CHAPTER-1

INTRODUCTION

We are persisting in an eminently hi-tech age. It is a period where lots of technologies are growing up to engender huge mass of data. These massive data in a tremendous velocity is a by-product of each domain of modern society, results in an accelerating Bigdata. Internet acceptance turns worldwide as a result many electronic devices such as smartphones, cell phones, laptops, sensors, smart kitchen, and household appliances provoked mammoth amount of digitized data. With the period of time data expanded in each domain. A domain like medical, credit card, banks etc. can be benefited from Bigdata analytics but extracting a useful pattern from these data's requires extraordinary skills. Usage of current technologies liberates huge amount of data which is helpful in many prospective for the various industries. These enormous historical transactional data produces as a by-product from several domains can be utilized for number of decision making purpose but only if it can be processed and utilize properly. The conventional old approaches of performing data analysis need to be altered with the augmentation of Bigdata. Therefore urgent necessitate of new tools and technologies for the data analytics in each domain is in demand.

Data mining is a technique for extracting essential patterns and pulling out knowledge from huge set of records. That extracted pattern from the massive quantity of data is advantageous for many areas such as fraud detection, disease detection, market analysis, customer retention, science exploration, etc. depending upon the nature of data. Data mining uses a machine learning algorithm to discover relevant information from the massive data set. Machine (ML) techniques includes a wide range of algorithms for learning predictive rules from historical data to build a model that can predict future data.

Machine Learning algorithms require huge set of clean data as a pre-requisite for training purpose but almost every dataset when gathered from different sources is unclean. Missing values, redundancy, outliers, integrity constraints violation, etc. are some of the problems which possess challenges for machine learning algorithms. Imbalance dataset is another complexity for machine learning algorithms. Imbalance dataset is consisting of the non-equal distribution of target class where majority class always dominates over minority class. ML algorithms does not go well with an imbalance data in terms of verdict patterns from

it. To overcome the complexity possess by imbalance dataset, modification of data is mandatory. Sampling is a technique which deals with the modification of dataset either by Under Sampling (US) or Over Sampling (OS). But both of these traditional sampling techniques suffers from several disadvantages.

Messy data and data imbalance is some of the major problems for ML algorithms and thus a focus area of this research. Above mentioned problems are extensively studied and taken care of in the current study. Many approaches to solve these problems are present but each one of them suffers from challenges and treat them separately. Therefore, there is a need of single solution for messy data and imbalanced data problem which can reduce the complexity of Bigdata. This research addressed two important parameter of data cleaning- Redundancy and outliers as because these exists in huge amount in almost every real world dataset making them dirty and thus degrading the performances of machine learning algorithm. Our proposed technique, Hybridization Preprocessing Resampling Technique (HPRT), proves as a single solution for both the problem and capable of removing redundancy and dropping outliers together, hence reducing the complexity of an imbalance dataset, which enhances performance of ML classification algorithm.

1.1. BigData : An Introduction

In 2001, Doug Laney illustrated remarkable changes in the volume of Data [1]. He predicted that data will flow like a flood in the coming year. He defines Bigdata in terms of 3 V's, (Volume, Velocity and Variety). In 2011, a report was presented by IDC [2], regarding massive growth of Bigdata over the period of time which recorded 130 Exabyte data produced in 2005, grew to 1,227 Exabyte till 2010 and estimated to grow almost 45% in another five to six year. Remarkable growth of data requires advancements in technologies during mining and analyzing it, which give rise to many new Bigdata processing tools.

Bigdata is a term associated with the dataset where volume is massive beyond the storage and analyzing capacity of traditional database management solution. Bigdata is a combination of 3V's. Volume is a term related to huge size of data [1]. There is much controversy related to size of Bigdata. For some organization few gigabyte or terabyte of data is considered as Bigdata but for some large organizations petabyte's or even Exabyte of data is considered as Bigdata.

With the arrival of Bigdata remarkable changes in the format of data has been observed giving rise to Variety of Bigdata. Data coming from different sources tends to form variety of data. Therefore, now-a-days with structured data, semi structured and un-structured data are

also observed [3]. Structured data is fit for tabular structure, in row and column format. Hence, traditional management systems can easily process these types of data but semi-structured and un-structured data are real challenges for traditional database management tools [17]. Weblogs and social media text are the example of semi-structured data whereas audio, video and an image comes under the category of un-structured data.

Velocity refers to speed at which data is coming from different sources and stored in a server [3].Earlier batch process was used to process the chunk of data. But now this approach is not suitable because of increased momentum of Bigdata. Now-a-days, Bigdata is streaming continuously as a real time into the server and these data is useful only if it can be processed in real time or near real time. These 3 V's of Bigdata are precious for an organization but needed extra-ordinary tools and techniques capable of dealing them.

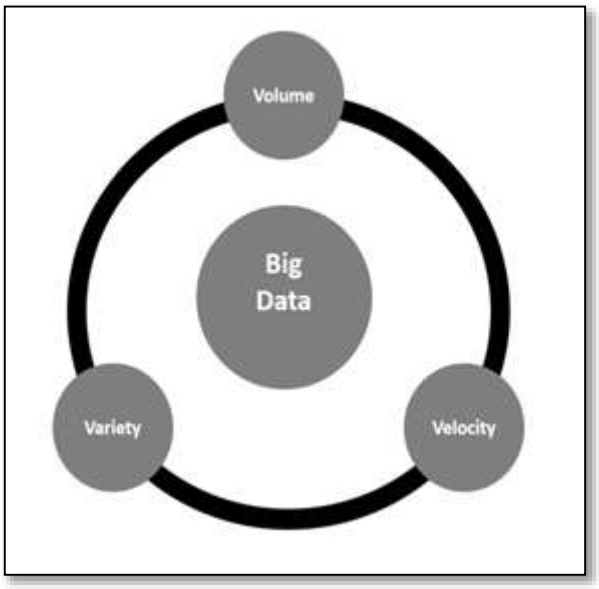


Figure 1.1: 3 V's of Bigdata [24]

Exponential growth in all the three V's of Bigdata continuously causes a real challenge for companies while storing and processing them. Another challenge occurs in extracting a new useful insight from huge volume of semi-structured and un-structured data coming from different sources in an incredible speed. Big organization and companies recognize power of Bigdata and realizes that it can do wonders pair with intelligent tools and techniques [4].

1.1.1. BigData Challenges and Technologies

With the rise in Bigdata, unstructured and semi-structured data grow and thus seized obstacles for management tools. Extensive variety of enormous volume of data such as video,

audio, log files, sensor data in an incredible velocity collectively shape three V's of Bigdata. These three V's should handle conscientiously to obtain precious insight from it. IBM reported 2.5 quintillion bytes of data to reveal per day [5], youtube spawn more than 300 videos every minute, 64,750,000 Twitter users are initiated 175 million tweets per day. Google having 1 billion accounts, 40 million photos are shared per day by Instagram. To process, this amount of data traditional application fails and therefore Hadoop comes into the picture.

Hadoop is a parallel processing system which runs on commodity hardware, providing scalability to its users to an extent not possible for standard SQL. Apart from this other technologies like NoSql, Spark, etc. are also use for Bigdata storage and processing. Hadoop has been projected for Bigdata to support parallel processing together with Google MR (Map Reduce) programming model and its eco system [6]. ML techniques with its immense computational and automation ability are capable of managing, using and analyzing Bigdata in a more systematic and efficient way than ever before [7].

1.2. Machine Learning : An Introduction

A Data Warehouse (DW) is termed discover by Bill Inmon in 1990 [8]. He stated data warehouse as a subject orientated, time variant, and an integrated and nonvolatile collection of data. A DW can be helpful to organize data in various ways but the traditional warehouse faces several issues with storage and management of unstructured Bigdata forming several data island and hence possesses complication during their use and therefore the cost of management of such dataset increases. Hence new and intelligent tools and techniques were highly in demand.

Data mining empower to un-turn essential patterns and set relationships out of massive volume of data [9]. It is a process of pulling out knowledge from records from the huge quantity of data and thus advantageous for many areas such as fraud detection, disease detection, market analysis, customer retention, science exploration, etc. depending upon the nature of data. Data mining is a branch of computer science containing machine learning, statistics, text mining and artificial intelligence etc. These combinational techniques of computer science are influential enough to dig out the precious pattern from the ocean of Bigdata. Data mining uses machine learning algorithm to discover relevant information from the massive data set.

ML includes wide range of algorithms for learning predictive rules from historical data and to build a model that can predict unseen future data. According to Arthur Samuel [10], machine learning is a field of computer sciences for providing an opportunity to a computer for learning from the data. Without programmed explicitly machines are fused with artificial

intelligence and is capable to act and think as human being accepting machine learning and Bigdata techniques. Therefore we can say that Bigdata is employing data mining and data mining is employing big data for surely favorable investigation.

In the presence of massive amount of data, machine learning and data mining are two esteemed topics of research. Bigdata consists of a large dataset, extracting information from which is very important. Data mining provides an opportunity to extracts important nuggets from gigantic data. It uses machine learning algorithms to perform several kinds of mining job. Machine learning and data mining are gaining popularity because of its usage in data analytics. Machine learning is a branch of computer science having the ability to self-train a system in an efficient way, so that system learns automatically from a data and also improves with the experiences without being programmed explicitly. It is consist of sets of algorithms that offer a possibility to a software application to predict the precise outcome. The fundamental goal of machine learning is to construct a model, receiving an input and using statistics formula for the prediction of an output when new data arrived in the system. Machine learning looks into the data examples to search patterns and construct decision rules for building a predictive model which can use to predict future data. These predictive models are capable enough of self-learning with experience without any human interruption and take a decision based on certain circumstances. Machine learning can broadly categorize into supervised and unsupervised learning containing wide sets of algorithms in each category.

1.2.1. Types of Machine Learning Techniques

Supervised machine learning algorithms are useful for a dataset which contains target class. Supervised models are functional to unseen data with the usage of past examples of the target class [10]. It acquires advantages of evaluating outcome prediction through comparison with that of original output and calculate the error and modify it accordingly. On the other hand, unsupervised machine learning algorithms are in use when the dataset is not having labels. It is basically applied to extract a hidden pattern from the dataset in an absence of target variable.

1.2.2. Classifier

ML algorithms uses mathematical function for mapping input to group it belongs and the functions are refers as classifier. A number of classification ML algorithms have been studied and applied during the current research. Classification algorithms are capable of building a predictive model upon a dataset. It discovers a pattern from the input dataset

examples and constructs rules which is able of predicting categorical labels, so that model can forecast unseen dataset in future based on historical data.

Figure 1.2: Classifier

The classifier can be mathematically expressed as: q = p (x)

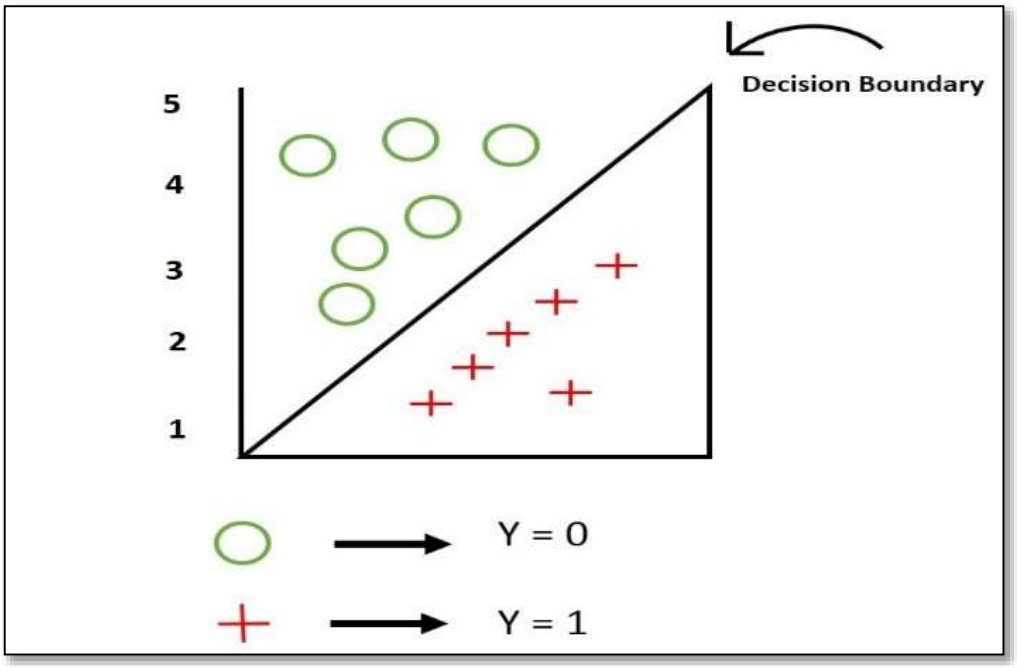


Figure 1.3: Classifier Forming Decision Boundary for Binary Output

A classifier helps in the classification of input data by developing a decision boundary which is clearly observed in figure 1.3. Input sample feature provided to a classifier so that it can predict test sample depending on the labels. The best classifier is one which produces less misclassification. Classifier performance is directly proportional to the quality of the dataset provided to it and therefore the performance of the classifier is data dependent. There is no such rule to select the type of classifier for a particular problem. Several classifiers should be applied upon a dataset and check the result of each one to decide best one. Confusion matrix, accuracy, misclassification, sensitivity, specificity, precision, F1-Score, Receiver Operating Characteristic Curve (ROC) and Area Under Curve (AUC) are some of popular matrices and

measures used to check the performance of the classifier during this study [11]. The confusion matrix is a table like structure consisting of 2 rows and 2 columns for a binary classification for displaying True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Accuracy measure is a total number of correct predictions over the total number of prediction. Misclassification is known as an error, is a measure of a total number of incorrect predictions over the total number of prediction. Sensitivity is the total number of correct positive predictions over the total number of positive. Specificity is a total number of correct negative predictions over a total number of negatives. Precision is a total number of correct positive predictions over a total number of positive predictions. F1-score is a weighted average of recall and precision. ROC curve is used for visualization of classifier performance at all possible threshold; whereas AUC is a measure of the complete region covered the ROC curve.

1.2.3. Supervised Classification Techniques

Supervised classification machine learning techniques are those where training is provided to computers based on input data and this training is utilized for classification of the new observation [12]. Classification can be of binary or multi type depending upon the number of labels present in the dataset. In order to answer classification problems numerous algorithms are available some of those are:

1.2.3.1. Logistic Regression
1.2.3.2. Naïve Bayes
1.2.3.3. Decision Tree
1.2.3.4. Random Forest
1.2.3.5. K-Nearest Neighbor
1.2.3.6. Support Vector Machine
1.2.3.7. Neural Network

1.2.3.1 Logistic Regression

Logistic regression (LG) follows statistical method for classification problem. Logistic regression is very much alike Linear Regression (LR) but linear regression is for number problem and logistic regression is for classification problem. Logistic regression aims to locate the coefficient value for output calculation. The output here is converted into 0 or 1 using a nonlinear function called the logistic function. LG predicts the output depending on one or more independent variable of a dataset. It is an effective and fast learning algorithm

1.2.3.2 Naïve Bayes

Naïve Bayes (NB) classifier as the name suggests based on Bayes theorem. NB presume dataset features are not related to each other, even if they are correlated. It requires a very small portion of a dataset to detect parameters necessary to predict the output. It is a fast classifier. NB predicts the output based on probability which is calculated by every feature individually. It is an easy going model for a large dataset.

1.2.3.3 Decision Tree

Decision Tree (DT) constructs a tree-like structure to solve both classifications as well as a regression problem. DT divides the large dataset into numbers of a small portion to the time division is possible. A DT contains a root node, decision node, and a leaf node. The root node is the topmost decision node which is consisting of more than one branch whereas the leaf node is meant for decision. DT is very easy to understand and is suitable for any types of data.

1.2.3.4 Random Forest

Random Forest (RF) is more or less similar to a DT, but in a RF, multiple trees are constructed to overcome the problem of overfitting of the training set. The RF can be used for regression, classification or any other machine learning task. RF performance is better in most of the cases when compared with the DT.

1.2.3.5 K-Nearest Neighbor

K-Nearest Neighbor (KNN) used during this study for classification problem, because of its accurate result. It is a supervised machine learning algorithm. It starts learning the process by observing a K-group of labeled data. In order to label a class for a new point, it calculates the distance between it with that of K-labeled data and looks for a maximum number of nearest neighbor to that point and new point belongs to the category of class having the maximum number of neighbors. K is a term used to denote the number of neighbors considers during the training process.

1.2.3.6 Support Vector Machine

Support Vector Machine (SVM) classify labels to their respective classes, through constructing hyperplane. It is widely used in various areas including credit card transaction dataset for fraud detection. SVM is capable of dealing high dimensional dataset. SVM is one

among some few machine learning algorithms capable of dealing overfitting. Therefore SVM is considered and compared in current research with some other machine learning technique to check its performance while detecting fraud in credit card transaction dataset and it produces a satisfactory result.

1.2.3.7 Neural Network

Neural Network (NN) is most powerful ML algorithm, which is now a day applied in most of data mining tools. Its working fundamental is just like human brain system. It is a set of numerous interconnected neurons for processing of data into meaningful information. The NN is capable of training a big amount of data very easily and efficiently. NN guarantees to extract meaningful information from a highly complex dataset and form pattern and trends for analysis which is not possible for general human being.

NN is utilized now-a-days for mining data, especially for decision support application. Human being learns from experience and computers are good in following instructions. NN takes best of both the systems i.e. it constructs a model behaves just like human brain neurons in a computer system. NN acquired attractiveness because of its precision during classification and its technique to distinguish complex pattern without difficulty.

NN sometime refers as Artificial Neural Network (ANN) models are extremely efficient and accurate when compared to other machine learning models [13]. NN fabricates the output by mapping multiple input node into numerous output node. NN learns from a set of examples provided to it as an input. Single layer architecture of the neural network is capable of solving the linear separable problem. But the exceedingly complex problem becomes a challenge for ANN. Therefore in order to conquer this obscurity multi-layered perceptron (MLP) was developed in 1950. MLP provides the user, a flexibility to add a number of the hidden layer between the input and output layer to solve the respective problem competently.

Feedforward is the essential architectural procedure of a neural network, intend propagation of the inputs from the input layer to the complete hidden layer present in between the architecture until it reaches the output layer. Back propagation neural network method calculates the error by propagating again from the output layer to hidden layer and then to the input layer. The process is repeated for a number of times until error is reduced. A number of neurons are connected together to form feed forward neural network architecture. The neural network architecture is consist of a number of the layers where the first layer is an input layer and the last layer is the output layer in between a hidden layer of any number. Below Neural network present feed forward neural network architecture of three layers.

The error is computed through back propagation method with the help of gradient descent. During back propagation, in order to minimize the error, weight is regularly updated through gradient descent in each epoch. Gradient descent minimizes loss occurred in training by comparing actual value with that of predicted output. As shown in figure 1.4 each layer connected to the other layer is represented through a weight coefficient. Input feature present in the dataset is received by the input layer. Hidden layer calculates weight sum of all the given input and redirect it to the activation function.

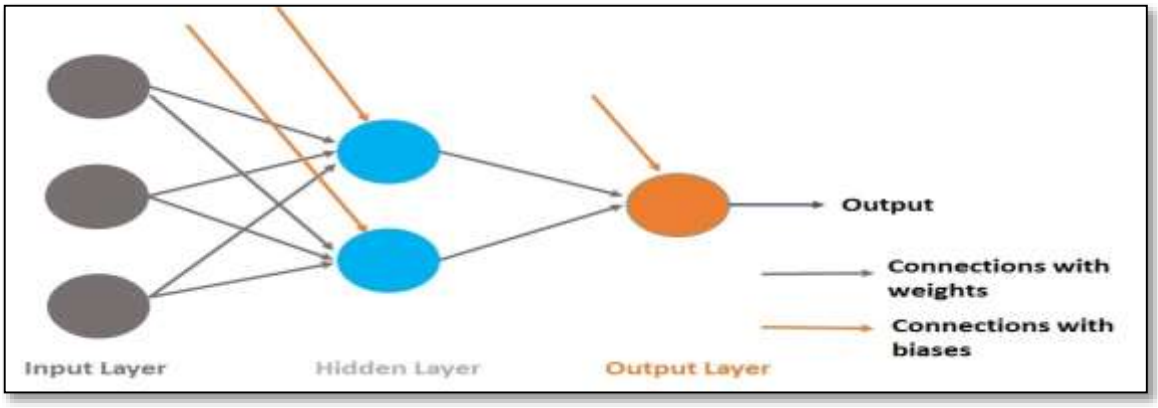


Figure 1.4: Three Layer NN Architecture [13]

$$Neural\ Net_x = \sum_{i=0}^{n} A_i w_{ji} + W_O \qquad \text{.....(1.1)}$$

In equation 1.1 i is an input layer unit, j is a hidden layer unit and w_{ji} is a weight at hidden layer j. Feed forward back propagation architecture can also be applied to (MLP) Multilayer Perceptron. NN multi-layer architecture has been used during this research because of its high performance when compared to other machine learning classification algorithms and its efficiency during processing of huge volume of data.

1.2.4. Performances Hindrances Challenges of Machine Learning Algorithms

ML performance is dependent on the quality of input data to construct a model. Lots of data examples are first and foremost essential obligation of ML algorithm to conduct training which becomes possible with the augmentation of big data. But chaotic data is ineffective for the machine learning algorithm thus creating detrimental circumstances for ML techniques. Almost all the dataset when gathering from diverse sources is unclean. Therefore data

preparation or data cleaning is the extremely significant procedure for cleaning the data before application of machine learning technique.

Data preparation makes dataset suitable for machine learning as messy data is a biggest challenge for machine learning and therefore it has to transform into clean and usable layout beforehand [34]. Poor quality of data generates a weak model have mystified decision-making ability for business.

Almost All real-world data is dirty. Missing values, redundancy, outliers, integrity constraints violation etc. are some of the problems which craft the data grubby [96]. Major and time consuming work is to clean the data by observing anomalies and treat them accordingly [50]. Data cleaning is thus gaining the attention of many researchers and scholars from several years.

Imbalance dataset is another dilemma for machine learning algorithms. Imbalance dataset is a term enforced for a set of data in which target class is consists of more number of negative samples compared to positive samples. Imbalance dataset is consisting of the non-equal distribution of target class where majority class always dominates over minority class [14]. Machine learning algorithm does not go well with an imbalance data in terms of verdict patterns from it. In many cases, it has been found that minority class is neglected and majority class dominates over the data. Models constructed on this imbalance class posse high accuracy but with a lot of misclassification. High accuracy is generated because the machine learning algorithm instead of extracting the pattern deviates much towards the majority class and minority class is treated as noise by them. This problem becomes more rigorous with the emergence of Bigdata. Due to the augment of Bigdata, inconsistency and imbalance property of the real-time dataset al.so expand. Due to the intensification of Bigdata every domain is producing an enormous amount of data. The colossal historical data can act as a key to machine learning and data mining technique for building decision rule but dataset should be clean and balanced. A huge amount of messy data is the root cause of data imbalance dataset thus causing complexity for ML technique during the learning process. Fraud detection, disease detection, network intrusion detection etc. are some of the area affected by the imbalance problem. Therefore lots of researches exemplify various solutions to these problems. Lots of tools and technique have been already presented by the researchers and scholars for dealing with the situation.

Redundancy and outliers along with data imbalance is a major problem of ML algorithm and thus a focus area of this research. Redundancy is a term given to multiple records present in the dataset. The manifold records thus demean the quality of a dataset by bringing

inconsistency within it. In many cases, redundant data causes challenges for machine learning technique for the training of a given dataset. Therefore removal of redundant data is extremely significant and should be taken care off in a pre-processing stage [17].

Outliers are extreme values in an observation of dataset that falls away from the range of other similar records [16]. Outliers occur in a dataset during the experiment, if any error is generated. Most of the machines learning algorithms are sensitive towards the outliers and unable to fabricate accurate result. Presence of enormous outliers in a dataset can generate poor model consuming extended training time. Outliers can occur at any of the phases of data collection and data manipulation because of several types of (human, instrument, experiment, data processing) error. Therefore to permit ML techniques to work in their way, outliers should be detected and removed or replaced in the preprocessing stages depending upon the problem.

