INTRUSION DETECTION IN SMART ENERGY METER

A PROJECT REPORT

submitted by

CHRISTINE MARIAM MAMMEN
(Reg. No. TKM20EEII10)

to

the APJ Abdul Kalam Technological University
in partial fulfillment of the requirements for the award of the Degree

of

Master of Technology
in
Electrical and Electronics Engineering

with specialisation in

Industrial Instrumentation and Control

Department of Electrical and Electronics Engineering
TKM College of Engineering
Kollam - 691005
KERALA

JULY 2022

DECLARATION

I undersigned hereby declare that the project report entitled "**Intrusion Detection In Smart Energy Meter**", submitted for partial fulfillment of the requirements for the award of degree of Master of Technology in Electrical and Electronics Engineering with specialisation in Industrial Instrumentation and Control, of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of *Dr. Sheeba .R*, Professor, Department of Electrical and Electronics Engineering. This submission represents my ideas in my own words and where ideas or words of others have been included. I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Kollam
July 01, 2022

CHRISTINE MARIAM MAMMEN

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

TKM COLLEGE OF ENGINEERING

Kollam - 691005

CERTIFICATE

This is to certify that the report entitled " **Intrusion Detection In Smart Energy Meter** " submitted by **CHRISTINE MARIAM MAMMEN** , **(Reg. No. TKM20EEII10)** of fourth semester to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Electrical and Electronics Engineering with specialisation in Industrial Instrumentation and Control, is a bonafide record of the project work done by her under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Dr. Sheeba. R
Project Supervisor
Professor
Department of EEE
TKM College of Engineering

Prof. Sumayya Jaleel
Project Coordinator
Assistant Professor
Department of EEE
TKM College of Engineering

Prof. Shanavas T N
PG Coordinator
Associate Professor
Department of EEE
TKM College of Engineering

Dr. Sabeena Beevi K
Head of the Department
Associate Professor
Department of EEE
TKM College of Engineering

Acknowledgement

A lot of effort and hard work has been put into this project in course of its presentation. However, it would not have been possible without the kind support and help of many individuals and other sources. I would like to extend my sincere thanks to all of them. I take this opportunity to express my deep sense of gratitude and sincere thanks to all who helped me to complete this project report successfully.

I express my sincere thanks to *Dr. T A Shahul Hameed*, Principal, TKM College of Engineering, for his encouragement in the completion of my project.

I thank *Dr. Sabeena Beevi K*, Head of the Department, Department of Electrical and Electronics Engineering, *Dr. Imthias Ahamed T P*, Professor, Department of Electrical and Electronics Engineering, and *Prof. Shanavas T N*, Associate Professor, PG Coordinator, Department of Electrical and Electronics Engineering for their support and cooperation.

I am greatly thankful to my project guide *Dr. Sheeba.R*, Professor, Department of Electricaland Electronics Engineering and *Mr. Naufal N*, for their supervision, assistance and helpful suggestions.

I am deeply indebted to *Prof. Sumayya Jaleel*, Assistant Professor,Project Coordinator,Department of Electrical and Electronics Engineering, *Prof. Amal A*, Assistant Professor, Department of Electrical and Electronics Engineering, and *Prof. Sumod Sundar*, Assistant Professor, Centre for Artificial Intelligence, for their excellent guidance, positive criticism and valuable comments.

Finally I thank my parents, friends near and dear ones who directly and indirectly contributed to the successful completion of my project project.

CHRISTINE MARIAM MAMMEN

Abstract

Smart metres and the Internet of Things (IoT) were increasingly used to replace conventional analogue metres in today's modern smart home. It converts collected data of meter readings into digital format. The data can be delivered wirelessly, which decreases the amount of human labour required. Smart Meters, on the other hand, bring a slew of new ways to steal electricity. Using advanced tools or cyberattack techniques, malicious users can break into smart meters. Every year, this illegal conduct results in a significant financial loss. Energy theft detection techniques face a difficult task as a result of this. The advance metering infrastructure (AMI) has the capability of monitoring each consumer's consumption details, tracking their patterns of consumption, billing them, and detecting variationsWith the help of the smart grid's communication capabilities, utilities have been able to save their customers' usage details. This database can be used to develop a theft detection model. Artificial intelligence based technologies are widely used in AMI, which deploys machine learning algorithm to detect prospective electricity thieves, frequentlyThe most common classification approaches involve utilising labels to identify unusual trends in customers' previous electricity usage data and then detecting possible electricity theft behaviors. In this work,the supervised learning techniques were used to detect electricity theft. To assess classification accuracy, comparison of several machine learning classifiers such as Support Vector Machine, Nave Bayes, Decision Tree, and RandomForest,is also presented and the classification is accurately done by Deision Tree Classifier(99.67 Percentage).Unsupervised learning techniques such as Kmeans and DBSCAN are also used to quickly spot irregularities in the readings.The effectiveness of the models were examined using the data from simulated attacking system.

Contents

List of Figures

Abbreviations

AMI Advanced Metering Infrastructure

DBSCAN Density-Based Spatial Clustering of Applications with Noise

GI Gini Index

IGR Information Gain Ratio

IOT Internet of Things

NTL Non Technical Losses

PCA Principal Component Analysis

SVM Support Vector Machine

TL Technical Losses

Chapter 1

Introduction

1.1 Overview

In today's modern world, the power grid has been a necessity. So many countries have been modernizing their existing grid system into smart grid, which facilitates bi-directional communication, high stability, feedbacking of real-time usage details, self-healing, and security, owing to the advancements in information and communication technology. AMI has become a critical component of the smart grid and is closely associated with people's daily lives[1]. AMI automates the electric metering system by replacing the traditional metres with smart metres that allow utility companies and energy users to communicate in real time.AMI integrates smart metres and IoT controlling devices that may collect enormous amounts of data in a short amount of time. AMI's rich information interchange and multilevel semi-open network structure, on the other hand, expand the security flaws for metering across entire public networks and offer numerous cyber-attack vulnerabilities. Data that varies from normal and predicted patterns are referred to as anomalies in IoT[2]. The electric utility experiences substantial income losses as a result of the electricity thieves. Technical Losses (TL) and Non-Technical Losses (NTL) are the two kinds of electricity losses in transmission and distribution (NTL). TL is caused by power losses in overhead power wires, transformers, and other substation devices. Electricity theft is the most common kind of NTL. Electricity theft is described as the use of energy without the permission of a power company. Direct hooking, bypassing the electricity metre, energy bribery of unauthorized connections, interfering with the metre reading, and bypassing the energy meter are all examples[3]. It is responsible for considerable revenue losses as well as a reduction

in electricity quality. According to a recent report, global power utility firms lose more than 20 billion dollar per year. Both industrialised and developing countries are affected by the NTL. In Pakistan, for example, losses of 17.5 percent in energy transmission and distribution were observed in the years 2017–2018. Each year, India loses roughly 4.5dollar billion due to electricity theft. According to a recent assessment, unlawful electricity use accounts for 20 percentage of total electricity consumption in India. Rich countries are also affected by this issue. Illegal electricity consumption costs the United States roughly 6 billion dollar per year, while power losses in the United Kingdom can cost up to 175 dollar million per year. Furthermore, electricity theft might have consequences on the power system's functioning and reliability. It reduces the quality of power by overloading transformers and voltage fluctuations[4].

1.2 Objective

The major consequences of electricity theft are financial losses for utilities, and imbalance in generation and Demand.This scenario arises the need for a smart theft detection system.Smart meters in an IoT based system allows the transfer of data as well as it is possible to store these data's.By utilizing the stored information from smart meter and Artifical Intelligence Techniques ,efficient theft identification model can be developed.Unsupervised learning methods were used to detect abnormalities from the plot obtained by plotting the consumption values.In supervised learning,these previous information can be used for training the model and new real time is given for the system to classify Fault and Normal values.

The remainder of this report is organized as follows. Chapter 2 provides the idea of already the existing theft detection methods. Chapter 3 contains problem statement and methodology employed. Chapter 4 describe about various supervised models and Chapter 6 explains about various unsupervised models used. Chapter 6 comprises of result analysis and validation. Finally, Chapter 7 draws conclusions.

Chapter 2

Literature Survey

2.1 Introduction

To provide the background regarding the currently using methos, this section first briefly describes about false data injection type attack in AMI . Then analysed various techniques proposed in other works.

Energy theft has been becoming a critical problem in the Smart Grid system. Many countries have suffered significant losses in the billions of dollars. A smart metre is now fixed at the end point of every distribution system to track energy utilization pattern and create energy record remotely. Invasion of intruders in smart home appliances and, more commonly, straight hooking in another families' electrical supplies are two prominent means of energy theft[5]. Mess around with the smart meter's software and mechanism, as well as modifying data through cloud storage, are some of the other tactics used. As a result, the invader can lower their own energy usage by exploiting other families' electricity usage via tampering and hacking, as long as the total cost for all consumers in the community remains the same. False data injection type attack comes under the electricity theft situation.The home with higher energy usage can minimize their own electricity usage by drawing on the power consumption of another household through energy theft. It raises the electricity bills of the other family victim while lowering the expenses of the energy theft perpetrator. Researchers have recently concentrated their efforts on the development of improved theft detection systems based on artificial intelligence approaches. In[6],the new modern structure and security consideration in AMI network was discussed.Briefing basics of theft detection scheme using three categories,state-estimation based,classification-based and game theory based.An architecture for theft detection in IoT data

streaming is presented in[7].[8] describes deep learning based system that utilizes Convolution Neural Network(CNN). Bad Data Detector(BDD) is employed to filter out low quality data.[9] proposes a two stage theft identifying system. At first, SVM is used to find out the theft. In further stage attack is confirmed using Temporal Failure Propogation Graph(TFPG). Long Short -Term Memory employed in [10] to detect intrusion and also for predicting consumption pattern based on previous data.Whereas a combine application of CNN and LSTM architecture is suggested in[11].In[12],theft detection using consumers usage pattern is presented and also addresses the data imbalance problem.The implementation of Support Vector Machine(SVM) for identifying attacks or detecting theft is proposed in[13]. The Principle Component Analysis(PCA) is utilised in reducing the complexity of the process.To find out honest and dishonest consumer,deep recurrent vector embedding is employed in [14].The Sequential grid search hyperparameter algorithm is used to enhance the performance of theft detection system.For the same purpose,application of Extreme Gradient Boosting (XGBoost) in AMI is presented in [15]. By exploiting the relation between Non-Technical Loss and missing value patterns, it is possible to diagonise thefts. This methodology is implemented in[16], with the aid of CNN model to locate missing pattern.[17] present an approach that detect theft using unsupervised learning method Firefly Algorithm based XGBoost. In this Visual Geometry Group(VGG-16) and normalization techniques were also applied to improve the performance of proposed system.Comparison of SVM, Logisstic Regresion(LR), and, CNN is also discussed. In [18], CatBoost Algorithm and SMOTETomek algorithm were employed in theft detection.The catBoost algorithm is a supervised method and utilizes feature engineering to choose appropriate feature for theft identification process.[19] proposes a hybrid CNN-RF(Random Forest)model for theft detection in power grids. All these methods are time consuming and require ,complex processing stages.This work proposes a methodology utilizing supervised learning algorithms to perform theft detection. Various models,such as,SVM,Random Forest ,Decision Tree, and Naïve Bayes, were analyzed.Unsupervised learning models, K-Means and DBSCAN, were also evaluated.

