EAST WEST UNIVERSITY

Study on Big Data Analytics: Breast Cancer Perspective

By

Md. Asif Newaz
2015-2-50-006

Jannat Laila Karim
2015-2-50-009

Hemal Majumder
2015-2-50-012

Supervised by
Dr. Anup Kumar Paul
Assistant professor
Department of Electronics and Communications Engineering
East West University.

This Project submitted in partial fulfilment of the Requirement for the Degree of
Bachelors of Science in B.Sc. in Information and Communications Engineering

To the

Department of Electronics and Communications Engineering
East West University
Dhaka, Bangladesh

Declaration

We, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by us under the supervision of Dr. Anup Kumar Paul, Assistant Professor, Department of Electronics and Communications Engineering, East West University. We also declare that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

……..

Dr. Anup Kumar Paul
Supervisor,
Assistant Professor,
Department of ECE,
East West University.

……..

Md. Asif Newaz
2015-2-50-006
Department of ECE
East West University.

……..

Jannat Laila Karim
2015-2-50-009
Department of ECE
East West University.

…….

Hemal Majumder
2015-2-50-012
Department of ECE
East West University.

Acknowledgements

First we are grateful to almighty Allah for blessing us with such opportunity of learning and ability to successfully complete the task.

A special thanks with honour to our supervisor Dr. Anup Kumar Paul, Assistant Professor, Department of Electronics and Communications Engineering, East West University, who was kind enough to allocate his valuable time provide us with is humble guidance motivating thought and encouragement.

……..

Md.Asif Newaz
2015-2-50-006
Department of ECE
East West University

……..

Jannat Laila Karim
2015-2-50-009
Department of ECE
East West University

……..

Hemal Majumder
2015-2-50-012
Department of ECE
East West University

Acceptance

This research report presented to the department of Electronics and Communications Engineering. East West University submitted to partial fulfilment to the requirement for the degree of B.Sc. in Information and Communications Engineering under complete supervision of the undersigned.

……...

Dr. Anup Kumar Paul
Supervisor,
Assistant professor,
Department of Electronics and Communications Engineering,
East West University.

Abstract

In these paper, recent developments on the impact of newly developed Big Data as on manufacturing information systems is especially discussed. There is a misconception that Big Data only used for business. It's not true. Big Data also has social and medical benefits. We consider breast cancer data for know about the death reasons on breast cancer. Second highest cause of cancer death among Women in the worldwide is breast cancer. We collect data of breast cancer patients. This data has volume, variety, velocity and veracity. From this 4V we generated the 5^{th} V (value). Big knowledge may be a new driver of the globe economic and social group changes. Though huge data is wide mentioned in theoretical manners, there's a deficiency in publications and sources dedicated to its sensible usage. We want to storage these large amounts of information for our future prediction. If we have a tendency to lose our knowledge we have a tendency to lost our necessary data. The information complexities are increasing together with data's volume, variety, rate and truthfulness, the important impact hinges on our ability to uncover the `value' within the knowledge through huge knowledge Analytics technologies. Big Data Analytics poses a grand challenge on the look of extremely scalable algorithms and systems to integrate the information and uncover massive hidden values from datasets that are various, complex, and of an enormous scale. Huge knowledge analytics (BDA) has been known as an essential technology to support knowledge acquisition, storage, and analytics in knowledge management systems in trendy manufacturing.

Keywords: Big Data, data analysis, python, jupyterlab.

Table of Content

List of Figures

List of Tables

Chapter 1: Introduction

1.1. Motivation

Cancer has been identify as most deadly disease in the world. Every year millions of people affected and died on cancer. Breast Cancer is a serious matter for developing and underdeveloped countries[1]. Though developed countries are also not free from it but breast cancer is a matter of concern there. In US, breast cancer is the 2nd highest cancer death rate. If we know the death factors of breast cancer then we can take steps to reduce the death rate of breast cancer. Big Data can help us to find out the real death factors of breast cancer.

1.2. Goal to achieve

From a data set, by using some tools of big data, we will try to find out the death risk factors of big data. We will consider gender, age, ER and PR status, Tumor, Node and Metastasis stages. It may be help doctors to create a proper treatment plan.

Chapter 2: Theoretical Framework

The objective of this theoretical framework is to provide a foundation for this thesis. In section -

2.1. An overview will be given about history of big data

2.2. An overview about big data

2.3. How data can be analyzed

2.4. Python is better than other language in data analysis

2.5. Data analysis is used for business development

2.6. Decision Making

2.1. Literature

In 1663, John Graunt was dealing with "overwhelming amounts of information". He studied the bubonic plague, which was currently damaging Europe. Graunt was the first person who dealt with the statistical analysis of data. The field of statistics expanded to include collecting and analyzing data in 1800s.

The growth of Big Data includes a number of fundamental steps for its foundation, and while looking back to 1663 isn't necessary for the expansion of data volumes today. Big Data to Amazon or Google is different than Big Data for medium-sized insurance organization[2].

Such basic steps to the modern thought of Big Data include the improvement of computers, smart phones, the internet, and sensory (Internet of Things) device to provide data. Credit cards also played a vital role, by providing vast amounts of data, and surely social media changed the nature of data volumes .The enlargement of modern technology is interlacing with the evolution of Big Data. 90% of the available data has been generated in the last two years. The term Big Data has been all over from 2005, when it was first introduced by O'Reilly in 2005. However, the use of Big Data and the urgency to figure out all available data has been around much longer in process. In fact, the earlier records of using data to recode and manage businesses date back

from 7.000 years ago when accounting was introduced first in Mesopotamia in order to note the growth of crops and herds. Accounting principles continued to raise, and in 1663, John Graunt examined all data about mortality roles in London. He wanted to gain knowledge and build a cautioning system for the ongoing bubonic plague. In the first recorded data of statistical data analysis, he accumulated his findings in the book Natural and Political.

2.2. Big Data

Present days the internet is being extensively used than it was used a few years back. It has become a basic element of our life. Billions of people are using social media and social networking sites every day all across the planet. Such a huge number of people generate a huge amount of data which have become really complex to manage[3].

So, what is this term called?
Yes, Big Data.

Big Data is the term which refers to this massive amount of data. The concept of big data is fast growing its weapons all over the world. It is a burning topic for thesis, project, research, and dissertation.

Bernard Marr characterize Big Data as the digital record that we are developing in this digital stage .This digital recoded data is made up from all the data that is stored when we are using digital technology. The basic concept behind the phrase Big Data is that everything we do is raising data which we can use and analyses to become smarter[3].

The research firm Gartner, defines Big Data:
Big Data is high-volume, high-velocity, and/or high-variety information resources that consider profitable, new forms of information processing that enable enhanced observation, result and process automation[3].

Ernst and Young offer this definition:
Big Data is the productive, large and distinct volumes of data being generated by people, tools and devices. It need new, inventive, and extensible technology to gather ,host and analytically process this huge amount of data stored in order to derive real-time business observation that connected to consumers loss, profit, performance, productiveness and increase shareholder value.

Big knowledge comes in 3 forms. Structured, unstructured, and semi-structured.

Structured knowledge is knowledge that's organized, labelled, and features a strict model that it follows.

Unstructured knowledge is claimed to create up concerning 80%of knowledge within the world, wherever the information is usually in a text type and doesn't have a predefined model or is organized in any manner.

Semi-structured knowledge could be a combination of the 2.It is kind of like structured knowledge, where it should have associate organized structure, but lacks a strictly-defined model. Some sources of structured massive Data are relative databases and spreadsheets.

