MISSING DATA PREDICTION FROM WIRELESS SENSOR NETWORKS

By

PC Nashrah Khan
GC Mansoor Ahmad
PC Aniqa Tariq
NC Ali Sultan

Submitted to Faculty of Department of Computer Software Engineering National University of Sciences and Technology, Islamabad in partial fulfillment for the requirements of a B.E Degree in Computer Software Engineering, June 2019

In the name of Allah, the Most Beneficent, the Most Merciful

ABSTRACT

The aim of this project Missing Data prediction is to predict the missing values of well-known organization called British Petroleum (BP) and to eradicate the noisy data from their datasets. This will aid to find out the future left over reservoirs and future oil prices . Various machine learning algorithms were implemented and the best one among them was chosen out which gives most accurate predicted values and lessens the noisy data. The algorithm implemented is molded into a module, which provides ease of use to the users and also predicts missing data in other data types such as Tensorflow Tensor, Pytorch Tensor, Numpy Array, MXNet Nd Array.A dataset of the same data types as of the input dataset by the user and with the missing values replaced by the predicted values of the prediction algorithm is the output of this module.

CERTIFICATE FOR CORRECTNESS AND APPROVAL

Certified that the work contained in the thesis –Missing data prediction carried out by PC Nashrah Khan, GC Mansoor Ahmad, PC Aniqa Tariq and NC Ali Sultan under the supervision of Dr Saddaf Rubab for partial fulfillment of degree of Bachelor of Software Engineering is Correct and approved.

Approved by

Dr Saddaf Rubab

Department of CSE, MCS

Dated: _____

DECLERATION

No portion of the work presented in this dissertation has been submitted in Support of another award or qualification either at this institution or elsewhere.

DEDICATION

In the name of Allah, the Most Merciful, themost Beneficent. To our parents, without whose unflinching support and cooperation, a work of this magnitude would not have been possible.

ACKNOWLEDGMENTS

We would like to thank Allah Almighty for His incessant blessings which have been bestowed upon us. Whatever we have achieved, we owe it to Him in totality. We are also thankful to our families for their continuous moral support which makes us what we are.

We areextremely gratefulto our project supervisorDr Saddaf Rubab from MCS who in addition to providing valuable technical help and guidance also provided us moral support and encouraged us throughout the development of the project.

We would also like to thank ourCo Supervisor Dr Yasar Ayaz (HoD R&AI, SMME) for the encouragement, motivation and support.We are highly thankful to all our teachers and staff of MCS who supported and guide us throughout our coursework. Their knowledge, guidance and training enabled us to carry out this whole work. Finally we are grateful to the Faculty of Computer Software Engineering of the Military College of Signals, NUST.

In the end we would like to acknowledge the support provided by all our friends, colleagues and a long list of well-wishers whose prayers and faith in us propelled us towards our goal.

Table of Contents

Table Of Figures

Table of Tables

CHAPTER # 1

INTRODUCTION

1.1 Introduction

Our aim is to devise a model which is going to predict the missing data as well as the noisy data accurately. After we develop a working model of data prediction, that same model is going to be implemented by British Petroleum (BP).

1.2 Purpose

In a huge dataset, data is lost during the acquisition phase and transmission phase. This loss reduced the efficiency of the Machine Learning (ML) model as well as cause problems in data visualization, thus analysis failure occurs. Such is the case with British Petroleum, which needs this lost data to predict future oil prices as well as the oil reserves left at a respective location.

Thus, the purpose of our project is to develop an efficient working ML- model which is going to predict the missing values in their dataset which contains complex numerical data of oil wells, for the specified company.

1.3 Scope

Prototype of software based upon the ML- model that we will develop as a deliverable. It would be beneficial for BP to get good insights of their company progress. As well as we can extend it to other companies by training our proposed ML model on their datasets in future perspective.

1.4 Document Conventions

1. British Petroleum (BP)	7. Classification
2. Python	8. Clustering
3. Statistical Techniques	9. Principle Component Analysis (PCA)
4. Machine Learning (ML)	10. Scatter plot
5. Deep Learning	11. Box plot
6. Regression	12. Scree plot

1.5 Intended Audience

The intended readers of the SRS are all the stakeholders, detailed as follows:

- **Project Supervisor:** It will help to supervise the project and guide the team in a better way.
- **Developers:** Project developers have an advantage of quickly understanding the methodology adopted and personalizing the product.
- **Testers:** The testers of the system can check user requirements from provided SRS and develop test scenarios accordingly.
- **Documentation writers:** The document can serve as a future reference for other versions of the Software Design Specification (SDS).
- **Project Testers:** Project testers can use this document as a base for developing test cases.

1. **British Petroleum:** A highly reputable multinational oil and gas company for whom we are basically designing a prototype.

2. **Developers Team:** The developers involved in the project will include Supervisor and the group members.

1.6 Learning Sources

- 30 Days of Python | Unlock your Python Potential (udemy.com/30-days-of-python)
- Machine Learning A-Z : HandsOn Python and R in Data Science (udemy.com/machinelearning)
- Deep Learning A-Z : HandsOn Artificial Neural Networks (udemy.com/deeplearning)

CHAPTER # 2
LITERATURE REVIEW

2.1 Data

Data are plain facts. Data themselves are fairly useless. When data are gone through some process and comes into organized and structured form that it provides some meaning full information. Then it is called processed data.

2.1.1 Types of Data:

Growth in this technological era is fast. Now a day's data is in the form of text, audio, video, log and as well as web activity records. Data is not always in structured form. It is mostly in unstructured form.

The term big data used to show data in the petabytes or may be more than that. The 3Vs representing volume, variety and velocity. Now a days the world is revolving around big data. The web-based e-commerce use models of big data to predict the future. Such trends make the society more dependable on big data and creating a social issues of data privacy.

Data in various fields has a meaning beyond its use.For instance, in electronics and network communication the term data is referred to as 'control information and control bits' and sometimes called as 'transmission unit'.Moreover,in science and related fields the term data is used in gathered body of facts.

2.1.2 Data management and use

As data is more in use in organizations so there is emphasis on quality of data. We get this by data cleaning and date pre-processing phase. In which we eradicating duplicated rows and noisy data. We choose the best accurate data. Data management of modern data includes cleansing, extraction, transformation, load and integration.The term metadata referred to data about data that helps administrators and users comprehend databases and other data.

As large amount of data is in unstructured form so analytics combine structured and unstructured data and make it useful. Systems for analytics trying harder to achieve real-time performance, so they are designed to handle high ingestion rates incoming data consumed and to process data streams for immediate use in operations.

Over time, the operations and transactions of data bases has led to idea of predictive data analytics. One of the main examples is the data warehouse which is optimized to solve queries about operations for business analysts. The focus is on the predicting outcomes and finding out the patterns which led to the path of data mining techniques.

2.1.3 Accuracy and Correctness of Data

- **Accuracy**

Data used to give insights such as predictions and make better decisions. Data in raw form is not useful so we processed the data and display it in appropriate levels so that it can be implemented properly. If the data is inaccurate the results will be poor and thus gives the poor insights. Therefore, organization is more focused on data quality and use different analytical tools to get better insights.

- **The danger of inaccurate data**

Analysts have observed that organizations and companies that used accurate data. Their results are efficient, proficient and profitable. On the other hand, inaccurate data causes real detriments to businesses.

This can be obviously perilous conditions as one individual using data presentation while a second spotlights rather on Excel or standard reports. It could even on a very basic level be under-reprimanded geniuses using hardly extraordinary conveying which ousts their data from encourage. As flawless and uncommon as this all sounds, the impact could be tremendous. Curry continued suggesting an examination which showed that one pound in every six that is spent from departmental spending plan is wasted. The best idea here

isthat these hardships are totally avoidable, by ensuring that data precision persistently remains a key idea.

2.2Data Latency

Data latency estimates to what extent it takes the clients to recover information from an information distribution center. It ought to be evident that the shorter the inactivity time frame, the defter and more responsive the business. In any case, once in a while the readiness and responsiveness may come at the expense of proficiency and dependability.

2.2.1 Three types of data latency

Data latency is categorized in three types;

- **Real-time data**

Real time data is the data that winds up accessible in the database when the business action happens, with zero or almost no idleness. For example, stock-cost or cash esteem information is continuous. The present costs become accessible to clients promptly every time there's a change. Although constant information is incredible to have, accomplishing it very well may be costly and testing. You need forms that record the exchanges progressively, and the database assets that consider concurrent account and recovery of information.

- **Near-time data**

Near time information is the data that ends up accessible to clients at set interims. It is unique in relation to ongoing information as in close time information isn't recorded consistently, yet just "as auspicious as required." The day by day money report or month to month deals report is a genuine case of close time information. Close time information might be savvier to give than constant information, however giving despite everything it requires organizing the executive’s methods and conquering framework confinements.

- **Some-time data**

Sometime information is regularly refreshed just "in some cases". For instance, merchant or client contact data is put away in the framework just once, and may should be refreshed whenever there is a change.

2.2.2 How much data latency needed?

Data latency relies upon how much information we required. In the event that we are gathering information from various sources, at that point by the progression of time we need to refresh it and erase the past data to make our model work appropriately. The money of information, past a specific point, winds up inefficient, devouring superfluous assets and filling no valuable need. It additionally corrupts the nature of data, influencing information unwavering quality.

2.2.3 Examples of data latency on data reliability

There are instances that will aid to tell about effects of data latency on data reliability.

- **The example of Stock exchange**

A securities exchange will crumple if there's an idleness of even a couple of minutes between the adjustment in stock esteem and its detailing.

- **The Patients in the CCU**

A clinic can't manage the cost of even several seconds delay on the revealing of the fundamental indications of the patients in the CCU. It the information inactivity is low then the information won't be given on time this will influence the unwavering quality of the patient's reports and treatment.

- **Businesses Purposes**

Organizations frequently need information on schedule for their key choice purposes. So, in the event that there is delay in information, at that point choice can't be set aside a few

minutes and this will influence the exhibition and unwavering quality of business choices.

- **Big Data used in Uber's platform**

Uber is focused on conveying more secure and increasingly solid transportation over our worldwide markets. To achieve this, Uber depends intensely on settling on information driven choices at each dimension, from estimating rider request amid high traffic occasions to distinguishing and tending to bottlenecks in our driver-accomplice join process. After some time, the requirement for more experiences has brought about more than 100 petabytes of scientific information that should be cleaned, put away, and presented with least dormancy through our Hadoop-based Big Data stage. Since 2014, we have attempted to build up a Big Data arrangement that guarantees information unwavering quality, adaptability, and convenience, and are presently concentrating on expanding our stage's speed and proficiency. In the event that information inactivity is low, at that point the client will be in disarray that driver has arrived or not and book for another ride. The driver ought to likewise confront trouble in identifying the courses and area of the client.Data Latency is very important for Data provision in real time. Because as we see through the examples that how it will affect the performance and accuracy of the data late in any field.

2.3 Sensors Used in Wireless Sensor System

A Sensor is a device that gives responses and perceives a type of commitment from both the physical or natural conditions, for instance, weight, warm, light, etc. The yield of the sensor is all around an electrical banner that is transmitted to a controller for further gettingready. A framework of wireless sensor can be explained as an arrangement of contraptions that can grant the data amassed from a checked field through distant associations. The data is sent through many centers, and with a gateway, the data is related with various frameworks like remote Ethernet. In a remote sensor mastermind, sensors have a basic effect, as identifying is one of its central employments. Advancement behind sensors, in any case, isn't of noteworthy interest while thinking

about sensor frameworks, with the emphasis being more on correspondence, organize the board, and data control. Most sensors used in WSN structures have been made self-sufficiently of WSN development, and these two fields continue developing somewhat openly.

2.3.1 Temperature Sensor

A boss among the most comprehensively saw and most perceptible sensors is the Temperature Sensor. A Temperature Sensor, as the name recommends, assets the temperature for instance it checks the adjustments in the temperature. Traits of temperature sensor are as seek after the following;

- **Temperature expands**

The temperature degree of a sensor depicts the temperatures at which the sensor is assessed to work securely and give exact estimations. Each kind of thermocouple has a predefined temperature widen dependent on the properties of the metals utilized in making that thermocouple. RTDs offer a progressively minute temperature go as a side-effect of better linearity and exactness, and thermistors give the most inconsequential temperature grows yet brilliant affectability.

- **Linearity**

A perfect sensor would have a brilliantly straight reaction: a unit change in temperature would result in a unit change in voltage yield over the whole temperature degree of the sensor. If all else fails, by the by, no sensor is superbly quick.

- **Sensitivity**

The affectability of a given sensor exhibits the percent change in quantifiable yield for a given change in temperature. An inexorably delicate sensor, similar to a thermistor, can more effectively perceive little changes in temperature than a less dubious sensor, similar to a thermocouple. This affectability, regardless, goes to the hindrance of linearity. This can be a fundamental factor while picking the perfect sensor decision for the

temperatures you are evaluating. On the off chance that you want to get division of-a-degree changes over a little temperature go, a thermistor or a RTD is progressively perfect.

- **Stability**

The nature of a temperature sensor infers that its capacity to keep up a foreseen yield at a given temperature. Material acknowledge a key work in the security of a given sensor. RTDs are occasionally worked of platinum along these lines and despite guarantee low reactivity. The substrate to which the platinum is propped, regardless, may turn under pulled in out introduction to high temperatures, which can cause extra and sudden strain that prompts a change in evaluated obstruction.

- **Accuracy**

Accordingly, comparably similarly as with any estimation application, understanding your precision needs is essential in guaranteeing dependable outcomes. the sensor and estimation equipment choices acknowledge a fundamental work in absolute estimation exactness, yet humbler central focuses, for example, cabling, relative locale to different mechanical assembly, guaranteeing, setting up, etc. would all have the alternative to affect accuracy additionally. While picking a sensor, we ought to watch the fated insurances and any segments that may impact that confirmation (for instance, pulled in out preamble to high temperatures). Besides, be mindful so as to pick a sensor and estimation contraption with equivalent exactness's. A tight hindrance RTD comes at a continuously significant expense, yet you may not accomplish the extra exactness on the off chance that you utilize a low-quality estimation contraption.

2.3.2 Proximity Sensor

A Proximity Sensor is a non-contact type sensor that sees the closeness of a request. proximity Sensors can be executed utilizing specific procedures like Optical, Hall Effect, Ultrasonic, Capacities, and so on. The sensor limits by perceiving an alteration in the electromagnetic properties of the volume forming its 'sensitive area', which is found

basically outside the mouth of the circle and structures an expected 3D shape with sides of a near width to the gathering device twist. An inquiry going through the sensitive area, or the proximity of a substitute material can cause such a change. 'Since the sensor uses an electromagnetic field instead of either a totally appealing or electric field,' illuminates Adams, 'it is tricky to both alluring and electric properties of materials interfering into the sensitive volume meanwhile, for example, the teeth of a rotating metallic pinion going through the fragile region.' Other models join homogeneous liquid spilling inside a pipe arranged in the tricky locale, which contains little air pockets or bits of solid rubbish passed on along in the stream. As the liquid experiences the sensor a banner is conveyed that is with respect to the size or mass of the garbage and for which the banner sort is typical for that particular material.

1. Vicinity Sensors identify an item without contacting it, and they hence don't make scraped area or harm the article.

2. No contacts are utilized for yield, so the Sensor has a more extended administration life (barring sensors that utilization magnets).

3. In contrast to optical recognition techniques, Proximity Sensors are reasonable for use in areas where water or oil is utilized.

4. ProximitySensors give rapid reaction, contrasted and switches that require physical contact.

5. Vicinity Sensors can be utilized in a wide temperature go.

6. Closeness Sensors are not influenced by hues.

7. In contrast to switches, which depend on physical contact, Proximity Sensors are influenced by encompassing temperatures, encompassing items, and different Sensors.

2.3.3 Accelerometer

An accelerometer is a contraption that assesses the vibration, or reestablishing of progress of a structure. The power acknowledged by vibration or a change being created (developing rate) influences the mass "to beat" the piezoelectric material which passes on an electrical blame that is for respect to the power related upon it. Since the accuse is differentiating of the power, and the mass is an ardent, by then the charge is other than as to the reviving

2.3.4 IR Sensor (Infrared Sensor)

An infrared sensor is an electronic device, that can measure the sparkle of a test and sees the headway. These sorts of sensors measure fundamentally infrared radiation, as opposed to overflowing it that is called as a detached IR sensor. All things considered in the infrared range, most of the articles overflow a type of warm radiations. These sorts of radiations are subtle to our eyes, that can be seen by an infrared sensor. The maker is basically an IR LED (Light Emitting Diode) and the locator is on a very basic level an IR photodiode which is delicate to IR light of vague wavelength from that transmitted by the IR LED. Right when IR light falls on the photodiode, the assertions and these yield voltages, change in degree to the level of the IR light got.

The infrared sensors undergo through following laws:

1.Stephan Boltzmann Law
2. Planck's radiation law
3. Wien's Displacement Law

2.3.5 Pressure Sensor

A weight sensor is a contraption that recognizes weight and changes over it into an electric banner where the aggregate depends on the weight associated. TE Connectivity (TE) structures and manufactures weight sensors going from the identifying segment to system packaging for ruthless circumstances. We are an industry pioneer for our extent of both standard and custom weight sensors, from board level fragments to totally increased and packaged transducers. In perspective on piezoresistive Micro-electromechanical structures (MEMS) and silicon strain check (Micro fused, Krystal Bond) development, our sensors measure everything from creeps of water area (<5 mbar) to 100K psi (7K bar). A significant number of the accompanying specialized determinations show up on datasheets for weight transducers and comparable gadgets. Here we survey their specialized definitions and connections.

Root Sum Squares (RSS): accuracy of a weight transducer is discovered by taking the square establishment of non-linearity + hysteresis + non-repeatability.