1.2.5. Popular Approaches for Data Imbalance Problem

Sampling is a technique which deals with the modification of dataset either by US or OS [15]. US reduces the majority class and OS increases minority class such as to fetch normal distribution for a dataset. But both of sampling possess several disadvantages where US cause information loss and OS performance is not good for high dimensional dataset with an added overfittng problem, which is again a challenge for ML algorithms.

1.3. Summary

Messy data and data imbalance is a real challenges for ML algorithms and should be taken care off during data preprocessing stages. Several solution is present for considering both the issues with certain limitations but there is a need of single solution capable of solving two major dataset complexity at a time for saving lots of effort and time - a technique capable of (a) cleaning the dataset (b) while balancing it. Therefore in order to fulfil the need intension here is to design a technique capable of removing redundancy and outliers from a majority sample of a dataset, which cleans and at the same time reduces it and then increasing minority sample for balancing the dataset.

CHAPTER-2

Literature Review

Several studies had been carried out during this research period, a summary of diverse study which are relevant to the topic are presented as a brief literature review.

2.1. Bigdata: Tools, Techniques and Architecture

Dough Laney (2001) observed significant change in the size of data. He noticed the mounting data mass, day after day and estimated that data will flow enormously in future. He defined Bigdata in terms of 3V's, Volume, Velocity, and Variety, where volume refers to the huge amount of data, velocity refers escalating rate at which data is generated and variety refers to the diverse format of data gathering from a different source [1].

Mehmed Kantardzic (2002) illustrated the concept of data mining as an abstraction or discovery of patterns in a huge dataset with the help of a combination of various technologies, such as machine learning, statistics etc. Data mining is one of the branches of computer science applied for information extraction with some smart technique from a dataset and converts the information in a format that can use for further analysis. He further revealed that patterns are extracted from the dataset from centuries. At that age pattern discovered manually using Bayes theorem (1700) and then regression analysis (1800). But with the rise of Bigdata, growth in the size of the dataset has been encountered with it mounting complexity too. Therefore a method for mining has to be revised. Automated data processing combination with other technique, such as neural network, cluster analysis, and generic algorithm, support vector machine, decision tree and decision rules replaced with old techniques of data mining has been suggested by the author. Data mining map up artificial intelligent and applied statistics which can act as a database management tool for process learning and identifying algorithms in a more efficient way and can utilize for huge datasets [20].

Doug Cutting and Mike Caferella (2005) designed Nutch to support a web crawler. It is based on google file system and Google Map Reduce (MR). Web crawler also known as wen indexing and is designed for storing and processing large amount of dataset. The framework were written in Java and consists of HDFS and MR [21].

According to Abdelrahman et al. (2007), MR, a software for processing large datasets. It divides into Mapper and Reducers task for Bigdata computation. Mapper task is responsible for splitting the data into key and value pair and output intermediate value. Those values are processed by reducer to generate an actual outcome [22].

Vidyasagar S.D (2007) suggested Hadoop as a Bigdata solution. It is an open source Apache framework divides into two part, HDFS, and MR. Hadoop mechanism work on the model of master and slave architecture consisting one name Node (Master) and a number of Data Node (slave). HDFS is primarily used by Hadoop applications because of its high fault tolerance, cost-effective, reliable, and data processing system, designed to run on reasonably priced commodity hardware. The data are stored in terabytes (TB) and petabytes of data distributed unstructured data effortlessly [23].

Shilpa and Kaur (2007) experiential challenges faced by Bigdata such as: a) unstructured data, b) real time analysis, c) fault tolerance. Almost 80% of data generated through internet was un-structured. Out of which storage capacity and processing was so crucial, that existing techniques and technologies like NoSql, RDBMS fails to handle. Authors recommended HADOOP, as one of the best technology as it can store and process un-structured data efficiently. At last they suggested that in order to avail benefits of Bigdata there must be support for fundamental research [24].

Dittrich (2008) concluded that adoption of Hadoop, became standard for many officialdoms, various techniques was built to boost up the performance of Hadoop. Main challenge of Hadoop's was physical data organization and data layout and indexes. Therefore he proposed HAIL to overcome those challenges. HAIL: indexing technique, improved 70% of Hadoop performance [25].

2.1.1 Bigdata: Issues and Solutions

B. Arputhamary et al. (2008) affirmed data integration as a main issue of large data sets which is managed by Extract, Transform and Load (ETL) tools such as DW. It is the procedure of transforming all manifold data formats into a solitary format and combining them in one place. Mostly data were generated from social networks, web server logs where sensors used to gather climate information, stock market data, e-mails, transaction records, web click streams, etc. Most of these data were present in unstructured or semi structured forms. Major challenges of data integration in Bigdata environment were a) schema mapping, b) Record linkage and c) Data fusion. Authors mentioned many existing techniques and framework but none of these techniques were appropriate. Therefore they suggested urgent need of new

frameworks, techniques and algorithms or any new mechanism to handle the data integration issue in Bigdata environment [26].

Facebook Data Infrastructure Team (2009) determined to bring the concepts SQL into Hadoop. As Hadoop exists as only solution for storage and processing of Bigdata but its programming was complex and time consuming. People were missing easy to write SQL-queries. Therefore Hive has been presented, as an open-source data warehousing solution which is built on top of Hadoop. It is just like SQL and support quries- HiveQL, which were compiled into map-reduce jobs executed on Hadoop. They further added hive as a new born technique and future work were towards making HiveQL subsume SQL syntax, exploring column as storage, welcome appropriate data structure for faster processing, enhancing the JDBC and ODBC drivers for Hive for integration with commercial BI tools, Exploring multi-based optimization techniques and performing generic n-way joins in a single map-reduce job were pending task that time. There were some limitations like hive only support equality in a join predicate, Hive does not support inserting or updating, according to them resource scheduling were weak in Hadoop [27].

Yahoo Team (2009) discovered dataflow language for Bigdata environment, between SQL and Map-Reduce, known as Pig. It offers SQL like data constructs, compiled into sequences of MR jobs, and implemented in the Hadoop MR environment. Pig generates the code in language known as Pig Latin. They shared the tests or experiments faced by them in developing Pig, which consist implementation problems as well as challenges in moving the project from a research team to improvement team and converting it to open-source. They further concluded that Query optimization, improving SQL interface, improvement in grouping and joining queries, skew handling, etc. are some of the holes which has to fix in order to obtain optimized data flow environment [28].

Clark Bradley & et al. (2013) published technical paper to present the need of good data structures in the DW environment. To prove the fact that structuring data properly in HIVE were important as in RDBMS, experiment was conducted, which was based on: 1) File format, 2) Compression of data from uncompressed text files to compressed sequence files, 3) Using Indexes, 4) Partitioning of table. As a result drastic improvement was observed in performance, after proper structuring of data [29].

Bo Li (2013) after in depth survey of storage and DW of Big Data technologies claimed that although progress had been made in DW research much has to be done: 1) Handling streaming high rate data in relational models remains an open problem, 2) Statistical analysis

and ML algorithms for Bigdata need to be more robust and easier to use, 3) RC File has adopted as default data placement method in HIVE and PIG still has to be optimize in future [30].

2.2 Machine Learning: Issues and Solutions

Christopher M. Bishop (2006) illustrated ML and data mining are more or less useful for a similar task with little difference. ML predicts the data depending on the known property present in the dataset, learns from training data, whereas data mining focuses mostly to discover unknown fact from data. Data mining uses many ML technologies to discover hidden pattern from the huge set and ML uses data mining techniques as unsupervised learning for prediction from an unknown set of data. Rising trend had been recognized in the demands and usage of ML in the last few years with the growth of Bigdata for analyzing big chunks of data. Conventional method using statistics for data extraction and interpretation had been modified by automatic generic methods sets through machine learning. The traditional method of data analysis based on trial and error becomes difficult with the large and heterogeneous dataset. Machine learning provides a smart solution in terms of fast and efficient tools and data-driven models for processing real-time data with accurate results [31].

S.B. Kotsiantis et al. (2006) explained the importance of data cleaning for machine learning algorithms. Redundant, irrelevant, noisy or unreliable data causes difficulties during the training phase. Thus performance of ML algorithms are deeply affected by these factors. Therefore before mining of such dataset preprocessing is an important and time-consuming approach. Steps includes during data preprocessing are data cleaning, normalization, feature extraction, transformation, data selection, etc. The output of a data preprocessing is a clean training set, on which machine learning algorithms can apply. Real-world datasets mostly consist of noises in terms of missing or null values, outliers, redundancy, etc. Supervised machine learning algorithm performance enhances after data preprocessing but eliminating noise during the preprocessing step is the most time-consuming problem [32]. Almost 80% of time is dedicated to data cleaning out of total data preprocessing time. Therefore there were a need of robust automated or semi-automated tools and techniques for data cleaning.

K. R. Seeja (2014) pointed data mining and machine learning is an emerging field for solving many real-world problems. Almost all domains can use various machine learning classification algorithms such as SVM, neural network, decision tree, random forest for prediction of future data. Insufficient data, messy data, imbalance data are some of the challenges arises mostly in the dataset and therefore these algorithms cannot achieve superior outcome during training of ML models [71].

Zena M.Hira (2015) marked feature selection as a significant step during data preprocessing. Mostly real-world datasets are consisting of many features. Some of them may not be required by ML algorithms during the training phase. Redundancy is another major challenge; there can be numerous features correlated to each other, removing such features from the dataset are an essential task. Feature selection is a process of identification of important feature and removing the irrelevant and redundant feature. A dimensionality reduction enhances ML algorithms performance and allows it to work faster and efficient manner and hence improves the accuracy of a classification algorithm [33].

Artur d'Avila Garcez (2010) developed an improved sparse oracle based adaptive rule extraction algorithm for constructing of an easily understandable rule with the help of the neural network. Proposed optimization technique uses a search algorithm for enhancing NN classifier performance through partial area optimization. In order to check the performance of an algorithm experiment has been conducted on large dataset and improvement has been noticed. But the framework created during this work was firm and need to improve so that performance of MLP classifier enhances [35].

Ong Shu Yee et al. (2018) suggested combination of two or more data mining and ML techniques for additional benefit during prediction. Authors combined K2 and tree augmented Naïve Bayes (TAN) and Naïve Bayes and logistic classifier and observed that these techniques brought accurate and precise result. But before applying these algorithm data should be clean and balance because ML algorithms are not good for complex dataset and therefore it should preprocessed [36].

Wei Feng et al. (2018) illustrated that due to the expansion of big data, data sizes also enlarged. Therefore most of the real-time dataset become inconsistent and exposed to form imbalance dataset. Imbalance dataset is a term given to a dataset when majority class samples are more and minority class is negligible compared to former which causes problems for the machine learning algorithm. According to them traditional ML techniques are not competent with large and complex Bigdata and therefore they should be alter. They further added, combination of more than one algorithm is an excellent alternative for constructing a good model. Boosting and stacking [37] ensembles method combine together to produce a better result.

Satyam Maheshwari (2007) presented important characteristic of imbalance class. He suggested various problems along with its solution related to an imbalance problem. He explained Sampling as an approach to deal with the imbalance problem of data. It is basically of two types US and OS. US is an approach of reducing majority class sample and

oversampling is an approach of increasing minority class by adding artificial features to a minority class. None of the methods alone is capable of generating satisfactory result because of US causes loss of important feature whereas OS causes over fitting within a dataset. He concluded no existing approach as an optimal solution and therefore combinational effect of both of these approaches can be a better approach to balance the dataset [38].

Satyam Maheshwari (2011) developed E-SVM, Evolutionary OS with a combination of SVM and clustering method for classification of an imbalance dataset. This algorithm uses the concept of OS where several GA algorithm is used to increase positive sample and then data clustering is used in training set to remove redundancy from both the classes. An experiment accompanied to verify the performance of the proposed algorithm on four diverse imbalanced datasets. Result yield confirms good performance of an algorithm, better than traditional approaches and improved computational efficiency [40].

Mikel Galar et al. (2011) observed that taxonomy of ensemble method as a significant approach to solve the problem of class imbalance, for this reason, ensemble methods were distributed into four families depending upon their base learning technique. The more precise result can obtain after data preprocessing techniques and training the dataset with single classifier because more number of the classifier can make the situation complex. RUSBoosT or under bagging although being simple technique capable of generating high performance than many complex processes. Author experiences the performance of RUSBoost as the most excellent, which is computationally simple than the other traditional approaches [41].

Nitesh V.Chawla et al. (2002) developed SMOTE as an oversampling technique. Synthetic minority over-sampling technique, as the name suggest adds synthetic features, to the minority class through calculation of probability distribution with an intention to regenerate minority class by constructing larger decision boundary so as to capture nearest minority class. But SMOTE oversampling suffers from various challenges such as overfitting and increase in computational time dealing with large dataset [42].

Zhuoyuan Zheng (2015) discussed different problems of Synthetic Minority Over sampling Technique (SMOTE) and therefore several variations of SMOTE has been proposed. SNOCC variation of SMOTE to overcome over-fitting challenges has been proposed. The experiment is conducted to compare the performance of SNOCC with that of SMOTE and found to produce a better result than SMOTE but for larger dataset SNOCC performance was not good. SNOCC cannot deal with all types of data. Therefore working with Boolean and categorical variable was left as a future work [39].

Rheza et al. (2018) implemented Ripple SMOTE, a novel oversampling approach to overcome the challenges faced by SMOTE. In Ripple SMOTE addition of synthetic sample starts from border whereas ripple change its position towards centroid such that new synthetic samples are added by calculating k-nearest neighbor sample and ripple. An experiment conducted to prove the efficiency of a proposed algorithm and observed that its application improve imbalance set performances which can handle minority class exceptionally. In future the algorithm can be modify to minimize noise occurred through Ripple SMOTE so that more realistic sample can be formed during the process of OS as Ripple SMOTE was not suitable for real dataset [43].

Mostafizur Rahman and D. N. Davis (2014) presented an approach to handle imbalance class simultaneously between and within the class based on cluster-based oversampling. Combination of boosting and data boost IM algorithm a (data generation algorithm) has been use to oversample minority class of an imbalanced dataset. It was also observed that cluster based under-sampling technique reduces the data, based on the clustering technique but reducing data sample causes loss of important information within the class. On the other side increase in examples causes duplication of data without emphasizing any extra information into it and hence can provide overfitting and also increases computational cost. Therefore best solution to achieve optimum result in still required [44].

Tasadduq Imam et al. (2006) constructed Z-SVM, an enhanced version of SVM, where parameter Z shifts hyperplane to maximize g- mean value or imbalanced dataset. They revealed disadvantages of SVM classifier during mining of an imbalanced class. Decision boundary of SVM is deviated towards the minority class due to which rate of misclassification increases. Therefore modification in classification algorithm is applied in such a way that modified algorithm can deal with an imbalance problem of the dataset in initial stage before learning the rule for model building. [45].

Yuxuan Li (2011) developed kENN, by combining KNN with exemplar generalization so that it maximizes the decision boundary of the class containing minority features of training examples. kENN select minority sample by calculating a group of positive pivot points and generalize through Gaussian ball. Then KNN compute the distance of each pivot positive instances nearest neighbors. This combinational effect proved to be significant for minority class and thus helps in reduction of misclassification rate and hence covered disadvantages of KNN algorithm. At last authors suggested that to resolve the complexity in a dataset combinational effect of an algorithm can be fruitful [46].

Siyang Zhang and Fangjun Kuang (2012) suggested hybridization is an efficient technique which gains popularity in the coming year. Hybridization is a combination of two or more algorithm to reduce imbalance problem of a dataset and increase the performance of classification. Apart from the imbalance problem hybridization solves various other problems in sampling such as cost matrix optimization, feature selection etc. Sampling using SMOTE oversampling approach and cost-sensitive learning, combined together to enhance the performance of SVM classification algorithm. PSO was used as an optimization of SVM feature selection and the neural network is used to build the classification model of a high imbalance data for fault diagnosis of the power transformer [47].

Vaishali Ganganwar (2012) observed recent research for an imbalance class stepped towards hybrid algorithm. A high hybridization technique, combining PSO, random oversampling decision tree and feature selection is presented for highly imbalanced Zoo datasets but here decision tree doesn't prove as effective thus possess complexity during parameter selection. This challenge was reduced through HDDT Hellinger Distance Decision Tree which was proposed to overcome the complexity mentioned above and Hellinger Distance is used for splitting [48].

Keshav Dahal et al. (2015) developed GAFNN, GA-based learning approach for FNN's which is consist of three steps. In the 1st step self-organization algorithm is used to initialize membership function of both input and output variables with the help of determination of centers and width. GA based algorithm is used in the 2nd step for identification of fuzzy rule. In the 3rd step parameter and architecture is tuned using back propagation algorithms [49].

Radha R and Murlidhara S (2016) proposed speedy feature selection method as a feature selection algorithm for removing redundancy based on unsupervised learning with entropy. This algorithm removes redundancy and thus train model in a less time when compares to existing feature selection methods like Relief and FSGai-ra. It is found to be more effective than FSCor, FSCon and FSGai-ra. Redundancy means the occurrence of multiple copies of same records in a dataset which causes inconsistency in a database. It can occur accidentally or during back up of data collection process. The classification model were deeply affected by redundancy, causes undignified performance in terms of exactness. Therefore many researchers from time to time proposed many algorithms for removal of irrelevant and redundant data. Many clustering algorithms like K-Means has been applied to a large dataset for removal of redundant and irrelevant data from training dataset [50].

Shuchu Han et al. (2017) proposed SFG, the graph-based method for the detection and removal of redundant features from high dimension dataset. The proposed algorithm were

based on sparse learning. SFG, sparse feature graph, which is unsupervised feature selection framework and capable of modeling two redundant features as well as detecting two redundant groups too. SFG divides entries to feature in a dataset into groups and then redundant data is detected and removed from the group, thus increasing the consistency of the data [51].

Qinbao Song et al. (2013) proposed FAST, an effective and efficient feature selection algorithm based on graph-theoretic clustering. The FAST algorithm consists of various steps. In the first step, the graph-theoretic clustering method were applied to divides entire feature into a number of clusters. In the next step, most significant features were selected from the cluster to make a subset of most relevant features. Prim's algorithm was used for managing huge dataset with reducing time complexity. It also deals with feature interaction which was very helpful during the feature selection process [52].

Annalisa Appice et al. (2004) proposed REDUCE algorithm to reduce the duplicate boolean feature from a dataset. The algorithm works on pairwise comparison of features. The problem here was that it was only restricted to two classes problem. REFER redundant feature reduction an extension to REDUCE, had been proposed which overcomes the challenges causes by REDUCE algorithm. Here the process of redundant detection method has been enhanced without losing consistency of reduces feature set. Secondly, the method has been enhanced to work with the multiclass problem [53].

Bharati Kamble and Kanchan Doke (2017) proposed SLOF, local outlier factor for the detection of an outlier. This method uses the local outlier factor of a given data based on local density which was measured with that of its neighbor. The outlier is an extreme value recorded in observation of a dataset that is far away from the other similar values. It occurs mostly due to the experimental or another type of errors. It is a noisy data which should be removed from dataset because large numbers of outliers in a dataset misguide machine learning algorithm at some stage in the training phase. Statistical based depth based and distance based were some of the popular techniques used for detection of outliers. As outlier detection is gaining popularity in recent year's lots of research is carried out to discover a new tool for this purpose [54]. SLOF performances was good but removal of lots of outliers causes loss of information.

Tze Siong Lau et al. (2018) Non-parametric composite outlier detection was proposed for the detection of outlier data stream. GLRT, generalized likelihood ratio test was developed for calculating data stream. GLRT found to be exponentially consistent with the knowledge of kullbacklier divergence between normal and outlier distribution. Knowing Chernoff distance between normal and outlier distribution was also observed to be consistent with GLRT but

without knowing the distribution distance exponential test was not exist consistent still, GLRT was consistent with a low threshold [55].

Manju Kaushik and Bhawana Mathur (2014) explained during past few years, with the rise of Bigdata processing and analyzing a large amount of dataset clustering is an important approach. Clustering is essential for insertion of similar items into one group, thus presenting a simple hence systematic and powerful approach to deal with big amount of data. At present clustering technique is approachable for many big data issues, starts from to health care departments to the financial department. It many research clustering is found to be used for identifying and grouping redundant data from the different dataset, so further removal of it from a dataset become easier. Millions of records and web pages are clustered with this approach. With the increase in usage of its usage demands also increases and hence a more powerful set of clustering algorithm capable of exploring big data for a number of times without any unnecessary checking step is today's expectation. For this purpose, several research papers with the modification of the number of clustering algorithm had been published [56].

Rishikesh Suryawanshi et al. (2016) presented modified version of K-Means approach for clustering Bigdata. K-Means algorithm extended to become more efficient, less time consuming, better clustering and reducing complexities. This approach lowers down the workload of the immense large dataset to a great extent. The algorithm searches for starting centroid and marks an interval between features that can change their cluster with that of feature which will not change their cluster during a number of iteration [57].

SK Ahammad Fahad et al. (2016) presented a novel approach for big data analysis based on modified K-Means algorithm which overcomes the problems of traditional K-Means algorithm. Traditional K-Means algorithm performs an uncertain number of iteration which was observed and eliminating through the proposed algorithm by fixing a number of iterations without losing its efficiency. The technique proposed were capable of generating high accuracy with reducing time complexity and found to be scalable [58].

Ankita Sinha et al. (2016) proposed K-Means++ algorithm to overcome the listed drawback of K- Means clustering algorithm. One drawback of the K-Mean algorithm was that, it takes a number of the cluster as an input parameter which is not always possible to know in advance with an extremely large dataset. Therefore, K-Mean ++ were proposed to automate clusters number depending upon the data. It were implementing on spark program framework for Bigdata. The proposed K-Mean ++ algorithm is an enhanced version of the K-Mean algorithm [59].

Sapna S (2016) presented the use of the neural network for processing of huge dataset to identify thyroid patient. Here disease can be detected using back propagation while training takes place through feed forward with a gradient descent optimization technique. She further explains Neural Network as a very popular among machine learning techniques for extracting valuable information from Bigdata due to its robustness. MLP provides lots of flexibility to its user to build robust architecture consist of numerous hidden layers. Lots of papers discuss the use of a neural network to solve various business problems. The neural network is being used in the health care department for the detection of various diseases. A paper presented the use of neural network based model in the early detection and prediction of cancer through discovering pattern between input and output of a dataset [60].

Aman Gulati et al. (2017) presented the Neural Network based fraud detection framework. In most cases, fraud is detected after the completion of the transaction but this approach was not suitable for fraud detection tools. Therefore, in order to overcome this challenge proposed solution deals with the detection of fraud on the spot during the transaction. Behavioral and location analysis was used which tracks the behavior of spending money on the cardholder. Whenever some deviated pattern is observed will notify the system about fraud transaction and will handle on the spot itself [61].

Raghavendra Patidar et al. (2011) presented combination of neural network with a genetic algorithm to identify the fraudulent transaction. A genetic algorithm is used for decision making purpose about various neural network architecture such as number hidden layer, a number of nodes, network design etc. feed forward back propagation supervised learning are used for the training purpose [62].

Massimiliano Zanin et al. (2018) used combinational effect of hybrid data mining network classification algorithm for the detection of an illegitimate feature in a dataset. It was influenced by the work of neural reconstruction algorithm which deals with the creation of a representation of deviated single instances from a group of reference. Based on this, propose work discusses the importance of integrated complex network and data mining as an extraordinary tool where complex networks provide a view to data in an efficient way [63].