With this sort of structure, we all know however knowledge is associated with different data, what the information means that, and the knowledge is straightforward to question, using an artificial language like SQL .Some sources of semi-structured massive Data are XML and JSON files. These sources use tags or different markers to enforce hierarchies of records and fields at intervals knowledge. A large multi-radio telescope project called sq. km Array, or SKA, produced concerning a thousand pet bytes, in 2011 a minimum of, of data every day. It is projected that it'll turn out about 20,000 pet bytes or twenty billion gigabytes of knowledge every day in 2020.Currently, there's associate explosion of data coming from web activity and particularly, video production and consumptions well as social media activities. These numbers

can simply keep growing as web speeds increase and as additional and additional people all over the globe have access to the web. Structured knowledge refers to any data that resides during a fastened field at intervals a record or file. It has the advantage of being simply entered, stored, queried, and analyzed. In today's business setting, most massive Data generated by organizations is structured and keep in knowledge warehouses. Highly structured business-generated knowledge is thought-about a valuable supply of information and so equally necessary as machine and people-generated data.

2.2.1. V's of Big Data

Velocity: The speed of the data is known as Velocity, or it refers to rate at which the data is generating. Velocity is that the concept knowledge is being generated extraordinarily quick, a method that never stops. Its characteristics include near or real-time streaming and local and cloud-based technologies, which can process the information very fast. Hours of footage are uploading to YouTube in Every 60 seconds. This bulk amount of data is produced each minute. So give some thought to what quantity accumulates over hours, days, and in years[4].

Volume: It is the scale from which we can analyses the data measurement or the increasing amount of collected data .Volume is the amount of developed data. Every day we are generating around 2.6 quintillion bytes of data... Approximately the world population is 7 billion and the majority of them are using digital devices. These devices all generate, capture, and store data. Mobiles, desktop computers, laptops are generating vast amount of data. The number of data can be expressed in Exabyte's, zettabytes, yottabytes, etc[3]. For example, Facebook gather images, videos. That statement doesn't begin to amaze you until you get to know that Facebook has more users than China has people. Each of those users stored a lot of images. Facebook is storing approximately 250 billion images. Try to cloak your head around 250 billion pictures. Try this one. As far in 2016, Facebook had 2.5 trillion posts posted. Seriously, that's a huge in number, so big it's pretty much impossible to figure it out. As we are moving ahead, we're going to have

huge collection of newer data. For example, as we added sensors to everything, that produces data .Which will add up to this. Consider how much data is generating .Suppose, one have a temperature sensor in his garage. Even with a one-minute level of granularity (one measurement a minute), that's still 525,950 data in a year, and that's just in one sensor. Now think you have a factory of thousand sensors, you're can see half a billion data, just for the temperature only. Then, of course, there are all the inner enterprise collections of knowledge, starting from energy business to attention to national security. All of those industries are generating and capturing large amounts of knowledge. That's the volume.

Variety: You have noticed that we have discussed about photos, sensor data, tweets, encrypted packet data and many other different data. Each of data is very different from each other. These data's are different from their application to application, and most of them are unstructured. It means that it isn't easy to fit into fields like in a database application or any other similar section[3]. For example, email messages can be considered. In a legal process it might requires thousands to millions of email messages in a collection. One of those messages is not going to be similar like another. Each message will consist an email address which indicates the sender, a right destination and a time. Each message will contain human-written text and may be attached some document. We can get unstructured data from photos, videos, recordings , email messages , documents and books, presentations, tweets , ECG strips but all of them are incredibly varied. All of the data assortment makes up the variety of big data.

Veracity: Veracity means the allegiance to facts and efficiency. With the large amount of data available, the debate rages on about the accuracy of data in the digital era. Is the information real, or is it false? Quality and origin of data is called Veracity. 80% of data is considered to be unstructured and we must arrange them to produce reliable, steady and proper for understanding. The data must be classified, analyzed and envisioned. Characteristics of veracity might be included ambiguity, consistency, completeness, integrity and drivers include cost and traceable.

Value: Last V is value. This V is has the ability to turn data into value. Value doesn't only mean

just profit. It may be medical or social gain, or customer, employee, or personal comfort. The main case of why people spend time to understanding Big Data is to collect value from it. And one will gain knowledge how to gain value from it.

2.3. Data Analysis

For one to become knowledgeable knowledge human, operating in data processing and business intelligence companies you've got to know the basics of information analytics. We study about analytics, common terminologies employed in analytics, tools and basic prerequisites for analytics and also the progress of information analytics. While not additional fuss, let's dive in to explore the fundamentals of information analytics[5].

2.3.1. Types of analytics:

Raw data isn't any totally different from rock oil. These days, individual or establishment with a moderate budget will collect giant volumes of information. However the gathering in itself shouldn't be the top goal. Organizations that may extract meanings from the collected information are those that may vie in today's advanced and unpredictable setting. At the core of any knowledge refinement method sits what's ordinarily named as "analytics". However completely different folks use the word "analytics" to imply various things. If you're in selling and would really like to know knowledge analytics, you ought to perceive the various styles of analytics. Below are samples of analytics:

- Descriptive
- Diagnostic
- Prescriptive
- Exploratory
- Predictive
- Mechanistic

- Casual
- Inferential

Let’s go to some further explanation.

1. Descriptive analytics: The main target of descriptive analytics is to summarize what is happening in an institution. Descriptive Analytics examines the data or content — that is manually performed — to answer queries such as:
What happened?
What’s happening?

Descriptive analytics is characterised by standard business intelligence and visualizations like the bar charts, pie charts, line graphs, or the generated narratives. A straightforward illustration of descriptive analytics will be assessing credit risk in a very bank. In such a case, past money performance will be done to predict client’s possible money performance. Descriptive analytics is beneficial in providing insights into sales cycle like categorizing customers supported their preferences.

2. Diagnostic analytics: As the name suggests, diagnostic analytics is employed to unearth or to work out why one thing happened. As an example, if you’re conducting a social media selling campaign, you'll have an interest in assessing the quantity of likes, reviews, mentions, followers or fans. Diagnostic analytics will facilitate your distil thousands of mentions into one read in order that you'll build progress along with your campaign.

3. Prescriptive analytics: While most information analytics provides general insights on the topic, prescriptive analytics offers you with a “laser-like” focus to answer precise queries. For example, within the care business, you'll use prescriptive analytics to manage the patient population by activity the quantity of patients World Health Organization are clinically fat. Prescriptive analytics will permit you to feature filters in avoirdupois like avoirdupois with polygenic

disorder and sterol levels to search out areas wherever treatment ought to be targeted.

4. Exploratory analytics: Exploratory associate degree a lyrics is an analytical approach that primarily focuses on characteristic general patterns within the information to spot outliers and options which may not are anticipated mistreatment different analytical varieties. For you to use this approach, you have got to grasp wherever the outliers are occurring and the way different environmental variables are associated with creating enlightened selections. For example, in biological watching of information, sites is littered with many stressors, therefore, agent correlations are important before you try to relate the agent variables and biological response variables. The scatter plots and correlation coefficients will offer you with perceptive info on the relationships between the variables. However, once analysing completely different variables, the fundamental ways of variable image are necessary to produce larger insights.

5. Predictive analytics: Predictive analytics is that the use of information, machine learning techniques, and applied mathematics algorithms to see the chance of future results supported historical knowledge. The first goal of prophetic analytics is to assist you transcend simply what went on and supply the simplest potential assessment of what's doubtless to happen in future. Predictive models use recognizable results to make a model which will predict values for various kind of knowledge or perhaps new knowledge. Modelling of the results is critical as a result of it provides predictions that represent the chance of the target variable — such as revenue — based on the calculable significance from a collection of input variables. Classification and regression models are the foremost in style models utilized in prophetical analytics. Predictive analytics are often utilized in banking systems to discover fraud cases, live the degree of credit risks, and maximize the cross-sell and up-sell opportunities in a corporation. This helps to retain valuable shoppers to your business.

6. Mechanistic analytics: As the name suggests, mechanistic to know clear alterations in procedures or perhaps variables which might lead to ever-changing of variables. The results of mechanistic analytics are determined by equations in engineering and physical sciences. Also,

they permit information scientists to work out the parameters if they grasp the equation.

