red arrows, each of which connects a pair of child-parent classes. The reason for this regressive chain of function calls is to ascertain the existence

and fulfilling of each of the initial conditions stated by each parent class.
Once all the function calls acting on the parent classes are done, the process reorients itself and proceeds down to the child classes, depicted in the above figure as a set of blue arrows. Once the lowest class in the structure is accessed, and the "Evaluate" procedure chain is finished, the entire calculation is finished, and the required data is prepared for another process, such as saving said data to a file.

The evaluation of the web mining technology and how it efficiently integrates data mining into Business Intelligence will be effectively accomplished by simulating three vital algorithms for each core element in MATLAB.

3.2 DATASET

The dataset in these simulations was collected from Kaggle.com (which is a database for educational datasets). The dataset contained 1098 phrases obtained from two news agencies, i.e., Reuters and Newswire [10]. The phrases were first weighted based on their importance. The weighting was from 0-6 for all the phrases which each phrase is placed in a relevant group [10]. The higher the value illustrated the importance of the phrase due to its links and backlinks. The table 1 below illustrates how each was treated

Table 3.1: The classification of the Datasets

Weight	Members	Representation in the algorithm
0	30	1
1	61	2
2	7	3
3	584	4
4	188	5
5	64	6
6	64	7

3.3 NAÏVE BAYES

The Naïve Bayes is a classification algorithm used for determining the density in data. This algorithm leverages Bayes Theorem based on the assumption that the predictor in the data is conditionally independent of the class elements [1]. The Naïve Bayes classifier yield posterior distribution that is robust, especially in biased classes, i.e., the decision boundary. This algorithm assigns observations to the most frequent class, and it explicitly estimates the densities of the prediction within the classes [2]. The design of the posterior boundary is based on the following probabilities in the Bayes rule for all classes k = 1,…, K, In this algorithm, the observation is classified through the estimation of the posterior probability for the classes [4]. The observation is assigned to the class yielding the optimal posterior probability.

$$P(\gamma = k|X_i, \dots, X_p) = \frac{\pi(\gamma = k)\prod_{j=1}^{p} P(X_j|Y = k)}{\sum_{k=1}^{k} \pi(\gamma = k)\prod_{j=1}^{p} P(X_j|Y = k)} \tag{3.1}$$

Where

Y is a random variable that coincides with the class index that is assigned the observation

(Y = k) is the prior probability

$X1, \dots, Xp$ are the random predictors

In this algorithm, the observation is classified through the estimation of the posterior probability for the classes [4]. The observation is assigned to the class yielding the optimal posterior probability. We pick the naïve bayes classification algorithm to make an inhabitance probabilistic classes, which is then tried later on set. The likelihood of a house being involved in a given weekday and timespan is registered by partitioning the quantity of involved periods by the absolute number of existent periods in the individual time and weekday of the arrangement set. For instance, if our characterization set has weeks of information, and if in this weeks the house was constantly delegated involved, at that point the likelihood of essence in this period during classification as shown in figure 3.2. The naïve bayes classification was then processed by applying the condition over every interim of the week from the classes.

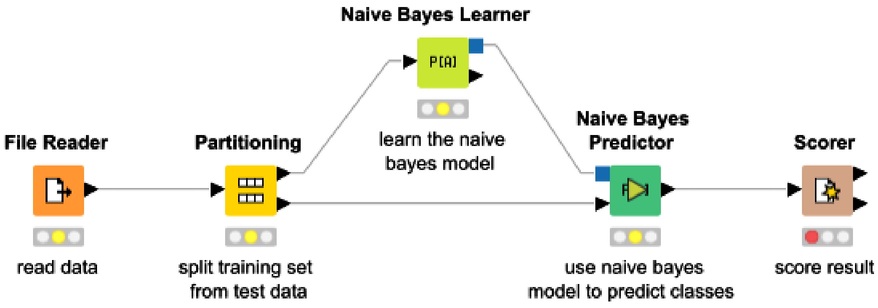
Naive Bayes Learner
P(A)
learn the naive bayes model
File Reader
read data
Partitioning
split training set from test data
Naive Bayes Predictor
use naive bayes model to predict classes
Scorer
score result

Figure 3.2: Naïve bayes algorithm for classification structure of intelligent business.

3.4 K-NEAREST NEIGHBORS

This algorithm categorizes the query points based on the distance to specific points in the training data sets [6]. The algorithm is vital in classifying new points in the dataset. In this research, we applied the Euclidean distance metric as illustrated below. The fuzzy K-NN methods which states that each instance has a certain degree of belonging to a group, even they can belong to more than one group; instead of assigning the instances to a specific group. The unsupervised learning- based K-NN; the supervised ones like the neural networks and some statistical based like the multivariate statistics. Other more recent techniques are the k-NN methods. Hence, it is observed that diverse techniques can be used to achieve the business pattern grouping as shown in figure 3.4, each technique has its own particular approach to reach the same final goal.

