# HEALTH INDEX ANALYSIS OF XLPE CABLE INSULATION USING MACHINE LEARNING TECHNIQUE

**A THESIS**
*Submitted by*

**MANISH KUMAR**

**(M200181EE)**

**MASTER OF TECHNOLOGY**

**IN**

**ELECTRICAL ENGINEERING**

**(High Voltage Engineering)**

Under the guidance of
**Dr. Preetha P**

**DEPARTMENT OF ELECTRICAL ENGINEERING**

NATIONAL INSTITUTE OF TECHNOLOGY CALICUT

NIT CAMPUS P.O., KOZHIKODE

KERALA, INDIA - 673601.

JULY 2022

# ACKNOWLEDGEMENT

I am extremely thankful to my project guide, **Dr. Preetha P** Associate Professor and Head, Department of Electrical Engineering, National Institute of Technology Calicut, whose valuable suggestions and cooperation helped me in the successful completion of my major project.

I would especially like to mention the support extended by **Dr. Preetha P**, Associate Professor and Head, Department of Electrical Engineering, National Institute of Technology Calicut, for providing the opportunity to materialize the thesis. I am also thankful to all the faculty of Electrical Engineering Department, National Institute of Technology Calicut for the support and cooperation.

I would like to acknowledge the help extended to me by my seniors and friends from my department without whose help my project could not be completed.

Date- 04/07/2022 MANISH KUMAR

# DECLARATION

*“I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.”*

|  |  |
|---|---|
| Place: NIT Calicut | Name: Manish Kumar |
| Date: 04/07/2022 | Roll No.: M200181EE |

# CERTIFICATE

This is to certified that the thesis entitled "**Health Index Analysis of XLPE Cable Insulation using Machine Learning Technique"** submitted by **Mr. Manish Kumar** (Reg No: M200181EE) to the National Institute of Technology Calicut towards partial fulfillment of the requirements for the award of the Degree of Master of Technology in **Electrical Engineering** (**High Voltage Engineering**) is a bona fide record of the work carried out by him under my supervision and guidance.

**Dr. Preetha P**

*Associate professor (Project Guide) &*
*Head of Department*
*Dept. of Electrical Engineering*

*Place: NIT Calicut*
*Date:04 /07/2022*

# CONTENTS

ABSTRACT ..... vii

LIST OF FIGURES ..... viii

LIST OF TABLES ..... x

LIST OF ABBREVIATIONS ..... xi

CHAPTER 1 ..... 1

INTRODUCTION ..... 1

1.1 INTRODUCTION ..... 1

1.2 MOTIVATION ..... 1

1.3 LITERATURE REVIEW ..... 2

1.4 RESEARCH GAP ..... 3

1.5 OBJECTIVE ..... 4

1.6 ORGANIZATION OF THE REPORT ..... 4

CHAPTER 2 ..... 5

MACHINE LEARNING TECHNIQUE AND ALGORITHM ..... 5

2.1 INTRODUCTION ..... 5

2.2 TYPES OF MACHINE LEARNING ..... 5

2.3 FLOWCHART OF MACHINE LEARNING ..... 6

CHAPTER 3 ..... 8

PARTIAL DISCHARGE AND NEUTRAL CORROSION MEASUREMENT ..... 8

3.1 PARTIAL DISCHARGE ..... 8

3.2 ONLINE PARTIAL DISCHARGE MEASUREMENT: ..... 9

3.3 NEUTRAL CORROSION ..... 11

3.4 NEUTRAL CORROSION MEASUREMENT ..... 12

CHAPTER 4 ..... 17

DATA COLLECTION AND VISUALIZATION ..... 17

4.1 INTRODUCTION ..... 17

4.2 MIN- MAX NORMALIZATION ..... 18

4.3 VISUAL REPRESENTATION OF DATA ..... 19

4.4 CORRELATION MATRIX ..... 22

CHAPTER 5 ..... 23

DATA PREPROCESSING ..... 23

5.1 INTRODUCTION ..... 23

5.2 ONE HOT ENCODING ..... 23

5.3 SPLITTING THE DATA INTO TRAIN-TEST ..... 24

5.4 FEATURE SCALING ....................25
CHAPTER 6 .................... 26
MACHINE LEARNING ALGORITHMS .................... 26
6.1 INTRODUCTION ....................26
6.2 CLASSIFICATION REPORT ....................27
6.3 CONFUSION MATRIX ....................28
6.4 HYPERPARAMETER TUNING ....................29
CHAPTER 7 .................... 31
RANDOM FOREST CLASSIFIER .................... 31
7.1 INTRODUCTION ....................31
7.2 REASON TO USE RANDOM FOREST ALGORITHM ....................32
7.3 MECHANISM OF RANDOM FOREST ....................32
7.4 IMPORTANT HYPERPARAMETERS OF RANDOM FOREST ....................33
7.5 CONFUSION MATRIX AND CLASSIFICATION REPORT BEFORE HYPERPARAMETER TUNING ....................34
7.6 HYPERPARAMETER TUNING OF RANDOM FOREST MODEL ....................35
7.7 CONFUSION MATRIX AND CLASSIFICATION REPORT AFTER HYPERPARAMETER TUNING ....................36
7.8 COMPARISON OF DIFFERENT MACHINE LEARNING MODEL WITH RANDOM FOREST ALGORITHM ....................37
CHAPTER 8 .................... 42
ARTIFICIAL NEURAL NETWORK MODEL (ANN).................... 42
8.1 INTRODUCTION ....................42
8.2 STEP WISE WORKING OF ANN MODEL ....................44
8.3 ACTIVATION FUNCTIONS....................45
8.4 OPTIMIZERS ....................47
8.5 LOSS FUNCTIONS ....................47
8.6 KERAS – DENSE LAYER ....................49
8.7 EPOCHS ....................49
8.8 PERFORMANCE LEARNING CURVE AND OPTIMIZATION LEARNING CURVE....................49
8.9 CLASSIFICATION REPORT AND CONFUSION MATRIX FOR ANN MODEL ....................51
CHAPTER 9 .................... 53
CONCLUSION.................... 53
REFERENCES .................... 54

# ABSTRACT

The main reason that leads to failure of cable insulation is degradation due to age and partial discharge. However, replacing and maintaining underground cable circuits throughout the excavation phase is very expensive. The severity of the condition of the cable insulation information aids in making better-informed judgments for system planning and maintenance forecasting. In the past, machine learning models such as the Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes were used to determine the insulation health condition of XLPE cable. The data which is provided by the Utility Analytics Network contains different parameters like Partial Discharge, Neutral Corrosion, Service Age, Visual Index and Health Index. The data needs preprocessing before applying Machine Learning models. In this work focus has been given on Random Forest Model and ANN model. The reason Random Forest was chosen over other algorithms is that it requires approximately less time to train. Additionally, it operates swiftly even with a large dataset and makes accurate output predictions. Even if some data are missing, Random Forest can still be effective. In this work Random Forest Model after doing the hyperparameter tuning provided the best results with accuracy of 98% and only 11 data were predicted wrong out of 500 data. ANN model gave an accuracy of 97.2%. A comparison has been made between different Models which shows that Random Forest produces the best results.

# LIST OF FIGURES

Fig. 2.1: Algorithm for Machine Learning Technique ........ 7

Fig. 3.1: Degradation of cable insulation ........ 9

Fig. 3.2: HFCT Sensor ........ 10

Fig. 3.3: Neutral corrosion of cable ........ 12

Fig. 3.4 Block diagram of TDR ........ 13

Fig. 3.5 Example of TDR trace indicating neutral corrosion ........ 13

Fig. 3.6: Level 1 Neutral Corrosion ........ 15

Fig. 3.7: Level 2 Neutral Corrosion ........ 15

Fig. 3.8: Level 3 Neutral Corrosion ........ 15

Fig. 3.9: Level 4 Neutral Corrosion ........ 16

Fig. 3.10: Level 5 Neutral Corrosion ........ 16

Fig. 4.1: Pie Chart Representation of Data ........ 19

Fig. 4.2: Bar Chart representation of Visual Condition of Datasets ........ 20

Fig. 4.3: Count of different Health Index Sample ........ 20

Fig. 4.4: Scatter Plot of Age and Partial Discharge(Normalized) ........ 21

Fig. 4.5: Scatter plot of Age and Natural Corrosion(Normalized) ........ 21

Fig. 4.6: Correlation Matrix of Dataset ........ 22

Fig. 5.1: Example of One Hot Encoding ........ 23

Fig. 5.2: Head of our data ........ 24

Fig. 5.3: Data after one hot encoding ........ 24

Fig. 5.4: After applying Feature Scaling ........ 25

Fig. 6.1: Confusion Matrix for Binary Classification ........ 28

Fig. 6.2: Confusion Matrix for Multi Class Classification ........ 29

Fig. 7.1: Working of Random Forest Algorithm ........ 31

Fig. 7.2: Visualization of a Random Forest Model Making prediction ........ 32

Fig. 7.3: Confusion Matrix for Random Forest Model Before Hyperparameter Tuning ........ 34

Fig. 7.4 Classification Report for Random Forest Model ........ 35

Fig. 7.5: Confusion Matrix after Hyperparameter Tuning ........ 36

Fig. 7.6: Classification Report after Hyperparameter Tuning ........ 37

Fig. 7.7: Performance comparison of Models ........ 41

Fig. 8.1: ANN model structure ........ 42

Fig. 8.2: Layers of ANN model ........ 43

Fig. 8.3: Working of ANN model ..... 44
Fig. 8.4: ReLU Function ..... 46
Fig. 8.5: Performance Learning Curve ..... 50
Fig. 8.6: Optimization Learning Curve ..... 50
Fig. 8.7: Classification Report for ANN Model ..... 51
Fig. 8.8: Confusion Matrix for ANN Model ..... 52

# LIST OF TABLES

Table 3.1: Corrosion Categories ........................................................................................ 14

Table 4.1:  Data Samples ................................................................................................ 17

Table 4.2: Health Index Code .......................................................................................... 18

Table 7.1: Best Hyperparameter Values for Random Forest .................................................. 36

Table 7.2: Comparison of Different Model before Hyperparameter Tuning .......................... 38

Table 7.3: Comparison of Different Model after Hyperparameter Tuning ............................ 39

Table 7.4: Hyperparameters used for KNN ........................................................................... 39

Table 7.5: Hyperparameters used for SVM ........................................................................... 40

Table 7.6: Hyperparameters used for Logistic Regression .................................................... 40

# LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| 1 | ML | MACHINE LEARNING |
| 2 | ANN | ARTIFICIAL NEURAL NETWORK |
| 3 | PD | PARTIAL DISCHARGE |
| 4 | HFCT | HIGH FREQUENCY CURRENT TRANSFORMER |
| 5 | HI | HEALTH INDEX |
| 6 | TDR | TIME DOMAIN REFLECTOMETER |

# CHAPTER 1

# INTRODUCTION

## 1.1 INTRODUCTION

XLPE cable is designed to withstand high and extra high voltages. These cables work best for long transmission routes at high voltages where significant dielectric losses are present. XLPE cable are easy to handle because it is light in weight in comparison to other cables.

The main reason behind failing of cable insulation is degradation due to ageing and partial discharge. Replacing and maintaining underground cable circuits throughout the excavation period is very expensive.

The use of machine learning (ML) and Deep Learning to evaluate the health index of high voltage XLPE cable is highlighted in this work. The test result revealed the degree of partial discharge of each sample, as well as its age, neutral corrosion, loading, and visual condition. The prediction of Health Index of the cable based on multiclass classification dataset with five different health index is emphasized in this work.

## 1.2 MOTIVATION

A variety of factors influence insulation performance, like load cycle changes, temperature changes, moisture invasion, mechanical stress and others. Although the insulation passed every mandatory test prior to installation and usage, but it may not maintain the same proportional functioning attribute throughout the many years expected. As a result, cable insulation deterioration which occur through partial discharge and the electrical treeing effect has become a major worry in today's society.

Partial Discharge can occur everywhere, independently of the kind of insulation used to separate electrical components: it can occur in a liquid insulation because of the presence of gas bubbles or in a solid insulation for example inside the dielectric of a cable. The effects of the Partial Discharges are different depending on the location where the discharge occurs.

It is possible to classify them in the following categories:

- Surface Discharge: occurring across the insulation surface.
- Internal Discharge: inside voids or defects of the solid insulation.
- Arcing Discharge: electrical breakdown of a gas producing a plasma discharge.
- Corona Discharge: ionization of the fluid or air surrounding a conductor.

The primary cause of cable insulation failure is degradation due to age and partial discharge; however, replacing and maintaining underground cable circuits throughout excavation stage is exceedingly expensive. The severity of the insulation level information aids in making better-informed judgments for system planning and maintenance forecasting.

## 1.3 LITERATURE REVIEW

In many regions, underground power cable systems are often employed, not only in transmission and distribution networks but also in industrial sectors. The quantity of cable installations is rising to enhance the efficiency, dependability, credibility, expense, and risk of power systems as well as to enhance the aesthetic appeal of a community. But this has also lead to increasing number of underground cable system failure. Installation defects, electrical, thermal, and mechanical loads, as well as potential harm to installation sites and the operational environment, might all result in both external and internal degradations of the cable system [1-2].

The main reason for failing of cable insulation is degradation due to age and partial discharge. However, replacing and maintaining underground cable circuits throughout the excavation phase is exceedingly expensive. The severity of the insulation level information aids in making better-informed judgments for system planning and maintenance forecasting. Health index (HI) is intended to be a straightforward indicator of an electrical apparatus's health. It is dependent on a variety of parameters, including age, atmospheric conditions, operating conditions, past apparatus maintenance, and more, in addition to diagnostic property measurements. The rating of in service insulation emphasizes the health index score. Assets can be allocated health index values (such as 1 to 5) based on their health status. The assets those are having high health index (very good health) requires less attention and are suitable for a run-to-fail strategy; an asset with a low health index (very bad health) should be replaced frequently to avoid losses; and assets those are having medium health index should be inspected and maintained more frequently than in starting stage [6]. The uses of neutrals in cable and also different methods to detect Neutral corrosion in Medium-Voltage Underground cable is studied [7]. In order to evaluate the state of high voltage insulation systems and ensure their integrity, partial discharge measurements offer useful information. In recent years, many PD measurement methods have been introduced specifically for use with online measurements. Online partial discharge measurement is done without interrupting service. The use of HFCT and UHF sensors is used to detect Partial discharge [11].