2.3 Research Gaps

With the advent of Bigdata nearly all the real world dataset posed several challenges due to which machine learning algorithm performances degraded. Out of many complexity exists in dataset, messy data and imbalanced data are two complexities exists almost in each real world dataset and hence focus area of this research. Several approaches to overcome above

mentioned problem premeditated and recorded but none of the technique proves to be a perfect solution .The point to note is, data cleaning and data sampling is the process which is applied as initial but as a different activity for almost every datasets to deal with each problem individually. In data cleaning redundancy, outliers, missing values, etc. is taken care of and in data resampling data is modified in such a way to reduce complexity given by data imbalance problem. Again, current approaches to balance the data suffer from several challenges. Both of these problems should be solve in initial phase before feeding dataset to ML classification algorithm. But there is no such single solution which solves all the discussed problem efficiently and concurrently. Many researchers suggested that combination or hybridization of more than one technique or process will provide proficient solution for data imbalance problem. Many of them developed hybridization techniques by combining one or more ML algorithm with some powerful approach to enhance it result but none of them developed a separate generalize hybridization tool as a solution which is applicable to enhance the performance of almost all ML techniques. Therefore, a single hybrid solution is needed with the capability of solving acknowledged problem competently, which automatically enhances performance of several ML classification algorithm while decreasing the complexity of Bigdata.

Data cleaning and data resampling must be perform for almost all the real world dataset. Current techniques studied for data resampling possess several disadvantages which are discussed above. A new single hybridization technique which can overcome those challenges and reduce the complexity of an imbalance Bigdata while cleaning it, can be a perfect solution for solving several purposes and outcome of this research. Expected outcome of this research should be a single technique capable of solving multiple issues listed below:

1. Cleaning messy Bigdata while reducing the redundancy and dropping outliers present within the dataset.
2. Reducing Bigdata imbalanced complexity with new improve approach capable to overcome challenges posed by traditional methods.
3. Capable of enhancing the performance of several ML algorithms.

2.4. Significance of Thesis

1. This research will provide an enhanced solution for managing the complexities posed by Bigdata.
2. This research will provide a single data management tool for Bigdata preprocessing.

3. This research will give enhanced version of data management solution, for improving Bigdata quality in terms of cleanliness (through redundancy and outlier removal) while reducing its complexity. Due to which several machines learning classification algorithm can work in their own way.

2.5. Thesis Statement

Traditional machine learning algorithms are not recommended for complex imbalance big dataset. Several studies and research paper presented the problems occur during classification of an imbalanced dataset. Therefore objectives to be achieved during this research are as follows:

1. To review several machine learning classification algorithms being used to handle Bigdata classification problem.
2. To identify and compare among best techniques for handling an imbalanced big dataset.
3. To implement an enhanced version of data management solution for Bigdata.

2.6 Thesis overview

- **Chapter 1** is an Introduction, of relevant and important topic used in this research.
- **Chapter 2** is a summary of diverse studies studied during the research phase. These studies are presented as a brief Literature Review.
- **Chapter 3** presents a comparative study of several classification machine learning algorithm tested on selected case study sample which was highly imbalanced. In this chapter challenges occur during mining imbalanced dataset has been observed and discussed. Thereafter presents different approaches to solve the problem caused by imbalance dataset. The experiment was conducted to check the result of popular sampling techniques and then balanced dataset is used during the experiment with several classification machine learning algorithms to construct the model.
- **Chapter 4** presents a new Hybrid Pre-processing and Resampling Technique (HPRT) for sampling and reducing the complexity of a dataset. This hybrid approach combined with several ML classification algorithms is used to construct predictive model and their result is compared during the experiment. HPRT shows positive impact on performance of ML algorithms when compared to old traditional sampling approaches.
- **Chapter 5** is a final chapter to conclude the research work. It also presents limitation, scope and future work of this study.

CHAPTER-3

Machine Learning Approaches for Complex Bigdata Processing

The chapter presents a comparative study of different classification ML algorithms to observe their performances on highly imbalanced complex dataset. Several challenges of ML algorithms occurs during mining imbalanced dataset has been discussed. Numerous popular techniques to reduce the complexity of an imbalanced dataset were reviewed along with their limitations to solve the problem. Some of them like US and OS is applied to sample imbalanced dataset and reduce its complexity. Number of experiments along with their results is produced to check the performance of US and OS to reduce dataset complexity along with it change in the behavior of ML algorithms has also been observed.

3.1 Machine Learning Impact on Bigdata

The term Bigdata, which was first coined in 1990 deals with the study of large and complex datasets. Data storage, data capturing, data analysis, data querying, data visualization and data transfer are some of the challenges of Bigdata [77] [85]. It can do wonder only if most important information can be extracted through it. In order to dig valuable information from the enormous peak of Bigdata use of predictive analytics, user behavior analytics and other Bigdata analytics are in trend [76]. Bigdata can prevent disease, detect crime, and help in business, financial services, etc. by analyzing new correlation and pattern.

ML is a branch of computer science that is used to uncover the hidden pattern from large and complex data. Machine learning is a technique through which model is trained to learn from data and hence it is widely used in almost every field in finding a valuable pattern from Bigdata [76]. This technique does not require human interruption for producing result. Modern time businesses are aware of the fact that, Bigdata is influential only if useful information is collected from it with help of an appropriate machine learning algorithm [102].

3.1.1 Issues and Challenges Related to ML Algorithms

Dataset Imbalanced classification has become popular from last few years. A dataset is said to be an imbalance when a positive and negative class is not equally distributed i.e. minority (or positive) class is negligible when compared to majority (or negative) class. There

can be many reasons for class imbalance; such as restriction during minority class data collection or very less existence of minority class because of security reason. With the evolution of Bigdata situation becomes more complicated. Bigdata added more problems, to the existing challenges of class imbalance. Volume, velocity and variety are some of the additional complexity added to this problem. Almost all real datasets are an imbalance in nature, creating difficulties for machine learning classifiers. Therefore, researchers and practitioner are taking interest in finding problems, and solutions associated with an imbalanced complex dataset.

In order to construct a precise model, the complexity of the dataset should be reduced. Redundancy and irrelevant information is another problem of ML during Bigdata mining. As data comes from a different source, lots of irrelevant and redundant data also accumulate. Removal of these types of data in a pre-processing stage can enhance the performance of a ML algorithm.

ML algorithm, when applied to the Bigdata, can extract meaningful information. ML algorithms construct rule from data provided and forecast future record. In order to build a model; data is divided into three different parts; training set, test set, and validation set. Model is initially fit on a training dataset, and then the validation set is used to compare the performance of classification parameter and decide best among them and finally, the test set is used to test the performance of the model based on certain characteristics such as accuracy, sensitivity, specificity, F-Measure, etc.

Overfitting is a term used for a situation, where model tests on same sets of training instance. As a result instead of performing calculation model repeat the label of the samples [74]. Therefore models in spite of giving well score fail to calculate accurately on an unseen data. In order to avoid this situation test set is separated and kept unseen from the rest of the dataset [75]. Cross-validation is a process which is used for evaluating or comparing the parameters used for the algorithm and selects the best parameter among them, in order to build a perfect model and thus avoid overfitting. In this study, Shufflesplit Iterator is used which can generate a number of independent train and test splits define by users. Dataset is first shuffled and then divided into train and test set. Then algorithms construct rule on training data, based on which prediction is done and evaluated using test data. Several ML algorithms used during research based on their popularity and importance are discussed below:

3.1.2 Selected Machine Learning Algorithms

ML Algorithms learn and improve automatically from past experience. The only requirement for it is a good amount of data for the classifier to learn and perform better than before with experience. During this study we will apply four selected classification algorithms (LG, KNN, DT and SVM) on an unbalanced sample dataset and compare their performances while constructing a predictive model for automatic prediction of unseen data. Performance evaluation of SVM, DT, LG, and KNN has been done through various matrices and measures.

3.1.2.1 Logistic Regression

LG [78] is simple and widely used classifier that classifies the target through returning its probabilities either as 0 or 1. It can be used for email spam detection, cancer detection, credit card fraud detection, and many such classification problems. Logistic regression is almost similar to linear regression, except linear regression predicts continuous number but logistic regression expects true and false. Logistic regression computes the correlation between the features variable and dependent variable based on probabilities output concluded logistic function. These probabilities must be transformed into either 0 or 1 by a sigmoid or S-shaped curve. It is also known as logistic function and a curve is constructed using natural logarithm of the odds of the target value rather than probability. The logistic function can be written as:

$$g(Z) = \frac{1}{1+e^{-x}} \quad \text{.....(3.1)}$$

Hypothesis representation of logistic regression can be defined as

$$h\theta(x) = g(\theta^T n) \quad \text{.....(3.2)}$$

Above function can be combined to form an equation

$$h\theta(x) = \frac{1}{1+e^{-\theta Tx}} \quad \text{.....(3.3)}$$

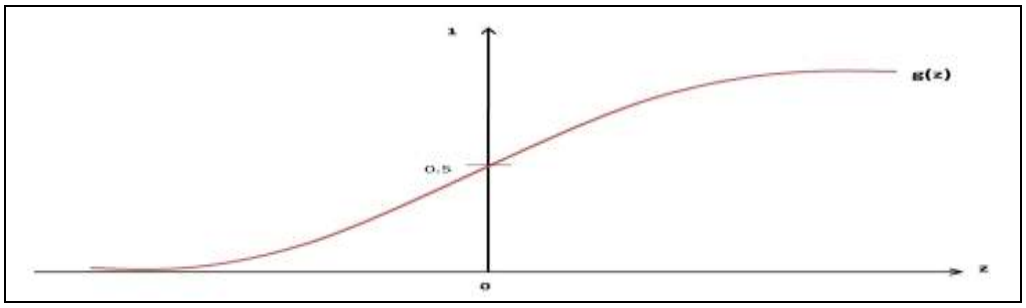


Figure 3.1: S-shaped Sigmoid Curve [78]

In figure 3.1. g(Z) calculates class output depending on input features. Value for sigmoid function ranges from 0 and 1, it predicts 0 when an input (Z) is -∞ and 1 when an input (Z) is ∞. The threshold value is kept at 0.5. It can be clearly observed in equation 3.4 and 3.5, when x is an input, passes to the sigmoid function and output is greater than 0.5 than class is equal to 1 (y=1) or output is less than 0.5 than class is equal to 0 (y = 0).

$$if\ h\theta(x) \geq 0.5\ then\ y = 1 \quad \text{.....(3.4)}$$

$$if\ h\theta(x) < 0.5\ then\ y = 0 \quad \text{.....(3.5)}$$

Optimization is needed to obtain the best fit line by using the right parameters for building models. Optimization algorithms are used to maximize the likelihood for accurate classification and known as maximum likelihood estimator. Gradient descent or gradient ascent, etc. are some of the different optimization algorithms can be used to maximize the likelihood in the case of LG.

3.1.2.2 Support Vector Machine

SVM is a supervised machine learning algorithm [79] used for individually in classifications as well as a regression problem. Support vector machine classifies target feature through constructing optimum hyperplane. Value of each feature is plotted in n-dimensional space as a particular coordinate value. Then separation hyperplane is calculated based on the coordinates which actually differentiate two classes precisely.

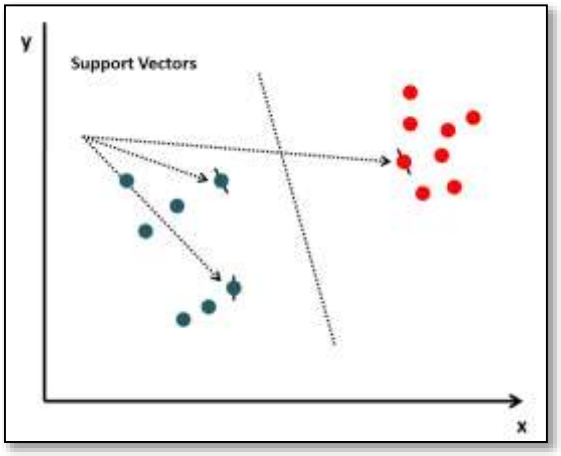


Figure 3.2: Support Vectors [79]

Support vectors are the feature separates two classes and helps in a calculating hyperplane. Many hyperplane can be constructed for target classification but best among them

should contain the widest margin [73]. Among all best hyperplane is chosen by get the most out of the distance between the support vectors of either class, which is known as margin. In real and complex dataset most of the time linear hyperplane is hard to construct then SVM uses a technique called kernel. A kernel can easily convert low dimension input space into high dimensionality space.

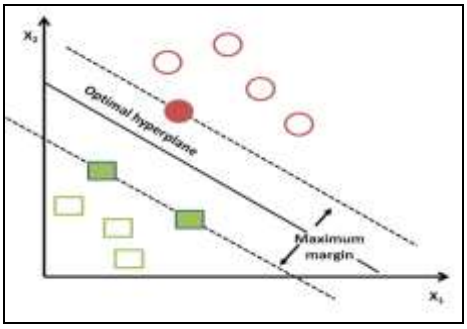


Figure 3.3: Hyperplane Constructed Through Support Vector [79]

Decision boundary is a term referred to separation line or hyperplane, helps to separate two classes. Decision boundary is calculated only for one time throughout the training process. Therefore SVM can classify a large amount of data. SVM classification is done first by calculating support vectors of the different class that are maximally separated by each other. Optimization is done by minimizing ||w|| and maximizing b where, b is a bias and ||w|| is the magnitude of a vector. Equation 3.6 is a formulae for constructing properly separable hyperplane.

$$Yi\left(X^i.W + b\right) \geq 1 \qquad \text{.....(3.6)}$$

Class (known feature w+b) = 1

3.1.2.3 K- Nearest Neighbor

K- Nearest Neighbor [80] is simplest and versatile classifier useful for both classification and regression problem. KNN takes plenty of computational space memory for all the training data. KNN classifier detects the unknown feature by calculating it with a nearest known feature in a dataset and predicts the output based on (K) number of features situated near to it.

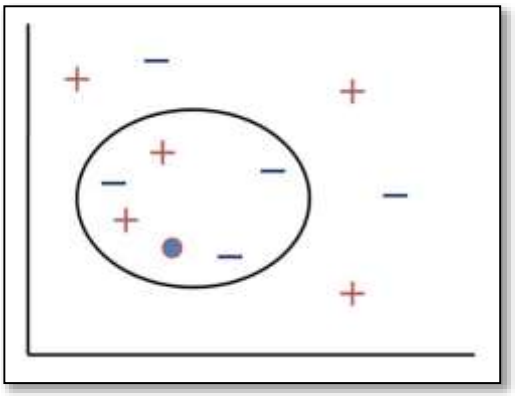


Figure 3.4: KNN with k = 5 Neighbors

In this study, the value of K is fixed to 5. It means KNN will try to find the distance of the unknown feature with five nearest neighbors. Distance is calculated between each of the row of training data and test data and then sorted in an ascending order. Among them predicted class belongs to the frequent number of features depending on k. Euclidean distance is a famous metrics which is used to calculate the distance between the unknown feature with that of K number of the nearest data sample. Euclidean squared, Chebyshev and cith block are other such matrices. Euclidean distance can be formulated as eq. 3.7:

$$Distance(P, Y) = \sqrt{\sum_{i=1}^{n} (P_i - Y_i)^2} \qquad \text{.....(3.7)}$$

3.1.2.4 Decision Tree

A decision tree can be used for solving classification as well as a regression problem [94]. Decision tree predicts the result, learning through decision-making rules constructed from the training data. Decision tree uses a tree-like structure, to accomplish the problem and flash the result. A decision tree is easily understandable when compared to other machine learning algorithm. A decision tree form tree-like structure where class labels are represented as a leaf node and attribute is represented as the internal node. In order to generate the tree-like structure from the training set, the best attribute is picked as the root of a tree. Then the training set splits into a different subset. The process of searching root and its subset is repeated at each branch of the tree to find leaf nodes in all the branches of the decision tree.

Graphical representation of a decision tree is easy to understand. A decision tree is less affected by outliers and missing values. Trees constructed by learning the decision rule is

consist of a root node, decision node, terminal node, branch, parent node, and child node. Splitting is a key process of separating nodes into one or more sub-nodes. Decision tree uses various algorithms such as the Gini index and information gain for a splitting process. Here, in an experiment, we use a decision tree with the Gini index to build the model.

3.1.2.4.1 Gini Index

Gini index is a type of metric used to measure, how randomly selected attribute as root node can identify incorrectly. It can be calculated from below mentioned formula and attribute with lower Gini index is preferable to be selected as the root node. Gini index is calculated by adding the weighted sum of squared possibilities of attributes and then subtracting it from 1. This study uses the Gini index for calculating the root node for a decision tree.

$$Gini\ Index = 1 - \sum_{x=1}^{n} P_x^2 \qquad \text{.....(3.8)}$$

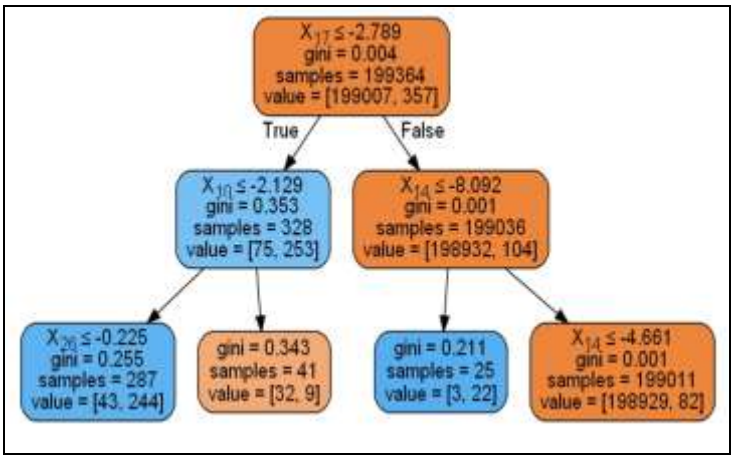


Figure 3.5: Decision Tree

3.1.2.4.2 Information Gain

Information gain is used to estimate the information contained by each attribute. Entropy is defined to measures the randomness of a random variable. In order to calculate information gain, it is required to measures entropy of each attribute. Entropy is a process to measure impurity. Information gain is calculated in two steps. In first step entropy of the target class is calculated:

$$Entropy = \sum_{x}^{class} - P\ log_2\ P_x \qquad \text{.....(3.9)}$$

In eq. 3.9, P_x is the probability of class x and entropy is calculated based on the proportion of target values**.** In the second step, information gain is calculated with the help of entropy. Information gain helps in making the decision about the best split in a training dataset and best variable to split a node. Counting of entropy of parent and child nodes is needed for calculation of information gain due to split. A variable with the highest information is selected for the split.

3.1.3 Performance Evaluation Matrices and Measures

Once an experiment had been conducted, performance evaluation is the most important task to be performed, for checking accuracies of different learning algorithms, used to construct the model. In this experiment different types of matrices and measures, to ensure the accurateness of an algorithm for prediction are, Confusion matrix, Accuracy, Classification Error, Sensitivity, Specificity, Precision, F1-score, ROC Curve and AUC. All of these matrices and measures are explained below:

3.1.3.1 Confusion Matrix

Confusion matrix [68] is a table like structure containing one value in each row and a column for the recitation of classifier performance in an experiment. If the dependent variable is consisting of two responses, confusion matrix construct - 2*2 size of the matrix for the comparison of the predicted value, with the test set value. Then the output of the evaluation is placed in respective boxes (TP, FP, TN, and FN) of the confusion matrix. In this study format of confusion matrix used are as follow:

Table 3.1: Confusion Matrix for Binary Classification

-Ve / +Ve (Predicted) / - Ve / +Ve (Actual)		Negatives	Positives
		Predicted (0)	Predicted (1)
Negatives	Actual (0)	True Negative (TN)	False Positive (FP)
Positives	Actual (1)	False Negative (FN)	True Positive (TP)

Each of the boxes in a confusion matrix has a special meaning which clearly explains, a number of times model is correct and the number of times model is incorrect during prediction. True negative (TN) is the total count of correct prediction of negative class. False

positive (FP) is a total count of the wrong prediction of positive class. False negative (FN) is a total count of the wrong prediction of negative class. It is also known as type error II. True positive (TP) is a total count of correct prediction of the positive class. Apart from a confusion matrix, there are numbers of performance measures, derived from confusion used here.

3.1.3.2 Classification Accuracy

Classification accuracy [68] is a measure of the total number of correct prediction over a total number of predictions as shown in eq. 3.10, detected by a classification model.

Table 3.2: Classification Accuracy

Test Data	Predictive Negative	Predicted Positive
Actual Negative	TN	**FP**
Actual Positive	**FN**	TP

$$Classification\ Accuracy = \frac{\boldsymbol{TN\ +\ TP}}{TP\ +\ TN\ +\ FP\ +\ FN} \quad \text{.....(3.10)}$$

3.1.3.3 Classification Error

Classification error of misclassification [68] is a measure of a total number of incorrect predictions as shown in eq. 3.11, detected by a classification model over a total number of predictions. It is also known as the misclassification rate.

Table 3.3: Classification Error

Test Data	Predictive Negative	Predicted Positive
Actual Negative	**TN**	FP
Actual Positive	FN	**TP**

$$Classification\ Error = \frac{FP + FN}{TP + TP + FP + FN} \quad \text{.....(3.11)}$$

3.1.3.4 Sensitivity

Sensitivity (SN) is an important measure derived from the confusion matrix [68] through the division of a total number of correct positive predictions over the total number of

positive as shown in eq. 3.12. True positive rate (TPR) or recall (REC) is a synonym for sensitivity. It responses the total number of times model prediction is correct for positive class.

Table 3.4: Sensitivity or Recall Rate

Test Data	Predictive Negative	Predicted Positive
Actual Negative	TN	FP
Active Positive	FN	**TP**

$$Sensitivity = \frac{TP}{TP + FN} \quad \text{.....(3.12)}$$

3.1.3.5 Specificity

Specificity (SP) is an important measure derived from a confusion matrix [68] through the division of a total number of correct negative predictions over a total number of negatives as shown in eq. 3.13. True Negative rate (TNR) is a synonym for specificity. It responses a total number of times model prediction is correct for negative class.

Table 3.5: Specificity

Test Data	Predictive Negative	Predicted Positive
Actual Negative	**TN**	FP
Active Positive	FN	TP

$$Specificity = \frac{TN}{TN + FP} \quad \text{.....(3.13)}$$

3.1.3.6 Precision

Precision (PR) is an important measure derived from the confusion matrix [68] through the division of a total number of correct positive prediction over the total number of positive predictions as shown in eq. 3.14. It is also known as Positive Prediction Value (PPV).

Table 3.6: Precision

Test Data	Predictive Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	**TP**

$$Precision = \frac{TP}{TP + FP} \quad \text{.....(3.14)}$$

3.1.3.7 F1 – Score

F1- Score is a weighted average of recall and precision as shown in eq. 3.15 [68].

$$F1 - Score = \frac{2 * Precision * Recall}{Precison + Recall} \quad \text{.....(3.15)}$$

3.1.3.8 Receiver Operatic Characteristic Curve

Receiver Operatic Characteristic Curve or ROC-curve [68] is used for visualization of performance of classifier at all possible thresholds. It is two-dimensional curves with True Positive Rate (TPR) or sensitivity on Y-Axis and False Positive Rate (FPR) or 1 – sensitivity on X-axis.

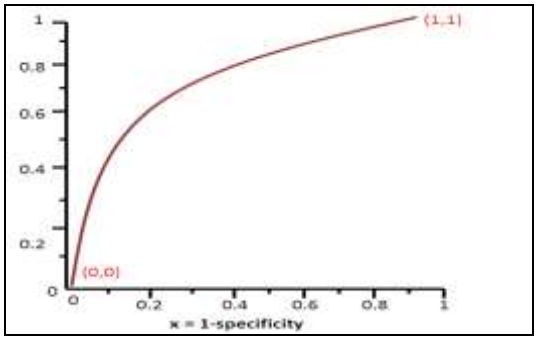


Figure 3.6: ROC Curve Showing Sensitivity (TPR) and 1- Specificity (FPR) at different Classification cut-off (Threshold).

3.1.3.9 Area Under Curve

Area Under Curve or AUC or area under the ROC curve is a measure of the complete region beneath the ROC curve [68].

