REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE IBN KHALDOUN – TIARET

MEMOIRE

Présenté à :

FACULTÉ MATHEMATIQUES ET INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : Génie Logiciel

Par :

Mr. Ayed Toufik

Sur le thème

Big Data Mining:

An overview on Machines Learning Algorithms

Soutenu publiquement le **08 / 11 / 2020** à Tiaret devant le jury composé de:

Mr. BOUDAA	Boudjemaa	Grade	MCB	Université	UIK-Tiaret	Président
Mr. DJAFRI	Laouni	Grade	MCB	Université	UIK-Tiaret	Encadreur
Mr. BEKKI	Khadhir	Grade	MAA	Université	UIK-Tiaret	Examinateur

2019-2020

Acknowledgment

I would like to thank my supervisor **Mr. DJAFRI Laouni.** *for his consistent support and guidance during the running of this project. I also wish to thank all the people whose assistance was a milestone in the completion of this project.*

I wish to acknowledge the support and great love of my family; especially my mother. They kept me going on and this work would not have been possible without their input.

Dedication

I dedicate this work:

To my lovely mother and My beloved brothers and sisters for their sincere love of the land and their ever-present support.

To my friends who encourage me.

الملخص

يعد تعدين البيانات الضخمة مصدرًا رائعًا للمعلومات والمعرفة من الأنظمة إلى المستخدمين النهائيين الآخرين. ومع ذلك ، فإن إدارة مثل هذه الكمية الكبيرة من البيانات / المعرفة تتطلب الخدمة التلقائية ، مما يؤدي مما يؤدي بين التفكير الى استخدام تقنيات وأساليب التعلم الآلي في معالجة البيانات الضخمة. يساعدنا التعلم الآلي على اتخاذ القرارات إذا لم يكن هناك "طريقة صحيحة" للمشكلة المحددة في قواعد المعرفة السابقة ، كما أنه يعد من أكثر الأدوات المستخدمة على نطاق واسع للتحليل والنمذجة. في هذا العمل، سوف نقدم فهمًا جيدًا لأفضل خوارزمية (خوارزميات) التعلم الآلي المطبقة لمعالجة تعدين البيانات الضخمة.

الكلمات المفتاحية: استنباط البيانات الضخمة، التعلم الآلي ،تقنيات البيانات الضخمة ، الحوسبة المتوازية والموزعة.

Abstract

Big Data Mining is a great source of information and knowledge from systems to other end users. However, managing such a large amount of data / knowledge requires automation, which is leading to a trend in data processing and machine learning techniques. Machine learning helps us make decisions if there is not a "right way" for the specific problem in previous knowledge bases, and it offers some of the most widely used tools for analysis and modeling. In this work we provide a good understanding of the best machine learning algorithm (s) applied for the processing of Big Data Mining.

Keywords: Big Data Mining, Machine Learning, Big Data Technologies, Parallel and Distributed Computing.

Résumé

Big Data Mining constitue une formidable source d’informations et de connaissances des systèmes vers d’autres utilisateurs finaux. Cependant, la gestion d'une telle quantité de données/connaissances nécessite une automatisation, ce qui conduit à une tendance en matière de traitement de données et de techniques d'apprentissage automatique. L’apprentissage automatique nous aide à prendre des décisions s’il n’existe pas de « bonne voie » pour le problème spécifique dans les bases de connaissances précédentes, et il offre certains des outils les plus utilisés pour l’analyse et la modélisation. Dans ce travail nous faisons bien comprendre le (les) meilleur(s) algorithme(s) d’apprentissage automatique appliqué(s) pour le traitement du Big Data Mining.

Mots clés: Exploration des données massives, apprentissage automatique, Technologies du Big Data, Calcul parallèle et distribué.

List of Figures

List of tables

Table of Contents

Chapter I

Generalities & Concepts of Big Data

Chapter II

Construction and Processing of Big Data Mining

Chapter III

Big Data Mining: Analytics and Technologies

Chapter IV\

Realization and Implementation

General introduction

1 General Introduction

1.1 Introduction

Machine learning and data mining are not the same, but cousins. Machine learning is a branch of artificial intelligence that provides systems that can learn from data. Machine learning is often used to classify data or make predictions, based on known properties in the data learned from historical data that's used for training. Data mining is sorting through data to identify patterns and establish relationships. Generally, data mining (sometimes called knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is the analysis of data for relationships that have not previously been discovered. It is an interdisciplinary subfield of computer science, the computational process of discovering patterns in large data sets ("Big Data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. So, data mining works to provide insights and discovery of unknown properties in the data. Machine learning can be carried out through either supervised learning or unsupervised learning methods. The unsupervised learning uses algorithms that operate on unlabeled data, namely, the data input where the desired output is unknown. The goal is to discover structure in the data but not to generalize a mapping between inputs to outputs. The supervised learning (It is the subject of our concern) use labeled data for training. Labeled data are datasets where the input and outputs are known. The supervised learning method works to generalize a relationship or mapping between inputs to outputs. There is an overlap between the two. Often data mining uses machine learning methods and vice versa, where machine learning can use data mining techniques.

1.2 Positioning of the problem

The field of big data mining is a hot topic for researchers in the fields of computer science mathematics. Modeling and predictive analytics can be of critical importance to organizations if they are properly aligned with their processes and business needs. They can also significantly improve their performance and the quality of their decisions increasing their business value. Every organization can statistically analyze their data and better understand their environment, but the greatest profit potential lies with those who are able to perform modeling and predictive analysis based on machine learning algorithms. Accordingly, the problem posed that we want to solve in our master thesis is to find (the) machine learning algorithm(s) very suitable(s) for Big Data Mining?

1.3 Objective

The objective of our work is first, to find a classifier that gives a good prediction result (precision) in the context of Big Data Mining. Second, our goal is to speed up the execution time (speed).

1.4 Organization of the master thesis

We have structured our master thesis according to the plan described below:

We start with a general introduction that briefly describes our topic, the problems we face in Big Data Mining, and the solutions that will be discussed.

From an organizational point of view, the remainder of the master thesis is structured around four chapters.

Chapter I: Generalities & Concepts of Big Data

This chapter describes the open doors and challenges of big data, highlighting the recognized characteristics of big data.

Chapter II: Construction and Processing of Big Data Mining

In this chapter, we present how to manage big data mining, from the data collection phase, through the analysis of this data, and how to visualize it.

Chapter III: Big Data Mining: Analytics and Technologies

In this chapter, we will take an in-depth look at the field of data analysis, and we will introduce the different types of analysis, we present various systems and technologies that allow us and that help us to process big data more efficiently, as well as a comprehensive overview of machine learning algorithms.

Chapter IV: Realization and Implementation

In this chapter, we will present the work that we have done as well as the results obtained. We end this chapter with a synthesis of comparison of the classifiers implemented, limits and open perspectives.

Finally, we will end our master thesis with a general conclusion and perspectives opened by the work presented in this end of study project

Chapter I

Generalities & Concepts of Big Data

Chapter I Generalities & Concepts of Big Data

I.1 Introduction

Big Data mining is creating a new generation of decision support data management. Businesses recognize the potential value of this data; they put in place the technologies, people and processes to take advantage of the opportunities. Using analytical data is essential to leveraging Big Data. This chapter introduces several key concepts that introduce big data like definitions, characteristics, types of data, applications of big data and big data architecture.

I.2 definitions

There are several definitions of big data from different points of view. For example, according to [Mills et al. 2012] Big Data is a term used to describe large data at high speed and / or large variety, it requires new technologies and techniques to capture, store and analyze it; it is also used to improve decision-making, provide information and insights, and support and optimize processes.

According to [NIST 2015] Big Data is a term where the volume of data, the speed of processing or the representation of the data determines the capacity to perform effective analysis using traditional approaches, Big Data requires significant scaling (more nodes) for efficient processing. On the other hand [Barker & Ward 2013] defines big data as a term describing the storage and analysis of massive and / or complex data sets using a series of techniques including: NoSQL, MapReduce and Machine Learning.

Big data researchers, however, remain puzzled as to how to effectively use all of this data. They seek to find a balance between the two equations for the analysis of Big Data; the first equation, if the volume of data increases, then the machine learning algorithms give very precise results, while the second equation, it is hoped that these algorithms will be able to give the results within acceptable time frames. Perhaps because of this inherent conflict, many experts in the field see big data not only as one of the greatest challenges, but also one of the most exciting opportunities over the past decade [Fan et al. 2012].

I.3 Big Data Characteristics

Since the advent of the Internet to this day, we have seen explosive growth in the volume, speed and variety of data created daily. This data comes from many sources including mobile devices, sensors, personal records, Internet of Things, government databases, software logs, public profiles on social networks, business data sets, etc.

In 2001, Gartner proposed a three-dimensional view (volume, variety and velocity) regarding the challenges and opportunities associated with data growth [Chen et al. 2014]. In 2012, Gartner updated this report as follows: Big data is high volume, high speed, and / or wide variety of information resources that require new forms of processing to improve decision making [Erl et al. 2016].

The characteristics that define big data often referred to as the three Vs: Volume, Variety and Velocity (as shown in figure I.1), from so that:

- **Volume:** How much data is there?
- **Variety:** How diverse are the different types of data?
- **Velocity:** How fast is new data generated?

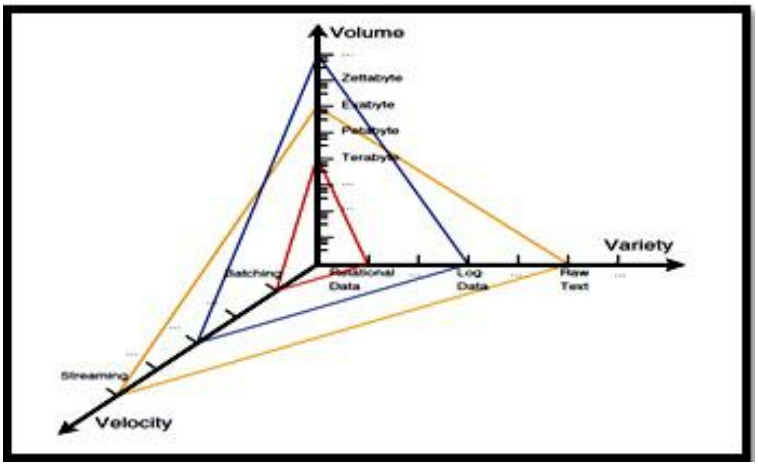


Figure 1: The characteristics (3 Vs) of Big Data.

- **Volume:** The first thing everyone thinks about big data is its size [Lyman et al. 2016] [Eaton et al. 2011]. In the age of the Internet, especially social networks producing streaming data whose volumes are increasing exponentially [Hota & Prabhu 2012] [DBTA 2013]. In 2000, eight hundred thousand (800,000) petabytes of data were stored worldwide [Eaton et al. 2012]. We expect this number to grow to thirty to forty (30-40) zettabytes (Zo) by 2020, it may reach 175(Zo) by the year 2025. For example, on Facebook, over five hundred (500) terabytes (TB) of data is created every day. [Grinter 2013]. Twitter alone generates over seven (7) terabytes (TB) of data every day, and some companies generate terabytes of data every hour of every day [Gantz & Reinsel 2012].

- **Variety:** Previously, all data needed by an organization to run its operations was structured data generated within the organization, such as customer transaction data, etc. Today, companies are looking to leverage much more data from a wider variety of sources, both inside and outside the

company, such as documents, contracts, machine data. , sensor data, social media, medical records, emails, etc. But, the problem is that a lot of this data is unstructured or has a complex structure that is difficult to represent in rows and columns in structured or semi-structured databases [Pattnaik & Mishra 2016] [Chmidt 2012].

- **Velocity:** Just like the volume and variety of data we collect and store, velocity refers to the speed at which data is generated and the time required to process it. Or in another way, it refers to the increasing speed of data generation, processing and use of that data [Power 2014].

Often these features are supplemented by a fourth V, veracity: how accurate is the data?

- **Veracity:** refers to the fact that the data must be credible, accurate, complete and fit for the task. Since big data comes from various sources beyond the control of organizations such as social media. Veracity has become a real problem. False messages or spam are very common, they make trust a major challenge [Chan 2013] [Roos et al. 2013].

We can extend this model to the dimensions of Big Data on ten Vs: volume, variety, velocity, veracity, value, variability, validity, volatility, viability and viscosity [Khan et al. 2018].

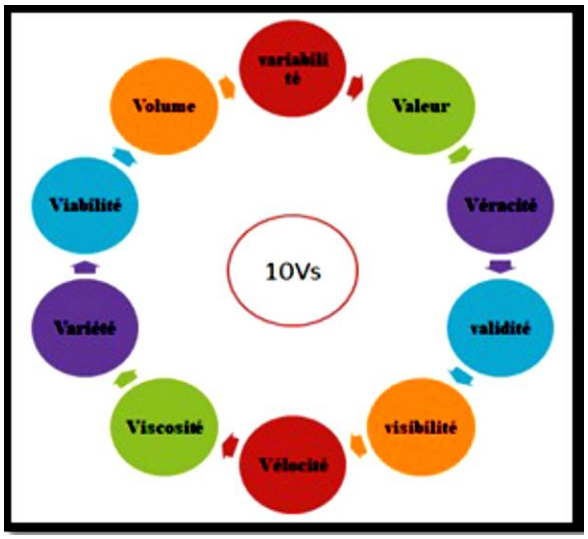


Figure 2: The 10 Vs of Big Data.

- **Value:** is a major factor that all organizations should consider when implementing big data, because the other characteristics of big data don't make sense if you don't derive business value from that data So, we can say that value helps companies better understand their customers [Kayyali et al. 2013].

- **Variability:** Variability in the Big Data context refers to different things. One is the number of inconsistencies in the data. Is the data consistent in terms of availability or reporting interval? Does it

accurately represent the reported event? [Katal et al. 2013] Data should be detected by anomalies and outliers detection methods in order to allow meaningful analysis. Therefore, proper treatment of the property of variability increases the utility of big data systems [Power 2014].

- **Validity:** The term refers to the validation of data. That is, check whether the data used is correct and accurate for the intended use, so that this data is therefore used to assess the performance of the forecast [Ferguson 2013]. Take the example of social media: unlike polls, market specialists use correct methods, but in fact do not have the same concepts and theories. For example, imagine that the meteorological moon indicates that the storm has started somewhere in the world, so how does this storm affect people? With around half a billion users, Twitter feeds can be analyzed to determine the impact of storms on local people. Therefore, using Twitter with data from a weather satellite can help researchers understand the validity of weather forecasts.

However, analyzing non-structural data from a social network still makes it difficult for reliable prediction. Correct input data followed by proper data processing should yield accurate results. With Big Data, you need to be more vigilant about validity.

- **Volatility:** Volatility is the nature of sudden, unstable changes, changed inadvertently or anonymously. The volatility of big data refers to the validity period of the data and its retention [Ripon & Arif 2016]. For example, some companies may keep the most recent data and transactions of their customers; this ensures rapid retrieval of this information when needed.

- **Viability:** Viability means big data needs to be active for a very long time. It must be able to grow, evolve and produce more data when needed. We can identify the characteristics and factors most likely to predict results, so that the most important point for companies is to generate additive value [Khan et al. 2018].

- **Viscosity:** Viscosity refers to the stability and resilience of the large data stream. Big Data offers a limited perspective by telling a certain storytelling. Viscosity measures the resistance to flow in the volume of data. This resistance can come from different data sources, friction resulting from integration rates and the processing required to turn data into information. Technologies for dealing with viscosity include complex event processing, agile integration, and streaming [IBM 2014].

I.4 Big Data Structuring

Big data is very diverse as it comes from different sources and different formats. There are many ways to categorize data types, but one of the most fundamental and relevant differences are between structured and unstructured data.

According to [Iafrate & Front 2015], around 80-90% of future data growth comes from unstructured data types such as media files (videos, images and sound), text files, geospatial and financial data. , which requires different techniques and tools to store, process and analyze. Figure I.3 shows four basic categories.

Big Data structuring can be assigned into four groups:

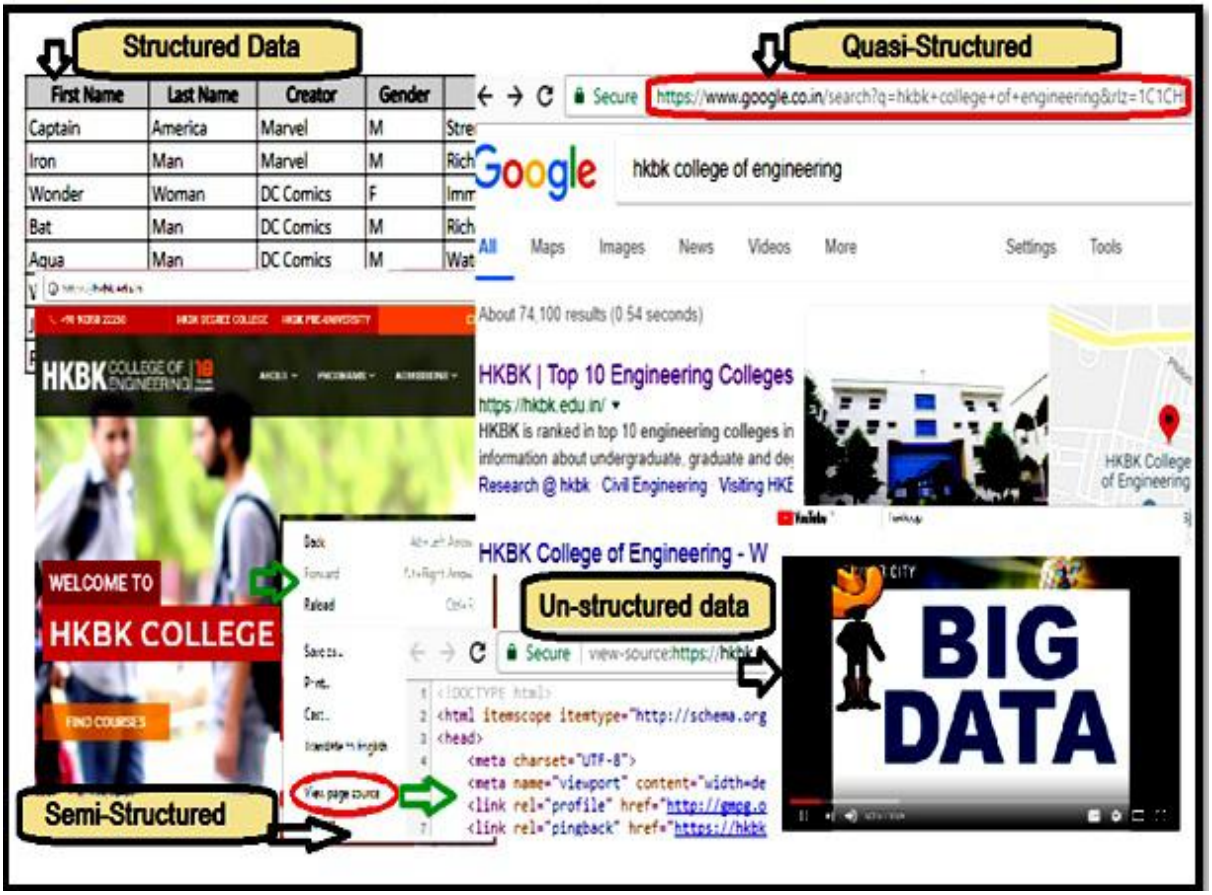


Figure 3: The four basic categories of Big Data Structuring

I.4.1 Structured data

Structured data is data stored in a structure that defines its format. Experts have assumed that structured data is between 5% and 10% of the total amount of data in the world [Kanimozhi & Venkatesan 2015]. Structured data is divided into two main categories; machine-generated data and human-generated data.

I.4.1.1 Machine-generated structured data includes:

- ***Point of sale data:*** are all transactional data generated each time a purchase is made [Hurwitz et al. 2013].
- ***Financial data:*** Financial data was mainly produced by humans and some of them are still produced this data. Most of the processes have been automated and take place without any human intervention [Hurwitz et al. 2013]. For example, trading in stocks that contain structured data such as company symbol and dollar or euro value, etc.

I.4.1.2 Human-generated structured data includes:

- **Input Data:** Humans enter different types of data into computers every day, and some of this information is structured, such as names, ages, and email addresses. One of the most useful types of data is qualitative survey data. They can be used to better understand clients [Hurwitz et al. 2013].
- **Clickstream data:** Another type of data to better understand consumer behavior is clickstream data. Every click of a person while surfing the web is recorded and used to search for patterns [Hurwitz et al. 2013].
- **Game-related data:** The same thing happens with every movement made by customers in a video game. With the growing popularity of video games, this data is gaining in volumes large enough to be considered a separate type [Hurwitz et al. 2013].

In general, structured data can be easily stored, processed, analyzed and queried using traditional analysis tools and software. A typical example can be thought of as a Relational DataBase Management System (RDBMS), transaction data, data files such as spreadsheets. Another example, Microsoft Excel is a great, relatively simple tool for working with structured data. Millions of rows with numbers and titles can be handled easily by knowing the right combinations of formulas and functions.