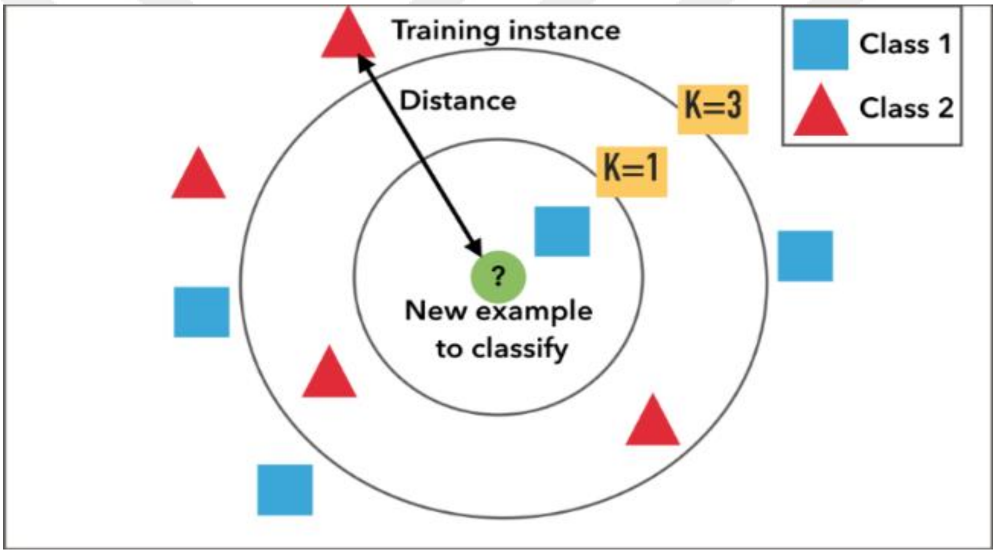

Figure 3.3: K-NN algorithm for classification structure of intelligent business.

3.5 PAGE RANK

This algorithm utilized the weights that were developed for the data set to describe the importance of each phrase. The algorithm used the equation below to determine the rank of the phrases. Different measurements were likewise utilized as a correlative measure, as depicted further in this in part as shown in figure 3.4. We assess the likelihood of foreseeing the inhabitance of a family unit dependent on its power utilization via preparing a model and testing it in a similar family. In any case, the fundamental business challenge is to utilize a conventional model that predicts with high exactness the inhabitance of any family unit [38]. To this end, we prepared a model in a solitary family unit and tried in numerous families in the given equation below.

$$\text{PR(A)} = \text{(1-d)} + \frac{d(PR(T1))}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)(1)} \tag{3.2}$$

$(T1)$ is the rank of a page

d is the damping factor 0 to 1

Ti is the link created between pages

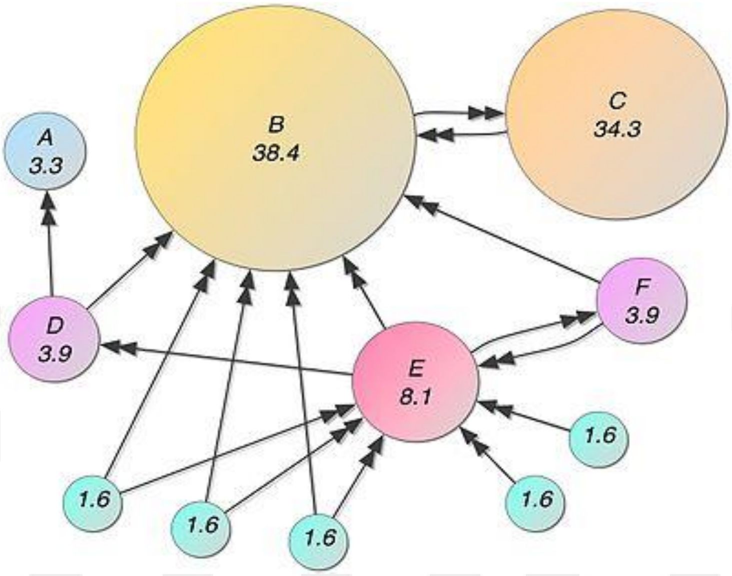

Figure 3.4: Page rank algorithm for classification structure of intelligent business [37].

1.6 FALSE POSITIVE AND FALSE NEGATIVE RATE

The false positive in the results represent the specificity and sensitivity of classification based on which classes have been made for business intelligence with data mining, represents a *false positive* and may reduce the energy efficiency of home heating systems due to unnecessary heating. The false positive rate (FPR) is a measure of these occurrences and is computed by dividing the number of false positives by the number of total slots.

$$\text{FPR} = \frac{fp}{fp+tn} \tag{3.3}$$

A low false positive rate indicates that the classification algorithm has a good performance on detecting absence. On the other side, in an interval, represents a *false negative* and may cause discomfort to the occupants due to the temperature lowering (in a thermostat application). The frequency of this type of errors can be measured with the false negative rate (FNR), which is calculated by dividing the false negatives by the total of occupied slots, as shown in equation.

$$\text{FPR} = \frac{fn}{fn+tn} \qquad (3.4)$$

A low false negative rate indicates that the classification algorithm has a good performance on predicting the consumption of usage.

1.7 TRUE POSITIVE AND TURE NEGATIVE RATE

The true positive rate (TPR) is defined as the percentage of positive instances that are correctly classified and is given by equation. It is also known as sensitivity or recall.

$$\text{TPR} = \frac{tp}{tp+fp} \qquad (3.5)$$

The true negative rate (TNR), or specificity, is the proportion of negatives that are correctly identified,   and is calculated by equation.

$$\text{FPR} = \frac{tn}{tn+fp} \qquad (3.6)$$

4. RESULTS

Measured performance of K-nearest neighbor, Page Rank method builds a hierarchy of classes by the help of a tree diagram. Two strategies can be followed when using KNN; the agglomerative (bottom-up) approach or the divisive (top-down) approach. In the current study, the agglomerative (bottom-up) approach is followed; in which each observation starts being a class, then the method aims to find the closest observations and put them together according to its similarity forming classes; the same procedure is followed successively until all the observations are part of the same class. At the end of the process the prediction illustrates how close the observations are to each other; however to reach this, a metric distance and a linkage approach from the ones need to be defined.