- **Non-Linearity**

The relationship of an alteration curve to a foreordained straight line.

- **Hysteresis**

The most outrageous refinement in yield at any weight a motivation inside the foreordained range, when the regard is moved closer with extending and reducing weight.

- **Non-Repeatability**

The limit of a transducer to reproduce yield readings when a comparative weight regard is associated with it successively, under comparable conditions, and in a comparative heading.

- **Long-Term Stability**

The limit of a transducer to copy yield readings got in the midst of its exceptional change at room conditions for a predefined time span.

- **Zero Offset**

Zero yield is modern office set to inside a particular % of full scale. Results in a climb or down of the arrangement twist.

- **Span Offset**

Range yield is modern office set to inside a particular % of full scale. Results in an alteration in the grade of the curve.

- **Thermal Effects**

The alteration in the zero and range yield that happens in view of temperature changes.

2.3.6 Light Sensor

Light sensors seem, by all accounts, to be extremely essential. They sense the light, much equivalent to a thermometer recognizes the temperature, and a speedometer identifies speed.Temperature and speed are anything but difficult to fathom since we sense them in a straight-forward manner. Be that as it may, light is exceptionally entangled. Temperature and speed are escalated properties, so they don't rely upon the mass or size of a question. Light can be estimated as a broad property, which means the aggregate light gathered relies upon the span of the gatherer (e.g. a landfill sun-based cluster gathers more light than a little sun based telephone charger), or seriously by separating by the zone.

- **Photo diode**

  Light sensors once in a while utilize a segment called a photo diode to quantify illuminate. At the point when light emissions strike a photo diode, they tend to thump electrons free, making an electric flow stream. The more splendid the light, the more

grounded the electric flow. The current would then be able to be estimated to restore the illuminate of the light. On the off chance that light-initiated electric flow sounds recognizable, it is on the grounds that this is the working guideline of the sun-based boards used to control street signs and homes. Sun powered boards are essentially enormous photo diode light sensors.

- **Photograph resistor**

Another sort of light sensor is the photograph resistor. A photograph resistor is light-needy resistor, implying that if there is an adjustment in the brilliance of the light sparkled on it, there will be an adjustment in opposition. Photograph resistors are less expensive than photograph diodes, however are substantially less exact, so they are for the most part used to think about relative light dimensions or essentially whether a light is on or off.

2.3.7 Ultrasonic Sensor

Ultrasonic Sensor is a sort device that is non-contact and can be used to look at speed of a test. It works subject to the properties of the sound waves with repeat more recognizable than that of the human detectable range. Using the season of excursion of the sound wave, Ultrasonic Sensor can check the fragment of the test, for example, SONAR. The Doppler Shift property of the sound wave is used to evaluate the speed of a distinction.

Ultrasonic sound vibrates at a repeat over the level of human hearing. Transducers are the intensifiers used to get and send the ultrasonic sound. Our ultrasonic sensors, additionally as different others, use a singular transducer to send a heartbeat and to get the resonate. The sensor picks the division to a goal by reviewing time goes between the sending and suffering of the ultrasonic heartbeat.

2.3.8 Humidity Sensor

Humidity Sensor is a champion among the most crucial contraptions that has been extensively in buyer, mechanical, biomedical, and environmental, etc applications for

evaluating and checking Humidity. Stickiness is described as the proportion of water present in the including air. This water content detectable all around is a key factor in the prosperity of mankind.

In any case, if the temperature is 100C and the dampness is high for example the water substance of air is high, by then we will feel extremely uneasy. Mugginess is moreover a fundamental thought for working unstable equipment like contraptions, mechanical rigging, electrostatic sensitive sensor and high voltage devices, etc. Such unstable apparatus must be worked in a stickiness space that is proper for the sensor.

2.3.9 Stream and level Sensor

Wide extent of sensors is open in the market and normally, they are collected dependent on the particular use of the sensor. Sensor utilized for surveying stickiness is named as soaked quality sensor, the one utilized for estimation of weight is called weight sensor, sensor utilized for estimation of dislodging is called position sensor, etc through every one of them may utilize the close distinctive standard. In like way, the sensor utilized for estimation of liquid estimations is known as an estimation sensor. Particularly clear from its name, level sensors are utilized to check the component of the free-spilling substances. Such substances solidify fluids like water, oil, slurries, and so on and furthermore solids in granular/powder shape (solids which can stream). These substances will when all is said in done get settled in the holder tanks because of gravity and keep up their estimation in rest state. Level sensors measure their estimation against a pre-set reference.

2.3.10 Tilt Sensor

Tilt sensors is a sensor that make an electrical pennant that changes with a precise improvement. These sensors are utilized to check tendency and tilt inside a restricted degree of advancement. Every once in a while, the tilt sensors are recommended as inclinometers in light of the manner in which that the sensors basically produce a

standard yet inclinometers make both readout and a standard.
To choose best tilt sensor we look at their properties such as Number of axes, resolution, sensitivity, noise tolerance, vibration measuring range and output.

2.4 Various Methodologies for Collecting Natural Data

Natural events and natural data always seem to be random, but on a higher level, there is always some pattern that it follows. The formulation of the law of probability and several techniques of data analytics have enabled mankind to truly analyze and extract important information from natural events. The first step of data analytics is of data acquisition i.e. to acquire accurate data from the target event/subject. Data analytics is applied on data of all kinds i.e. seismic, oceanic, atmospheric and of forests which have till now given us key information e.g. about the rate of deterioration of ozone layer, the rate of deforestation, the effect of deforestation on the atmosphere and in turn have directed mankind to make the necessary changes in order to preserve nature.

Several methodologies are used to acquire data from nature which are going to be discussed here one by one.

2.4.1 Seismic Data

Seismic techniques are the most usually led geophysical studies for designing examinations. Seismic refraction gives architects and geologists the most fundamental of geologic information by means of basic systems with basic hardware.

Seismic waves are the vibrations that are produced from mechanical waves. They are started from the source and undergo the area where its sound is noted. These vibrations are weight state by virtue of disturbance ratio.Further these vibrations go all over the mediumi.e. liquid solid and gas but do not flow in vacuum.

There are two main types of seismic waves:

- **Body Waves.**
  They experiences the volume of a material

- **Surface Waves.**

  These waves are the waves that flow across the surface of the earth.

2.4.2 Wave Theory

In seismic activity wave theories can be defined as the propagation of waves through molecules. Those molecules can be of air, gas, liquid or solid. These molecules support the wave motion and wave vibration of the radiant body.

2.4.2.1 Data Acquisition

- **Source**

The source of seismic data can be an aluminum plate, rifle shots, developing sizes of drop heaps, oscillator or weighted board A geophysical definitive worker normally ought to be given degree in changing or picking the source critical for the assignment. The customer ought not vacillate in setting limits on the temporary worker's strange use of two or three sources. In private or present-day zones, maybe the most uncommon risky ge ought to be constrained. The centrality of debilitating shot gaps for explosives or rifle shots may should be bound; legitimately restricting pros ought to be careful so as not to outflank prerequisites of stipends, utility easements, and contract assent.

- **Geophones**

The sensor getting seismic centrality is the geophone otherwise called phone. These sensors can be speed transducers or accelerometers, and convert ground improvement into a voltage. Ordinarily, the hoisting of the ground is various offers of enormity, yet talented on a relative reason. The overall estimation of particle invigorating can't be settled, aside from if the geophones are balanced.

Most geophones have vertical, single-center guide response toward get the pushing toward waveform from underneath the surface. Some geophones have level turn response

for S-wave or surface wave assessments. Triaxial phones, orchestrated looking over absolutely response, are used expressly follows. Geophones are picked for their repetitive band response.

The line, spread, or game-plan of phones may contain one to scores of sensors depending on the sort of study. The individual station of record routinely will have a lone phone. Various phones per station may help in lessening breeze commotion or air impact or in improving gigantic reflections.

- **Seismographs**

The rigging that records input geophone voltages in a masterminded development is the seismograph. Current practice uses seismographs that store the channels' signs as cutting-edge data at discrete time. Earlier seismographs would record undeniably to paper or photographic film. Stacking, contributing, and dealing with the enormous volumes of data and recording the information for the client basically require pushed seismographs. The seismograph structure may be a confounding amalgam of mechanical assembly to trigger or recognize the source, digitize geophone signals, store multichannel data, and give some segment of getting ready show up. Moved seismograph gear isn't usually required for structure and environmental frameworks. One key striking case is the rigging for sub-base examinations or nondestructive testing of dim tops.

2.4.3 Atmospheric Data:

Atmosphere gauges are better than anything they ever have been. As shown by the World Meteorological Organization (WMO), a 5-day atmosphere figure today is as reliable as a 2-day gauge was 20 years earlier! This is in light of the fact that forecasters by and by use forefront developments to amass atmosphere data, close by the world's most pivotal PCs. Together, the data and PCs make complex models that even more definitely address the conditions of the atmosphere. These models can be altered to envision how the atmosphere and the atmosphere will change. Despite these advances, atmosphere gauges

are still often incorrect. Atmosphere is significantly difficult to anticipate in light of the way that it is an incredible and confounded system.

To make an atmosphere figure, the conditions of the earth must be known for that region and for the incorporating zone. Temperature, pneumatic power, and various characteristics of the earth must be evaluated and the data assembled.

For atmospheric Data thermometers, indicators, climate stations, radios nodes, radar, satellites, numerical weather predictions and weathers maps (isotechs, isotherms, isobars) are used.

2.4.4 Oceanographic Data

The strategies through which maritime information is gathered, are quickly expressed beneath after which more detail is examined about that acquired data.

- **CTD Rosette**

CTD remains for Conductivity, Temperature, and Depth recorder. It is an electronic instrument that ceaselessly records the saltiness (by surveying conductivity), temperature, and hugeness (by assessing weight) as the instrument is brought down on a hydro wire from the ship. The CTD is joined to an edge fitted with various wide water-gathering bottles called Niskins; the instrument together with the compartments is known as the CTD rosette. The Niskin bottles are made with the target that they have covers at the two terminations. They are sent down open, and instituted to close electronically at the noteworthiness that water should be gathered. Subsequently, water can be attempted at various profundities all through the water piece, and kept separate from water amassed in different compartments at different profundities. Aces separate the water that is amassed from the CTD for a wide extent of things, for example, oxygen, supplements, plant shades, and infinitesimal life frames wealth.

- **Microscopic fish Nets**

Minor fish, Greek for "vagabond", are little plants (phytoplankton) and creatures (zooplankton) that coast with the sea's streams. They shape the base of the created way of

life in the ocean and are essential in sea sustenance frameworks. Oceanographers use nets to get these little animals and study them. The nets are of a liberally best work over edge nets, as the work openings must be adequately insignificant to think the little fish while up to this point permitting water through. Phytoplankton nets have a little work opening (around 36/1000 of a mm) and zooplankton nets have more noteworthy cross zones (around one 1/3 to 1/2 of a mm). The nets are related with the hydro wire and towed behind the ship. Little fish tows ought to be possible at any importance or time of day, and can be utilized with opening/shutting instruments to connect with them to collect at a pined for essentialness.

- **Dregs/Residue Traps**

Development traps are cone-confined or barrel formed gatherers that get debris that sinks down from the surface sea to the remote ocean. This material is incorporated dead phytoplankton and zooplankton, the fecal matter of zooplankton and edge, and different various sorts of reject. This material, reliably named marine snow, is a fundamental sustenance hotspot for creatures that live in the remote ocean and besides a system for transporting material from the surface waters to the remote ocean where it is unquestionably debilitated by minor living things. The development traps are joined to a line that has skims at first look and a weight at the base to keep it vertical. A couple of remains traps have subsurface buoys and a base weight that really lays on the ocean base. Following two or three days or weeks, oceanographers recuperate the catches, measure the particulate material in that, and separate the material's science. The proportion of material isolated by the get-together locale and the time the gadgets were sent gives the molecule advance.

- **Fundamental Production Array:**

One essential estimation that oceanographers make is the rate of plant photosynthesis in the ocean, by and large called the rate of key creation. Amidst photosynthesis, phytoplankton take up carbon dioxide that is separated in ocean water. These little marine life outlines in this way give the section point to carbon into the marine created way of

life. Knowing how fast the phytoplankton make gives oceanographers a thought of how much carbon they take up, and how quickly.

To gauge the rate of plant photosynthesis, the authorities amass water tests just before dawn at different profundities in the major 140 meters of the sea. (Underneath this hugeness there is everything viewed as insufficient light for photosynthesis.) They by then void these water tests into direct compartments and consolidate a little extent of radioactive tracer. Adding the compartments to a line, they discharge them over the edge at the criticalness from which the water in the holder was at first gathered. At sunset, they recover this drifting showcase of holders, and channel the water to collect any little phytoplankton. They by then utilize a radiance counter to assess the extent of radioactive tracer that the phytoplankton cells have participated amidst the day. The entirety distributed the measure of hours that the holders were sent gives the rate of photosynthesis or rate of essential age.

- **The Auto-sub**

For a few purposes, mechanical vehicles give a dynamically beneficial, progressively reasonable, and progressively secure strategies for social event authentic information. The Auto-sub is a test ocean robot expected to gather information from the sea. Organizers are building up the Auto-sub for a basic number of equivalent reasons that they caused unmanned rocket to amass information from Earth to drift or from the outside of different planets. The sea surface, similar to the outside of different planets and moons, can demonstrate stunning impediments to human examination. Devouring or below zero temperatures, high breezes, and gigantic waves can now and again plot to make standard sea investigate awkward and even risky. Simply envision endeavoring to gather information from the deck of a ship while fighting 30-hitch winds, 5-meter waves, and ocean torment! The sea is in like way colossal and critical. Automated vehicles give a way to deal with research liberal districts of the sea's surface and profundities generally rapidly and monetarily. Satellites can just observe the upper couple of meters of the sea's surface.

2.4.4.1 In depth of data

There is a wide extent of oceanographic data forms. This region will base on the data that describes the physical characteristics of the ocean condition limit layer and the subsurface scattering of sea water properties. Upper air observations taken from ocean islands or pontoons can be seen as conventional meteorological discernments.

Oceanographic data is assembled using both in situ systems and remote identifying. The most apparent remote identifying stages are satellites, anyway legitimate flying machine, some novel buoys, and even a couple of pontoons use instruments (e.g., radiometers) to remotely test the ocean surface. A segment of the basic remote identifying instruments and coming about oceanographic data are: radiometers which check sea surface temperatures (SSTs), disperse o-meters which measure wave disrupting impacts and yield surface breeze paces and headings, and high precision altimeters that measure ocean surface twisting. The surface winding is used to evaluate sea surface inclinations and ocean streams. Satellite data are a vital asset for oceanographic ask about. In situ inspecting from boats and buoys does not, when all is said in done, yield satisfactory spatial or common data objectives over the colossal ocean regions that spread over 70% of the planet. Meticulously adjusted and adjusted satellite data, now and again blended with in situ observations as a data taking care of technique, give our best assessment of overall ocean conditions.

Supportive in situ ocean recognitions begin from different sources, with contrasting degrees of significant worth. The most surprising quality data are accumulated in the midst of consistent research programs, by instrumented skims (both secured and free gliding), by boats especially planned to assemble natural data, and via ocean side or island stations that limit in a manner like standard land stations. Lower quality data, yet before long critical, are oftentimes assembled on board seller sends as they cross conveyance courses, and by calculating task force vessels in the midst of business calculating exercises.

Sensible research programs assemble the broadest collection of in situ data. Common dispatch board activities will assemble sea surface data (SST, saltiness, wave stature, wave course, etc.), close surface meteorological conditions (air temperature, wind speed, wind heading, dew point temperature, barometric weight, haziness, etc.) and, routinely,

subsurface sea water characteristics (e.g., vertical profiles of temperature, saltiness, separated enhancements, deteriorated gases, anthropogenic tracers, ocean streams, and ocean base significance). Some investigation programs moreover pass on surface drifting buoys whose zones are seen by satellite. These give drift bearings (that construed surface ocean stream), and conventionally several other geophysical variables (e.g., SST, barometric weight, etc.). To a lesser degree, some free skimming buoys are arranged underneath the ocean surface. These buoys are pursued acoustically or they once in a while climb to the surface for satellite after. Buoys of this create are used to screen subsurface sea stream and likewise subsurface sea water properties. Secured surface buoys with subsurface instruments underneath are in like manner used by science programs. The surface instrumentation assembles various sorts of data appropriate to ocean condition limit layer frames, while the subsurface instruments routinely revolve around water temperature, saltiness, weight, and ocean streams.

One logical program that is contributing significant information for atmosphere examines is called TOGA/Tropical Atmosphere Ocean (TAO). A littler program started around 1980 and has developed and ventured into the present-day TOGA/TAO which covers the central belt of the Pacific Ocean with surface and subsurface instrumentation fixed to 30-50 secured floats (Fig. 3.1). In close constant mode, information from these instruments are gathered by satellite and are utilized for worldwide climate and sea condition gauges. Amid occasional instrument administration and fix, this informationis likewise gathered in a deferred mode. Following post testing adjustment of the instruments, the deferred mode information is quality checked and rectified. Both the close continuous and postponed mode information are significant for checking commonplace sea conditions, for example, El Nino.

Information gathered on vendor and angling vessels are a huge hotspot for surface oceanographic information. Commonly, this informationis assembled at succinct climate watching times on board dispatches that are in travel. Sailors have done this through history, obviously, with generally shifting strategies and degrees of exactness. The run of the mill estimations are wind speed and heading, barometric weight, air temperature, SST, and neighborhood climate conditions. In the most punctual occasions, this

informationwas recorded by submit logbooks. Some verifiable logbook information has been digitized and now the most punctual advanced records are from the mid 1800's. Present day vessels utilize computerized frameworks whereby the information is gathered carefully and transmitted by means of satellite to arrive based accumulation organizations. This information gives basic data to introduce day climate and sea condition anticipating. All things considered; huge measures of information are still just recorded in logbooks. A few information documented projects are in advancement to digitize more logbook information. Given the tremendous locales of the sea, and the generally meager inspecting that happens, practically any accessible information is viewed as helpful.