7. Causal analytics: Causal analytics enable huge knowledge scientists to work out what's doubtless to happen if; one part of the variable is modified. Once you use this approach, you must depend on variety of random variables to see what's doubtless to happen next while you'll use non-random studies to infer from causations. This approach to analytics is suitable if you're coping with massive volumes of knowledge.

8. Inferential analytics: This approach to analytics takes completely different theories on the planet under consideration to see the sure aspects of the massive population. After you use inferential analytics, you'll be needed to require a smaller sample of knowledge from the population and use that as a basis to infer parameters concerning the larger population.

2.4. Python

Python is a more and more in style tool for knowledge analysis. In recent years, variety of libraries have reached maturity, permitting R and Stata users to require advantage of the wonder, flexibility, and performance of Python while not sacrificing the practicality these older programs have accumulated over the years.

The best reason to be told Python is additionally the toughest to articulate to somebody UN agency is simply setting out to work with Python: in terms of structure and syntax, it's a superbly designed, intuitive, however passing powerful general artificial language. Python was expressly designed (a) therefore code written in Python would be simple for humans to scan, and (b) to reduce the quantity of your time needed to write down code. Indeed, its simple use is that the reason that consistent with a recent study, eightieth of the highest ten cesium programs within the country use Python in their intro to engineering categories. At the identical time, however, it's a true, all-purpose artificial language. Major firms like Google and Dropbox use Python in their core applications.

This sets Python other than “Domain Specific Languages” languages like R that are extremely tuned to serve solely a selected purpose – like statistics – and work for a selected audience. John Chambers created R with the goal of creating a language that non-programmers may start with quickly, however that may even be utilized by “power users”. To an oversized degree he succeeded, as is proven by R’s uptake. However in attempting to form the language thus accessible to non-programmers, several compromises were created within the language. R solely very serves one purpose – applied mathematics analysis – and also the language syntax has all styles of oddities and warts that return from this original cut price. Python will need a bit a lot of coaching to urge started with (though not that a lot of more), however as a result there’s no ceiling to what you'll do with Python. If you learn Python, you’re learning a full artificial language. This suggests if you ever must add a distinct language like Java or C for a few reason, perceive code some other person has written, or otherwise cope with a programming downside, your background during a real artificial language can provide you with a decent abstract foundation for no matter you bump into. Indeed, this is often the explanation prime Cs programs teach in Python. Of all the explanations to settle on Python[6], I believe this can be out and away the foremost compelling. Python sets you up to grasp and operate within the broader programming world. And if you’re in the least inquisitive about doing process scientific discipline, building a generalizable programming ability simply causes you to additional versatile. R is nice if you wish to only run regressions or do things that completely match the mildew somebody has created with Associate in Nursing R operate. However as social scientists keep finding new sources of information (like text) and new ways in which to research it, the additional literate you're generally programming, the additional ready you the additional ready you'll be to steal tools from alternative disciplines and to write down new tools yourself.

2.5. Business intelligence

A vital part for the success of a contemporary organization is its ability to require advantage of all obtainable information” - Cody et al. (2002). In fact, the flexibility to collect and timely

remodel of all data in effective business data isn't solely essential to succeed, however conjointly necessary to survive (Lönnqvist & Pirttimäki, 2006). For instance, a casino may gather data of a special event or the usage of a coin machine to trace the preferences of a client or the chance of assorted games and shut unpopular, unprofitable, or unknown games quickly (Watson & Wixom, 2007). However, the challenge to rework all this data to effective business data becomes tougher because the data keeps growing exponentially and also the increasing quantity of workers UN agency want access to the current data Organizations deploy, to support these information savvy staff, information warehouses and frontend applications that may access, analyses, summarize and visualize all on the market info (Rivest et al., 2005). For instance, organizations produce frontend applications with a visible dashboard that enable a call maker to trace key performance indicators of their operations (Chaudhuri et al., 2011). These frontend applications that organizations are making and deploying also are illustrious on the market as "Business Intelligence" applications (Rivest et al., 2005). Several organizations use these Business Intelligence applications to make a data centrically approach (Cody et al., 2002). In fact, Business Intelligence not solely has the power to boost the structure data, however additionally to decrease info Technology prices by deleting duplicated information and eliminating inessential information But what's "Business Intelligence"? Inside the literature there are plenty of definitions and opinions concerning Business Intelligence. Consistent with Duan & Xu (2012) Business Intelligence is that the method of changing information into info that offer associate degree organizations with new insights and edges choices creating. Watson & Wixom (2007) outline Business Intelligence as a method with 2 primary activities – the primary activity is to induce the info into a knowledge warehouse and also the second activity is to induce the info out of the info warehouse and use it run a question, to perform associate degree analysis, or use it for reportage. Chaudhuri (2011) argues that Business Intelligence could be an assortment of various technologies that alter associate degree worker to create higher and quicker choices. However, during this thesis Business Intelligence refers to the applications, methodologies, practices, systems, techniques, and technologies that analyze knowledge to assist a corporation perceive their operations and market and build timely choices. The landscape of Business Intelligence applications is growing and organizations are quickly adopting these applications (Chaudhuri et

al., 2011). However, an important question is what benefits are achieved by organizations that use Business Intelligence applications (Elbashir et al., 2008). For instance, Business Intelligence applications alter organizations to spot profitable customers and build long run relationships with these profitable customers (Lee & Park, 2005). What is more, Business Intelligence applications can be wont to consistently analyse the structure external surroundings (Chung et al., 2005). as an example, a Business Intelligence application that runs on a weekly basis and helps to extract valuable market info of all competitors and establish new business opportunities (Chen et al., 2012). Business Intelligence applications may even be used for time period knowledge – a call centre may use some screens to show the performance or associate degree airline will establish passengers UN agency are in danger of missing their flight.

However, some organizations cannot directly see the opportunities of Business Intelligence applications, as a result of the benefits of Business Intelligence applications are largely nonfinancial and intangible (Lönnqvist & Pirttimäki, 2006). Most Business Intelligence applications that claim to try and do analysis solely offer some completely different views of data (Chung et al., 2005). Additionally, |half of the prices and eighty percent of the time of a Business Intelligence application is thanks to poor knowledge quality, gift systems, and issues with knowledge possession (Watson & Wixom, 2007). For instance, Business Objects – a Business Intelligence.

Application – needs a selected data Technology infrastructure so as to perform properly (Elbashir et al., 2008). Lastly, thanks to the restricted visual capabilities and therefore the new opportunities oxyacetylene by the net, organizations need new and smarter Business Intelligence applications (Chen et al., 2012). According to Watson & Wixom (2007) organizations has a lot of probably to own success with business Intelligence once the subsequent conditions exist:

1. Management of a corporation ought to have a vision for Business Intelligence and believe information-based deciding.

2. The utilization of Business Intelligence and analytics ought to be a part of the structure culture and counter deciding supported intuition or “gut feelings”.

3. Alignment between business methods, business model, and Business Intelligence methods allows a corporation to make structure modification and new business opportunities.

4. A corporation ought to have a robust and effective Business Intelligence governance and infrastructure, as a result of it'll address business alignment, funding, project prioritization, and knowledge quality.

5. Lastly, a corporation must give users with applicable Business Intelligence tools for his or her wants and provides effective coaching and support to those users.

Thus, Business Intelligence applications will change organizations to spot profitable customers, facilitate a company to research their external setting, and counter higher cognitive process supported intuition or “gut feelings” (Chung et al., 2005; Lee & Park, 2005; Watson & Wixom, 2007).

2.6. Decision Making

For most businesses and government agencies, lack of knowledge isn’t an issue. In fact, it’s the opposite: there’s typically an excessive amount of data offered to form a transparent call.