I.4.2 Semi-structured data

Semi-structured data is data that does not reside in a rational database or other forms of data tables, it may contain tags or other markers to separate semantic elements and apply nested information hierarchies, therefore, this structure is also referred to as a self-descriptive structure.

Although this is a semi-structured data type, it does not have a strict model structure. The best example is textual data files that can be parsed, such as XML -Extensible Markup Language-. These data files are described by an XML schema. JSON format also meets the same specifications as XML. XML and JSON are used to manage semi-structured data, and to convert semi-structured data into structured data [Kanimozhi & Venkatesan 2015].

I.4.3 Quasi-structured data

Although not commonly mentioned, this group can be added to all three categories (structured, semi-structured and unstructured data), namely the one between semi-structured and unstructured. This type of data represents click-stream data from websites, URLs -Uniform Resource Locators-, web applications, etc. These data are used for service level agreements or to forecast security breaches [Hurwitz et al. 2013].

URLs define a flow of clicks that can be analyzed and leveraged by data scientists to uncover usage patterns and highlight relationships between clicks and areas of interest on websites. Then the quasi-structured data can be defined as rather unstructured data and in an erratic format that can be manipulated with special tools. Data flow data may contain inconsistencies in data values and formats [EMCES 2015].

I.4.4 Unstructured data

It is hard to classify data that makes it the opposite of structured data. They have no inherent structure. They currently represent 90% of all data in the world [Kanimozhi & Venkatesan 2015].

Unstructured data can also be divided into two categories: Machine-generated data and Human-generated data.

I.4.4 .1 The unstructured data generated by the machine includes:

- ***Satellite images:*** *meteorological data, landforms, military movements.*
- ***Scientific data:*** *oil and gas exploration, space exploration, seismic imagery, atmospheric data.*
- ***Digital surveillance:*** *Photos and surveillance videos.*
- ***Sensor data:*** *traffic, weather, oceanographic sensors.*

I.4.4 .2 The unstructured data generated by the Human includes:

- ***Text files:*** *Word documents, spreadsheets, presentations, emails, newspapers.*

- ***Email:*** *thanks to its metadata, electronic mail has an internal structure. Sometimes scientists call them semi-structured. However, its message field is unstructured and traditional analysis tools cannot analyze it.*
- ***Social media:*** *Facebook, YouTube, Twitter, Instagram LinkedIn ...*
- ***Website:*** *Wikipedia, encyclopedia, google Map ...*
- ***Mobile data:*** *SMS, MMS ...*
- ***Communications:*** *Chat, instant messaging, phone recordings, collaboration software.*
- ***Media:*** *MP3, digital photos, audio and video files.*
- ***Professional applications:*** *MS Office documents, productivity applications.*

Another big difference between structured and unstructured data is that the latter cannot be analyzed through traditional tools and services. The first obstacle is the volume of data. You cannot store such large amounts of data using the same storage systems as for structured data. It's easy to understand the difference in size: an Excel document with 1,000 rows and 8 columns filled with information at a size of 100 kilobytes, while a single image in JPEG format can easily be about 2 megabytes or more. The next problem is the problem of the data format which is very uncertain; it switches from one unstructured data type to another. Excel works great with numbers, but it can't handle images, videos, Facebook profiles, and long text. Sometimes it is necessary to process all of these types of data simultaneously to gain truly valuable insight.

I.5 Conclusion

The availability of big data, inexpensive basic hardware and new data analysis software created a unique moment in the history of big data analysis (It can be said data mining. These massive data require the development of techniques that can be used to facilitate their analysis. In the second chapter, we will discuss how to prepare this data and how to extract information (move from Big Data to Big Data mining) from it and see it clearly.

Chapter II

Construction and Processing of Big Data Mining

Chapter II Construction and Processing of Big Data Mining

II.1 Introduction

In this chapter, we will see how to collect and process big data and how to derive knowledge and values from this data, knowing that Big data is generated through the internet, such as social networks, international scientific societies, commercial organizations, as well as remote control.

II.2 Data collection

Data collection plays the most important role in the Big Data cycle. The Internet provides almost unlimited sources of data for a variety of subjects. The importance of data collection depends on the type of business, but traditional industries can acquire a diverse external data source and combine it with their transactional data [Benfield & Szlemko 2006].

Data collection is similar to collecting the ingredients for a recipe. If you want to create a great dish, you have to start with the right ingredients, which mean you will have to make a series of decisions up front. For example, if you need honey, do you want generic honey or a specific variety such as orange blossom? Does the brand of honey matter? Does it need to be raw or organic? Do you prefer to get the honey from your local farmer or from a supermarket? And who is going to have all of these ingredients? If you don't have the time, are you willing to pay someone to get it for you, even if that means you might not get exactly what you want? So, it is best to collect different types of data, in different ways, and in different places. Sometimes you don't have enough time or money to collect the data you need on your own. In this section, we'll talk about data collection methods to help us determine where and how to get the best information we're looking for, and how to determine the information we actually need.

II.2.1 Methods of data collection

The way of handling big data is not quite the same as dealing with traditional information. There are various methods that are used by organizations; they are as follows [Paradis et al. 2016]:

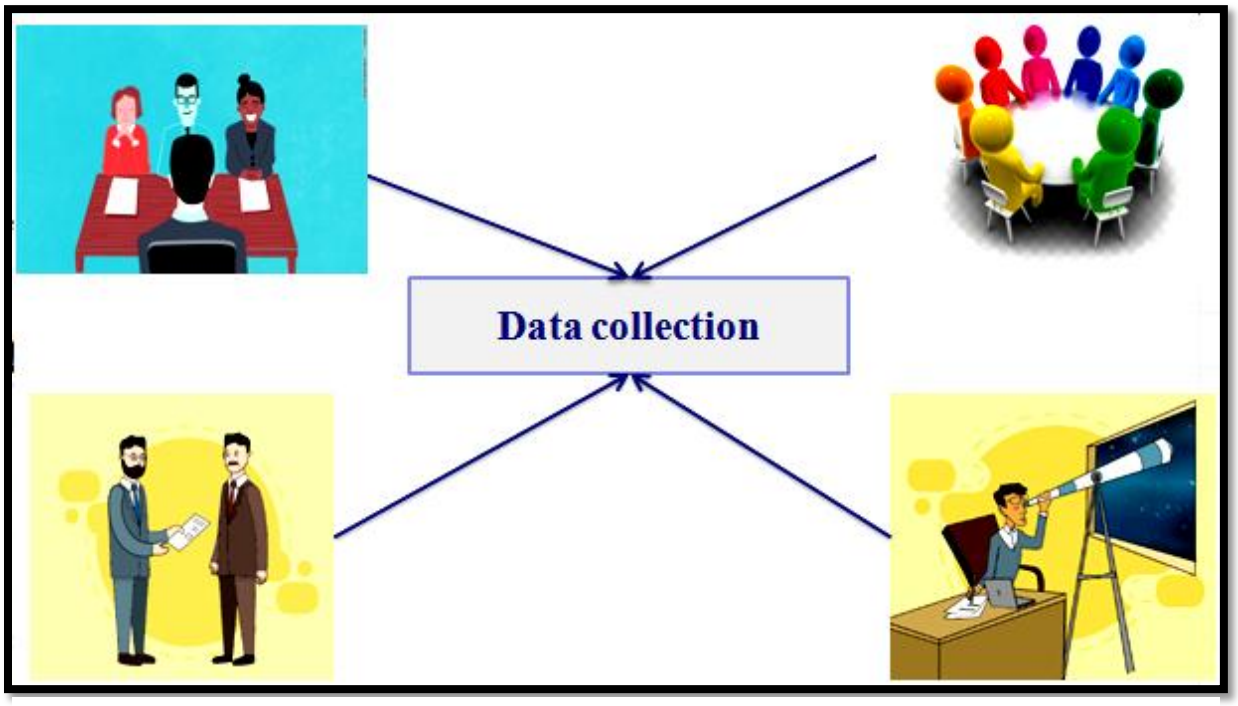


Figure 4: Methods of data collection

II.2.1.1 Observation

The purpose of observation is to gather evidence of achievement by observing a learner's performance while participating in an activity, but without interfering with their work. The activity can be a real situation or a simulated situation. Observation allows you to see knowledge put into practice, it is best used to assess and demonstrate learning based on active citizenship or skills. There are two categories of observations; the first is direct observation, and the second is indirect observation. Direct observation is carried out personally by a teacher, instructor or supervisor in the workplace. Indirect observation occurs when using appropriate technology such as video recording [Mila 2018].

II.2.1.1.1 Direct observation

For example, the teacher's follow-up sheet is completed when the student starts the activity (participation). The teacher records what the student does (work to be done), and how the student behaves and interacts during the session (attendance). Another example, evaluation by monitors, can take the form of a discussion, a question-and-answer session or the recording of information on a platform. The instructor will be another learner who took part in the activity alongside the assessed

learner [Mila 2018]. The instructor will record or provide verbal feedback on what the learner did during the activity. Witness testimony, this is a statement by a "third party" who saw the learner take part in the activity. The witness can be a community worker, a work supervisor or a member of the public. Comments can be given in oral or written form [Kawulich 2005].

II.2.1.1.2 Indirect observation

In indirect observation, the observer does not need to be on site and does not even need to watch a real-time stream of the event. It can be relayed by technological means, such as a recorded event that is relayed back to the viewer from transcriptions of audio recordings, or from verbal behaviors in natural environments [Anguera et al. 2018]. For example, the use of social media, blogs, e-mail or other online archival media are methods of indirect observation [Johnson 2007].

II.2.1.2 Questionnaires

The questionnaire provides the fastest and easiest technique for gathering data on groups of individuals scattered over a large and wide field. The use of questionnaires as a measurement tool depends on the type and duration of the activity. Questions to test or measure learning can be presented in two formats: verbal interrogation, for example a question and answer session at the start and end of a session, or in written form, for example tests or exams. The format chosen should be suitable for the intended use depending on whether the learner is at the start, middle or end of the activity. Questions can be asked to identify knowledge, experience, skills, and accomplishments [Kothari 2004]. Questionnaires may or may not have answers, if there were answers, these are inevitably agreements or disagreements. These responses often go through five levels ranging from "strongly agree" to "strongly disagree". Nothing is more confusing, frustrating and potentially embarrassing than a poorly designed or poorly written questionnaire. Fortunately, with thought and planning these problems can be easily avoided. The design of the questionnaire is a logical process that can be broken down into simple steps. We will take the following steps to help us develop a valid, reliable and efficient tool [Robson 2002].

II.2.1.3 Interviews

The interview is a method of direct investigation. It is simply a social process in which a person known as the interviewer, that interviewer asks questions face-to-face with one or more so-called interviewees to discover perspectives, experiences, feelings and opinions. ideas about a phenomenon [Kothari 2004]. In addition, interviews can reveal success stories that can be used when communicating evaluation

results. Participants may be reluctant to include their results in a questionnaire, but they will voluntarily provide information and answer questions. We can now say that we have more information.

Interviews fall into two basic categories: structured interviews and unstructured interviews. The first category is very much like a questionnaire; in this case the interviewers ask specific questions with little leeway to deviate from the desired answers. One of the main advantages of structured interviews over questionnaires is that questionnaires ensure that all participants answer all questions and that the interviewer understands the participant's responses. Whereas, the second category allows interviewer to search for additional information. This type of interview uses a few general questions that can lead to more detailed information when important data is discovered; in this case, the interviewer should be skilled at asking follow-up questions, as well as seeking additional information if necessary [Dipboye 1994]. The interviewee answers these questions, and the interviewer gathers various information from these answers through very healthy and friendly social interaction. In The Interview, we base some crucial points are as follows:

- We prepare an interview form with questions corresponding to the objective of the assessment;

- We use open and clear questions with the guests;

- We do not test knowledge, but we explore it at through questions of experience and description;

- We do not direct respondents with biased questions and loaded with assumptions;

- We record the conversation with permission (if tape recording is not possible, then we take abbreviated notes).

The main objectives of the interviews were to ensure that the assessor identified all the interviewees relevant to the assessment, and that all questions were prepared in a clear and concise manner. Some questions to assess whether this goal has been achieved are as follows:

- Were we able to identify all the people we needed question?

- Are there any people who were excluded from the interviews with the people we need to include in the next iteration? If so who?

- Did we receive valuable information during the interviews?

- Were there people included in the interviews who did not provide really not much value to the process? Can they be excluded from future interviews? If yes, why?

- Were our questions clear and relevant? Is there anything we need to add to our questions or change to make them clearer?

The interview uses a basic method of communication; it eliminates the limitations and artificiality of writing and filling out a questionnaire. It collects in-depth and detailed data. In addition, it is flexible, open to ensure follow-up. But, it requires a lot of effort and time.

II.2.1.4 Focus groups

Focus groups are useful for exploring norms, beliefs, attitudes, practices and for examining how social knowledge is produced etc., so that, the facilitator stimulates discussion to examine how knowledge and ideas develop and work in a given group. The group is generally made up of six to twelve people [Nyumba et al. 2018].

II.2.1.4.1 Focus group applications

Focus groups are particularly useful when qualitative information is needed on the success of a program. For example, focus groups can be used in the following situations:

- Evaluate reactions to exercises, situations, simulations or other specific components of the program.

- Evaluate the overall effectiveness of the program implementation.

- Evaluate the impact of the program afterwards.

Essentially, focus groups are useful when evaluation information is needed, so that this information cannot be adequately collected using questionnaires, interviews or quantitative methods. Compared to questionnaires and interviews, focus groups have many advantages:

- Focus groups are very close to everyday forms of communication.

- Focus groups can be used to 'explore the ground".

- The researcher obtains information on a particular subject; he can use this information to generate ideas and develop more structured methods such as questionnaires.

On the other hand, this method also has drawbacks:

- Focus groups provide information on a group and not on individuals, they do not provide any information on the frequency or distribution of beliefs in the population.

- A lot of effort and time is required.

In addition, there are other methods used to collect the data, we cite some of these methods:

II.2.1.5 Documents

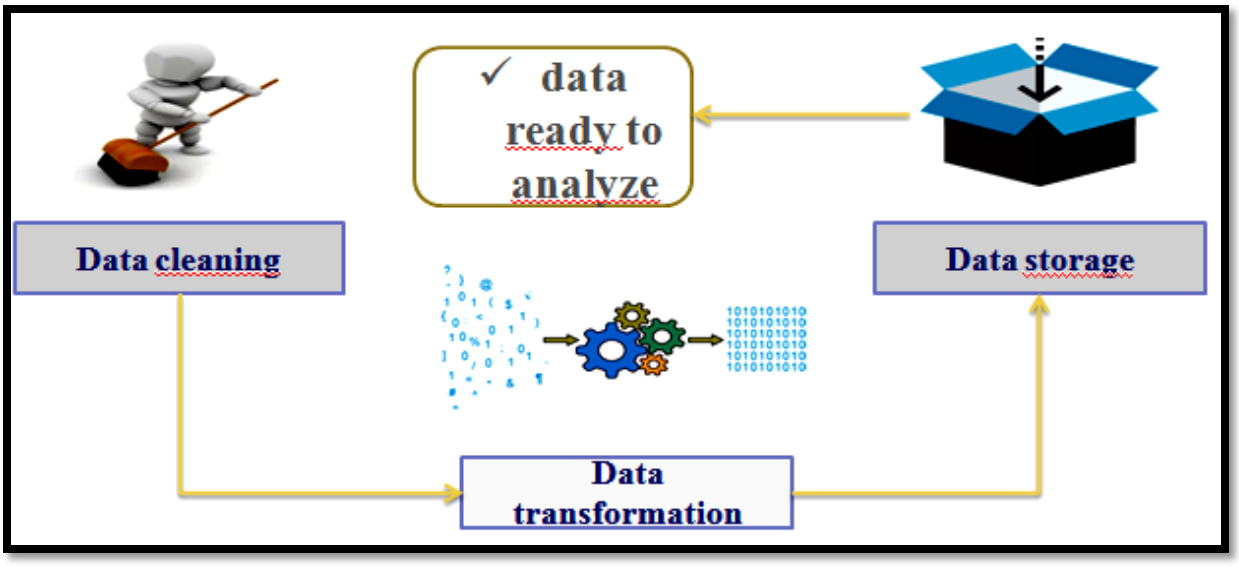
There is a considerable amount of material describing the information of citizens at various levels of social organization. Examples include: The national register of civil status (national center in Algiers), registers kept by the movement of persons (at the level of the wilaya DRAG), registers which are found at the level of the ministerial administrative services to follow the accession ownership, migration, energy production and consumption, employment, taxation, hospitals and schools, records maintained by formal and informal organizations, such as corporations, political parties.

II.2.1.6 Concept map

The concept map is a diagram designed to clarify the understanding of the relationships between the concepts contained in a particular area. A list of words describing important aspects of the topic is compiled. The words in the hierarchy are classified from the most general to the private. They are arranged so that similar terms are close to each other. Links are then created between the words of the concept and the instructions written to describe or explain the links [Faubert & Wheeldon 2009].

II.3 Data preparation

Data preparation (or data pre-processing) is the manipulation of data into a form suitable for



further analysis and processing. The goal of this phase is to provide a dataset that will be used primarily in data analysis. It consists of several general tasks mainly aimed at cleaning and transforming data [Barapatre & Vijayalakshmi 2017].

Figure 5: Data preparation process.

II.3.1 Data cleaning

Data cleaning is a step-in preparing data for analysis by removing or modifying incorrect, incomplete, irrelevant, or repetitive data, because such data is generally neither necessary nor useful for data analysis [Barapatre & Vijayalakshmi 2017]. There are several ways to clean the data depending on how it is stored with the required responses.

Data collection and entry are error-prone processes. They often require human intervention and, since they are just human beings, they make typos or lose focus for a second and introduce errors. But data collected by machines or computers is not error free either. Errors can be caused by human negligence, while others are due to machine failure. But errors must be corrected as quickly as possible for several reasons:

- Decision-makers can make costly errors in information based on incorrect data from applications that do not correct erroneous data.

- If errors are not corrected early in the process, cleanup will need to be done for each project using this data.

- Errors can indicate faulty equipment, such as broken transmission lines and faulty sensors.

- Errors can indicate that a business process is not working as expected.

The main data cleaning tasks [den Broeck et al. 2005] [Krause & Lipscomb 2016] include:

- ***Coding of missing values:*** Data is often missing in some cases. Maybe the family data is collected by the electronic version, but the questions are not asked on the paper version.

- ***Standardization:*** It is good practice to have all data of a similar type in a similar format.

II.3.2 Data transformation

Data transformation is a process of converting or merging data from one format or structure to another format or structure [Kumar & Wu 2007]. In big data field, generally we use XML-Extensible Markup Language- technology to convert any data format into semi-structured data, we can also transform it into structured data. At this stage of the transformation process, there should be a good example of transforming data into a clean and well-formatted format, this data will be mainly used during the data analysis phase.

II.3.3 (Big) Data mining and analysis

Big data isn't just big, it's fast too. This amount of data is sometimes created by a large number of constant streams which usually send data records simultaneously and in small sizes (order of kilobytes). Data streaming includes a wide variety of data such as click stream data, financial transaction data, log files generated by web or mobile applications, sensor data from Internet of Things (IoT), in-game player activity and telemetry from connected devices.

Data analytics is the scientific and statistical tool for analyzing raw data that enables the updating of the information necessary for the acquisition of knowledge. Data analytics works with data to formulate complex decisions from different perspectives to meet real-world challenges. The role of data analytics is to assemble, store, process data in order to put empirical methods in the real world for decision making. It is broadly classified into three types: descriptive analytics, predictive analytics, and prescriptive analytics. Big data analytics now covers almost all aspects of modern society, such as manufacturing, retailing, financial services, etc [Guerra et al. 2015].

Data mining is considered a sub-step of the process known as Knowledge Discovery in Databases. There are the following processes:

- The choice of database;
- Preprocessing, in order to initiate data cleaning;
- Their transformation into the form suitable for their treatment;
- The process of mathematical analysis (data mining);
- Interpretation of the results of the analysis

Chapter II Construction and Processing of Big Data Mining

The knowledge that may have been acquired through KDD (Knowledge Discovery in Databases) is an integral part of the strategic positioning of any online business model, as well as of the marketing decisions that result from it. The fields of application are characterized by their multiplicity.

So, data mining represents the work of processing, graphically or numerically, large amounts or continuous streams of data, with the aim of extracting information useful to those who possess them.

The metaphor "data mining" means that there are treasures or nuggets hidden under mountains of data that can be discovered with specialized tools (classification, Kohonen maps, visualization with methods such as PCA for Principal Component Analysis-, FCA for -Factorial Correspondence Analysis- and MCA for -Multiple correspondence analysis, neural networks, etc.).