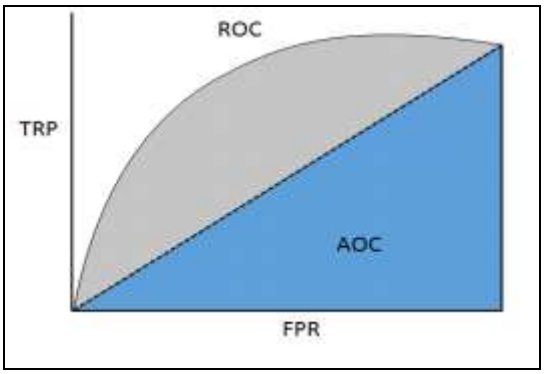


Figure: 3.7 Area Under Curve (AUC)

3.1.4 Machine Aided Detection of Credit Card Fraud

Credit card fraud is a form of financial frauds growing every year and causing losses to financial multinationals as well as government sector. Frauds can be defined as illegal actions for the purpose of making money without the consent of the proprietors. In the present day, credit card and debit card have become a frequent mode of payment. The growth of the internet is one of the prime reasons for increasing frauds. Credit card fraud is classified as inner and external card fraud. Inner card fraud occurs as a result of consent between cardholder and bank by using a false identity to commit fraud while external card fraud involves the use of stolen credit card to obtain cash through it. Traditional methods like manually detecting credit card frauds are feasible only for small datasets, but with the rise of big data, these methods are of no worth, hence has been proved as an unrealistic and time-consuming solution [64]. Therefore the finance industry is looking for help from data mining and machine learning methodologies in order to handle problems.

Global Payment Report 2015 [66] states that credit card is the highest used payment method globally in the year 2014, compared to another method. According to the released by Federal Trade Commission the overall, credit card fraud accounted for 32.7% and expected to continue increasing attempt to credit card crime. Credit card frauds are increasing from the last few years. Nilson report states that loss caused globally through credit card fraud reached $16.51 billion in 2014 and it's estimated that it will exceed $35 billion in 2020. Therefore there is an urgent need of credit card fraud detection techniques for identification of genuine and fraud transaction correctly. Researchers and scholars are putting their interest in data mining and machine learning techniques for the credit card fraud detection system.

ML and Data Mining (DM) is a field of computer science, which can dig out precious information to construct analytical models out of big data [67]. DM consists of a powerful algorithm, which uses statistical and mathematical approaches to extract productive information from large-scale data. ML techniques are nowadays successfully applied in each and every domain starting from spam detection to spam filtering, web searching, drug design, stock trading, fraud detection, and much such application as it does not pursue outdated static programming rule. In contrast, it has a powerful algorithms which can build models by learning from dataset features. Different types of algorithms are used for solving different problems. Supervised, semi-supervised, unsupervised, reinforcement, etc. are different types of majorly used machine learning techniques. In the current study, we will compare several selected supervised classification algorithm for credit card fraud detection. Performances of classifiers

are measured by using confusion, accuracy, misclassification, sensitivity, precision, f1-score metrics and ROC and AUC curve [68].

DM and ML techniques have been widely used in the detection process of fraud and non-fraud from the target class of credit card dataset through a pattern that separates two classes (fraud and legitimate). Genetic algorithm, SVM, DT, NB, LR, NN, frequent item set mining, migrating bird optimization algorithm are some of the DM techniques, which is being used in the credit card fraud detection method [65]. Machine learning approach is powerful because it does not require expert knowledge but it learns from the data itself.

3.1.5 Unauthorized use of Credit Card

Today's finance businesses are captivating their customers with abundant services such as credit card, debit card, ATM, internet and mobile banking. A credit card is plastic money which is supplied by the bank to their customers, and widely utilized for making payment anywhere at any time just like currencies. It is an alternative mode of payment for any kind of purchase we make, starting from restaurant bills, buying travel tickets, online shopping, purchasing on an e-commerce platform, buying petrol, online payment, online financial transaction, and many others. Cash can be withdrawn from ATM with the help of the card; therefore it is one of the safest modes to carry cash. With the increase of digitization, usage of credit card has increased and with it increasing credit card frauds. There are lots of benefits of using a credit card, thus making life easier by providing an alternative and convenient payment system other than cash [69]. A credit card can be used anywhere and anytime without any restriction of time and location in order to make payment. It can provide the customer with another level of security during the purchase of goods in case the original receipt is misplaced. The customer can claim their authority on a particular item by showing their credit card details. Some companies are also providing insurance on the purchase of an expensive item from their card. Credit cards are playing an important role to maintain the credit history of the customer. Its historical details are very important to judge the customer's loyalty before offering him a loan or some other financial facility.

Unauthorized use of credit card by a person or a group of a person without the knowledge of its owner can be referred to as credit card fraud. Credit cards break-in or hacking sensitive information, while making payment with the purpose of getting financial gain or causing loss to the owner of a credit card without his knowledge by some tricks are some of the credit card frauds. Demand and use of credit cards are growing with the augmentation of the internet users, thus playing an important role in today's financial world. Usage of a credit

card in day to day activities is increasing and has become a significant part of modern lifestyle. Along with all the facilities provided by credit card, an increase in the frauds during the transaction of it is also becoming a serious issue. With the rise of e-payment facility offered by the companies, a risk of frauds also increases. Billions of dollars are lost every year due to the fraud caused by general purpose cards (credit, debit etc.) thus creating a severe problem for finance industry as well as to the owner of the card. Various news, articles, and studies are pointing toward the rising and fast-growing trend of credit card fraud influencing the credit card industry.

Credit card fraud can be categorized as an internal or external [70]. False identity to commit fraud is an internal fraud and withdrawing cash by various mean from the stolen credit card are external fraud. This fraud occurs with no awareness of cardholders and violates public laws where fraudster causes loss to the card owner. This causes a huge loss of money which indirectly affects not only individual but also the society. With the expansion of modern technology, frauds rates are increasing at a higher level. Internet crime complaints center estimates; online fraud causes $3.4 billion loss in 2011 at the US alone [66]. In order to identify promptly fraud among legitimate transaction, fraud detection method is emerging rapidly.

Fraud detection tools should be smarter and powerful than ever before. Lots of criteria should be kept in mind while developing these techniques. These tools should be capable of detecting frauds from highly imbalanced datasets. Mostly credit card fraud datasets are extremely complex containing a negligible amount of fraudulent transactions in comparison to the genuine transaction. Therefore detecting frauds appropriately from these high imbalance datasets is complicated [71]. A powerful fraud detection model should be capable of addressing misclassification importance (model predicting innocent transaction as fraud is not as much dangerous as the model predicts fraud as a normal transaction). False positive and false negative rate should be very low for classifying accurate detection of fraud.

Although lots of tools and software's are made available in the market place to get rid of credit card and other financial frauds, which include retail, insurance, e-commerce, etc.). Data mining and machine learning are one of the promising fields and can be fully utilized for the detection of frauds these days [65]. Several classifications like artificial neural network, SVM, decision tree, random forest etc. have been used to solve the problem. But there are many complexities in these datasets, which creates lots of difficulties for traditional machine learning in finding the patterns, that discriminate legit and fraud transactions. Here are some of the drawbacks while finding credit card fraud are as such lack sufficient data, unavailability of credit card fraud detection datasets, due to security reason are some of the major challenges for

researchers when finding co-relation and patterns among the features of a dataset. An enhanced approach is needed for reducing the complexity of such datasets, so that machine learning algorithm can be used on it in order to build a powerful model which can precisely and automatically detect credit card frauds.

Lots of work has been done to detect fraud from the credit card dataset in order to minimize fraud rate. Credit card fraud detection is a binary classification problem consisting of two classes; negative class (legitimate) and positive class (fraudulent). Challenge is to appropriately distinguish legitimate and fraudulent class from the credit card transaction dataset. A fraudster uses an exceptionally smart way for processing payment routine of the cardholder and pretends as an authentic card owner during the fraud. This leads to a decrease in the number of true fraudulent cases and a highly skewed distribution dataset towards the negative class. With the increase in usage of a credit card as common payment mode, fraud rate is also increasing and thus traditional methods of detecting frauds are deteriorating, because of their inadequacy and time taking performance. Manually recognition of frauds from large datasets becomes quite impossible with the emergence of big data. Therefore, in order to automate this entire fraud detection process, the finance industry is shifting towards computational approaches such as machine learning and data mining.

3.1.6 Data Mining Techniques as Fraud Detection Tool

Several predictive classification data mining techniques become current trend of fraud detection. The Bayesian classification model to detect fraud in automobile insurance has been observed. In order to analyze and interpret the classifier prediction naive Bayesian visualization were used. ROC curve explains intuitive analysis of the models.

Two unsupervised algorithms, principal component analysis, and SIMPLEKMEANS algorithm was used to develop an anti-fraud project. Currently, everyone is using a Smartphone and so geographical position of the client and operations was compared to check the accuracy of the model [72]. The proposed model obtained a good result on test datasets through correctly classifying possible fraud. LG, KNN, NB on highly skewed datasets has been used and compared. A hybrid technique of under sampling and oversampling was carried out in python, by stepwise addition and subtraction of data point with interrelated existing data points unless the overfitting threshold is reached. Comparative results show KNN performs better than the other two above mentioned algorithm. Performance is evaluated based on accuracy, sensitivity, specificity, precision, Matthew's correction, and balanced classification rate.

NN [35] is also noticed as an efficient tool for fraud detection. A method to improve sparse oracle based adaptive rule extraction algorithm is presented which is used to construct an understandable rule from a neural network. Experiments have been performed on a large and real dataset. Proposed optimization uses an evolutionary search algorithm for the improvement of performance of NN classifier by optimizing a partial area under a domain specific area. Hence improvement in fraudulent detection rate has been noticed.

DT, RF, SVM and LG has been used on highly skewed credit card fraud data [65]. The performance of the above mention techniques is evaluated based on accuracy, sensitivity, and specificity, precision. Among all the techniques, logistic regression performance was best with 97.7% accuracy.

Online customers is effected mostly because of frequent fraud cases, therefore, a combination of data mining and machine learning techniques for the classification of genuine and fraudulent transactions [75]. Discussion on supervised classification using Bayesian network classifiers namely k2, Tree augmented, Naïve Bayes (TAN) and Naïve Bayes, Logistic and J48 classifies have been identified. It is clearly mentioned that algorithms perform accurately and precisely after preprocessing only.

Researchers conclude banks and financial industry are more looking for machine learning and AI tools for the prevention and detection of frauds in order to save their customer from fraud. Outlook of the model predicting credit card frauds are that, they should be capable of avoiding misclassification (i.e. fraud treated as genuine or genuine treated as fraud). Because of model misclassification, the genuine customer can be treated as fraud and find their card block while making payment. At the same time Increase in the level of customer dissatisfaction will create a huge loss to the financial industry. Therefore, the classification model should be powerful enough to classify fraud as well as non-fraud precisely. An interesting and accurate pattern drawn from the credit card processed dataset can be useful for a financial institution to detect credit card fraud and losses caused to banks and finance industry through it. The enhanced credit card fraud detection technique, data mining and machine learning approaches are considered as a clever fraud detection tool so that genuine customer do not have to pay for items which they do not purchase.

3.1.7 Dataset Description

Credit card transaction dataset is selected as a case study during the research because complexities within the dataset are relevant to the topic. The dataset used is captured from university libre de Braxelles website [103], during research collaboration of machine learning

group and world line on credit card fraud detection. It contains details of European card holder (2013),), credit card transaction for two days. It is very large dataset consisting almost 300,000 transactions out of which only 0.17% is fraud cases. The dataset is highly complex and imbalanced, containing very less fraudulent class compare to highly distribution genuine class. The aim is to detect credit card frauds such as to reduce fraud rate so that genuine customers do not have to pay for the item which is not purchased by them. Dataset is consist of numerical data having 30 features. Much information about the dataset is not obtained due to the security issues. 'Class' is the dependent variable, consisting of two features -1 (fraud) and 0 (genuine).

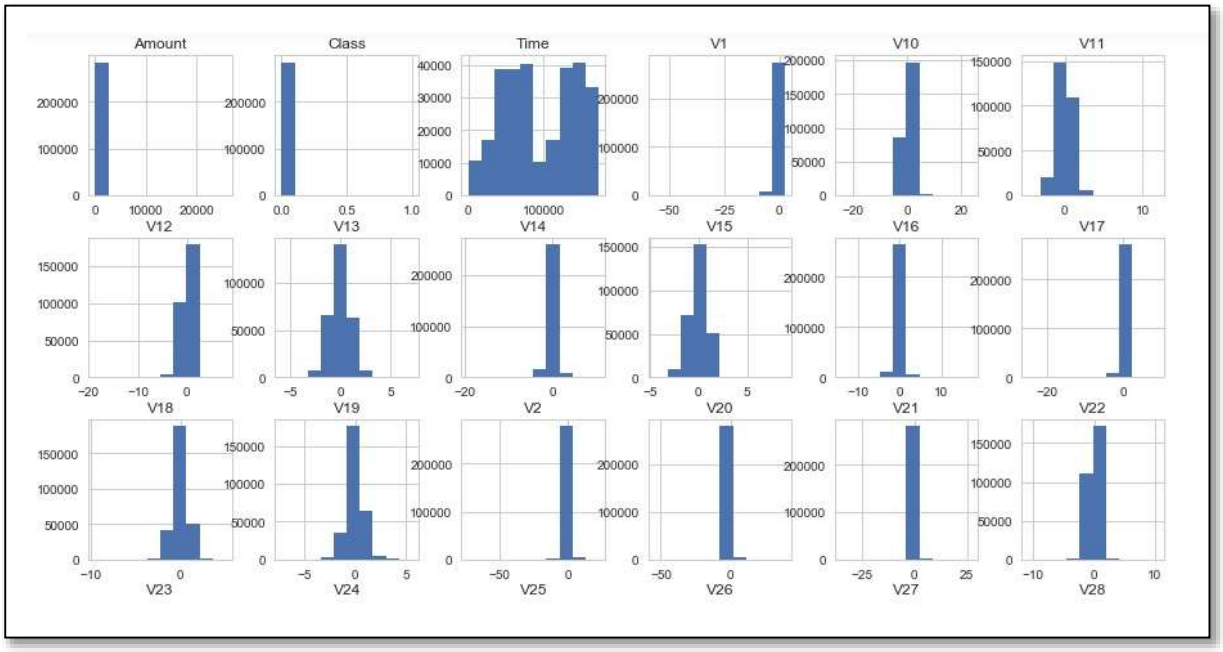


Figure 3.8: Dataset Representation

3.1.8 Experiment I Design : Imbalanced Sample Dataset

Scikit learns [19] is a freely available machine learning software consisting of a wide range of learning tools and packages. It is written in python language and used here for the classification of fraudulent and non-fraudulent transactions of European cardholders, used as sample dataset with the application of four classification ML algorithms. These algorithms are selected due to their high-performance rate yet simplicity. The dataset is highly complex because of its imbalanced feature, where out of 284807 transactions only 492 cases are frauds. The experiment is performed on the original data set including all the features. Much information about these features is not provided because of security issues.

StratifiedShuffleSplit cross-validation technique with 10-splits is used for validation and circumvent overfitting of a predictive model. It divides original dataset k-times into train

and test splits after shuffling of data. Here, original dataset is partitioned into a training set and test set where the training set helps training the data and test set is used for validation purpose. Cross-validation technique is repeated for k-times where each of randomly selected subsamples used for validation. The result generated for k-times are combined together to measure average to produce a result with the help of training and test set depending upon mentioned size. It is just like randomly selecting features from the dataset for validation. Thus all the features in the data having a fair chance to behave both as a validation set for testing and training set for the training of the model, where each observation behaves as a test for only once. StratifiedShuffleSplit returns a sample and stratified folds which is very important for imbalance dataset because stratification helps in distributing the target class sample in correct proportions. GridSearchCV algorithm is used to discover the best parameters for classifiers to acquire a correct predictive score. A number of matrices such as confusion matrix, accuracy, misclassification, sensitivity, specificity, recall, F1-score etc. are used for the analysis of classifiers performance.

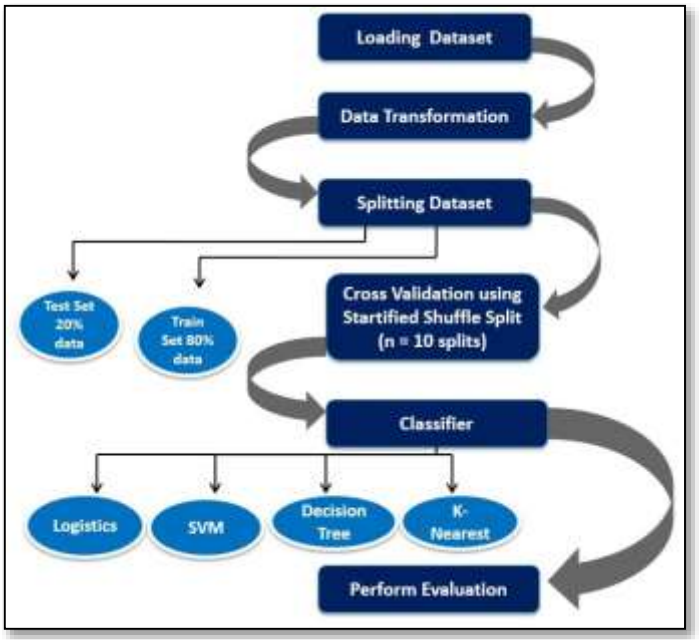


Figure 3.9: Workflow during Experiment I

Original dataset divided into training and test set sample where total number of test sample taken 85443 out of which only 135 is detected fraud and rest 85308 as genuine.

Table 3.7: Results of ML Model during Experiment I

Confusion Matrices	LR	DT	SVM	KNN
TP	78	98	97	4
TN	85293	85281	85283	85308
FP	15	27	28	0
FN	57	37	35	131

Metrics	LR	DT	SVM	KNN
Accuracy	99(%)	99(%)	99(%)	100(%)
Recall	58(%)	73(%)	74(%)	3(%)
Specificity	99(%)	99(%)	99(%)	100(%)
Precision	84(%)	78(%)	79(%)	100(%)
F1 – Score	68(%)	35(%)	34(%)	.06(%)
Misclassification	.08(%)	.07(%)	.06(%)	.05(%)

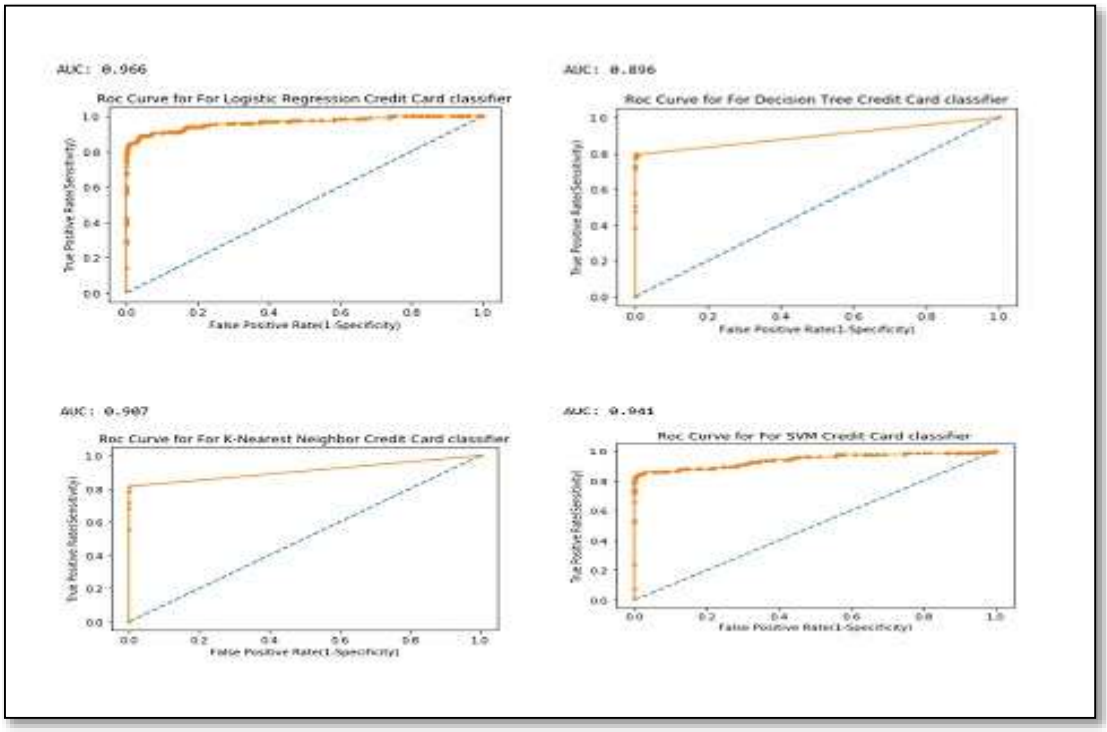


Figure 3.10: ROC Curve and AUC Score of ML Models during Experiment I

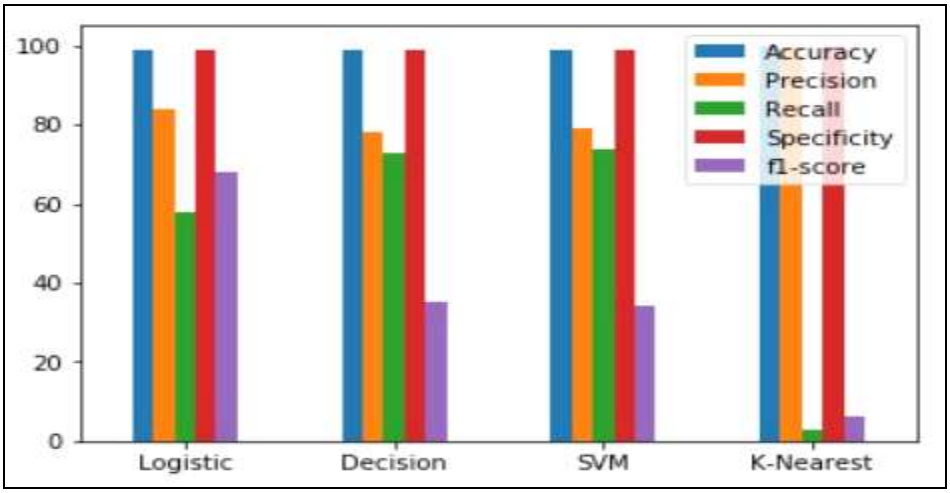


Figure 3.11: Performance Comparison of ML Models during Experiment I

3.1.9 Discussion

This experiment is carried on an original dataset consisting near about 3, 00,000 transactions. The result shows high accuracy for all the four classifiers. But if we scrutinize the confusion matrix, it can be observed that all the models prediction are highly biased. Dataset having few fraudulent (0.17% only) compares to genuine transaction creates confusion for machine learning algorithms and therefore these classifiers built biased models, treating the minor class as noise and building rules only for majority classes. Due to the reason, most fraud transactions are predicted as non-fraud. The confusion matrix is the best evaluation matrix for a model built by the imbalanced dataset. Test data consist of 85443 transactions where 85308 is non-frauds and only 135 is fraud. In confusion matrix it is clearly seen among 135 frauds, LG predicted correct result for 78 times, KNN for 4 times, SVM for 97 times and DT for 98 times while prediction of fraud i.e. in spite of the high accuracy of the classifiers, models are not able to predict fraudulent accurately. The result shows logistic regression having accuracy (99%), Recall (58%), Specificity (99%), Precision (84%), F1-Score (68%) perform better than other three classifiers. KNN performs worst with very low Recall (3%) and F1-Score (6%). Decision tree and SVM performance are moderately producing almost the same results. ROC and AUC curves are high for all classifiers but for such an imbalance dataset high ROC and AUC curve must not be considered for evaluation purpose. Experiment result possesses high accuracy, ROC, specificity, and precision for all classifiers and AUC curves but with low recall and F1-Score values i.e. models are assuming the transaction as the negative or non-fraud class. But models should calculate the pattern of the class and then construct a rule for both the class, instead of assuming the prediction. In this case, optimization for F1-Score and recall rate is

required. The accuracy and ROC curves of highly imbalance dataset are misleading with very high percentage value; therefore we trust confusion matrix, recall, F1- Score for the performance evaluation of classifiers.