With such a lot information to type through you wish one thing a lot of from your data:

- you wish to grasp it's the proper information for respondent your question;
- you wish to draw correct conclusions from that data; and
- you wish information that informs your higher cognitive {process} process

In short, you wish higher information analysis. With the proper information analysis method and tools, what was once an awesome volume of disparate data becomes a straightforward, clear call purpose?

To improve your information analysis skills and modify your selections, execute these 5 steps in your information analysis process:

Step 1: Define Your Questions

In your structure or business information analysis, you want to begin with the correct question(s). Queries ought to be measurable, clear and terse. Style your inquiries to either qualify or disqualify potential solutions to your specific downside or chance.
For example, begin with a clearly outlined problem: A government contractor is experiencing rising prices and is not any longer ready to submit competitive contract proposals. One in every several inquiries to solve this business downside would possibly include: will the corporate cut back its workers while not compromising quality?

Step 2: Set Clear Measurement Priorities

This step explains with two sub-steps:
A) Decide what to measure
B) Decide how to measure it

A) Decide What to Measure: Using the Government contractor example, take into account what reasonably information you'd have to answer your key question. During this case, you'd have to understand the quantity and price of current employees and therefore the share of your time they pay on necessary business functions. In responsive this question, you probably have to answer several sub-questions (e.g., Are employees presently under-utilized? If thus, what method enhancements would help?).
Finally, in your call on what to live, make certain to incorporate any cheap objections any stakeholders might need (e.g., If employees are reduced, however would the coeporate reply to surges in demand?).

B) Decide How to Measure It: Thinking about however you live your knowledge is simply as vital, particularly before the information assortment section; as a result of your activity method either backs up or discredits your analysis in a while.

Key inquiries to evoke this step include:

• What's it slow frame? (e.g., annual versus quarterly costs)

• What's your unit of measure? (e.g., USD versus Euro)

• What factors ought to be included? (e.g., simply annual remuneration versus annual remuneration and price of employees benefits)

Step 3: Collect Data

With your question clearly outlined and your measure priorities set, currently it's time to gather your knowledge. As you collect and organize your knowledge, bear in mind to stay these details in mind:

• Before you collect new knowledge, verify what data may well be collected from existing databases or sources there. Collect this knowledge initial.

• Verify a file storing and naming system previous time to assist all tasked team members collaborate. This method saves time and prevents team members from assembling the identical data doubly.

• If you wish to collect knowledge via observation or interviews, then develop associate interview guide previous time to make sure consistency and save time.

• Keep your collected knowledge organized in an exceedingly log with assortment dates and add any supply notes as you go (including any knowledge normalisation performed). This observe validates your conclusions down the road.

Step 4: Analyse Data

After you've collected the proper knowledge to answer your question from Step one, it's time for deeper knowledge analysis. Begin by manipulating your knowledge in a very range of various ways in which, like plotting it out and finding correlations or by making a pivot table in stand out. A pivot table permits you to type and filter knowledge by completely different variables and permits you to calculate the mean, maximum, minimum and variance of your knowledge – simply take care TO AVOID THESE 5 PITFALLS of applied math knowledge analysis.

As you manipulate knowledge, you will notice you've got the precise knowledge you would like, however additional possible, you would possibly have to revise your original question or collect additional knowledge. Either way, this primary analysis of trends, correlations, variations and outliers helps you FOCUS YOUR knowledge ANALYSIS ON higher responsive YOUR QUESTION and any objections others may need.

During this step, knowledge analysis tools and computer code are extraordinarily useful. Visio, Minitab and Stata are all smart computer code packages for advanced applied math knowledge analysis. However, in most cases, nothing quite compares to Microsoft stand out in terms of decision-making tools. If you would like a review or a primer on all the functions stand out accomplishes for your knowledge analysis, we have a tendency to suggest this HARVARD BUSINESS REVIEW category.

Step 5: Interpret Results

After analysing your knowledge and presumably conducting more analysis, it's finally time to interpret your results. As you interpret your analysis, confine mind that you simply cannot ever prove a hypothesis true: rather, you'll solely fail to reject the hypothesis. That means that irrespective of what quantity knowledge you collect, likelihood might continuously interfere together with your results.

As you interpret the results of your knowledge, raise yourself these key questions:

• will the info answer your original question? How?

• will the info facilitate your defend against any objections? How?

• Are there any limitation on your conclusions, any angles you haven't considered?

If your interpretation of the info holds up underneath all of those queries and concerns, then you doubtless have return to a productive conclusion. The sole remaining step is to use the results of your knowledge analysis method to choose your best course of action.

By following these 5 steps in your knowledge analysis method, you create higher selections for your business or office as a result of your selections are backed by knowledge that has been robustly collected and analysed. With observe, your knowledge analysis gets quicker and a lot of correct – that means you create higher, a lot of au fait selections to run your organization most effectively. Want to draw the foremost correct conclusions from your knowledge? Click below to transfer a free guide from huge Sky Associates and find out however the correct data analysis drives success for your organization.

Chapter 3: Methodology

3.1. Python

Guido van Rossum created Python and it absolutely was 1st obtainable in 1991. Python is a taken, high-level, general artificial language. Python incorporates a style philosophy that emphasizes code readability, notably victimization vital whitespace. It provides constructs that modify clear programming on each tiny and huge scales. Now a days, it's changing into a decent selection for knowledge analysis.

It's some characteristic to use it for knowledge analysis. This options are – giant online Community, Open supply & Free, and straightforward to find out. However it would take a lot of mainframe time compare to alternative artificial language. Python IDE is for write and execute python program. Python has several IDE versions. From obtainable IDE versions 3.7 has latest options and higher optimization than alternative versions. Python IDE is for write and execute python program.

3.2. Anaconda

Anaconda is ASCII text file platform for knowledge science, machine learning applications, massive scale processing, prognosticative analytics, etc. It supports Python and R programing language. Readying and package management is extremely simplified in Eunectes murinus. Here around 1400 common data-science package obtainable that is appropriate in Windows, UNIX and MacOS. It's around VI million users. On the opposite hand, period of time collaboration is extremely vital in knowledge science.
Eunectes murinus permits to period of time collaboration for contemporary knowledge analysis and build commonplace that others will follow.

It's a virtual setting manager that is thought as Eunectes murinus Navigator. Actually, Eunectes

murinus Navigator could be a user interface, that that permits you to launch applications and simply manage conda packages, environments and channels while not the requirement to use statement commands. JupyterLab, Jupyter Notebook, QtConsole, Spyder, Glueviz,Orange, Rstudio, Visual Studio Code this application are obtainable in Eunectes murinus Navigator. It conjointly has cloud platform that permits user notice, access, store and share public and personal notebooks, environments, and PyPI packages. Cloud hosts helpful Python packages, notebooks and environments for a large kind of applications.