4.1 RESULT FOR THE NAÏVE BAYES

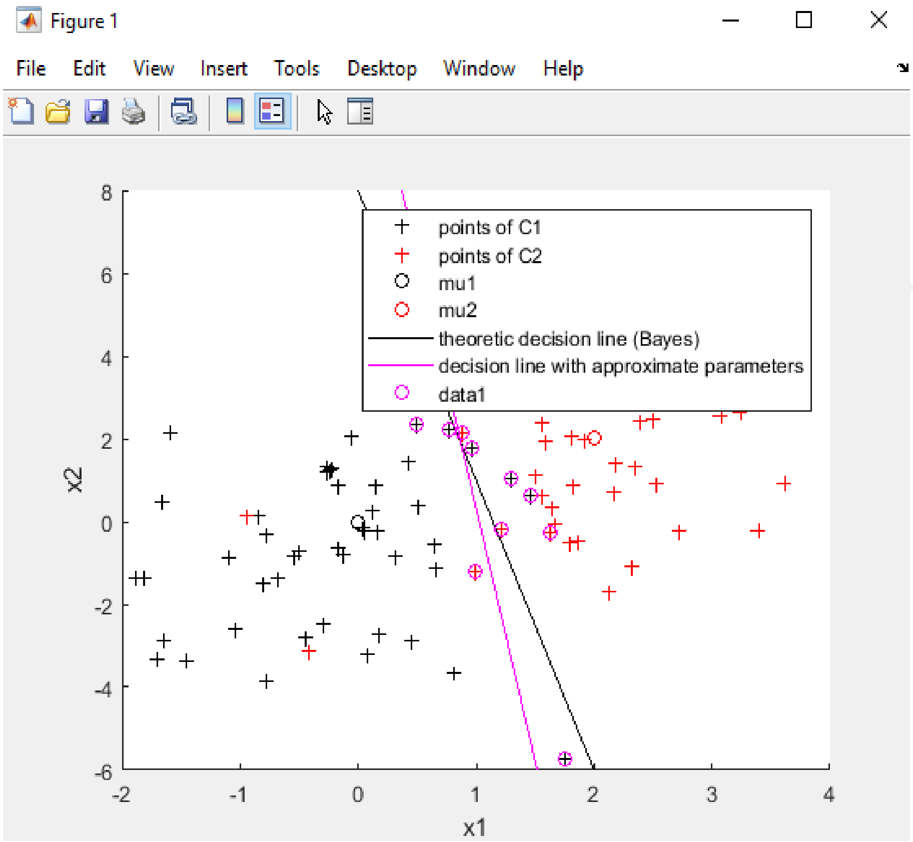

Figure 4.1: The differentiation of the phrases based on Bayesian decision and the real point of differentiation

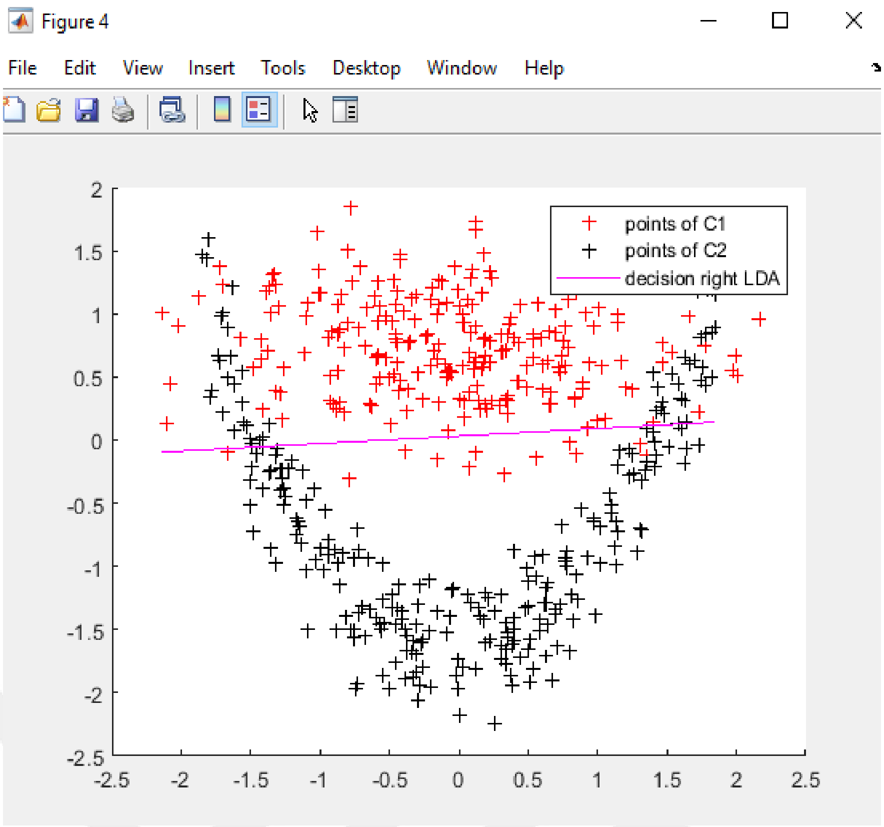

Figure 4.2: The division of the phrases based on the LDA conditions.