2.2 Conclusion

In this chapter we discussed about the basic concepts of false data injection type attacks in smart energy meter.Literature review regarding currently used methods is also presented.

Chapter 3

Problem Statement

3.1 Introduction

This chapter deals with motivation behind the work as well as the methodology developed for the problem identified.

Energy meters are widely used in various industries, such as residential, commercial, and industrial sectors, to calculate the energy consumption of a consumer. Readings from the energy meter are used to generate the electricity bills of the consumer depending on their tariff schemes.Recent advancements in Communication systems enable the electric utilities to modernize the metering system by replacing old analog meters with smart meters. The unique feature of the smart energy meter is the bi-directional communication between utilities and consumers. It also facilitates various demand-side management strategies. However, numerous data exchanges and the semi-open structure of the smart meter system make the metering process vulnerable to various physical as well as cyber-attacks. These attacks have adverse effects on the entire power system like reduction in revenue of electric utility, altering load forecasting, pricing strategy, and the stability of the system.The early detection of these type of intrusions in metering infrastructure is a huge burden for the power industry.Machine learning techniques can resolve these problems more effectively using the stored information.

3.2 Methodology

In machine learning, computers impulsively understand and make smart decisions based on a given set of information. Supervised Learning(Classification), Unsupervised Learning(Clustering),

Semi-supervised learning and, Reinforced learning comes under machine learning process. Classification is a type of supervised learning in which supervision is accomplished by a set of labelled data. There are mainly two steps in classification process. First step is to generate a classifier model using a set of labelled data called training set. In second step , an unlabelled data is given to the classifier for classification.In case of unsupervised learning data is not labelled .When a set of data input into a unsupervised model, it will create clusters or groups based on similarities.

In this work ,focus area is utilization of supervised learning techniques for finding electricity theft by false data injection. Unsupervised learning methods were also analysed. The Block Diagram and Flowchart of the proposed methodology is shown in Fig. 3.1 and Fig. 3.2. The historical information gathered from smart energy meter is passed into a pre-processing stage to remove the insignificant or noise data. Then it is given to each supervised models for the purpose of training.After training,each model performance is analyzed with a set of values for validation.In this work, The data's from a simulated attacker circuit were also utilized to validate the model.The new unlabelled data is classified into either fault or normal data.Whereas in case of unsupervised models,the given data is distributed among different classes,ie,the graphical visualization of the data is obtained.

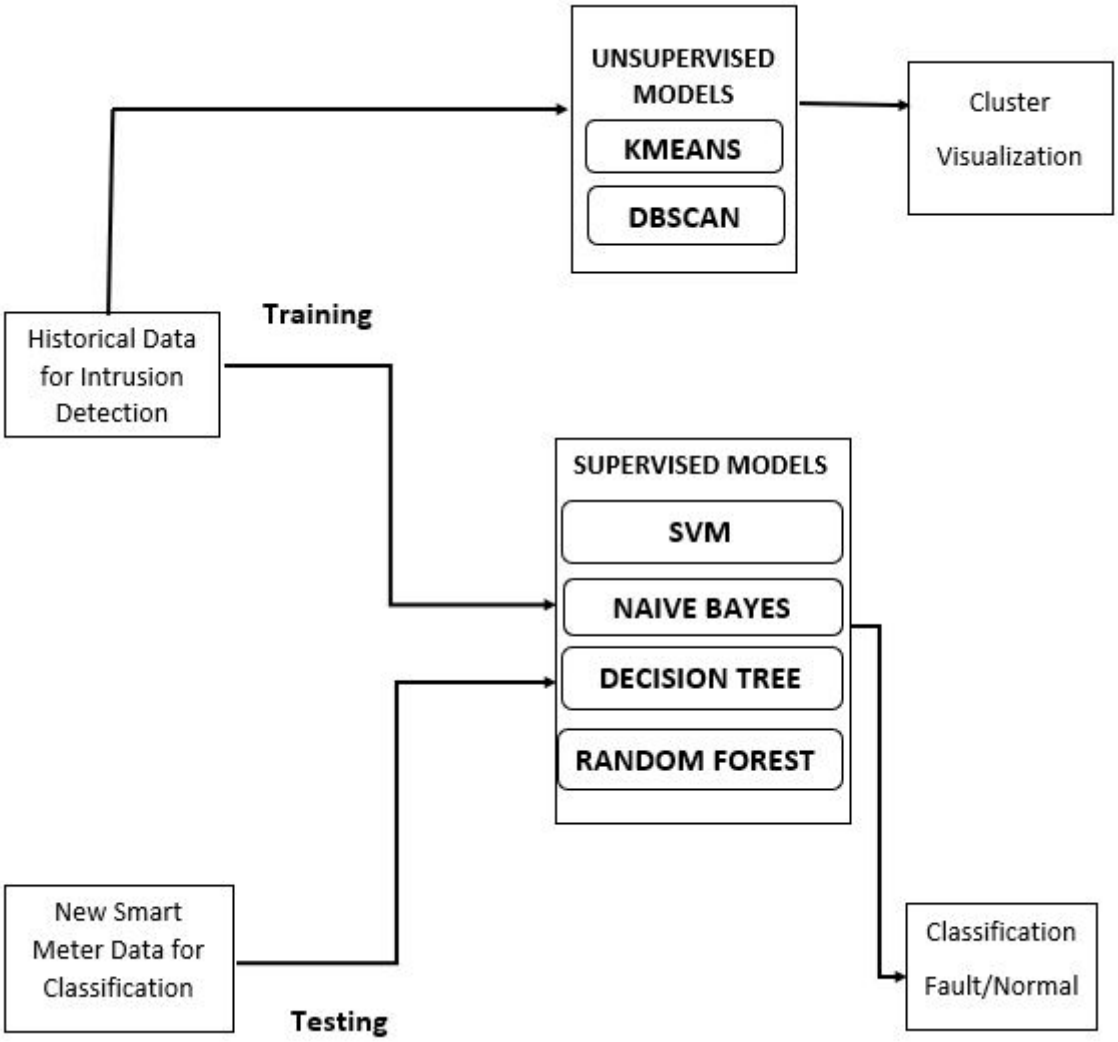


Figure 3.1: Blockdiagram of Proposed Methodology

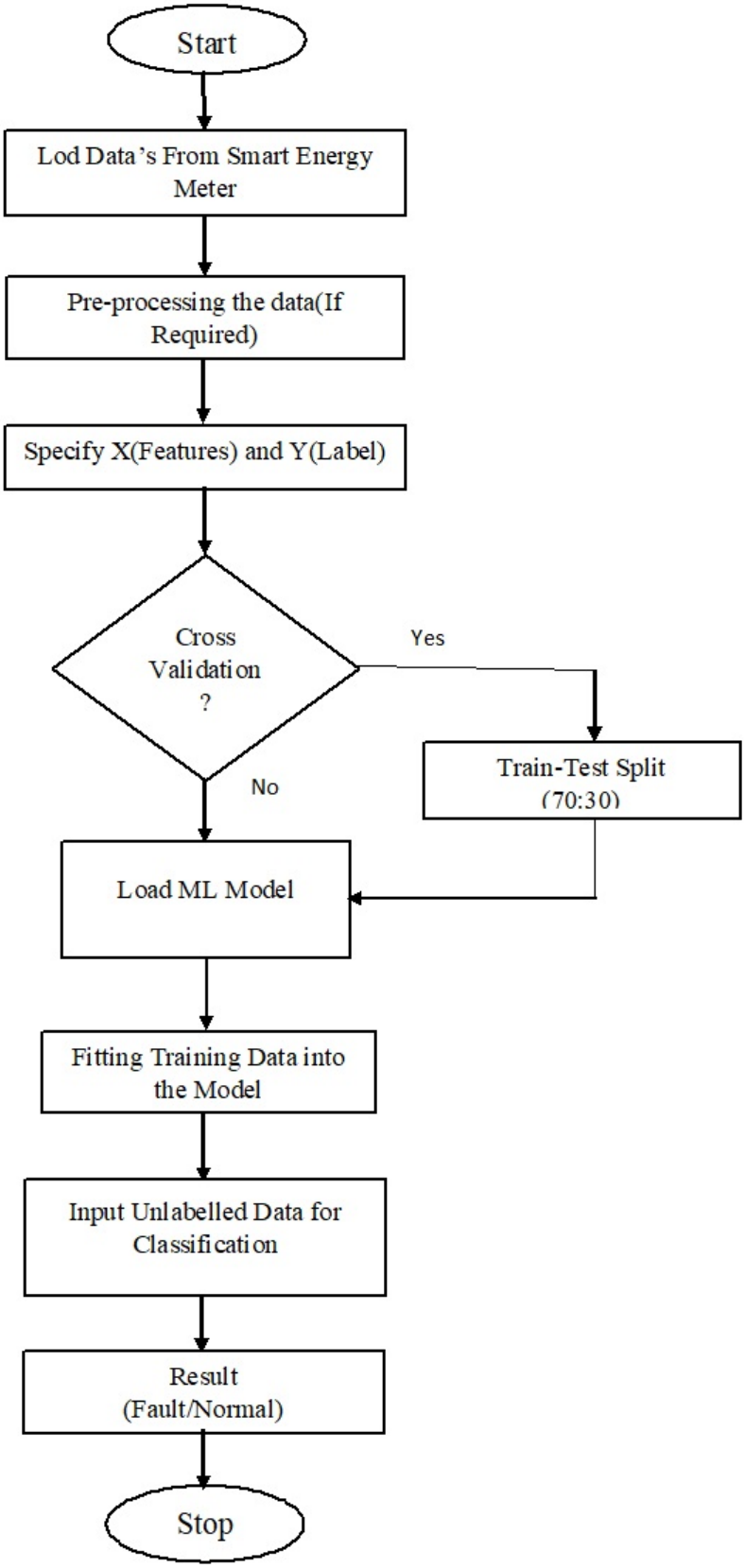


Figure 3.2: Flowchart of Proposed Methodology

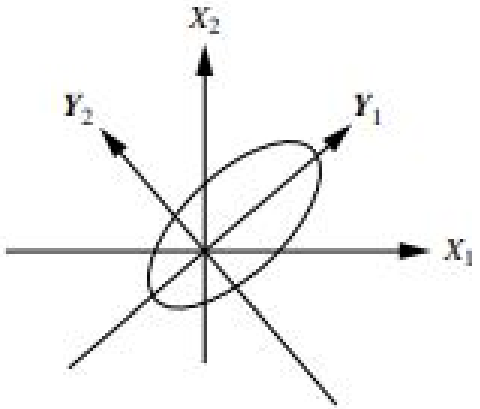


Figure 3.3: Principal Components

3.3 Principal Component Analysis(PCA)

Principal Component Analysis or Karhunen-Loeve or K-L method is a dimensionality reduction technique.Assume,the input data vector consist of 'n' feature is required to reduce to a data vector with 'k' feature such that $k \leq n$. Dimensionality reduction is achieved by projecting the actual data vector onto a smaller space.The PCA identifies the important aspects regarding the original data and create equivalent smaller set of features.The general procedure is given below,

- Normalize the input data.By this avoiding dominance of the large domain features over the small domain feature.So that every attributes comes in the same range.
- PCA calculates k orthonormal vectors,unit vectors that are perpendicularly projected,called Principal Components.
- These principal components are arranged in the descending order of strength.
- Data size is reduced by removing weaker component,ie,component with lower varience.