There are numerous auxiliary oceanographic information types that are significant. Ocean level, ocean ice focus, and sea base geography are a couple of models.

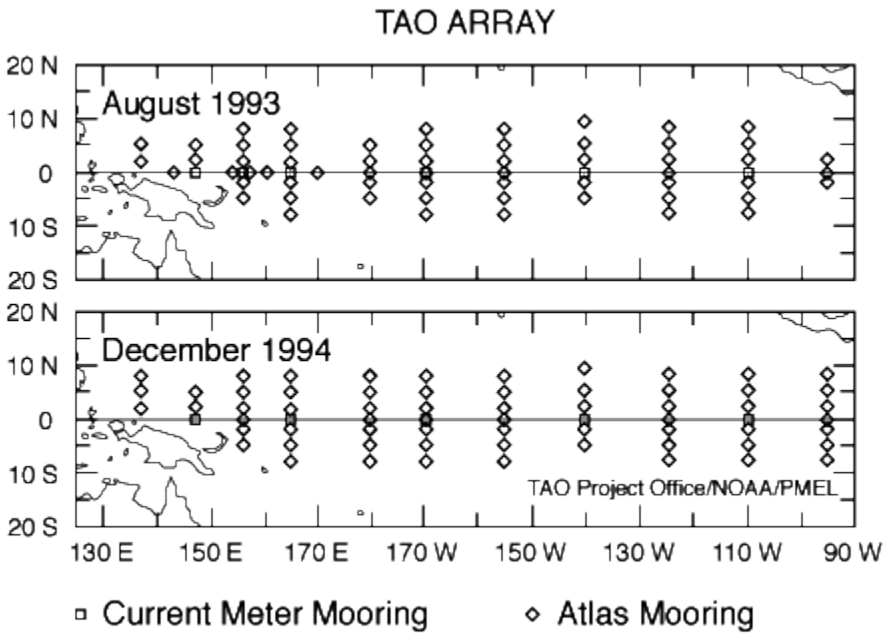


Figure 1 The TOGA/TAO array buoys in August 1993 and the final configuration in December 1994. (From McPhaden, 1993)

2.4.5 Analysis of various methodologies for collecting natural data

In the light of above all discussion, it can be concluded that given the present technology, data can be gathered in any kind of environment, whether it be the data of glaciers, mountains, weather, population or genetics. Moreover, data analytics is not only confined to natural things, it is also actively applied in stock trading, forex trading, Machine Learning and other technological advancements which are breaking through into a new, more sophisticated era of technology. It has become a helping hand for the present human being, helping him in making tough and complicated decisions for the betterment of environment and of the whole mankind.

2.5 Stratigraphy

A part of geography which is worried about investigation of shake layers implies strata and layering which mean stratification is known as Stratigraphy. Stratigraphy is utilized in the investigation of sedimentary and layered volcanic rocks. It has two comparable branches:

- Litho-stratigraphy
- Bio-stratigraphy

A logical order which is worried about portrayal of progressions and its understanding as far as their general time scale is known as Stratigraphy. Oil and archaic exploration are the fields wherein its standards and strategies have found and it likewise gives premise to chronicled geography.

The investigation of Stratigraphy manages the sedimentary shakes yet it might likewise encompass by layered molten rocks like which are coming about because of successive magma streams or transformative rocks which are shaped from extrusive volcanic material or sedimentary rocks.

Objective of stratigraphic think about is the subclass of an arrangement of shake strata into mappable units which decide time connections included and corresponding units of grouping or all succession with shake strata. The International Union of Geological Sciences (IUGS; established 1961) built up a Commission on Stratigraphy to work after

some fizzled endeavors amid the last 50% of nineteenth century of International Geological Congress (IGC;founded 1878) to systematize a stratigraphic scale.

Stratigraphic plan relies upon these two scales:

- Time scale (utilizing eons,periods,ages,eras and chrons), and every unit is characterized by its start and completion focuses.
- Correlated size of shake arrangements which use systems,stages,chronozones and arrangement. All these when are utilized related to some other dating procedure like radiometric dating (which is the estimation of radioactive decay),paleoclimatic dating and paleomagnetic judgments which were created in last 50% of twentieth century and they prompted some less disarray of classification and to progressively substantial data on which ends are base about Earth history.

Because gaseous petrol and oil, they nearly happen each time in stratified sedimentary shakes so the method to find oil repository traps has been encouraged with the utilization of stratigraphic ideas and information. Law of superposition – In this rule in any undistributed store the most established layer is situated at least dimension and it is normal that remaining parts of each succeeding age are left on trash of last. This guideline (law of superposition) is viewed as significant in the use of stratigraphy to archaic exploration.

2.5.1 Big data

Now a day's data is in petabytes or in exa-bytes. This large amount has data should be treated in a proper way to make it useful for human beings and for all sectors. The fields which includes data cleaning pre- processing analyzing of the data and converting complex problems into simpler ones is Big Data.

Challenges which are include for Big data are data storage, querying, transfer, visualization, data analysis, capturing data, search, data source, updating and information privacy.

Big data is related with these three characteristics:

- Volume
- Velocity
- Variety.

2.5.2 Cloud computing

Circulated processing is generally what makes the PC system resources, figuring power and particularly storing get open on enthusiasm with no prompt unique organization by its customer. Appropriated figuring is used to portray server cultivates that are open to a huge number of customers on the Internet. Huge fogs are noteworthy and transcendent today and they habitually have those limits that are dispersed on various different regions from the central servers. It is mark as an Edge server if the relationship with the customer is correspondingly close.

Fogs can be bound to a singular affiliation (like endeavor fogs), get available to a huge amount of affiliations (like open cloud) or a blend of these both which is known as mutt cloud. Amazon AWS is the greatest open cloud.To achieve soundness and economies of scale, conveyed registering relies upon the sharing of benefits.Conveyed registering empower firms to maintain a strategic distance from or to reduce direct front IT structure costs, this was noted by the support of open and mutt fogs. Supporters moreover articulate that dispersed figuring in like manner offered firms to get their applications an opportunity to completely operational faster which are with the improved reasonableness and less upkeep so on account of which it enables IT gatherings to modify even more quickly the benefits so that to fulfill changing and unforeseeable need. "pay-as-you-go" model is being used by the cloud providers which lead to working costs which are not expected if the regulators are not acclimate with cloud-esteeming models.For the advancement in dispersed figuring, the availability of these is required: ease PCs, high-limit frameworks, wide gathering of gear virtualization, organization orchestrated structure, amassing devices and autonomic and utility enlisting.

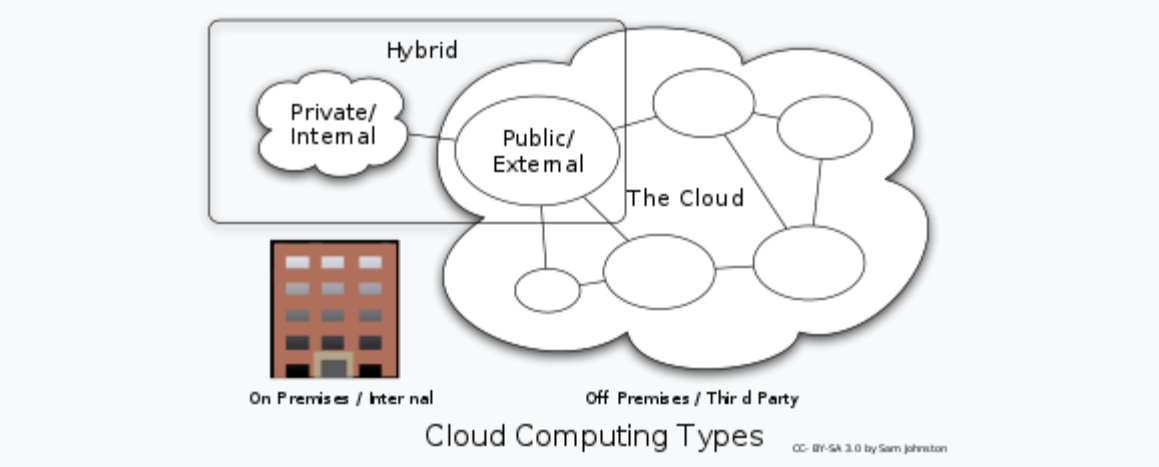


Figure 2 Deployment Model

There are kinds of Cloud Computing that are listed below:

- Private Cloud
- Public Cloud
- Hybrid Cloud
- Community Cloud
- Distributed Cloud
- Multi Cloud
- HPC cloud
- Big Data cloud

2.5.2.1 How does cloud computing work?

Cloud computing works in a way where there are bulk of servers and data bases arranged to get maximum utilization from the internet. There are sites and organizations such as Amazon Web Services that updates the system daily so that customers do not face problems and keep the business scale grow. They also deliver static files through CDN which is Content Delivery Network.

2.5.2.2 Benefits of cloud computing

Consumers and organizations have many different reasons for choosing to use cloud computing services. They might include the following:

- Convenience
- Scalability
- Security
- High availability
- Low costs
- Reliability

2.5.3 Different wellsprings of mistake in stratigraphy

Stratification happens in light of the fact that the surface or type of the minerals, called silt, changes as they settle down into shake. Weight, warmth, and compound responses change the dregs into rocks. This procedure is critical to the stone cycle. Strata can be various hues as a result of it, as well.

Mistakes in stratigraphic estimation might redress or non-redressing. Repaying mistakes don't influence the mean estimation of the trait under examination, yet they do expand the clear inconstancy of the observational information.

Non-repaying mistakes influence the obvious mean esteem, and they may mutilate facies inclinations or trends. The significance of operational definitions in stratigraphic estimation is pushed, and the properties of the subsequent numbers are inspected

regarding the sizes of estimation included. Exactness and accuracy, as they apply to stratigraphic information, are addressed. A sober minded strategy for assessing facies maps as foreseeing gadgets is portrayed. This technique considers all wellsprings of inconstancy in the first guide, including mistakes of estimation and judgment, yet the common huge scale and little scale stratigraphic varieties too. A precedent is utilized to represent the normal size of the vulnerability in forecasts from a particular facies map.

2.5.4 Importance of correct data obtained in sensor networks

Sensor systems are by and large progressively utilized in a few application regions, especially to gather information and screen physical procedures. Non-useful necessities, similar to unwavering quality, security or accessibility, are regularly significant and must be represented in the application improvement. For that reason, there is a huge assemblage of learning on trustworthiness strategies for disseminated frameworks, which give a decent premise to see how to fulfill these non-utilitarian necessities of Sensor Network based observing applications.

The information driven nature of checking applications, it is of specific significance to guarantee that information is solid or, all the more conventionally, that they have the fundamental quality.

CHAPTER # 3
REQUIREMENT ANALYSES

3. Overall Description

3.1 Product Perspective

We are developing software prototype in Python which would work on specific ML-model that can predict the missing data values using techniques of statistics and software engineering. While prediction of missing data from a data set is easy if the data is linear or two dimensional. But in real life scenarios, data is nonlinear and complex, thus making the prediction of missing values more challenging. Our methodology can be applied to many other problems in the similar domain.

3.1.1 Product features

The following are the basic *features* that will be provided to the users by our software prototype.

- **Acquisition of Dataset**

BP will be providing us the required data which is 32 GB in size.

- **Analyzing of Datasets**

We will analyze the dataset using our specific ML techniques.

- **Predict the Missing values**

Prediction of lost data or noisy data so that we can visualize dataset.

- **Visualize Datasets**

We can visualize datasets using Scatter plot, Box plot and Scree plot. It will help us to classify/group the data, which will produce better and meaningful results.

- **Deliverables**

Prediction model will be implemented for BP upon completion.

3.1.2 Characteristics

Our prototype software aims to elucidate on several approaches available for handling missing values as in real world datasets provided by BP are rarely clean and homogeneous. Data can either be missing during data acquisition through Wireless Sensor Networks (WSN) or transformation. Missing values need to be handled because they reduce the reliability, effectiveness and quality data visualization. Such is the case with BP, which needs this lost data to predict future oil prices as well as the oil reserves left at a respective location. Missing data values can lead us to wrong prediction or classification and can also cause a highly biased results for any applied prediction model. While our proposed model intends to reciprocate the ambiguities of data predictions and tends to give unbiased and transparent insights.

3.1.3 Design and Implementation

- **BP Dataset**

Look at a dataset that has known missing values.

- **Marking of Missing Values**

Marking of missing values from dataset

- **Missing Values Causes Problems**

Missing values causes problems by fling the machine learning algorithm

- **Ruling Out Rows Having Missing Values**

Rows that contain missing values. They can be removed from the dataset.

- **Imputation of Missing Values**

Replacing the missing values with some other value(mean etc.)

- **Algorithms that Support Missing Values**

Not all algorithms support missing values but there are some that supports.

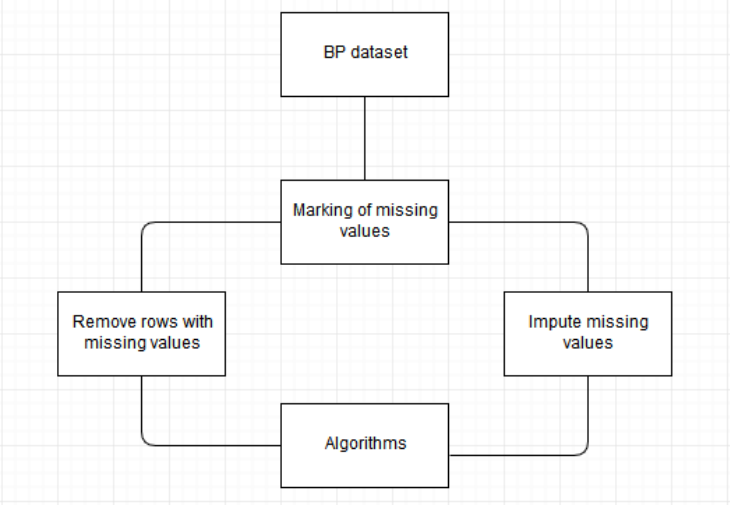


Figure 3 Design and Implementation

- ❖ Following techniques will be used to explore which one is fit for prediction in our case. (Move to Appendix B)
    a. Regression (Simple linear, Multiple linear, Polynomial, Support Vector Regression)
    b. Classification (K-Nearest Neighbor , Support Vector Machine)
    c. Clustering (K-Means)
    d. Principle Component Analysis (Non Linear Principle Component Analysis)

1. We shall be able to visualize the processed data in following formats: (Move to Appendix C)
    a. Scree Plot
    b. Box Plot
    c. Scatter Plot

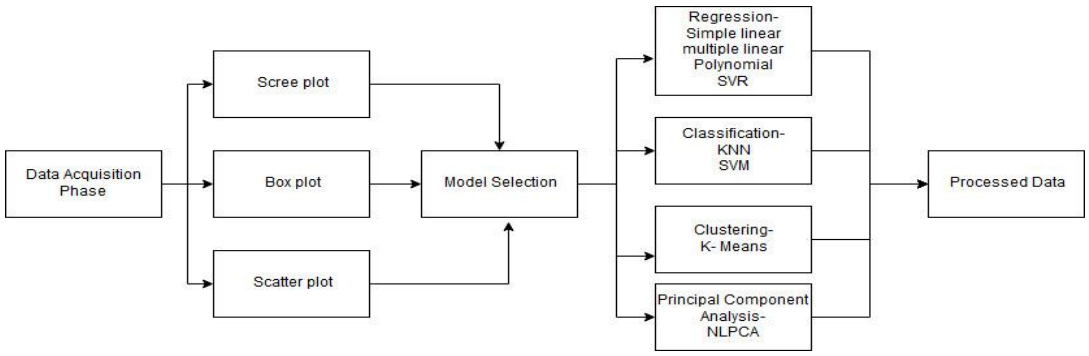


Figure 4Design and Implementation

3.2 External Interface Requirements

- Fetch dataset having missing values from user interactively.
- Prediction technique with the most accuracy will be implemented in software prototype.
- Give graphical visualizations of BP dataset.
- Optional – if extensive computational power is made available, we may deal with big datasets.

3.3 System Features

3.3.1 Functional Requirements

1. The prototype shall be able to fetch the dataset having missing values.
2. The model shall be able to predict the missing values from the dataset provided by BP.
3. Functional Requirements for the project Missing Data Prediction from WSN by system features:-
    - **Dataset Acquisition**

The system to acquire dataset provided by BPwhich is 32 GB in size.

- **Check Missing Values**

The system will check for the missing values in the given dataset.

- **Visualize dataset**

The system will visualize dataset by using scatter plot, box plot, and scree plot, which will help us to classify/group the data for meaningful and better results.

- **Data Prediction Model**

The prediction of missing values will be done using the data prediction model and software prototype.

3.4 Other Non-Functional Requirements

- **Computation**

If extensive computational power is made available, we may deal with big datasets.

- **Reliability**

Our ML-models results would be reliable

- **Accuracy**

Our ML-models results would be accuracy

- **Optional Generic**

We can extend our proposed model or apply it on other datasets.

CHAPTER # 4

SYSTEM DESIGN

4.1 Introduction

Our aim is to devise a model which is able to predict the missing data as well as the noisy data accurately. After we develop a working model of data prediction, that same model is going to be implemented for dataset provided by British Petroleum (BP).

4.1.1 Purpose

The purpose is that we get to know about the components and modules and their relations in our developed prototype. It will further tell us the interrelationships between subsystems. We will get a clear understanding of system architecture and structures. This document also serves the reusability in terms of product design, manufacturing and implementation. In addition to that, design decision and pseudo code for components are also mentioned in detailed.