In the paper “Health Index Analysis of XLPE Cable Insulation using Machine Learning Technique” different Machine Learning Algorithms has been implemented to predict the health index of the cable. In comparison to other classifiers, SVM with hyper parameter tuning produced the greatest results, 98 percent accuracy or 2 percent inaccuracy. Deep Learning can also be implemented to find out the health index of cable. With the help of ANN model 96% accuracy is achieved [3-4].

## 1.4 RESEARCH GAP

Health insulation of cable is done with the assistance of many machine learning algorithms like Support vector machine (SVM), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), and Naive Bayes [1]. However, the use of algorithm ‘Random Forest Classifier’ is missing. Random Forest gives prediction by combining a number of decision tree on different subsets of dataset and finally gives result by taking averages to increase dataset predicted accuracy. The random forest checks each tree's prediction rather than relying solely on one, and ultimately decides what to produce based on the majority vote of projections. The "Random Forest" algorithm is employed in this study.

Artificial Neural Network is used to give the prediction of health index of cable. The accuracy achieved is 96% [2]. However, more accuracy can be achieved if model is trained in certain manner.

## 1.5 OBJECTIVE

The main objective of this work is to apply Machine Learning Algorithms for the prediction of insulation health of high voltage XLPE cable to prevent breakdown like phenomenon. Special attention has been given to Random Forest and ANN model to predict the HI of the XLPE Cable Insulation.

## 1.6 ORGANIZATION OF THE REPORT

The remainder of the report is laid out as follows. Different Machine Learning Technique and flow chart of Machine Learning Model is discussed in **Chapter 2. Chapter 3** describes the Partial Discharge and Neutral Corrosion phenomena in cables. In **Chapter 4**, how data is collected and how visualization is done based on the data collected is explained. Different preprocessing technique that has to be done after obtaining the raw data is explained in **Chapter 5**. Basic terms related to Machine Learning Algorithms and different algorithms used are discussed in **Chapter 6**. In **Chapter 7** and **Chapter 8** working procedure of Random Forest and ANN model is explained.

# CHAPTER 2
# MACHINE LEARNING TECHNIQUE AND ALGORITHM

## 2.1 INTRODUCTION

Machine learning is the usage and development of computer systems that have the ability to learn and adapt without being given explicit instructions, by employing statistical models and algorithms to analyze and infer conclusions from patterns in data. Machine learning forecasts future output values using historical data as an input. Machine learning is crucial because it enables businesses to identify patterns in consumer behavior and internal business processes while also assisting in the development of new products. Today many successful organizations like Amazon, Google, and Facebook uses Machine Learning for their business. For many firms today, machine learning has emerged as a critical point of distinction.

## 2.2 TYPES OF MACHINE LEARNING

On the premise that an algorithm learns to make predictions, machine learning is categorized. Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are basic types of Machine Learning. The algorithm which have to be applied depends on type of data we are working to forecast.

1) Supervised Learning:

   In this form of machine learning, labelled training data is provided to algorithm and we need to specify the variables between which we are looking for connections. Input and output both are provided to algorithm.

2) Unsupervised Learning:

   Unlabeled data are provided in this type of Machine Learning Algorithms. The program finds for inter relations between data sets. All of the data utilized to train algorithms, as well as the forecasts and recommendations we provide, is predetermined.

3) Semi-supervised Learning:

   The different approaches mentioned above are combined in this machine learning technique. The algorithm is permitted to explore the data and develop its own understanding of the set, even though data scientists may feed it largely labelled training data.

4) Reinforcement Learning:

   Data scientists utilize the method of reinforcement learning to train a machine to perform multiple processes with well-defined criteria. To finish a task, data scientists develop an algorithm and give positive or negative feedback to it as it learns how to do so. The algorithm, though, generally decides for itself what steps to take at each junction.

## 2.3 FLOWCHART OF MACHINE LEARNING

The first part in Machine Learning is to collect the data. The procedure of collecting data from the world around us so that it may be processed, presented and stored in a computer for further use is known as data acquisition. Data preprocessing is procedure where we convert our raw data into a form which machine learning algorithms and computers can comprehend and do evaluation.

After successful preprocessing of data, the data is split into train and test dataset. For this work the data is split into 80% training and 20% test data respectively. The training dataset is used to train the classifier, and then the 'test data set' is used to evaluate its performance. It's worth noting that only the training set is available to the classifier during training. The test data set must not be utilized meanwhile the classifier is being trained. The test set will only be available while the classifier is being tested. After that accuracy is found on the test data set and checked whether the model is performing good. If the model is not performing good, hyper parameter tuning is done on it and attempt to achieve better accuracy is tried. Once better accuracy is achieved from the model that model is used to find prediction on our test dataset.

The complete algorithm for Machine Algorithm Technique is shown in Fig. 2.1.

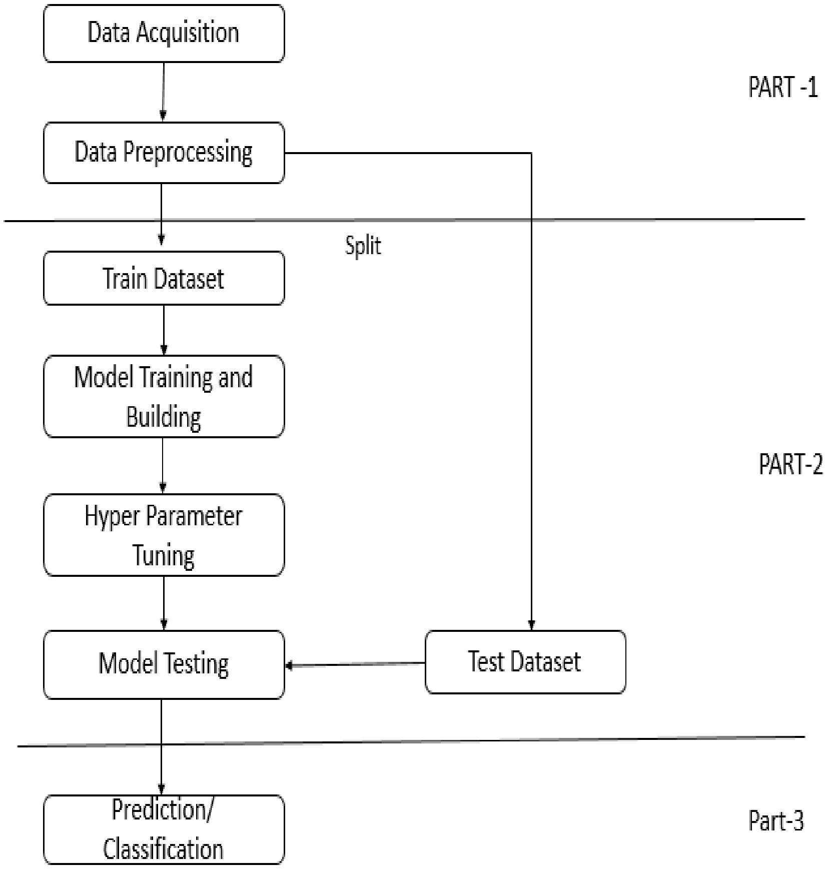


Fig. 2.1: Algorithm for Machine Learning Technique

# CHAPTER 3
# PARTIAL DISCHARGE AND NEUTRAL CORROSION MEASUREMENT

## 3.1 PARTIAL DISCHARGE

A partial discharge (PD) is a kind of electrical discharge in which the gap between two conducting electrodes is not entirely filled. The discharge might take place in a gas-filled vacuum in a solid insulator, around an electrode in a gas, or a gas bubble in a liquid insulator. Corona happens whenever partial discharge occurs in gases. Partial discharge is widely acknowledged as the leading cause of electrical insulation deterioration and failure over time. As a result, measuring it in the production is normal operation for many types of high-voltage equipment. Additionally, partial discharge activities on in-service equipment can be checked for or monitored to warn of impending insulation failure.

Partial discharge can occur in high-voltage insulation faults or gas-filled cavities. The degradation of cable insulation is shown in Fig. 3.1. These flaws might appear in a variety of process:

1) While Manufacturing:

   Solid insulators are made to ensure that electrical stress is distributed evenly across the conducting electrodes. In practice, however, errors in the manufacturing process might result in tiny holes or gaps in the insulating bulk.

2) Installation of Equipment:

   When electrical equipment is factory constructed or installed on site, faults might occur, causing damage to the insulation and therefore weakening it, or increasing electrical stress across the insulation.

3) Aging and Deterioration:

   As internal chemical linkages break down, most insulating materials naturally degrade with time. When exposed to the electrical forces that occur under typical working circumstances, this process renders the insulation fragile and less strong.

4) Over stressed in-service:

   The insulation is likely to experience stress from a short circuit fault or lightning impulse because of a fault current or an overvoltage. Although such episodes are normally brief, the increased electrical stress or warmth caused by the current overload might damage the insulation permanently.

5) In service damage:

   External influences can cause physical harm to electrical equipment while it is in use. Third-party damage to underground cables is particularly vulnerable, as seen by roadworks near buried wires or due to accumulative effect of heavy vehicles driving over them.

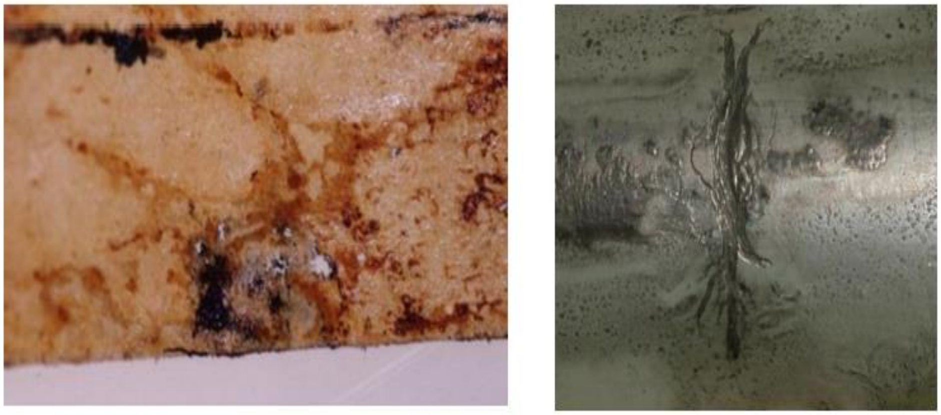


Fig. 3.1: Degradation of cable insulation

## 3.2 ONLINE PARTIAL DISCHARGE MEASUREMENT:

The process for performing partial discharge monitoring in power cables while the cable is powered under normal operating conditions is known as online partial discharge measurement. When safe access to the power line metallic sheath is not accessible, a brief de-energization may be necessary for testing purposes. For checking the insulation state of installed HV equipment, on-line PD measurements have become popular. This form of testing is done when the electrical system is in regular operation.

The most appealing feature of on-line PD measurements for utilities is that the electrical supply is not interrupted during the measurements after the sensors are put in the power grid. Another advantage of on-line tests is that PD activity may be evaluated in temporal or permanent monitoring apps under varied load conditions, which is extremely useful for diagnosing certain types of disorders and tracking their progression over time.

- **High Frequency Current Transformer Sensor**

High Frequency Current Transformers (HFCT) are inductive sensors that may be clamped around the metallic sheath of a power wire that connects to the substation ground. The polarity of HFCT sensors should be pointed towards the ground when they are installed. This will aid in the analysis of partial discharge signals and their direction in the future. Positive pulses travelling in the polarity direction of the High Frequency Current Transformer will produce positive pulses, and vice versa.

The HFCTs are commonly clamped at cable terminations where access to the cable's inner and outer conductors is possible. It comprises of a ferromagnetic core induction coil for measuring transient signals such as partial discharge or pulse-shaped noise interferences. The current of the PD pulses flowing through it is the input and the output is the induced voltage which is measured over the input impedance of the measuring instrument. The layout of HFCT sensor is shown in Fig. 3.2.

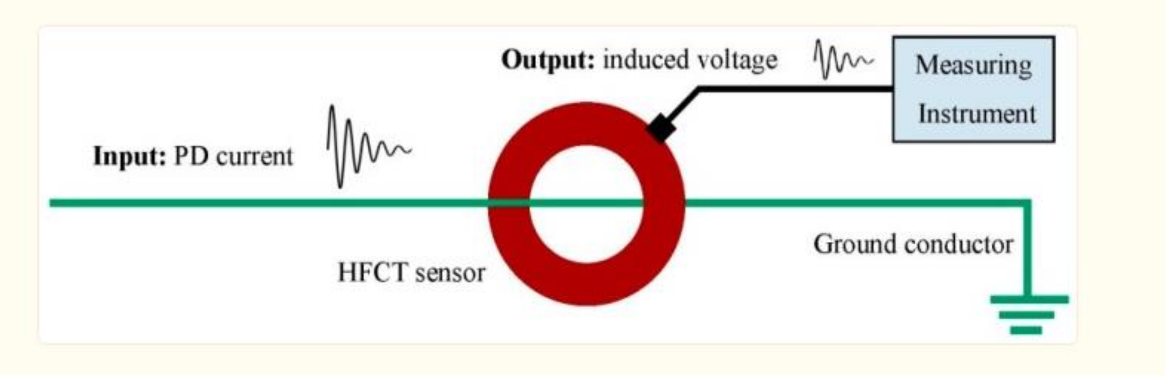


Fig. 3.2: HFCT Sensor

The transfer function of these magnetic sensor can be expressed as:

$$e = -n\frac{d\Phi}{dt} = -n.A.\frac{dB}{dt} = -\mu_0.n.A.\frac{dH}{dt} \quad (3.1)$$

where Φ is the magnetic flux that is passing through the coil of the secondary side which is formed by a number of turns n and presents an area A.