In Fig. 3.3, Y1 and Y2 are the principal components of actual data mapped on X1and X2 axes.

3.4 Cross Validation

In a model,overfitting or underfitting problem may arise as a result of using same data set for training and testing. In case of underfitting, it is difficult for the model to identify the differences in each class.It occurs when the model is too simple.Whereas in case of overfitting the model is

complex and can't generalize.When testing a model with same data used for training,the model perform satisfactorily.But ,if a set of unseen data is given to the model,the accuracy is poor and fails to generalize pattern and is called overfitting.Cross validation techniques are employed to evaluate the performance of a model by using a part of data for training and holding the remaining part,unseen data,for validation.Train-Test split method can be used if the data set is large enough and data's are equally distributed. In this technique ,the data is randomly selected for training and testing.Common splitting ratios are 70:30 and 80:20.The accuracy is same in both cases.Suppose the splitting ratio is 70:30,70 percentage of the data is used for training the model and 30percentage is kept as unseen data for evaluating the performance of model with new data.

3.5 Conclusion

This section briefly described about the need for a intrusion detection system and methodology implemented .The PCA is used for plotting decision boundaries . To overcome issues due to Overfitting,Tran-Test Split is used.

Chapter 4

Supervised Models

In this section, fundamentals of applied supervised learning model were discussed.

4.1 Support Vector Machine(SVM)

SVM is a classification model which involve transformation of actual data to a higher dimension using nonlinear mapping. With this new dimension, a decision boundary or hyperplane is created.The data's are separated into two classes by the hyperplane. SVM is generating the hyperplane using support vectors and margins. Training data on decision boundary is called support vectors. Hyperplane with large margin offer accurate classification of new data as compared with hyperplane having small margin. Therefore,SVM aims to find out Maximum Marginal Hyperplane(MMH), which is depicted in Fig. 4.1[20]. For linearly separable data,Decision Boundary or hyperplane can be defined as ,

$$w.x + b = 0 \tag{4.1}$$

Where w is weight vector given by$w = \{w_1, w_2...w_n\}$,n is the number of attribute,b is bias and x is training input,$X = \{x_1, x_2...x_n\}$.Consider linearly inseparable data with input 3D vector X=(x1,x2,x3) which is mapped into 6D place Z by$\phi_1(x) = x1, \phi_2(x) = x2, \phi_3(x) = x3, \phi_4(x) = (x1)2, \phi_5(x) = x1x2and, \phi_6(x) = x1x3$.

Decision boundary is given by,

$$d(z) = wZ + b \tag{4.2}$$

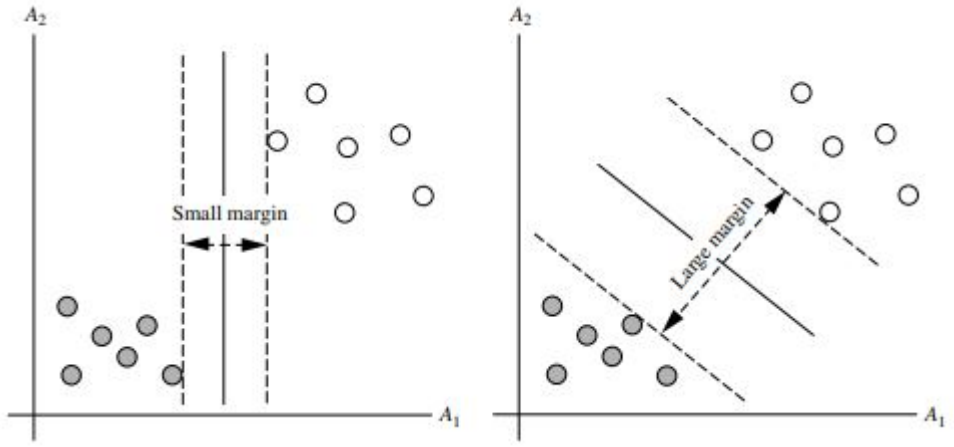


Figure 4.1: SVM Decision Boundary

Kernel function is equivalent to the dot product calculation which solves the computational complexity. This eliminate the necessity of mapping.

$$K\left(X_i, X_j\right) = Xi.(Xj) \tag{4.3}$$

Three popular Kernel's are,

- Polynomial Kernel with degree ,h

$$K\left(X_i, X_j\right) = \left(X_i.X_j + 1\right)^h \tag{4.4}$$

- Gaussian Radial Basis Function (RBF)Kernel,

$$K\left(X_i, X_j\right) = e^{||Xi-Xj||2}/2\sigma 2 \tag{4.5}$$

- Sigmoid Kernel

$$K\left(X_i, X_j\right) = \tanh\left(kX_i.X_j - \delta\right) \tag{4.6}$$

Various kernels give different SVM classifiers. There are no specific criteria for selecting appropriate kernel.Accuracy of SVM using different kernel comes in the similar range.

4.2 Naive Bayesian Classifier

These are statistical classifiers and arrange given data into different classes based on class membership probability.Naive Bayesian classifier is a simple Bayesian classifier.It categorize data on the assumption that impact of a feature score on a particular class/group is distinct,ie,not

depending on the other feature value and is called Class Conditional Independence.Let X be a data and H is the hypothesis for X being a member of a particular class,then Bayes theorem is given by,

$$p\left(H|\ X\right) = \frac{p\left(X|\ H\right)p(H)}{p(X)} \tag{4.7}$$

Where $p\left(H|\ X\right)$ is termed as posterior probability of H and condition is X, $p\left(X|\ H\right)$ is the posterior probability of X and condition is H,p(H) and p(X) are prior probability of H and X respectively.Given D is a training dataset consist of n number of data,$X = \{x_1, x_2 ... x_n\}$ characterized by n attributes A1,A2,...An.Assume the number of class is m,c1,c2,...cm.The data X is the member of a class which has largest volume for posterior probability,that is,X is a member of class ci,if it satisfies the following condition,

$$p(ci|X) > P(Cj|X) for 1 \leq\ j \leq\ m,\ j \neq\ i \tag{4.8}$$

Maximum posteriori hypothesis is given by ,

$$p\left(c_i|\ X\right) = \frac{p(X|c_i)}{p(X)} \tag{4.9}$$

4.3 Decision Tree Classifier

A decision tree is a tree like structure generated from a set of labelled data.Tree begins from the topmost node called root node.Internal node represent a test on feature ,result is indicated by branch and leaf node has class label. When a new data is given to the model ,feature values are tested and a class is predicted by going through root node to terminal node. Decision trees are capable of dealing with multidimensional data.The best feature for proper classification can be selected using attribution selection measures namely Information Gain Ratio(IGR), Gini Index(GI)[21]. The tree begins from a single node,N, that represent training data D. If all the data's belongs to same class,then the node becomes terminal node.Otherwise splitting criterion is determined by calling attribution selection. In this way identifies the feature on which the test is going to perform. Pure separation is obtained if all the data's are of same class. This process is repeated until no data is coming from partitioning. Following are the stopping conditions,

1. All the data comes into the same class.

2. No more attributes are remaining for partitioning.

3. No data is available from partitioning.

Decision path is traced from these steps. For theft detection Gini Index is used as attribution selection measure.

$$Gini\,(D) = 1 - \sum_{i=1}^{m} P_i^2 \tag{4.10}$$

where $P_i = \frac{|C_{i,D}|}{|D|}$ is the probability of p being a member of class C_i and m is the number of classes.

4.4 Random Forest

The performance of ensemble models are much better than base classifier models. There are various ensemble methods such as Boosting,Bagging and Random Forest. The given data set D is splitted into$D_1, D_2, \ldots, D_k$ to generate $M_1, M_2, \ldots, M_k$ base classifier model.A composite classifier model is created from these k series of base models.For a given unknown data,each model makes predictions and is called as voting.The ensemble classify the new data by analysing the votes from the base models.In random forest, the base classifier model is a decision tree classifier and cluster of these decision tree is termed as 'Forest'.Each model generate decision trees from the randomly selected features.Whenever a new input is given,classification task is performed based on the voting from base model.General procedure is as follows,

- From each iteration, i=1,2,3,. . . ,k,training set Di consist of d data's is created.
- Some of the data's repeatedly appear in training set whereas some other data's may be eliminated.
- Assume,at each node,there are F number of features which will be smaller than the total number of features.
- By randomly selecting attributes,decision tree classifier model,Mi,is generated.

4.5 Results

The performance evaluation of a classifier model is an important aspect.The matrix evaluation method or confusion matrix were most widely used to compute the accuracy of classification.Suppose number of class is m,such that ,$m \geq 2$,then confusion matrix will be a $m \times m$

Figure 4.2: Confusion Matrix of SVM

matrix.In ideal case,all the tuples are distributed in diagonal elements of the matrix.If the classifier predicts all the positive case as positive ,then it is a True Positive (TP) case.Otherwise if the classifier predict negative value as positive,then it is False Positive(FP).The classification of actual negative case as negative implies True Negative(TN).If not,ie, the model groups positive data as negative,it is False Negative(FN).In confusion matrix ,number of predictions belongs to each case(TP,TN,FP,FN) is given by the matrix elements.Accuracy can be computed using the equation,

$$Accuracy = \frac{TP + TN}{P + N} \tag{4.11}$$

FP and FN are errors in prediction since it refers to the number of miss classification.The evaluation of each models with confusion matrix is presented.For training the model 10,000 data's were utilized.By employing Train-Test Split method,7000 data's used for training and 3000 is used for validation. The figure 4.2 gives the confusion matrix of SVM with RBF kernel. Here $C_{11} = 1492$,which is the number of true fault(TN) class members.$C_{12} = 1$ represent False Normal(FP), $C_{21} = 354$ refers to the False Fault (FN),and $C_{22} = 1153$,is the True Normal (TP) values.In the false identification scenario,TP is the number of actual normal data,TN refers to the number of faulty values.Whereas FN and FP indicates the number of miss-classification of normal data as fault and fault value as normal value respectively.In SVM classifier error from

FP is higher than FN and testing accuracy is 88.167 percentage.Decision boundary generated using different kernels, Linear, RBF, and Sigmoid is depicted in Fig. 4.3,Fig 4.4 and Fig. 4.5 respectively.To plot these boundaries, PCA is performed. As a result number of feature is reduced to two equivalent feature from the original five features.The SVM model with RBF kernel gives comparatively accurate classification.In Naïve Bayes Classifier Model, using three types, Gaussian, Multinominal and Bernoulli models.For Gaussian naïve bayes classifier in Fig .4.6, True Normal(TP)=790 and True Fault(TN)=1407 and both errors are present in the model(Fales Normal(FP)=56, False Fault(FN)=747). Testing Accuracy of the model is 73.2 percentage.