4.2 RESULT FOR THE KNN

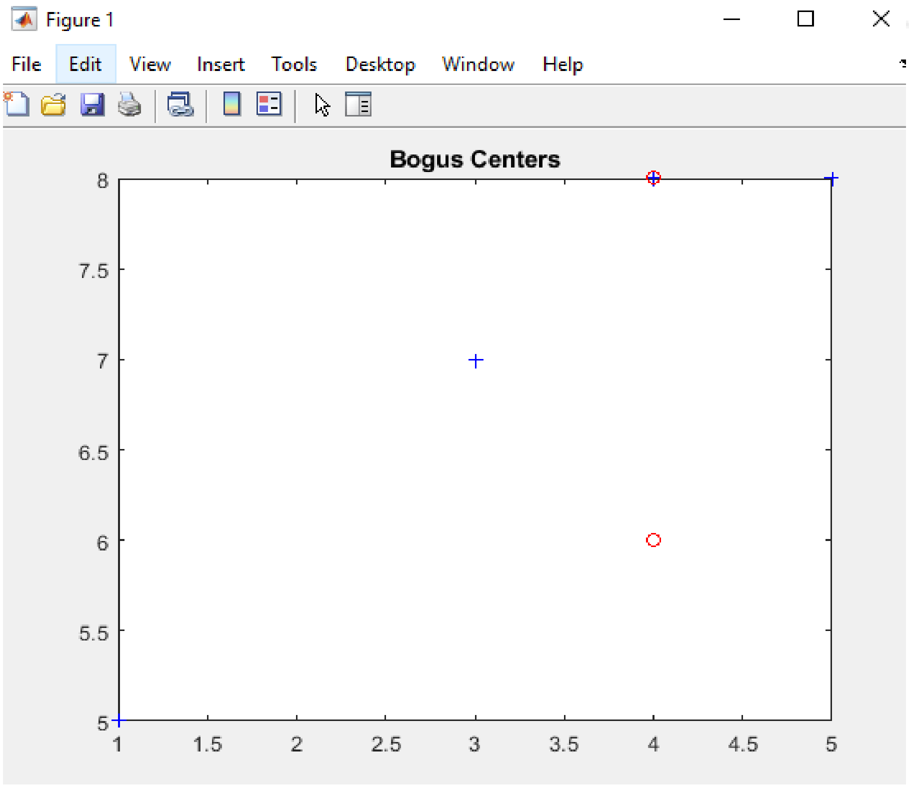

Figure 4.3: The unreal point in the training data for the KNN

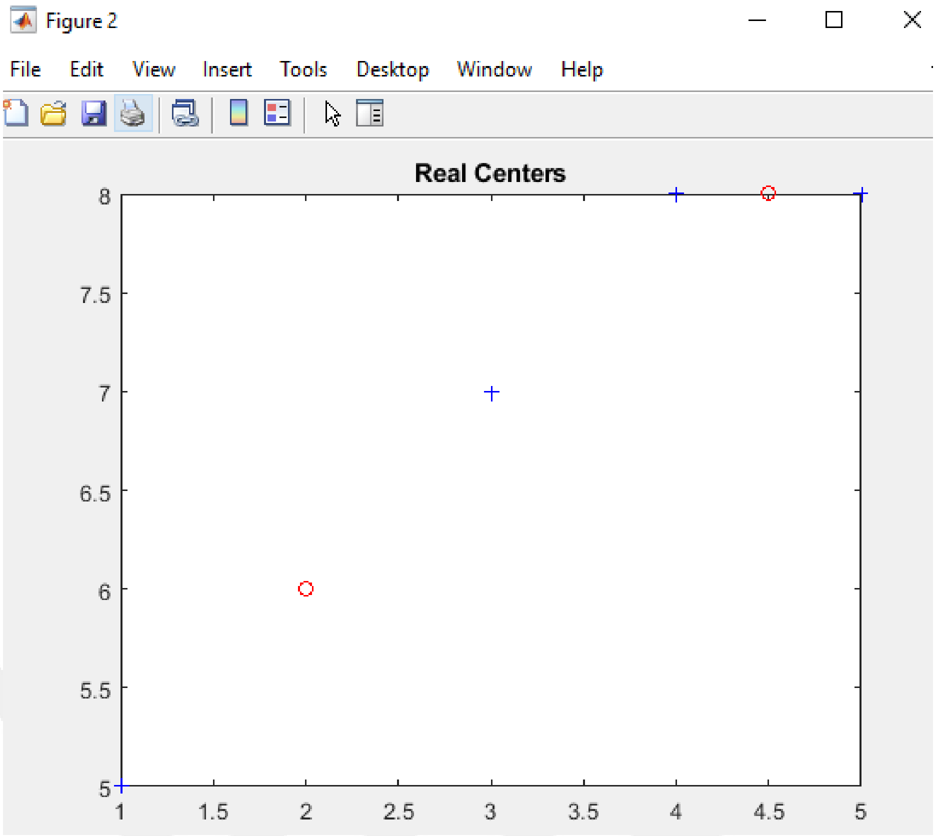

Figure 4.4: The Real centers for the nearest neighbors for the important phrase