In the case of a coil which has ferromagnetic core,

$$e = -\mu_0.\mu_r.n.A.\frac{dH}{dt} \quad (3.2)$$

The induced voltage in the secondary is proportional to the rate of change of current in the primary, being the mutual inductance between the earth conductor and the secondary M, the proportional constant.

$$e = M\frac{di}{dt} \quad (3.3)$$

## 3.3 NEUTRAL CORROSION

Metallic corrosion may be described as the return of a metal to its natural state in the environment. The degree of corrosion and location can be used by a utility to determine whether to repair, rejuvenate, or replace a cable. Corrosion happens when a chemical reaction occurs in the environment that surrounds something, causing the substance to deteriorate and weaken over time. The pace and type of corrosion are ultimately determined by the gases in an atmosphere that come into contact with a metal. Corrosion can create many problems in the cable. It can lead to discontinuity in the power cable. Unbalanced line-to-neutral voltages, inappropriate functioning, and even damage to underground communication infrastructure can all result from a loss or discontinuity in the power cable. It might also present a threat to one's safety. The neutral corrosion of cable is shown in Fig. 3.3.

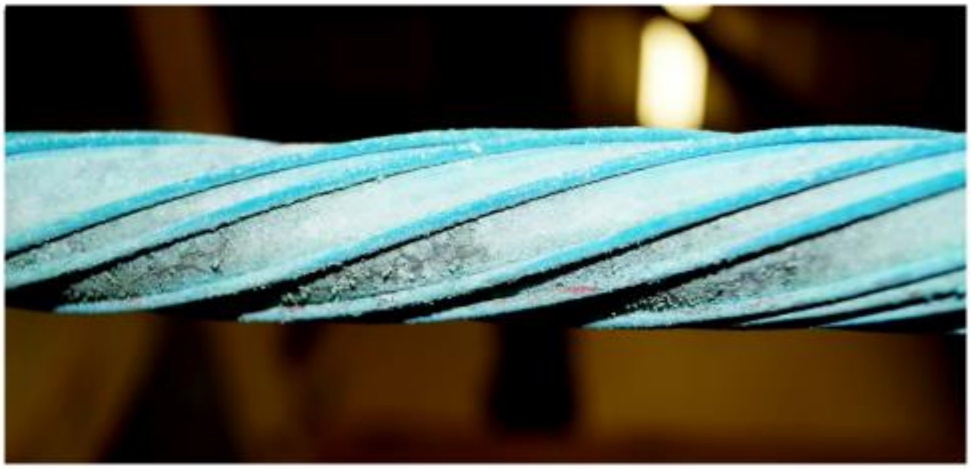


Fig. 3.3: Neutral corrosion of cable

### 3.4 NEUTRAL CORROSION MEASUREMENT

A Time Domain Reflectometer (TDR) is a tool that may be used to locate defects in transmission lines and coaxial cables. A low-voltage pulsed signal is transmitted along the transmission line by the TDR to look for reflections caused by impedance mismatch. No reflections will happen if there is no impedance mismatch along the line; but, if the transmission line ruptures at a particular location, some of the pulsed signal will be reflected back to the TDR. The TDR can pinpoint the precise position of the defect and its nature, such as open circuit, short circuit, or impedance mismatch, by monitoring the timing and propagation velocity of the received pulse.

The TDR (Time domain reflectometer) sends a pulse of energy (difference in potential between two conductors) into a cable being tested. Any time this pulse encounters anything that changes the impedance of the cable, some of the energy is sent back as a reflection. The magnitude of the impedance change determines the reflection's amplitude. The block diagram of TDR is shown in Fig. 3.4.

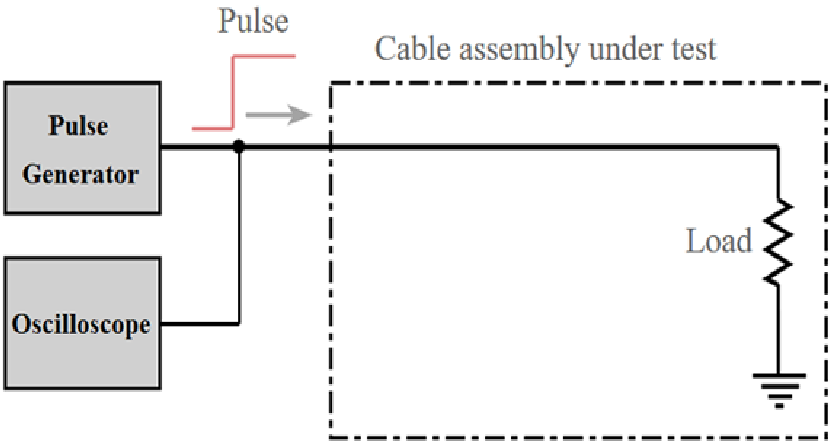


Fig. 3.4 Block diagram of TDR

The time domain reflectometer is made up of a pulse generator and a sampler, as can be shown. An oscilloscope shows waves on the line. In reality, a bit extra signal processing is frequently used to help in the detection of line difficulties and concerns. Only a small portion of the pulse's energy is carried by each neutral wire. If the pulse hits a single severed neutral wire, just that wire's energy is reflected as a positive reflection. The others will continue on their way unaffected by the solitary damaged wire. The reflection from a single damaged wire is so tiny that it is typically undetectable. If more than one neutral wire is broken at the same time, the TDR receives the reflection from each at the same time. The energy emitted by each wire's reflection will add up. The anomaly gets significant enough to be identified on the TDR if enough finally returned at the same time. The form of these abnormalities is highly distinctive. Fig 3.5 shows the example of TDR indicating Neutral Corrosion.

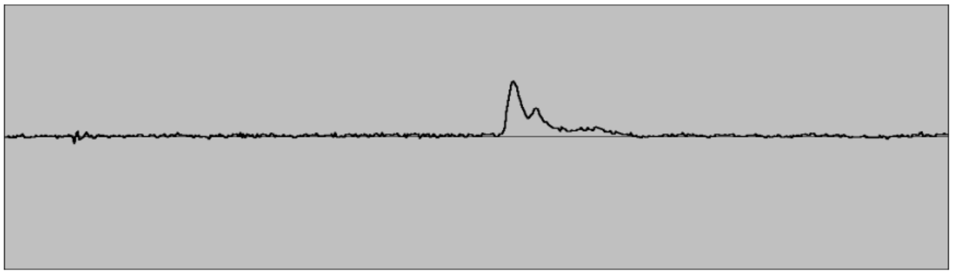


Fig. 3.5 Example of TDR trace indicating neutral corrosion

Determining neutral corrosion in a power wire is a necessary requirement. A utility can examine the degree of corrosion and the location of a cable to determine whether to repair, rejuvenate, or replace it. The neutral wires of the cable may be corroded beyond the utility's permitted limit in some situations. A method for using a TDR to determine the existence and severity of neutral corrosion has been devised.

The severity of corrosion on each cable may be classified using a TDR by comparing it to other cable reflections. The corrosion reflection's amplitude is compared to the reflection of the cable's far end and splices. According to studies, a cable's neutral wires can be destroyed up to 25% of the time before any observable abnormality emerges on the TDR. This is referred to as a level 1 condition. The table 3.1 shows the different corrosion categories.

Table 3.1: Corrosion Categories

| | Wires Broken | Reflection Size |
|---|---|---|
| Level 1 | 0% to 25% | No recognizable reflection |
| Level 2 | 25% to 50% | Recognizable but smaller than a splice. |
| Level 3 | 50% to 75% | Greater than a splice but less than the cable termination. |
| Level 4 | 75% to 100% | Larger than the end of cable reflection |

The fig (3.6-3.10) shows different level of Neutral Corrosion.

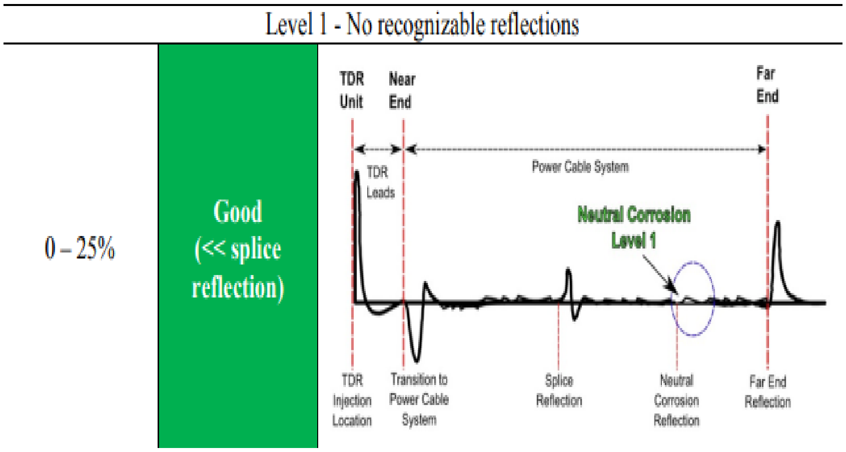


Fig. 3.6: Level 1 Neutral Corrosion

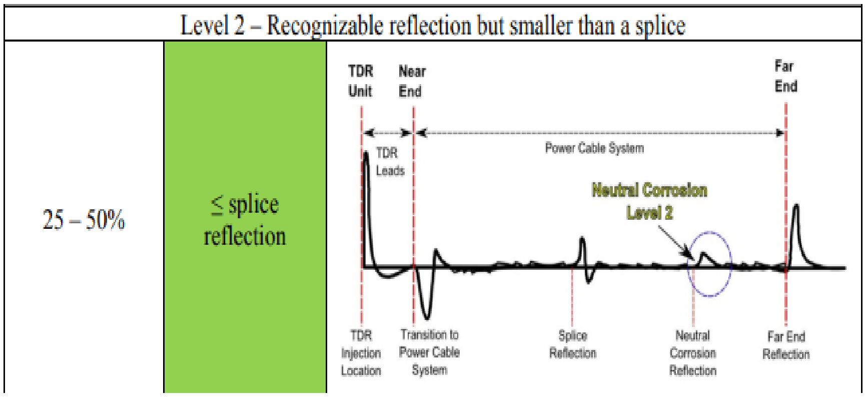


Fig. 3.7: Level 2 Neutral Corrosion

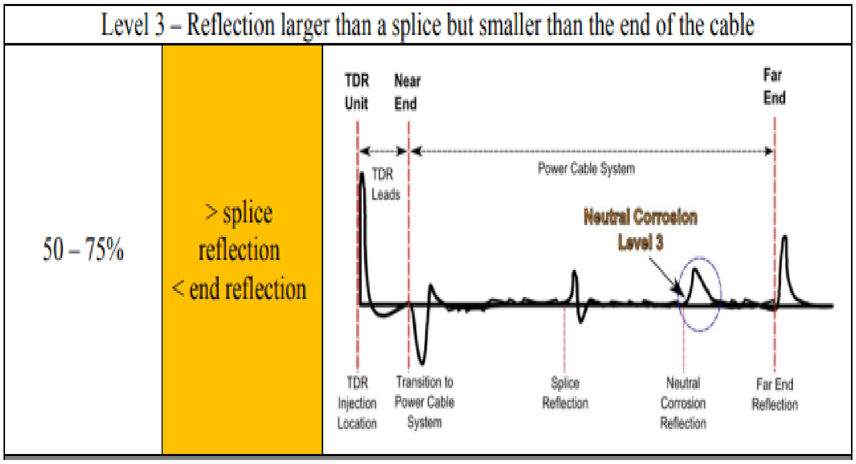


Fig. 3.8: Level 3 Neutral Corrosion

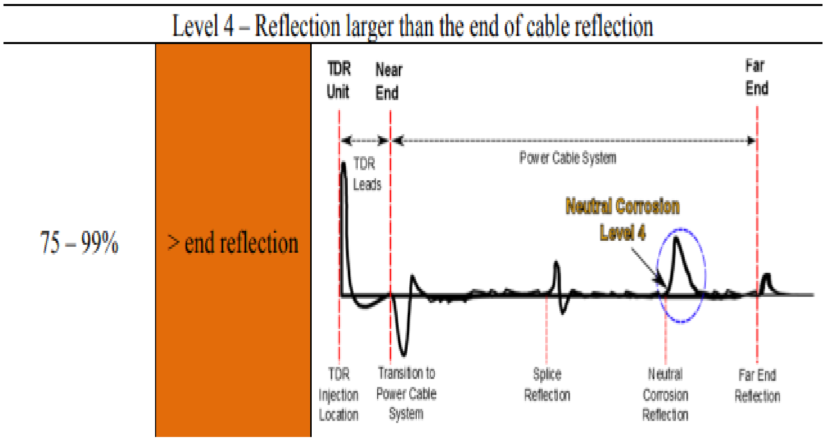

Level 4 – Reflection larger than the end of cable reflection
75 – 99%
> end reflection
TDR Unit
Near End
Far End
TDR Leads
Power Cable System
Neutral Corrosion Level 4
TDR Injection Location
Transition to Power Cable System
Splice Reflection
Neutral Corrosion Reflection
Far End Reflection

Fig. 3.9: Level 4 Neutral Corrosion

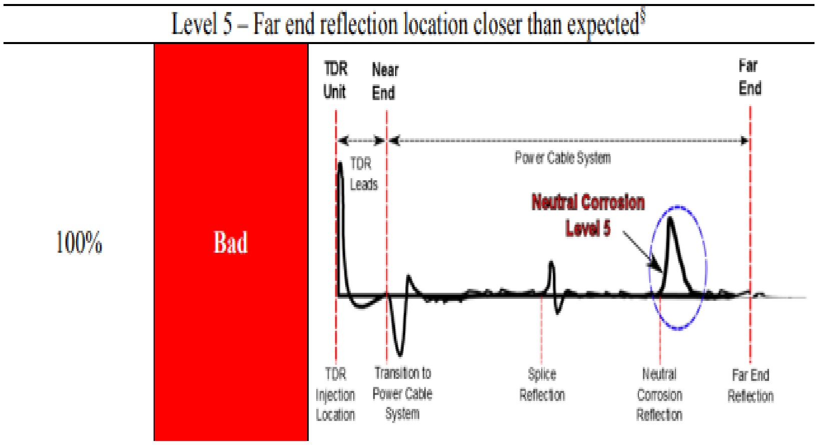

Level 5 – Far end reflection location closer than expected§
100%
Bad
TDR Unit
Near End
Far End
TDR Leads
Power Cable System
Neutral Corrosion Level 5
TDR Injection Location
Transition to Power Cable System
Splice Reflection
Neutral Corrosion Reflection
Far End Reflection

Fig. 3.10: Level 5 Neutral Corrosion

In Level 1 it can be seen that there is no recognizable reflection detected by TDR which means cable condition is good whereas in Level 4 and Level 5 it can be seen that the reflection received is larger than end of cable reflection which means cable condition is not good.