4.1.2 Scope of the development project

Prototype of software based upon the ML- model that we will develop as a deliverable. It would be beneficial for BP to get good insights of their company progress. As well as, we can extend it to other companies by training our proposed ML-model on their datasets in future perspective.

4.1.3 Definitions, acronyms, and abbreviations

1. British Petroleum (BP)	7. Classification	8- Simple Linear Regression
2. Python	8. Clustering	9-Support Vector Regression
3. Statistical Techniques	9. Principle Component Analysis (PCA)	10-K-Means
4. Machine Learning (ML)	10. Scatter plot	11-K Nearest Neighbor
5. Deep Learning	11. Box plot	12- Kalman Filter
6. Regression	12. Scree plot	

4.1.4 Overview

This document tells us the about a model which is able to predict the missing data as well as the noisy data accurately. The working model we built will later be used by British Petroleum (BP). Its first section tells us about the introduction of the document. Section 2 tells us the modules and components of how our prototype will be structured. Its relationship and its working are further being explained. Section 3 further add more detailed overview and explanation of the prototype and its components. Section 4 tells the reuse and relationship to other products. It will provide us the answer that is the model being reused or not. Section 4 tells us the design and trade- offs and last section will represent the pseudo code for the components.

4.2 System architecture description

4.2.1 Overview of modules / components

Following are the five modules/ components of how our prototype will be structured.

- **Acquisition of Dataset**

BP will be providing us the required data which is 32 GB in size.

- **Analyzing Datasets**

We will analyze the dataset using our specific ML techniques.

- **Predict the Missing values**

Prediction of lost data or noisy data so that we can visualize dataset.

- **Visualize Datasets**

We can visualize datasets using Scatter plot, Box plot and Scree plot. It will help us to classify/group the data, which will produce better and meaningful results.

4.2.2 Deliverables

Prediction model and software prototype will be implemented for BP upon completion.

4.3 Structure and relationships

After acquiring the data, we will visualize it in following formats. (Appendix A)

a. Scree Plot
b. Box Plot
c. Scatter Plot

These visualizations give us better idea of our data and it further help us in choosing a technique. The techniques that we are working on are following. (Appendix B)

a. Regression (Simple linear, Multiple linear, Polynomial, Support Vector Regression)
b. Classification (K-Nearest Neighbor, Support Vector Machine)
c. Clustering (K-Means)

d. Kalman Filter

e. Principle Component Analysis (Non-Linear Principle Component Analysis)

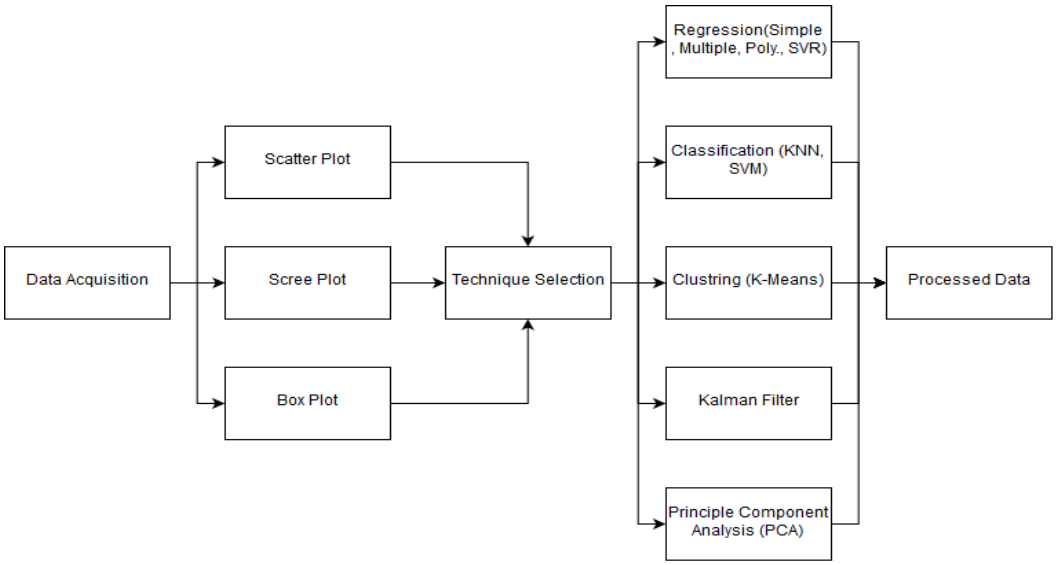


Figure 5 Structure and relationships

4.3.1 Data flow diagrams of the missing data prediction prototype

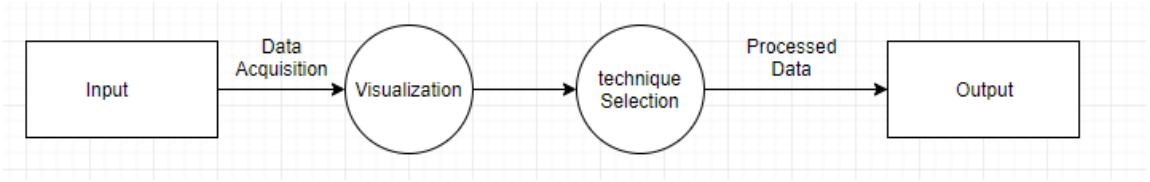


Figure 6 DFD Level 0

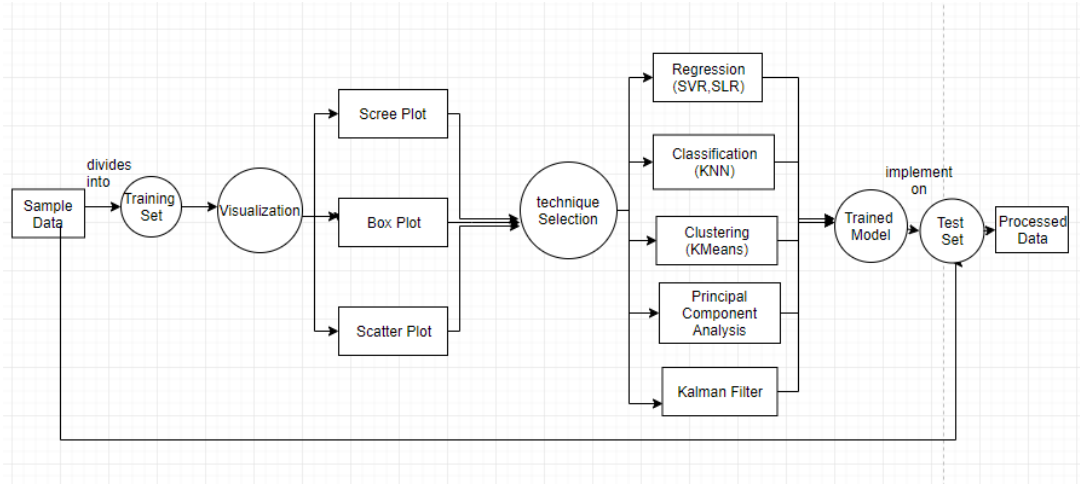


Figure 7 DFD Level 1 (Overall)

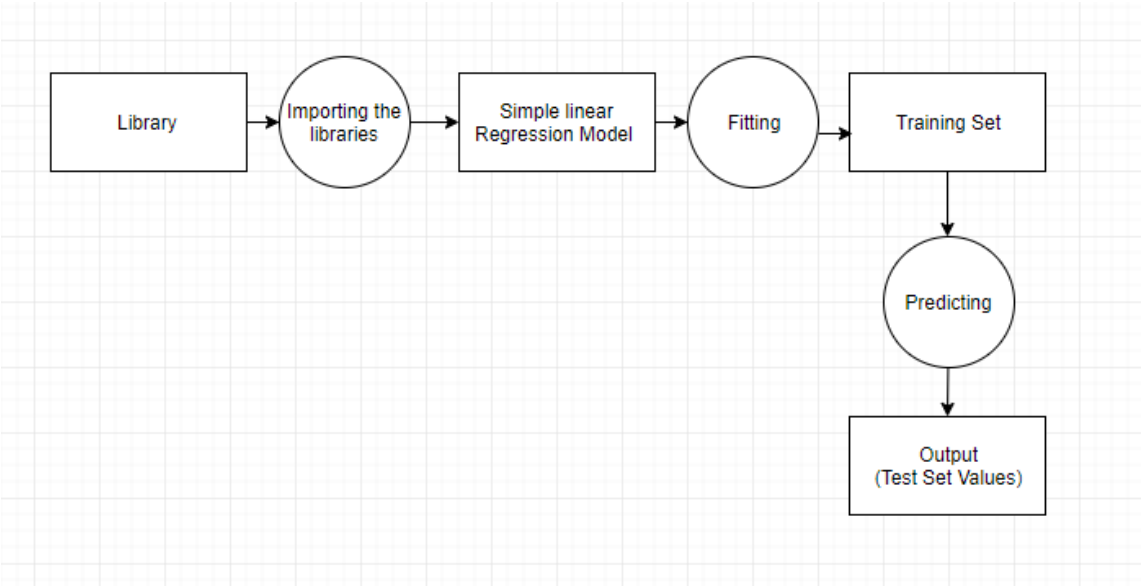


Figure 8 DFD level 1(Simple linear regression)

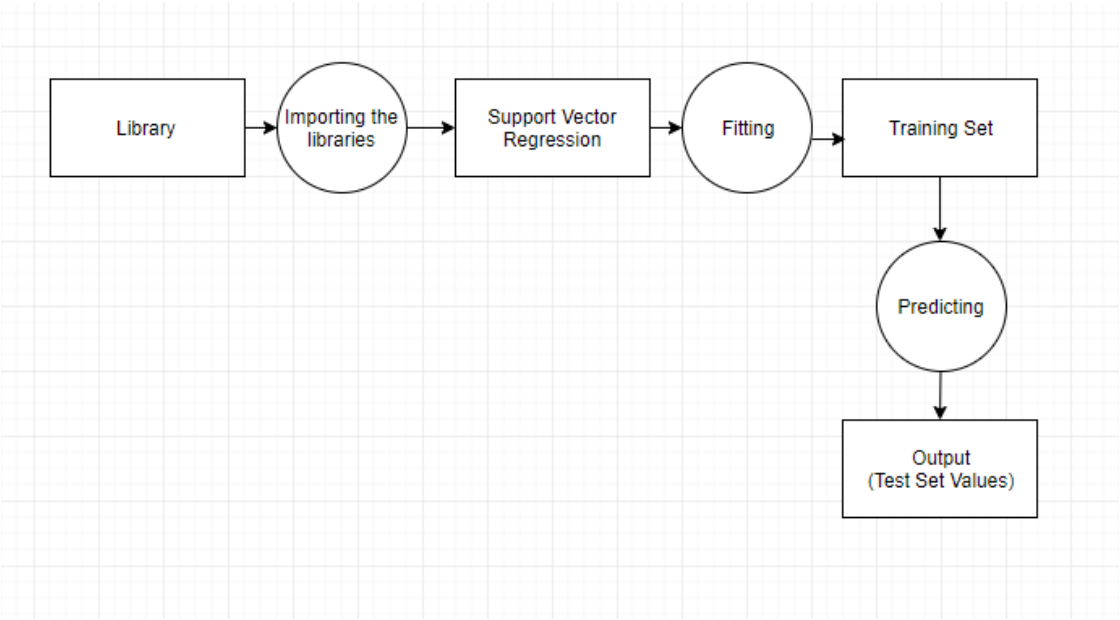


Figure 9 DFD level 1(Support Vector Regression)

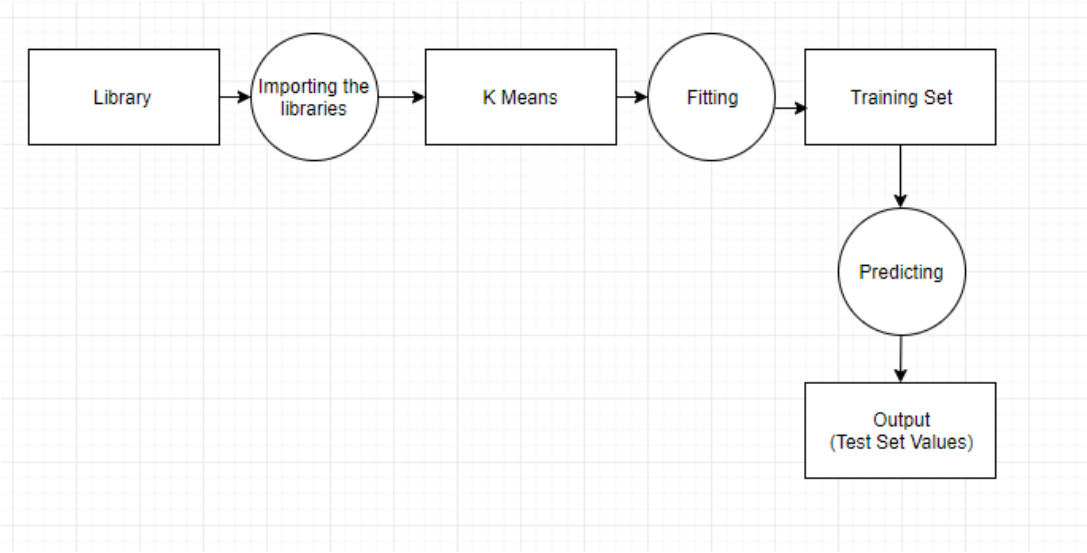


Figure 10 DFD Level 1 (K-Means Clustering)

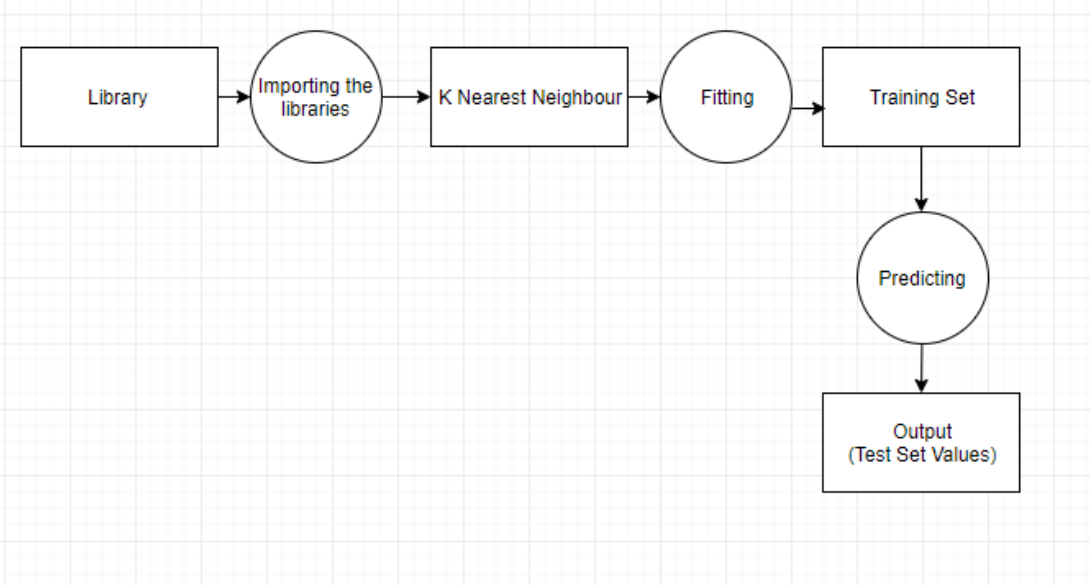


Figure 11 DFD Level 1 (K-Nearest Neighbor)

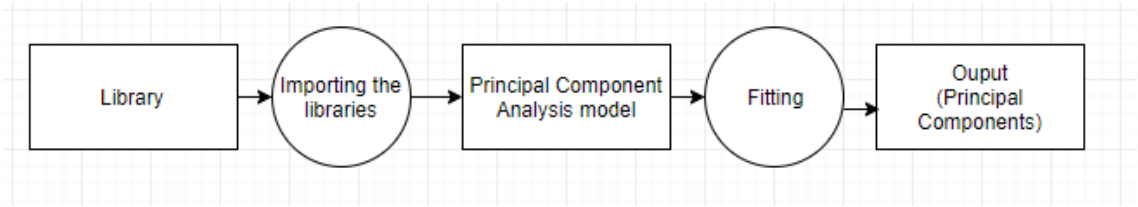


Figure 12 DFD Level 1 (Principal Component Analysis)

4.4 User Interface

To begin the process, we can import any file in ".csv" format. Our model will work with that uploaded file and give us the visualization of the uploaded data file. Our output will be the visualization of the predicted value and the estimated error.

Our model will generate the data (including predicted values) with minimal error and it would help the British Petroleum to get the better insights and analytics, which leads to more appropriate decision making.

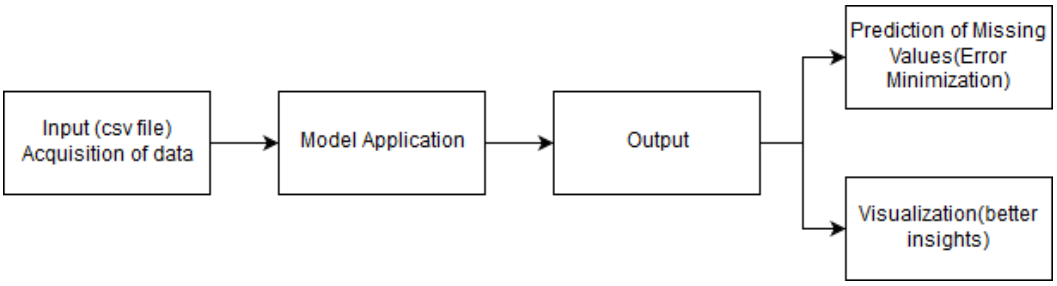


Figure 13 User Interface

4.5 Detailed description of Module/Components

4.5.1 Description of Components

The acquisition of data, that is in the form of ".csv" file is obtained using our input component and processed by our Model Application Component, that apply our model to predict missing values and pass it to our output component to produce the end results.