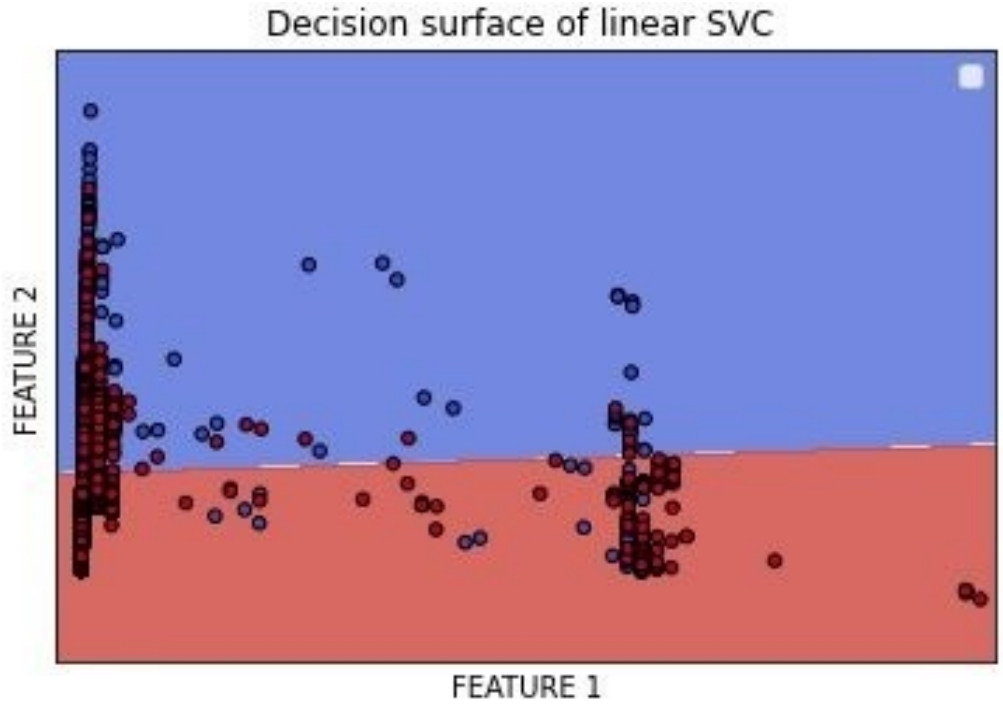


Figure 4.3: Decision Boundary of SVM using Linear kernel

For Multinominal NB in Fig. 4.7, TP or True Normal=1446 and True Fault or TN=1370.Error due to wrong classification of normal value as faulty value is zero,ie, False Fault(FN)=0,while type of error due to False positive or False Normal is present,number of such classification is equal to 184. An accuracy of 93.86 percentage is obtained for this model. In case of Bernoulli NB in Fig. 4.8, Accuracy of classification is lesser compared to other models. Both errors incurred in this model, FP=265 and FN=955.Testing accuracy is only 59.33 percentage.True predictions TF and TP are 1224 and 556 respectively.

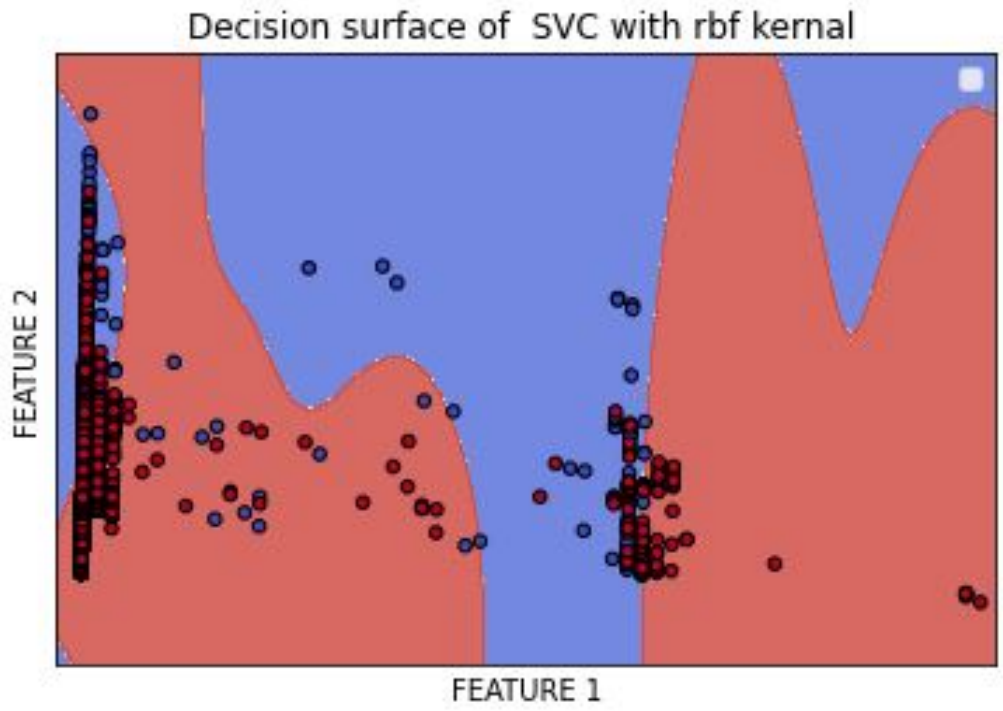


Figure 4.4: Decision Boundary of SVM using rbf kernel

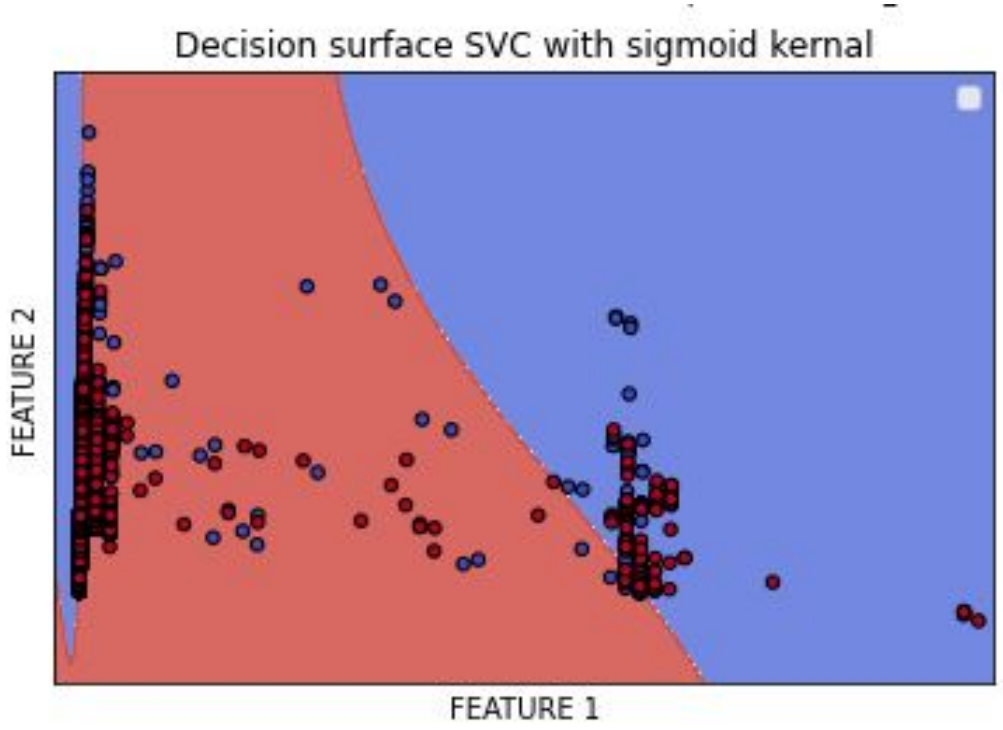


Figure 4.5: Decision Boundary of SVM using Sigmoid kernel

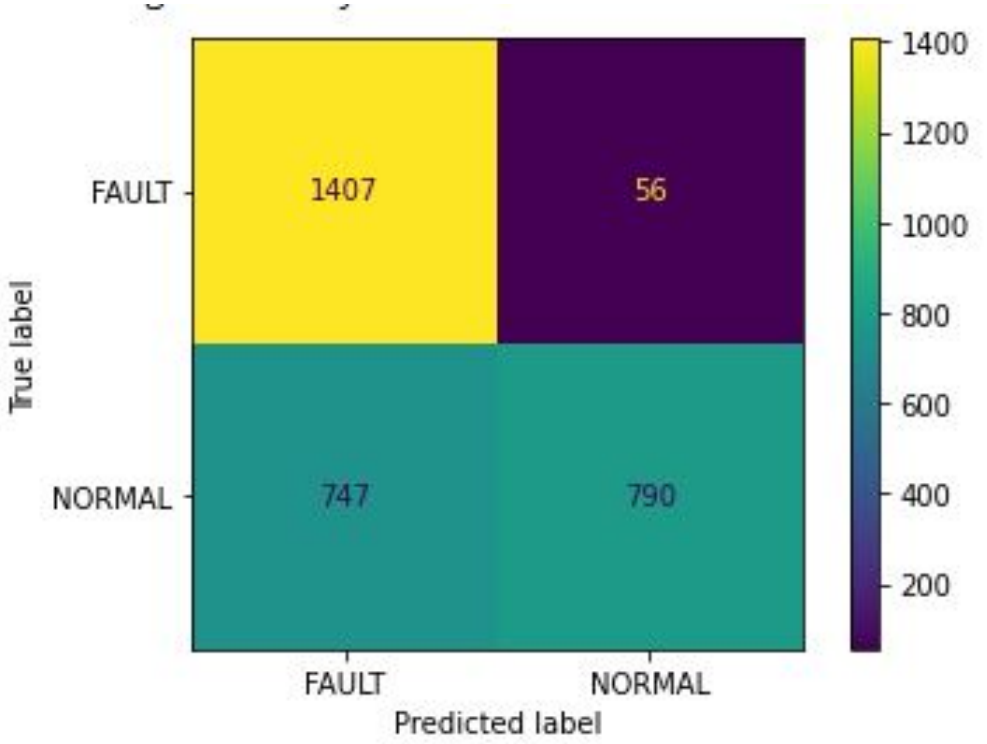


Figure 4.6: Confusion Matrix of Gaussian NB

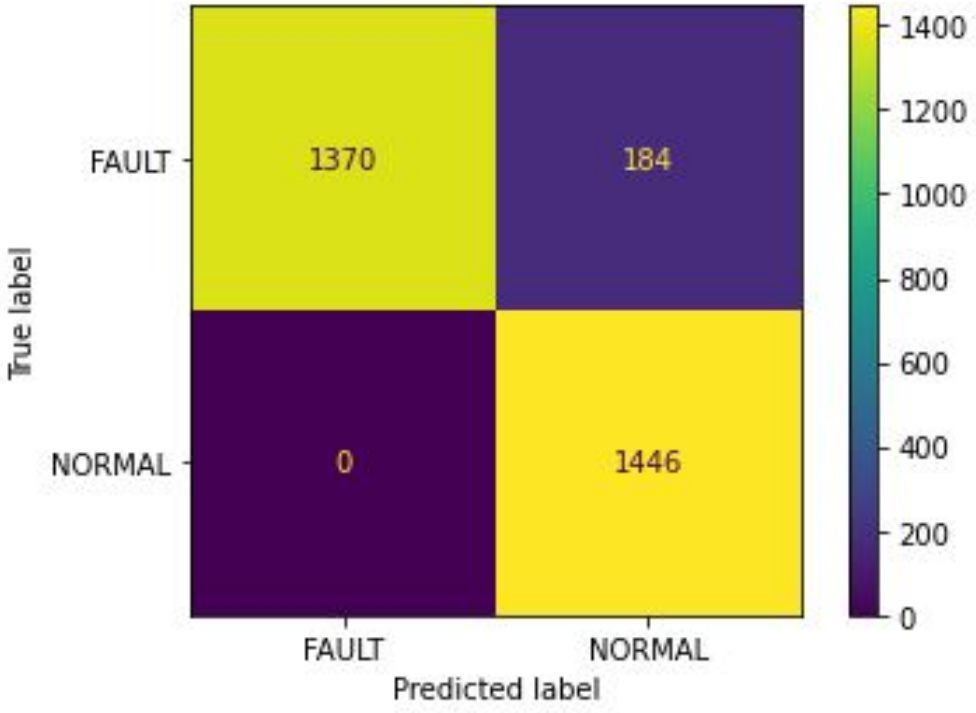


Figure 4.7: Confusion Matrix of multinominal NB

Decision tree classifier is a best model for the fault or theft detection. The confusion matrix is given in Fig. 4.9. It gives an accuracy of 99.76 percentage.Errors due to miss clas-