# CHAPTER 4
# DATA COLLECTION AND VISUALIZATION

## 4.1 INTRODUCTION

The data is provided by the Utility Analytics Network, which is a group sharing data and encouraging research and application of utility data analytics for North American utility companies. Data can be found on Kaggle link [12]. It contains 2500 sample cable data points with various XLPE insulation parameters such as partial discharge, ageing, neutral corrosion, loading effect, visual-index and so on. The different parameters are recorded over different service age of cable and total 2500 samples has been created. The sample of data is shown in following Table 4.1:

Table 4.1: Data Samples

| Id | Age | Partial Discharge | Visual Condition | Neutral Corrosion | Loading | Health Index |
|---|---|---|---|---|---|---|
| 1 | 18 | 0.08 | Medium | 0.53 | 646 | 4 |
| 2 | 28 | 0.21 | Medium | 0.71 | 131 | 4 |
| 3 | 27 | 0.19 | Medium | 0.69 | 552 | 4 |
| 4 | 18 | 0.07 | Medium | 0.53 | 155 | 4 |
| 5 | 16 | 0.06 | Good | 0.50 | 349 | 5 |
| . | . | . | . | . | . | . |

Id: Id of the cable

Age: Service age of cable mentioned in years.

Partial Discharge: Partial Discharge of cable which is min-max normalized.

Visual Condition: Visual Condition of XLPE Cable (Poor, Medium, Good)

Neutral Corrosion: Neutral Corrosion of Cable which is min-max normalized.

Loading: Recording Peak Loading of Cable in Amps, Not Affecting Health Index.

In the above data values of Partial Discharge and Neutral Corrosion is min-max normalized.

The Index used in Data is shown in following Table 4.2:

Table 4.2: Health Index Code

| Health Index Code | Health Index Overview |
|---|---|
| Index-1 | Very poor, Red, Need an immediate replacement |
| Index-2 | Poor, Brown, Replacement within the predefined period |
| Index-3 | Moderate, Light green, Maintenance required concerning the time |
| Index-4 | Good, Yellow-green, Normal Maintenance |
| Index-5 | Very good, Green, Wait up to the scheduled maintenance period |

## 4.2 MIN- MAX NORMALIZATION

It is among the most often used methods for normalizing data. The smallest value of each characteristic is turned into a 0, the highest value is transformed into a 1, and all other values are transformed into a decimal between 0 and 1. If a feature's lowest value is 30 and its maximum value is 50,then 40 is changed to around 0.5 because it is halfway between 30 and 50. In our data values of Partial Discharge and Neutral Corrosion is Min-Max Normalized which means that if the values of Partial Discharge and Neutral Corrosion is very high it will approach to 1 or else if the values are less it will approach to zero. The formula is given as

$$\frac{value-min}{max-min} \tag{4.1}$$

Here, min denotes the lowest value and max denotes the highest value.

## 4.3 VISUAL REPRESENTATION OF DATA

Data visualization is the process of presenting information and data graphically. Using visual components like charts, graphs, and maps, data visualization tools make it simple to see and understand trends, outliers, and patterns in data. The complete visual representation of Data can be done through different plots such as Pie chart, Bar Chart and Scatter plots etc. The Fig. 4.1 shows the pie chart representation of Data:

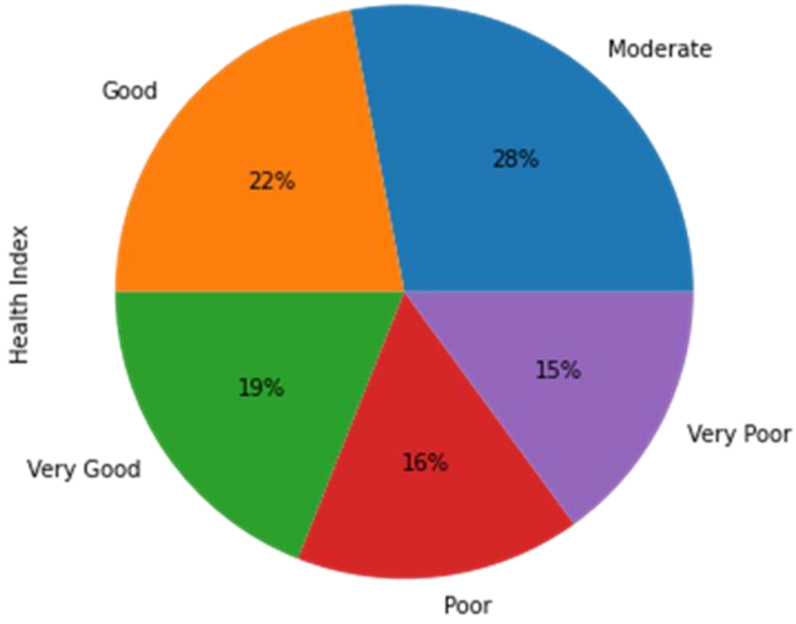


Fig. 4.1: Pie Chart Representation of Data

From the pie chart shown in Fig 4.1, it is analyzed that 28% of dataset consist of Moderate Type Health Index,22% are of Good Type Health Index and similarly 16% of datasets are of Poor Type Health Index.

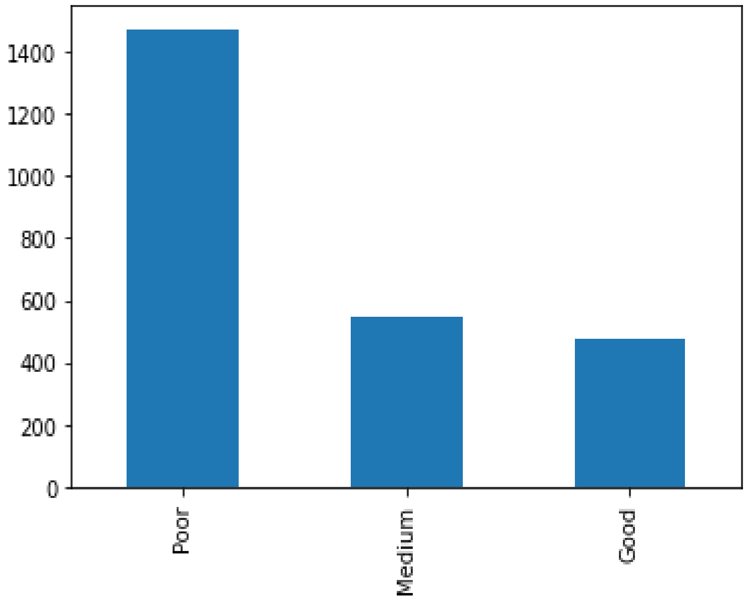


Fig. 4.2: Bar Chart representation of Visual Condition of Datasets

From the bar chart shown in Fig 4.2, it can be seen that out of 2500 samples in the data 1475 samples are with poor visual condition and similarly 550 samples are of medium visual condition and around 475 are of good visual condition.

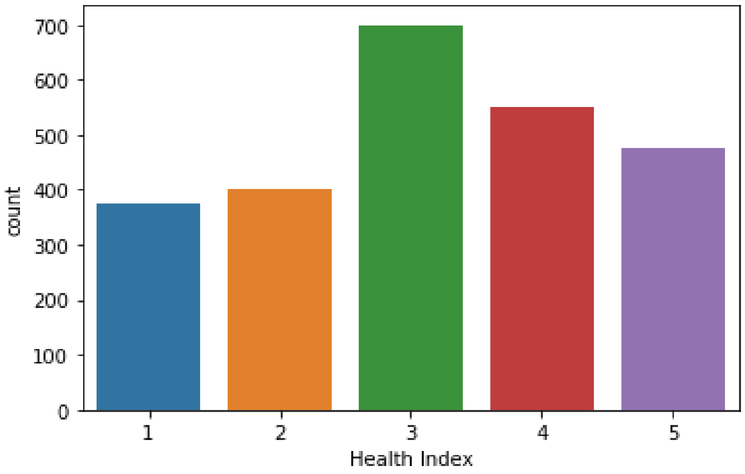


Fig. 4.3: Count of different Health Index Sample

From the bar graph shown in Fig 4.3, it can be analyzed that out of 2500 samples, datasets with Health Index as 3 are mostly present in Data and datasets with Health Index 1 are least present in Data.

Fig. 4.4 and Fig. 4.5 shows some scatter plots that can be inferred from data:

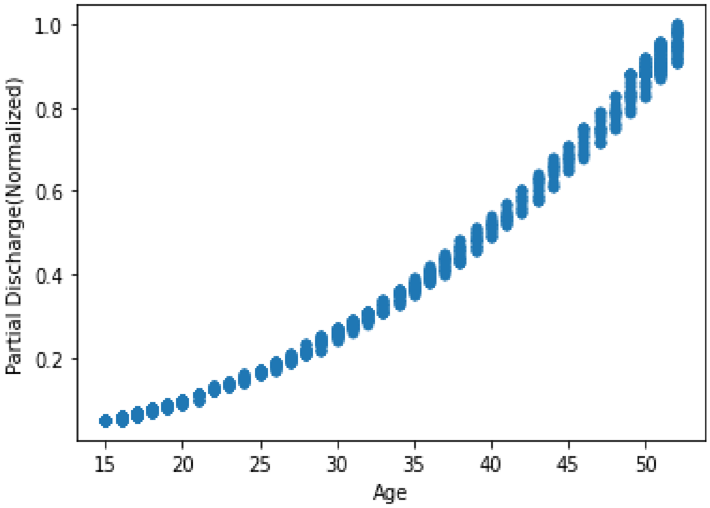


Fig. 4.4: Scatter Plot of Age and Partial Discharge(Normalized)

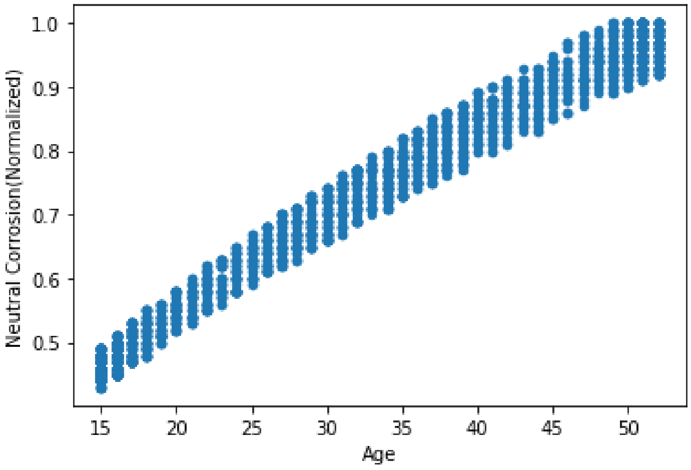


Fig. 4.5: Scatter plot of Age and Natural Corrosion(Normalized)

From the scatter plots shown in Fig. 4.4 and Fig. 4.5, it is interpreted that as the service age of the cable is increasing, Partial Discharge of Cable is also increasing. Similarly, it can be seen that Neutral corrosion also increases with increment of service age of cable which is very obvious.

## 4.4 CORRELATION MATRIX

A table that displays the correlation coefficients between various variables is called a correlation matrix. A matrix represents the correlation between all possible value pairings in a table. The tool is useful for analyzing and visualizing patterns in large datasets.

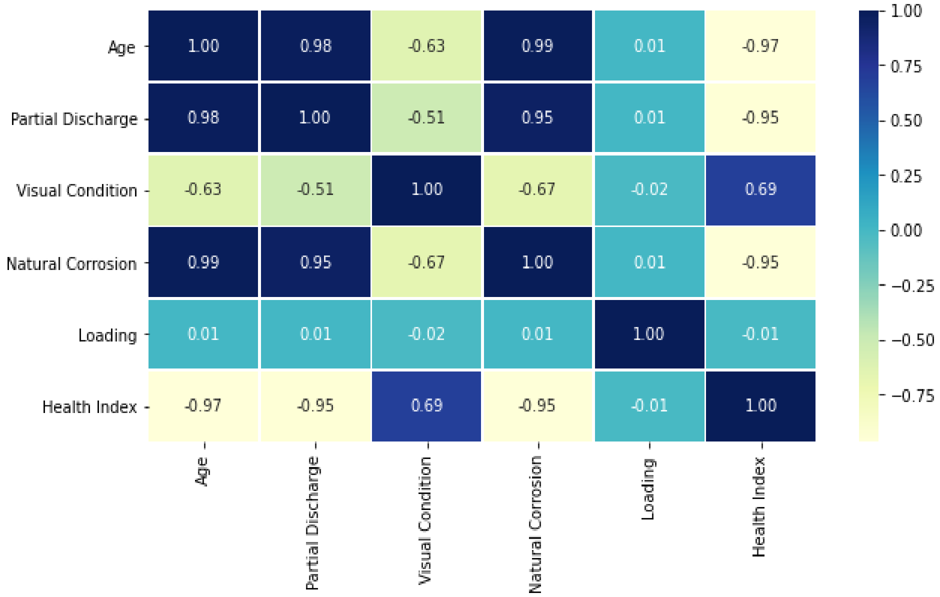


Fig. 4.6: Correlation Matrix of Dataset

Correlation matrix of the entire dataset is plotted in Fig 4.6. From that it can be analyzed that Health Index is strongly correlated with Age, Partial Discharge, Neutral Corrosion and Visual Index of Cable. Basically there is two type of correlation. One is known as positive correlation and the other is negative correlation. A positive correlation is a two-variable correlation in which both variables move in the same direction. When two variables move in opposing directions, such as when one grows, the other lowers, and vice versa, this is known as negative correlation or inverse correlation.