- **Input Component**

Input Component can acquire the data in ".csv" file format and pass it to the Model Application Component.

- **Model Application Component**

Model Application Component analyses the data and predict values for us to minimize noisy measurements.

- **Output Component**

The Output Component generates the visuals of data that is being predict by our model and provide us better insights with much accuracy.

4.6 Reuse and relationships to other products

Our model can be reused in terms of datasets. In first phase, there is acquisition of data. We are taking an input by importing the .csv file. We then apply different models on it and predict the values. The best predicted values and estimated errors are selected.
These machine learning models once developed, can be used with any dataset by providing the respective ".csv" file. We can reuse our model for any dataset according to our requirement.
This model will later be used by the British Petroleum for their datasets to predict oil rigs and prices of the oil in later stages.

4.7 Design decisions and tradeoffs

Our aim is to devise a model which is able to predict the missing data as well as the noisy data accurately. After we develop a working model of data prediction, that same model is going to be implemented by British Petroleum (BP). According to the market point of view, Data Science is a vast field and it's in demand. Companies are focusing primarily and investing large portions on creating better data analytics teams and professions. This will help to analyze the data effectively so that the better strategies can be carries out. Demand of data science is very high because it aid companies and organizations to give enhanced insights as well as make better decisions via analyzing the historic data. It will reduce the cost and make improved strategies for future growth.

4.8 Why We Prefer PYTHON over MATLAB

As there are advantages given below that will clearly tell about the importance of python over Matlab.

- **PYTHON Code id readable and compact**

a. PYTHON use indentations to determine the scope of block statement.
b. For indexing and brackets of function and method calls. PYTHON uses square brackets. This helps the code in readability.

c. PYTHON uses libraries of Numpy for arrays methods of min, mean, max etc. and doesnot use arbitrarily limits the usage of literals in expressions.

- **PYTHON uses zero based indexing**

PYTHON uses NUMPY library for arrays and to mark first value lets sat x it will be x[0] ,not x[1].So it uses zero based indexing.

- **PYTHON offers dictionaries- hashes**

Dictionaries are used to implement a symbol table. They are used in various institutions of engineering as well as scientific programming. Example of dictionary is given below.

dictionary =

```
{
  "brand": "Mehran",
  "model": "Suzuki",
  "year": 1964
}
```

- **PYTHON is Object oriented programing language**

PYTHON is Object oriented programing language and it is simple and elegant.

a. When class or method is set to private then there is a property isSet Access and Get Access.
b. In all super classes the classes having same definition.

- **PYTHON is free and opensource**

a. Python is free. From its software to its libraries they are all open source.
b. The licenses of Matlab are expensive. However, PYTHON is free and open source.There are libraries such as NUMPY, SCIPY, MATPLOTLIB, IPYTHON and various others that are available in python.

- **Any number of functions can be packaged in one file (module)**

MATLAB program documents can contain code for more than one capacity. The primary capacity in the document (the principle work) is unmistakable to capacities in different records, or you can call it from the direction line. Extra capacities inside the document are called neighborhood capacities. Nearby capacities are just noticeable to different capacities in a similar record.

Since neighborhood capacities are not available outside of the record where they were characterized, designers dealing with complex Matlab-based undertakings will in general be overpowered by a plenty of small files.

- **Python's import Statement**

Python enables one to sort out classes and capacities into modules and bundles, with the module or bundle name being utilized to determine any name clashes. Python's import order gives one exact power over what segments are utilized by any program.

- **Python offers more choices in graphics packages and toolsets**

a. Matplotlib delivers fantastic non-intelligent 2-D and 3-D diagrams, and is more than sufficient for most designing and logical illustrations. Creating production prepared yield with Python and matplotlib requires less tweaking than with Matlab.
b. Graphical UIs can be made utilizing Qt, Traits, or Wx. (Wx is bit by bit being eliminated).
c. Chaco gives an API to making intuitive diagrams.

The yield it will provide for us will be tasteful with having less blunder. Information representation is and information examinations will be founded on information inputs.

4.8 Pseudo code of components

- **Simple linear regression**

We have used the dataset in which we have different x and y values. Screenshot of some of it is as:

	A	B
1	X	Y
2	6.2	29
3	9.5	44
4	10.5	36
5	7.7	37
6	8.6	53
7	34.1	68
8	11	75
9	6.9	18
10	7.3	31
11	15.1	25
12	29.1	34
13	2.2	14
14	5.7	11
15	2	11

Figure 14 Dataset Sample 1

We have imported different libraries in python as shown below:

```
In [22]: import numpy as np
    ...: from scipy.stats import norm
    ...: import matplotlib.pyplot as plt
    ...: import pandas as pd
    ...: import seaborn as sns

In [23]:
```

Figure 15 Output of import statements

Then data is imported and divide as training and test sets:

Name	Type	Size	Value
X1	float64	(42, 1)	array([[6.2], [9.5],
X_test	float64	(14, 1)	array([[9.], [3.4],
X_train	float64	(28, 1)	array([[4.8], [7.2],
Y_test	int64	(14, 1)	array([[39], [17],
Y_train	int64	(28, 1)	array([[19], [29],
dataset	DataFrame	(42, 2)	Column names: X, Y
y1	int64	(42, 1)	array([[29], [44],

Figure 16 Training and test sets

Before applying any model will visualize it:

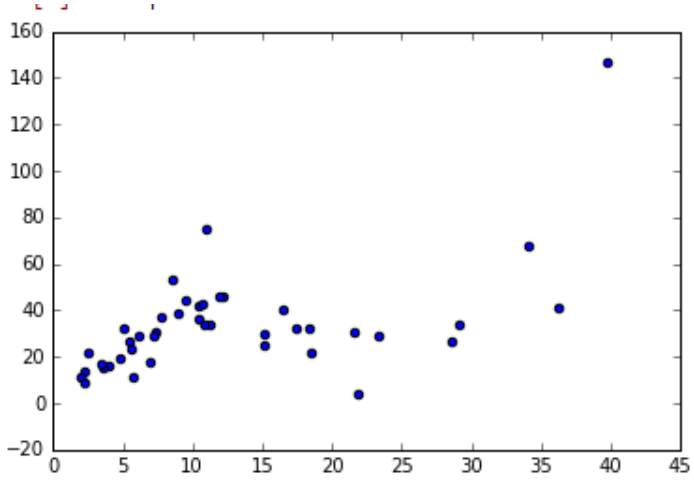


Figure 17 Scatter Plot

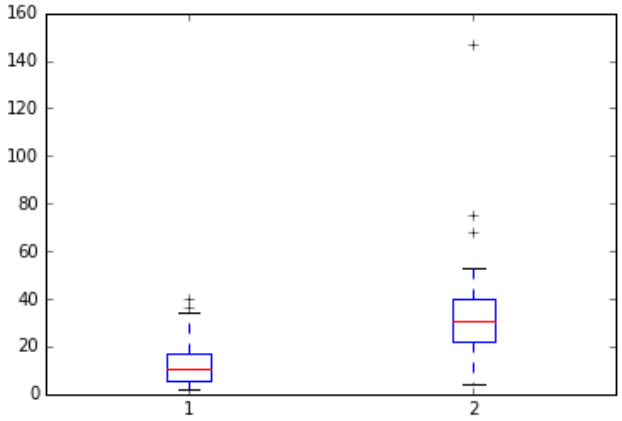


Figure 18 Box Plot

a. Implementation of Simple Linear Regression

We applied Simple linear regression model on it:

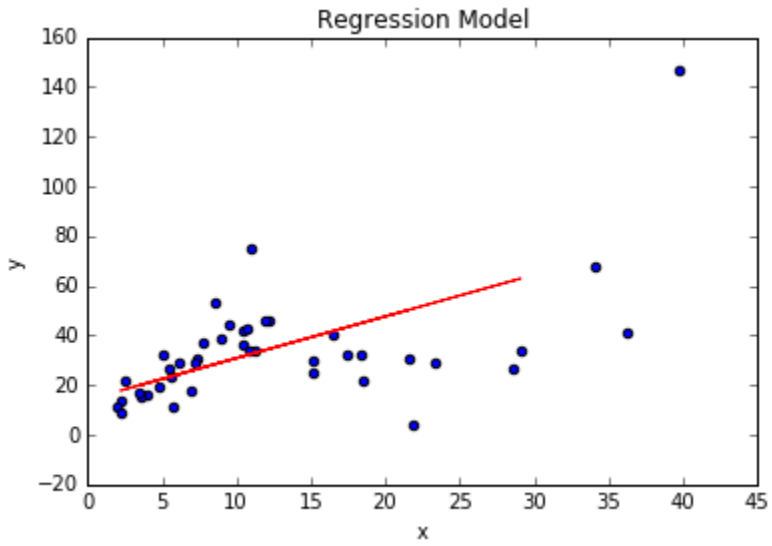


Figure 19 Simple Linear Regression Model

b. Comparison of actual and predicted values

Here **Y_test** is our actual values and **Y_pred1** is our predicted values.

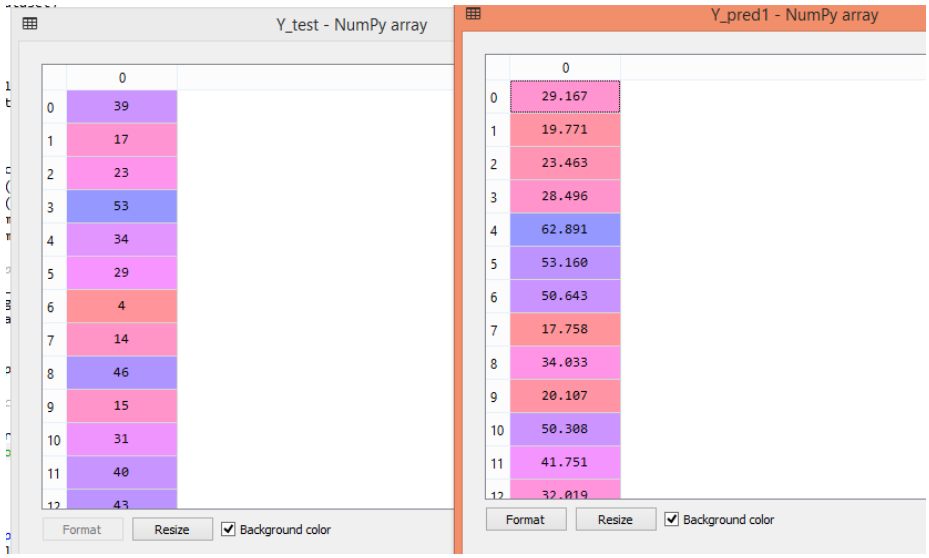


Figure 20 Comparison of Actual and Predicted Values

- **Principal component analysis**

Here we take the data of which screenshot is attached below.

	A	B	C	D	E	F	G	H
1	Relative Measures	Iter-1	Iter-2	Iter-3	Iter-4	Iter-5	Iter-6	Iter-7
2	Code Size	15	28	13	26	72	79	87
3	Execution Time	-6	-14	-19	-50	-66	-73	-80
4	EnergyConsump	-1	-8	-4	-14	-19	-21	-23
5	Slotutilization	17	19	54	45	64	70	77
6	SchedulingFactor	4	4	10	17	36	40	44
7	Highwayusage	94	182	221	319	327	359	395
8	InstruCacheMiss	-6	-13	-9	-18	-28	-30	-33
9								

Figure 21 Data Sample 2

We have used the following libraries.We have imported the dataset and make all iterations as our X and Relative measure as our Y.

Before visualizing it in scree plot and applying PCA on the dataset, we have performed Feature Scaling such that they appear on one scale.

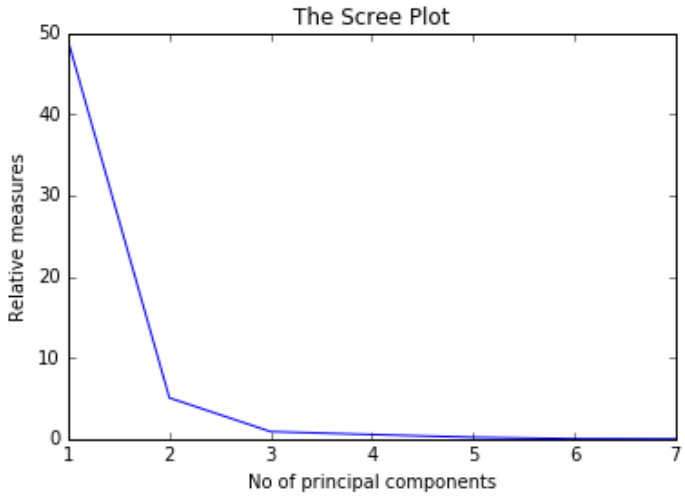


Figure 22 Scree Plot

a. Explained_variance

We get 2 principal components. PCA can also be find out by explained variance ratio **explained_varance=pca.explained_variance_ratio_**

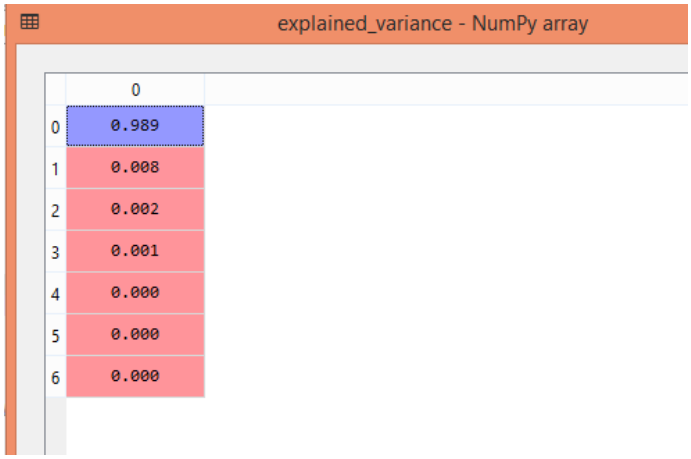


Figure 23 Explained Variance

This shows principal component 1 has 0.989 effect and principal component 2 has 0.008+0.989 impact others PC3, PC4, PC5, PC6 has no impact.

b. Principal components

There are two principal components that we have find out.

Index	principal component 1	principal component 2	0
0	-0.0514	0.394	Code Size
1	-2.25	-0.359	Execution Time
2	-1.52	-0.0923	EnergyConsump
3	0.103	0.0302	Slotutilizat…
4	-0.7	0.25	SchedulingFa…
5	6.14	-0.145	Highwayusage
6	-1.72	-0.0782	InstruCacheM…

Figure 24 Principal Components

- **Support vector regression (SVR)**

SVR is a regression algorithm that is used for continuous values. It is kind of SVM (Support Vector Machine)that distinguish between linear and non-linear regressions.
In SVR we try to fit the error within a certain threshold or those points which have least error rate. Thus, giving us a better fitting model.
Goal in SVR is to make sure that errors do not exceed the threshold.

a. Aim of using SVR

We are using SVR for regression problems. SVR tries to fit a line to data by minimizing a cost function. The interesting part of SVR is that we can deploy a non-linear kernel. In this case we end up making non-linear regression, i.e. fitting a curve rather than a line.

Our Aim when we are moving on with SVR is to basically consider the points that are within the boundary line. Our best fit line is the line hyperplane that has maximum number of points.

b. Applying SVR practically

We have created data in which there are X and Y values to apply Support Vector Regression on it. We have performed our task in Python. First of all, we imported different libraries in Python to proceed further.The next step is: Data is imported and Feature scaling is applied on it. This is so because SVR is a less common model, so we used feature scaling here. Following is the screenshot of after imported data and applying feature scaling on our dataset after successfully execution in variable explorer. Before applying Support Vector Regression model on our dataset, we visualize the data The Scatter plot applied on our dataset and the plot we obtained in Python is as shown on the next page.

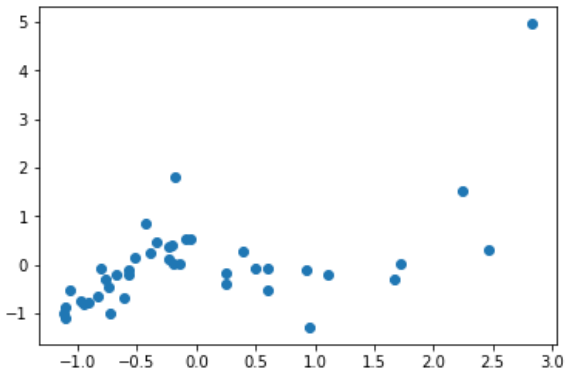


Figure 25 Scatter Plot of data sample 3

The Box plot applied on our dataset and the plot we obtained in Python is as:

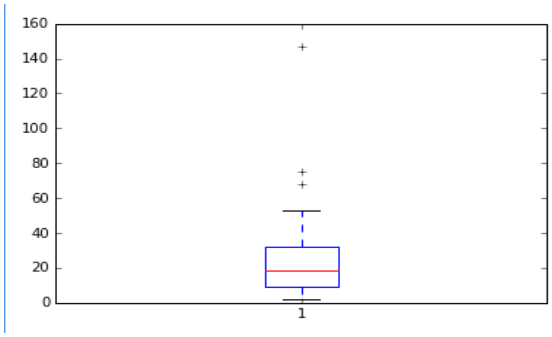


Figure 26 Box Plot of data sample 3

c. Implementation of SVR

Here we want to fit a model to predict a quantity for future and we want the data point (observation) to be as close as possible to the hyperplane.

Applying SVR on the dataset using “SVR” class and “svm” library by making objects of SVR class and using the “rbf” kernel.