Dataset presented in the study is highly imbalanced with very less number of features in minority class. Due to the complexity of the dataset traditional machine learning algorithms fails to produce accurate results i.e. low specificity, recall rate, and F1-score with high misclassification rate. Thus produce costly models where most of the times fraud transactions are considered as genuine transactions. Therefore optimization for recall rate is needed because false negatives (frauds that are not detected by the model) costly and should be low for high performing models. In order to achieve this objective complexity of the dataset should reduce first by decreasing imbalance property of the dataset. Therefore in the next section different contemporary approaches for reducing the complexity of an imbalanced dataset is presented.

3.2 Imbalanced Data Classification: An Introduction

Dataset Imbalance classification has become popular since last few years. A dataset is said to be imbalanced when a positive and negative class is not equally distributed i.e. positive class is very minor when equated to negative class [81]. There can be numerous reasons for class imbalance; such as limit during minority class data collection or very less existence of minority class because of security reason. With the expansion of Bigdata, the situation becomes more complicated. Big data added surplus problems, to the existing challenges of class imbalance. Volume, velocity, variety, veracity, and value are some of the additional complexity added to this problem. Almost all real datasets are imbalance in nature thus increasing difficulties for machine learning classifiers [82]. Therefore, researchers and general practitioner are captivating attention towards researching, problems, and solutions associated with an imbalanced dataset.

Traditional machine learning algorithm cannot perform well in terms of finding patterns for classification or prediction of imbalanced class. In such cases because of mystification, minority classes are ignored and only majority class is taken care of, hence produces weak models with high accuracy and lot of misclassification for a given dataset [83]. This type of classifications tends to give a very high number of false negatives, which is very costly for some problem like credit card fraud detection, medical diagnose etc. Considering a case of credit card fraud detection where false negative is riskier than false positive (i.e. false negative means frauds detected as genuine). These kinds of models are biased towards the majority class during the entire learning process and therefore minority class is weakly modeled. Due to which

in spite of high accuracy, a model could not distinguish positive and negative class accurately. Therefore biased and inaccurate models are constructed with such traditional machine learning algorithm. Lots of techniques have been proposed to deal with the class imbalance problem. Several algorithms modification had been seen in order to reduce the problem of imbalance dataset. Sampling is one of the popular techniques of balancing the class of a dataset. Sampling is majorly categorized into oversampling and under sampling [84].

The dataset used for current experiment is highly imbalanced, consisting of only 0.17% of a minority class. The main objective of the current study is to balance the dataset with the help of under sampling and oversampling methods for reducing the complexity of a dataset. Hence performance of machine learning algorithms improved when applied to balance dataset, mounting a powerful predictive fraud detection model. Experiments had been conducted to find the best predictive model for credit card fraud dataset. The result of both sampling method had been compared. LG, SVM, DT and KNN classifier are used to build a classification model for a balanced credit card transaction dataset. The model evaluation shows improvement in the performance of these classifiers on the balanced dataset.

3.2.1 Issues Related To Data Imbalance

With the emergence of big data, sizes of records have also increased. At present because of Bigdata real-time dataset become more disordered and inconsistent, thus forming imbalance dataset. A large volume of data creates an additional complication to the machine learning algorithms [82]. Lots of research determines that data together with, large volume and high imbalance poses challenges for machine learning predictors. Implementation of more than one algorithm should be combined to form good models for such kind of big and highly imbalance dataset [83].

A huge volume of data is accumulating almost in every sector because of escalation use of the internet. These data are much recommended as it provides a lot of information which can help in decision making [86]. Proper analysis of such a huge volume of a dataset can predict unknown facts based on past data and this machine learning approach is known as supervised learning. In this type of classification, models are trained based on target classes and can predict future data. Availability of good quality of data in sufficient amount is the only requirement of these ML algorithms. Data insufficiency and huge volume cause an imbalance dataset thus creating the learning process more difficult. Almost every domain data [87], for example, fraud detection, weather forecast, network intrusion detection are affected with this problem, hence the solution is needed for proper analytics.

The method of OS and US are very popular [81] for reduction of imbalanced dataset complexity. Lots of hybrid method based on oversampling and under sampling presented to balance a dataset. Solution based on the support vector machine had been proposed and this classifier is sensitive to cost and rough set based on minority class rule oriented.

Types of matrices and measures chosen for the performance evaluation of imbalanced data are very important. Accuracy matrix for an imbalanced dataset is not considered as good performance evaluator of such kind of datasets [84]. Most of the real datasets are imbalanced such examples can be found in various domain finance, Medical, network intrusion, telecommunication, natural disaster, etc. are some of them. Lots of challenges occur during the processing of imbalanced learning [90], some of which are listed below:

1. When the dataset is dealing with imbalance problem generalized matrix such as accuracy cannot evaluate the performance of classifier properly because in such cases costs of different errors are diverse. Different matrices such as confusion metrics, F-Score, etc. can be used for better performance.
2. Another problem associated with class imbalance is the unavailability of informative training data for machine learning. Very less number of features related to minority class possesses challenges for classifier during the process of learning from such dataset and hence produces the results with weak learning models where minority features are misclassified.
3. Most of traditional machine learning algorithm uses greedy search and divide and conquer rule during the learning process. But these rules perform inadequately for a minor class of imbalance dataset. Therefore further researchers are looking up this matter very seriously for development of an enhanced version of traditional machine learning algorithm which can deal with the imbalanced type of dataset.

3.2.2 Several Techniques for Dealing an Imbalanced Bigdata and Limitation

There are many different methods for dealing with problem occurred during the learning phase of the imbalanced dataset. Sampling techniques, kernel-based techniques; cost sensitive techniques etc. are some of the approaches used for this purpose. Among all of these techniques, sampling methods are most general methods used for learning of imbalance type of dataset al.though it has many faults.

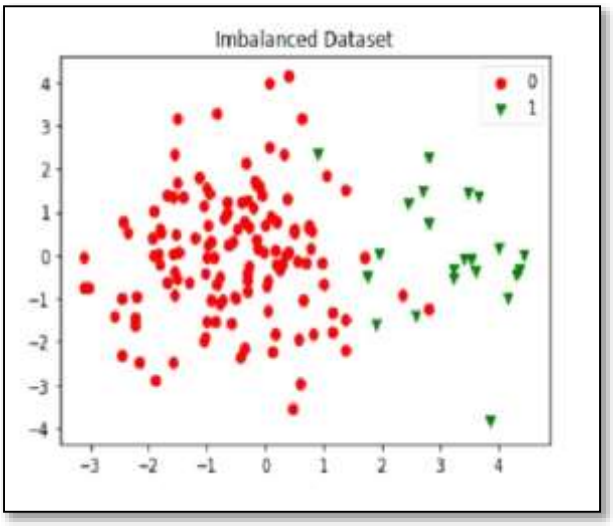


Figure 3.12: Imbalanced Dataset

Sampling is the most common approach when dealing with class imbalance. The process of sampling is applied to the dataset in order to convert it as a balanced class distribution. Sampling is further classified into under sampling and oversampling. US is the method, where a feature of majority class is reduced and oversampling improves new artificial features to the minority class [15].

OS is a process of adding artificial features to minority either randomly or by some calculation. Newly added samples help to balance of the dataset but can increase the risk of overfitting. Random OS is a very familiar way of balancing the dataset. In this technique, features are added randomly by selection of the set of examples from minority class. Although, this approach increases features of a minority class and hence balance the dataset the extra headache of overfitting and an increase in the time of training data are some of the drawbacks of random oversampling.

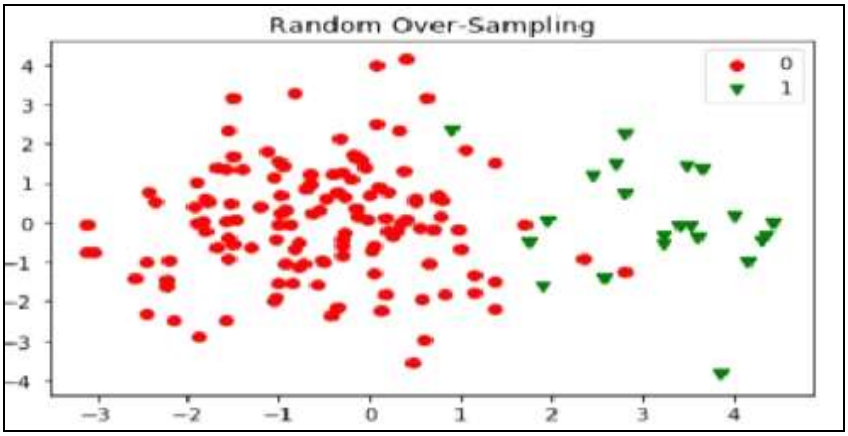


Figure 3.13: Random Over-Sampling

SMOTE is a type of oversampling for increasing the features of minority class [42] by adding new artificial features similar to it. KNN algorithm is used by smote for adding new features to the minority class, by choosing its neighbor randomly depending on the amount of oversampling needed. This is proved to be a better sampling approach because here data points similar to the original features are added to a minority class. Many other algorithms derived from the SMOTE, but it gives poor results with high dimensional data as the calculation time of SMOTE increases thus decreasing its performance.

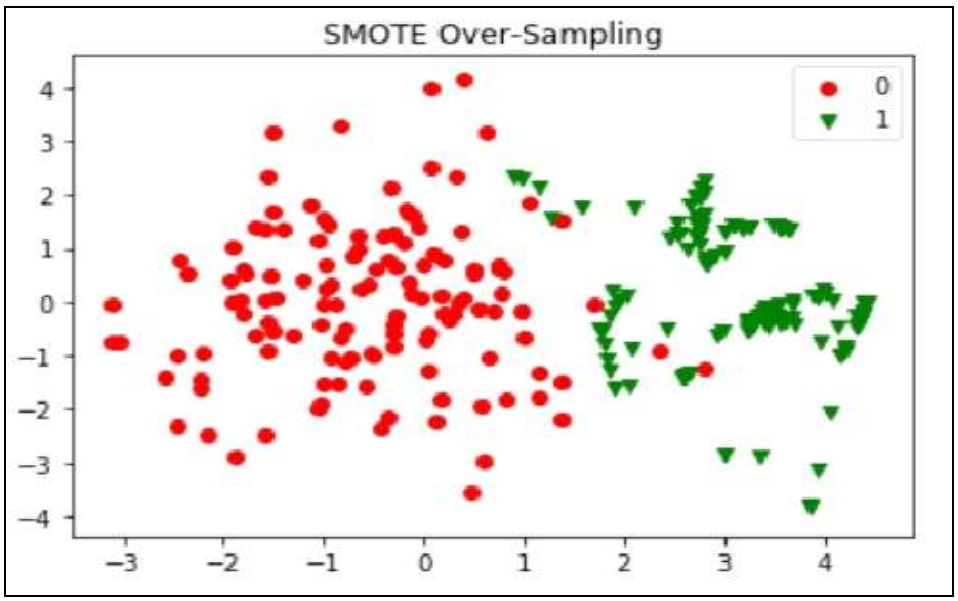


Figure 3.14: SMOTE Over-Sampling

Borderline SMOTE [93] is an oversampling algorithm inspired by SMOTE. It has two versions, borderline SMOTE1, and borderline SMOTE2. These methods add features only near the borderline using KNN. Borderline SMOTE2 is advance than Borderline SMOTE1, as it selects both positive and negative nearest neighbors.

Adaptive Synthetic (ADASYS) is also derived from SMOTE, using the weighted distribution of minority class. Here data points are added using KNN method depending upon majority NN. It cannot deal with outlier well and thus not perform well with noisy data.

Random US is simplest among all the methods of reducing the majority class randomly such that minority class becomes equal or near to equal with majority class. But major drawback with this type of sampling is a huge information loss caused by deleting lots of features from the dataset. The features discarded from the dataset might contain useful information for building pattern, the absence of which can generate week models.

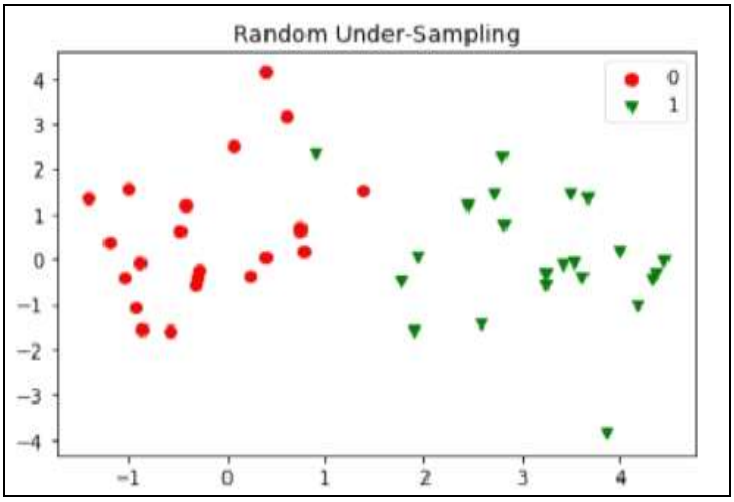
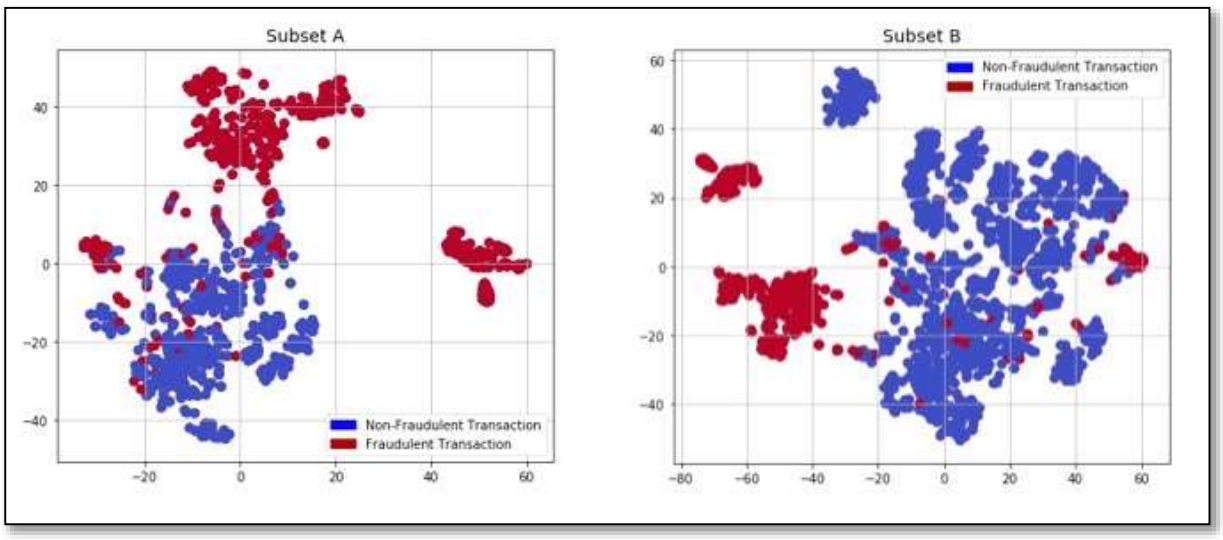


Figure 3.15: Random Under-Sampling

3.2.3 Comparative Analysis of Under-Sampling and Over-Sampling

The current section is dedicated to compare US and OS techniques on a highly imbalanced sample dataset. The experiments were carried out to balance the sample dataset and then built the model on top of it applying ML algorithms. Both the experiment uses credit card transaction of European card holder (2013) as a sample data.

During Random US experiment original credit card transaction dataset is converted into three subset- subset A (50% minority, 50% majority), subset B (75% majority, 25% minority) and subset C (80% majority, 20% minority). The reason of segregation of dataset into different ratios of minority and majority class is to check, how performance of the model are affected with the changes in distribution of minority and majority class of the dataset.



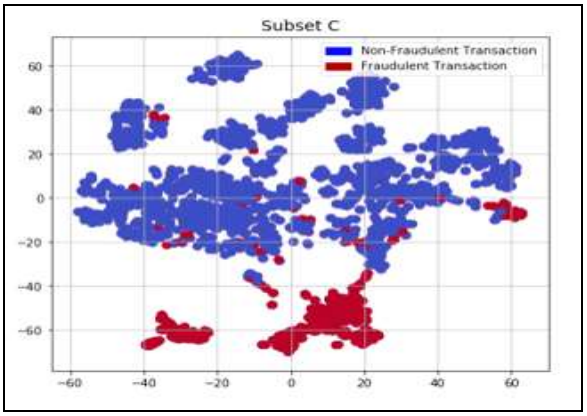


Figure 3.16 Application of Random Under-Sampling Subset A, Subset B and Subset C

During second experiment SMOTE oversampling is used to balance the dataset. In this process features in minority class are increased by adding artificial sample to the minority class with the help of the KNN algorithm and Euclidean Distance. During the process difference between minority feature vector and its nearest neighbor is evaluated and then the output is multiplied by any random number between 0 and 1, to create a new synthetic feature. The synthetic data point created is added randomly at any point on the same calculated line segment. SMOTE oversampling is performed during cross-validation to avoid overfitting because, SMOTE if applied earlier there can be the possibility of adding features exactly the same in the validation set, hence causing the problem of data leakage. To avoid this, validation set must be excluded first with other training set and then during cross-validation oversampling should be applied on rest of the dataset. This will avoid overfitting as well as data leakage problem because during testing phase validation set will be completely unseen.

StartifiedShuffle split with n = 10 (number of the split) is used in both the experiment. After re-sampling of data, binary classification algorithm, such as logistic regression, K-nearest neighbor, SVM, and decision tree had been applied on the balanced dataset, so as to construct credit card fraud detection models. The models are evaluated with the help of a confusion matrix, sensitivity, specificity, precision, accuracy, and misclassification. Confusion matrix gives a total count of true negative (TN), true positive (TP), false negative (FN) and false positive (FP). TN is a correct count for negative class, TP is a correct count of positive class, FN is a wrong prediction for negative class and FP is a wrongly classified positive class.

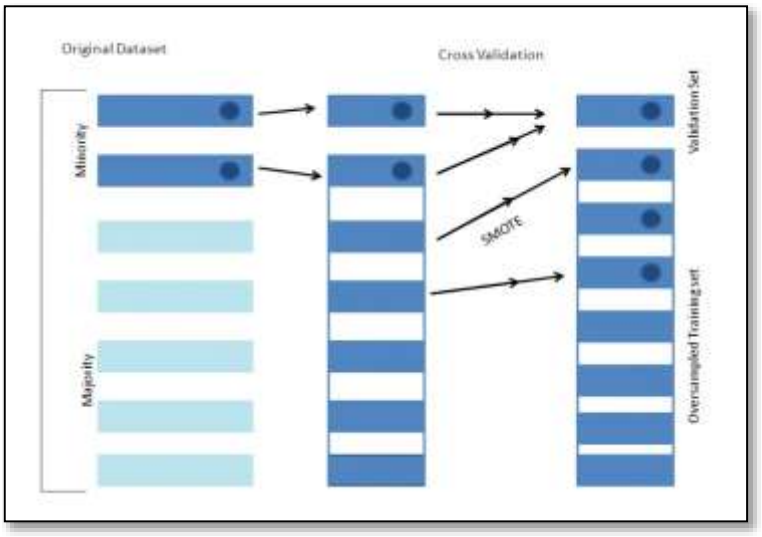


Figure 3.17: Application of SMOTE Over-Sampling Process during Cross-Validation

LG predicts the result based on probabilities and sigmoid curve value. The value of sigmoid curve ranges from 0 and 1. Probabilities calculated from the input feature during logistic regression process passes through the sigmoid curve. If the output is greater than or equal to 0.5, then the class is predicted as 1 (positive class) and if the output is less than 1 then the class is predicted as 0 (negative class).

SVM calculates the result based on an optimal hyperplane. The hyperplane is the line between two classes separating one from another. It is calculated depending on coordinates which differentiates two classes. Support Vector is featured helps to detect hyperplane. Many hyperplanes are constructed during the process of SVM out of which best is chosen by maximizing the distance of the support vector with that of hyperplane of both the class.

KNN predicts the result after calculating the distance of unknown features with that of given features distance. Euclidean distance is used for distance calculation here. The result is predicted based on a number of K-nearest neighbors which is provided by the user. In this study parameter for k is 5 i.e., the algorithm will consider five nearest data point with that of the unidentified feature, and then output class for that feature will calculate by counting class of the data points appears for a maximum number of time.

DT builds models, by defining rules for training examples. During this process, trees are constructed to solve the classification problem. Generation of the tree starts with the selection of the root node from the training dataset and then the process continues to find branches, sub-branches etc. The process of searching roots and its subset is repeated for each branch of the tree. Gini index is a matrix used in this experiment for selection of roots during

tree construction. It identifies root node by measuring how often attribute chosen randomly can be incorrect. In the next section experiments conducted during study to check the performance of above mentioned ML algorithms after application of random US and SMOTH oversampling on an highly imbalance sample dataset is discussed along with its result.

3.2.3.1 Experiment II Design: Using Random Under-Sampling

During this experiment, we have considered credit card transaction dataset as a sample data downloaded from University Libre De Bruxell's. The intention is built a precise model for classification of frauds and non-fraud transaction. But the dataset is highly skewed towards the majority (negative or non-fraud) class, with a very low percentage (0.17%) of minority (positive or fraud) class. Challenge here is to balance the dataset before application of machine learning algorithms on it. US and OS are the popular methods for resolving imbalance problem. Random under sampling is useful for randomly reducing the majority class from a dataset, such that the distribution of two (minority and majority) classes becomes normal or close to normal. With this idea, the dataset is divided into three different subsets. Subset A ratio (50:50) of majority & minority followed by subset B with ratio (75:25) of majority & minority, and subset C ratio (80:20) of majority & minority. The figure 3.18 illustrates the process of random US clearly. Logistic regression, SVM, KNN, and DT algorithms applied to all the subset individually. Confusion matrix, precision, recall, F1-Score, Misclassification, false positive rate, AUC is the matrices and measures used for the evaluation of the performance of classifiers with different subset. Several steps incorporated during the experiment is as follow.

Step I: Dataset (DS) is divided to into two subsets, one contains minority class (fraud) and other majority class (non-fraud) and random US is applied to majority (non-fraud) class.

Step 2: Dataset is again concatenated and divided to form dataset X and dataset Y, where dataset X is consist of entire dataset except the target variable and Y is consist of only target variable.

Step 3: Stratifiedshufflesplit is applied to X and Y after division of training and test data- X_{Train}, Y_{Train}, X_{Test}, Y_{Test}.

Step 4: GridSearchCV is applied to find the best parameter and model is constructed using ML classifier.