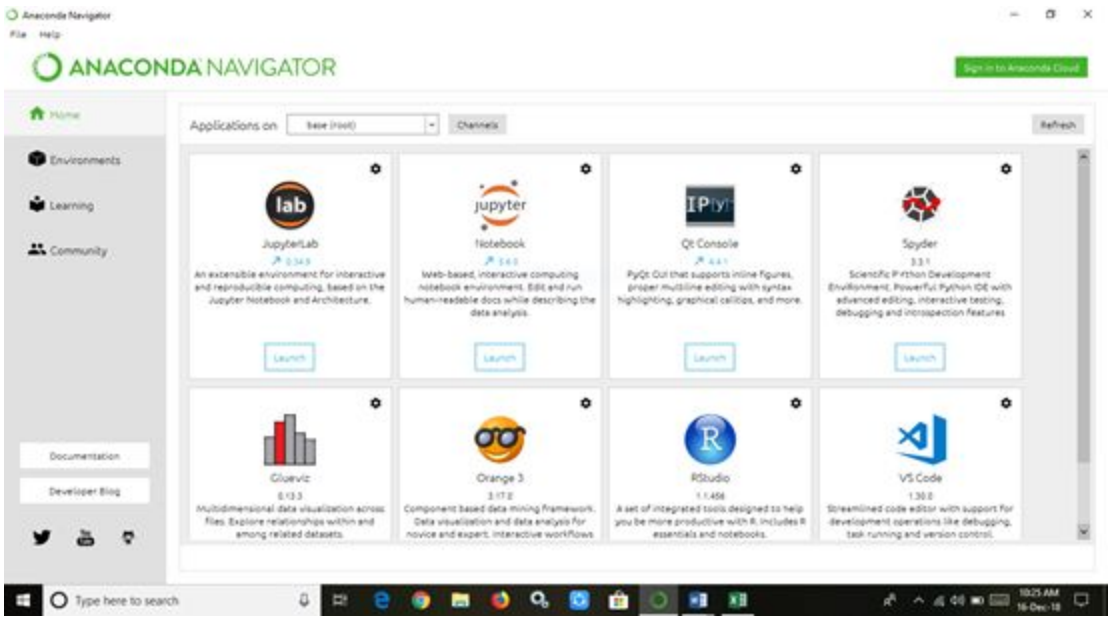

Figure 1 – Anaconda Navigator

3.3. JupyterLab

JupyterLab is a part of non-profit organization Jupyter. Jupyter is an open-source software, open-standards, and services for interactive computing across dozens of programming languages. JupyterLab is a web-based user interference (UI) for Data Science. It is supported on Chrome,

Firefox and Safari. It allow us to data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. We used Anaconda Navigator for launch JupyterLab. For programming language we used Python.

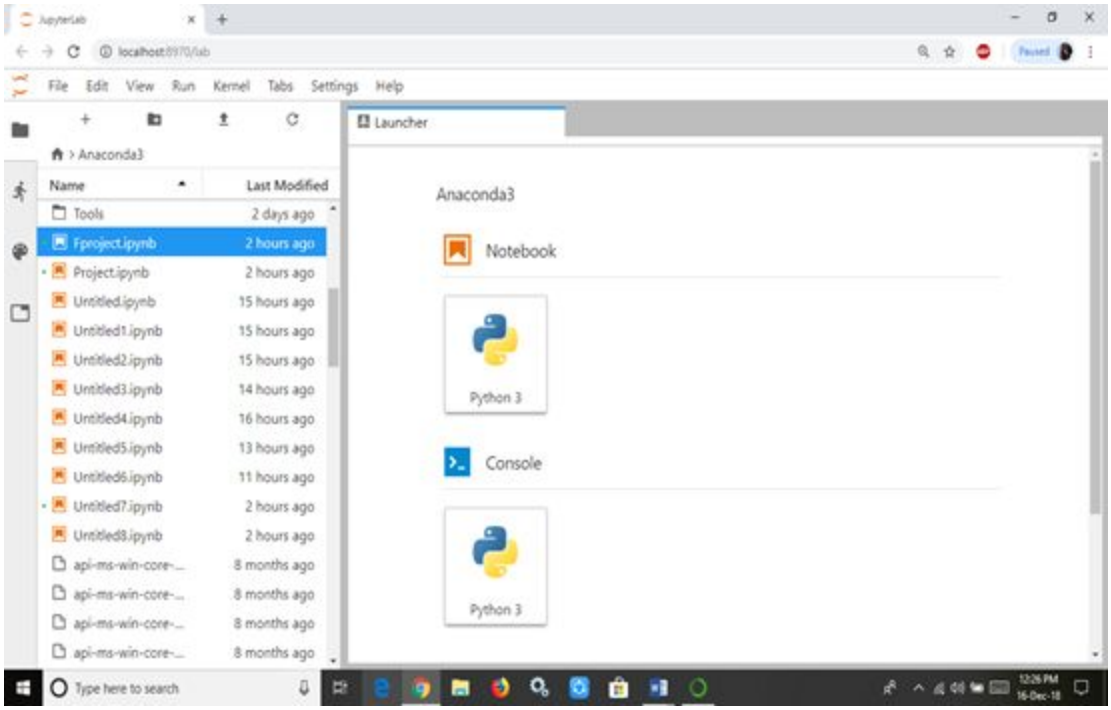

Figure 2: JupyterLab on Chrome Browser.

3.4. Python Libraries

1. NumPy

NumPy (short for Numerical Python) is one in every of the highest libraries equipped with helpful resources to assist knowledge scientists flip Python into a strong scientific analysis and modelling tool. The favoured open supply library is out there beneath the BSD license. It's the

foundational Python library for playing tasks in scientific computing. NumPy is an element of a much bigger Python-based system of open supply tools known as SciPy.

This library empowers Python with substantial knowledge structures for effortlessly playing multi-dimensional arrays and matrices calculations. Besides its uses in resolution algebra equations and different mathematical calculations, NumPy is additionally used as a flexible multi-dimensional instrumentation for various varieties of generic knowledge.

Furthermore, it integrates cleanly with different programming languages like C/C++ and FORTRAN. The flexibility of the NumPy library permits it to simply and fleetly coalesce with an in depth vary of databases and tools. As an example, let's examine however NumPy (abbreviated np) will be used for multiplying 2 matrices.

2. Pandas

Pandas is another nice library that may enhance your Python skills for information science. Rather like NumPy, it belongs to the family of SciPy open supply code and is accessible beneath the BSD free code license.

Pandas offers versatile and powerful tools for mugging information structures and performing arts in depth information analysis. The library works well with incomplete, unstructured, and unordered real-world data—and comes with tools for shaping, aggregating, analyzing, and visualizing datasets.

There are 3 sorts of information structures during this library:

Series: single-dimensional, unvaried array.

Data Frame: two-dimensional with heterogeneously typewritten columns.

Panel: three-dimensional, size-mutable array.

3. Matplotlib

Matplotlib is additionally a part of the SciPy core packages and offered below the BSD license. It’s a well-liked Python scientific library used for manufacturing straight forward and powerful visualizations. You’ll use the Python framework for knowledge science for generating artistic graphs, charts, histograms, and different shapes and figures—without worrying concerning writing several lines of code. For instance, let's have a look at however the Matplotlib library will be wont to produce an easy chart.

3.5. Logistic Regression Model

During this Model, dependent variables are binary. It's wont to describe correlation between one variable and one or a lot of nominal, ordinal, interval or ratio-level freelance variables. It’s employed in machine learning, knowledge analysis, medical fields etc.

3.6. Decision Tree Model

Call tree models may be a develop classification systems that predict or classify future observations supported a group of call rules from on the market knowledge. Knowledge is employed to make rules which will use to classify recent or new cases with most accuracy. Several rule is on the market is employed to try and do call tree model.