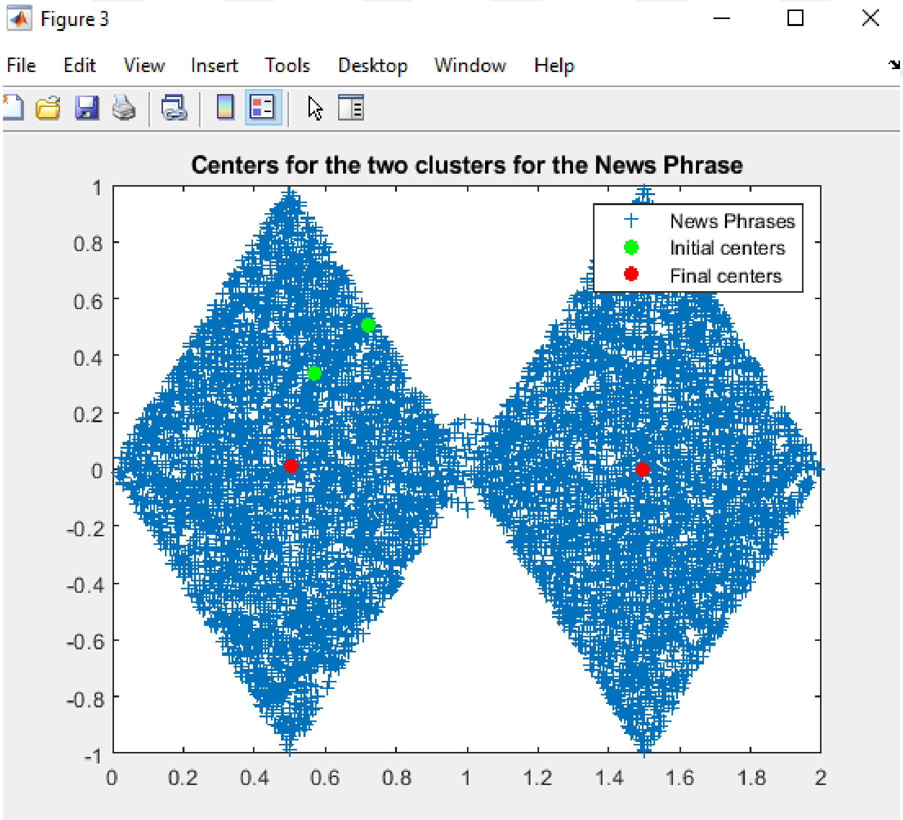

Figure 4.5: The cluster for the probability of the new phrases.

4.3 RESULTS FOR THE PAGERANK

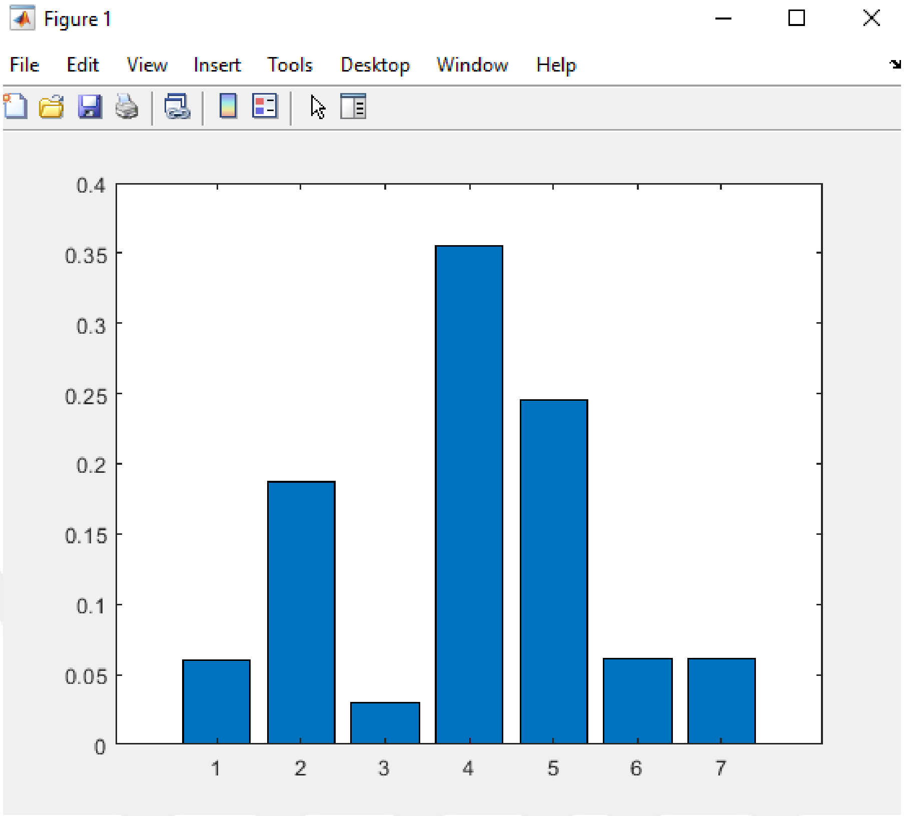

Figure 4.6: The Importance of the phrase based on the weights.