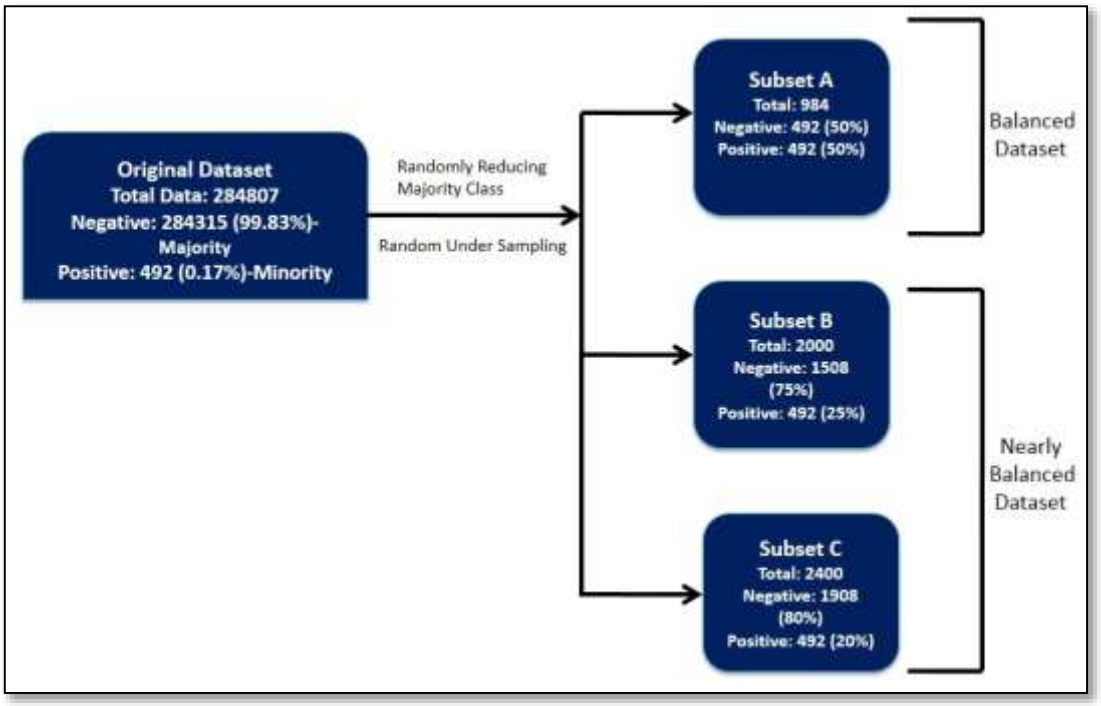


Figure 3.18: Experiment II using Random Under-Sampling

Table 3.8: Subset A (50%: 50%) results of ML Model during Experiment II

Matrices & Measures	L R	KNN	SVM	DT
Accuracy	0.94%	0.91%	0.93%	0.93%
Precision	0.98%	0.96%	0.98%	0.92%
Recall	0.91%	0.88%	0.90%	0.93%
F1 – Score	0.94%	0.92%	0.94%	0.92%
Misclassification	0.06%	0.08%	0.07%	0.07%
False Positive Rate	0.02%	0.04%	0.02%	0.05%
AUC	0.97%	0.95%	0.98%	0.91%
Confusion Metrics				
Total Test Samples 197 / Predicted Negative / Predicted Positive				
Actual Negative: TN / FP	85 / 2	83 / 4	85 / 2	78 / 9
Actual Positive: FN / TP	10 / 100	13 / 97	11 / 97	8 / 102

Table 3.9: Subset B (75%: 25%) results of ML Model during Experiment II

Matrices & Measures	L R	KNN	SVM	DT
Accuracy	0.94%	0.94%	0.94%	0.93%
Precision	0.97%	0.97%	0.95%	0.89%
Recall	0.90%	0.76%	0.79%	0.79%
F1 – Score	0.86%	0.86%	0.86%	0.84%
Misclassification	0.05%	0.05%	0.05%	0.06%
False Positive Rate	0.01%	0.01%	0.01%	0.02%
AUC	0.99%	0.96%	0.98%	0.90%
Confusion Metrics				
Total Test Samples 400 / Predicted Negative / Predicted Positive	293 3 9 95	292 4 11 93	292 4 10 94	280 16 15 89
Actual Negative / TN / FP				
Actual Positive / FN / TP				

Table 3.10: Subset C (80%: 20%) results of ML Model during Experiment II

Matrices & Measures	L R	KNN	SVM	DT
Accuracy	0.94%	0.94%	0.94%	0.93%
Precision	0.97%	0.97%	0.95%	0.89%
Recall	0.90%	0.76%	0.79%	0.79%
F1 – Score	0.86%	0.86%	0.86%	0.84%
Misclassification	0.05%	0.05%	0.05%	0.06%
False Positive Rate	0.008%	0.005%	0.01%	0.02%
AUC	0.84%	0.86%	0.89%	0.83%
Confusion Metrics				
Total Test Sample 480 / Predicted Negative / Predicted Positive	376 3 22 79	377 2 24 77	375 4 21 80	369 10 21 80
Actual Negative / TN / FP				
Actual Positive / FN / TP				

3.2.3.2 Discussion

During the experiment, the highly imbalanced dataset is alienated into three different subsets keeping the data sample of minority class unchanged. i.e. 492 majority class of dataset compact randomly by application of random under sampling. The motive was to reduce the complexity of the dataset to an extent where traditional machine learning algorithm can perform well.

In subset A Logistic regression performance was outstanding with high accuracy (94%), precision (0.98%), recall (92%), F1-Score (94%), AUC (97%) with very less misclassification (0.6%) and False Positive Rate (0.02%), confusion matrix shows that out of 197 test samples, logistic regression misclassified only 12 times. In subset B performance of logistic regression are better than other algorithms, where logistic regression misclassification rate was even smaller than subset A. According to the confusion matrix, out of 400 test samples, it was only 12 times when logistic regression was not correct. In subset C although accuracies and other measures show very high percentages the rate of misclassification also increases compares to other two subsets.

Comparison between subset A, B and C concludes subset B performs better than the other two. So subset B, logistic regression can be concluded as the best model with the accuracy (97%), precision (97%), recall (91%), F1-Score (94%), AUC (99%), misclassification (0.03%) and False Positive Rate (0.01%).It can be observed that the misclassification happens more in subset C than A and B. Therefore it can be concluded that more the dataset will be skewed; more the result will be invalid or inaccurate. But before finalizing best model for credit card fraud detection, we have to look into the negative aspects of random under sampling. In the experiment, it had been seen that large dataset of almost 3 billion data, had been reduced drastically to a very negligible amount (1% or 2%) of the entire dataset. This means 99% volume of data is discarded. It can happen that all useful information for making rules through classifiers had been flushed off. Therefore, results of random under sampling can be biased, and the model performing great with test data may produce inaccurate prediction with unseen real-time data. Due to this reason, we conducted the next experiment to check the SMOTE oversampling method for balancing the dataset with the application.

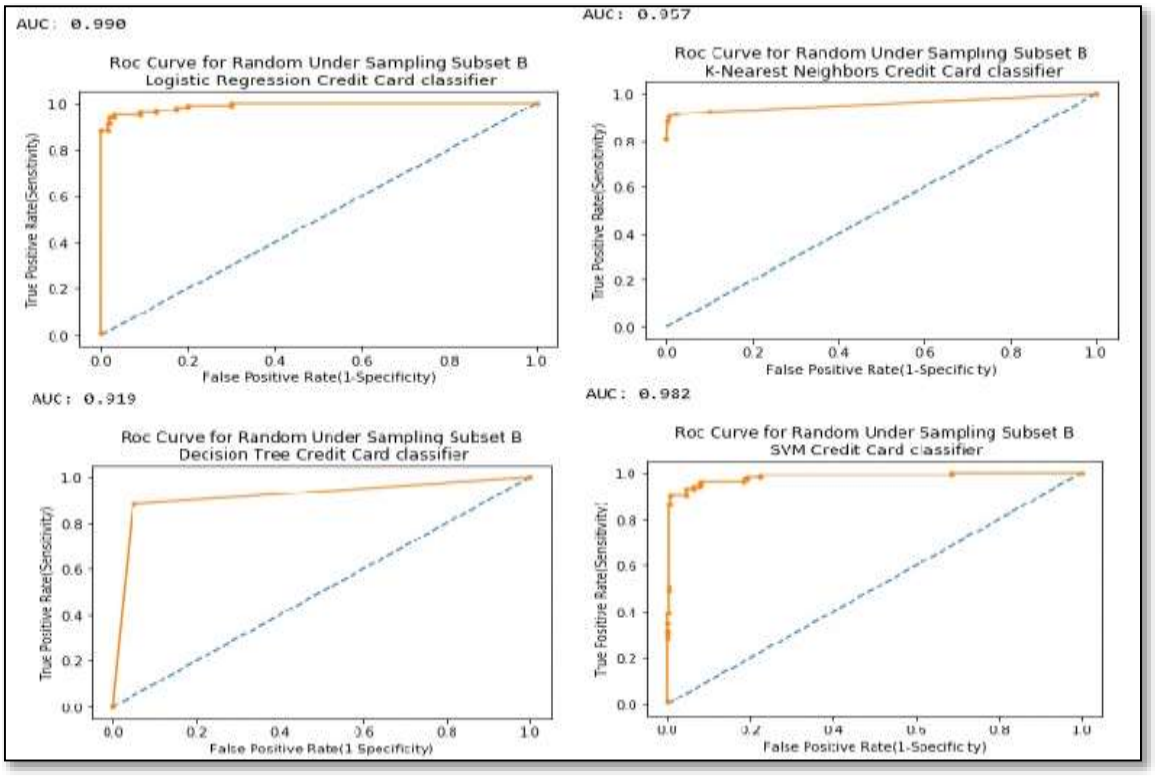


Figure 3.19: ROC Curve and AUC Score of ML Models (Subset B)

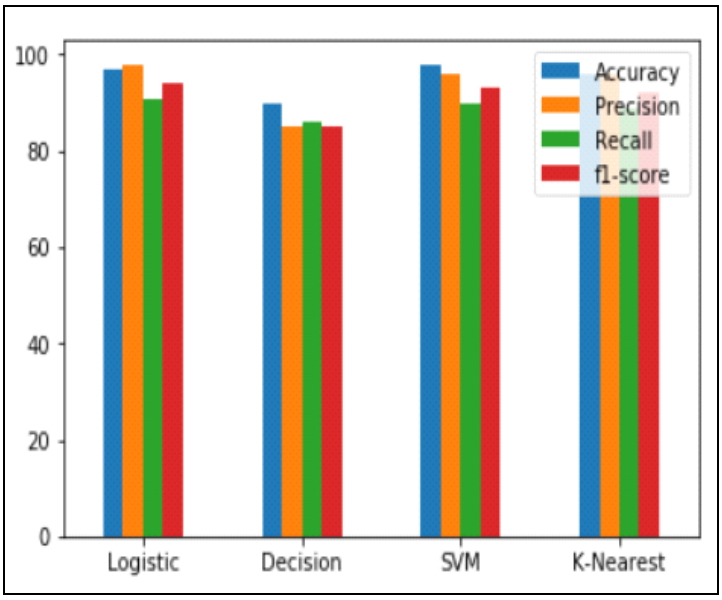


Figure 3.20: Performance Comparison of ML Models (Subset B)

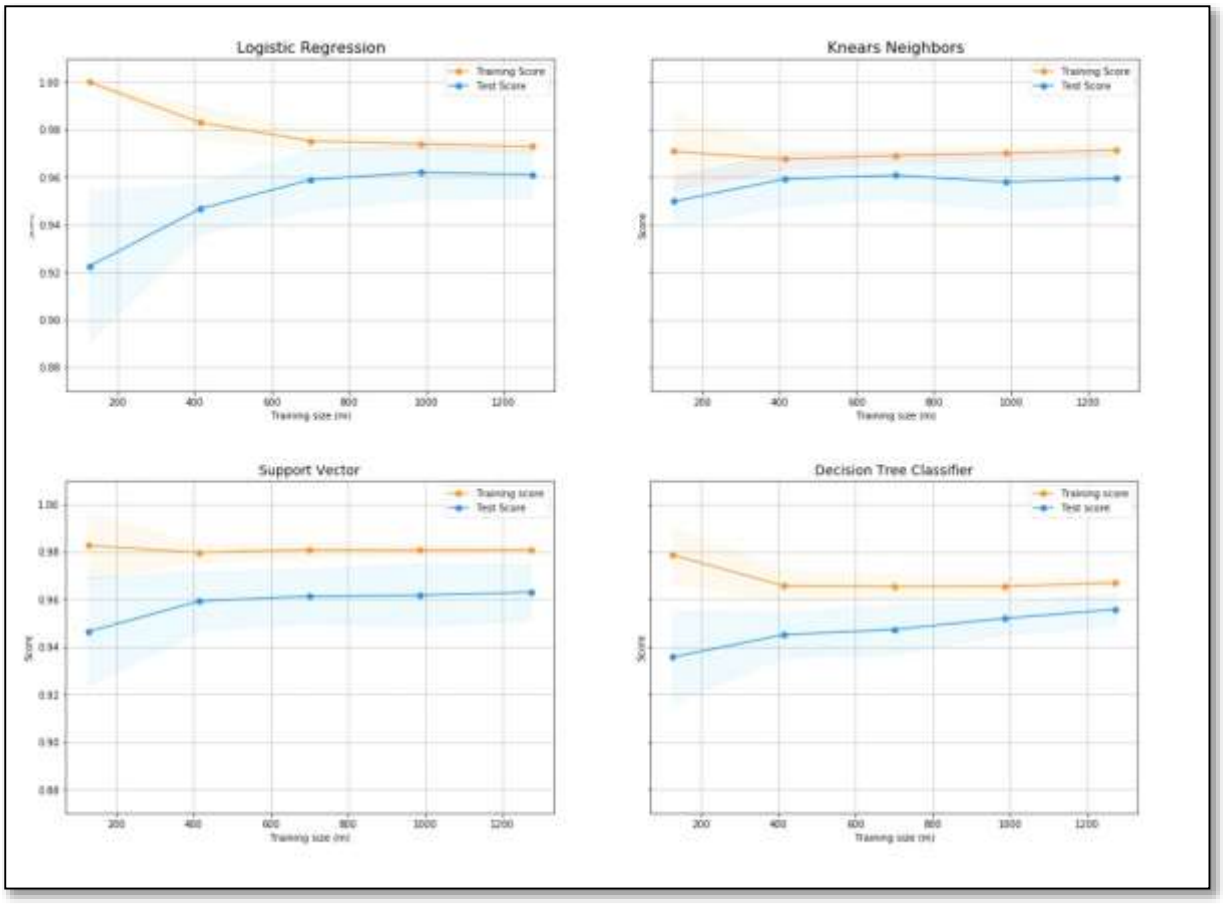


Figure 3.21: Learning Curve of ML Models (Subset B)

3.2.3.3 Experiment III Design: Using SMOTH Over-Sampling

In experiment III, SMOTE is applied to a dataset for increasing minority class features synthetically. As discussed earlier, SMOTE is an oversampling method, use to balance the dataset through the addition of minority class features by means of KNN algorithm. The experiment is performed on the original dataset (284,807), containing 99.83% majority and 0.17% minority class. Train set (227,845) and test set (56962) had been constructed by splitting original dataset. SMOTE is applied on minority class during cross-validation to avoid overfitting. After performing SMOTE three different machine algorithms are applied on the normally distributed dataset to construct the model for binary classification. Logistic regression, KNN and decision tree are used in this experiment for the formation of a model. Prediction of the model is compared with the original test data to attain accuracy. Model evaluation is performed through different matrices and measures.

Dataset is provided as an Input: C {V1, V2, V3,….,V_n} to get an output as a Model constructed for a SMOTE generated training set through learning rules from machine learning classifier. Algorithm used during the experiment is illustrated in several steps:

Step 1: Split C into X_1, Y_1 such that X_1 is consist of an entire set of data excluding target class and Y_1 is consist of target class only.

X_1 = Entire C features excluding target class

Y_1 = Target class of dataset C

Step 2: Generate train and test set splits randomly using StarifiedShuffleSplit with a number of split = 10.

S = StarifiedShuffle (Split = 10, test-size = 0.2)

For train-index, test-index in S (X_1, Y_1)

X_{Train}, Y_{Train} = X_1 [train-index], X_1 [test-Index]

X_{Test}, Y_{Test} = Y_1 [train-index], Y_1 [test - Index]

X_{Train}, Y_{Train} is 80% of data from the dataset used for training a model

X_{Test}, Y_{Test} is 20% of data from the dataset for a testing model prediction after training. Above set of the algorithm will create 10 sets of X_{Train}, Y_{Train}, X_{Test}, Y_{Test} for validation purpose wherein each iteration, from 10 sets 1 will act as the test set and other nine will remain as the training set.

Step 3: Using GridSearchCV for searching the best parameters for the classifiers.

Step 4: Using SMOTE during cross-validation for adding synthetic features for the minority class

For train in split X_{Train}, Y_{Train}

P = SMOTE (X_{Train}, Y_{Train})

M = P.Classifier (X_{Train},[train] , Y_{Train}[train])

Where M is a model constructed using P (contains artificial data point for majority class) with the help of ML Classifier.

Step 5: Model Prediction for credit card fraudulent, through SMOTE generated training set and original test set.

Algorithm mentioned above takes the original dataset as input. Original dataset splits in the first step, the intention was to separate the target data from the entire dataset. 10 fold stratified shuffle split perform for the validation of training and test set with 10 splits. In order to construct the model and find the best-suited parameter for a given classifier GridSearchCV is used. SMOTE is applied for adding synthetic features to a minority class of training set to balance the dataset. Here, SMOTE process is applied during the cross-validation and not before it. The idea was to avoid the problem of "data leakage" i.e. to keep test data unseen from rest of the data so that during the training process, classifier does not affect by the overfitting problem. After the process of cross-validation, classifier built the model for prediction and then the model was tested on original test data i.e. data not affected by synthetic SMOTE data. Performance evaluation of the model is done through different matrices and measures.

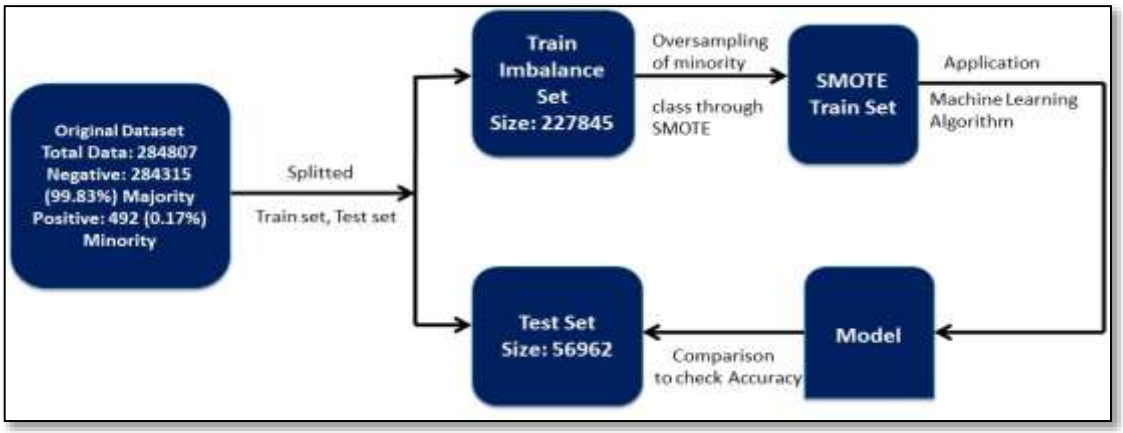


Figure 3.22: Experiment III using SMOTE Over-Sampling

Table 3.11: SMOTE results of ML Model during Experiment III

Matrices & Measures	L R	KNN	DT
Accuracy	0.97%	0.98%	0.99%
Precision	0.065%	0.24%	0.032%
Recall	0.87%	0.87%	0.72%
F1 – Score	0.12%	0.34%	0.45%
Misclassification	0.002%	0.002%	0.002%
Confusion Metrics			
Total Test Sample 56962 / Predicted Negative / Predicted Positive	55567 | 1297	56739 | 125	56734 | 130
Actual Negative / TN / FP	22 | 79	24 | 77	25 | 73
Actual Positive / FN / TP			

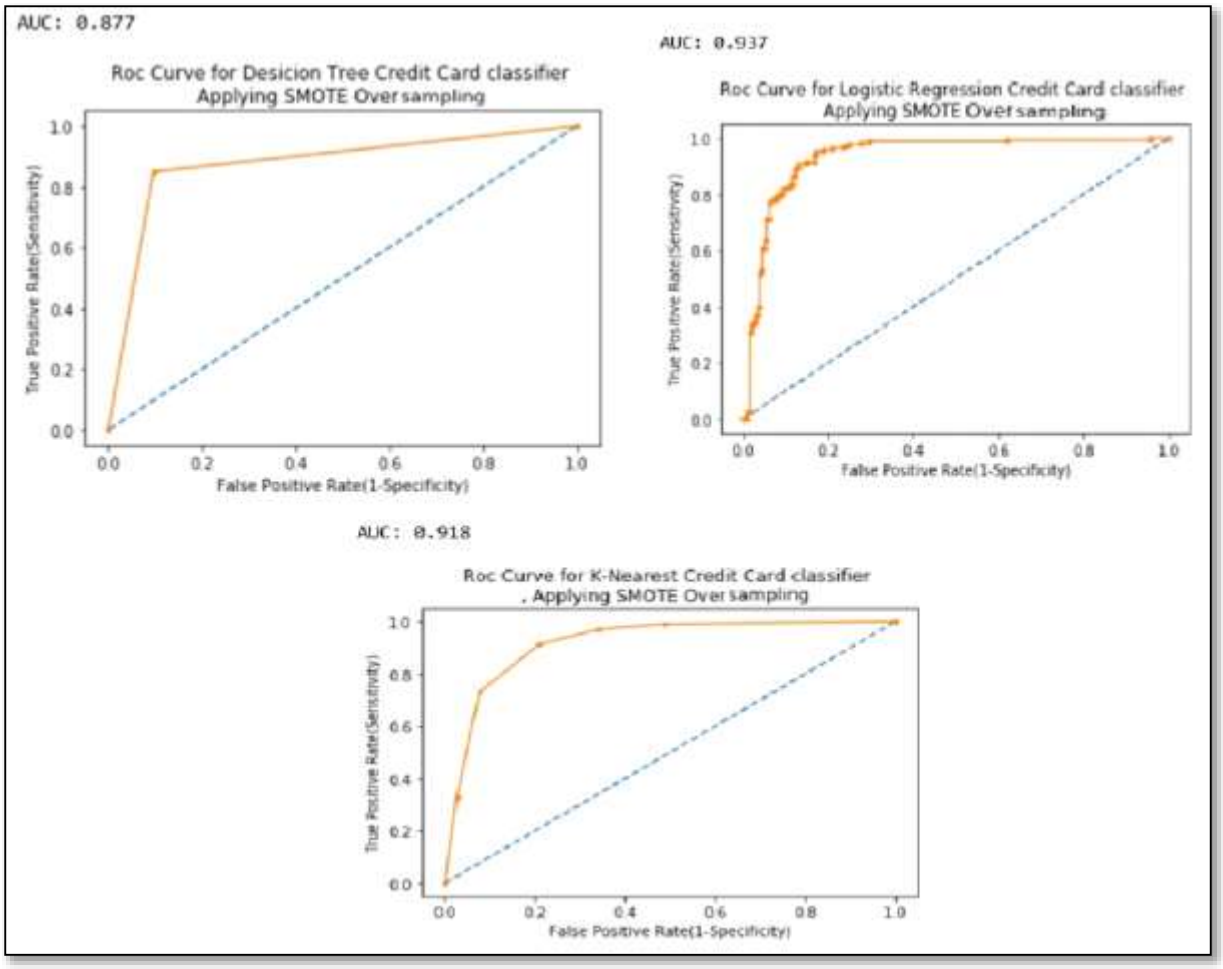


Figure 3.23: ROC Curve and AUC Score of ML Models during Experiment III

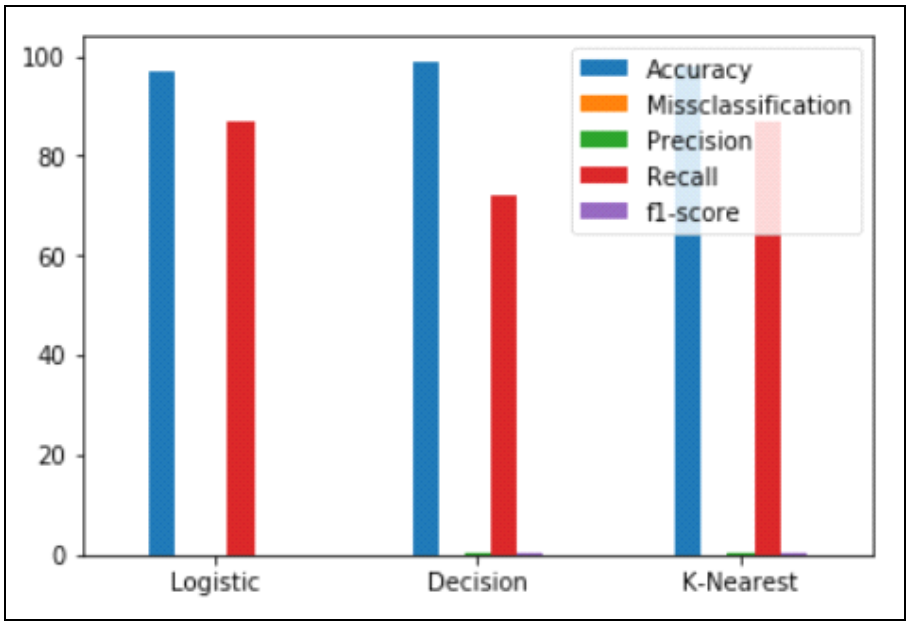


Figure 3.24: Performance Comparison of ML Models during Experiment III

3.2.3.4 Discussion

The current experiment was conducted to ensure the performance of SMOTE oversampling method for balancing dataset, so that binary classification algorithm such as LG, KNN, DT can build powerful models to detect credit card fraud detection. An experiment conducted with original credit card transaction dataset later divides into a training set (80%) and test set (20%). SMOTE was applied on training set during the process of cross-validation and not before it to avoid the problem of overfitting as well as data leakage and to keep test data untouched so that during the process of model validation data leakage does not occur. Performance of a model was evaluated with the help of a confusion matrix, precision, recall, F1-Score, and misclassification.