Here is a simple decision tree for buying a car:

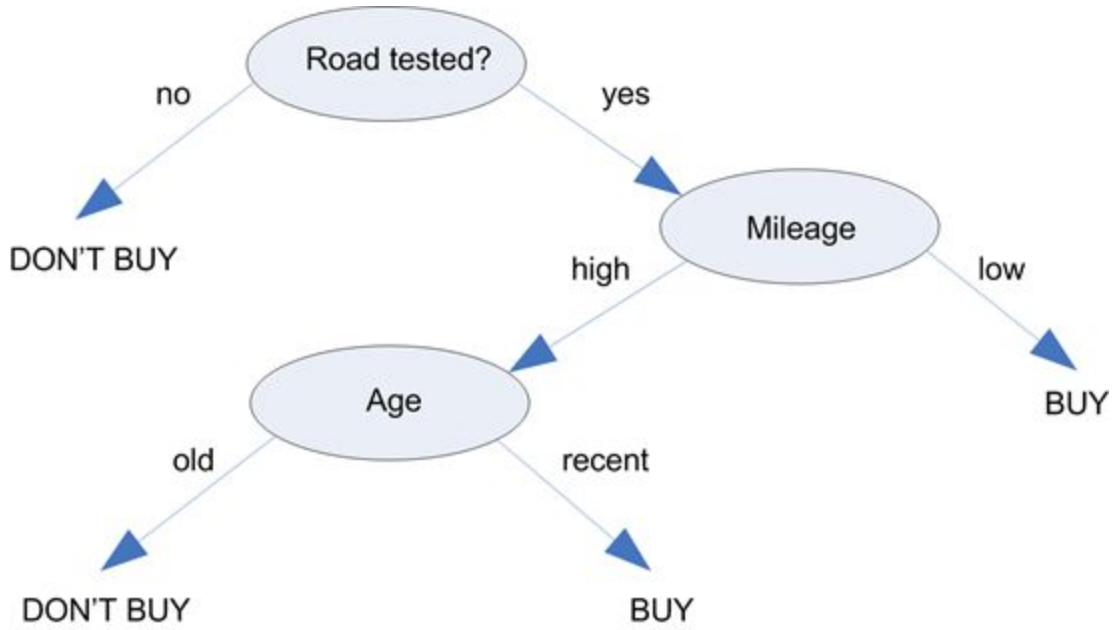

Figure 3 – Decision Tree Model

3.7. Random Forest Model

Random forest is additionally referred to as random call forests.

This methodology is initial planned by holmium in 1995. It builds multiple call trees like many trees within the forest, then merge and compare those trees to urge a lot of correct and stable prediction. It's employed in each classification and regression model. Random Forest increase prediction power, increase model speed. This methodology is employed available market prediction, E-Commerce, Banking Sector, Medical sector etc.

3.8. Cross Validation

Cross-validation a model validation techniques for assessing however the results of an applied mathematics associate degree alias can generalize to a freelance knowledge set.

It’s principally utilized in settings wherever the goal is prediction, and one desires to estimate however accurately a prognostication model camper form in follow. The goal of cross-validation is to outline a knowledge set to check the model within the coaching introduce order to limit issues like overfitting, underneath fitting and obtain associate degree insight on however the model can generalize to associate degree freelance knowledge set. It’s vital the validation and also the coaching set to be drawn from the identical distribution otherwise it might create things worse.

Chapter 4: Experiment

4.1. Import Libraries

Python library could be an assortment of functions and strategies that permits you to perform plenty of actions while not writing your own code. For instance, if you're operating with information, numpy, scipy, pandas, etc. are the libraries you want to apprehend. Here we import numpy, paandas, matplotlib and sklearn for our experiment in JupyterLab.

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# keeps the plots in one place. calls image as static pngs
%matplotlib inline
import matplotlib.pyplot as plt # side-stepping mpl backend
import matplotlib.gridspec as gridspec # subplots

#Import models from scikit learn module:
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.cross_validation import KFold   #For K-fold cross validation
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn import metrics
```

Figure 4 – Import Libraries

4.2. Load Data

Breast cancer refers to malignant tumors that developed from cells in the breast. Tumors in the breast can be benign (non-cancerous) or malignant (cancerous), the latter one being harmful[7]. CSV files are a typical file format for transferring and storing knowledge. the power to browse, manipulate, and write knowledge to and from CSV files victimization Python could be a key ability to master for any knowledge mortal or business analysis. Python Pandas read_csv - Load knowledge from CSV Files. Here we import clinical_data_breast_cancer.csv file for analyze.

```
df = pd.read_csv("F:\\clinical_data_breast_cancer.csv")
```

Figure 5 – Python Code for Load Data

4.3. Explore Data

By using df.head() we explore the dataset load from clinical_data_breast_cancer.csv. Here we see a dataset of breast cancer patient details. We are going to analyze this data.

```
df.head()
```

Figure 6 – Python Code for Explore Data

Complete TCGA ID	Gender	Age at Initial Pathologic Diagnosis	ER Status	PR Status	HER2 Final Status	Tumor	Tumor--T1 Coded
TCGA-A2-A0T2	FEMALE	66	Negative	Negative	Negative	T3	T_Other
TCGA-A2-A0CM	FEMALE	40	Negative	Negative	Negative	T2	T_Other

Table 1.1 - Data Sample(Part 1)

Node	Node-Coded	Metastasis	Metastasis-Coded	AJCC Stage	Converted Stage	Survival Data Form	Vital Status
N3	Positive	M1	Positive	Stage IV	No_Conversion	Followup	DECEASED
N0	Negative	M0	Negative	Stage IIA	No_Conversion	Followup	DECEASED

Table 1.2 - Data Sample(Part 2)

Days to Date of Last Contact	Days to date of Death	OS event	OS Time	PAM50 mRNA	RPPA Clusters
240	240	1	240	Basal-like	Basal
754	754	1	754	Basal-like	Basal-like

Table 1.3 - Data Sample(Part 3)

4.4. Attribute Details

In clinical_data_breast_cancer.csv file, there is 105 row and 22 columns. This 22 columns has different attribute name. This attribute defines data's information type.

Complete TCGA ID	ID number of patient.
Gender	Divide in Male and Female categories. It defines the patient gender.
Age at Initial Pathologic Diagnosis	Patient age when he/she get pathological diagnosis.

ER Status	It shows estrogen Receptor hormone is positive or negative.
PR Status	It shows Progesterone Receptor is positive or negative.
HER2 Final Status	HER2 is a gene that can play a role in the development of breast cancer. This attribute represent the HER2 status.
Tumor	4 kind of tumor is available – T1, T2, T3, T4. This attributes represent which type of tumor it is.
Tumor--T1 Coded	It represent the tumor is T1 or not.
Node	4 types of node available – N0, N1, N2, N3. N0: There is no cancer in nearby lymph nodes. N1, N2, and N3: Refers to the number and location of lymph nodes that contain cancer. The higher the number after the N, the more lymph nodes that contain cancer. This Node attribute represent lymph nodes cancer status.
Node-Coded	It represent node it N0 or not.
Metastasis	*This attribute has two part – M0 and M1. M0*: Cancer has not spread to other parts of the body. *M1*: Cancer has spread to other parts of the body.
Metastasis-Coded	Here shows the M0 and M1 in to negative and positive terms.
AJCC Stage	Initial Stage of cancer.
Converted Stage	Patient present stage of cancer
Survival Data Form	Survival Data of patient.
Vital Status	Patient is dead or alive.

Days to Date of Last Contact	Total treatment days.
Days to date of Death	Total treatment days before death.
OS event	Patient dead or alive in binary values.
OS Time	Total Days of treatment.
PAM50 mRNA	PAM50 is a tumor profiling test that helps determine the benefit of using chemotherapy in addition to hormone therapy for some estrogen receptor-positive.
RPPA Clusters	Types of PAM50.

Table 2 – Attribute Details

4.5. Prepare Data

In clinical_data_breast_cancer.csv data set, all data is not in numerical values. It is difficult to analysing text data. So we converted our wanted data in to numerical values.

Code for prepare data:

```
df['Gender'] = df['Gender'].map({'FEMALE':1,'MALE':0})
df['ER Status'] = df['ER Status'].map({'Negative':0,'Positive':1})
df['PR Status'] = df['PR Status'].map({'Negative':0,'Positive':1})
df['HER2 Final Status'] = df['HER2 Final Status'].map({'Negative':0,'Positive':1})
df['Tumor'] = df['Tumor'].map({'T4':3,'T3':2,'T2':1,'T1':0})
df['Tumor--T1 Coded'] = df['Tumor--T1 Coded'].map({'T_Other':1,'T1':0})
df['Node'] = df['Node'].map({'N3':3,'N2':2,'N1':1,'N0':0})
df['Node-Coded'] = df['Node-Coded'].map({'Negative':0,'Positive':1})
```

```
df['Metastasis'] = df['Metastasis'].map({'M0':0,'M1':1})
df['Vital Status'] = df['Vital Status'].map({'DECEASED':0,'LIVING':1})
df['AJCC Stage'] = df['AJCC Stage'].map({'Stage I':0,'Stage II':1, 'Stage III':2,'Stage IV':3,'Stage IA':4,'Stage IIA':5,'Stage IIIA':6,'Stage IB':7,'Stage IIB':8,'Stage IIIB':9,'Stage IC':10,'Stage IIC':11,'Stage IIIC':12})
df.head()
```

4.6. Describe Data

By using df.describe() method we count the total data, mean, std, min and max values.
We also know 25%, 50%, 75% data condition from here also.