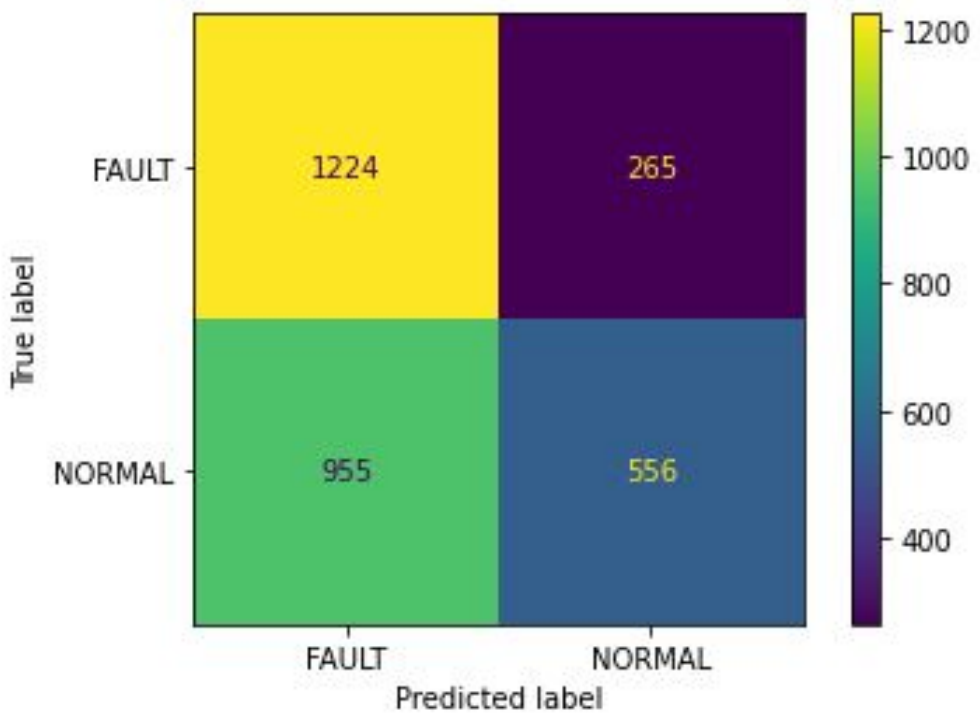


Figure 4.8: Confusion Matrix of Bernoulli NB

sification is much lower when comparing with other models(FP=3 and FN=4). True fault and True Normal values are 1519 and 1474 respectively. For Random Forest Classifier in Fig 4.10, number of correct fault predictions is equal to 1363 and normal case classification is 961. Miss classification of fault and normal cases are 459 and 157 respectively. Classification accuracy of the model is 78.13 percentage.

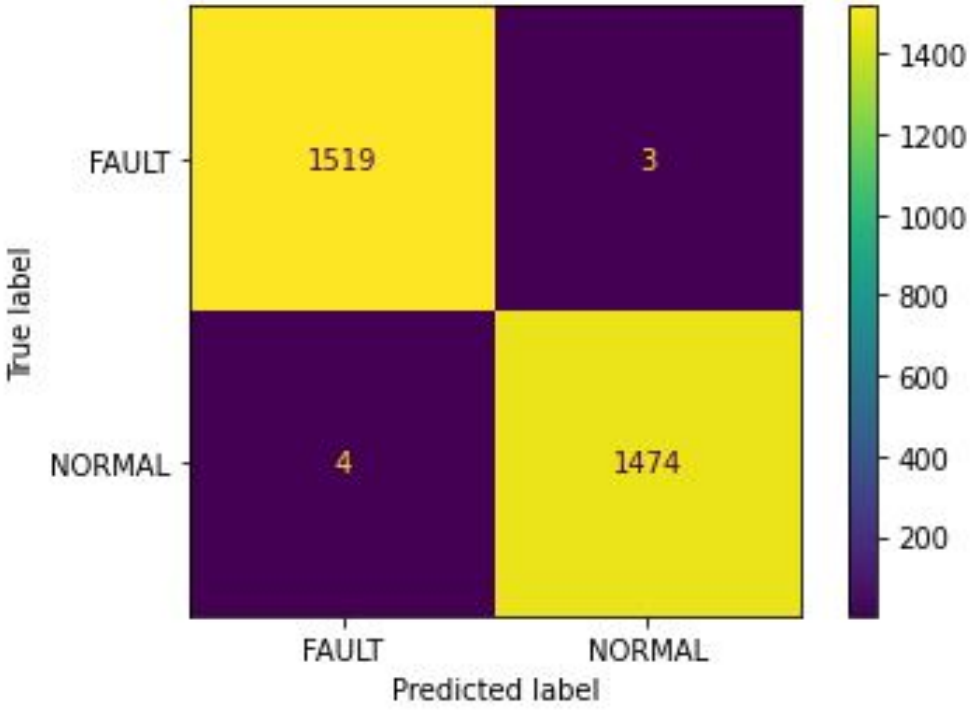


Figure 4.9: Confusion Matrix of Decision Tree

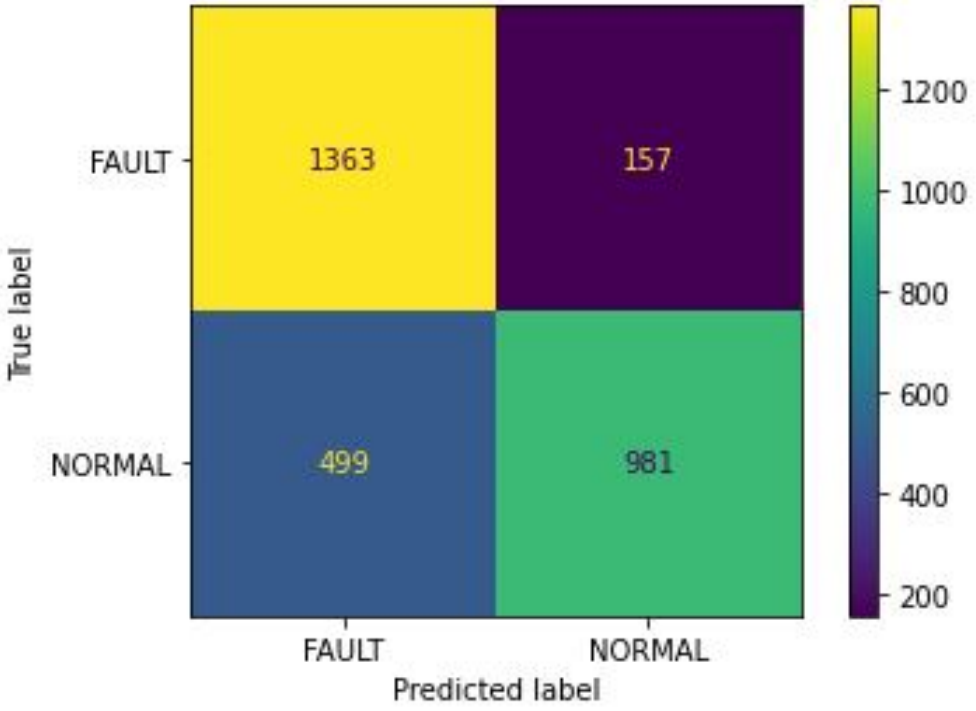


Figure 4.10: Confusion Matrix of Random Forest

4.6 Conclusion

Fault identification using supervised learning is presented in this section.In supervised learning, SVM, Naive Bayes, Decision Tree, and Random Forest models were developed. Analysed the classification task performed by each model using confusion matrix.

Chapter 5

Unsupervised Models

5.1 Introduction

The technique of grouping a set of data points (or observations) into subsets is known as cluster analysis or simply clustering. All subsets are cluster, with member in one cluster resembling those in another but differing from those in other clusters. A clustering is the collection of clusters that results from a cluster analysis. The cluster center, centroid Ci, is used to point out that cluster in a centroid-based partitioning technique.

5.2 K-Means

In k-Means algorithm,the mean value of the points in a cluster gives the centroid of that cluster. It begins by selecting k of the items in D (Data set) at random, at first,each of the selected item indicate a cluster mean or centroid. Depending on the object's Euclidean distance from the cluster mean, each of the remaining elements in the data set is allocated to that cluster to which it is more relatable. The within-cluster variation is then improved iteratively using the k-means algorithm. It calculates the new mean for each cluster by utilizing points set to the cluster in the preceding iteration. The updated new means are then used to reassign all of the elements as a new cluster centres. The iterative process is continued until the allocation is permanent, which means that the clusters developed in the last iteration matched with those created in the foregoing iteration(Fig. 5.1).

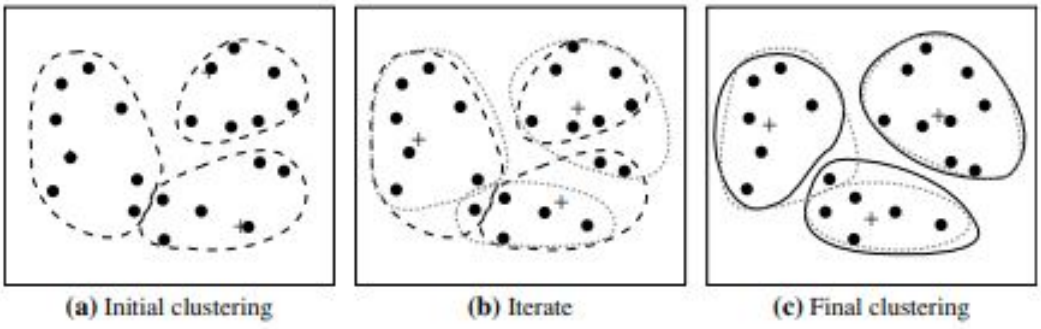


Figure 5.1: Steps in K-Means Clustering

5.3 DBSCAN

The number of elements close to an element o can be used to determine its density.DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an approach for locating core items, or objects with dense surroundings. The radius of a neighbourhood we evaluate for each object is set by a user-provided parameter $\epsilon > 0$.The region confined in a radius , centred at the point o is called its -neighborhood. It is possible to easily assess the solidity of a neighbourhood by the number of objects in the neighbourhood due to the fixed neighbourhood size parameterized by ϵ. At first, each elements present in the available data set, D, are tagged as unvisited. DBSCAN randomly chooses an unvisited element p , name it as visited,then verifies if the neighborhood of p holds at least MinPts elements, the density threshold for dense regions.If this is not the case, then marking p as a disturbance point. If not, a new group or cluster C is formed for p, and remaining clusters are removed.A candidate set, N, is created from the objects in p's ϵ -neighborhood. DBSCAN put on elements in N which can't be add to any one of the cluster C, iteratively. DBSCAN labels a point p^* in the group N, that has the tag unvisited ,as visited, and also verifies their neighborhood in this process. If the p^* 's neighborhood contains at the minimum of MinPts elements,then that items are also included to N. The DBSCAN keeps entering points to Cup to the condition that it can no longer expand it, at which point N becomes empty. Cluster C is now complete, and output is now available. DBSCAN chooses an unvisited item from the remaining ones at random to identify the next cluster. The clustering procedure is repeated until all items have been visited(Fig. 5.2).