4.4 COMPARATIVE ANALYSIS

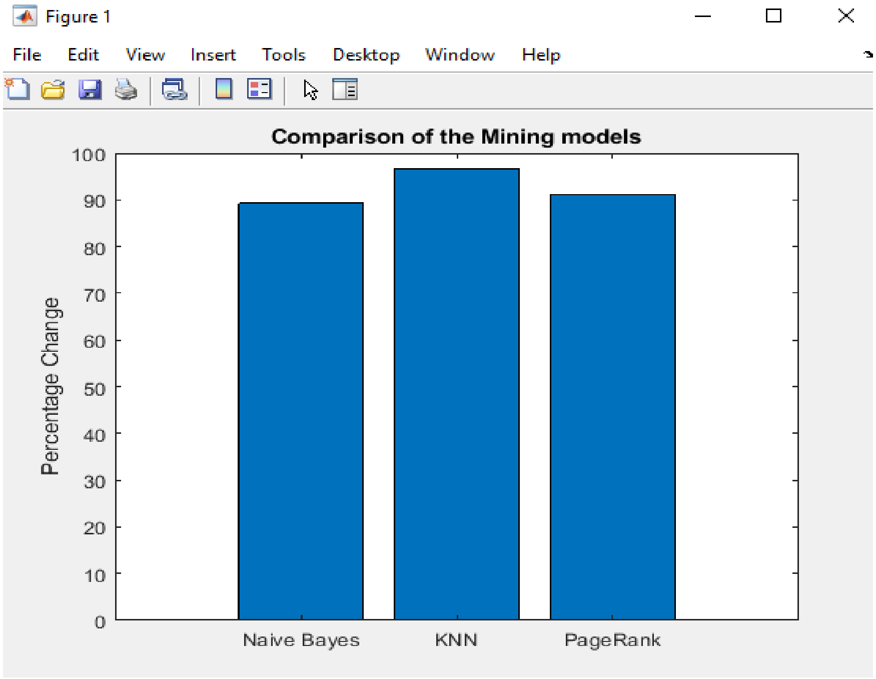

Figure 4.7: The comparative analysis of the mining models.

The results illustrate that the algorithms selected for the three different core elements of web mining have performed exceptionally well [1]. The Naïve Bayes algorithm was utilized for content mining, which entailed relating the phrases based on relevance to each other. The data was divided based on the theoretical posterior boundary of the Bayes Theorem and real points defined by the Latent Dirichlet Allocation (LDA) [3]. The Naïve Bayes produce a boundary that favors the imaginary values compared to the actual values, whereas the decision line based on the approximate parameters favor the real values as illustrated in figure 4.1. Further, the results in figure 4.2 show that the real values are differentiated from the imaginary data which tend to simulate greater importance than the real value of the phrase [3]. Data quality for the business intelligence usage is the key to perform a proper analysis and extract reliable conclusions from it. In this study, the statistical dataset was the optimal quantitatively and qualitatively. A much larger dataset, for instance counting for thousands of users, would have been more adequate to the purpose of the study than the small sample studied of only users.

In addition to that, the business intelligence data from the different web mining presented some issues that difficult the analysis at some point and demanded a pre-mining stage for treating and cleaning the data. As the communication between the sub-techniques device and the server is done through the machine learning algorithms of the our study, when it is switched-off the business analysis data is lost and counted like if it had been no techniques used. The observation was then assigned to the real values in the Naïve Bayes because it is made up of the most frequent values. The KNN algorithm was used to determine the structure of the hyperlinks in the dataset, especially based on the weights assigned to each class. So, when the possibility to access to the hourly business intelligence consumption data becomes a reality, the current study purposes will find a suitable framework since the numbers of users could be much higher and the data quality should be almost perfect as are the measures from the business intelligence to bill their users using machine learning known algorithms. The training data led to the generation of the bogus centers in figure 4.3 before the data became probabilistic and illustrated that the real center was for the most common weights such as 3 and 4 in this case. Considering that the weights 3 and 4 have more individuals 584 and 188 respectively, making them more significant. Finally, the PageRank algorithm was used to characterize the web usage based on the frequency of the phrase weight. The most frequent class make the class the most important. Similar to the KNN technique, class 3 and 4 were the most important class. The comparative analysis in this study has illustrated that the KNN technique was better compared to the PageRank and the Naïve Bayes techniques due to pruning and learning of data. The KNN has an accuracy of 96.7%, the Naïve Bayes has 89.3% and the PageRank has 91.1% respectively as shown in figure 4.7. More advanced web mining techniques could be used in order to improve some stages of the procedure, together with a deeper literature review regarding smart meter data treatment and segmentation for business intelligence.

5. DISCUSSION

After having data mining as a major part of this work the business intelligence and predicting the web mining effectiveness for mining by their load profile, the next step is to find the more likely characteristics and features of the business and stakes inside each class. In order to be able to associate some specific features to a specific classes load profile. We choose the publically available dataset from website, it is used a labeled dataset which is a dataset containing both input and output data, used to train a model. The labeled data allows the model to compare and adjust its parameters so that the performance is maximum. and in this field extract few features and uses them all in the classification process or there is an another way to select the optimal feature set through a brute-force evolution of all possible combinations, however, this process complexity grows exponentially with the number of features that's why we stick to the classification process of selecting the features [19-24].

After selecting the dataset we start by cleaning erroneous data, including missing values and we also delimited the number of records used in our experiments based on the minimum amount of samples verified in the five supervised learning algorithms. The final step consists on scaling the features using the standardization process, which a common method used in these type of problems, the choice of the right features plays and important role for the performance of any classifier [26]. For this reason, it is important to understand which features are relevant and may have a high correlation in the dataset.