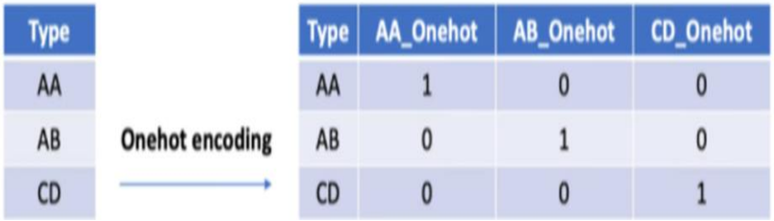
# CHAPTER 5
# DATA PREPROCESSING

## 5.1 INTRODUCTION

Data preprocessing is a step in the data mining and analysis process that transforms unstructured data into a form that computers and machine learning algorithms can understand and interpret.

## 5.2 ONE HOT ENCODING

Machine learning algorithms are unable to work directly on label data. It demands the use of numeric values for all input and output variables. Data that is categorical must be transformed into numerical data. One hot encoding is the conversion of categorical data variables into variables that machine learning algorithms may use to enhance predictions.



| Type |
|---|
| AA |
| AB |
| CD |

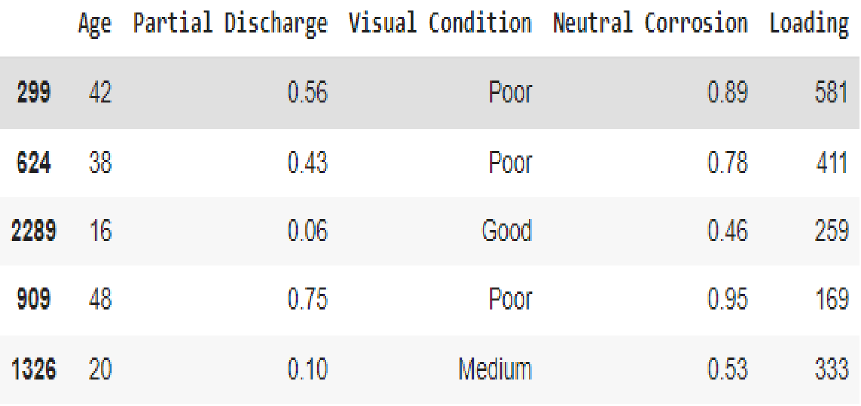
Onehot encoding

| Type | AA_Onehot | AB_Onehot | CD_Onehot |
|---|---|---|---|
| AA | 1 | 0 | 0 |
| AB | 0 | 1 | 0 |
| CD | 0 | 0 | 1 |

Fig. 5.1: Example of One Hot Encoding

Fig 5.1 shows how one hot encoding works. Here 'AA' is coded as 1,0,0. Similarly 'AB' is coded as 0,1,0 and 'CD' is coded as 0,1,0. This way it will be easy to train Machine Learning Model.

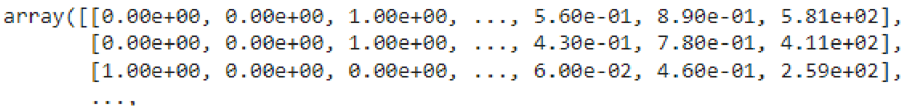
The data shown in Fig 5.2 is the head of the data used in this work:



| | Age | Partial Discharge | Visual Condition | Neutral Corrosion | Loading |
|---|---|---|---|---|---|
| **299** | 42 | 0.56 | Poor | 0.89 | 581 |
| **624** | 38 | 0.43 | Poor | 0.78 | 411 |
| **2289** | 16 | 0.06 | Good | 0.46 | 259 |
| **909** | 48 | 0.75 | Poor | 0.95 | 169 |
| **1326** | 20 | 0.10 | Medium | 0.53 | 333 |

Fig. 5.2: Head of our data

Here it can be seen that ‘Visual Section’ part of our dataset consist of categorical names such as Poor, Good, Medium. It is needed to convert “Visual Section” into numerical so that Machine could understand. After applying one hot encoding on the above data, converted data will look like shown in Fig 5.3:



```
array([[0.00e+00, 0.00e+00, 1.00e+00, ..., 5.60e-01, 8.90e-01, 5.81e+02],
       [0.00e+00, 0.00e+00, 1.00e+00, ..., 4.30e-01, 7.80e-01, 4.11e+02],
       [1.00e+00, 0.00e+00, 0.00e+00, ..., 6.00e-02, 4.60e-01, 2.59e+02],
       ...,
```

Fig. 5.3: Data after one hot encoding

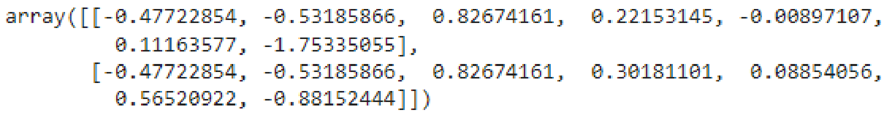
It can be clearly seen that ‘Poor’ has been encoded as 0,0,1 and ‘Good’ and ‘Medium’ as been decoded as 1,0,0 and 0,1,0 respectively.

## 5.3 SPLITTING THE DATA INTO TRAIN-TEST

The train-test split is done on the data. We train our model on the training data and then check the performance of our model on test data. We can quickly and easily use this method to compare the effectiveness of several machine learning algorithms for our problem of predictive modelling. The method involves splitting a dataset into two separate groups. The initial subset used to fit the model is the training dataset. The model is not trained on the second subset; rather, it is given the dataset's input element and asked to make predictions before comparing the actual values to those predicted. The second dataset is referred to as the test dataset. In this work 80% of dataset has been kept for training purpose and 20% of dataset has been kept for test dataset. This means out of 2500 samples, 2000 are for training purpose and 500 for testing purpose.

## 5.4 FEATURE SCALING

A method for standardizing a collection of independent variables or data elements is feature scaling. In data processing, it is sometimes referred to as data normalization and is typically carried out during the data preprocessing stage. Feature selection aids in the rapid computation of algorithms. It is a crucial stage in the data preprocessing process. The machine learning model would give more weightage to higher values and lower weightage to lower values if we didn't use feature scaling. In addition, training the machine learning model takes a long time. For example, we can say that if we have multiple independent variables like age, salary and height with values varying from (18-100 age), (25000-75000 euros), (1-2 meters) respectively, Feature Scaling will convert all in the range of (0,1) depending on scaling technique used. For this work the Standard Scaler library from Scikit-Learn has been used.
After applying Feature Scaling our data will look like shown in Fig 5.4:



```
array([[-0.47722854, -0.53185866,  0.82674161,  0.22153145, -0.00897107,
         0.11163577, -1.75335055],
       [-0.47722854, -0.53185866,  0.82674161,  0.30181101,  0.08854056,
         0.56520922, -0.88152444]])
```

Fig. 5.4: After applying Feature Scaling

It can be clearly seen that all parameters have been scaled.

# CHAPTER 6
# MACHINE LEARNING ALGORITHMS

## 6.1 INTRODUCTION

Machine learning algorithms are the means by which an AI system carries out its objective, which is typically to predict output values from given input data. Regression and classification are the two fundamental stages of machine learning algorithms. The two types of machine learning (ML) algorithms are supervised and unsupervised, respectively. While supervised learning algorithms receive both their input data and their desired output data through labelling, unsupervised algorithms deal with data that has neither been classified nor labelled. In this work different Algorithms has been used but focus has been made on Random Forest Algorithm and ANN model. Following algorithms are used in this work:

- Gaussian Naïve Bayes
- K Nearest Neighbor(KNN)
- Support Vector Machine(SVM)
- Decision Tree Classifier
- Logistic Regression
- Random Forest Classifier
- Artificial Neural Network(ANN)

The accuracy is found through each model and a comparison of accuracy is made through each model. Before applying Machine Learning Algorithm, we should have knowledge of Classification Report and Confusion Matrix and Hyperparameter Tuning.

## 6.2 CLASSIFICATION REPORT

A classification report is used to evaluate how well a classification algorithm predicts the desired output values. It tells about how many predictions are correctly done by our model. True Positives, False Positives, True Negatives, and False Negatives are the parameters that are used to estimate the metrics of a classification report.

True Positive- If the case is positive and it is predicted positive it is called True Positive.

False Positive- If the case is negative but it is predicted as positive it is known as False Positive.

True Negative- If the case is negative and is predicted as negative it is known as True Negative.

False Negative- If the case is negative but it is predicted as positive it is known as False Negative.

Important terms for classification report are:

- Precision
- Recall
- F1 score
- Support

Precision: The ratio of true positives to the sum of true positives and false positives is what it's called. Precision counts the number of positive class predictions that really fall into the positive class.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (6.1)$$

Recall: The ratio of true positives to the sum of true positives and false negatives is what it's called. A metric called recall counts the number of accurate positive predictions that were made out of all available positive predictions.

$$\text{Recall}=\frac{TP}{TP+FN} \qquad (6.2)$$

F1 Score: The weighted harmonic mean of precision and recall is called F1 Score, with 1.0 being the maximum and 0.0 being the minimum.

$$\text{F1 Score} = \frac{2*Recall*Precision}{Recall+Precision} \qquad (6.3)$$

Support: Support is the quantity of instances of the class that actually occur in the dataset that is provided.

## 6.3 CONFUSION MATRIX

An evaluation tool for the performance of machine learning for classification is a confusion matrix. This particular sort of table allows us to assess how well a classification model works when applied to a set of test data for which the real values are known. It gives a comparison between actual and predicted values.

The confusion matrix shows both the types and the number of errors that our classifier is making. This breakdown helps us get over the limitations of depending only on classification accuracy. The instances of that predicted class are shown in each column of the confusion matrix. Each row of the confusion matrix corresponds to an instance of the true class. Confusion Matrix for Binary Classification is shown in below Fig 6.1:

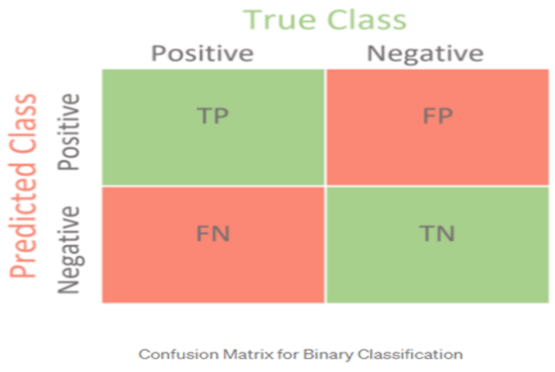


Fig. 6.1: Confusion Matrix for Binary Classification

Confusion Matrix for Multi class Classification is shown in below Fig 6.2:

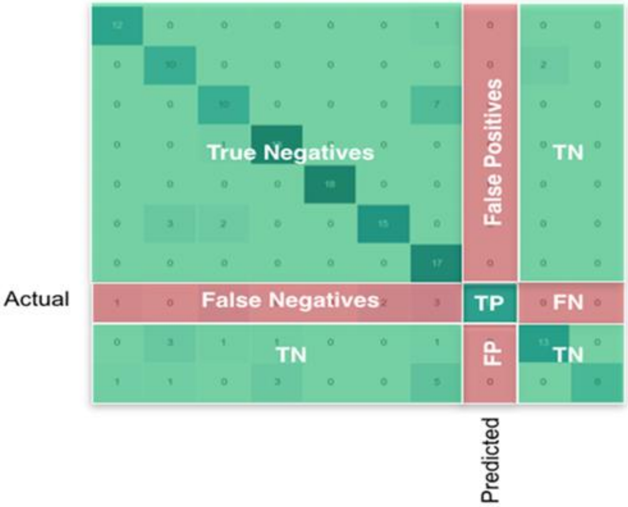


Fig. 6.2: Confusion Matrix for Multi Class Classification

## 6.4 HYPERPARAMETER TUNING

A model argument that has a value predetermined before the learning process begins is referred to as a hyperparameter. Hyperparameter tuning is the fundamental to machine learning algorithms. Hyperparameters are controllable factors that allow us to fine-tune the training process for our models. A key component of controlling the behavior of a machine learning model is hyperparameter tuning. If we don't properly change our hyperparameters, our estimated model parameters produce unsatisfactory results, because they don't minimize the loss function. Therefore, our model commits more errors. The process of determining the optimal configuration of hyperparameters is known as hyperparameter tuning. Finding the parameters that provide the model the highest performance—or the best performance with the lowest error rate—is the aim of hyperparameter tuning.

Steps for doing Hyperparameter tuning:

- Select the model we want to use.
- Checking the parameters of that model.
- Selecting the methods to see hyperparameters of that Model
- Selecting the cross validation approach
- Evaluating the accuracy on those hyperparameters

The two way of doing Hyperparameter tuning is:

- Grid Search CV

- Random Search CV

Grid Search CV: Based on the cross validation score, we check every combination of the current list of hyper-parameter values and pick the best one. It takes more time as it tries different combinations. The advantage is that it gives the best hyperparameters.