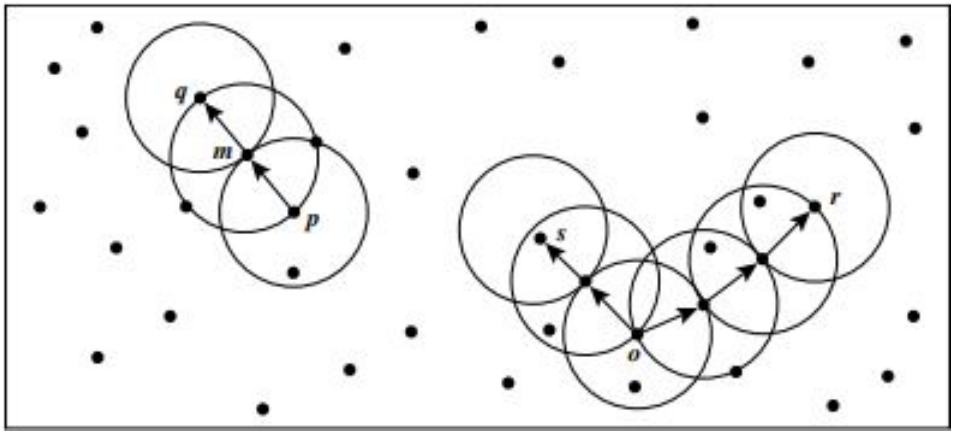


Figure 5.2: DBSCAN Cluster Formation

5.4 Results

This chapter comprises of results obtained from various unsupervised models(K-Means, DBSCAN). Model formulation and evaluation is carried out using Python programming.These models created clusters based on similarities of the input data's.The clusters formed in both clustering techniques is shown below.In Fig. 5.3, K-Means classifier creates two clusters from the given data,red colour indicates Fault value while blue indicates normal data.Whereas in Fig. 5.4, DBSCAN developed three clusters which will leads to miss-classification.The original data has two class only, DBSCAN distribute data in three groups, implies clustering using DBSCAN is not suitable for this scenario.In clustering methods,if a new data is given to the model it is very difficult to separate the new data from the visual representation,ie,labelling of unseen data is not possible.This can be used only for identifying possibility of theft.In case of supervised learning technique,labelling of a new unseen data is possible and benefits for future training of the model.

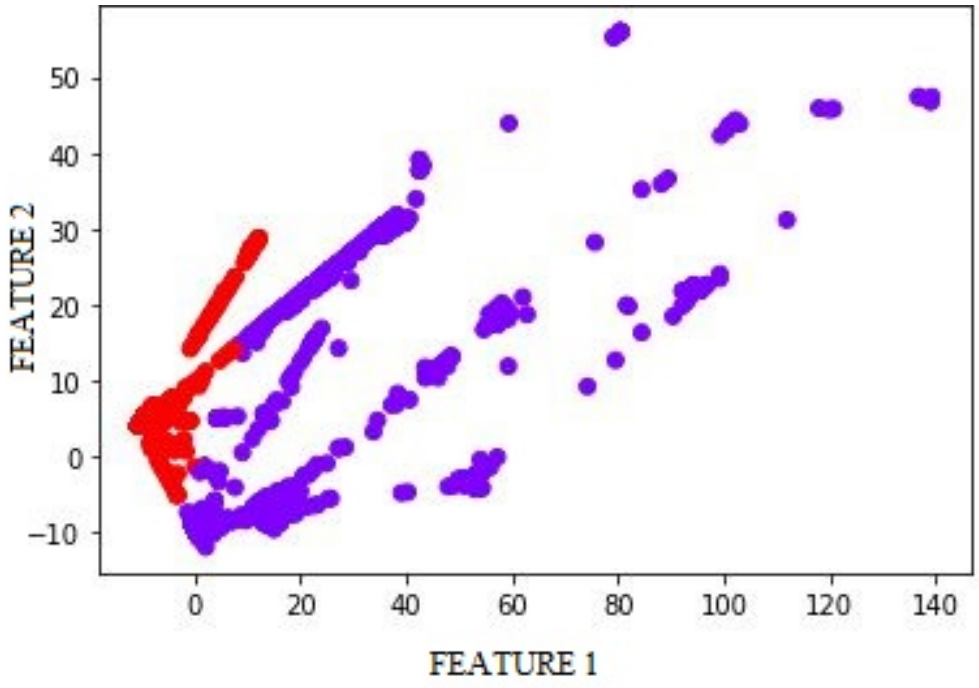


Figure 5.3: Clusters in K-Means

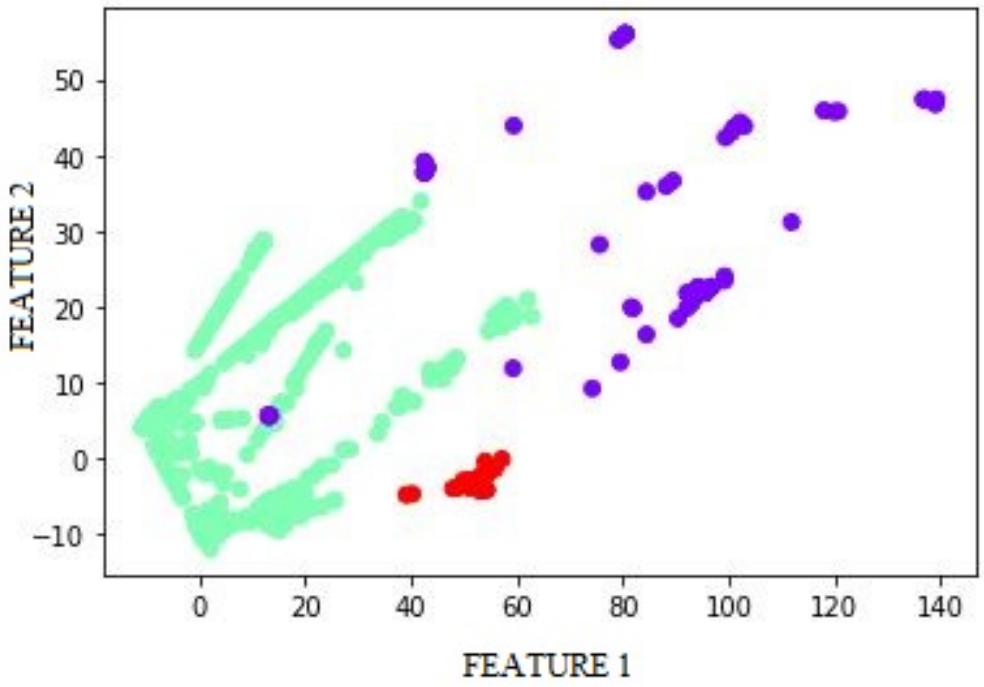


Figure 5.4: Clusters in DBSCAN

5.5 Conclusion

In this chapter, discussed about various unsupervised techniques(K-Means and DBSCAN) employed for visualization of given data.Based on the clusters created, abnormalities present in the input data can be identified.The output obtained from each model is also presented.

Chapter 6

Results And Analysis

6.1 Introduction

This chapter comprises of comparison of various models based on classification accuracy. Model formulation and evaluation is carried out using Python programming.The data's from an attacker circuit is used to validate the model.

6.2 Data Collection

Energy consumption data of domestic consumers for both normal and faulty conditions were collected from smart meters. A part of these datasets is given to the classifier model for training and the remaining is used for the model for performance evaluation.The consumption details of 20000 consumers collected from dataset of State Grid Corporation of China, were used. The dataset consists of time series data during 2016 -2017.Details include global active power which is the sum of power consumption other than sub metering. Sub metering is used to find out consumption of warehouse like areas. There were three sub metering in the home area.The total consumption is calculated with these data.In case of faulty consumers,there is a mismatch in the aggregate value and the sum of consumption from each area.whereas in normal case, the aggregate value is equal to the sum of individual areas.Data's of 10000 consumers is used for training the model.Among these 5000 is honest consumer and remaining 5000 is faulty consumers.the remaining 10000 data is used for validating the model. In Fig. 6.1 consumption pattern when no theft is encountering is depicted. Whereas Fig. 6.2 shows the consumption pattern during theft.The theft is a fault data injection type attack.In which faulty consumption

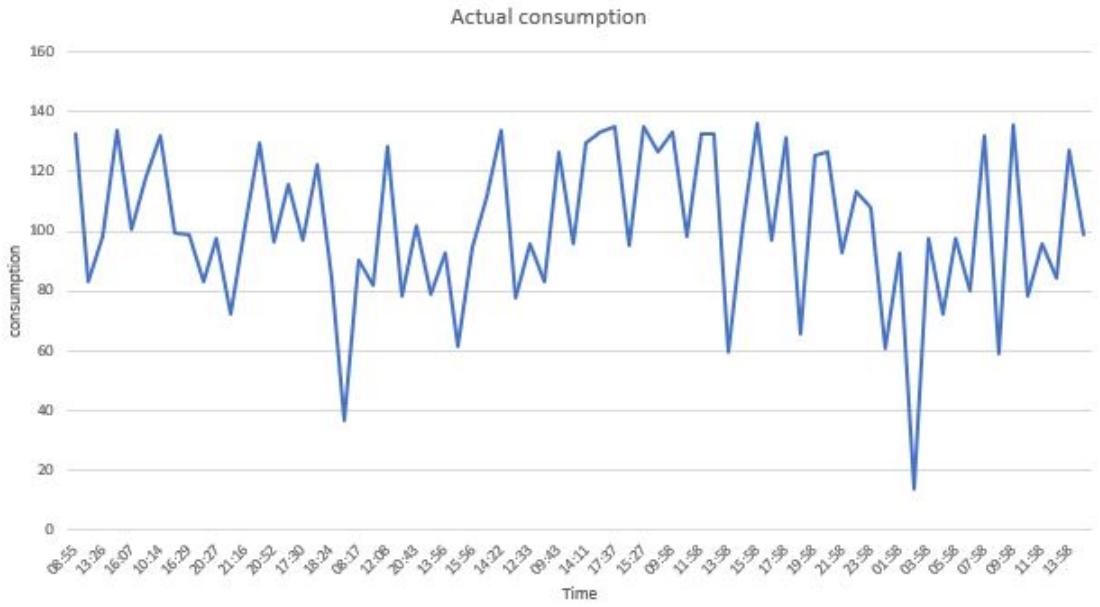


Figure 6.1: Actual Consumption

value is entered into the AMI system.Thereby consumption is made far away from the actual value and greedy attacker earns certain profit by reduction in the electricity bill.