In absolute value for power or energy consumption *min*, *max* and *mean* features were extracted to measure the absolute value of temperature and represent, respectively, the minimum, maximum and average temperature. The inference of occupancy through the electricity consumption data represents a supervised classification problem which include three algorithms for mining naïve bayes, k-NN and page rank model. Supervised classification is a machine learning technique typically used for pattern recognition. A supervised machine learning algorithm is an algorithm that required label data to learn the patterns and recognize. In this work, the label data represents the ground truth temperature based on which classification has been done. Splitting the dataset for each classifier, we repeat a 5-fold cross- validation ten times over our training data in order to obtain the best feature combination and model parameters. Training set 60% of the data is used to train the classification models through cross-validations. Testing set 40% of the data is used to test the classification models and to measure the accuracy of the forecasting the usage of electricity usage. Despite the simplicity of this study, we concluded that the algorithm performed well specially the hybrid model combining the random forest and decision tree both for the classification process for evaluation accuracy. We selected this five models since they are commonly used in similar problems and have proven to provide good results. They are very different from each other. For example, naïve bayes may require more data to provide good results and takes more time to run while support vector machines runs faster and may provide better results with a smaller amount of data (comparing to naïve bayes). K-NN is an ensemble algorithm that is based on decision trees to perform the classification with hybrid model combining both for forecasting the usage as low, medium and high, the mean square error also comes as mining for all models. Each iteration generates a mining error, which, in our case, represents the accuracy, and the performance of the classifier is the average. This part is very important since the wrong choice of features or model parameters can severely reduce the performance of a classifier.

6. CONCLUSION

The research has illustrated that data mining is a concept that is vital in the collection of information, knowledge, and patterns form large sets of data. Web mining is a subset of data mining that focusses on collecting information from the web. In this study, we evaluated web mining techniques and how they are vital in powering business intelligence. The research has shown that the frequency of a phrase is important in extraction, identification, and characterization of trends. Further, web mining shows that these techniques can be used to identify trends that will be vital for business development.

In addition to that, the business intelligence data from the data mining presented some issues that difficult the analysis at some point and demanded a pre-mining stage for treating and cleaning the data. As the communication between the different algorithms and the server is done through the machine learning algorithms of the our study, when it is switched-off the data is lost and counted like if it had been no mining for data.

So, when the possibility to access to the hourly business intelligence data becomes a reality, the current study purposes will find a suitable framework since the numbers of users could be much higher and the data quality should be almost perfect as are the measures from the business intelligence utility to bill their customers using machine learning known algorithms [33-35]. The main objective of obtaining data mining segmentation for business intelligence by the similarity of their load profiles was satisfactorily achieved; however complementary objective regarding the features analysis of each class couldn’t add value to the project. To obtain the final mining it was necessary an iterative process. The mining techniques used (Naïve Bayes, K-NN and Page Rank) have given similar outputs which facilitates the mining visualization and the partitioning possibilities.

6.1 FUTURE WORK

The comparative analysis in this study has illustrated that the KNN technique was better compared to the PageRank and the Naïve Bayes techniques due to pruning and learning of data. Research is recommended for further evaluation of web mining as part of data mining. Little has been done in terms of experimental analysis therefore we would like to recommend research in all the three aspects of web mining. The above research has formed a roadmap for future experimental analysis.

However, it was seen that data was enough for the purpose of mining profiles, but if what is sought is to determine the cause of the curve shape and the business intelligence data won’t be enough. So that, it will make necessary to use business intelligence data and features data. By providing that, this will improve the effectiveness of business intelligence efficiency programs and also it will have a positive effect on the resource reduction recommendations. As well as, improving the effectiveness of the demand response programs aiming to shift the time of business intelligence usage to avoid peaks or reduce the peaks duration using web mining.

REFERENCES

[1] A. Kumar and R. K. Singh, "Web Mining Overview, Techniques, Tools and Applications: A Survey," International Research Journal of Engineering and Technology, vol. 3, no. 12, pp. 1543-1547, 2016.

[2] A. K. Sharma and P. C. Gupta, "Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining," International Journal of Advanced Research in Computer Engineering & Technology, vol. 1, no. 8, pp. 287-293, 2012.

[3] M. J. H. Mughal, "Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview," International Journal of Advanced Computer Science and Applications, vol. 9, no. 6, pp. 208-215, 2018.

[4] S. Vidya and K. Banumathy, "Web Mining- Concepts and Application," International Journal of Computer Science and Information Technologies, vol. 6, no. 4, pp. 3266-3268, 2015.

[5] S. Vijiyarani and E. Suganya, "Research Issues in Web Mining," International Journal of Computer-Aided Technologies, vol. 2, no. 3, pp. 55-64, 2015.

[6] S.-T. Wu and Y. Li, "Pattern-Based Web Mining Using Data Mining Techniques," International Journal of e-Education, e-Business, e-Management and e-Learning, vol. 3, no. 2, pp. 163-167, 2013.

[7] T. A. Al-asadi, A. J. Obaid, R. Hidayat, and A. A. Ramli, "A Survey on Web Mining: Techniques and Applications," International Journal on Advanced Science Engineering Information Technology, vol. 7, no. 4, pp. 1178-1184, 2017.

[8] J. L. K. Grace, V. Maheswari and D. Nagamalai, "Effective Personalized Web Mining by Utilizing The Most Utilized Data," International Journal of Database Management Systems, vol. 3, no. 3, pp. 100-108, 2011.