Random Search CV: It attempts a variety of value combinations at random (we have to define the number of iterations). It is capable of evaluating a large range of values and, in most cases, quickly achieves a very excellent combination; nevertheless, it cannot promise that it will provide the optimal parameter combination since not all parameter values are tested. This method is recommended on big datasets. The disadvantage is that it does not guarantee we have the best parameters. On the other hand, advantage is that it can give result fast as all the parameters are not tried.

# CHAPTER 7
# RANDOM FOREST CLASSIFIER

## 7.1 INTRODUCTION

The supervised learning technique is used by the well-known machine learning algorithm Random Forest. It can be applied to both classification and regression applications in machine learning. It is based on ensemble learning, a technique for combining various classifiers to tackle a complex problem and improve the performance of the model. Random Forest is a classifier that increases the projected accuracy of a dataset by averaging the outcomes of numerous decision trees applied to different subsets of the dataset. The random forest gathers the predictions from each decision tree and predicts the ultimate result based on the majority votes of predictions, as opposed to relying just on one decision tree. The accuracy increases with the size of the forest and overfitting becomes less of an issue. The working mechanism of Random Forest Algorithm is shown in Fig 7.1.

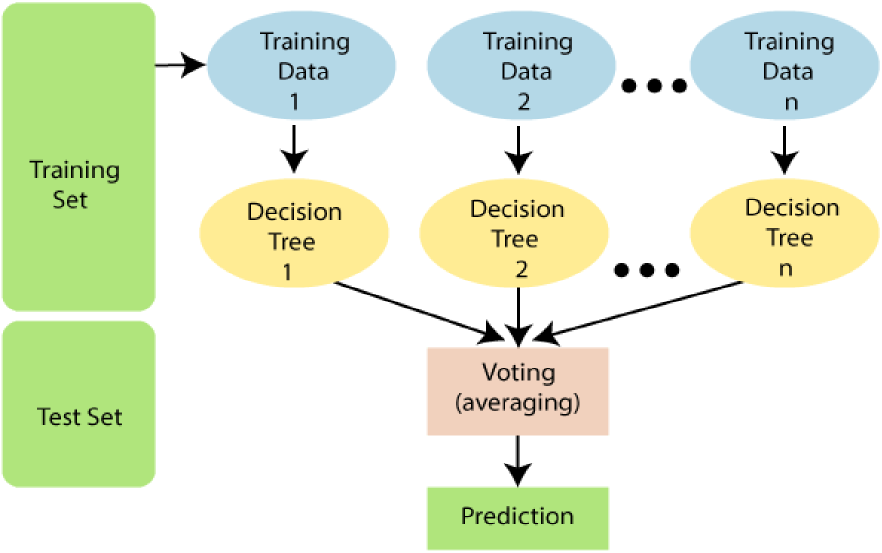


Fig. 7.1: Working of Random Forest Algorithm

A Random Forest is largely comprised of different decision trees. The class with the highest votes becomes the prediction of our model. In the random forest, each tree produces a class prediction. The Random Forest prediction is shown in Fig. 7.2.

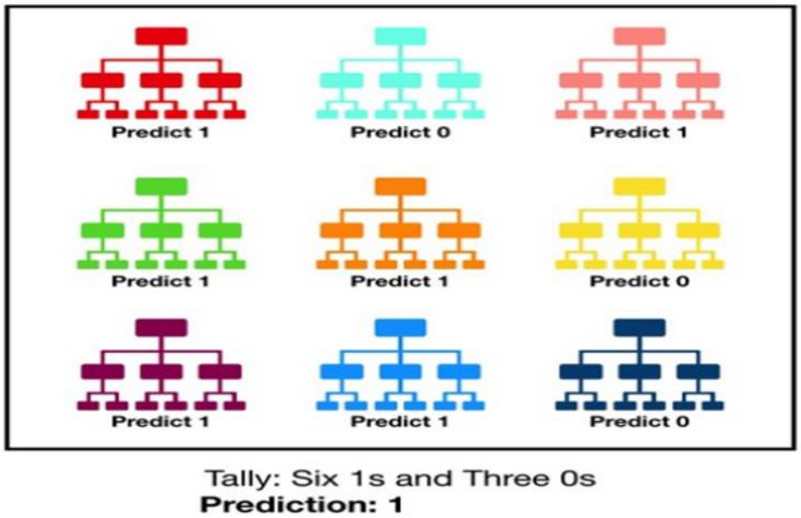


Fig. 7.2: Visualization of a Random Forest Model Making prediction

The figure shown in Fig 7.2 shows that there are number of decision trees that Random Forest make, out of them some are giving prediction as 0 and some are giving prediction as 1. Since the majority of them are giving prediction as 1, the final prediction is predicted as 1.

## 7.2 REASON TO USE RANDOM FOREST ALGORITHM

- Compared to other algorithms, it trains faster.
- It runs rapidly even with a large dataset and makes accurate output predictions.
- Even when a sizable portion of the data is missing, it can still be accurate.

## 7.3 MECHANISM OF RANDOM FOREST

Random Forest uses two methods:

- Bagging (Bootstrap Aggregation)
- Feature Randomness

**Bagging**: Negligible modifications to the training set can produce drastically different tree structures. Decision trees are quite sensitive to the data they are trained upon. In order to create distinct trees, random forest makes use of the ability for each tree to sample randomly from the dataset with replacement. Bagging is the term used to describe this process.

**Feature Randomness:** Each tree in a random forest can only select from a randomized subset of features. This increases the variance among the model's trees and decreases correlation while increasing diversity.

### 7.4 IMPORTANT HYPERPARAMETERS OF RANDOM FOREST

Hyperparameters are utilized in random forest to either speed up the model or increase prediction accuracy. Some important hyperparameters of Random Forest are:

- n_estimators
- max_features
- min_sample_leaf
- n_jobs
- random_state
- oob_score

n_estimators: It's simply the amount of trees the algorithm creates before doing maximum voting or averaging predictions. Although using more trees speeds up computation, it also improves performance and increases the stability of predictions.

max_features: It represents the greatest number of features that a random forest can take into account while splitting a node.

min_sample_leaf: It defines how many leafs are needed to separate an internal node.

n_jobs: It details the number of processors that the engine is allowed to utilize. If the value is one, just one processor may be used.

random_state: It allows the model's output to be reproduced. When the hyperparameters and training data are kept constant and the random state is assigned to a fixed value, the model will consistently deliver the same results.

oob_score: It is a random forest-based cross-validation technique. In this sample, about one-third of the data is not used to train the model but can be used to evaluate how well it performs. They are known as out-of-bag samples.

## 7.5 CONFUSION MATRIX AND CLASSIFICATION REPORT BEFORE HYPERPARAMETER TUNING

The figure shown in Fig 7.3 and Fig 7.4 shows the confusion Matrix and Classification Report before doing Hyperparameter tuning.

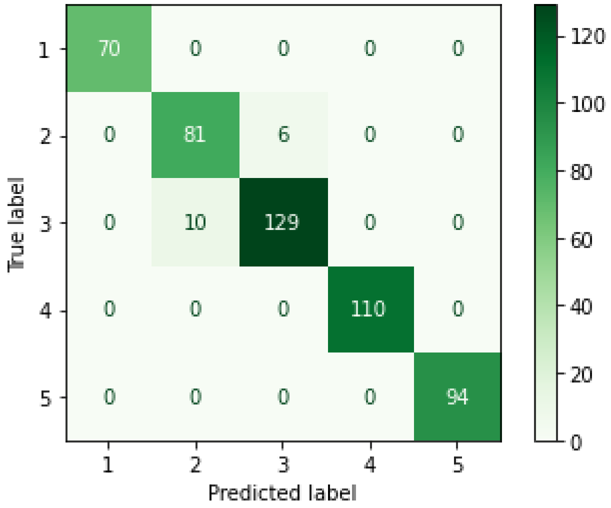
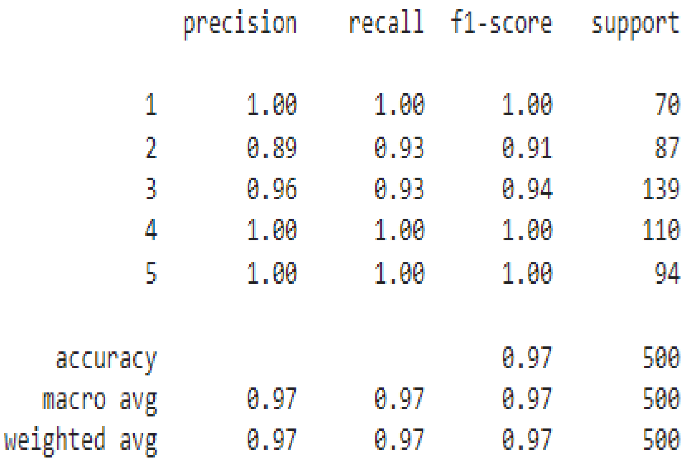


Fig. 7.3: Confusion Matrix for Random Forest Model Before Hyperparameter Tuning

In the figure shown in Fig 7.3 it can be seen that out of 500 test samples only 16 are wrongly predicted and rest are correctly predicted by the model. The diagonal represents the values those are correctly predicted by our model.

Below is the figure for Classification Report for our Random Forest Model:



```
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        70
           2       0.89      0.93      0.91        87
           3       0.96      0.93      0.94       139
           4       1.00      1.00      1.00       110
           5       1.00      1.00      1.00        94

    accuracy                           0.97       500
   macro avg       0.97      0.97      0.97       500
weighted avg       0.97      0.97      0.97       500
```

Fig. 7.4 Classification Report for Random Forest Model

In the Fig. 7.4 it can be clearly seen that achieved accuracy is 97% without doing hyperparameter Tuning. To improve this accuracy, help of Hyperparameter Tuning is taken.

## 7.6 HYPERPARAMETER TUNING OF RANDOM FOREST MODEL

After doing the Hyperparameter Tuning of Random Forest Model using GridSearchCV best Hyperparameters are found. Those Hyperparameters are listed in the Table 7.1:

Table 7.1: Best Hyperparameter Values for Random Forest

| HYPERPARAMETERS | VALUES |
|---|---|
| ➢ MAX_DEPTH | 80 |
| ➢ MIN_SAMPLES_LEAF | 8 |
| ➢ MIN_SAMPLES_SPLIT | 4 |
| ➢ N_ESTIMATORS | 10 |
| ➢ RANDOM_STATE | 0 |
| ➢ CRITERION | ‘gini’ |
| ➢ MAX_FEATURES | ‘auto’ |

## 7.7 CONFUSION MATRIX AND CLASSIFICATION REPORT AFTER HYPERPARAMETER TUNING

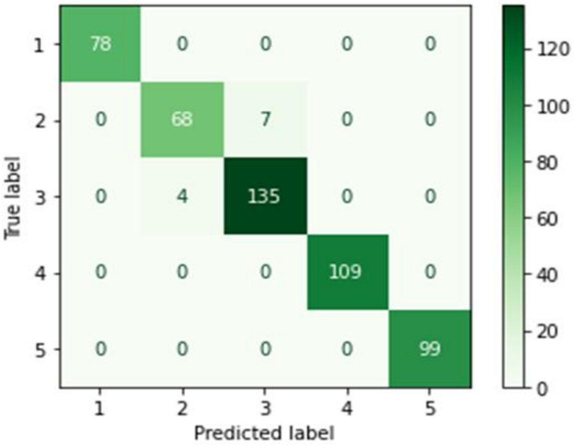
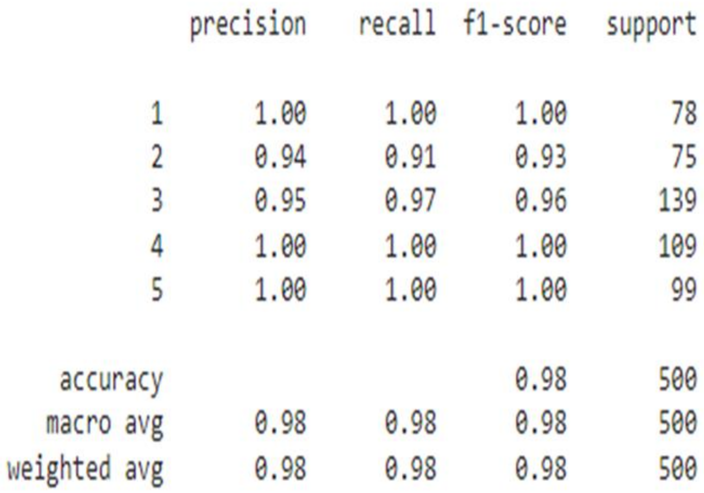


Fig. 7.5: Confusion Matrix after Hyperparameter Tuning

In the confusion matrix shown in Fig 7.5, it is clear that out of 500 samples only 11 are predicted wrong by our model after doing Hyperparameter Tuning. The accuracy can also be checked from the classification report which is shown in figure 7.6:



```
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        78
           2       0.94      0.91      0.93        75
           3       0.95      0.97      0.96       139
           4       1.00      1.00      1.00       109
           5       1.00      1.00      1.00        99

    accuracy                           0.98       500
   macro avg       0.98      0.98      0.98       500
weighted avg       0.98      0.98      0.98       500
```

Fig. 7.6: Classification Report after Hyperparameter Tuning

From the Classification Report it is clear that achieved accuracy through Random Forest Model is 98% after doing proper tuning of Hyperparameters.

### 7.8 COMPARISON OF DIFFERENT MACHINE LEARNING MODEL WITH RANDOM FOREST ALGORITHM

Different Machine Learning Algorithms apart from Random Forest is also implemented in this work to compare the accuracy with other models. The other models are KNN, Naïve Bayes, Decision Tree, SVM, and Logistic Regression.