The graph obtained after applying SVR on the dataset is as shown on the next page.

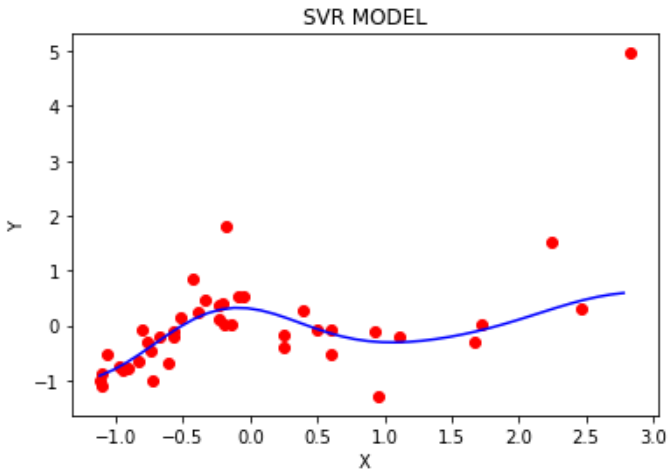


Figure 27 Support Vector Regression Model

The red dots obtained are the observation points and blue line obtained shows the predicted points.

d. Predicted value

Here we predict the value of X1=29.1 using SVR. Below is the one line code for y_pred which we have applied in Python.

Below is the output of y_pred, shown in variable explorer.

Figure 28 Comparison of Actual and Predicted Values of SVR

- **K-Nearest neighbors**

KNN is a calculation that is helpful for coordinating a point with its nearest k neighbors in a multi-dimensional space. It tends to be utilized for information that are continuous, discrete, ordinal and categorical which makes it especially valuable for managing all sort of missing information.

The assumption behind utilizing KNN for missing qualities is that a value can be approximated by the estimations of the values that are nearest to it, in light of different factors.

a. Advantages

- ✓ Robust to noisy training data.
- ✓ Effective if training data is large.

b. Disadvantages

- ✓ K must be defined by us.
- ✓ Distance method must be defined by us e.g. Euclidian Distance, Minkowski distance etc.
- ✓ Computation cost is high as we need to compute distance of each query instance to all training samples.

c. Basic understanding:

To get a basic understanding of how KNN works we can summarize the whole process in 4 simple steps:

- ✓ Assume a value for k, but it should be kept in mind that;

$$k\ \alpha\ \frac{1}{Noise}$$

- ✓ Take k = 5 nearest neighbors of the new data point, according to Euclidian distance.
- ✓ Among these K neighbors, count the number of data points in each category.

- ✓ Assign the new data point to the category where you counted the most neighbors.

d. Implementation of model

- ✓ **KNN Parameters Calibration:**

  Following are the parameters that are taken into consideration for our KNN classifier in python;

  There are number of neighbors to use by default for queries. The algorithms we used areBall Tree, KD Tree, Brute Force and Auto (determine the algorithm itself based on the dataset)

- ✓ **P:**

  Power parameter for the Minkowski metric. When p = 1, this is equivalent to using Manhattan Distance, and Euclidian Distance for p = 2. For arbitrary p, Minkowski Distance is used.

- ✓ **Metric:**

  The distance metric to use for the tree. The default metric is minkowski, and with p=2 is equivalent to the standard Euclidean metric.

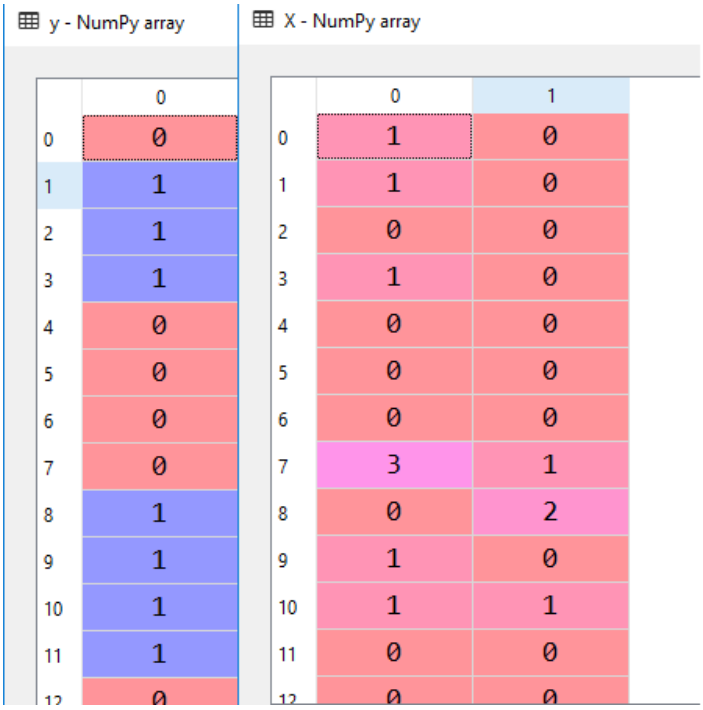
e. Implemented example:

Here in this example, we used the "Titanic" dataset to predict which of the passengers survived the incident and vice versa. We have a test data set which is 100% complete so that we can compare the predicted results with a confusion matrix. Our dataset has the following variables;

- ✓ Passenger ID
- ✓ Survived (0 or 1)
- ✓ PClass
- ✓ Name
- ✓ Sex
- ✓ Parch
- ✓ SibSp
- ✓ Ticket
- ✓ Fare

- ✓ Cabin
- ✓ Embarked

Now we are going to split this dataset into 2 subsets. X contains the independent variables (SibSp and Parch) and we have assumed "Survived" to be the dependent variable y. Both the datasets are shown as below:



	0
0	0
1	1
2	1
3	1
4	0
5	0
6	0
7	0
8	1
9	1
10	1
11	1
12	0

	0	1
0	1	0
1	1	0
2	0	0
3	1	0
4	0	0
5	0	0
6	0	0
7	3	1
8	0	2
9	1	0
10	1	1
11	0	0
12	0	0

Figure 29 Sample dataset 4

These two datasets are further split into 2 subsets each, a training set and a test set. We have named them X_train, X_test, y_train, y_test. It is to be noted that our test size here is 25% of the values. After splitting into testing and training sets and transforming the values of dependent variables i.e. X_train, X_test through feature scaling, we get the following datasets:

f. X training and test sets:

X_train - NumPy array

	0	1
0	-0.460372	-0.47721
1	2.98532	1.9562
2	0.401052	-0.47721
3	-0.460372	-0.47721
4	-0.460372	-0.47721
5	-0.460372	-0.47721
6	0.401052	0.739493
7	-0.460372	-0.47721
8	0.401052	1.9562
9	-0.460372	-0.47721
10	-0.460372	-0.47721
11	-0.460372	-0.47721

X_test - NumPy array

	0	1
0	-0.460372	-0.47721
1	-0.460372	-0.47721
2	2.98532	0.739493
3	0.401052	-0.47721
4	-0.460372	1.9562
5	-0.460372	-0.47721
6	-0.460372	-0.47721
7	-0.460372	1.9562
8	0.401052	-0.47721
9	-0.460372	-0.47721
10	0.401052	-0.47721
11	-0.460372	-0.47721

Figure 30 Training and test set of Sample dataset 4

g. Y training and test sets:

y_train - NumPy array

	0
0	0
1	1
2	0
3	0
4	1
5	1
6	0
7	0
8	1
9	1
10	0

y_test - NumPy array

	0
0	0
1	0
2	0
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	0

Figure 31 Training and test set of Sample dataset 4

Here comes the main step i.e. the importing and creation of our KNN classifier object and fitting it onto our X_train and y_train datasets. With this classifier, we predict our values, i.e., if the passenger survived (1) or he/she didn't (0). The predicted values are stored in

the y_pred.We compare the predicted values with the real values which were in y_test to measure the accuracy of our used process. We use a confusion matrix for this purpose.

h. Scatter plot of KNN

Here below is the representation of dataset in form of scatter plot
In the first scatter plot we see the correlation between the two variables of X, on the basis of which we are going to predict values of y;

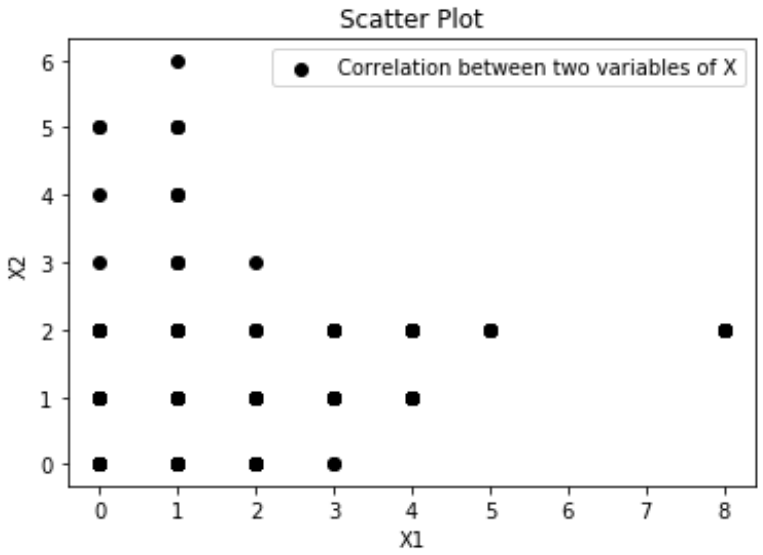


Figure 32 Scatter Plot of X1

Similarly, we are going to see the correlation between 1st variable of X i.e. X1 with Y and X2 with y;

- ✓ **X1 with Y;**

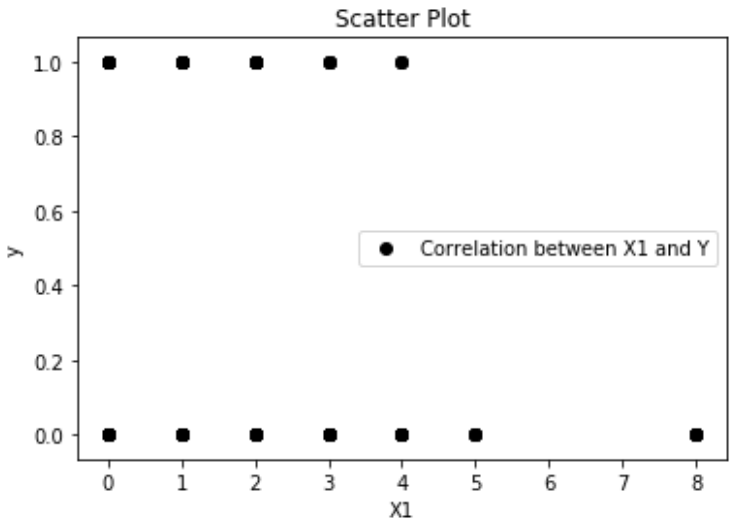


Figure 33 Scatter plot of X1 with Y

- ✓ **X2 with Y;**

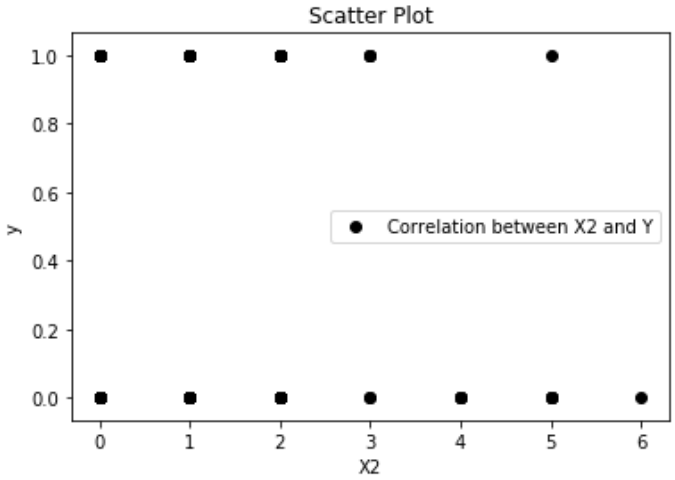


Figure 34 Scatter plot of X2 with Y

And finally, we plot the predicted set with the test set but the actual number of accurate results can only be seen through the confusion matrix;

✓ **Y_test, Y_pred;**

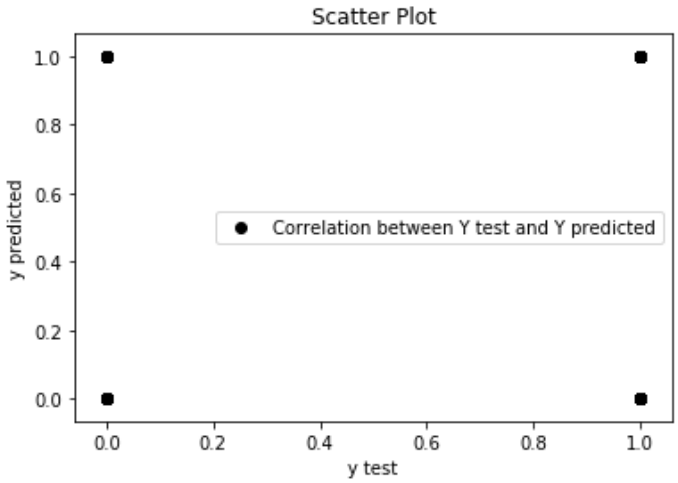


Figure 35 Scatter plot of Y_test with Y_pred

i. Box plot of KNN

✓ **X1:**

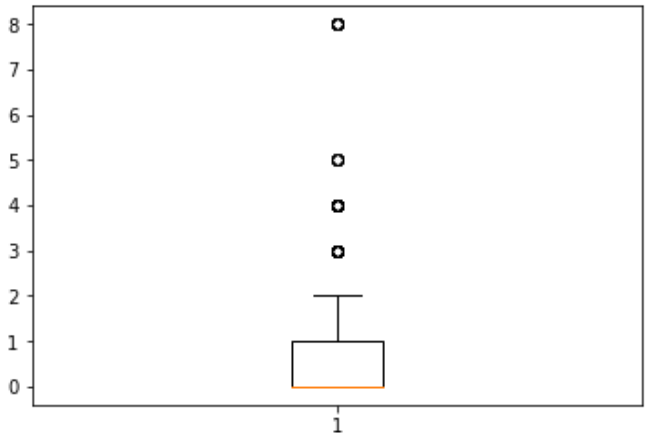


Figure 36 Box plot of X1

✓ **X2:**

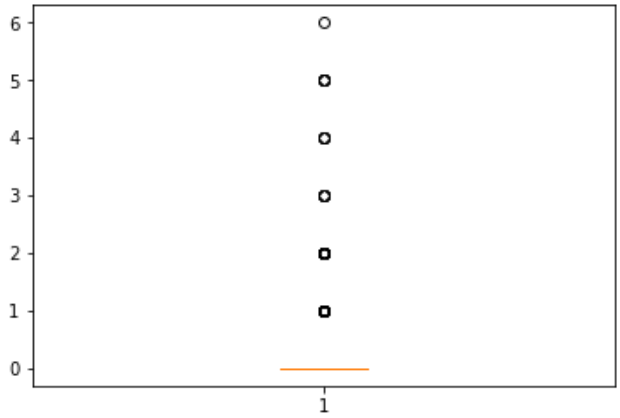


Figure 37 Box plot of X2

✓ **Y:**

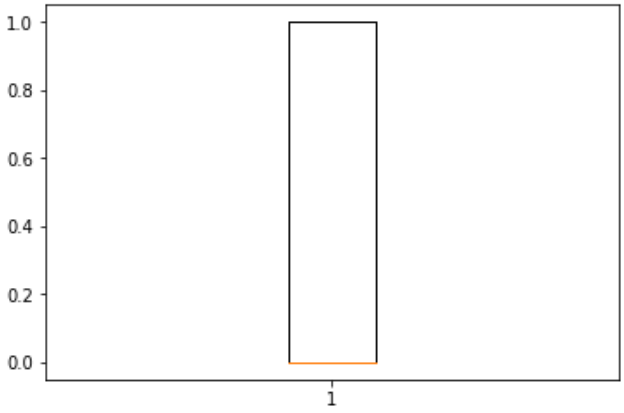


Figure 38 Box plot of X2

j. Resulting confusion matrices:

Note: The black circles enclose the number of wrong predictions, e.g., in "Auto" algorithms' case, in a set of 223 values, a total of 59+17 = 76 prediction errors occurred. It is to be noted that the number of errors vary with changing the algorithm while defining our classifier. We shall compare the no. of errors occurring in different algorithms later on.

- **Auto:**

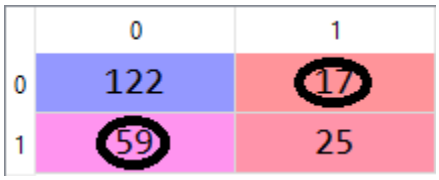


Figure 39 Confusion Matrix of Auto Algorithm

- **Ball tree:**

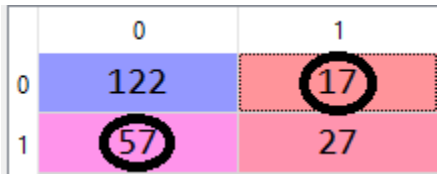


Figure 40 Confusion Matrix of Ball Tree Algorithm

- **KD Tree:**

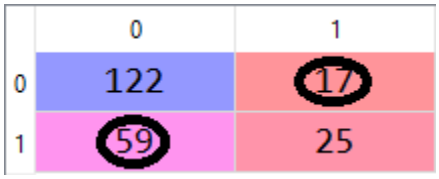


Figure 41 Confusion Matrix of KD Tree Algorithm

- **Brute Force:**

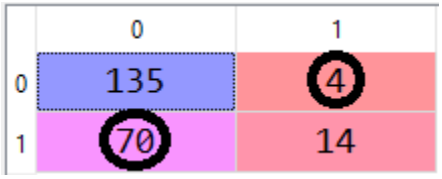


Figure 42 Confusion Matrix of Brute Force Algorithm

Visualization:

Below are shown the visualized test set and the training set. Those points which are not of the same color as their background represents an error. By comparison, we can see that with KNN we have obtained better results than the techniques that we used previously.