[9] F. Johnson and K. G. Santosh, "Web content mining techniques: a survey," International Journal of Computer Applications, vol. 47, no. 11, 2012.

[10] Y. E. Gündoğmuş, "Reuters Categorized Wire Data," http://dataraccoons.com/, 2017. [Online]. Available: https://www.kaggle.com/yemregundogmus/reuters-categorized-wire-data. [Accessed 22 May, 2019].

[11] K. C. ARMEL, A. GUPTA, G. SHRIMALI and A. ALBERT, "Is disaggregation the holy grail of business intelligence? The case of business intelligence," Energy Policy, vol. 52, p. 213–234, 2013.

[12] M. C. Mozer, L. Vidmar and R. H. Dodier, "The Neurothermostat: Predictive Optimal

Control of Residential business intelligence Systems," Advances in neural information processing systems, 2017.

[13] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," in Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, 2007.

[14] J. Scott, A. J. Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges and N. Villar, "PreHeat: Controlling Home business intelligence Using Occupancy Prediction," in Proceedings of the 13th international conference on Ubiquitous computing, 2011.

[15] J. Krumm and A. J. Bernheim Brush, "Learning Time-Based Presence Probabilities," in Pervasive Computing, 2011.

[16] X. Liang, T. Hong e G. Q. Shen, “business intelligence data analytics and prediction: A case study,” em Building and Environment, Elsevier, 2016, pp. 179-192.

[17] kdnuggets, "CRISP-DM, still the top methodology for analytics, data mining, or data science projects," [Online]. Available: http://www.kdnuggets.com/2014/10/crisp-dm-top- methodology-analytics-data-mining-data-science-projects.html. [Accessed 25 03 2017].

[18] A. Fleury, M. Vacher and N. Noury, "SVM-Based Multimodal Classification of Activities of Daily Living in Health Smart Homes: Sensors, Algorithms, and First Experimental Results," in IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, IEEE, 2010, pp. 274 - 283.

[19]S.Raschka,"MachineLearningFAQ,"[Online].Available:https://sebastianraschka.com/aq/docs/evaluate-a-model.html. [Accessed 04 04 2017].

[20] M. Swain, S. K. Dash, S. Dash and A. Mohapatra, "An Approach for IRIS Plant Classification Using Neural Network," International Journal on Soft Computing, vol. 3, 2012.

[21] T. Durieux, "Exploring the use of artificial neural network based business intelligence scale models in a variational multiscale formulation," 2015.

[22] S. K e S. Sasithra, “REVIEW ON CLASSIFICATION BASED ON ARTIFICIAL NEURAL NETWORKS,” International Journal of Ambient Systems and Applications (IJASA), vol. 2, 2014.

[23] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network

classification," Journal of Biomedical Informatics, p. 352–359, 2003.

[24] Y. LeCun, L. Bottou, G. Orr and K.-R. Müller, "Effiicient Back-Propogation," in Neural Networks: Tricks of the Trade, Springer-Verlag, 1998, pp. 9-50.

[25] A. Ben-Hur and J. Weston, "A User’s Guide to Support Vector Machines," in Methods in molecular biology, 2010, pp. 223-239.

[26] stackexchange, "Use Gaussian RBF kernel for mapping of 2D data to 3D," [Online]. Available: http://stats.stackexchange.com/questions/63881/use-gaussian-rbf-kernel-for-mapping-of-2d- data-to-3d. [Accessed 06 04 2017].

[27] E. Goel and E. Abhilasha, "K-NN: A Review," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 1, pp. 251-257, 2007.

[28] Anuradha and G. Gupta, "A Self Explanatory Review of Naïve bayes," in IEEE International Conference on Recent Advances and Innovations in Engineering, 2014.

[29] S. B. Kotsiantis, "Page rank: a recent overview," Artificial Intelligence Review, vol. 39, no. 4, p. 261–283, 2011.

[30]"Diving into data," 19 10 2014. [Online]. Available: http://blog.datadive.net/interpreting- random-forests/. [Accessed 17 5 2017].

[31] G. Biau, "A K-NN Guided Tour," TEST, vol. 25, no. 2, p. 197–227, 2016.

[32]ArcToolbox, "Fit Random Forest Model," [Online]. Available: http://code.env.duke.edu/projects/mget/export/HEAD/MGET/Trunk/PythonPackage/dist/TracOnlineDocumentation/Documentation/ArcGISReference/RandomForestModel.FitToArcGIST able.html. [Accessed 27 03 2017].

[33] J. Weiss, "Lecture 22—Wednesday, November 10, 2010," [Online]. Available: https://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture22.htm. [Accessed 02 04 2017].

[34] J. Kelly and W. Knottenbelt, "Neural NILM: Deep Neural Networks," in Proceedings of the 2nd ACM International Conference on Embedded Systems for Efficient Business Environments, 2015.

[35] A. Jain, K. Nandakumar and A. Ross, "Score normalization in multimodal biometric systems," in Pattern Recognition of Businesses, Elsevier, 2015, p. 2270–2285.

[36] I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, 2015.

[37] Cost-benefit analyses & state of play of smart business deployment in the EU-27, 2014. *Benchmarking smart business deployment in the EU-27 with a focus on businesses.* [Online] Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014SC0189&from=EN

[38] Country fiches for electricity smart business, 2014. *Benchmarking smart business deployment in the EU-27 with a focus on businesses.* [Online] Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014SC0188&from=EN