The goal of data mining is therefore to discover unknown and useful structures, and networks in large databases. This objective makes it possible to add value to the databases contained in the management information systems of companies and the processing of gigabytes of data. Data mining is therefore all the methods and techniques intended for the exploration and analysis of (large) databases, automatically or semi – automatically, with a view to detecting in these data rules, associations, unknown or hidden trends, particular structures restoring most of the useful information while reducing the quantity of data.

Data mining is either descriptive or predictive analysis: descriptive (or exploratory) techniques aim to highlight information that is present but hidden by the volume of data; predictive (or explanatory) techniques aim to extrapolate new information from present information [Poornima & Pushpalatha 2016]. Descriptive techniques often include constructing quintile tables, measures of dispersion such as variance or standard deviation, and crosstabs that can be used to examine many disparate hypotheses. These assumptions often relate to the differences observed between subgroups. Specialized descriptive techniques are used to measure segregation, discrimination and inequality. Discrimination is also measured using audit studies or decomposition methods. Greater segregation by type or by inequality of outcomes does not necessarily have to be entirely good or bad, but it is often seen as a marker of inequitable social processes. Accurate measurement of levels in time and space is a prerequisite for understanding these processes [Evans & Lindner 2012] (see more details in chapter III.2.1.1).

In our work we are interested in data mining in its cleft predictive analysis; the predictive analysis is a type of analysis that allows you to predict the occurrence of particular events in the future based on data from the past. Predictive analysis is widely integrated into decision-making organizations [Davenport & Kim 2013] (see more details in chapter III.2.1.3**)**.

II.4 Data visualization

Data visualization explains the importance of data by placing it in terms of visual context [Olshannikova et al. 2015]. It is a visual representation of the data. Data visualization allows the user to gain more knowledge about the raw data collected from various sources.

II.4.1 Data visualization in the past

Data visualization is nothing new. Visual data communication has been around in various forms for hundreds, if not thousands, of years.

Abo El-Rayhan Mohamed Ibn Ahmed El-Bayrouni, born in Khouarizm, is one of the oldest researchers of circular representations of the phases of the moon. El-Bayrouni was known in the Islamic world and was one of the most acclaimed scholars of his time. Al-Bayrouni was known from his main book titled "The Mas'udi Canon", concerning astronomy, geography and engineering, one of these pages appearing

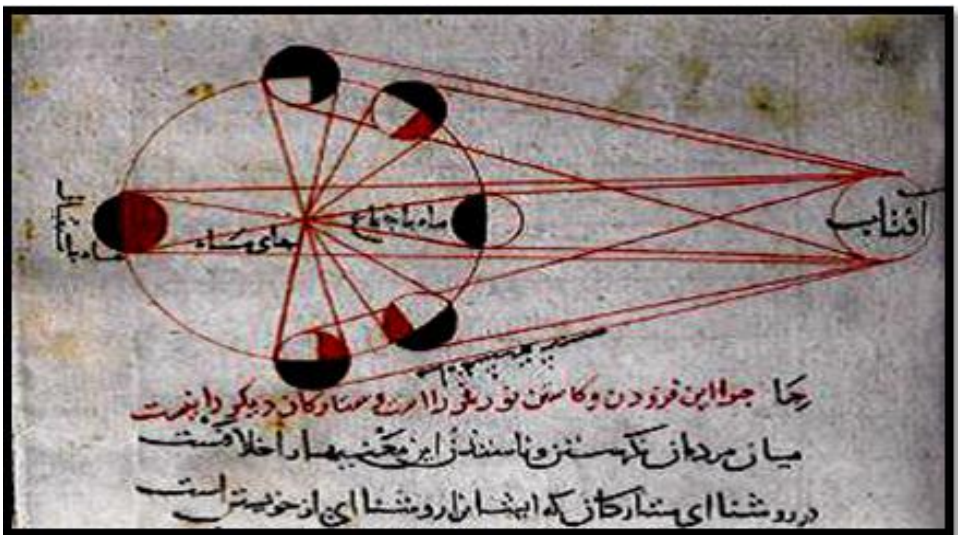


Figure 6: Data visualization in the past.

in the figure below:

In the eighteenth century, common styles still dominated the corporate boards across the country including bars and pie charts.

The appetite for data visualization has increased dramatically in recent years. Data visualization has become a mainstream consciousness over the past decade, it is catalyzed by powerful new technological capabilities, as well as a culture shift towards greater transparency and accessibility of data, the field has experienced rapid growth in enthusiastic participation. While the practice of data visualization was once the preserve of statisticians, engineers and specialist experts, the field of globalization that exists today constitutes a very active, informed, inclusive and innovative community of practitioners who advancing the profession in fascinating directions.

II.4.2 Visualization of (big) data today

Big Data Visualization allows you to visualize a large data set in a format that is easily understood by any user. In addition, the presentation of the data is simple and straightforward, so that the users will easily understand the message [Alhadad 2018]. Simply put, data visualization is the representation of

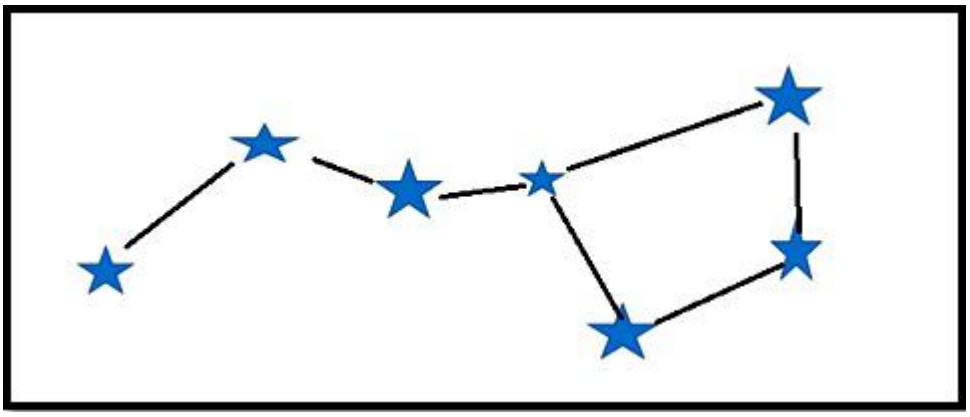


Figure 7: Data visualization today

information in a graphical form as shown in the figure below:

A simple example that can be used to define data visualization, we can draw lines between stars in the night sky to create an image. Imagine certain stars as data points that interest you (among the

billions of other stars visible in the sky) and connect them in a certain order to create an image to visualize the constellation. Currently, it is said in the industry that many disciplines regard data visualization as a modern equivalent of visual communication. Okay, so what is the main purpose or purpose of visual communication or visualizing your data?

When using data visualization, the main goal (although there are other goals) is to make something complicated look simple. Looking at tables and graphs are generally much easier on the eyes. What's more, the point is that we humans are able to process even very large amounts of data much faster when the data is presented in graphical form. Therefore, data visualization is a way to convey concepts universally allowing your target audience or target to quickly get your point of view.

A well-designed visualization allows our eyes to quickly discern patterns of data, allowing us to better understand the underlying characteristics and phenomena of our data. Visual inspection brings us new insights into our data by helping us make assumptions about next priorities. Interaction allows us to explore these hypotheses further, either by showing other parts of the data or by showing the same data from a different perspective. These features make data visualization a valuable tool for exploration, analysis and communication. Data visualization is widely used in companies to communicate their data, and to deepen their understanding of data [Padilla et al. 2018].

Data visualization can guide decision-making, it becomes a tool for conveying information essential to any data analysis [Kinkeldey et al. 2017]. However, to be truly actionable, data visualization must contain the right amount of interactivity. As the volume of data continues to increase, on the one hand, a growing number of vendors and communities are developing tools to create clear and impactful graphics for use in presentations and applications. And on the other hand, many companies believe that understanding this data requires the use of some form of data visualization [Padilla et al. 2018], because it's virtually impossible to visualize a million rows of data and understand their meaning.

II.4.3 Traditional concepts of data visualization

First, we clarify what we mean when we say the word "traditional", i.e., we are referring to ideas and methods that have been used with some success in data visualization over the year’s time. Additionally, most traditional visualization systems cannot handle the size of many contemporary datasets. They are limited to dealing with small data sets that can be easily manipulated and analyzed

with conventional data management techniques. Although it seems that every day new technologies and practices are discovered, developed and deployed offering new and different options. Real-time data visualization is always more ingenious, in addition to understanding the basic concepts for data visualization is also still essential [Bikakis 2018].

At this point, it is essential to fully understand how to choose the most appropriate or effective visualization method. To make this choice, we usually have to answer the following questions:

- What is the volume of data to visualize?
- What are we trying to communicate?
- What is the point we want to communicate?
- Who is our audience? Who will consume this information?
- What type of data visualization could best convey our message to our audience?

We have also been realistic that sometimes the approach or method used is based solely on your time and budget. And you probably already know these more common visualization methods include:

- The tables.
- Histograms.
- Point clouds.
- Line, bar, pie, area, flow and bubble charts.
- Data series or combinations of graphs.
- Timelines.
- Data flow diagrams.

II.4.4 Interactive data visualization

Data visualization is an interesting species. Scientists often know all about data, but visualization issues can be difficult for them. UI designers and graphic designers dominate the visual aspects, but data processing is beyond their reach. Data visualization allows UI designers to familiarize themselves with unfamiliar elements. Data visualization is therefore both a science and an art [Dormehl 2014].

Companies that can analyze real-time data from many forms of data have a great advantage. For example, they can know the feelings of their customers by looking at their purchasing habits. Businesses that leverage their available data by integrating it into interactive data visualizations enjoy the following benefits:

- Users can manipulate the data to find specific information they need.

- Users can be alerted to situations requiring immediate attention.

- When everyone on a team is looking at the same data, they can solve problems more easily.

- Users present only the key elements that allow them to get both the big picture and the details in a single visualization.

- Users can get important insights into business performance from a good interactive visualization.

Today's data visualization has moved to interactive web-based presentations. The increase in data has also led to increasingly visual presentations. Visual interfaces today face difficult challenges, as analysis must be done visually. As a result, an increasing number of user interfaces will appear in dashboard form. These user interfaces have a visual purpose and the data has to be adjusted on the fly [Koppal 2017].

II.4.5 Very important advice in visualization methodology

- Data visualization should be attractive. The advent of more sophisticated and high-quality visual creation tools have, for example, placed it at the level of mobile applications. This will only increase with the evolution of technology.
- We ensure that our data visualization is built on a system that is scalable and accessible for future maintenance and modifications. Because, if our conception of data visualization is successful, others will want to use and exploit it.
- The user must obtain the correct information. Because it's a problem when users are focusing on visualization or on a particular feature that they don't really need.
- Before creating visualization, we define exactly their use in self-service, deep analysis or general presentation.

- The data visualization should be accessible and easy to use; it can be easily changed if needed. In addition, data should be accessible on any device, anytime, anywhere. This functionality is essential for user adoption.

II.5 Conclusion

The availability of big data, inexpensive basic hardware, and new data analytics software created a unique moment in the history of big data analytics. It has been said most often that using this big data with machine learning algorithms results in better prediction results. It is also possible to find these same results in the shortest time if one uses sophisticated technologies; this will be detailed in the next chapter, and will also be verified in the last chapter.

Chapter III

Big Data Mining: Analytucs and technologies

III.1 introduction

The purpose of this chapter is to establish an understanding of data analysis focusing on how predictive analysis forecasts future values, getting familiar with Hadoop and how its ecosystem implements big data analytics and seeing some of the machine learning algorithms that are most compatible in building models in the big data world.

III.2 Data Analytics

Analytics is a broad term that encompasses the processes, technologies, frameworks and algorithms to extract meaningful insights from data. Raw data in itself does not have a meaning until it is contextualized and processed into useful information. Analytics is this process of extracting and creating information from raw data by filtering, processing, categorizing, condensing and contextualizing the data. This information obtained is then organized and structured to infer knowledge about the system and/or its users, its environment, and its operations and progress towards its objectives, thus making the systems smarter and more efficient [A. Bahga et al.2019].

III.2.1 Types of (Big) data analytics

Analysis of data is a vital part of running a successful business. When data is used effectively, it leads to better understanding of a business's previous performance and better decision-making for its future activities. We all know there are different analytics to process the data, but many businesses don't know when to leverage what. In fact, what distinguishes the best data scientist or data analyst from others is their capability to identify the right kind of analytics that best fits the business needs to maximize outcomes. All types of analytics offer distinct insights. In our work, we will explore the four

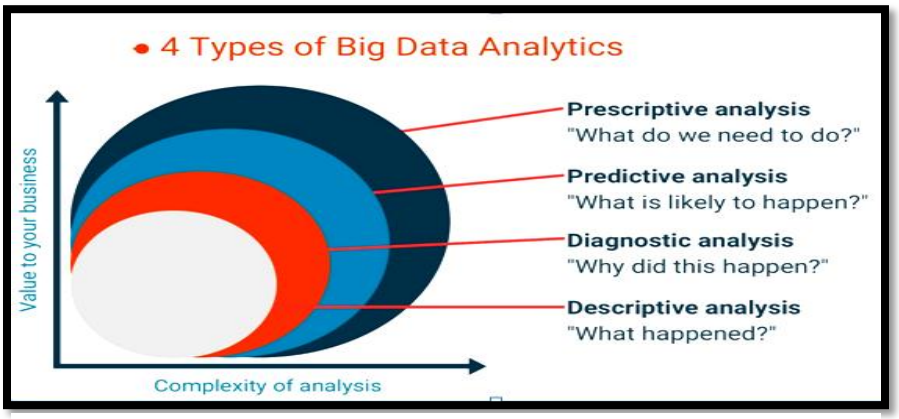


Figure 8: types of big data analytics

different types of data analytics (see figure below).

III.2.1.1 descriptive analytics

Descriptive analysis does exactly what the name implies it "describes", or summarizes raw data and makes it something that is interpretable by humans. It is used to describe the basic characteristics of the study data. It limits the generalization to a particular group of individuals observed. No conclusions extend beyond this group and no similarities to those outside the group can be assumed. The data describes one group and that group only. Very simple research involves descriptive analysis and provides valuable information about the nature of the particular group of individuals [Best & Kahn 2003]. For that reason, descriptive analytics comprises analyzing past data to present it in a summarized form which can be easily interpreted. Descriptive analytics aims to answer - What has happened?

Descriptive analysis provides simple summaries of the sample and measurements, along with simple graphical analysis; it is the virtual basis for any quantitative data analysis. With descriptive analysis, you simply describe what the data shows. Description of the data is necessary to determine the normality of the distribution. The description of the data is necessary, because the nature of the techniques to be applied for the inferential statistics of the data depends on the characteristics of the data [Bhaskar & Zulfiqar 2016]. Descriptive analysis is used to summarize data, such as mean, standard deviation for continuous data types (such as age), while frequency and percentage are useful for categorical data (such as than the genre). The descriptive analysis does not provide details on why certain events occurred or what can be said in the future [Davenport & Kim 2013].

III.2.1.2 diagnostic analytics

Diagnostic analytics is a more advanced form of analytics which examines data or content to answer the question “Why did it happen?”, and is characterized by techniques such as drill-down, data discovery, data mining and correlations. You can see how the human input in this type of analytics remains high. The goal of the diagnostic analytics is to help you locate the root cause of the problem. To do so, the algorithms use owned proprietary data, and leverage outside information to understand what exactly happened and help you find a quick fix[1].

[1] https://www.logianalytics.com/predictive-analytics/comparing-descriptive-predictive-prescriptive-and-diagnostic-analytics/

III.2.1.3 predictive analytics

Predictive analytics is used to make predictions about unknown future events. It uses many techniques, such as statistical algorithms, data mining, statistics, modeling, machine learning and artificial intelligence, to analyze current data and make predictions about the future. It aims to identify the likelihood of future outcomes based on the available historical data [A. Zhang. 2017]. Predictive analytics is the process of taking historical data (the past), identifying patterns in the data that are seen though some methodology (the model), and then using the model to make predictions about what will happen in the future (scoring new data) [J. Dean. 2014].

So, predictive analytics is the art of building and using models that make predictions based on models extracted from historical data. This includes empirical predictive models (statistical models such as data mining algorithms) that predict future scenarios and evaluation methods to assess the predictive power of a model [Schmueli & Koppius 2011]. Predictive analysis identifies relationships between variables and then, based on those relationships, it predicts the likelihood of a certain event occurring. Although predictive analysis relies on strong relationships between data, sometimes ill-defined relationships can be expected [Evans & Lindner 2012] [Davenport & Kim 2013].

In predictive analytics, we use a broad definition of the word prediction. In current practice, the word prediction has a temporal aspect: we predict what will happen in the future. However, in data analysis, a prediction is the assignment of a value to an unknown variable. This can be predicting the price at which a product will be sold in the future or, conversely, predicting the type of document. So in some cases the prediction has a temporal aspect but not at all. In our work, we use Machine Learning algorithms to train predictive models.

III.2.1.3.1 Predictive Data Analytics Project Lifecycle

Like any other significant project, the chances of a predictive data analytics project being successful are greatly increased if a standard process is used to manage the project through the project lifecycle. One of the most commonly used processes for predictive data analytics projects is the Cross Industry Standard Process for Data Mining (CRISP-DM).Key features of the CRISP-DM process that make it attractive to data analytics practitioners are that it is non-proprietary; it is application, industry, and tool neutral; and it explicitly views the data analytics process from both an application-focused and a technical perspective [J. D. Kelleher et al. 2015].

- **Business Understanding:** Predictive data analytics projects never start out with the goal of building a prediction model. Instead, they are focused on things like gaining new customers, selling more products, or adding efficiencies to a process. So, during the first phase in any analytics project, the primary goal of the data analyst is to fully understand the business (or organizational) problem that is being addressed, and then to design a data analytics solution for it.
- **Data Understanding:** Once the manner in which predictive data analytics will be used to address a business problem has been decided, it is important that the data analyst fully understand the different data sources available within an organization and the different kinds of data that are contained in these sources.
- **Data Preparation:** Building predictive data analytics models requires specific kinds of data, organized in a specific kind of structure known as an Analytics Base Table (ABT).
- **Modeling:** The modeling phase of the CRISP-DM process is when the machine learning work occurs. Different machine learning algorithms are used to build a range of prediction models from which the best model will be selected for deployment.
- **Evaluation:** Before models can be deployed for use within an organization, it is important that they are fully evaluated and proved to be fit for the purpose. This phase of CRISP-DM covers all the evaluation tasks required showing that a prediction model will be able to make accurate predictions after being deployed and that it does not suffer from overfitting or underfitting.
- **Deployment:** Machine learning models are built to serve a purpose within an organization, and the last phase of CRISP-DM covers all the work that must be done to successfully integrate a machine

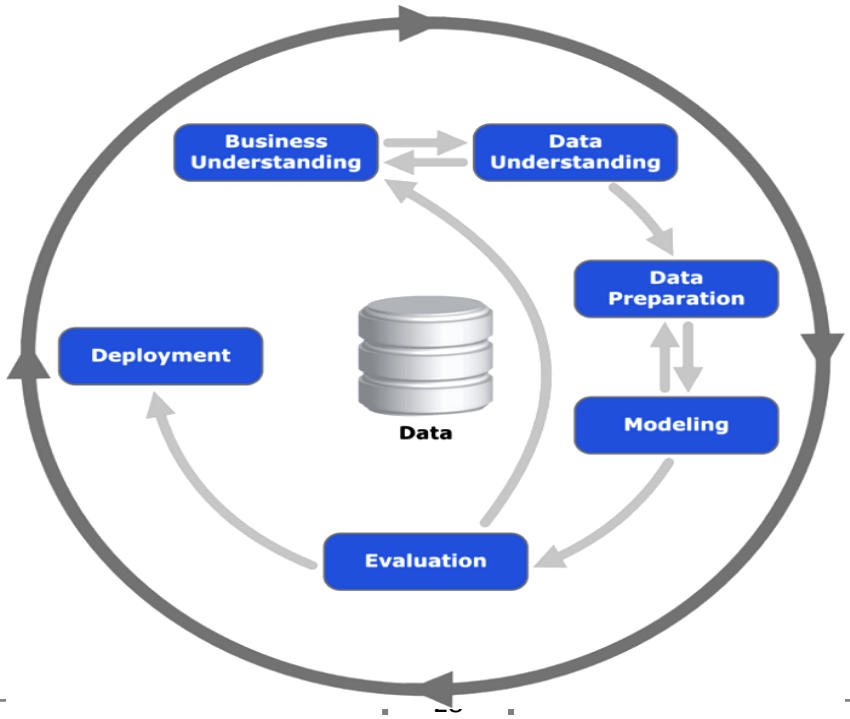


Figure 9: The predictive data analytics project lifecycle by the CRISP-DM.

learning model into the processes within an organization.

III.2.1.4 perspective analytics

Prescriptive analytics has been one of the big buzzwords of recent years. Being able to automatically prescribe actions to achieve a goal would mean a huge step forward in automatic help or decision making in all areas. It suggests the best course of action to optimize the results of your business. Typically, prescriptive analytics combines a predictive model with business rules (such as declining a transaction if the probability of fraud is above a certain threshold). This technique is used to more effectively support decision-making based on diverse ideas emanating from decision makers such as technical directors and general managers analyze and predict complex situations. So, it is considered nirvana in analysis, it is often used by the most analytically sophisticated organizations [Gim et al. 2018].