Model performances before Hyperparameter tuning is shown in Table 7.2

Table 7.2: Comparison of Different Model before Hyperparameter Tuning

| MODELS | ACCURACY |
|---|---|
| ➢ GAUSSIAN NB | 96.4% |
| ➢ KNN | 85.4% |
| ➢ SVM | 30% |
| ➢ DECISION TREE CLASSIFIER | 95.8% |
| ➢ LOGISTIC REGRESSION | 95.2% |
| ➢ RANDOM FOREST CLASSIFIER | 96.8% |

It can be seen that before doing Hyperparameter Tuning Random Forest Classifier gives the best result of 96.8%. Model performances after Hyperparameter Tuning is shown in Table 7.3:

Table 7.3: Comparison of Different Model after Hyperparameter Tuning

| MODELS | ACCURACY |
|---|---|
| ➢ GAUSSIAN NB | 96.4% |
| ➢ KNN | 90% |
| ➢ SVM | 95.8% |
| ➢ DECISION TREE CLASSIFIER | 95.8% |
| ➢ LOGISTIC REGRESSION | 96% |
| ➢ RANDOM FOREST CLASSIFIER | 98% |

After doing the Hyperparameter Tuning there is increase in accuracy in Model performance. Random Forest Algorithm gives the best result with the accuracy of 98%. The other models also performed well after tuning them to certain hyperparameters.

KNN: Parameter used for Hyperparameter Tuning are given in Table 7.4:

Table 7.4: Hyperparameters used for KNN

| HYPERPARAMETERS | VALUES |
|---|---|
| Leaf_size | 1 |
| p | 1 |
| N_neighbors | 1 |

SVM: Parameters used for Hyperparameter Tuning are given in Table 7.5:

Table 7.5: Hyperparameters used for SVM

| **HYPERPARAMETERS** | **VALUES** |
|---|---|
| Kernel | ‘rbf’ |
| Gamma | 0.0001 |
| C | 1000 |

Logistic Regression: Parameters used for Hyperparameter Tuning of Logistic Regression are given in Table 7.6:

Table 7.6: Hyperparameters used for Logistic Regression

| **HYPERPARAMETERS** | **VALUES** |
|---|---|
| C | 100 |
| Penalty | ‘l2’ |
| solver | ‘newton-cg’ |
| max_iter | 10000 |

A graphical comparison of different model has been done which is shown in the Fig. 7.7 listed below:

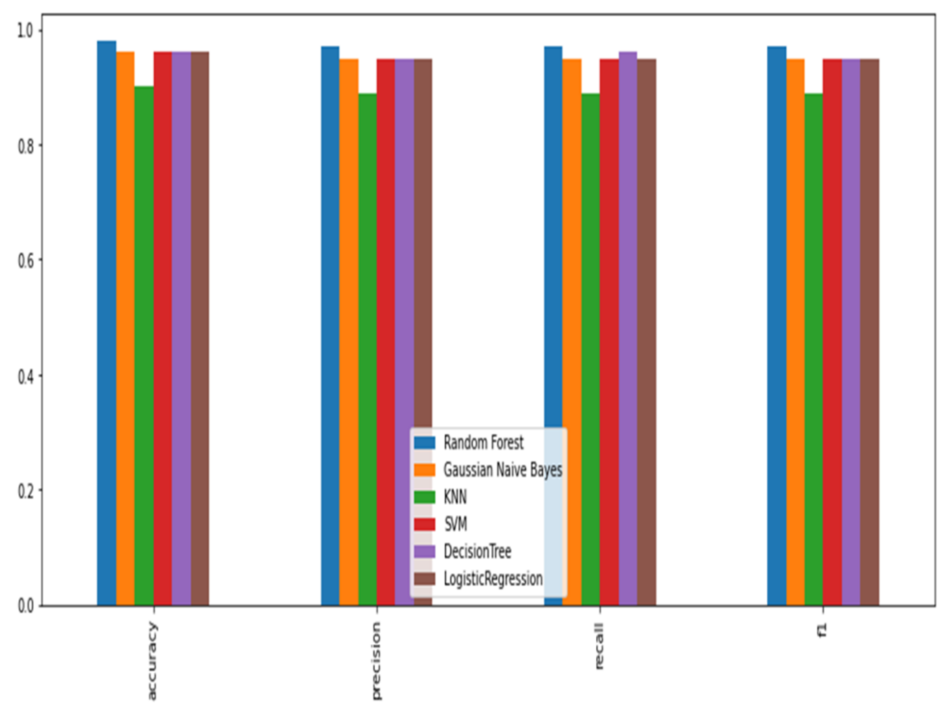


Fig. 7.7: Performance comparison of Models

From the graph plotted in Fig.7.7 it is very clear that Random Forest Model gives better Accuracy, Precision, Recall and F1 score in comparison to other models.

# CHAPTER 8
# ARTIFICIAL NEURAL NETWORK MODEL (ANN)

## 8.1 INTRODUCTION

An artificial neural network is a computer network based on biological neural networks that form the framework of the human brain. Similar to real brains, artificial neural networks contain neurons that are connected to one another at different levels. These neurons are referred to as nodes. Deep learning techniques are based on a subset of machine learning called artificial neural networks (ANNs) and simulated neural networks (SNNs). Their structure and nomenclature are inspired by the human brain, and they behave similarly to how actual neurons interact with one another. The standard Artificial Neural Network resembles the illustration below in Fig. 8.1:

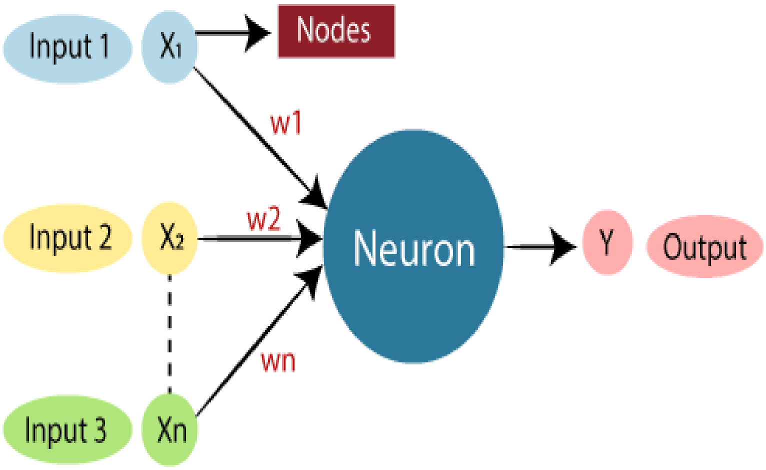


Fig. 8.1: ANN model structure

An Artificial Neural Network is made up of three layers as shown in Fig. 8.2. Those are termed as Input Layer, Hidden Layer and Output Layer.

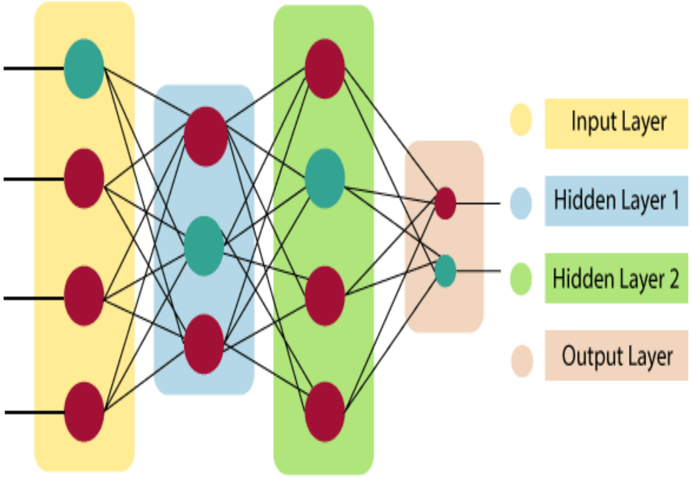


Fig. 8.2: Layers of ANN model

Input Layer: It takes the real dataset provided by the programmer.

Hidden Layer: The hidden layer is located between the input and output layers. To find hidden characteristics and patterns, it performs all the math. In the above figure it can be seen that two hidden layer is used.

Output Layer: The input undergoes a series of modifications via the hidden layer, culminating in output that is sent by this layer. The output layer may contain a single node or more. In a binary classification problem, the output node is 1, however in a multi-class classification problem, there may be more than one output node. The artificial neural network computes the weighted sum of the inputs and also includes a bias. This calculation is expressed in terms of a transfer function.

$$\sum_{i=1}^{n} Wi * Xi + b \qquad (8.1)$$

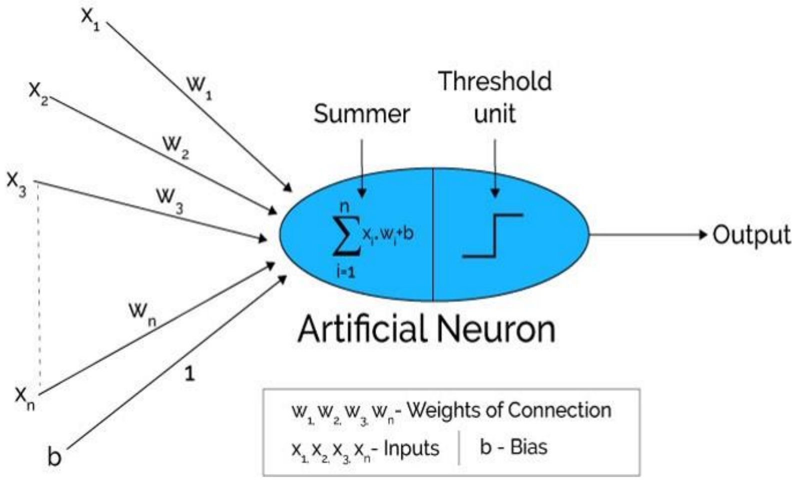


Fig. 8.3: Working of ANN model

The working of ANN Model is shown in Fig. 8.3. The weighted total is provided as an input to an activation function to produce the output. Every node, or artificial neuron, is interconnected with the others and associated with a weight and threshold. A node is activated and data is delivered to the network's upper tiers if its output rises over a predetermined threshold. If not, nothing is transmitted to the network's next tier. In this work a deep neural network with 7 neurons in first and second hidden layers consecutively is implemented.

## 8.2 STEP WISE WORKING OF ANN MODEL

- The units of input are passed. i.e. data is transmitted to the hidden layer with certain weights attached.
- Neurons comprise each hidden layer. Each neuron is linked with all of the inputs.
- After the inputs have been transmitted, all calculations are carried out in the hidden layer.
- Computation is done in hidden layer in two steps:
    - All of the inputs are first multiplied by their corresponding weights. Each variable's weight is its gradient or coefficient. It displays how powerful a specific input is. After the weights have been assigned, a bias variable is added. Bias is a constant which helps in the model's best potential fit.

Z1=W1*X1+ W2*X2 +W3*X3+………+b (8.2)

Here W1, W2, W3 are weights assigned to input X1, X2, X3 and b is the bias.

- In the second stage, the linear equation Z1 is subjected to the activation function. Prior to being transmitted to the next layer of neurons, the input is adjusted nonlinearly by the activation function. In order to introduce nonlinearity into the model, the activation function is essential.

- After passing through each hidden layers it moves to the output layer which gives us the final output. All the process that we did above is known as Forward Propagation.
- After getting predictions from output layer we calculate the error which is difference between input and output layers.
- If the error is significant, actions are made to reduce the error, and Back Propagation is used for the same reason. It is the process of updating and determining the best weights or coefficients to assist the model reduce errors. With the aid of optimizers, the weights are modified. Optimizers are mathematical formulas or methods for changing the properties of neural networks, such as weights, in order to reduce error.

## 8.3 ACTIVATION FUNCTIONS

Activation functions are mathematical equations that decide whether or not a neuron should be triggered based on whether or not the neuron's input is important to the model's prediction. The activation function's objective is to inject nonlinearity into the data.

Different types of Activation Function used are:

- Sigmoid Activation Function
- TanH/ Hyperbolic Tangent Activation Function
- Rectified Linear Unit Function (ReLU)
- Leaky ReLU
- Softmax

In this work Rectified Linear Unit Function (ReLU) is used in the hidden layers and ‘Softmax’ function is used in the output layer.

ReLU - Rectified Linear Units are the most often used activation function in deep learning models. When the function is given a negative value as input, it returns 0, but when it is given a positive value, it returns x.

Graphical representation of ReLU is shown in Fig. 8.4:

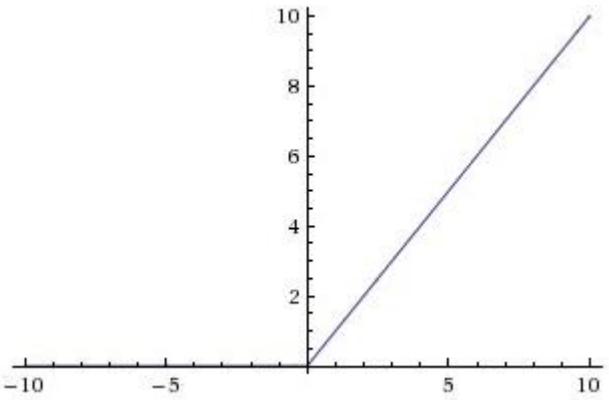


Fig. 8.4: ReLU Function

It's amazing how well such a simple function (made up of two linear parts) can accommodate for non-linearity and interactions in our model. However, the ReLU function performs admirably in the vast majority of situations, and as a result, it is extensively employed.

Softmax - The activation function of a neural network is a crucial element. A basic linear regression model without an activation function is what makes up a neural network. This suggests that the activation function produces the neural network non-linearity. Softmax activation function is used for multi-class classification problem. The equation for Softmax function is given as:

$$\sigma\ (\mathrm{Zi}) = \frac{e^{Zi}}{\sum_{j=1}^{k} e^{Zj}},\ \text{for i=1,2,……N} \qquad (8.3)$$

$\sigma$ = Softmax

Z = input vector

$e^{Zi}$= standard exponential function for input vector k = number of classes in multiclass classifier

$e^{zj}$= standard exponential function for output vector

The exponential is utilized as a non-linear function. By dividing these values by the total of exponential values, these values are then normalized and transformed into probabilities.