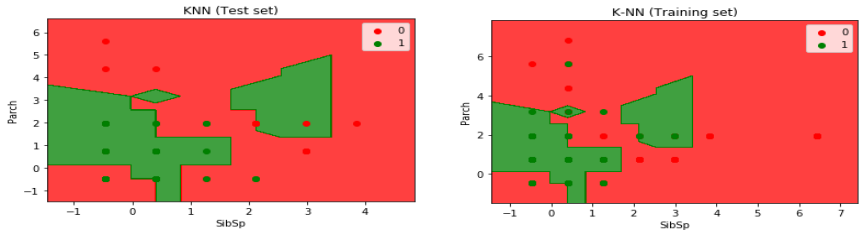


Figure 43 Visualization of Test set and Training set in KNN

Now finally we compare the accuracy of different algorithms that we can use for determining the nearest neighbors on the basis of the confusion matrices obtained in each algorithm;

Table 1 algorithms and their accuracy

Algorithm	**Accuracy**
Auto	65.9%
Ball tree	66.8%
KD tree	65.9%
Brute force	66.8%

- **K-means clustering**

a. Introduction

Data clustering, or cluster analysis, is the process of grouping data items so that similar items belong to the same group/cluster. There are many clustering techniques. In this section, we will explain how to implement the K-means technique.

There are many variations of the K-means technique. The basic version is sometimes called **Lloyd's heuristic**.

b. Advantages

- ✓ The K-means technique has an element of randomness, so the random seed is set so that results are reproducible.
- ✓ The K-means technique typically stabilizes very quickly, often within 10 iterations.

c. Applications

- ✓ **Importing Libraries:**

```
import matplotlib.pyplot as plt
import numpy as np
from sklearn.cluster import KMeans
```

- ✓ **Feeding the values from given .csv file:**

```
file = r'124.csv'
dataset = pd.read_csv(file)
X = dataset.iloc[:, [0, 1]].values
```

- ✓ **Plotting Graph:**

```
plt.scatter(X[:,0],X[:,1], label='True Position')
```

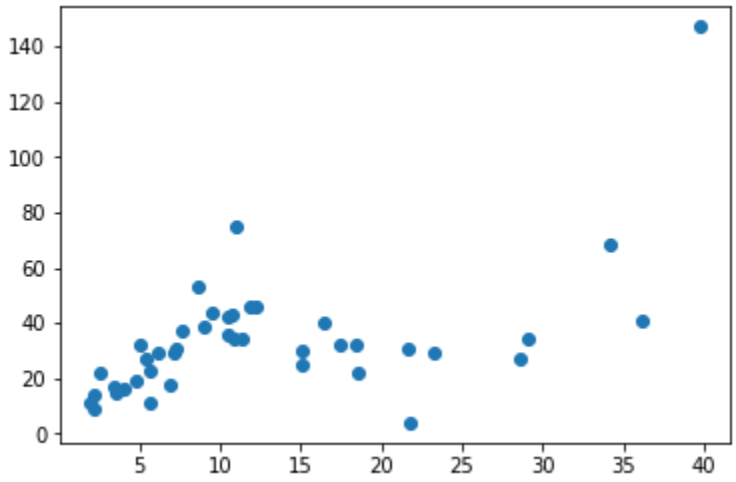


Figure 44 Scatter Plot of Sample Data 5

- ✓ **Evaluating the number of clusters:**

```
alist = []
for iteration in range(1, 11):
    kmeans = KMeans(n_clusters = iteration, init = 'k-means++')
    kmeans.fit(X)
```

```
    alist.append(kmeans.inertia_)
plt.plot(range(1, 11), alist)
plt.show()
```

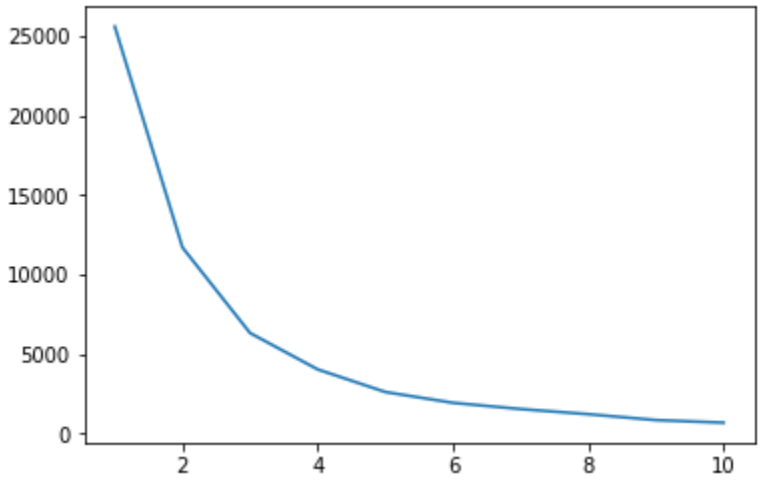


Figure 45 Scree Plot of Sample Data 5

- ✓ **Visualization of given data with specified number of clusters:**

```
kmeans = KMeans(n_clusters=5)
kmeans.fit(X)
plt.scatter(X[:,0],X[:,1], c=kmeans.labels_, cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:,0] ,kmeans.cluster_centers_[:,1],
color='black')
```

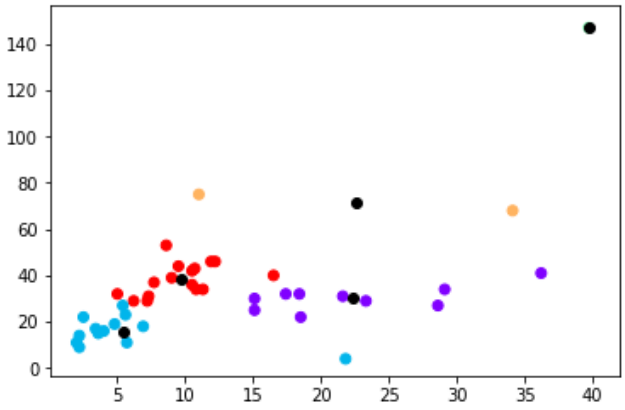


Figure 46 K-means model of Sample Data 5

CHAPTER # 5
TESTING AND EVALUATION

5.1 Introduction

This test plan section portrays the fitting systems, procedure and philosophies used to design, execute and oversee testing of the Missing Data Prediction of WSN project. The test plan will guarantee that the application meets the client's necessities at an authorized level.

Manual Testing will be sought after which consolidates testing an item physically, i.e., without using any automated instrument or any substance. In this sort, the analyzer accept command over the activity of an end-customer and tests the item to perceive any astounding behavior or bug. Each module will be attempted freely and, from that point onward, will be composed with various modules. Thusly, Unit Testing and Integration testing will be sought after. For each module, Black box Testing is done and for merged modules Acceptance Testing is done.

The test scope incorporates the Testing of every single utilitarian prerequisite, application execution and use cases necessities recorded in the necessity report.

Programming testing, dependent upon the testing methodology used, can be executed at whatever point in the improvement system. Regardless, most of the test effort occurs after the requirements have been portrayed and the coding method has been done.

This report fuses the course of action, expansion, approach and arrangement of Missing information expectation model of WSN test. The pass/miss the mark criteria of the experiments are moreover described. The Test Plan report records and tracks the significant information required to satisfactorily portray the best approach to manage be used in the testing of the thing.

5.2 Test Items

In light of the Missing Data Prediction Model necessities and plan portrayal, application modules of versatile data prediction model and non-functional situation will be tried. The Requirements Defined in Software Requirements Specification and the Design substances as clarified in Software Design Document will be tried.

5.3 Features Tested

Following Features are Tested:

- Ability to import the necessary libraries for the importing of dataset, processing, mathematical operations, dataset manipulation and exporting of the resultant dataset.
- Ability to detect the noisy and missing data instances.
- Ability to extract individual features from the dataset.
- Ability to extract subsets from each feature where data needs to be predicted.
- Ability to test the Kalman Filter's accuracy by intentionally removing the datapoints and predicting them, which ultimately deduce the Filter's accuracy after comparison of the values with the predicted values.
- Ability to apply Kalman Filter to the subsets for predicting the marked datapoints.
- Ability to update the dataset with the correct values.
- Ability to predict missing values in a Pytorch Tensor and give an output of the same datatype for ease of use.
- Ability to predict missing values in a TensorFlow Tensor and give an output of the same datatype for ease of use.
- Ability to predict missing values in a MXNet ND Array and give an output of the same datatype for ease of use.
- Ability to predict missing values in a Numpy Array and give an output of the same datatype for ease of use.
- Ability to predict missing values in a Pandas DataFrame Tensor and give an output of the same datatype for ease of use.
- Ability to automate this whole model.

5.4 Approach

- **Acceptance test**

  It will be executed by this acknowledgment test plan. Besides, after all investigations are executed, a test report will be dense to exhibit the idea of

Missing Data Prediction Model. Following test procedures will be used in test execution:

- **Unit test**

  Planners are responsible for unit test as white-box testing. The utilization of each module and individual part will be affirmed autonomously.

- **Integration test**

  After the unit test is disregarded the described quality point of confinement, analyzers will execute the mix tests. After all of the modules are composed, it's fundamental to test the thing as a black-box. All the way circumstances will be endeavored to ensure the correspondence convenience.

- **Regression test**

  After specialists fix the bug in one segment, backslide test will be executed by analyzers to ensure that various limits are not impacted.

- **Field test**

  Initially, untrained end customers replicate in any event one existing (yet restricted) mass discernment events in the Missing Data Prediction Model. Different onlookers will be free to help with appraisal. Starting there ahead, post event surveys will be used to accumulate quantitative usage data similarly as emotional data and further improvement will be thought about. This model has additionally been sent to our customers for example English Petroleum for testing and we are anticipating input.

- **Positive and negative testing plan system**

  This procedure will be united with unit test and consolidation test. Trials are arranged in clear circumstances, which ensure that each valuable essential are satisfied. Moreover, phenomenal analyses will in like manner be verified to exhibit how the structure reacts with invalid assignments.

5.5 Item Pass/Fail Criteria:

Nuances of the trials are demonstrated in region Test Deliverables. Following the principles illustrated beneath, an experiment would be closed as pass or fall flat.

- Preconditions are met
- Inputs are completed as determined
- The result functions as what determined in yield => Pass
- The framework doesn't work or not equivalent to yield particular => Fail

5.6 Suspension Criteria and Resumption Requirements

Any bugs found can be fixed by engineers rapidly and no compelling reason to begin the testing procedure from the earliest starting point. Be that as it may, when significant bugs will obstruct some experiments, as they are reliant, the testing process must be delayed. The test will restart from the earliest starting point until the real mistake is settled.

5.7 Test Deliverables

Table 2 Importing Libraries

Test Case Name	Importing Libraries
Test Case No	1
Description	Testing feature import libraries into the system.
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system.
Input Values	Python statements for the relevant libraries i.e. numpy, scikit learn, math, pandas, csv.
Valid Inputs	Syntactically correct python statements.
Steps	Select the python statements and execute them.
Expected Output	Execution Successful. Libraries imported.
Actual Output	Execution Successful. Libraries imported.

Table 3 Import the valid dataset (with invalid data)

Test Case Name	Import the valid dataset (with invalid data)
Test Case No	2
Description	Testing feature to import the target dataset into the system
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and the user must not have the dataset in .csv located in the specified folder.
Input Values	Function with dataset name with .csv extension as argument.
Valid Inputs	Enter invalid filename and execute the function.
Steps	Write the dataset name within the argument section of the import function and execute.
Expected Output	Dataset not found.
Actual Output	Dataset not found.

Table 4 Import the valid dataset (with valid data)

Test Case Name	Import the valid dataset (with valid data)
Test Case No	3
Description	Testing feature to import the target dataset into the system
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and the user must have the dataset in .csv located in the specified folder.
Input Values	Function with dataset name with .csv extension as argument.
Valid Inputs	Enter valid filename and execute the function.
Steps	Write the dataset name within the argument section of the import function and execute.
Expected Output	Dataset imported.
Actual Output	Dataset imported.

Table 5 Detect noisy/missing datapoints (valid input)

Test Case Name	Detect noisy/missing datapoints (valid input)
Test Case No	4
Description	Testing feature to detect noisy/missing datapoints in the dataset
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and the dataset must have missing/noisy values.
Input Values	Target dataset.
Valid Inputs	Pass the imported dataset into the argument of the detection function.
Steps	Write the dataset name within the argument section of the detection function and execute.
Expected Output	Missing values are replaced by NaN.
Actual Output	Missing values are replaced by NaN.

Table 6 Test Case NameDetect noisy/missing datapoints (full dataset)

Test Case Name	Detect noisy/missing datapoints (full dataset)
Test Case No	5
Description	Testing feature to detect noisy/missing datapoints in the dataset
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and the dataset must not have missing/noisy values.
Input Values	Target dataset.
Valid Inputs	Pass the imported dataset into the argument of the detection function.
Steps	Write the dataset name within the argument section of the detection function and execute.
Expected Output	All present.
Actual Output	All values present.

Table 7 Extract individual features/columns from dataset.

Test Case Name	Extract individual features/columns from dataset.
Test Case No	6
Description	Testing Feature to separate each column from the dataset.
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and the dataset must have columns allocated for each variable.
Input Values	Target dataset.
Valid Inputs	Dataset having columns allocated for each variable.
Steps	Write the dataset name within the argument section of the extraction function and execute.
Expected Output	Each column with all rows is stored in a different data frame.
Actual Output	Each column with all rows is stored in a different data frame.

Table 8 Extract subsets from each column where data is missing. (Valid)

Test Case Name	Extract subsets from each column where data is missing. (Valid)
Test Case No	7
Description	Testing feature to extract subset from each column where data is missing
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and the columns must have at least one value present.
Input Values	Column index passed into the argument of the subset creator function.
Valid Inputs	Column having at least one present value.
Steps	Write the column name within the argument section of the subset creator function and execute.
Expected Output	Subsets are created such that the first value of the subset is the first valid value in the complete column.
Actual Output	Subsets are created such that the first value of the subset is the first valid value in the complete column.

Table 9 Extract subsets from each column where data is missing. (Invalid)

Test Case Name	Extract subsets from each column where data is missing. (Invalid)
Test Case No	8
Description	Testing feature to extract subset from each column where data is missing
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and the columns must have no value present.
Input Values	Column index passed into the argument of the subset creator function.
Valid Inputs	Column having not a single present value.

Steps	Write the column name within the argument section of the subset creator function and execute.
Expected Output	Subsets are not created and an output of "All missing" is shown.
Actual Output	Subsets are not created and an output of "All missing" is shown.

Table 10 Testing accuracy of the Kalman Filter on our dataset

Test Case Name	Testing accuracy of the Kalman Filter on our dataset
Test Case No	9
Description	Testing Feature to predict the already known data to test the Kalman Filter's applicability on our dataset.
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and the testing range must have known values.
Input Values	Observation, transition and state matrices along with the targeted range of values of the subset.
Valid Inputs	Correct observation, transition and state matrices and the targeted range must have values present.
Steps	Remove some values from the known range intentionally and apply the Kalman Filter function to predict those values. Then compare the known and predicted values through an overlapped graph of both values.
Expected Output	Overlapped time graph of both the values in which both the curves fit each other.
Actual Output	Overlapped time graph of both the values in which both the curves fit each other.

Table 11 Applying Kalman Filter to predict the marked missing values in Pytorch Tensor.

Test Case Name	Applying Kalman Filter to predict the marked missing values in Pytorch Tensor.
Test Case No	10
Description	Testing Feature to apply Kalman Filter to predict the marked missing values.
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and the extracted subset.
Input Values	Pytorch Tensor with missing values
Valid Inputs	Correct observation, transition and state matrices along with the targeted subset with the missing values.
Steps	Apply the predict function to predict the missing values in the subset. Only the Pytorch tensor is given as argument.
Expected Output	Missing values are predicted and a complete pytorch tensor is returned as output
Actual Output	Missing values are predicted and a complete pytorch tensor is returned as output

Table 12 Applying Kalman Filter to predict the marked missing values in TensorFlow Tensor.

Test Case Name	Applying Kalman Filter to predict the marked missing values in TensorFlow Tensor.
Test Case No	11
Description	Testing Feature to apply Kalman Filter to predict the marked missing values.
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on

	the system and the extracted subset.
Input Values	TensorFlow Tensor with missing values
Valid Inputs	Correct observation, transition and state matrices along with the targeted subset with the missing values.
Steps	Apply the predict function to predict the missing values in the subset. Only the TensorFlow tensor is given as argument.
Expected Output	Missing values are predicted and a complete TensorFlow tensor is returned as output
Actual Output	Missing values are predicted and a complete TensorFlow tensor is returned as output

Table 13 Applying Kalman Filter to predict the marked missing values in MXNet ND Array Tensor.