In this case, trained classifiers using all 10,000 data's. Validation is done using another set of 10,000 values,which will be unknown data for the model.Overfitting problem is detected in this stage,ie,training accuracy is higher while validation accuracy is reduced considerably.Accuracy comparison of various model in this situation is depicted in Fig. 6.3. To overcome these overfitting problem, utilized Train-Test Split method.The total training data is divided in the ratio 70:30.,ie,70 percentage is used for training and 30 percentage is used for testing .By this both training and testing accuracy comes in the same range.Accuracy of different model after cross validation is shown in Fig. 6.4.

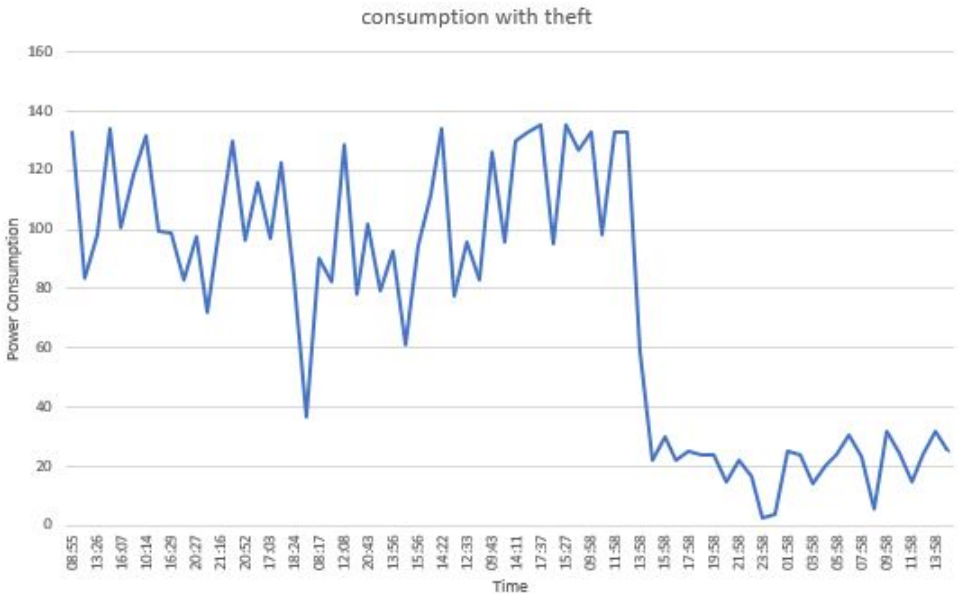
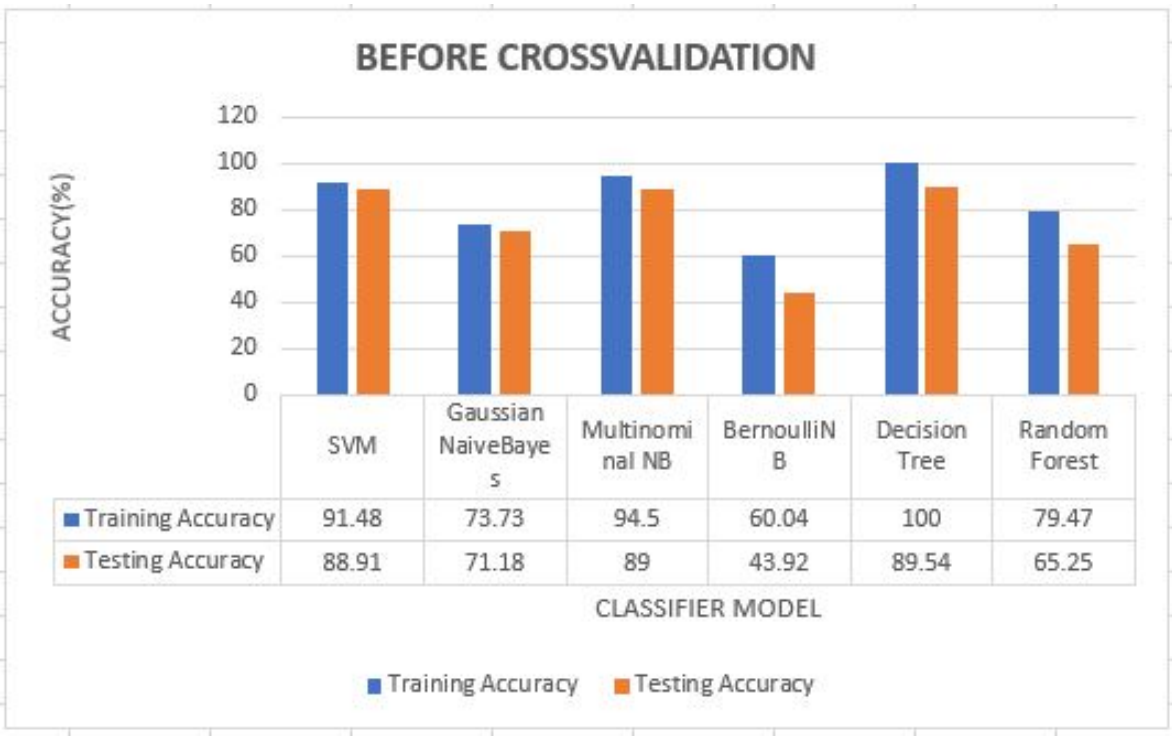


Figure 6.2: With Theft



	SVM	Gaussian NaiveBayes	Multinomial NB	BernoulliNB	Decision Tree	Random Forest
Training Accuracy	91.48	73.73	94.5	60.04	100	79.47
Testing Accuracy	88.91	71.18	89	43.92	89.54	65.25

Figure 6.3: Accuracy Comparison of models before cross validation

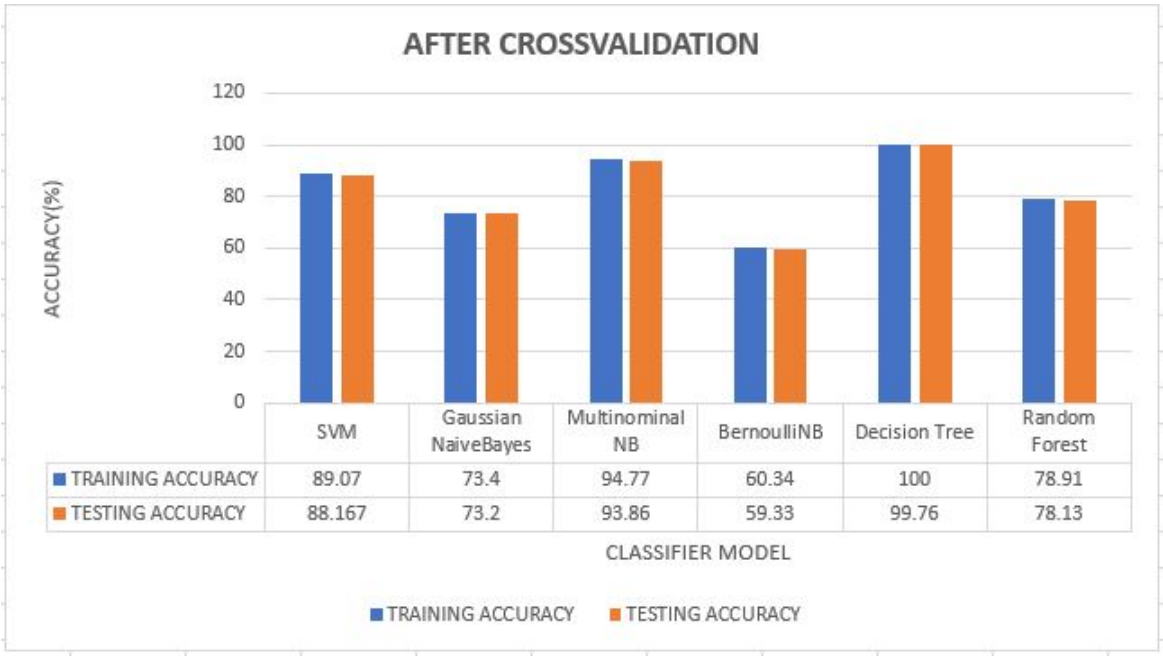


Figure 6.4: Accuracy Comparison of models After cross validation

6.3 Validation of Model

For verifying the effectiveness of the model,data's collected from a simulated false data attack system.In this attack , the intruder insert an IGBT switching circuit between the measuring meter and load. The Gate signal is transmitted to the IGBT gate to turn off the circuit while the meter is measuring the current reading.In this manner, the measured power consumption is made different from the actual consumption.

For controlling the switching action, an MCU is used, from which the controlling digital signal is sent to the IGBT. At the time of sampling the current reading, the MCU will send a gate-OFF signal and make the circuit open-circuited. During this period measured current value is zero. Hence power consumption at this period is also zero. The MCU will send a gate-ON signal in a non-sampling period and energy consumed by the load during this time only is considered for estimating the energy cost. In this manner attacker manipulate the actual value and introduce computational error in the metering mechanism. In simulation, signal generator is used for sending the gate signals. The performance of various classifiers with cross validation also analysed with a set of new unknown data obtained from the simulated attacking system. The plot of data from such an attacking system is shown in Fig. 6.5. During last period of time,the actual consumption value is reduced to lower value. Hence it is a false data injection type attack. The accuracy of various models in this fault identification scenario is depicted in

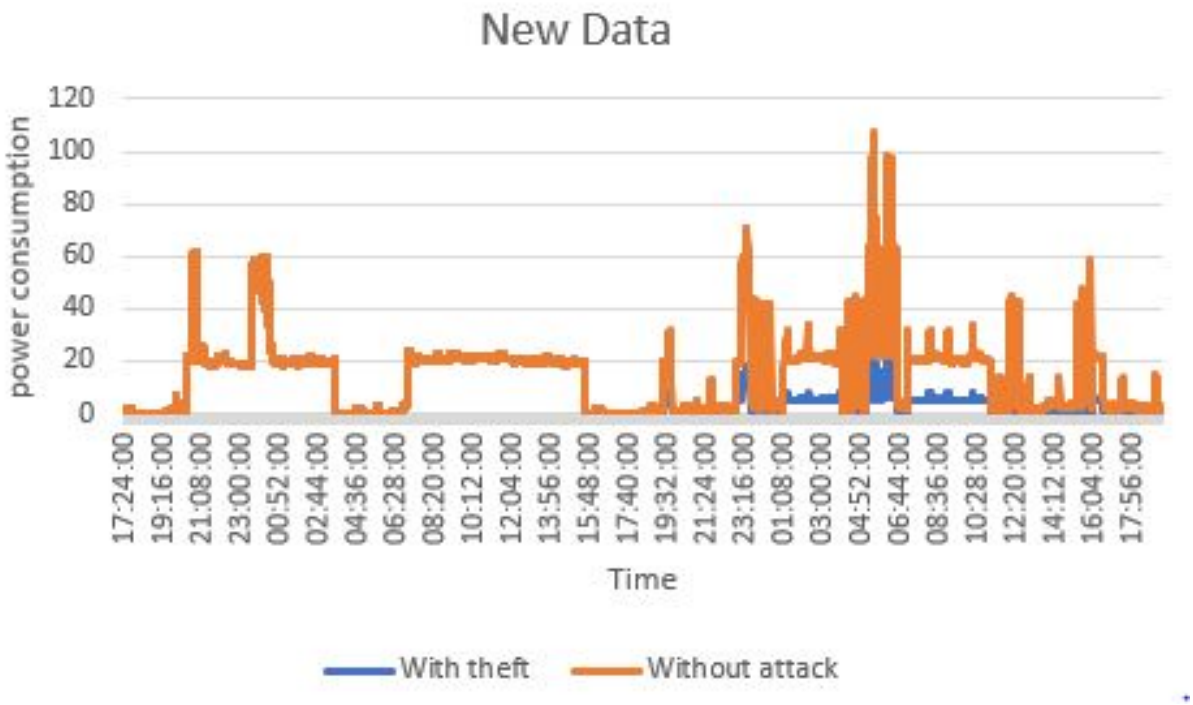


Figure 6.5: Data Representation

Fig. 6.6. In this case also ,models were performed the classification with cross-validation.The validation accuracy of decision tree model is 90.7 percentage and that of Bernoulli is only 49.46 percentage.In all the cases Decision Tree classifier has higher accuracy.Multinominal Navie Bayes classifier is better than SVM classifier with an accuracy of 94.77 percentage.Among all these classifier Bernoulli NB offer lowest accuracy(59.33 percentage).