## 8.4 OPTIMIZERS

In order to reduce losses, optimizers are methods or tactics that change the properties of our neural network, such as its weights and learning rate.

There is different type of Optimizers:

- Gradient Descent
- Stochastic Gradient Descent
- Mini Batch Gradient Descent
- Momentum
- Nesterov Accelerated Gradient
- Adagrad
- AdaDelta
- Adam

In this work Adam optimizer is used. The method converges too quickly because it is too fast. It rectifies high variance and vanishing learning rate. The method is particularly effective for dealing with large situations with a lot of data or parameters. It uses less memory and is effective. Optimizer used in this work is ‘Adam’.

## 8.5 LOSS FUNCTIONS

The loss function is a mathematical formula used to determine the discrepancy between the algorithm's actual and expected output. In order to assess how well our algorithm models the data, we use this method.

The commonly used loss functions to train Neural Network are:

- Cross-Entropy

- Log loss
- Exponential Loss
- Hinge Loss
- Kullback Leibler Divergence Loss
- Mean Square Error
- Mean Absolute Error
- Huber Loss

Since our problem is of classification Type, Cross- Entropy is used in this work.

**Cross-Entropy** - The purpose of this function is to quantify the difference between two averages of the amount of bits in a distribution of information, which originates from information theory. It determines how two probability distribution functions differ from one another. Since it produces superior generalization models and faster training, cross-entropy is a type of Loss function that is frequently used in machine learning. It is employed in binary and multiclass problems.

Types of cross-entropy:

- Binary cross-entropy
- Categorical cross-entropy
- Sparse cross-entropy

Binary cross-entropy: This is used for classification problems which deals with 0 or 1.

Categorical cross-entropy: It is used for binary and multiclass problems. But the labels need to be encoded as categorical i.e. One hot encoding representation. (for 4 classes: [0, 1, 0,0], [1,0,0,0] …)

Sparse cross-entropy: It is used for binary and multiclass problems. Here the label is an integer (0 or 1 or……, n, depends on number of labels).

In this work sparse cross-entropy is used as our labels are of integer type.

## 8.6 KERAS – DENSE LAYER

The dense layer is the typical deep layer of a neural network. It is the most well-known and frequently used layer. A dense layer is one that is closely coupled to the layer above it, which means that every neuron in the layer is related to every other neuron in the layer above it. In a model, each neuron in the preceding layer sends information to the neurons in the dense layer, which then performs matrix-vector multiplication. The dense layer applies the following operation to the input, and the output is then returned.

$$\text{Output} = \text{activation (dot(input, kernel) +bias)} \qquad (8.4)$$

input –It represents input data

kernel- represents the weight data

dot- it represents numpy dot product of all input and its corresponding weight

bias – it represents a bias value used in machine learning to optimize model

activation – it represents activation function

## 8.7 EPOCHS

In machine learning, an epoch is a whole iteration of the algorithm through the training dataset. The method's most important hyperparameter is the number of epochs. It specifies the number of full runs or epochs for the entire training dataset during the algorithm's learning phase. Each epoch modifies the internal model parameters of the dataset. In this work the dataset is trained on 400 epochs.

## 8.8 PERFORMANCE LEARNING CURVE AND OPTIMIZATION LEARNING CURVE

A learning curve is simply a graph that depicts the progression of a given learning measure over time during the training of a machine learning model. They resemble a mathematical illustration of the learning process. In this, the x-axis represents time or progress, and the y-axis represents error or performance.

- Performance Learning Curve: Those learning curves are determined by the metrics, such as accuracy, precision, recall, or F1 score, that will be used to evaluate and choose the model.

- Optimization Learning Curve: Learning curves based on the metric used to optimize the model's parameters, such as loss or mean square error.

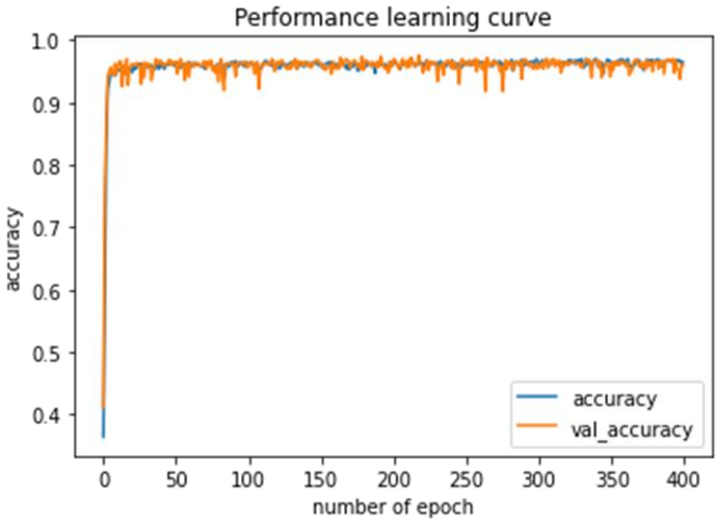


Fig. 8.5: Performance Learning Curve

From the performance learning curve shown in Fig. 8.5 trained on 400 epochs it can be seen that the model is learning very well on training dataset.

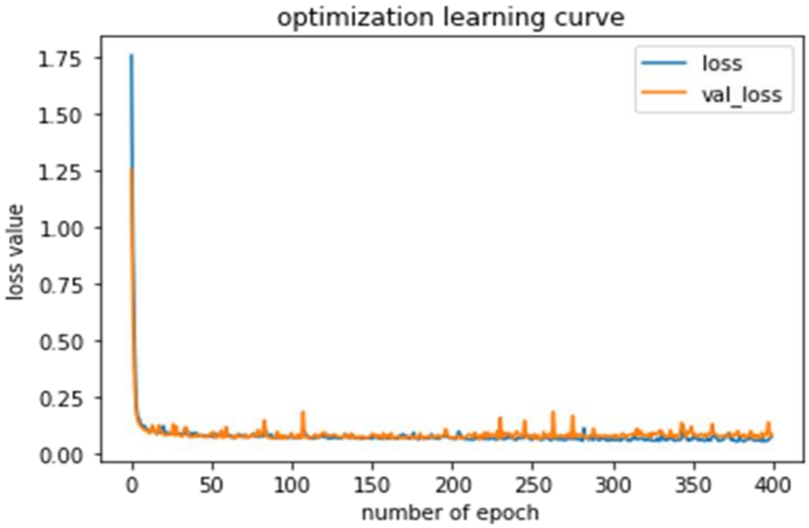


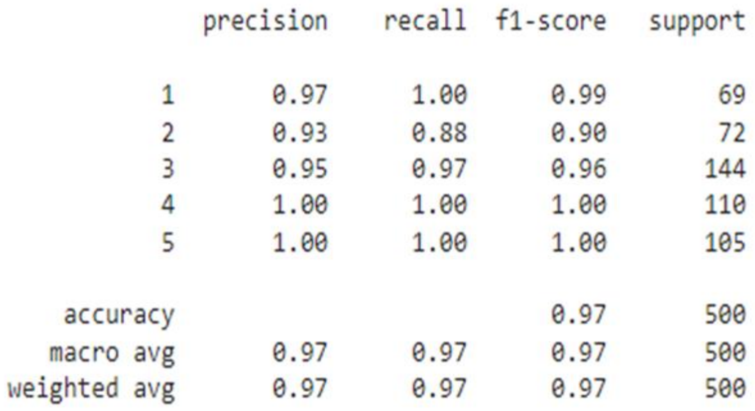
Fig. 8.6: Optimization Learning Curve

From the Optimization Learning Curve, shown in Fig. 8.6 it can be understood that the loss is decreasing as the number of epochs is increasing on which data is trained. There is two graph plotted. One is for train learning curve and other is for validation learning curve. Learning curve gives the idea of how well our model is learning on training dataset whereas validation curve tells how model is generalizing from hold-out validation dataset.

It is regarded as a good match when the training and validation loss slowly lowers to a stable value with a negligible difference between the two final loss values. Here it can be clearly seen that there is minimal gap between training and validation loss hence giving a good fit.

### 8.9 CLASSIFICATION REPORT AND CONFUSION MATRIX FOR ANN MODEL

The classification report for ANN Model is plotted in Fig. 8.7:



```
              precision    recall  f1-score   support

           1       0.97      1.00      0.99        69
           2       0.93      0.88      0.90        72
           3       0.95      0.97      0.96       144
           4       1.00      1.00      1.00       110
           5       1.00      1.00      1.00       105

    accuracy                           0.97       500
   macro avg       0.97      0.97      0.97       500
weighted avg       0.97      0.97      0.97       500
```

Fig. 8.7: Classification Report for ANN Model

From the Classification Report it can be concluded that accuracy obtained from ANN model is 97%.

The confusion matrix for ANN model is plotted in Fig. 8.8:

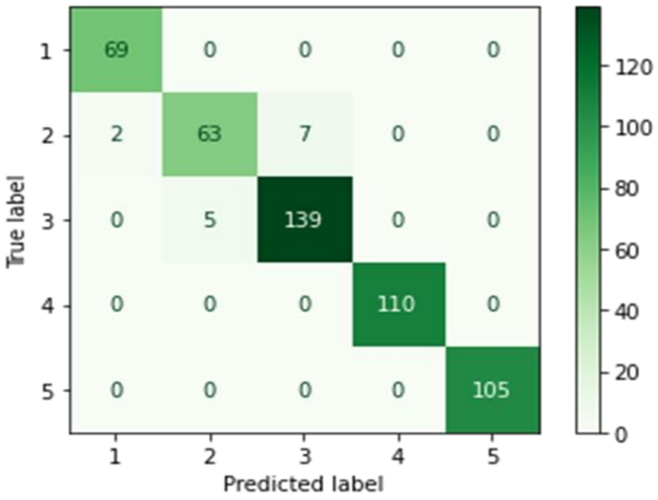


Fig. 8.8: Confusion Matrix for ANN Model

From the confusion matrix it can be see that our Model has predicted only 14 wrong results out of 500 samples. The accuracy is found to be 97.2%.

# CHAPTER 9
# CONCLUSION

The health index score is one of the best ways to determine how long an underground cable's insulation will last after it is put to use. Additionally, it demonstrates the probability of insulation failure, or how the status changes from routine maintenance to urgent replacement. Maintenance procedures can be optimized, and replacement plans can be implemented as a result of this prediction. HI category identification and performance evaluation for each class are provided through a comparison of various machine learning classifiers. This study primarily focuses on the different classes of the insulation dataset so that, based on the observed health index score, relevant prediction steps can be made early on to prevent unforeseen breakdown-like events.

In this work focus has been laid especially on Random Forest Algorithm and ANN Model. The reason behind using Random Forest Model is due to following reasons:

- It requires less time to train than other algorithms.
- It runs rapidly even with a large dataset and makes accurate output predictions.
- Even when a sizable portion of the data is missing, it can still be accurate.

With the Random Forest Algorithm, accuracy of 98% is achieved with only 11 values predicted wrong out of 500 in Confusion Matrix. Also ANN Model is implemented which gives the accuracy of 97.2%. Predictive decisions can be made early on with the help of the predicted Health Index, allowing us to prevent cables from breakdown-like occurrences.

# REFERENCES

[1] Klerx, M.H.P.; Morren, J.; Slootweg, H. “Analyzing Parameters That Affect the Reliability of Low-Voltage Cable Grids and Their Applicability in Asset Management”. IEEE Trans. Power Deliv. **2019**, 34, 1432–1441.

[2] Chimunda, S.; Nyamupangedengu, C. “A Reliability Assessment Model for an Outdoor 88kV XLPE Cable Termination. Electr. Power Syst”. Res. **2019**, 177, 105979.

[3] R. Sahoo, S. Karmakar and S. Panigrahy, "Health Index Analysis of XLPE Cable Insulation using Machine Learning Technique," 2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)”, 2020, pp. 1-6

[4] R. Sahoo and S. Karmakar, "Health Index Prediction of Underground Cable System using Artificial Neural Network," 2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology(ODICON), 2021, pp. 1-4

[5] Dong, Ming, Wenyuan Li, and Alex Nassif, “Long-term Health Index Prediction for Power Asset Classes Based on Sequence Learning” arXiv preprint, arXiv:2006.14193, 2020.

[6] G. C. Montanari, P. Seri and R. E. Hebner, "A scheme for the Health Index and residual life of cables based on measurement and monitoring of diagnostic quantities," 2018 IEEE Power & Energy Society General Meeting (PESGM), 2018, pp. 1-5

[7] "IEEE Guide for Detection, Mitigation, and Control of Concentric Neutral Corrosion in Medium-Voltage Underground Cables," in IEEE Std 1617-2007, vol., no., pp.C1-18, 18 Feb. 2007, doi: 10.1109/IEEESTD.2007.4454717.

[8] Khan, N. Malik, A. Al-Arainy and S. Alghuwainem, "A review of condition monitoring of underground power cables," 2012 IEEE International Conference on Condition Monitoring and Diagnosis, 2012, pp. 909-912, doi: 10.1109/CMD.2012.6416300.

[9] G. C. Montanari, P. Seri and R. E. Hebner, "A scheme for the Health Index and residual life of cables based on measurement and monitoring of diagnostic quantities,"

2018 IEEE Power & Energy Society General Meeting (PESGM), 2018, pp. 1-5, doi: 10.1109/PESGM.2018.8585858.

[10] M. Dong, W. Li and A. B. Nassif, "Long-Term Health Index Prediction for Power Asset Classes Based on Sequence Learning," in IEEE Transactions on Power Delivery, vol. 37, no. 1, pp. 197-207, Feb. 2022, doi: 10.1109/TPWRD.2021.3055622.

[11] Fernando Alvarej, Fernando Garnacho, Javier Ortigo, “Application of HFCT and UHF Sensors in On-Line Partial Discharge Measurements for Insulation Diagnosis of High Voltage Equipment.”

[12] Link :- https://www.kaggle.com/datasets/utilityanalytics/utility-underground-cable-dataset1. (Online)