Test Case Name	Applying Kalman Filter to predict the marked missing values in MXNet Nd Array.
Test Case No	12
Description	Testing Feature to apply Kalman Filter to predict the marked missing values.
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and the extracted subset.
Input Values	MXNet Nd Array with missing values
Valid Inputs	Correct observation, transition and state matrices along with the targeted subset with the missing values.
Steps	Apply the predict function to predict the missing values in the subset. Only the MXNet Nd Array is given as argument.
Expected Output	Missing values are predicted and a complete MXNet Nd Array is returned as output
Actual Output	Missing values are predicted and a complete MXNet Nd Array is returned as output

Table 14 Applying Kalman Filter to predict the marked missing values in Numpy Array

Test Case Name	Applying Kalman Filter to predict the marked missing values in Numpy Array.
Test Case No	13
Description	Testing Feature to apply Kalman Filter to predict the marked missing values.
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and the extracted subset.
Input Values	Numpy Array with missing values
Valid Inputs	Correct observation, transition and state matrices along with the targeted subset with the missing values.
Steps	Apply the predict function to predict the missing values in the subset. Only the Numpy Array is given as argument.
Expected Output	Missing values are predicted and a complete Numpy Array is returned as output
Actual Output	Missing values are predicted and a complete Numpy Array is returned as output

Table 15 Applying Kalman Filter to predict the marked missing values in Pandas DataFrame

Test Case Name	Applying Kalman Filter to predict the marked missing values in Pandas Dataframe.
Test Case No	14
Description	Testing Feature to apply Kalman Filter to predict the marked missing values.
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on

	the system and the extracted subset.
Input Values	Pandas Dataframe with missing values
Valid Inputs	Correct observation, transition and state matrices along with the targeted subset with the missing values.
Steps	Apply the predict function to predict the missing values in the subset. Only the Pandas Dataframe is given as argument.
Expected Output	Missing values are predicted and a complete Pandas Dataframe is returned as output
Actual Output	Missing values are predicted and a complete Pandas Dataframe is returned as output

Table 16 Update dataset with estimated values

Test Case Name	Update dataset with estimated values
Test Case No	15
Description	Testing Feature to update the dataset with the predicted/estimated values resulting from application of the Kalman Filter.
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and Kalman filter must be applied on the subset.
Input Values	Target dataset and estimated values from the filter.
Valid Inputs	Target dataset and estimated values from the filter.
Steps	The estimated values and their respective indexes are saved in a temporary list and then the corresponding index values in the target dataset are replaced.
Expected Output	Updated dataset.
Actual Output	Updated dataset.

Table 17 Plot the new values and the original values to visualize result.

Test Case Name	Plot the new values and the original values to visualize result.
Test Case No	16
Description	Testing feature to plot each column of the new data set, overlapped with the original dataset to visualize the error.
Testing Technique Used	Unit Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and must have both the new dataset and the original dataset.
Input Values	New dataset with the predicted values and the original dataset.
Valid Inputs	New dataset with the predicted values and the original dataset.
Steps	Iterate over each row of both the columns and plot them overlapped with different colors and against T=total number of rows.
Expected Output	Overlapped graphs visualizing error in predicted and original values.
Actual Output	Overlapped graphs visualizing error in predicted and original values.

Table 18 Automate the whole process

Test Case Name	Automate the whole process
Test Case No	17
Description	Testing to automate this whole process which requires only one operation from the user.
Testing Technique Used	Integration Testing
Preconditions	The user must have python 3 or later and Windows 10 OS installed on the system and must have the dataset whose values are to be predicted.
Input Values	Target dataset.
Valid Inputs	Target dataset.
Steps	Execute the .py file

Expected Output	Two files i.e. PredictedValues.csv, Errors.csv and 36 plots (.png) will be generated in the output folder.
Actual Output	Two files i.e. PredictedValues.csv, Errors.csv and 36 plots (.png) is generated in the output folder.

5.8 Environmental Needs

5.8.1 Hardware

- A PC with i3-3110M or greater processor, RAM: 4 GB or greater, HDD with 2GB of free space.

5.8.2 Software

- Windows 10 OS.
- Python 3 or greater.
- Microsoft Excel.

5.9 Responsibilities, Staffing and Training

Responsibilities

- Nashrah Khan is responsible for Acceptance Testing
- Mansoor Ahmad is responsible for Integration Testing
- All of the group members are responsible for testing each of their prepared separate unit that is Unit Testing.

Skills

- Skills that were required to create this model are;
    a. Python language
    b. Spyder IDE.
    c. ML Algorithms for data science.
    d. Kalman Filtering.

5.10 Risks and contingencies

Although this model predicts the values with a higher accuracy, but it only works when we have at least one value preceding the missing values. If we start from the missing value as the first input, we will get totally false and inaccurate data, the more preceding values, the better. Moreover, the data is stratigraphic and it is collected at an interval of just a few seconds, so it must be collected and saved into .csv form before we can predict the missing values. Live prediction of data cannot be done by this model, which would've been a more favorable outcome of this project. This leaves room for further research and improvement in this project. More techniques and algorithms can be explored, even created to bring the above-mentioned improvement to this project.

CHAPTER 6

USER MANUAL

6. User Manual

6.1 Step-By-Step Guide

- Import the module i.e. “missingDataPredictor” into the program file. The statement would be like the following;

  import missingDataPredictor as mdp

- Call the “predict” function of the module and pass the dataset with the missing values as the function’s argument and store the output in a variable for further use.

  Output = mdp.predict(dataset)

- The “Output” will now have the complete dataset with the missing values replaced by the predicted values resulting from the prediction algorithm.

  NOTE: If the dataset passed is a pandas data frame, the output will also be a dataframe. Similar is the case of Pytorch Tensor, Tensorflow Tensor, Mxnet ND Array and Numpy Array. We have provided functionality for all of the above datatypes to maximize ease of use i.e. the whole algorithm is now easily usable by just writing one simple programming statement as shown above, provided our module is imported into the program file.

Conclusion

Missing Data Prediction is a funded project by a renowned company "British Petroleum" (BP). In this final year project, the missing values have been predicted with palatable accuracy figures for the client successfully. Moreover, the extra functionality of reading datasets via various file formats has been added in the trained model and developed a final module for data predictions. There are various people using Machine Learning through different frameworks i.e TensorFlow, Pytorch, Numpy, and MXNet.Their training datasets have missing values in a different file format and to overcome it, they have to write a bunch of lines of code to do data preprocessing, data cleaning and apply different algorithms to predict the missing values with least error. This final year project has provided them with ease by developing a module that can resolve this issue of missing data. It is now only required to import our module and provide their dataset, alter just one line of code, i.e., call the developed prediction function and as an output, they will get the dataset with missing values replaced by the predicted values. The accuracy was calculated through mean absolute error and root mean square error which gave us acceptable error values.

References

[1] Mathias Scholz, Fatma Kaplan, Charles L.Guy, Joachim Kopka 2005. "Non-linear PCA: a missing data approach". Bioinformatics , Vol. 21 no. 20 2005, pages 3887–3895

[2] Zhou,X. et al. (2003) Missing-value estimation using linear and non-linear regression with bayesian gene selection. Bioinformatics, 19, 2302–2307.

[3] Simeka,K. and Kimmel,M. (2003) A note on estimation of dynamics of multiple gene expression based on singular value decomposition. Math. Biosci., 182, 183–199.

[4] Honkela,A. and Valpola,H. (2005) Unsupervised variational bayesian learning of nonlinear models. To appear in Advances in Neural Information Processing Systems 17.

Appendix A: Project Proposal

Missing Data Prediction from WSN

Extended Title:[if required]: N/A

Brief Description of the Project/Thesis with Salient Space:

Prediction of missing data from a data set is easy if the data is linear and two dimensional. But in real life cases, data is nonlinear and complex, thus making the prediction of missing values difficult. In this project, the missing data values are predicted using techniques of statistics and software engineering. Our methodology can be applied to many other problems in the same domain.

Scope of Work:

The data prediction will be achieved by exploring Machine Learning algorithms and then picking the best one for our application.

Academic Objectives:

1. Develop a working data prediction method that is going to be applied directly in the industry and research.
2. Exploring new dimensions of Software Engineering i.e. Deep Learning
3. Practically applying our gathered knowledge from the whole academic program since 2015.
4. Working effectively as team.
5. Improving writing and oral presentation skills.
6. Successfully develop our FYP.

Application/End Goal Objectives:

To devise a method which predicts the missing values in non-linear data set to its nearest and most precise value.

Previous Work Done on The Subject:

Statistics, Python

Material Resources Required:

Test Data Set, Research Papers on Inverse NLPCA, online learning resources on Deep learning

No of Students Required: 4

Nashrah Khan
Ali Sultan
Aniqa Tariq
Mansoor Ahmad

Special Skills Required:

Python, Deep learning, Statistics

Appendix B: Techniques

- **REGRESSION**

  Regression models (both linear and non-linear) are used for predicting a real value, like salary for example. If your independent variable is time, then you are forecasting future values, otherwise your model is predicting present but unknown values. Regression technique vary from Linear Regression to SVR and Random Forests Regression.

In this part, you will understand and learn how to implement the following Machine Learning Regression models:

a. Simple Linear Regression
b. Multiple Linear Regression
c. Polynomial Regression
d. Support Vector for Regression (SVR)

- **CLASSIFICATION**

  To work on a continuous data , Classification technique is used.It predicts a categories in the continuous data.There are also any types of classification models such as Logistic Regression, SVM, and nonlinear ones like K-NN, Kernel SVM and Random Forests.

In this part, you will understand and learn how to implement the following Machine Learning Classification models:

a. K-Nearest Neighbors (K-NN)
b. Support Vector Machine (SVM)

- **CLUSTERING**

Clustering is like order, yet the premise is extraordinary. In Clustering you don't have the foggiest idea what you are searching for, and you are endeavoring to recognize a

few fragments or bunches in your information. When you use bunching calculations on your dataset, unforeseen things can all of a sudden spring up like structures, groups and groupings you would have never thought of something else.

In this part, you will comprehend and figure out how to execute the accompanying Machine Learning Clustering models:

a. K-Means Clustering

- **PRINCIPLE COMPONENT ANALYSIS(PCA)**

  A standard system for imagining high dimensional information and for information pre-handling. PCA diminishes the dimensionality (the quantity of factors) of an informational collection by keeping up however much difference as could reasonably be expected.

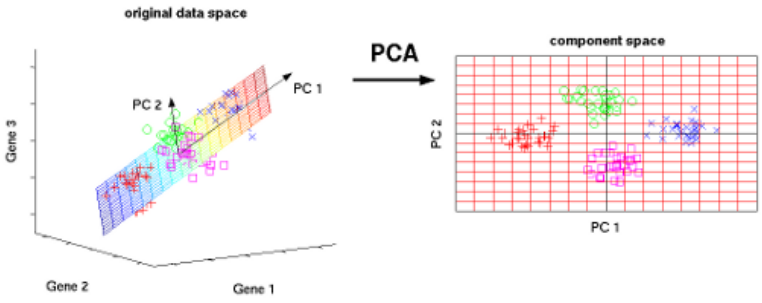


Figure 1. Principle Component Analysis

Matthias Scholz, Ph.D. thesis (http://phdthesis-bioinformatics-maxplanckinstitute-molecularplantphys.matthias-scholz.de/)

- **KALMAN FILTERS**

  Kalman filter can be used where we have **uncertain information** about some dynamic system, and we can make an **educated guess** about what the system is going to do next. Even if messy reality comes along and interferes with the clean motion we can guess about, the Kalman filter will often do a very good job of figuring out what actually happened. And it can also solve complex correlations problems.

Kalman filters are ideal for systems which are **continuously changing**. They are light on memory (they don't need to keep any history other than the previous state), and they are very fast, making them well suited for real time problems and embedded systems.

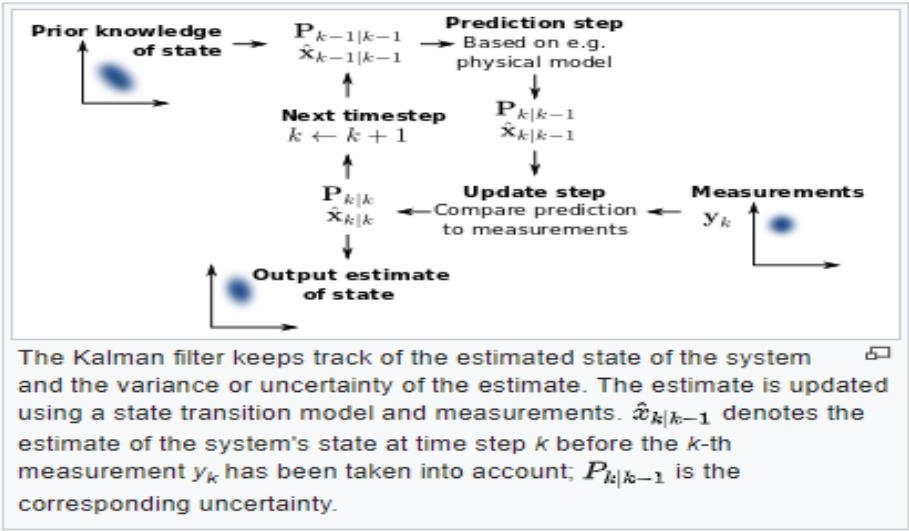


Figure 2. User Kalman Filter

https://en.wikipedia.org/wiki/Kalman_filter

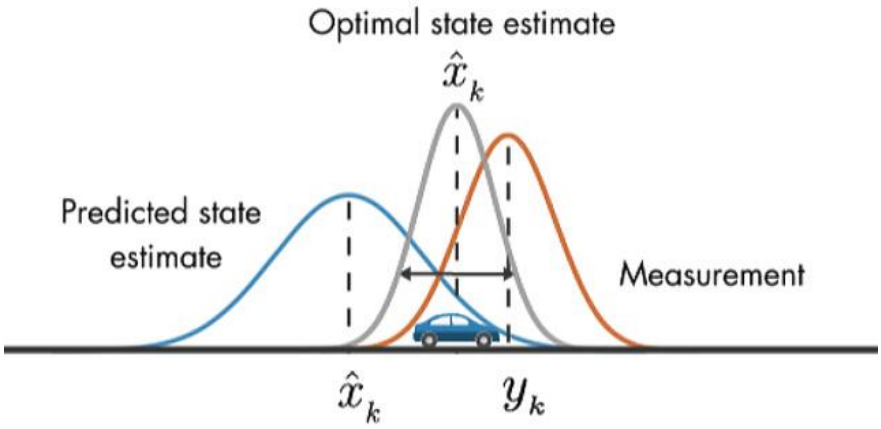


Figure 3. Kalman Filter

https://ch.mathworks.com/content/dam/mathworks/videos/u/Understanding-Kalman-Filters-Part-3-Optimal-State-Estimator.mp4/jcr:content/renditions/S2E3_Thumbnail.jpg

Appendix C: Plotting Tools

- **SCREE Plot**

  A Scree Plot is a straightforward line portion plot that demonstrates the division of complete change in the information as clarified or spoken to by every Principle Component.

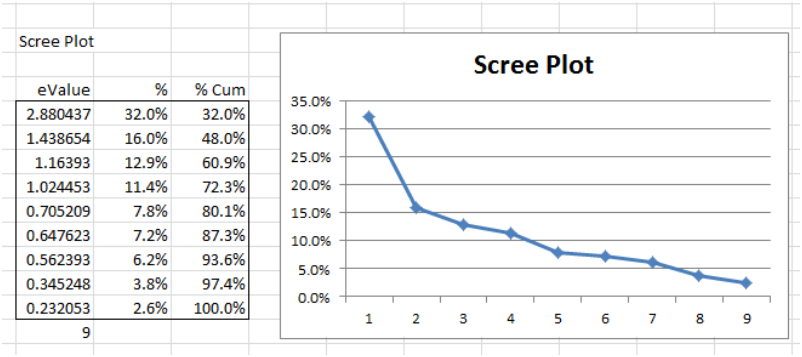


Figure 4. SCREE Plot

https://i0.wp.com/www.real-statistics.com/wp-content/uploads/2013/09/scree-plot-factor-analysis.png

- **Box Plot**

  A box plot is four quartile rectangular plots. In which each shows some value. The id line shows mean value and variance of data from the mean from high to low. The dot shows the outliers.

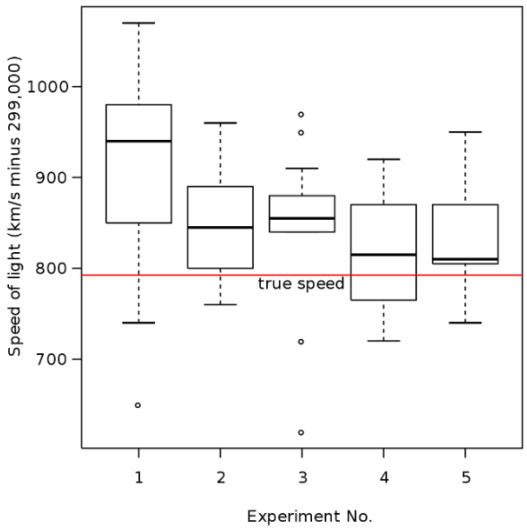


Figure 5. Box Plot

https://upload.wikimedia.org/wikipedia/commons/thumb/f/fa/Michelsonmorley-boxplot.svg/1200px-Michelsonmorley-boxplot.svg.png

- **Scatter Plot**

The scatter plot lattice is intended to show the connection between all sets of a few factors. The plot framework comprises of plot cells containing little disperse plots shaped from a couple of factors. The factors are spoken to by the X-pivot and Y-hub of each plot cell. The watched qualities on the two factors are spoken to by focuses in the little dissipate plot.

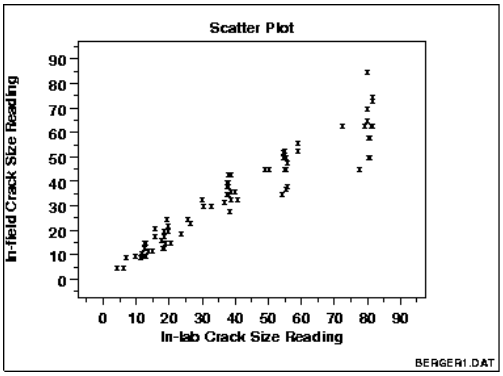


Figure 6. Scatter Plot

https://www.itl.nist.gov/div898/handbook/eda/gif/scatterp.gif