According to the result of this experiment, KNN was the best performer among three classifiers, with very less misclassification and high accuracy (98%) and recall (87%). Logistic regression performance was worst with lots of misclassifications. KNN and decision tree percentage for all measures and matrices shows almost equal result but we can affirm KNN superior to decision tree because in this case, KNN predicted fraud more accurately also rate of False Negative (FN i.e. number of times when the transaction was fraudulent was detected as genuine) was lesser in case of KNN than that of decision tree.

3.2.4 Summary

Experiment carried on with a highly imbalanced dataset fails to produce accurate result. Therefore several experimentations is conducted to check the result of random sampling and SMOTE oversampling in an imbalanced credit card transaction dataset. Both the technique aims to minimize the complexity of dataset. Random under sampling results show a high percentage of accuracy, sensitivity, precision, and F1-score with very low misclassification whereas in case SMOTE oversampling although accuracy was high still F1-score, precision was very low. Misclassification was higher than that of random sampling. Although random sampling performs better that SMOTE oversampling having lots of disadvantages. As random under sampling reduces the size of the majority class randomly, good amount of Information loss happens and thus generate bias result. Although SMOTE result is dependable it is very slow with high dimensions dataset, therefore none of both the solution perfect fits for all the real-time dataset. Enhanced and optimize solutions for optimizing the complexity of large and big datasets is still needed.

CHAPTER-4

Hybridization Pre-processing and Resampling Technique (HPRT) : A Solution

This chapter presents a new Hybrid Preprocessing and Resampling Technique (HPRT) for reducing the complexity and redundancy of a Bigdata. This hybrid approach combined with several ML classification algorithms is used to construct predictive model and to check result which is compared during the experiment. HPRT shows positive impact and enhances performance of ML algorithms when compared to old traditional sampling approaches.

4.1 HPRT: An Enhanced Technique for Optimizing Complexity and Redundancy of Bigdata

Dataset state of imbalance occurs when more numbers of examples for one class is present in it. A class having examples more in number are called majority class and other is called minority class. This problem is not new during mining the data. Since past years the researchers are continuously looking towards finding a solution for learning from imbalanced data but still, this issue is an important topic for research. With the alighting of big data, both data mining and machine learning technology over- emphasizing to grow into one of the extensive tools for eradicating deep insight from imbalance dataset. Dataset imbalance problem, on the other hand, gained lots of importance with the advancement of big data [95] and with it also bought much threat while gathering valuable information from it [39]. Up gradation of data level and algorithm level methods have been noticed with many new hybrid approaches [97]. Many domains do not have balanced dataset. Minority class consists of the most important information and extracting that piece of information from the big dataset is an actual challenge. During the processing of these datasets minority class does not get importance and therefore classification shows inaccurate results. These real-life imbalance dataset challenges galvanize researchers and scholars to focus an effective and real-time solution for such a problem. Fraudulent credit card transactions, intrusion detection, hardware fault detection, insurance risk modeling etc. are some of the examples of real-world imbalance dataset.

Before classification of imbalance complex datasets, resampling methods are headed for balancing such dataset. Old resampling methods like over-sampling and under-sampling techniques balance the dataset but have lots of disadvantages. OS when applied can cause problem of over-fitting and under-sampling cause's information loss. HPRT, an algorithm which will automatically balance the data set without much information loss and at the same time, the classifier will not cause any over-fitting problem. Our method uses both US and OS, where US occurs at majority class through reduction of redundancy and extreme outliers. This step can reduce majority class, almost up to 40% and then in the next step SMOTE OS is applied to a minority class. This two-step approach can balance the dataset while cleaning it and as a result machine learning algorithm performance can enhance.

HPRT consists of several steps; each step is designed to reduce the complexity of the dataset automatically. This algorithm is basically designed for large and big datasets, where once an imbalance and the complex dataset is input to an algorithm; HPRT will reduce the complexity of the dataset and convert it to form balance state, in several steps. Machine learning techniques are applied in this algorithm thus the process is automated. In the first step, PCA reduces the dimensionality of a big dataset. Minority class is then separated as we do not want to lose any kind of information consisting in it after which K-Mean unsupervised clustering algorithm divides majority class into several clusters depending upon the business problem. This step divides the similar kind of data feature in the same cluster and hence minimizes comparison between features of majority class. Redundancy is reduced using divide and conquer rule within each cluster and then extreme outliers have been dropped from each cluster with the help of Tukey method. In this way almost 30%, 40%of the complexity had been removed. In the next step synthetic sample is added to minority class during cross-validation to avoid data leakage problem and reduce the chances of over-fitting. Once the dataset passes through all these steps of the proposed algorithm automatically balance dataset and then we can apply machine learning binary classification for construction of the predictive model. The HPRT is an enhanced and new resampling preprocessing approach for reducing the complexity of imbalance large and big dataset through reduction of redundancy and extreme outliers from majority class and then sample is increased in minority class through SMOTE. This enhanced algorithm is applied to the credit card transaction imbalance dataset used as a sample dataset. Thus, it successfully reduces the complexity of the dataset and proved to be enhancing resampling technique for balancing of the dataset.

4.1.1 PCA & K-Mean Clustering in HPRT

PCA is a dimensionality reduction method which extracts the set of features from a very large or big dataset and converts it into low dimension dataset. In other words, it extracts all the information from a high dimensional dataset with an intention to capture almost all the information from a big and high dimensional dataset. Increased size of data is not only hazards for storing but processing it also becomes problematic for many traditional machine learning approach. Nowadays, sizes of data anticipating from different sources have been enlarged containing in numerous features in it [98]. Out of these numerous features, many are redundant and convey the same piece of information. Therefore, this type of redundant features should be removed so that dataset contains less but very important and meaningful information. PCA checks the variance of each property into a dataset and collects all the high variance data to form a new set of examples based on original one while absorbing all the information consists in it. PCA linearly transformed high dimensional original data using the algebraic calculation of principal component.

PCA is very popular dimensionality reduction algorithm [98], which is used in most of the areas, such as for gene expression analysis data, stock market prediction, a medical dataset for predicting disease, and many such places where a dataset is big. In general, there are two types of linear transformation method: Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA). LDA make use of class information whereas PCA uses the variance of each feature to find new features so that its separation should maximize. PCA perform a mathematical calculation in several steps to get the desired low dimensions. Data is standardized in the first step after which covariance matrix from the dataset is generated. In the next steps, Eigen vector and Eigen values are formed and sorted in decreasing order. N number of Eigen vector having largest Eigen values is selected to form a new feature sub place where n is a number of dimensions. Let the dimension of the original set of data is X where N will always remain less than X, hence forming N Eigen vectors X projection matrix is constructed. Therefore transformation of original dataset D is done from X having N subspace feature dimension. PCA technique is also used to reduce redundant feature in the dataset. But PCA works for numeric data and therefore before applying PCA we should all the features of the dataset should be converted into a number. PCA can be used for high dimension data for reducing its dimensionality. Due to several reason discussed above PCA is used in our new proposed technique HPRT.

Clustering is a technique of assembling a group of similar elements. Dataset divides into different groups by placing identical items in a similar group to form a cluster. Features in the same cluster contain the same properties as much as possible. Features in the different cluster will be different altogether. Clustering is an unsupervised technique which is used in almost every domain. Now a day's big companies like Amazon, Netflix, Facebook, etc. are using clustering to manage the data appropriately [99]. Banks are using clustering technique for analyzing customer probabilities and their credit score. Finance companies are using clustering, for credit card detection, fraud detection, and identification of risk factor. It is used in market segmentation, image segregation and also in grouping web pages. In the business problems, clustering helps to formulate rewarding decision by analyzing customer behavior for shopping.

With the augment of big data expansion in the volume of data occurs speedily. Big data is a term given to the large volume of the complex dataset, processing of which is not possible using traditional methods. Therefore clustering is very important data mining method used in big data analytics. When data is too large gather from different source traditional approaches and old technology are not sufficient for managing it, even if a very high-end machine is used. Therefore, to overcome this challenge, new machine learning and data mining based technology is used for processing and analyzing of such vast dataset. Data clustering is an important data mining approach [100] for processing and analyzing big data. Big data sets consist of large volumes of different varieties of messy data having many groups in it but in disorganized format. It is very important to group them in a systematic approach with similar items, therefore data mining based data clustering method is used to form clusters for combining different items in the dataset. Nowadays, clustering is used in many big data problems, such as in healthcare department, it is used to discover important pattern from patients records, identification of duplicate records easily and elimination of it in a different dataset, clustering millions of web pages or web documents.

A cluster can be divided into two different groups, (a) hard clustering and (b) soft clustering. Hard cluster data points belong to one cluster exclusively. In soft cluster data points can belongs to several clusters, depending upon the probability of given data point presents in those assigned clusters. Although there are more than 100 clustering algorithms, for this study, we are interested in K-Means clustering because of its extensive and easy to use characteristic. K-Means algorithm is a significant machine learning iterative technique where the closeness of a data point is calculated based on the centroid of the cluster.

K-Means clustering is an unsupervised machine learning approach as it is used for classification when a label is not present in the data. K-Means algorithm explicitly accepts the value of K, as it divides the dataset into K number of groups by placing a similar item in one group. The algorithm performs a number of iterations, to assign each data point to groups depending upon its feature. K-Means define centroid for each group K, with the help of which new data can be labeled. It classifies training dataset by defining labels for them. K-Means algorithm allows analyzing the cluster which is formed systematically. The centroid is the center point of the group depending on which other data points are placed on a cluster accordingly. K-Means algorithm is used in many domains for segregation of facts into groups. It is very useful in business problems for grouping purchase history, several activities, clustering inventories depending on various modes, finding groups for images, it also benefits in health care by clustering patient records into a different category. K-Means are used in the detection of anomalies also. It helps to discards, noisy data, like outliers and redundant items more efficiently by tying them together. In a dataset containing a huge amount of data, redundancy reduction or outlier's reduction and other cleanup process is a big challenge if data is not categorized properly because a number of comparisons are required in such situation. Therefore in order to make the task easier and simpler, data should be huddled in the same group based on similarity of an item.

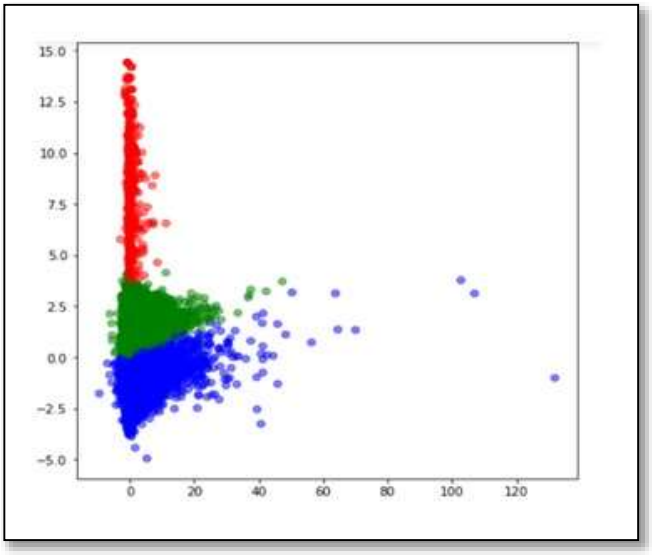


Figure 4.1: K-Mean Clustering in HPRT

K-Means unsupervised clustering algorithm is an iterative process which accepts dataset (on which clustering has to be performed) and K (the number of a cluster into dataset has to divide) as an input. K-Means algorithm initiates one centroid randomly for each of the

K group from the items in a dataset. Euclidian distance is calculated between centroids and each of the data points and then those data points are placed to their nearest centroid calculated in the previous step. Let ds is a dataset, Cen_j is a collection of centroids in ds, x is a data point in ds. Then each Items x in ds, assigned in a cluster is given as shown in eq. 4.1,

Items x in ds, assigned in cluster = Euclidian distance $(Cen_j, x)^2$ …..(4.1)

Where, Cen_j∈ds

In the next step centroid of a particular cluster is recalculated based on the mean of each of the item present on that cluster. Above steps are repeated till centroid becomes constant. K-Means require few calculations and hence it is very fast and applicable for large volumes of data having linear complexity O (n). The intention of K-Means is to minimize the squared error function as shown in eq. 4.2.

$$f(j) = \sum_{j=1}^{C} \sum_{k=1}^{n} || x_k^j - Cen_j ||^2 \quad \text{…..(4.2)}$$

$where, || x_k^j - Cen_j ||^2$ → Distance function

N → number of iteration

I → i^{th} iteration

f(j) → function to search centroid

C → number of cluster

Due of several advantages of K-Mean clustering, which is discussed above we used it in HPRT, it divided datasets into several groups of similar items [101] which help in reducing numbers of comparison which reduces both time and complexity. Therefore, it is suitable for Bigdata processing as it is powerful enough to explore such vast dataset for several numbers of without much complication.

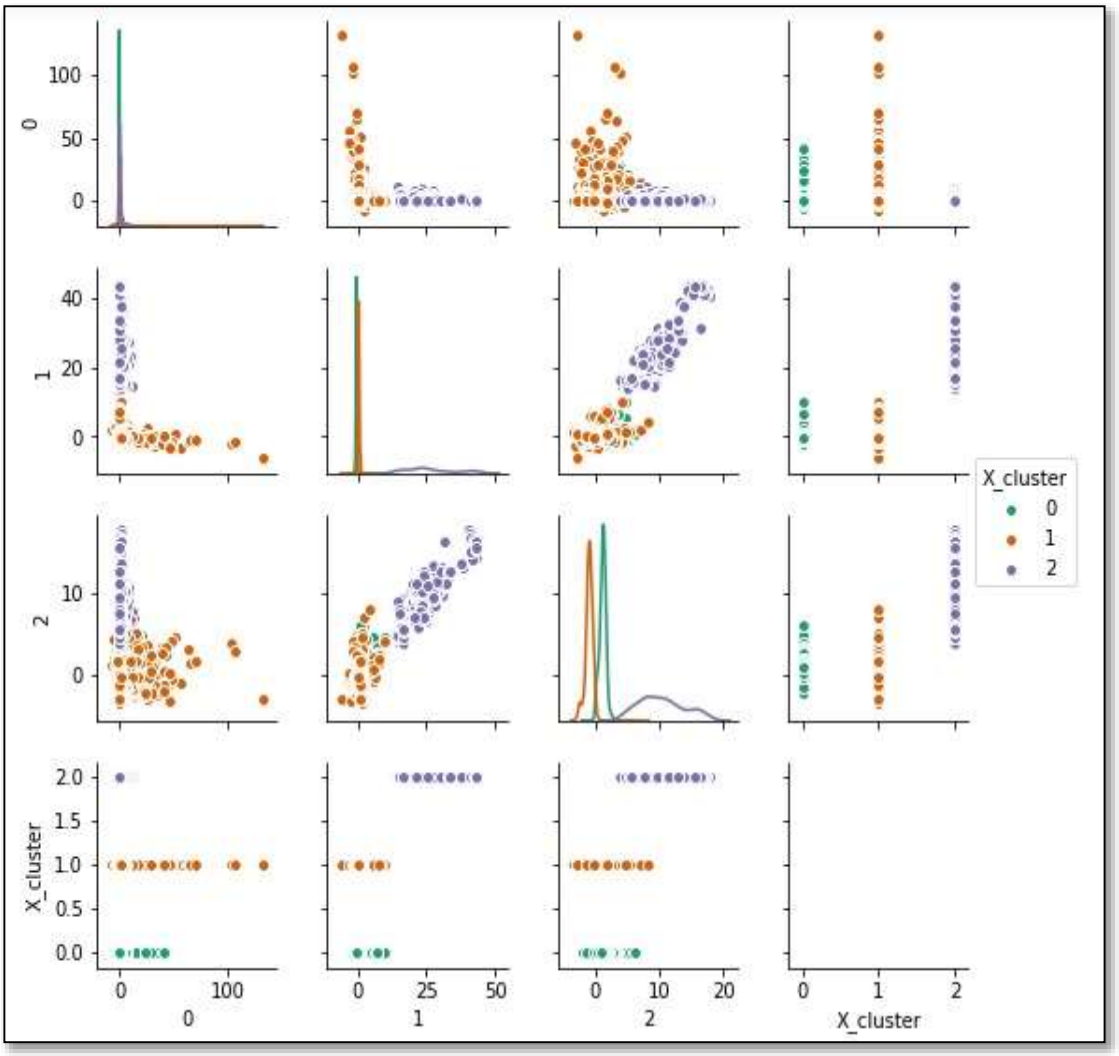


Figure 4.2: Columns K-Mean Clustering Representation of Sample Dataset

4.1.2 Outlier: A Problem for ML Algorithms

Outliers are extreme values in an observation of dataset that falls beyond the range from other records that are more or less similar to each other [16]. Outliers can occur in a dataset due to some experiment error variation in measurement etc. Outliers in a dataset should be taken care of before application of machine learning algorithm to the dataset in order to get an appropriate result because machine learning algorithm is sensitive towards the distribution of the dataset. A massive amount of outliers in a dataset can cause poor and inaccurate models with longer training times by misguiding machine learning algorithms. In a dataset, outliers can be part of records during the collection, processing or analyzing records. Machine learning algorithm like linear and logistic regression is very much affected in its training process with the presence of outliers. Human error, instruments error, experimental errors, data processing

errors, etc. are some of the causes of outliers. Outliers can be broadly classified into two types' univariate and multivariate.

Univariate outliers occur when extreme values are searched on one feature of the dataset. Box plots fall under the category of a univariate method. It is one of the simplest and popular techniques for the detection of outliers. Box plot describes a feature of the dataset, by using statistics calculation such as lower quartiles, upper quartiles, and median. Box plot uses the format of the box for specification of data distribution. Box plot, known as Tukey's method, introduced in 1977 is a very important visualization tool for the detection of outliers by displaying univariate feature as lower quartile, upper quartile, lower extreme, upper extreme and median of the records present in the dataset. IQR is an interquartile range, which is a distance between Q1 and Q3 quartiles. Inner fences are the place at 1.5 IQR below Q1 and above Q3 distance. Outer fences are at the place of 3 IQR below Q1 and above Q3. Features detected between the inner and outer fences are the possible outliers and values beyond outer and inner fences can be identified as extreme outliers.

Multivariate method occurs when extreme values are searched throughout the dataset. Multivariate method built a model for outlier detection. It uses all the data at once and solves the problem through cleaning the instances having errors above a value specified.

Presence of outliers in a dataset during construction of some machine learning models Increases misclassification [16]. In many cases, recognition of outliers is vital for the extraction of useful information. During the process of data mining, unknown information is extracted from a large amount of noisy dataset. But this process may bring many challenges such as data redundancy, incomplete and missing values, outliers etc. Now a day’s outlier detection is gathering importance while mining the data because detection of such deviated outlier in a dataset can uncover much-hidden information and had importance in many domains such as health care, weather, location-based domains, transport department etc. Many methods for the detection of outliers is based on statistics such as a Gaussian mixture model, used by vaminish et al. Many researchers proposed many different statistical approaches for the detection of outliers.

In this study, we use the Tukey method with K-Means clustering to detect extreme outlier. Care is taken to drop only extreme outliers without losing important information. Dropping extreme outliers present in majority class of our large dataset; will help in balancing the dataset, to some extent, without any fear of vital information loss which creates positive effects on the accuracy of the model. We determine threshold by multiplying IQR with 1.5. This value act as a threshold for detecting extreme outliers and therefore not much vital

information will discard from the dataset. Value beyond threshold range will consider as outliers and can be dropped from the dataset. Threshold value lesser than 1.5 can able to detect more outliers and on the other hand higher than 1.5 will detect comparatively fewer outliers. Box plot is used for the visualization of outliers.

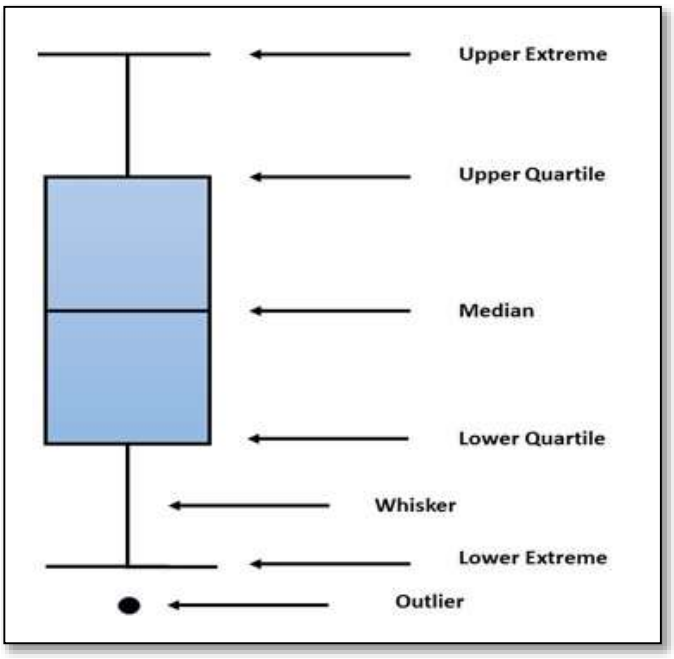


Figure 4.3: Outlier Detection Using Box Plot (Tukey Method)

4.1.3 Redundancy: A Problem for ML Algorithms

Redundancy is a phase where more than one copy of data is present in the database. The same piece of more information can cause inconsistency in a database. Data redundancy occurs either circumstantially during storage or deliberately during backup purposes [50]. Predictive model accuracy depends upon the quality of training and pattern used while learning process. Therefore features present in a dataset play a very important role for the training of model as well as in its accuracy. In such condition redundant and non-relevant features in a dataset end with the construction of a weak model resulting poor accuracy during prediction [104]. However many researchers have been conducted to overcome this situation. With the invention of big data, the clustering algorithm has been used to a great extent to check for redundant data and remove it from a dataset.

During the experiment, redundant data is removed from the majority class of given imbalance dataset for reducing complexity and enhancing its performance. Removing redundancy from imbalance dataset will remove the irrelevant feature which does not make any sense during the process of machine learning model building. Therefore removal of

redundant data from the dataset does not cause information loss. So we separated minority and majority classes in a dataset and reduce the majority class by discarding redundancy and outliers from the dataset for reducing the complexity of an imbalanced dataset.

4.1.4 Towards Solution: HPRT as a Tool to Enhance ML Algorithms

Two popular resampling methods, Random US and SMOTE OS, cause certain challenges during its application to an imbalanced dataset. Random US, abandon the majority class features randomly causing enormous information loss. Consider the scenario where random under sampling is applied to big data for business decision making or medical big data for detecting any severe disease or credit card fraud detection dataset and during re-sampling features reduced randomly, can discard immense amount of vital information and leave the dataset with not as much of informative data, hence the traditional machine learning algorithm cannot learn a strong predictive pattern. SMOTE cannot perform well with high dimensional big data, and application of it during re-sampling big data will cause several complications. Therefore, the goal of this research is to develop an enhanced and hybrid machine learning resampling approach, which can act as a perfect solution for reducing the complexity of big data because of its imbalance nature.