Prescriptive analysis, on the other hand, consists of suggesting a number of actions; it includes methods of experimental design and optimization. The experimental design explains the reasons why a phenomenon occurs by performing experiments in which independent variables are manipulated, superfluous variables are controlled, and therefore conclusions are drawn from the actions that need to be taken by the decision maker.

III.3 Hadoop and its ecosystem

III.3.1 Hadoop

Apache Hadoop is a framework that performs distributed processing of massive datasets across clusters of computers that scale up from a single server to thousands. Apache Hadoop is an open source framework for reliable, scalable and distributed computing over a massive amount of data developed in Java and consists of four main sub projects: MapReduce, Hadoop Distributed File System (HDFS), YARN, and common Hadoop utilities like Hbase, Zookeeper, Avro and some Other [N. M. F. Qureshi et al. 2005].

Hadoop is a platform that provides both distributed storage and computational capabilities. Hadoop was first conceived to fix a scalability issue that existed in Nutch[2], an open source crawler and search engine. At the time, Google had published papers that described its novel distributed file system, the Google File System (GFS), and MapReduce, a computational framework for parallel processing. The successful implementation of these papers concepts in Nutch resulted in it being split into two separate projects, the second of which became Hadoop, a first-class Apache project[A. Holmes. 2015].

III.3.2 Hadoop ecosystem

Handling huge volume of data generating from billions on online activities and transactions require continuous up gradation and evolution of Big data. Hadoop ecosystem is a framework of various type of complex and evolving tools and techniques. MapReduce and Hadoop Distributed File System (HDFS) are two components of Hadoop ecosystem which manages big data.

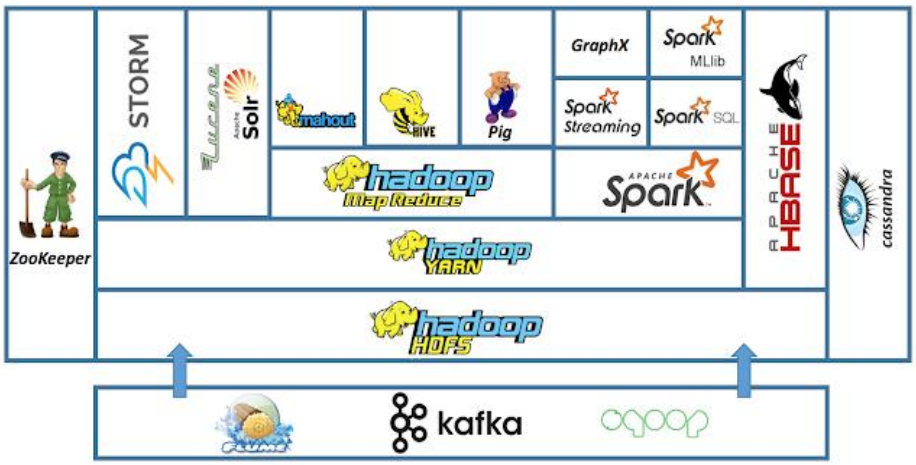


Figure 10: Hadoop ecosystem

III.3.2.1 MapReduce programming model

The computation takes a set of input key/value pairs, and produces a set of output key/value pairs. The user of the MapReduce library expresses the computation as two functions: Map and Reduce. Map, written by the user, takes an input pair and produces a set of intermediate key/value pairs. The MapReduce library groups together all intermediate values associated with the same intermediate key I and passes them to the Reduce function. The Reduce function, also written by the user, accepts an intermediate key I and a set of values for that key. It merges together these values to

[2]The Nutch project, and by extension Hadoop, was led by Doug Cutting and Mike Cafarella.

form a possibly smaller set of values. Typically, just zero or one output value is produced per Reduce invocation. The intermediate values are supplied to the user's reduce function via an iterator. This allows us to handle lists of values that are too large to fit in memory [H. Yang. et al. 2007] [S. Sakr et al. 2011]:

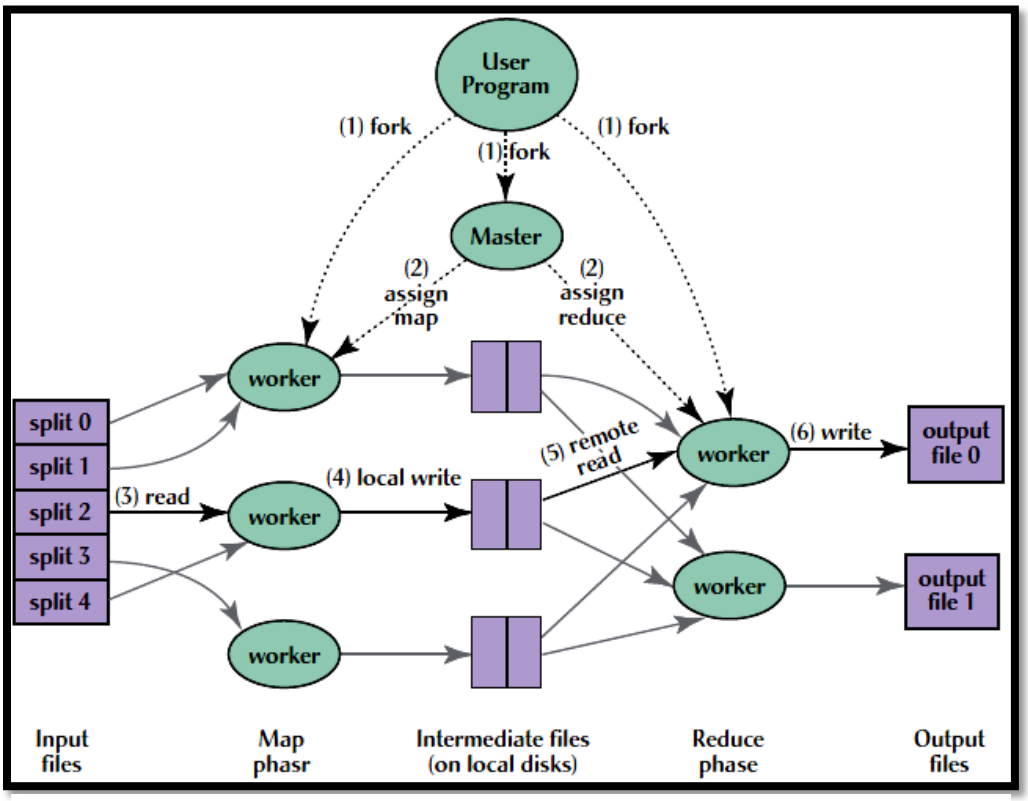


Figure 11: An Overview of the Flow of Execution a MapReduce Operation.

1) The input files of the MapReduce program is split into M pieces and starts up many copies of the program on a cluster of machines.
2) One of the copies of the program is elected to be the copy while the rest are considered as workers that are assigned their work by the master copy. In particular, there are M map tasks and R reduces tasks to assign. The master picks idle workers and assigns each one a Map task or a reduce task.
3) A worker who is assigned a map task reads the contents of the corresponding input split and parses key/value pairs out of the input data and passes each pair to the user defined Map function. The intermediate key/value pairs produced by the Map function are buffered in memory.

4) Periodically, the buffered pairs are written to local disk, partitioned into R regions by the partitioning function. The locations of these buffered pairs on the local disk are passed back to the master, who is responsible for forwarding these locations to the reduce workers.
5) When a reduce worker is notified by the master about these locations, it reads the buffered data from the local disks of the map workers which is then sorted by the intermediate keys so that all occurrences of the same key are grouped together. The sorting operation is needed because typically many different keys map to the same reduce task.

6) The reduce worker passes the key and the corresponding set of intermediate values to the user's Reduce function. The output of the Reduce function is appended to a final output file for this reduce partition.
7) When all map tasks and reduce tasks have been completed, the master program wakes up the user program. At this point, the MapReduce invocation in the user program returns back to the user code.

III.3.2.2 Hadoop Distributed File System (HDFS)

HDFS is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is now an Apache Hadoop subproject[3].

III.3.2.2.1 HDFS architecture

[3] https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

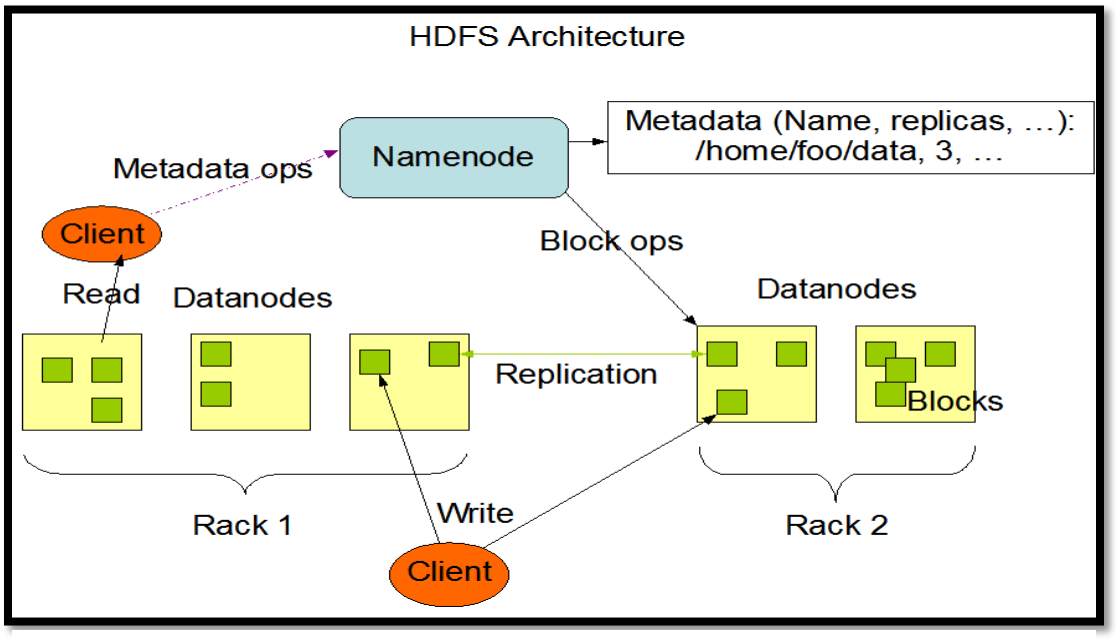


Figure 12: HDFS Architecture.[4]

HDFS is based on the GFS and has master/slave style architecture. The TCP/IP stack protocol supports HDFS, in which a client uses the Client-Protocol with the Name-Node through a port. Data-Nodes adopt the Data-Node-Protocol with the Name-Node, and these protocols execute a Remote Procedure Call (RPC).

An HDFS cluster comprises a unique Name-Node, a master-server which is responsible for the system file management and controls access to all client files. On the other hand, there is a set of Data-Nodes, usually one per cluster node, which acts as manager for local storage. The Name-Node operates all commands in the file systems, such as open, close, and renames (file or directory).

We mention below, but are not limited to some of the most used hadoop ecosystems in the Big Data context:

III.3.2.3 Cassandra

Apache Cassandra is a distributed NoSQL database system based on Amazon's Dynamo and Google's Big table. Cassandra is a fast, distributed database that's highly fault tolerant as well as scalable. It provides high availability and linear scalability, twin goals that traditional relational databases cannot satisfy when handling very large data sets. Cloud applications require highly scalable back-end databases that are capable of distributed, massive workloads across clusters of servers. These applications require very fast access to data to satisfy interactive usage of the data stores by various

applications, as well as ad-hoc queries. Cassandra is expressly designed for high-volume, low-latency cloud applications [S. R. Alapati. 2018].

III.3.2.4 HBase

HBase is a column-oriented non-relational database management system that runs on top of Hadoop Distributed File System (HDFS). HBase provides a fault-tolerant way of storing sparse data sets, which are common in many big data use cases. It is well suited for real-time data processing or random read/write access to large volumes of data[4].

Unlike relational database systems, HBase does not support a structured query language like SQL; in fact, HBase isn't a relational data store at all. HBase applications are written in Java™ much like a typical Apache MapReduce application. HBase does support writing applications in Apache Avro, REST and Thrift.

III.3.2.5 Zookeeper

Apache ZooKeeper is an effort to develop a highly scalable, reliable, and robust centralized service to implement coordination in distributed systems that developers can straightaway use in their applications through a very simple interface to a centralized coordination service. It enables application developers to concentrate on the core business logic of their applications and rely entirely on the ZooKeeper service to get the coordination part correct and help them get going with their applications. It simplifies the development process, thus making it nimbler[5].

Zookeeper mitigates the need to implement coordination and synchronization services in distributed applications from scratch by providing simple and elegant primitives through a rich set of APIs.

III.3.2.6 Pig

Pig is a platform for analyzing large data sets with a sophisticated environment for optimization and debugging. It introduced a scripting-based language called Pig Latin that is used for data processing. Pig Latin is data flow language that follows a step-by step process to analyze data. Pig Latin can launch MapReduce, Tez, and Spark jobs. Pig Latin can call Java, JavaScript, Python, Ruby, or Groovy code through UDFs [B. Vaddeman. 2016].

[4] https://www.ibm.com/analytics/hadoop/hbase
[5] https://hbase.apache.org

III.3.2.7 Apache Hive

Hive is a standard for SQL queries over peta-bytes of data in Hadoop. It provides SQL-like access to data in HDFS, enabling Hadoop to be used as a data warehouse. The Hive Query Language (HQL) has similar semantics and functions as standard SQL in the relation al database, so that experienced database analysts can easily get their hands on it. Hive's query language can run on different computing engines, such as MapReduce, Tez, and Spark [D. Du. 2018].

III.3.2.8 Flume

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application[6].

III.3.2.9 Storm

Apache Storm has emerged as the platform of choice for industry leaders to develop distributed real-time, data processing platforms. It provides a set of primitives that can be used to develop applications that can process a very large amount of data in real time in a highly scalable manner [A. Jain. 2017].

III.3.2.10 apache Spark

Apache Spark is a cluster computing platform designed to be fast and general purpose. On the speed side, Spark extends the popular MapReduce model to efficiently support more types of computations, including interactive queries and stream processing. Speed is important in processing large datasets, as it means the difference between exploring data interactively and waiting minutes or hours. One of the main features Spark offers for speed is the ability to run computations in memory, but the system is also more efficient than MapReduce for complex applications running on disk [H. Karau et al. 2015].

[6]https://flume.apache.org

On the generality side, Spark is designed to cover a wide range of workloads that previously required separate distributed systems, including batch applications, iterative algorithms, interactive queries, and streaming. By supporting these workloads in the same engine, Spark makes it easy and inexpensive to combine different processing types, which is often necessary in production data analysis pipelines. In addition, it reduces the management burden of maintaining separate tools. Spark is designed to be highly accessible, offering simple APIs in Python, Java, Scala, and SQL, and rich built-in libraries. It also integrates closely with other Big Data tools. In particular, Spark can run in Hadoop clusters and access any Hadoop data source, including Cassandra [H. Karau et al. 2015].

III.3.2.11 Kafka

Kafka is a popular high-volume, low-latency messaging system for handling real-time data feeds. Kafka partitions the data streams and spreads them over a cluster, which also allows for multiple coordinated "consumers" to consume the data, which is generated by "producers" [S. R. Alapati. 2018].

III.4 Machine learning

Machine learning is a subfield of computer science that is concerned with building algorithms which, to be useful, rely on a collection of examples of some phenomenon. These examples can come from nature, be handcrafted by humans or generated by another algorithm [A. Burkov. 2019].

Machine learning can also be defined as the process of solving a practical problem by gathering a dataset, and algorithmically building a statistical model based on that dataset. That statistical model is assumed to be used somehow to solve the practical problem.

III.4.1 Objectives and uses of machine learning

Machine learning has evolved from the broad field of artificial intelligence. It provides scientists with a way to explore learning models and learning algorithms that can help machines (for example computers) learn the system from data; therefore, the main goal of machine learning is to equip machines with human intelligence; so that it is able to provide predictions based on a huge amount of data, which is an almost impossible task for a human being [Burhan et al. 2014]. The goal of machine learning is to find the predictive model that generalizes the best. In order to find this best model, a machine learning algorithm must use certain criteria to choose among the candidate models it considers in its research.

III.4.2 types of machine learning

Machine learning can be thought of as a set of algorithms capable of automatically recognizing data patterns, while recognized data patterns are used to predict new observed values. Machine learning algorithms are broadly classified into three categories: supervised, unsupervised and reinforcement learning [Dasgupta & Nath 2016]. Each learning method has its meaning and its dimensionality [M. Mohri. 2018].

III.4.2.1 Supervised learning

Supervised learning algorithms vary from application to application; however, they fall into three categories: learning phase, validation phase and testing phase [Kotsiantis 2007]. The algorithms developed for a particular problem in these categories depend on the designer and developer of the algorithm. Supervised learning algorithms for big data are more complex. They have to take into account the physical operation. The volume of data is unmanageable, the number of class types is large, and the speed required to process data is high. Therefore, it needs distributed file sharing, parallel processing technology, lifelong learning techniques, and multi-domain representation learning techniques [Chiliang & Wenting 2012]. In the supervised learning method, a model from labeled learning data is interpreted to allow the prediction of test data [Caruana et al. 2008]. Supervised learning refers to known labels (predicted classes are known beforehand) as a set of samples to achieve the desired result. We mention some known algorithms in supervised learning:

- **Random forest:** according to [J. Mueller et al. 2016] RF is a classification and regression algorithm that uses a large number of decision tree models built on different sets of bootstrapped examples and sub sampled features, and according to [S. Shalev-Shwartz et al. 2014] a random forest is a classifier consisting of a collection of decision trees, where each tree is constructed by applying an algorithm A on the training set S and an additional random vector, θ, where θ is sampled independent and identically distributed from some distribution. The prediction of the random forest is obtained by a majority vote over the predictions of the individual trees.
- **Support Vector Machine:** Support Vector Machines is an efficient and effective pattern recognition technique, which is based on Vapnik-Chervonenk is structural risk minimization theory.

SVM learning is based on mapping the sample points into a high-dimensional feature space in order to search and obtain an optimal separating hyperplane, which maximizes the sum of the distances between two classes in this space [Demidova et al. 2016].

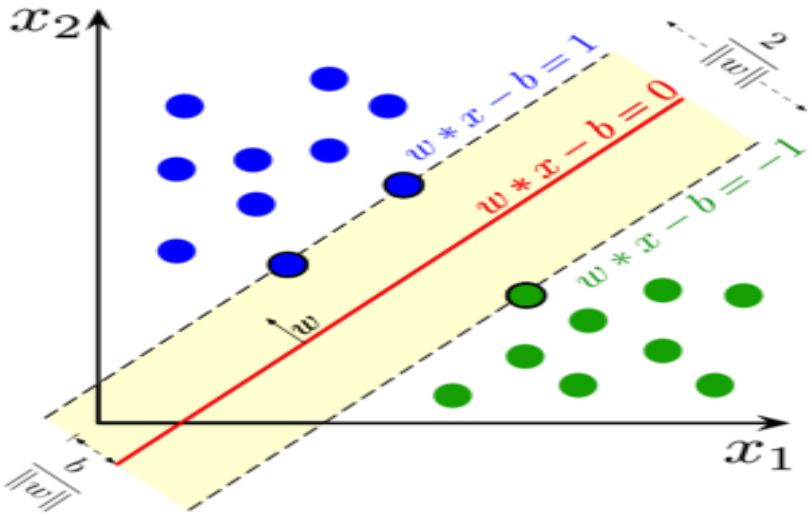


Figure 13: hyperplane and margins for an SVM trained with samples from two classes.

- **K-nearest neighbor (KNN):** The k-NN algorithm is a non-parametric method that can both classification and regression problems as one of the simplest of all machine learning algorithms. a sample is classified by estimating the majority vote of its neighbors, with the new object assigned to the class that is most common among its nearest neighbors (k being a positive integer, and typically small). It belongs to the family of lazy learning which means that it does not carry out an explicit training phase (it does not need to build a model) and new unseen cases are classified on-the fly by comparing them against the entire training set. In spite of its simplicity, the k-NN is known because it usually offers a good performance in a wide variety of problems. However, this method becomes very sensitive to the local structure of the training data (that needs to be kept stored on a drive). Thus, the classical k-NN algorithm suffers from a number of weaknesses that affect its accuracy and efficiency [I. Triguero et al. 2019].
- **Naïve Bayes:** The Naïve Bayes algorithm uses Bayes theorem to train a classifier. The model trained by the Naïve Bayes algorithm is a probabilistic classifier. For a given observation, it calculates a probability distribution over a set of classes. The Naive Bayes algorithm assumes that all the features or predictor variables are independent. That the reason it is called naïve. In theory, the Naive Bayes algorithm should be used only if the predictor variables are statistically independent; however, in

practice, it works even when the independence assumption is not valid. Naïve Bayes is particularly suited for high dimensional datasets. Although it is a simple algorithm, it often outperforms more sophisticated classification algorithms [M. Guller. 2015].