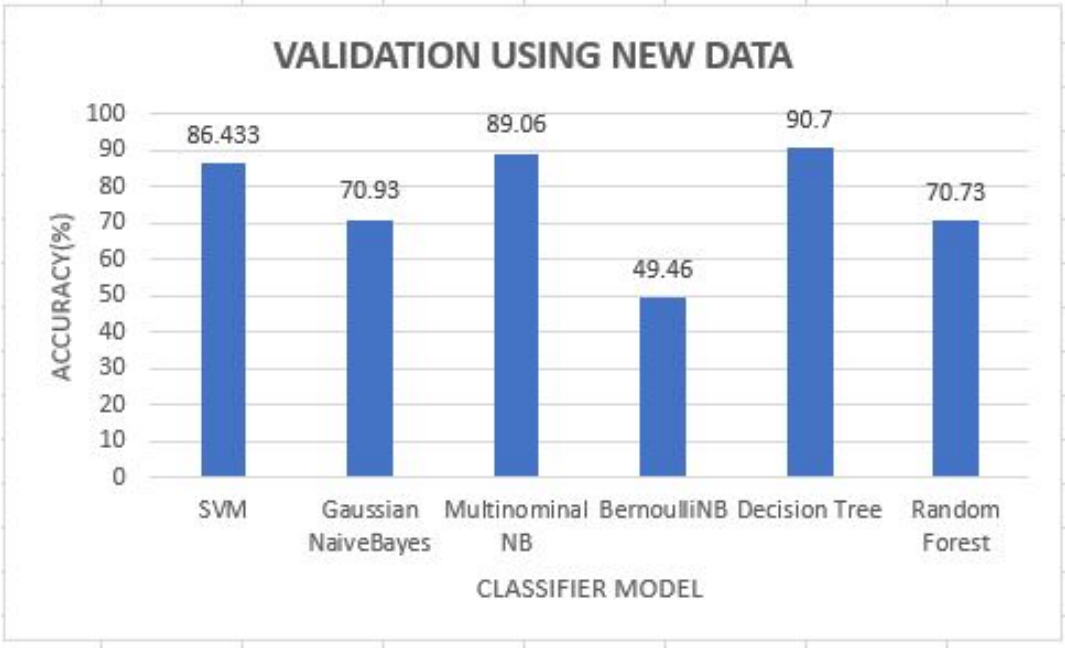


Figure 6.6: Accuracy Comparison of models with unknown dataset

6.4 Conclusion

In this chapter,comparison of classification of data's using various supervised learning models were presented.When a new data set is given to the trained models,Decision Tree model classified the data's form smart energy meter in an efficient manner.The graphical analysis of the result is also included in this chapter.

Chapter 7

Conclusion And Future Scope

One of the major challenge faced by the energy sector is development of an efficient energy theft detection system. Now a days more research is going on this area. Currently available methods are becoming inapt in the era of advanced technologies. With the aid of changes in the AMI structure and development of adequate tools intruders can easily attack smart metering system. The emerging Artificial Intelligent techniques provides effective solution to this problem. Theft detection strategy based on Machine Learning techniques is proposed in this work and also evaluated the efficiency of these models. From unsupervised learning, K-Means and DBSCAN is used.Performance of K-Means clustering is comparatively higher than DBSCAN. But using unsupervised models, unable to identify the state of the new unseen data through visual realization. With the help of supervised models, labelling of a new unseen data into fault or normal is possible. SVM, Gaussian Naïve Bayes, Multinominal Naïve Bayes, Bernoulli Naïve Bayes, Random Forest, and Decision Tree are the classifiers used in this work.Among these classifiers Decision tree classifier achieved higher validation accuracy of 99.67 percentage with crossvalidation and Bernoulli Naïve Bayes has the least accuracy (59.33 percentage). Train-Test split method is used to overcome the overfitting problem. This work deals with only false data injection type attack. In which intruder perform by manipulating actual consumption data y various means. In future we can improve the model by identifying various type of attack . For this ,more data under different type of theft , should be collected from the real smart metering system both in normal and theft cases. However, the effectiveness of the use of supervised learning techniques in theft detection system is verified with available data from simulated attacking system.

References

[1] S. Sahoo, D. Nikovski, T. Muso and K. Tsuru, "Electricity theft detection using smart meter data," *2015 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2015, pp. 1-5, doi: 10.1109/ISGT.2015.7131776.

[2] H. Zubi and A. Alrmaih, "Smart Energy Meter System Design Simulation Presenting Electricity Theft Methods, Detection and Protection," *2021 IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA*, 2021, pp. 533-538, doi: 10.1109/MI-STA52233.202 -1.9464447.

[3] C. Richardson, N. Race and P. Smith, "A privacy preserving approach to energy theft detection in smart grids," *2016 IEEE International Smart Cities Conference (ISC2)*, 2016, pp. 1-4, doi: 10.1109/ISC2.2016.7580882.

[4] Assia Maamar and Khelifa Benahmed. 2018. "Machine learning Techniques for Energy Theft Detection in AMI". *2018 International Conference on Software Engineering and Information Management (ICSIM2018). Association for Computing Machinery*, New York, NY, USA, 57–62. https://doi.org/10.1145/3178461.3178484

[5] W. Li, T. Logenthiran, V. Phan and W. L. Woo, "A Novel Smart Energy Theft System (SETS) for IoT-Based Smart Home,"*in IEEE Internet of Things Journal* , vol. 6, no. 3, pp. 5531-5539, June 2019, doi: 10.1109/JIOT.2019.2903281.

[6] A. Yahyaoui, H. Lakhdhar, T. Abdellatif and R. Attia, "Machine learning based network intrusion detection for data streaming IoT applications," *2021 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*, 2021, pp. 51-56, doi: 10.1109/SNPDWinter52325.2021.00019.

[7] Jiang, Rong Lu, Rongxing Wang, Ye Luo, Jun Shen, Changxiang Shen, Xuemin. (2014). Energy-Theft Detection Issues for Advanced Metering Infrastructure in Smart Grid. *Tsinghua Science Technology*. 19. 105-120. 10.1109/TST.2014.6787363.

[8] S. Wang, S. Bi and Y. -J. A. Zhang, "Locational Detection of the False Data Injection Attack in a Smart Grid: A Multilabel Classification Approach," *in IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8218-8227, Sept. 2020, doi: 10.1109/JIOT.2020.2983911.

[9] C. -C. Sun, D. J. Sebastian Cardenas, A. Hahn and C. -C. Liu, "Intrusion Detection for Cybersecurity of Smart Meters," *in IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 612-622, Jan. 2021, doi: 10.1109/TSG.2020.3010230.

[10] X. Wang, T. Zhao, H. Liu and R. He, "Power Consumption Predicting and Anomaly Detection Based on Long Short-Term Memory Neural Network," *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 2019, pp. 487-491, doi: 10.1109/ICCCBDA.2019.8725704.

[11] R. U. Madhure, R. Raman and S. K. Singh, "CNN-LSTM based Electricity Theft Detector in Advanced Metering Infrastructure," *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225572.

[12] D. Syed, H. Abu-Rub, S. S. Refaat and L. Xie, "Detection of Energy Theft in Smart Grids using Electricity Consumption Patterns," *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 4059-4064, doi: 10.1109/BigData50022.2020.9378190.

[13] R. N. Toma, M. N. Hasan, A. -A. Nahid and B. Li, "Electricity Theft Detection to Reduce Non-Technical Loss using Support Vector Machine in Smart Grid," *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 2019, pp. 1-6, doi: 10.1109/ICASERT.2019.8934601

[14] A. Takiddin, M. Ismail, M. Nabil, M. M. E. A. Mahmoud and E. Serpedin, "Detecting Electricity Theft Cyber-Attacks in AMI Networks Using Deep Vector Embeddings," *in IEEE Systems Journal*, vol. 15, no. 3, pp. 4189-4198, Sept. 2021, doi: 10.1109/JSYST.202 -0.3030238.

[15] Z. Yan and H. Wen, "Electricity Theft Detection Base on Extreme Gradient Boosting in AMI," *in IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1-9, 2021, Art no. 2504909, doi: 10.1109/TIM.2020.3048784.

[16] J. Yang et al., "Non-technical Loss Detection using Missing Values' Pattern," *2020 International Conference on Smart Grid and Clean Energy Technologies (ICSGCE)*, 2020, pp. 149-154, doi: 10.1109/ICSGCE49177.2020.9275601.

[17] Khan, Zahoor A., Muhammad Adil, Nadeem Javaid, Malik N. Saqib, Muhammad Shafiq, and Jin-Ghoo Choi. 2020. "Electricity Theft Detection Using Supervised Learning Techniques on Smart Meter Data" *Sustainability* 12, no 19: 8023.

[18] Saddam Hussain, Mohd. Wazir Mustafa, Touqeer A. Jumani, Shadi Khan Baloch, Hammad Alotaibi, Ilyas Khan, Afrasyab Khan,A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection,*Energy Reports*,Volume 7,2021,Pages 4425-4436,ISSN 2352-4847,https://doi.org/10.1016/j.egyr.2021.07.008.

[19] Shuan Li, Yinghua Han, Xu Yao, Song Yingchen, Jinkuan Wang, Qiang Zhao, "Electricity Theft Detection in Power Grids with Deep Learning and Random Forests",*Journal of Electrical and Computer Engineering*, vol. 2019, Article ID 4136874, 12 pages, 2019. https://doi.org/10.1155/2019/4136874.

[20] Han, J.W., Kamber, M. and Pei, J. (2012) Data Mining Concepts and Techniques. *3rd Edition, Morgan Kaufmann Publishers*, Waltham.

[21] S. O. Tehrani, M. H. Y. Moghaddam and M. Asadi, "Decision Tree based Electricity Theft Detection in Smart Grid," *2020 4th International Conference on Smart City, Internet of Things and Applications (SCIOT)*, 2020, pp. 46-51, doi: 10.1109/SCIOT50840.2020.9250-194.

List of Publication

[1] Christine Mariam Mammen, R Sheeba and N Naufal,"False data Injection Attack on Smart Energy Meter",*International Conference on Communications and Cyber-Physical Engineering (ICCCE)*,2022. (Accepted)