```
df.describe()
```

Figure 7 – Python Code for Describe Data

4.7. Vital Status Analysis

Here we analysis the Vital Status data. By analysis this we are going to know when a patient is dead and when penitent is survived.

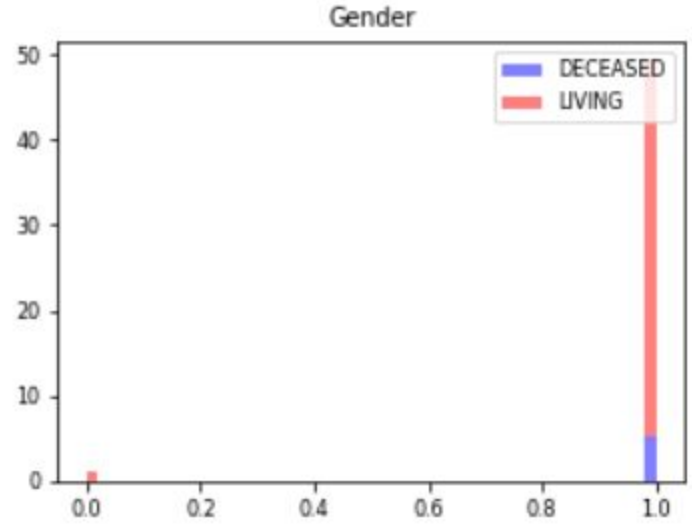

Figure 8 – Gender

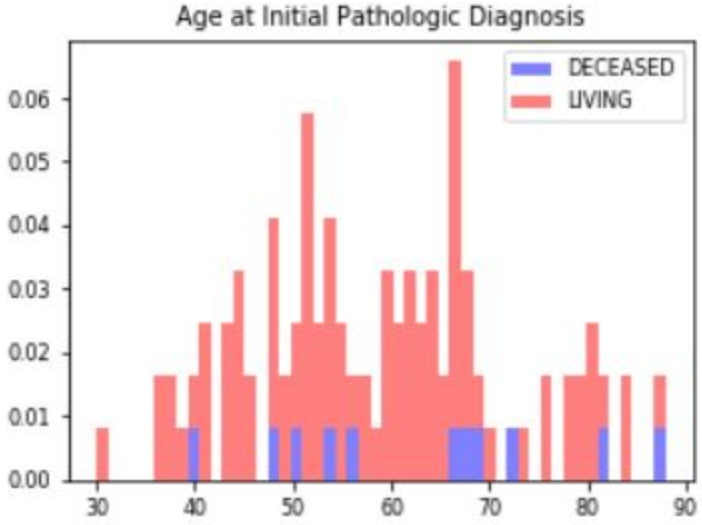

Figure 9 - Age

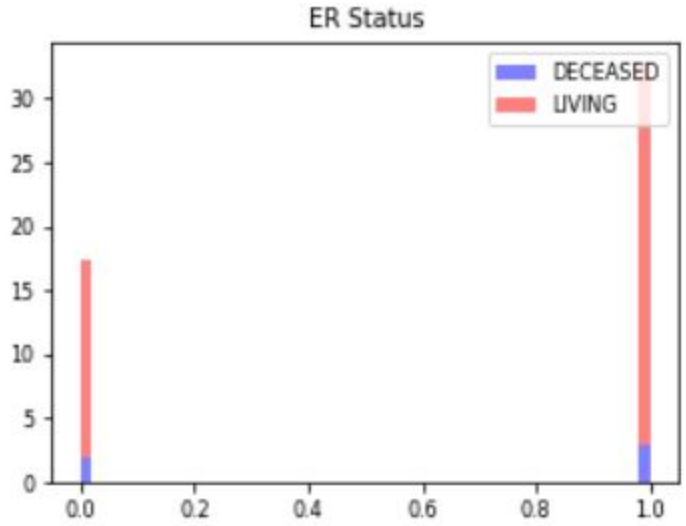

Figure 10 – ER Status

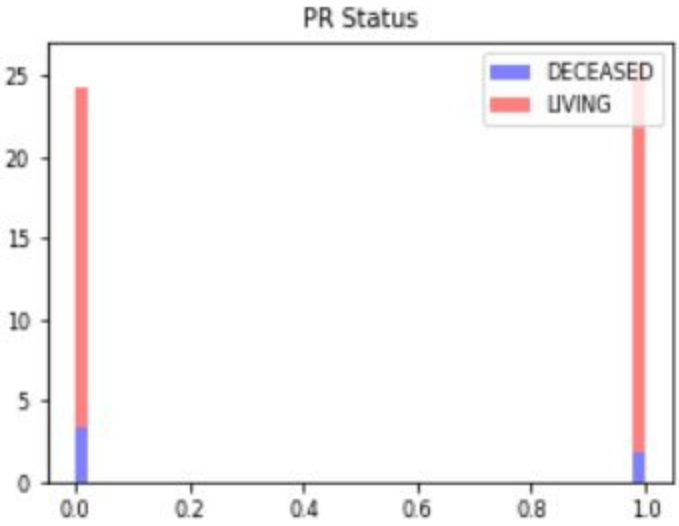

Figure 11 – PR Status

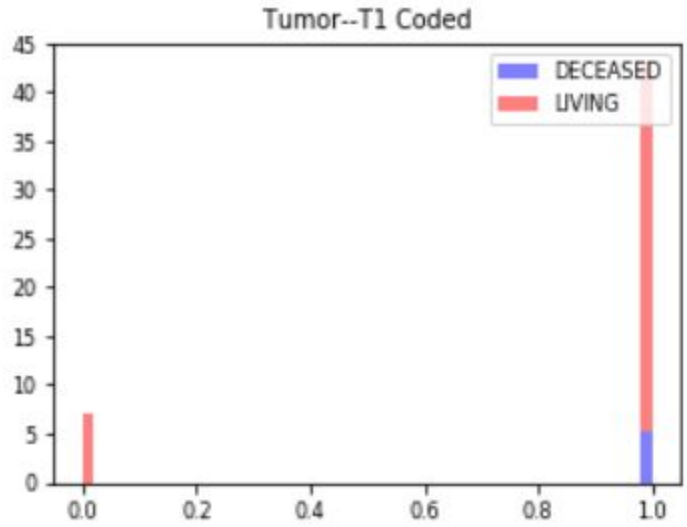

Figure 12 – Tumor—T1 Coded

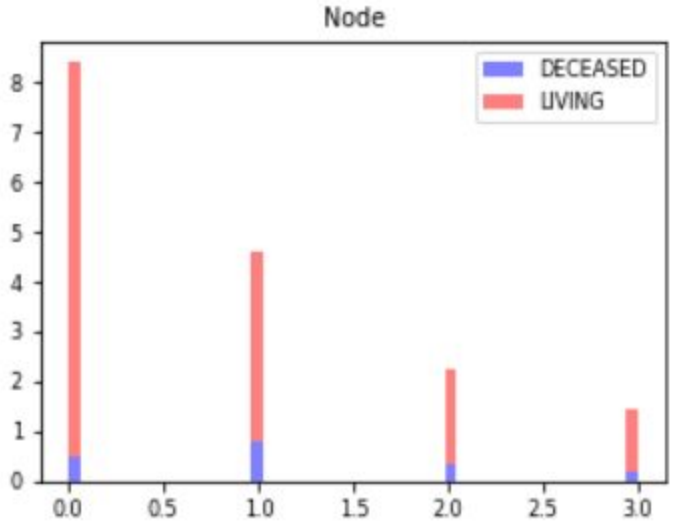

Figure 13 – Node

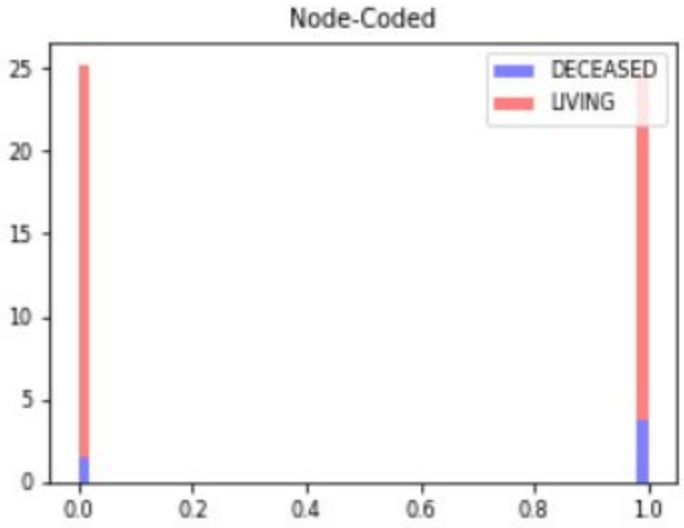

Figure 14 – Node – Code

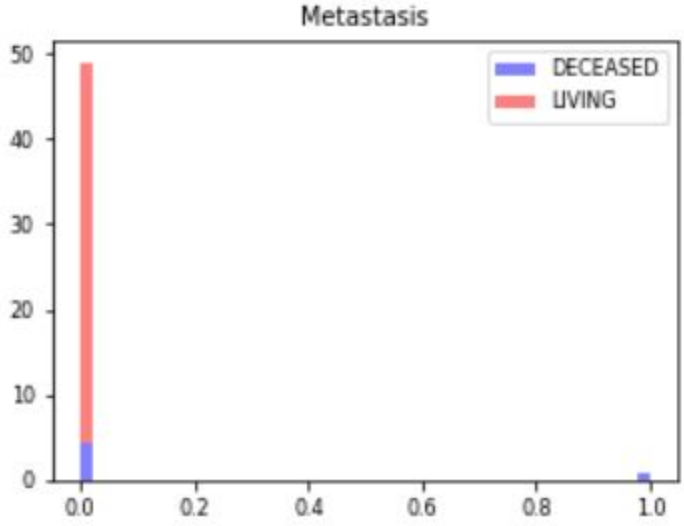

Figure 15 – Metastasis

4.8. Observation

1. Male breast cancer is very rare case. In our data there is no record of male death due to breast cancer. Large number of female is diagnosed with breast cancer. Death ratio is average for female breast cancer patient.

2. Breast Cancer under age 35 is very low. Age 50 to age 70 has a very high chance of affected with breast cancer. Age doesn't matter in death due to breast cancer.

3. ER & PR status is not matter in breast cancer. Death % of breast cancer patient with ER & PR with any stage is almost same.

4. Breast cancer with tumor stage T1 has no death record. That means if any patient have breast cancer with tumor stage T1, patient may get well soon easily. But for tumour with T2, T3, T4 has a risk of death.

5. Node has no relation with breast cancer patient death. Patient can die in any stage of Node-N0, N1, N2, and N3.

6. Metastasis shows an important information. If any breast cancer patient attacked with M1 type metastasis, then data shows 100% death record. But it also positive news that, very few number of breast cancer patient is attacked with M1 metastasis. Large number of patient attacked with M0 type metastasis but death rate is normal here.

4.9. Logistic Regression Model

The accuracy of the predictions are good but not great. The cross-validation scores are reasonable.

Figure 16 – Logistic Regression Code and Output.

4.10. Decision Tree Model