- **Artificial neural networks (ANNs):** Artificial Neural Networks make up an integral part of the Deep Learning process. Are computational networks which attempt to simulate, in a gross manner, the decision process in networks of nerve cell (neurons) of the biological (human or animal) central nervous system. This simulation borrows from the neuro-physiological knowledge of biological neurons and of networks of such biological neurons. It thus differs from conventional (digital or analog) computing machines that serve to replace, enhance or speed-up human brain computation without regard to organization of the computing elements and of their networking [D. Graupe. 2013].

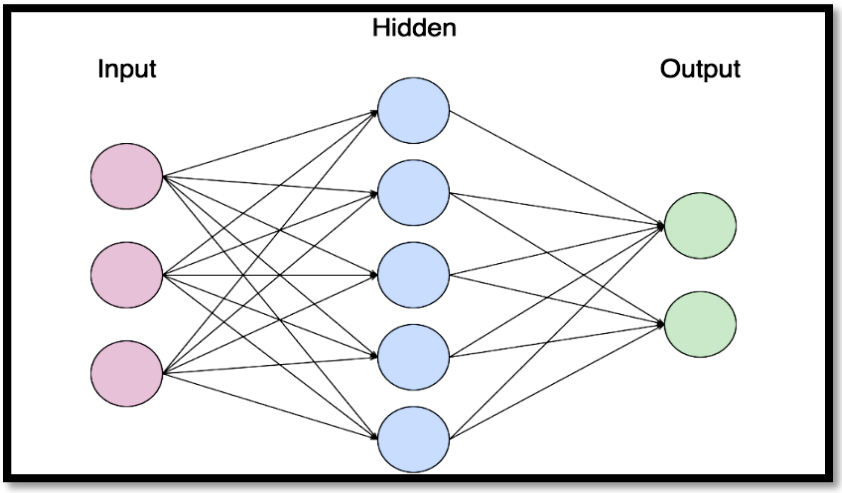


Figure 14: a simple example of a neural network.

III.4.2 Unsupervised learning

The unsupervised learning method processes data without labels or of an unknown structure. In this type of learning, algorithms learn on their own without supervision or any target variables provided. It is about finding patterns and relationships hidden in the data provided, so that, the learner receives exclusively unlabeled learning data, he makes predictions for any unseen points. The lack of direction of the learning algorithm in unsupervised learning can sometimes be advantageous, as it allows the algorithm to search for patterns that had not yet been considered [Kohonen & Simula 1996]. It can be difficult to quantitatively assess a learner's performance.

Further, clustering and dimensionality reduction are categories of unsupervised learning problems [Dayan 1999]. We mention some known algorithms in unsupervised learning:

- **K-means:** The k-means algorithm finds groupings or clusters in a dataset. It is an iterative algorithm that partitions data into k mutually exclusive clusters, where k is a number specified by a user. The k-means algorithm uses a criterion known as within-cluster sum-of-squares for assigning observations to clusters. It iteratively finds cluster centers and assign observations to clusters such that within-cluster sum-of-squares is minimized. The number of clusters in which data should be segmented is specified as an argument to the k-means algorithm [M. Guller. 2015].

- **Principal Components Analysis (PCA):** PCA is used for dimensionality reduction. It is a statistical method for reducing a large set of possibly correlated variables to a smaller set of uncorrelated variables, known as principal components. The number of principal components is less than or equal to the number of original variables. The goal of PCA is to find the fewest number of variables responsible for the maximum amount of variability in a dataset. The first principal component is the variable with the largest variance. The second component is the variable with the second largest variance and it is statistically independent with respect to the first component. Similarly, the third component is the variable with the third largest variance and orthogonal to the first two components. This is true for each succeeding principal component. Thus, each principal component has the largest variance possible under the constraint that it is uncorrelated to the previous components [M. Guller. 2015].

III.4.3 Reinforcement learning

Reinforcement learning is between supervised learning and unsupervised learning. It is often considered a branch of artificial intelligence; it has been one of the central subjects in a wide range of scientific fields for the past two decades. In reinforcement learning, sequential decisions have to be made rather than a single decision making, which sometimes makes the learning phase a bit difficult; so, the algorithm is informed when the answer is wrong, but does not tell how to correct it. He must explore and experiment with different possibilities until he finds the right answer. Reinforcement learning mimics how humans learn by interacting with environment, repeating actions for which the reward that is received is higher, and avoiding risky moves for which there is a low or negative reward as an outcome of their actions. In reinforcement learning, we do not have a target variable. Instead we

have reward signals, and the agent needs to plan the path on its own to reach the goal where the reward exists [P. Dangeti. 2017]. We can solve reinforcement learning problems using:

- **Markov decision process:** The Markov property states that the future depends only on the present and not on the past. The Markov chain is a probabilistic model that solely depends on the current state to predict the next state and not the previous states, that is, the future is conditionally independent of the past. The Markov chain strictly follows the Markov property. MDP is an extension of the Markov chain. It provides a mathematical framework for modeling decision-making situations. Almost all Reinforcement Learning problems can be modeled as MDP [S. Ravichandiran 2018].

- **Dynamic programming-DP:** The term dynamic programming refers to a set of algorithms that can be used to calculate optimal policies from a perfect model of the environment in the form of a Markov Decision Process (MDP). Dynamic programming makes it possible to solve large complex problems by breaking them down into smaller sub-problems which are then solved independently, the results of which are combined to form the solution to the initial problem [Dreyfus 2002].

- **Monte Carlo methods:** Monte Carlo methods are one way to solve the problem of reinforcement learning based on the mean of sample returns. They make random selections from samples based on an assumed model. They only require experiences through sequences of sample states, actions and rewards drawn from real or simulated interactions with an environment [Nutini 2017].

III.5 Conclusion

In this chapter we have learned that Machine learning alongside with Hadoop and its ecosystem gives a powerful instrument to tackle massive data and gain insight. In the fourth chapter we will see an implementation of the most famous and approved supervised machine learning algorithms for handling big data. Then we will compare these algorithms and see which one works best and which one adapts and gives better results in the Big Data world.

Chapter IV
Realization and implementation

Chapter IV Realization and Implementation

IV.1 Introduction

In this chapter, we will approach the practical part of our project which consists of describing the technicalities of our proposed study, and we will mention the different tools and datasets used in our application.

IV.2 Used tools versions

We developed our software with the tools and APIs located in the table below under the Linux Ubuntu operating system 20.04. The following table shows all the tools and APIs used to build our application.

N°	Tool	Type	Version
1	Pycharm CE	IDE	2020.2
2	Java development kit	plateforme	11.0.8
3	Apache Spark	API	3.0.1
4	Python	Programming language	3.8
5	PyQt5 desinger	GUI designer	5

table 1: tools and APIs

IV.3 Datasets

We have used multiple datasets in our study with characteristics listed down below:

Data-set	Type	N° of classes	N° of rows	N° of features	Source
Iris	classification	3	150	4	Spark data examples
mnist(testing)	classification	10	10000	778	YL98a
mnist	classification	10	60000	778	YL98a
mnist8m	classification	10	101250	782	GL07b
Sample binary classification data	classification	2			Spark data examples

Svmguide1	classification	2	3089	4	CWH03a

table 2: datasets characteristics

Why did we choose the five classifiers in our study?

We have selected the five classifiers because they are the most widely used and are highly suitable for the analysis of large data and we cite the following:

SVM: Over the past decade, SVM has been gradually integrated into the Big Data field. It solves big data classification problems. In particular, it can help multi-domain applications in a big data environment [Demidova et al. 2016].

ANN: Artificial neural networks constitute a realistic criterion in the field of Big Data, knowledge of this field is therefore of paramount importance for those who wish to extract significant information from the large databases available to date [Ossa 2017].

RF: Random Forests seem insensitive to over-fitting,this method generally does not require a lot of parameter optimization efforts. Random forests therefore avoid one of the main pitfalls of Big Data approaches in machine learning [Bei et al. 2018]

LR: logistic regression model gives better result for analyzing the Big data [Dhamod haravadhani 2019].

Naïve Bayes: Finally, a very interesting feature of the Naïve Bayes algorithm is that it is extremely useful for generating synthetic data when the actual data is insufficient. This classifier can also be used in the Big Data field [Prabhat & Khullar 2017].

IV.4 Experimentation

IV.4.1 Multiclass experimentation

In this experiment, we performed three tests (Test1, Test2 and Test3) to test the strength of classifiers regardless of the data size used.

Test 1: presents the classification of the **Small** data set (Iris) using the four classifiers to see which classifiers work well in such cases.

Test 2: presents the classification of the **Medium** data set Mnist(test) and Mnist using the four classifiers to see which classifiers work well in such cases.

Test 3: presents the classification of the **Big** data set (Mnist8m) using the four classifiers to see which classifiers work well in such cases.

We start with the first test which represents the **Small** data set; the results obtained for the four

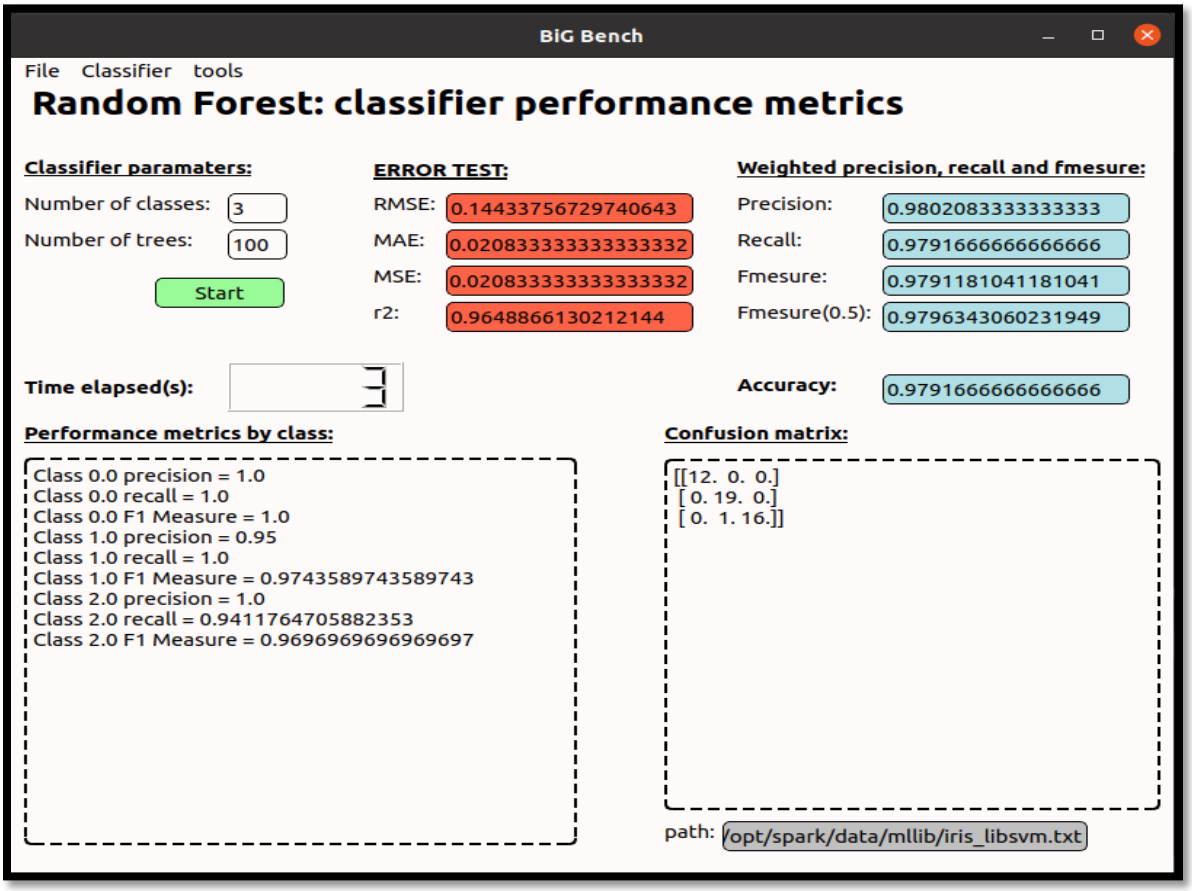


Figure 15 : Random forest classifier performance metrics

classifiers are shown in the figures below:

Chapter IV Realization and Implementation

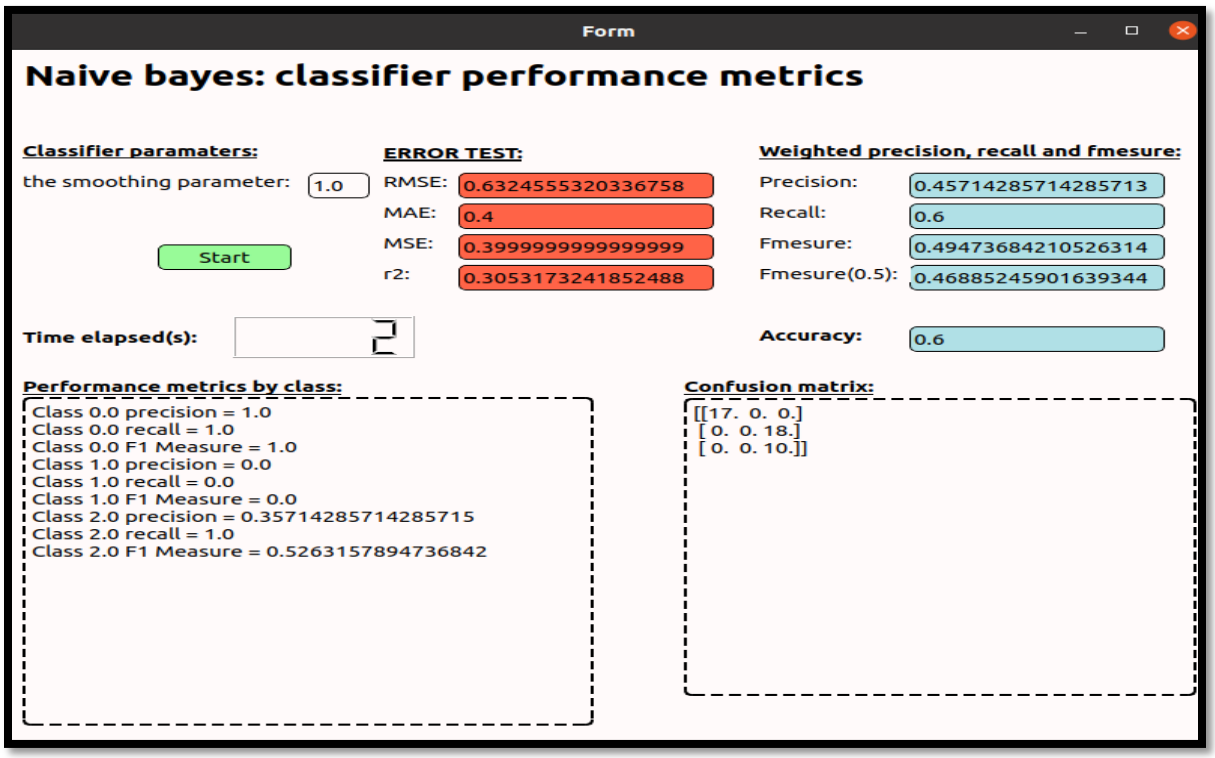


Figure 17: naïve bayes classifier performance metrics

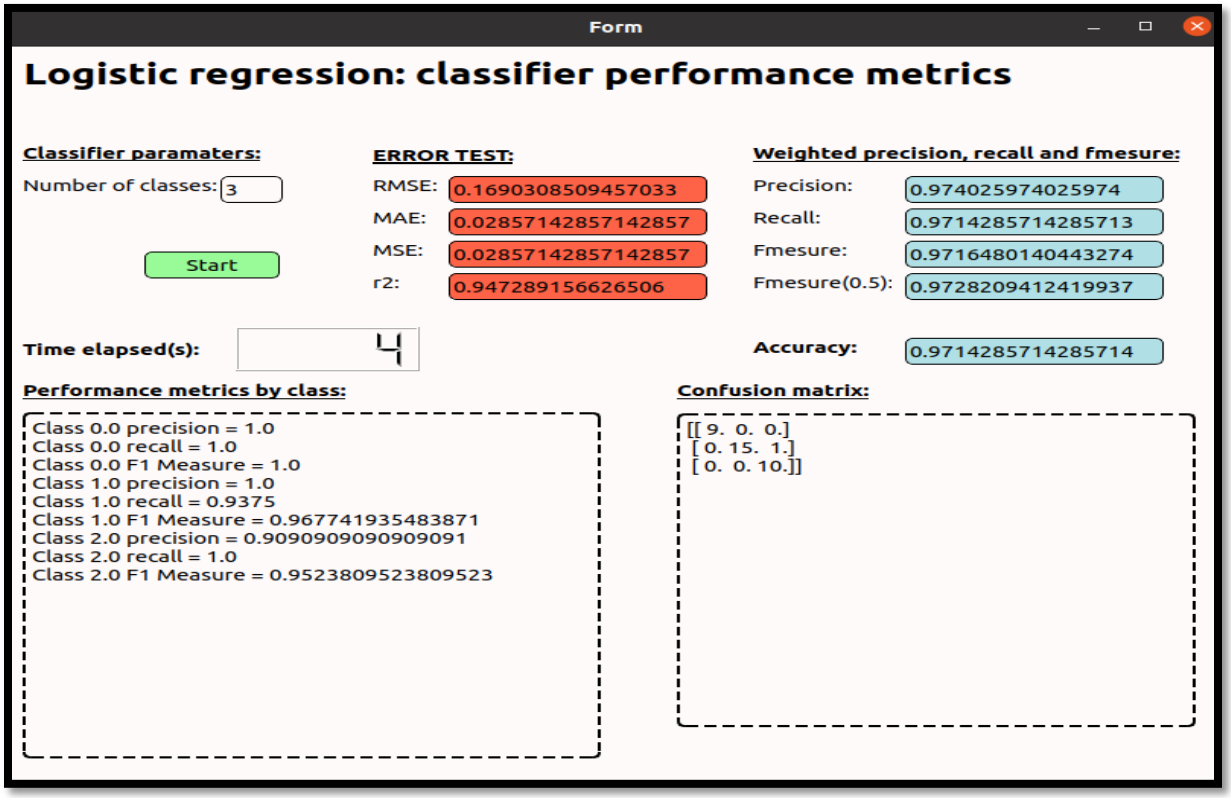


Figure 16: logistic regression classifier performance metrics

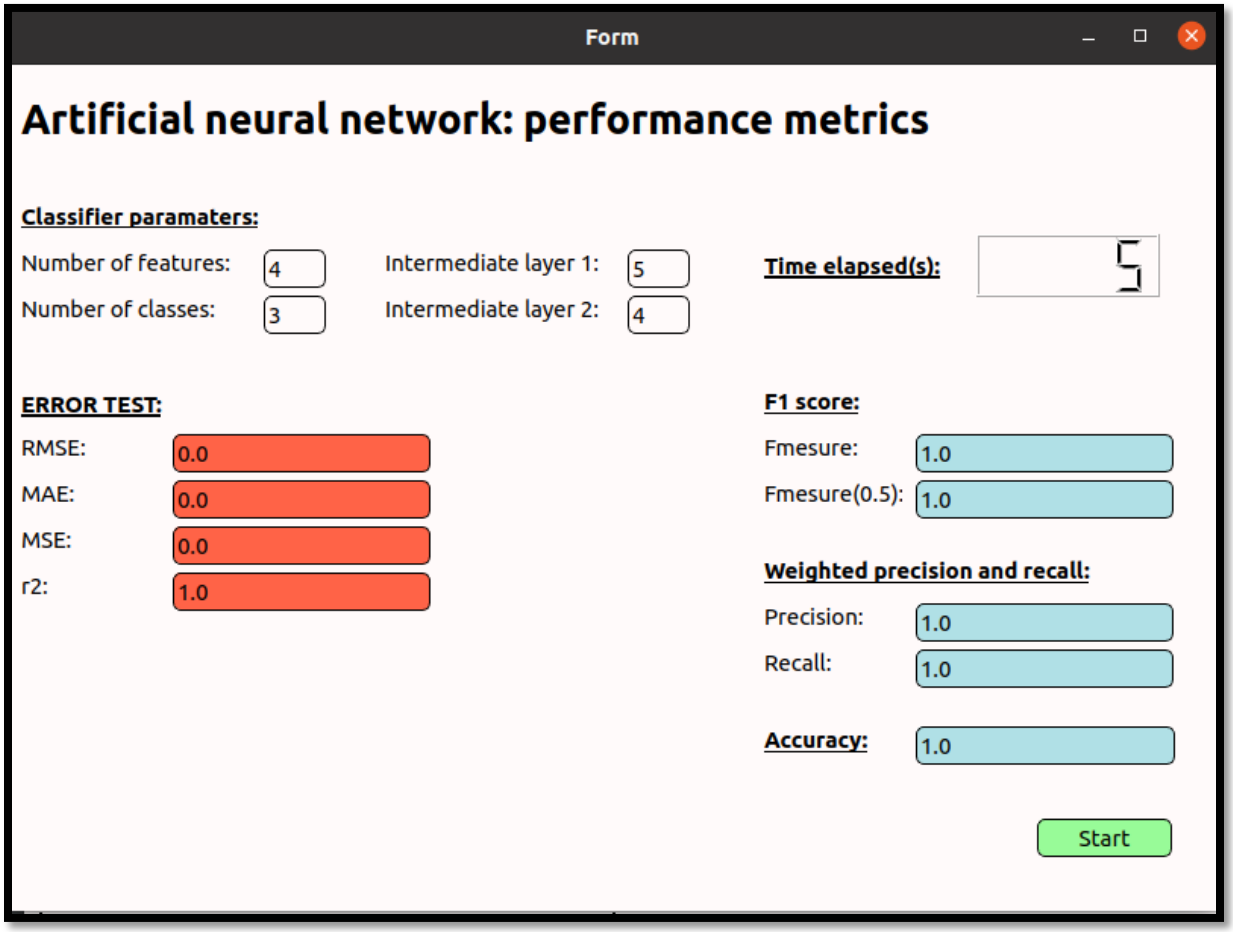


Figure 18:artificial neural network classifier performance metrics

We now go to the second test which represents the **Medium** data set; the results obtained for the four