```
[71]: predictor_var = ['Gender','Age at Initial Pathologic Diagnosis','ER Status','PR Status','HER2 Final Status','Tumor','Tumor--T1 Coc
      model = DecisionTreeClassifier()
      classification_model(model,traindf,predictor_var,outcome_var)
```

```
Accuracy : 100.000%
Cross-Validation Score : 86.667%
Cross-Validation Score : 80.000%
Cross-Validation Score : 84.444%
Cross-Validation Score : 84.762%
Cross-Validation Score : 86.381%
```

Figure 17 – Decision Tree Code and Output.

4.11. Random Forest Model

```
predictor_var = features_mean
model = RandomForestClassifier(n_estimators=100,min_samples_split=25, max_depth=7, max_features=2)
classification_model(model, traindf,predictor_var,outcome_var)
```

```
Accuracy : 93.151%
Cross-Validation Score : 93.333%
Cross-Validation Score : 93.333%
Cross-Validation Score : 93.333%
Cross-Validation Score : 91.429%
Cross-Validation Score : 93.143%
```

Figure 18 – Random Forest Code and Output.

Chapter 5: Conclusion & Future Work

5.1. Conclusion

Big Data still a new technology. Lots of research is happening on this topic. Big Data analysis can be combine with machine learning, artificial intelligence. The trio combination of big data, machine language, and artificial intelligence can make human life smarter. Work which is time consuming or risky, can be solved by this technologies. Big companies like google, walmart, amazon, facebook already invested lots of money on big data. Because they understand that data is the next generation gold. The much data you have, much stronger you are. Big data is not only benefited for business world, it is also benefited for social gain and medical treatment. In our project, we wants to use big data for benefited our social life. We are working with breast cancer data. Breast cancer is a burning topic for women's in the world. More and more research should done on this topic by using Big Data. More observation on breast cancer data make a change in death ratio of breast cancer patients. This observation also can help government and other organization to make awareness people about breast cancer.

5.2. Future Work

In the world, especially women are seriously threatened by breast cancer with high morbidity and mortality [8]. For our analysis we taken breast cancer patient data from www.kaggle.com [9]. It is a very short amount of data. We also doesn't the origin of this data. We will continue our work on this topic. We wants to collect real patient data of Bangladeshi breast cancer patient and discover the death factors of breast cancer patient. We wants to apply machine learning, artificial technology which can help early detection of breast cancer. We also wants to build a website where anyone can post if they see any difference on their breast, and our automatic data analysis system can predict the current condition by analysis previous patient data.

Bibliography

1. Hossain, Mohammad & Ferdous, Shameema & Karim-Kos, Henrike. (2014). Breast cancer in South Asia: A Bangladesh perspective. Cancer Epidemiology. 38. 10.1016/j.canep.2014.08.004.

2. https://datafloq.com/read/big-data-history/239

3. https://cognitiveclass.ai/courses/what-is-big-data/

4. https://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/6

5. S. M. Shamsuddin and S. Hasan, "Data science vs big data @ UTM big data centre," 2015 International Conference on Science in Information Technology (ICSITech), Yogyakarta, 2015, pp. 1-4. doi: 10.1109/ICSITech.2015.7407766

6. https://opensource.com/article/18/9/top-3-python-libraries-data-science

7. H. Hasan and N. M. Tahir, "Feature selection of breast cancer based on Principal Component Analysis," 2010 6th International Colloquium on Signal Processing & its Applications, Mallaca City, 2010, pp. 1-4. doi: 10.1109/CSPA.2010.5545298

8. B. Fu, P. Liu, J. Lin, L. Deng, K. Hu and H. Zheng," Predicting Invasive Disease-Free Survival for Early-stage Breast Cancer Patients Using Follow-up Clinical Data," in IEEE Transactions on Biomedical Engineering.

9. www.kaggle.com/piotrgrabo/breastcancerproteomes#clinical_data_breast_cancer.csv