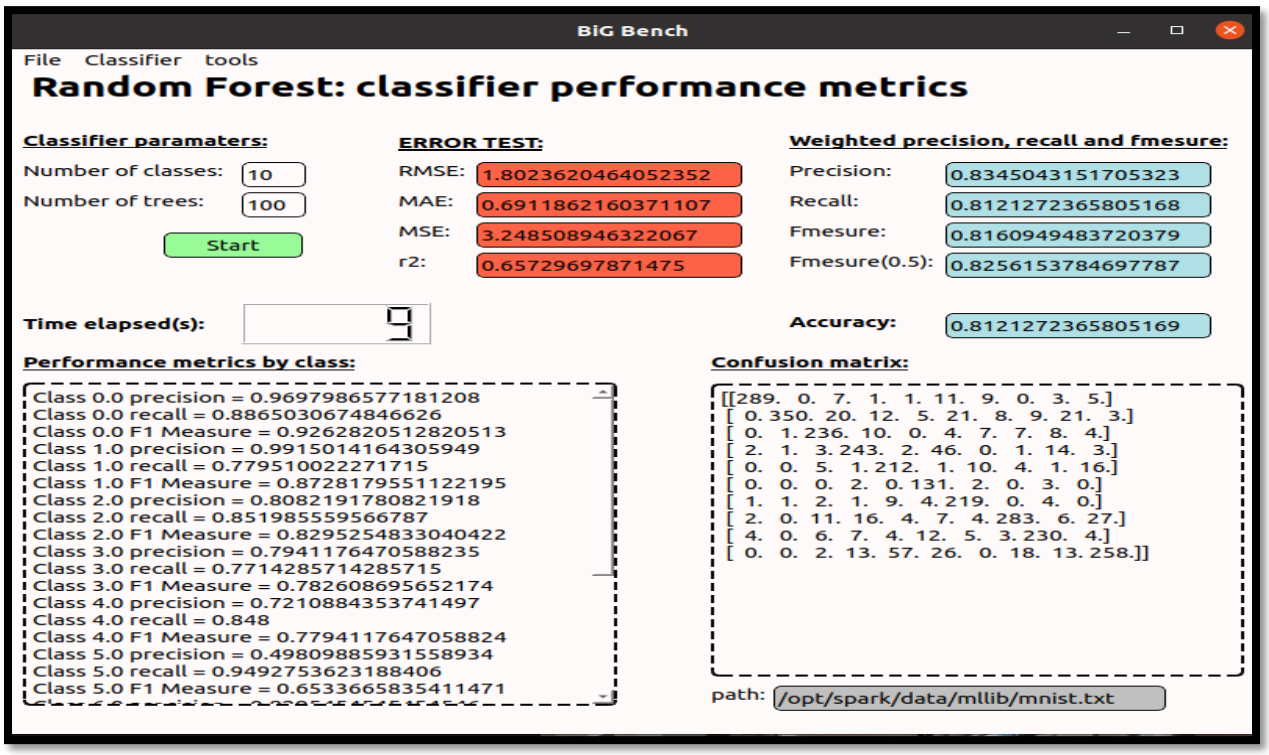


Figure 20: random forest classifier performance metrics

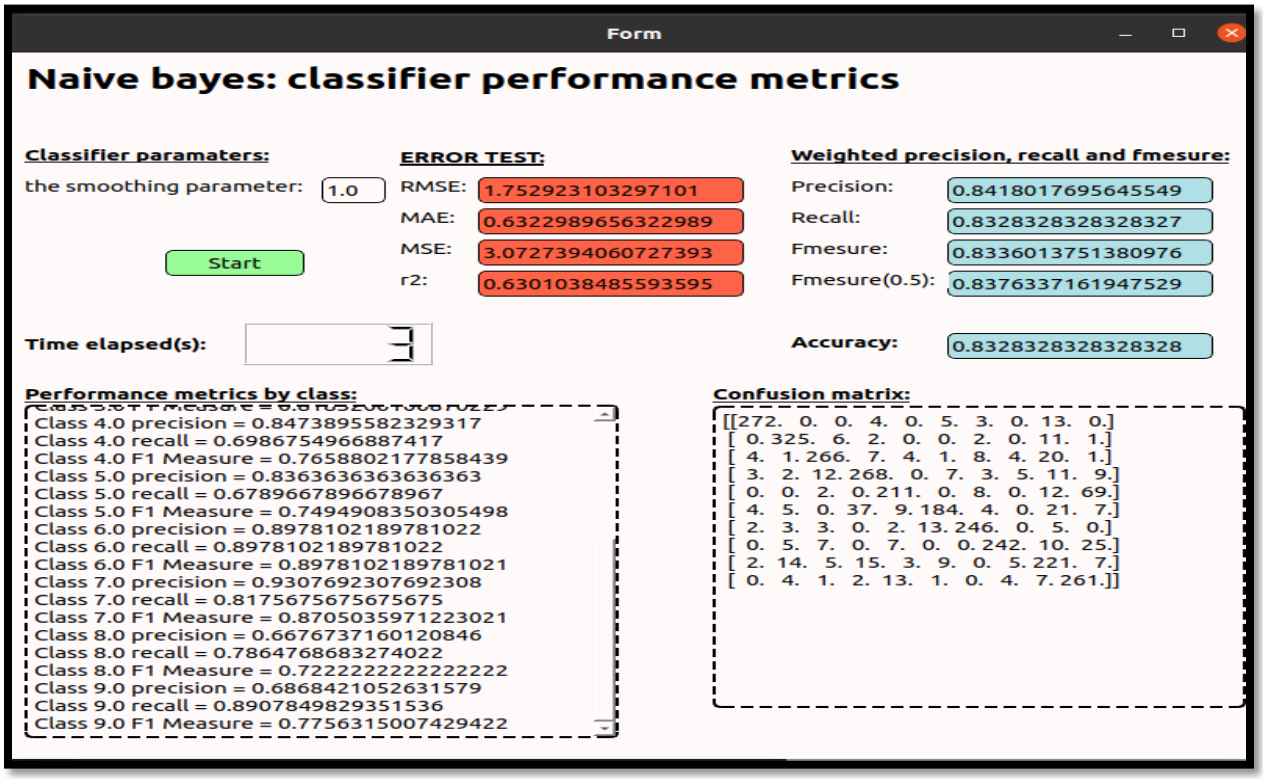


Figure 19: Naïve Bayes classifier performance metrics

classifiers are shown in the figures below:

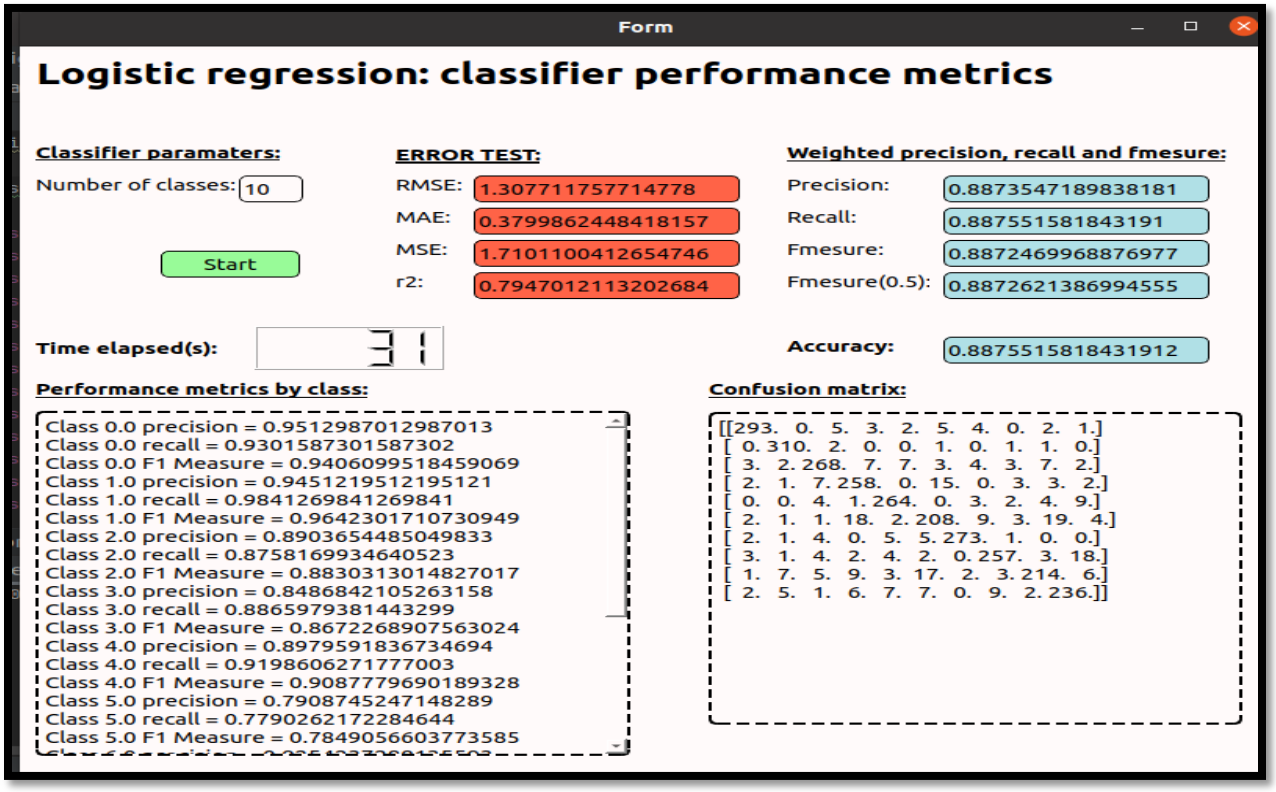


Figure 221: logistic regression classifier performance metrics

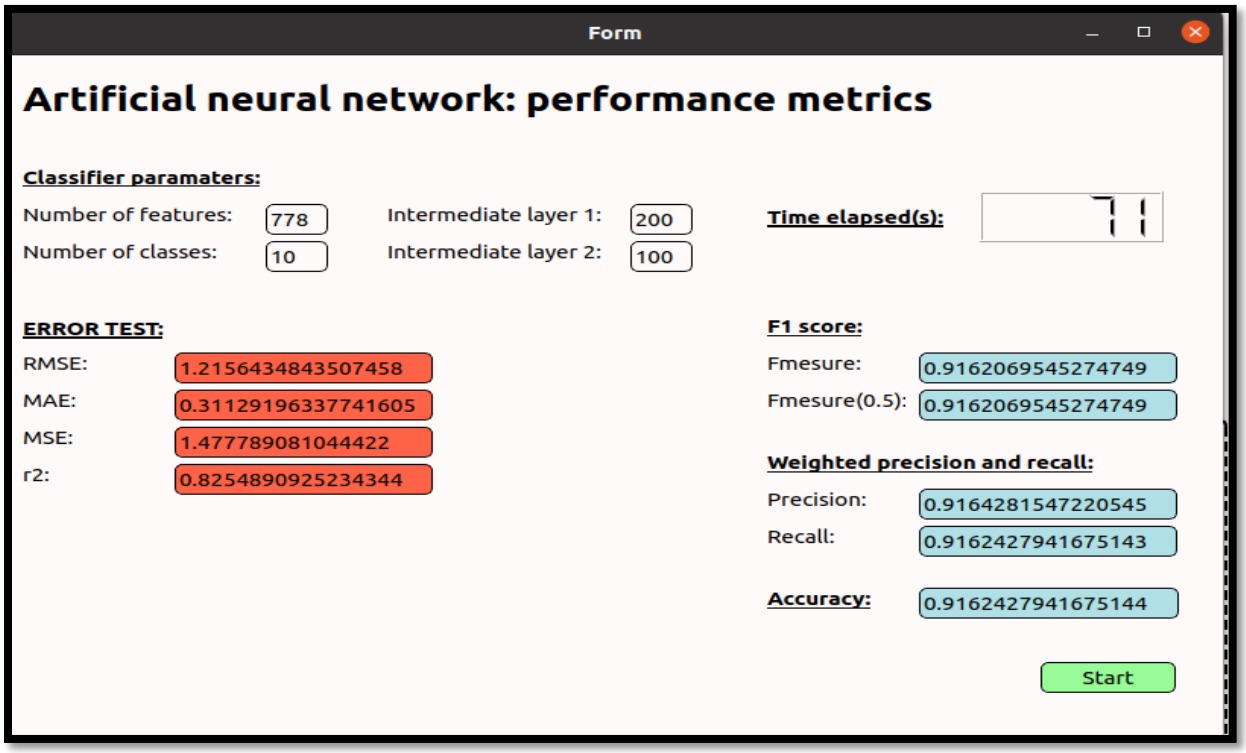


Figure 212:artificial neural network classifier performance metrics

The last test (test 3) that we ran that interests us most is the classification of **Big** data set. The ANN classifier had no result in this test case due to memory saturation, the results obtained for the rest

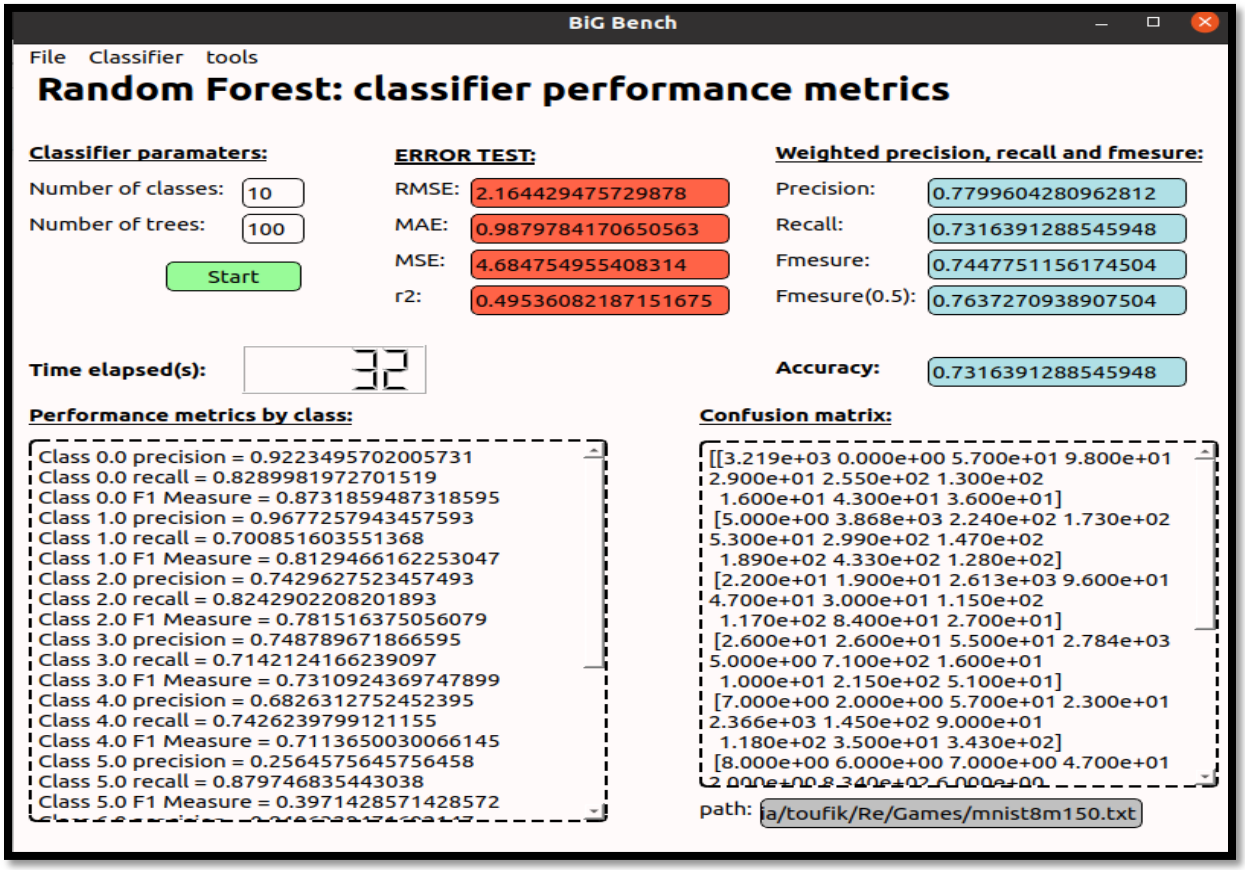


Figure 23:random forest classifier performance metrics

three classifiers are shown in the figures below:

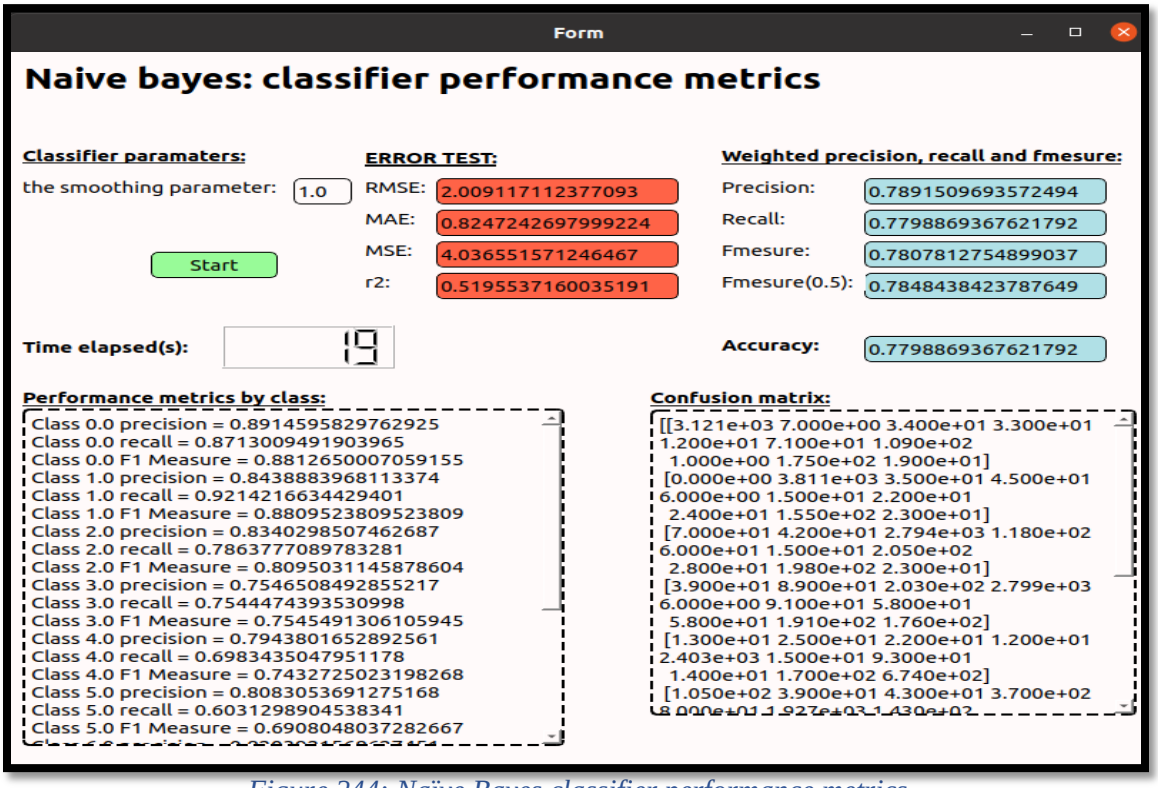


Figure 244: Naïve Bayes classifier performance metrics

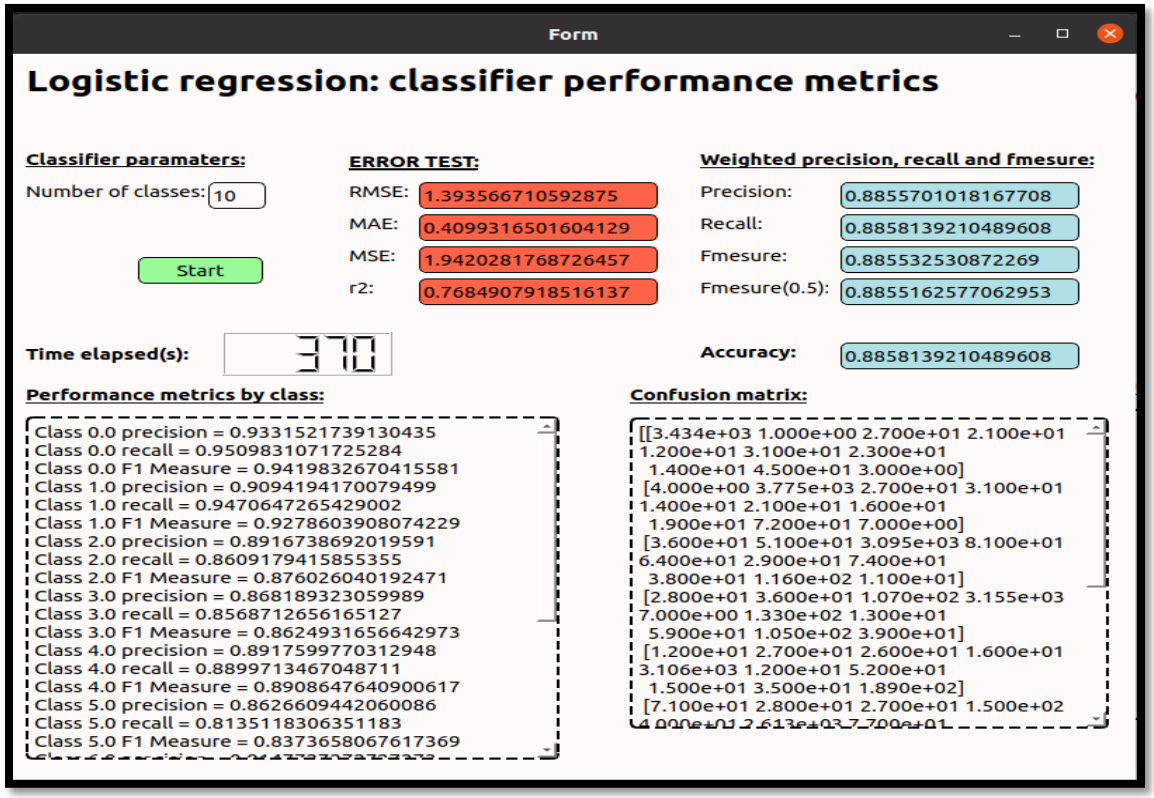


Figure25: Logistic regression classifier performance metrics

IV.3.2 Binary experimentation

In this experiment, we performed two tests (Test1, Test2) to test the strength of classifiers classifier regardless of the data size used.

Test 1: presents the classification of the **Small** data set (svmguide1) using two classifiers (SVM and Logistic regression) to see which classifiers work well in such cases

Test 2: presents the classification of the **Big** data set (kdd12) using two classifiers (SVM and Logistic regression) to see which classifiers work well in such cases

We start with the first test which represents the **small** data set; the results obtained for

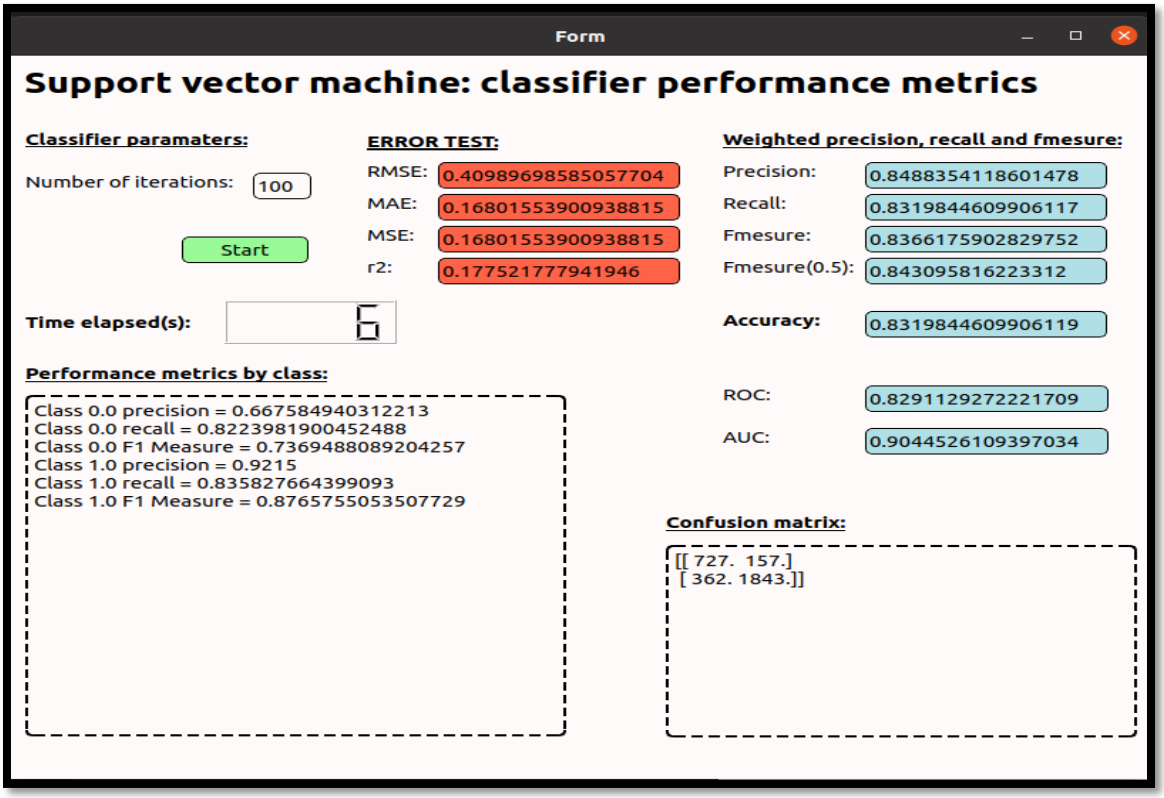


Figure 25: SVM classifier performance metrics

the two classifiers are shown in the figures below:

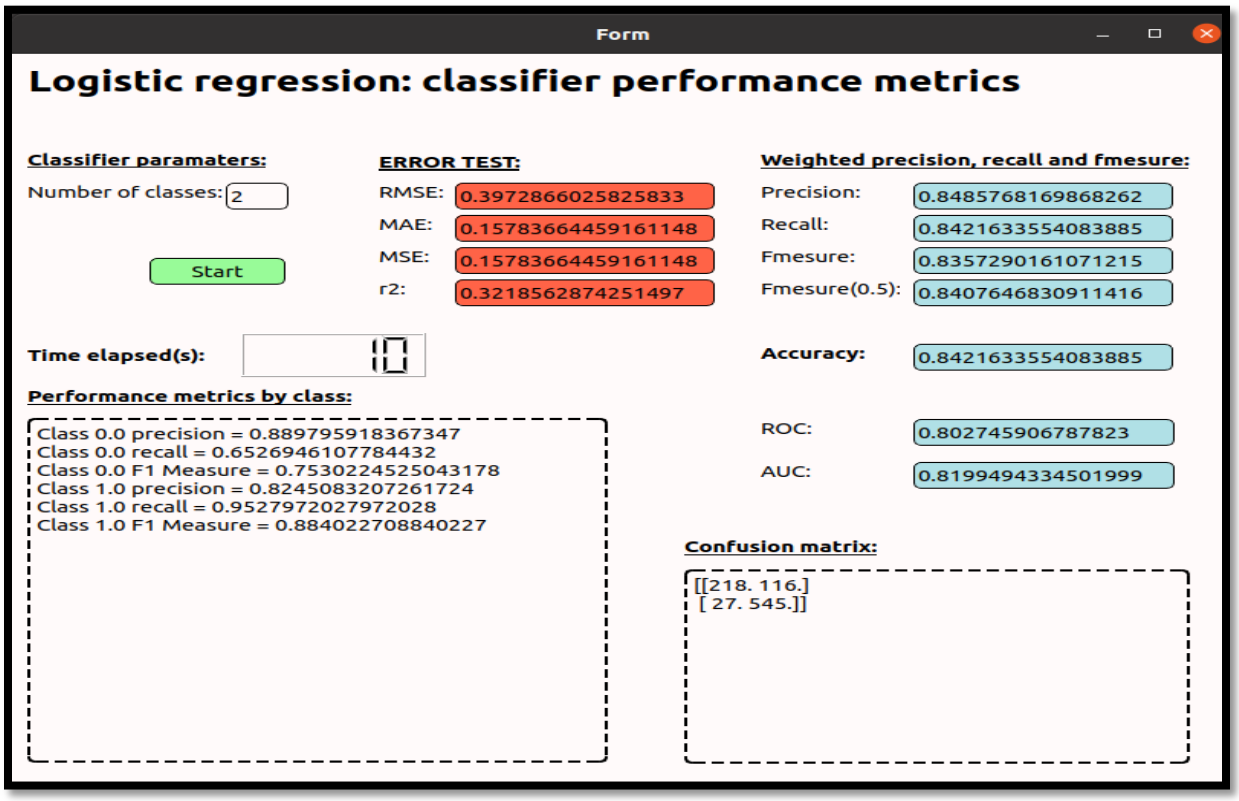


Figure 27: Logistic regression classifier performance metrics

We now go to the second test which represents the **BIG** data set; the results obtained for the two

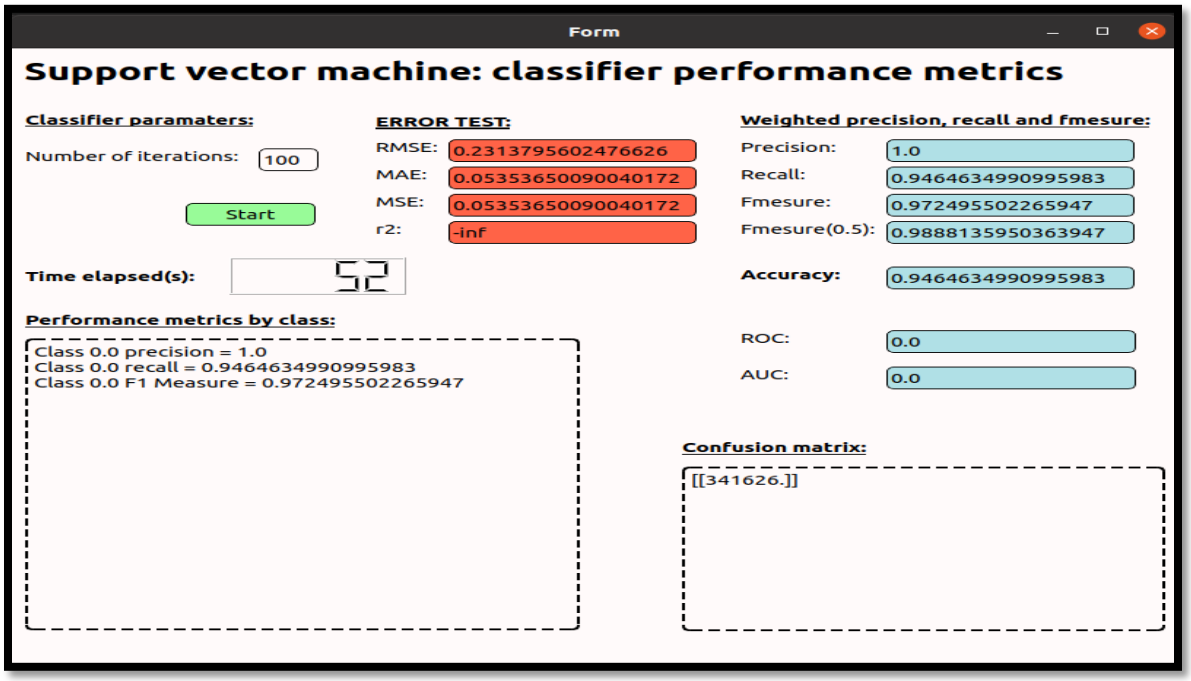


Figure 268: SVM classifier performance metrics

classifiers are shown in the figures below:

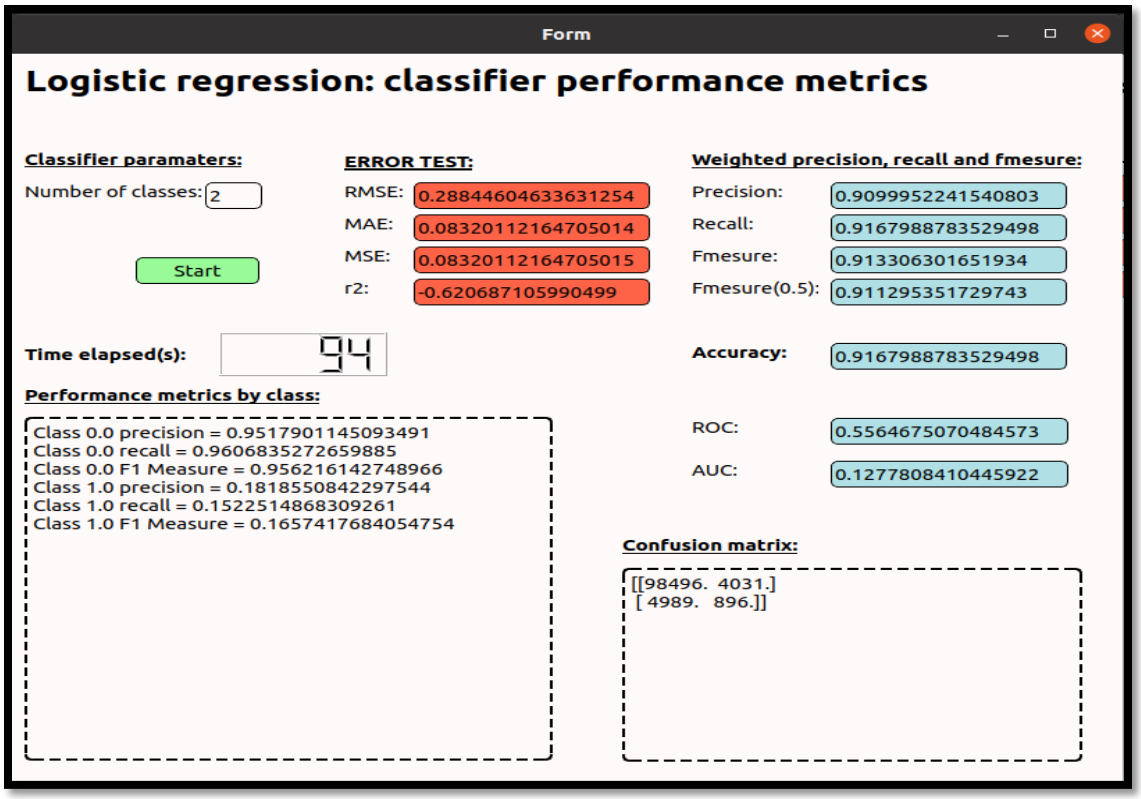


Figure 27: Logistic regression classifier performance metrics

IV.4 Evaluation Metrics

When our algorithms are applied to build machine learning models, we need to evaluate the performance of the model on some criteria, which depends on the application and its requirements. Specific machine learning algorithms fall under broader types of machine learning applications like classification, regression, clustering, etc. Each of these types has well-established metrics for performance evaluation and those metrics that are currently available in spark mllib, definitions of some metrics here:

Root mean-squared error: This measure of error mainly concerns predictors. Square root of the mean squared error: with the same notations as above, it corresponds to:

$$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_1-a_1)^2}{n}}$$

Mean Absolute Error: is a model evaluation metric generally used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set. Each prediction error is the difference between the true value and the predicted value for the instance.

$$\text{mean absolute error} = \frac{|p_1-a_1|+|p_2-a_2|+..+|p_n-a_n|}{n}$$

Mean Squared Error: The mean squared error of a model with respect to a test set is the mean of the squared prediction errors over all instances in the test set. The prediction error is the difference between the true value and the predicted value for an instance.

$$\text{MSE} = \sqrt{\frac{(p_1-a_1)^2+\cdots+(p_1-a_1)^2}{n}}$$

R^2 Error: Coefficient of Determination or R^2 is another metric used for evaluating the performance that helps us to compare our current model with a constant baseline and tells us how much our model is better. The constant baseline is chosen by taking the mean of the data and drawing a line at the mean. R^2 is a scale-free score that implies it doesn't matter whether the values are too large or too small, the R^2 will always be less than or equal to 1.

Accuracy: Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.It works well only if there are equal number of samples belonging to each class.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ predictions\ made}$$

Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$precision = \frac{True\ Positives}{True\ positives + False\ positives}$$

Recall: It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

$$recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Fmesure: F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances) High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model.

$$F1 = 2 . \frac{precison\ .\ recall}{precision + recall}$$

ROC: The Receiver Operator Characteristic curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR (True Positive Rate) against FPR (False Positive Rate) at various threshold values and essentially separates the 'signal' from the 'noise'

AUC: The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

IV.5 Performance metrics result comparison

Our performance comparison is divided into two categories multiclass and binary classification of small, medium and big data set. The tables below represent performance metrics results of all our chosen classifier for each dataset.

IV.5.1 Multiclass classification:

Classifier	Iris dataset								
	RMSE	MAE	MSE	R2	Accuracy	Precision	Recall	Fmesure	Time(s)
Random forest	0.14433	0.0208	0.0208	0.9648	0.97916	0.9802	0.97916	0.97911	3
Naïve bayes	0.2156	0.0465	0.0465	0.9333	0.9534	0.9596	0.9534	0.9534	2
ANN	0.0	0.0	0.0	1.0	1.0	1.0	1.0	/	5

Logistic regression	**0.1690**	**0.0285**	**0.0285**	**0.94**	**0.9714**	**0.9740**	**0.9714**	**0.9716**	**4**

table 3: performance result using Iris dataset

The first conclusion: According to the results obtained by our system, and according to Table 3, we conclude that ANN is a stronger classifier compared to other classifiers in the processing of small data set, it gives very good classification results, but it gives a long time.

Classifier	Mnist(test)								
	RMSE	MAE	MSE	R2	Accuracy	Precision	Recall	Fmesure	Time(s)
Random forest	**1.8023**	**0.6911**	**3.2485**	**0.6572**	**0.8121**	**0.8345**	**0.8121**	**0.8160**	**9**
Naïve bayes	**1.8035**	**0.6645**	**3.2528**	**0.6071**	**0.8253**	**0.8340**	**0.8253**	**0.8262**	**6**
ANN	**1.2156**	**0.3112**	**1.4777**	**0.8254**	**0.91624**	**0.9164**	**0.91622**	**/**	**71**
Logistic regression	**1.3077**	**0.3799**	**0.7101**	**0.7947**	**0.8875**	**0.8873**	**0.8875**	**0.8872**	**31**

table 4: performance result using mnist(test) dataset

Classifier	Mnist								
	RMSE	MAE	MSE	R2	Accuracy	Precision	Recall	Fmesure	Time(s)
Random forest	**1.8265**	**0.7347**	**3.3362**	**0.6193**	**0.7883**	**0.8253**	**0.7883**	**0.7968**	**17**
Naïve bayes	**2.0208**	**0.8297**	**4.0836**	**0.5102**	**0.7780**	**0.7876**	**0.7780**	**0.7789**	**11**
ANN	**0.9726**	**0.1934**	**0.9461**	**0.8864**	**0.9486**	**0.94862**	**0.94864**	**/**	**312**

Logistic regression	**1.1797**	**0.2972**	**1.3918**	**0.8341**	**0.9157**	**0.9155**	**0.9157**	**0.9156**	**180**

table 5: performance result using mnsit dataset.

The second conclusion: According to the results obtained by our system, and according to Table 4 and 5, we conclude that ANN is still a stronger classifier compared to other classifiers in the processing of medium data set, it gives very good classification results, but it gives a long time.

Classifier	Mnist8m								
	RMSE	MAE	MSE	R2	Accuracy	Precision	Recall	Fmesure	Time(s)
Random forest	**2.1644**	**0.9879**	**4.6847**	**0.4953**	**0.7316**	**0.7799**	**0.7316**	**0.7477**	**32**
Naïve bayes	**2.009**	**0.8247**	**4.0365**	**0.5195**	**0.7798**	**0.7891**	**0.7798**	**0.7807**	**19**
ANN	/	/	/	/	/	/	/	/	/
Logistic regression	**1.3935**	**0.4099**	**1.9420**	**0.7684**	**0.88581**	**0.88557**	**0.88551**	**0.88853**	**370**

table 6: performance result using mnist8m dataset.

The third conclusion: According to the results obtained by our system, and according to Table 6, we conclude that logistic regression is a stronger classifier compared to other classifiers in the processing of big data set, it gives very good classification results, but it gives a long time. In this case, ANN was unable to perform classification, but the problem could be solved by powerful machines.

IV.5.2 binary classification:

Classifier	Svmguide1 dataset										
	RMSE	MAE	MSE	R2	Accuracy	Precision	Recall	Fmesure	ROC	AUC	Time(s)

Logistic regression	0.3972	0.1578	0.1578	0.3218	0.8421	0.8485	0.8421	0.8357	0.8027	0.8199	10
SVM	0.4098	0.1680	0.1680	0.1775	0.8319	0.8488	0.8319	0.8366	0.8291	0.9044	6

table 7:performance result using svmguide1 dataset

The fourth conclusion: According to the results obtained by our system, and according to Table 7, we conclude that SVM is a stronger classifier compared to other classifiers in the processing of small data set

Classifier	Kdd12 dataset										
	RMSE	MAE	MSE	R2	Accuracy	Precision	Recall	Fmesure	ROC	AUC	Time(s)
Logistic regression	0.2884	0.0832	0.0832	-0.6206	0.9167	0.9099	0.9167	0.9133	0.5564	0.1277	94
SVM	0.2313	0.0535	0.0535	NA	0.9464	1.0	0.9464	0.9724	0.0	0.0	52

table 8: performance result using kdd12 dataset

The fifth conclusion: According to the results obtained by our system, and according to Table 8, we conclude that Logistic regression is better in binary classification than SVM in the processing of big dataset.

Generally, we conclude the following:

From the results obtained in the implementation part we deduce that:

1. Apache spark distributed in-memory cluster-computing framework is better for tackling massive data.
2. the chosen five classifiers can digest and produce accurate machine learning models, however in an overall manner in respect of time and other metrics, ANN classifier have higher performance for Multiclass classification.
3. when it comes to binary classification of a big dataset, SVM fails to perform properly

Chapter IV Realization and Implementation

IV.6 Conclusion

In this chapter, we presented the comparison implementation of the selected machine learning algorithms under spark framework in purpose of generating valuable insights and accurate models in big data world.

General Conclusion

General Conclusion

In the field of big data mining, data is automatically recorded via digital processing medium with a heterogeneous property created from different data sources or in different contexts. Machine learning algorithms can help extract patterns from the dataset. Over the years, they have had great success due to their ability to deal with problems associated with predictive analytics. But the problem, they are not capable for big data mining. In some cases, the simple logic is: know when a predictive model will not be sufficient to solve a particular problem (in the field of big data mining), especially if all cases are dealt with quickly and efficiently. We have also shown that the field of big data mining needs to use robust machine learning algorithms and big data solutions to develop good predictive models for users.

In this master thesis, we explored how to use machine learning algorithms (Random Forest, SVM, ANN, Logistic regression) in distributed systems (Apache spark) for big data mining.

Sometimes, machine learning algorithms cannot be efficient in dealing with big data, because they often face the problem of over-fitting, even the problem of hardware (GPU, HPC, Clusters ...).

Several avenues of research are emerging and worth exploring, the game-changing machine learning boom of which will be. Machine learning, which plays an important role in big data mining, is expected to flourish in the near future. In addition, we can use prediction techniques or methods, including deep learning.

References chapter I

[Chan 2013] Chan. An architecture for big data analytics. Communications of the IIMA, vol. 13, no. 2, page 1–13, 2013.

[Chen et al. 2014] Chen, Mao et Liu. Big Data : a survey. mobile networks and application, vol. 19, no. 2, page 171–209, 2014.

[Chmidt 2012] Chmidt. Data is exploding : the *3* versus of big data. Bus Comput World, vol. 15, 2012.

[DBTA 2013] DBTA. Big Data Sourcebook. Unisphere Media., 2013.

[Eaton et al. 2011] Eaton, Deutsch Zikopoulos DeRoos et Lapis. Understanding Big Data. McGraw-Hill, USA, 2011.

[EMCES 2015] EMC Education Services EMCES. Data Science and Big Data Analytics. Indianapolis : John Wiley Sons, vol. 978-1-118-87613-8,2015.

[Erl et al. 2016] Erl, Khattak et Buhler. Big Data Fundamentals: Concepts. Prentice Hall Press, Drivers Techniques, 2016.

[Fan et al. 2012] Fan, Gondek Kalyanpur et Ferrucci. knowledge extraction from documents. IBM Journal of Research and Development, vol. 56 (3.4), no. 5, pages 1–10, 2012.

[Ferguson 2013] Mike Ferguson. Enterprise Information Protection- The Impact of Big Data. IBM, 2013.

[Gantz & Reinsel 2012b] Gantz et Reinsel. The digital universe in *2020* : Big data, bigger digital shadows, and biggest growth in the far east. IDC : Analyze the Future, 2012.

[Grinter 2013] Grinter. A big data confession. Interactions, vol. 20, no. 4, page 10–11, 2013.

[Hota & Prabhu 2012] Hota et Prabhu. No problem with Big Data. What do you mean by Big ? Journal of Informatics, page 30–32, 2012.

[Hurwitz et al. 2013] Hurwitz, Halper Nugent et Kaufman. Big Data For Dummies. John Wiley Sons, Inc. Hoboken, vol. NJ 07030-5774,2013.

[Iafrate & Front 2015] Fernando Iafrate et Matter Front. From Big Data to Smart Data. John Wiley Sons., 2015.

[IBM 2014] IBM. The top five ways to get started with big data. 2014.

[Kanimozhi & Venkatesan 2015] Kanimozhi et Venkatesan. Unstructured Data Analysis -A Survey. International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 3, pages ISSN (Online) 2278–1021, 2015.

[Katal et al. 2013] Avita Katal, MohammadWazid et RH Goudar. Big data : issues, challenges, tools and good practices. In In Contemporary Computing (IC3) Sixth International Conference, page 404–409. IEEE,2013.

[Kayyali et al. 2013] Basel Kayyali, David Knott et Steve Van Kuiken. The big-data revolution in us health care : Accelerating value and innovation. Mc Kinsey Company, vol. 2, no. 8, page 1–13, 2013.

[Khan et al. 2018] Nawsher Khan, Habib Shah, Gran Badsha, Aftab Ahmad Abbasi, Mohammed Alsaqer et Soulmaz Salehian. *10* Vs, Issues and Challenges of Big Data. In International Conference on Big Data and Education ICBDE ’18, pages 203–210. March 9–11, 2018, Honolulu, HI, USA, 2018.

[Lyman et al. 2016] Lyman, Strygin Varian Dunn et Swearingen. How much information? Counting-the-Numbers, vol. 6, no. 2, 2016.

[Mills et al. 2012] Mills, Lucas, Irakliotis, Ruppa, Carlson et Perlowitz. Demystifying Big Data: A Practical Guide to Transforming the Business of Government. Washington: TechAmerica Foundation, 2012.

[NIST 2015] NIST. Definitions and Taxonomies Subgroup. National Institute of Standards and Technology, vol. 1, http ://dx.doi.org/10.6028/NIST.SP.1500-1,, 2015.

[Pattnaik & Mishra 2016] Pattnaik et Mishra. Introduction to big data analysis. In : Techniques and Environments for Big Data Analysis, vol. Springer ,doi :10.1007/978-3-319-27520-8, page 1–20, 2016.

[Power 2014] Power. Using 'Big Data' for analytics and decision support. J. Decis. Syst, vol. 23, no. 2, page 222–228, 2014.

[Ripon & Arif 2016] Ripon et Arif. Big Data: The V's of the Game Changer Paradigm. In IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems, volume DOI 10.1109/HPCC-SmartCity-DSS.2016.8. IEEE, 2016.

[Roos et al. 2013] Roos, Deutsch Corrigan Zikopoulos Parasuraman et Giles. Harness the Power of Big Data: The IBM Big Data Platform. New York: McGraw-Hill, 2013.

References chapter II

[Alhadad 2018] Sakinah Alhadad. Visualizing Data to Support Judgement, Inference, and Decision Making in Learning Analytics: Insights from Cognitive Psychology and Visualization Science. The Journal of Learning Analytics, vol. 5, no. 2, pages 60–85, http ://dx.doi.org/10.18608/jla.2018.52.5, 2018.

[Anguera et al. 2018] Anguera, Chacón-Moscoso Portell et Sanduvete-Chaves. Indirect observation in everyday contexts: Concepts and methodological guidelines within a mixed methods framework. Frontiers in Psychology, vol. 9, page Article ID 13. http://dx.doi.org/10.3389/fpsyg.2018.00013, 2018.

[Barapatre & Vijayalakshmi 2017] Darshan Barapatre et Vijayalakshmi. Data preparation on large datasets for data science. Asian Journal of Pharmaceutical and clinical research (AJPCR), pages ISSN: 2455–3891, DOI https ://doi.org/10.22159/ajpcr. 2017.v10s1.20526, 2017.

[Benfield & Szlemko 2006] Benfield et Szlemko. Internet-based data collection: Promises and realities. Journal of Research Practice, vol. 2, no. 2, pages Article D1. Retrieved from, http ://jrp.icaap.org/index.php/jrp/article/view/30/51, 2006.

[Best & Kahn 2003] Best et Kahn. Research in Education. 9th Edition, Prentice-Hall of India Private Limited, New Delhi., 2003.

[Bhaskar & Zulfiqar 2016] Bhaskar et Zulfiqar. Basic statistical tools in research and data analysis. Indian J Anaesth, vol. 60, no. 9, pages 662–669. doi : 10.4103/0019–5049.190623, 2016.

[Bikakis 2018] Nikos Bikakis. Big Data Visualization Tools. ATHENA Research Center, Greece, vol. arXiv :1801.08336v2 [cs.DB], Springer, 2018.

[Davenport & Kim 2013] Davenport et Kim. Keeping Up with the Quants. Harvard Business Review Press, USA, 2013.

[Dipboye 1994] Dipboye. Structured and unstructured selection interviews. Research in Personnel and Human Resources Management, vol. 12, pages 79–123, ISBN : 1–55938–733–5, 1994.

[Dormehl 2014] Dormehl. The Five Best Libraries For Building Data Visualizations. Fast Company, 2014.

[Evans & Lindner 2012] Evans et Lindner. Business Analytics : The Next Frontier for Decision Sciences. Decision Line, vol. 43, no. 2, pages 4–6, 2012.

[Faubert & Wheeldon 2009] Jacqueline Faubert et Johannes Wheeldon. Framing Experience : Concept Maps, Mind Maps, and Data Collection in Qualitative Research. International Journal of Qualitative Methods, vol. 8, no. 3, 2009.

[Guerra et al. 2015] Guerra, Simonini et Vincini. Supporting image search with tag clouds : a preliminary approach. Advances in Multimedia, pages 1–10, https ://doi.org/10.1155/2015/439020, 2015.

[Johnson 2007] Johnson. Technology, Indirect Observation Yield Insights. Marketing News Journal, 2007.

[Kinkeldey et al. 2017] Kinkeldey, Riveiro MacEachren et Schiewe. Evaluating the effect of visually represented geodata uncertainty on decisionmaking : Systematic review, lessons learned, and recommendations. Cartography and Geographic Information Science, vol. 44, no. 1, pages 1–2, doi : 10.1080/15230406.2015.1089792., 2017.

[Kawulich 2005] Kawulich. Participant Observation as a Data Collection Method. FQS, vol. 6, no. 2, page Art. 43, 2005.

[Koppal 2017] Koppal. Trends in Data Visualization. vol. Available at : http ://www.labmanager.com/ask-the-expert/2017/05/trends-indata-visualization.WXsTDISLSpo [Accessed 28 July 2017]., 2017.

[Kothari 2004] Kothari. Research methodology : methods techniques. New Delhi : New Age International (P) Ltd, 2004.

[Kumar & Wu 2007] Kumar et Wu. Survey Paper on Top *10* Algorithms in Data Mining. London : Springer-Verlag Limited, 2007.

[Mila 2018] Steele Mila. The SAGE Handbook of Qualitative Data Collection. Book,ISBN 978-1-4739-5213-3, pages 314–322 and 587–589, 2018.

[Nyumba et al. 2018] Nyumba, Christina Derrick Kerrie Wilson et Nibedita Mukherjee. The use of focus group discussion methodology : Insights from two decades of application in conservation. Methods in Ecology and Evolution, vol. 9, pages 20–32, DOI : 10.1111/2041–

210X.12860, 2018.

[Olshannikova et al. 2015] Ekaterina Olshannikova, Yevgeni Koucheryavy Aleksandr Ometov et Thomas Olsson. Visualizing Big Data with Augmented and virtual reality : challenges and research agenda. Journal of Big Data, vol. 2, no. 22, 2015.

[Padilla et al. 2018] Lace Padilla, Mary Hegarty Sarah Creem-Regehr et Jeanine Stefanucci. Decision making with visualizations : a cognitive framework across disciplines. Cogn Res Princ Implic, vol. doi : 10.1186/s41235-018-0120-9, 2018.

[Paradis et al. 2016] Elise Paradis, Glen Bandiera Bridget O'Brien Laura Nimmon et Maria Athina. Selection of Data Collection Methods. J Grad Med Educ, vol. 8, no. 2, pages 263–264, doi : 10.4300/JGME–D–16–00098.1, 2016.

[Poornima & Pushpalatha 2016] Poornima et Pushpalatha. A journey from big data towards prescriptive analytics. arpn Journal of Engineering and Applied Sciences, vol. 11, no. 19, pages ISSN 1819–6608, 2016.

[Robson 2002] Robson. Real World Research. UK : Blackwell Publishing, vol. (2nd), 2002.

[Schmueli & Koppius 2011] Schmueli et Koppius. Predictive analytics in information systems research. MIS Quarterly, vol. 35, pages 553–572,2011.

References chapter III

[A. Bahgaet al.2019] A. Bahga and V. Madisetti. Big Data Analytics A Hands-On Approach.2019.

[A. Zhang. 2017] A. Zhang, Data Analytics: Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, and Predictive Analysis to Improve Business, Work, and Life. 2017.

[Burhan et al. 2014] Khan Burhan, Rashidah et Hunain. Critical Insight for MapReduce Optimization in Hadoop. International J of Computer Science and Control Engineering, vol. 2, no. 1, pages 1–7, 2014.

[Dasgupta & Nath 2016] Ariruna Dasgupta et Asoke Nath. Classification of Machine Learning Algorithms. International Journal of Innovative Research in Advanced Engineering (IJIRAE),ISSN : 2349-2763,vol. 3, no. 3, 2016.

[Gim et al. 2018] Jangwon Gim, Sukhoon Lee et Wonkyun Joo. A Study of Prescriptive Analysis Framework for Human Care Services Based On CKAN Cloud. Hindawi Journal of Sensors, pages Article ID 6167385,https ://doi.org/10.1155/2018/6167385, 2018.

[Kohonen & Simula 1996] Kohonen et Simula. Engineering Applications of the SelfOrganizing Map. In Proceeding of the IEEE, volume 84 of *10*, page 1354 – 1384, 1996.

[Dreyfus 2002] Dreyfus. Richard Bellman on the birth of dynamic programming. MLRG - Winter Term 2, vol. 50, no. 1, page 48–51, 2002.

[Nutini 2017] Julie Nutini. Monte Carlo Methods (Estimators, On-policy/Offpolicy Learning). MLRG - Winter Term 2, 2017.

[Dayan 1999] Peter Dayan. Unsupervised Learning. In Wilson, RA Keil, F, editors. The MIT Encyclopedia of the Cognitive Sciences, 1999.

[Kotsiantis 2007] Kotsiantis. upervised machine learning : A review of classification techniques. Informatica, vol. 31, page 249–268, 2007.

[Chiliang & Wenting 2012] Chiliang et Wenting. Cross-domain representation-learning framework with combination of class-separate and domain-merge objectives. Proceedings of the CDKD 2012 Conference, page 18–25, 2012.

[Caruana et al. 2008] Rich Caruana, Nikos et Ainur. An Empirical Evaluation of Supervised Learning in High Dimensions. Conference on Machine Learning, ACM, 2008.

[J. Dean. 2014] J. Dean. Big Data, Data Mining, and Machine Learning Value Creation for Business Leaders and Practitioners.2014.

[V. Kumar. 2018] Kumar, Vaibhav & L., M. Predictive Analytics: A Review of Trends and Techniques. International Journal of Computer Applications. Vol 182. 31-3710.5120/ijca2018917434. 2018.

[J. D. Kelleher et al. 2015] J. D. Kelleher, B. Mac Namee, and A. D'Arcy. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. Cambridge, Massachusetts: The MIT Press. 2015.

[N. M. F. Qureshi et al. 2005] N. M. F. Qureshi, D. R. Shin, I. Farah, and A. Abbas, PREDICTIVE ANALYSIS OF LOCALITY-AWARE STORAGE- TIER DATA BLOCKS OVER HADOOP. Vol., no. 12, 2005.

[A. Holmes. 2015] A. Holmes. Hadoop in practice: includes 104 techniques. 2nd ed. Shelter Island, NY: Manning, 2015.

[H. Yang. Et al. 2007] H. Yang, A. Dasdan, R.-L. Hsiao, and D. S. Parker. Map-Reduce-Merge: Simplified Relational Data Processing on Large Clusters. 2007.

[S. Sakr et al. 2011] S. Sakr, A. Liu, D. M. Batista, and M. Alomari. A Survey of Large-Scale Data Management Approaches in Cloud Environments. IEE communications surveys and tutorials. Vol 13. 2011.

[S. R. Alapati. 2018]S. R. Alapati. Expert Apache Cassandra Administration. Berkeley, CA: Apress, 2018.

[B. Vaddeman. 2016] B. Vaddeman. Beginning Apache Pig. Berkeley, CA :Apress, 2016.

[D. Du. 2018] D. Du.Apache Hive Essentials. 2018.

[A. Jain. 2017] A. Jain. Mastering apache storm: processing big data streams in real time. Birmingham Mumbai: Packt, 2017.

[H. Karau et al. 2015] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia. Learning Spark. First edition. Beijing; Sebastopol: O'Reilly, 2015.

[A. Burkov. 2019] A. Burkov.The Hundred-Page Machine Learning Book. 2019.

[M. Mohri.2018] M. Mohri, A. Rostamizadeh, and A. Talwalkar.Foundations of machine learning. Second edition. Cambridge, Massachusetts: The MIT Press, 2018.

[J. Mueller et al. 2016] J. Mueller and L. Massaron.Machine learning for dummies. Hoboken, New Jersey: John Wiley & Sons, Inc, 2016.

[S. Shalev-Shwartz et al. 2014] S. Shalev-Shwartz and S. Ben-David. Understanding Machine Learning: from theory to algorithm. 2014.

[A. Géron. 2019] A. Géron. Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow Concepts, Tools, and Techniques to Build Intelligent Systems. 2019.

[D. Graupe. 2013] D. Graupe.Principles of artificial neural networks. 3rd edition. New Jersey: World Scientific. 2013.

[I. Triguero et al. 2019]I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, and F. Herrera. Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data.Wiley Interdiscip. Rev. Data Min. Knowl. Discov.vol. 9. 2019.

[M. Guller. 2015] M. Guller. Big Data Analytics with Spark:A Practitioners Guide to Using Spark for Large Scale Data Analysis. 2015.

[Sudharsan Ravichandiran 2018] S. Ravichandiran. Hands-On Reinforcement Learning with Python Master reinforcement and deep reinforcement learning using OpenAI Gym and TensorFlow. 2018.

[P. Dangeti. 2017] P. Dangeti. Statistics for machine learning: build supervised, unsupervised, and reinforcement learning models using both Python and R. 2017.

Refs chapter IV

[Demidova et al. 2016] Demidova, Nikulchev et Sokolova. Big Data Classification Using the SVM Classifiers with the Modified Particle Swarm Optimization and the SVM Ensembles. International Journal of Advanced Computer Science and Applications, vol. 7, no. 5, 2016.

[Ossa 2017] Chinedu Ossa. Integrated Big Data Analytics Technique for Real-Time Prognostics, Fault Detection and Identification for Complex Systems. Infrastructures, vol. 2, no. 20, page doi :10.3390/infrastructures2040020,2017.

[Bei et al. 2018] Zhendong Bei, Chuntao Jiang Chengzhong Xu Zhibin Yu Ni Luo et Shengzhong Feng. Configuring in-memory cluster computing using random forest. Future
Generation Computer Systems, vol. 79, pages 1–15, http://dx.doi.org/10.1016/j.future.2017.08.011, 2018.

Dhamodharavadhani S. and Rathipriya R., Enhanced-Logistic-Regression-(ELR)-Model-for-Big-Data, DOI: 10.4018/978-1-7998-0106-1.ch008,IGI Global,2019

[Prabhat & Khullar 2017] Anjuman Prabhat et Vikas Khullar. Sentiment classification on big data using Naïve bayes and logistic regression. International Conference on Computer Communication and Informatics, page doi : 10.1109/ICCCI.2017.8117734, 2017.