With the rise in big data complexities of the datasets also increases, thus traditional methods like random US, random OS, OS through SMOTE suffers from various challenges. Therefore intention is to propose enhanced machine learning approach and masquerade the limitation of traditional resampling methods. Here in this enhanced hybrid resampling approach both US and OS is applied to a dataset for balancing it but in a different approach. In first step feature reduction of the majority class is done by removing extreme outliers and redundant data. Redundancy in a dataset occurs when a piece of information exists for more than one time, either intentionally or unintentionally. Redundancy is consisting of repetition the variable, the existence of which brings inconsistency in dataset. Outliers are the extreme value that does not match with other observations in a dataset. Occurrences of redundant as well as outlier's data in a dataset will make it noisy and removing it will not generate much information loss. In the second step minority class is increased through adding data point synthetically to it through SMOTE. Keeping this in mind HPRT is proposed which is an enhanced automated machine learning based resampling approach, which can automatically convert the imbalance and complex structure of a dataset. Our enhance technique automates

the process of resampling for big datasets by detecting the imbalanced nature of a dataset and converting it into balanced dataset automatically.

HPRT accepts original imbalance dataset as an input. The dataset than undergoes through several steps, wherein each step complexity of the large dataset has been tried to reduce. In the first step, an algorithm accepts, big or large dataset as an input and then divide it into two subparts, such that 1st part contains all the features from majority class and another part contains all the features from minority class. In this way two subsets are derived from the original dataset are N and F where N is consist of only majority features and F is consist of minority features. In the next step, we individually apply PCA algorithm on both the subset to form $P_1$and P_2. This is a first phase where the complexity of the dataset is reduced through the dimensionality reduction. In the next step, we will take only subset P1 $\{V_1, V_2, V_3, \ldots V_n\}$ and divide it into various clusters with the help of K-mean unsupervised machine learning algorithms. K-means is capable of dividing observations in datasets into K different clusters. Clusters can be defined as the entities of similar group i.e. features in one cluster are identical that the features in the other clusters. With the help of K-mean clustering algorithm P1 $\{V_1, V_2, V_3 \ldots V_n\}$ is divided into $P_1\{k_i, k_j, \ldots k_n\}$.The cluster is developed to tie similar items in one group so that the number of comparisons will decrease to an enormous extent and which is helpful in reducing the processing time of big and large dataset. Redundancy is checked in each clusters using divide and conquer rule. Duplicate record detected is immediately discarded from the cluster.

After the removal of unwanted redundant data from the clusters $P_1\{k_i, k_j, \ldots..k_n\}$, algorithm proceed with the detection of outliers. IQR method is used for the detection of extreme outliers. The threshold is calculated by multiply 1QR with 1.5, upper bound and lower bound is calculated. A feature is considered an outlier if its value is less than lower bound or more than the upper bound. All the outlier is stored in temporary list O and discarded at last from the cluster. At this stage, our clusters P1 $\{V_1, V_2 \ldots Vn\}$ is free from outliers as well as duplicates data and then the cluster is again converted into a data frame. Half of our intension to reduce the majority class is achieved through these steps. Hence this reduction in the majority class is achieved without any type of information loss. In the next step, we concatenate both subset P1 $\{V_1, V_2, V_3 \ldots V_n\}$ and P2 $\{V_1, V_2, V_3 \ldots V_n\}$ to form D $\{V_1, V_2, V_3 \ldots V_n\}$.Then, In the next step SMOTE over sampling is used to balance minority class of the dataset. In this process features in minority class are increased by adding artificial sample to the minority class with the help of the KNN algorithm and Euclidean Distance. SMOTE oversampling is performed during cross-validation to avoid overfitting because, smote if applied earlier there

can be a possibility of adding features exactly the same in the validation set, hence causing the problem of data leakage. To avoid this, validation set must be excluded first with other training set and then during cross validation over sampling should be applied on rest of the dataset. This will avoid over fitting as well as data leakage problem because during testing period validation set will be completely unseen.

Above technique balances the dataset thus reducing its complexity and then in next part binary classification algorithm, such as logistic regression, K-nearest neighbor and decision tree had been applied on the balanced dataset, so as to construct credit card fraud detection models. The models are evaluated with the help of a confusion matrix, sensitivity, precision, accuracy, and misclassification. Confusion matrix will give a total count of true negative (TN), true positive (TP), false negative (FN) and false positive (FP). TN is a correct count for negative class, TP is a correct count of positive class, FN is a wrong prediction for negative class and FP is a wrongly classified positive class.

Logistic regression predicts the result based on probabilities and sigmoid curve value. The value of sigmoid curve ranges from 0 and 1. Probabilities calculated from the input feature during logistic regression process passes through the sigmoid curve. If the output is greater than or equal to 0.5, then the class is predicted as 1 (positive class) and if the output is less than 1 then the class is predicted as 0 (negative class).

KNN predicts the result after calculating the distance of unknown features with that of given features distance. Euclidean distance is used for distance calculation here. The result is predicted based on a number of K-nearest neighbor which is provided by the user. In this study parameter for k is 5 i.e., the algorithm will consider five nearest data point with that of the unidentified feature, and then output class for that feature will be calculated by counting class of the data points appears for a maximum number of time.

Decision tree builds models, by defining rules for training examples. During this process, trees are constructed to solve the classification problem. Generation of the tree starts with a selection of root node from the training dataset and then the process continues to find branches, sub-branches etc. The process of searching roots and its subset is repeated for each branch of the tree. Gini index is a matrix used in this experiment for selection of roots during tree construction. It identifies root node by measuring how often attribute chosen randomly can be incorrect.

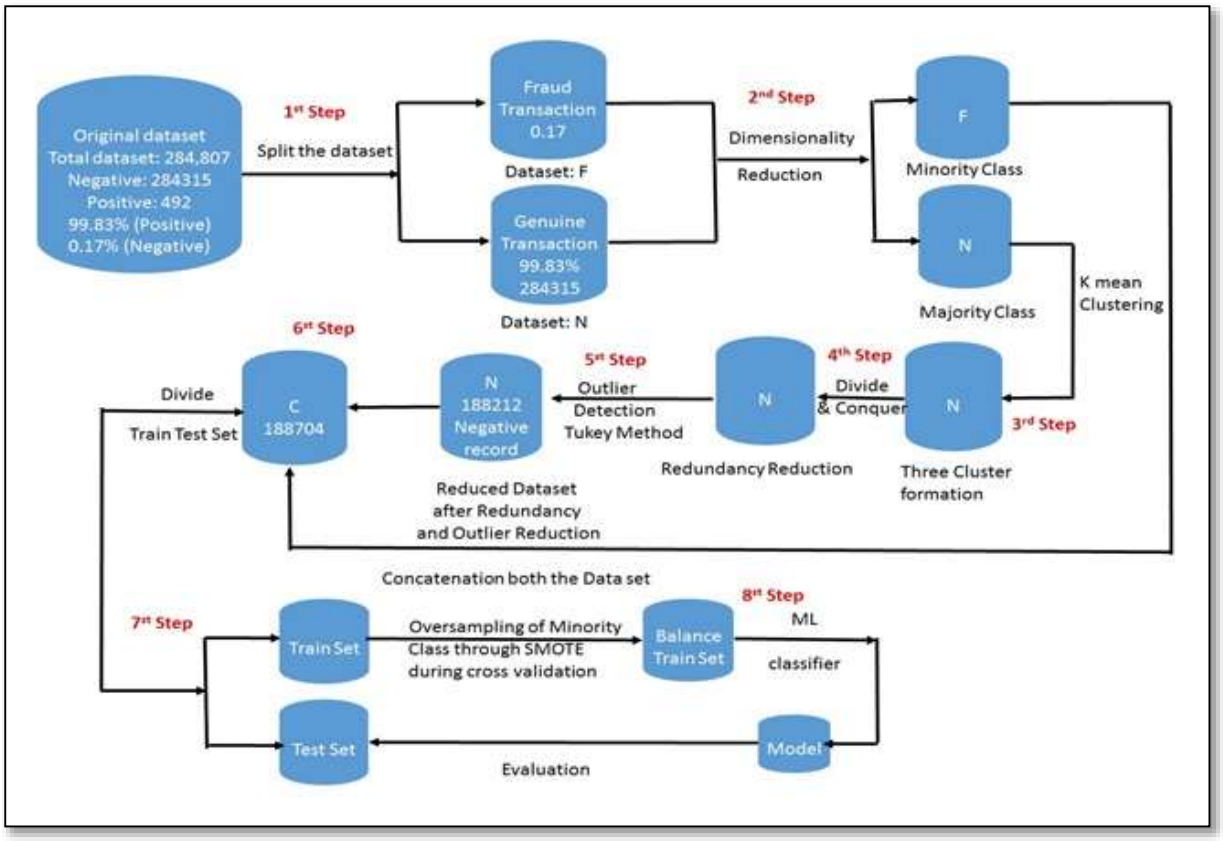


Figure 4.4: Application of HPRT on a Sample Dataset for Enhancing ML Models

4.1.5 Designing of HPRT to Enhance ML Algorithms

A new tool HPRT, for optimizing complexity of an Imbalanced Big Data. It is an enhanced resampling and preprocessing technique for balancing the dataset while cleaning it, such that enhancing performance of ML classifier.

Input: Imbalanced Big Dataset C {V_1, V_2, V_3…Vn}

Output: HPRT based ML Model constructed on top of balanced dataset

➔ **Step 1:** Load dataset C, divide it into two parts (N, F)

F consist only Minority Class

$F \in C$

N consist only Majority Class

$N \in C$

➔ **Step 2:** Reduce the dimensionality of both N and F separately through the application of PCA

1. Standardize dataset N and F so that values of all the features V1, V_2, V_3 …Vn exists within 0 to 1 without changing its original meaning.

$N = Standardize (N \{V_1, V_2, V_3 ...V_n \})$

$F = Standardize (F \{V_1, V_2, V_3...V_n\})$

2. Calculate covariance matrix for given set of data

$V_1 = Cov (N \{\{V_1, V_2, V_3 ...V_n\})$

$V_2 = Cov (F \{\{V_1, V_2, V_3 … V_n\})$

3. Calculate Eigen values and Eigen vectors for matrix V_1 and V_2

   $E_1, E_2 = Eigen (V_1)$

   $E_3, E_4 = Eigen (V_2)$

   Z_1 = Select K_1, Eigen values having K, largest Eigen values (E_1, E_2)

   Z_2 = Select K_2, Eigen values having K, largest Eigen values (E_3, E_4)

4. Project data by multiplying the original matrix with its transpose.

   $P_1 = Z_1^T . N$

   $P_2 = Z_2^T . N$

   Where,

   P_1= N {V_1, V_2, V_3, …V_n}

   P_2 = F {V_1, V_2, V_3, …V_n} undergone through PCA process.

➔ **Step 3:** Apply K- Means clustering on P_1 to divide it into k_n number of cluster.

1. P_1= P_1 {k_1, k_2,…k_n}
2. Select centroid for P_1 {C_1, C_2 … C_n} for each of the cluster P_1{k_1, k_2 … k_n} in P_1.
3. Calculate the distance between each vector in a cluster P1 {k_1, k_2 … k_n } and search for closest centroid.
4. Evaluate new centroid for each cluster k_1, k_2 … k_n applying divide and conquer rule.
5. Repeat above process 2, 3, 4 till centroid values for P_1 {k_1, k_2 …k_n} Becomes constant.

➔ **Step 4:** Dropping redundant data from each of the clusters in P_1 {k_1, k_2, ….. k_n}

1. Comparison for detection of redundant data points within cluster P_1{k_1, k_2, ….. k_n}

For i = 1 to k_n

For j = 1 to number of data points X in k_i

Search for redundant records using divide and conquer rule and store in R.

R = Duplicate features X_i

Discard R form the clusters

End For

End For

Note: $P_1 \{k_1, k_2 \ldots k_n\}$ after redundancy reduction is left from which outliers is removed in next step.

➔ **Step 5:** Dropping extreme outliers feature X from each cluster $P_1 \{k_1, k_2, \ldots.k_n\}$ using IQR

1. Sort each cluster $P_1\{k_1, k_2, \ldots.. k_n\}$ in ascending order
2. Calculate inter quartile range within each of cluster P_1 { k_1 { $X_1, X_2, x_3, \ldots X_n\}$, $k_2\{ X_1, X_2, X_3 \ldots X_n\} \ldots k_n \{ X_1, X_2, X_3 \ldots X_n\}\}$

For i = 1 to k_n

For j = 1 to Xn number of data point

Note: Calculate 25 %(q1) and 75% (q2) of data in each cluster k_i with in for loop

Iqr = q2 – q1

Threshold = iqr * 1.5

Note: Calculating upper bound (u) and lower bound (l) for each data point Xj in cluster ki

l = q1 – threshold

u = q2 – threshold

End For

For j = 1 to Xn number of data point

If Xj < l or Xj > u

Note: Store extreme outliers Xj in O

O = Xj

Drop O

End For

End For

Note: P1 {k1, k_2 … k_n} is free from outliers at this stage. Good amount of redundant and outliers feature is reduced from P_1 {k_i, k_j … k_n} without losing much information.

➔ **Step 6**: P_1 {k_i, k_j … k_n} is converted to data frame again to form P_1 { V_1, V_2, ..V_n}.D is dataset after concatenation of P_1{V_1, V_2, V_3 …V_n} and P_2 { V_1, V_2, V_3 …V_n}

$$D \Xi P_1 + P_2$$

Note: SMOTE Over Sampling is applied to minority class in the next step

➔ **Step 7**: Split D {V_1, V_2, V_3, ……V_n} into X_1 and Y_1 such that X_1 is consist of an entire set of data excluding target class and Y_1 is consist of target class only.

X_1 = D. Drop {target class}

X_2 = D. Drop {target class}

➔ **Step 8**: Generate train and test set splits randomly using stratified split with number of splits = 10

S = StarifiedShuffle (Split = 10, test-size = 0.2)

For train-index, test-index in S (X_1, Y_1)

X_{Train}, Y_{Train} = X_1 [train-index], X_1 [test-Index]

X_{Test}, Y_{Test} = Y_1 [train-index], Y_1 [test - Index]

Note: X_{Train}, Y_{Train} is 80% of data from dataset used for training a model X_{Test}, Y_{Test} is 20% of data from the dataset for testing model prediction after training. Above set of algorithm will create 10 sets of X_{Train}, Y_{Train}, X_{Test}, Y_{Test} for validation purpose where in each iteration, from 10 sets 1 will act as test set and other nine will remain as training set.

➔ **Step 9**: Using GridSearchCV for searching best parameters for the classifiers.

➔ **Step 10**: Using SMOTE during cross validation for adding synthetic features in minority class

For train in split X_{Train}, Y_{Train}

P = SMOTE (X_{Train}, Y_{Train})

➔ **Step 11**: HPRT based ML model for classification future data

$$M = P.Classifier\ (X_{Train,}[train]\ ,Y_{Train}[train])$$

Note: M is a model constructed using P (HPRT applicant subset) with ML Classifier

4.1.6 Experiment I: Comparison of HPRT Based Selected ML Classifier

HPRT is applied to highly imbalanced credit card transaction dataset. It is consists of two days transaction of European cardholders having two classes, Fraud (0.17%) and Non-Fraud (99.83%). The proposed algorithm can be applied for large and big dataset as a preprocessing algorithm for reducing the complexity of an imbalanced dataset. This algorithm automatically detects the level of imbalance in a dataset and then reduces it automatically in various steps. Then three different machine learning classifiers (LR, KNN and DT) are used to construct a model for the balanced dataset. Confusion matrix and various other measures evaluate the model which shows KNN as the best model with (99%) accuracy, precision (0.48%), Recall (0.91%), F1-Score (0.58%) and misclassification (0.014%). Confusion matrix confirms that out of 37741 test data only 55 times model predicted wrong result. Performance of decision tree is also satisfactory with 64 wrong predictions having an accuracy (99%), precision (35%), Recall (92%), F1-Score (48%), Misclassification (0.016%). Logistic regression perform worst among them, having an accuracy (98%), Precision (11%), Recall (92%), F1-Score (21%) and misclassification (0.018%) which is higher can that of other two model. Confusion matrix shows 689 wrong predictions among 37741 test data. For all the models constructed during this study proves application of proposed as a preprocessing step enhances their performances in terms for predicting frauds and non-frauds accurately.

Table 4.1: Results of HPRT based ML Models during Experiment I

Matrices & Measures	L R	KNN	DT
Accuracy	0.98%	0.99%	0.99%
Precision	0.11%	0.48%	0.38%
Recall	0.92%	0.91%	0.92%
F1 – Score	0.21%	0.58%	0.48%
Misclassification	0.018%	0.014%	0.016%
Confusion Metrics			
Total Test Samples 37741 / Predicted Negative / Predicted Positive			
Actual Negative / TN / FP	36962 / 681	367603 / 4	37592 / 51
Actual Positive / FN / TP	8 / 90	15 / 83	13 / 85

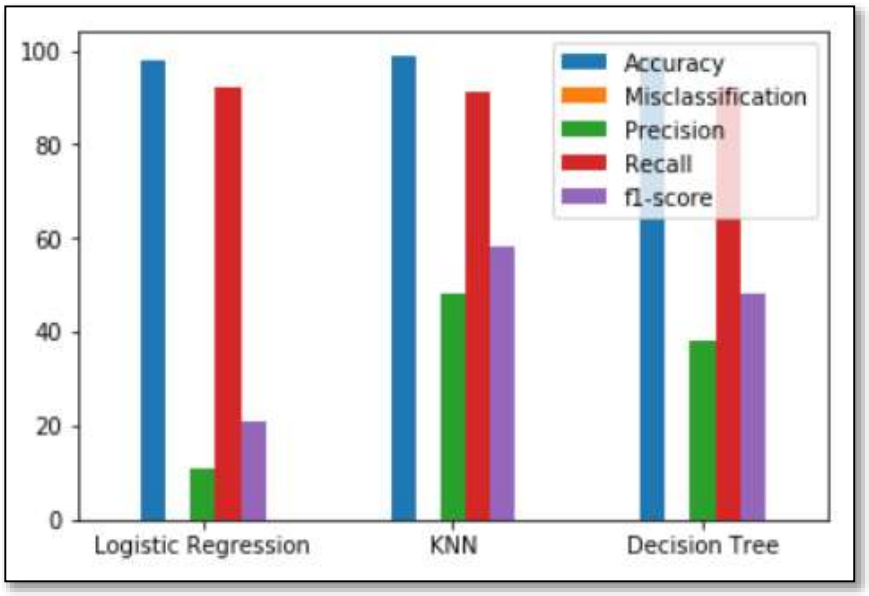


Figure 4.5: Performance Comparison of HPRT based ML Models during Experiment I

HPRT, An enhance Hybridization Pre-processing and Resampling machine learning approach for optimizing the complexity of an imbalanced dataset by dropping redundancy and outliers from majority class and then adding synthetic feature using SMOTE OS in minority class has been proposed in this study. HPRT successfully overcomes the disadvantage of

traditional sampling methods i.e. random US, OS and SMOTH OS. Random US causes huge loss of information, whereas random OS causes over fitting. HPRT is applied to imbalance sample dataset to balance it automatically. Among all the three machine learning algorithm, applied to the balanced dataset. The KNN algorithm performance was better but it is not competent for big dataset because it is an in-memory algorithm. KNN fails to construct good models with dataset having high volume and high dimensionality. Bigdata computation cost is very high with KNN algorithm, as the distance is calculated for every data points present in the dataset. Therefore, the model constructed by the KNN algorithm is not at all suitable for Bigdata. HPRT based robust machine learning algorithm can be perfect combination for optimizing imbalance and complex Bigdata. Therefore HPRT based neural network in used in a next section for of Bigdata classification.

4.2 Effectual Classification of Bigdata with an HPRT based Neural Network

An Imbalanced class is one of the foremost problems found in almost every real-world dataset and aggrandized since the rise of big data. If a dataset is an imbalance, then one class data dominates other and hence most of the traditional machine learning techniques, including neural network [106] fails to construct accurate learning rules for these datasets due to complexity existed in them. Traditional ML algorithms are strongly biased towards the majority class when two classes are poorly distributed. Here in this study, for classification of such datasets, we built a model by conjoining neural network with an enhanced integrated machine learning algorithm, which boosts up the performances of the neural network based classifier for such highly imbalanced datasets. For this purpose, we use a neural network to build classification model but before it, we optimize the complexity of the dataset with help of a HPRT. This methodology is a hybridization preprocessing and resampling technique which balances the dataset by reducing majority samples and increasing minority samples in a several steps. In the first step, algorithm divides majority and minority class. Then minority class is reduced by dropping redundancy and extreme outliers and minority class are increased by adding synthetic features through SMOTE, to form normal distribution between both the classes and then the algorithm concatenate both the classes again converting an imbalanced dataset to balanced. Hence balance dataset is provided to the neural network algorithm so as to construct a neural network based model for the binary classification.

We conducted two experiments for constructing Neural network model for classification of frauds and non-frauds from credit card transaction dataset – (a) in a first experiment an imbalanced dataset is a feed to traditional neural network classifiers. The first

model is consisting of imbalance dataset as input to traditional MLP architecture (1 input layer, 1 hidden layer, and one output layer. Due to imbalance nature of the dataset model does not perform well during classification of frauds and non-frauds. The model achieves high accuracy with low F1-Score. Confusion matrix present lot many false negative i.e. many fraudulent transactions detected as non-fraudulent. This type of model possess loss to finance industry. Traditional neural network model results are: accuracy (99%), Recall (78%), Precision (75%) and recall (39%) rate. Confusion matrix results display 71 misclassifications out of that 60-time model predict frauds as genuine which is very costly for any model. Therefore model was not further analyzed and dropped with an intention to construct efficient MLP based classifier. (b)In a second experiment, the neural network was combined with the proposed model. Our proposed model was first applied to an imbalance dataset to make it balance and then the neural network is used for building a predictive classification model. The second model result was outstanding, predicting both frauds and non-frauds correctly with very less misclassification. Accuracy matrix shows almost 99% accuracy with good precision (100%), recall (85%) and F1-score (89%) rate. Confusion matrix produces a good result with very less number of misclassification (21 times).

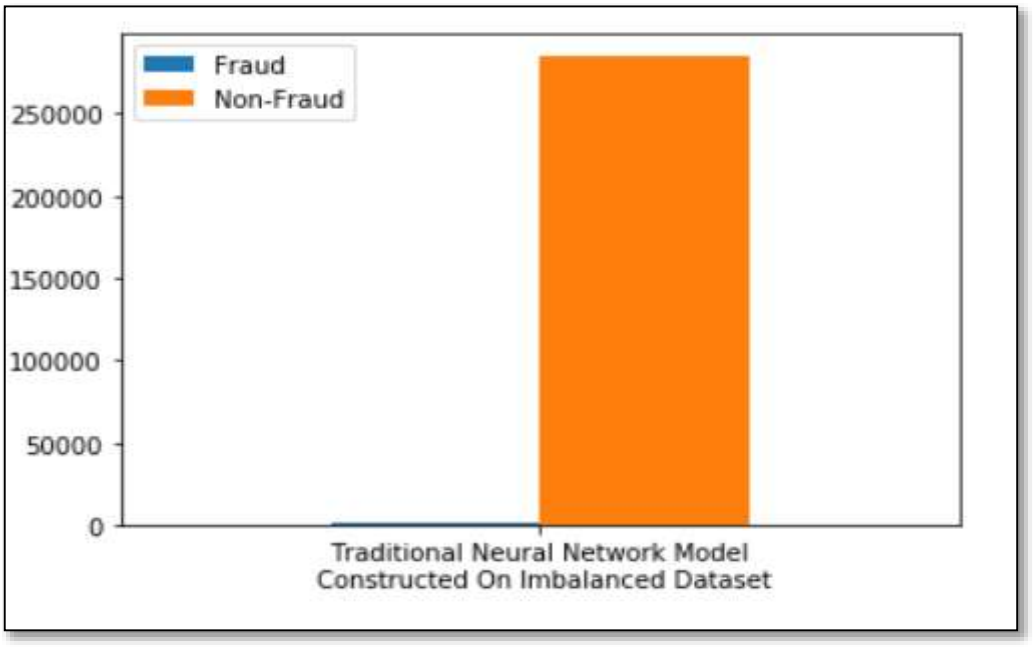


Figure 4.6 Imbalance Dataset -Experiment I

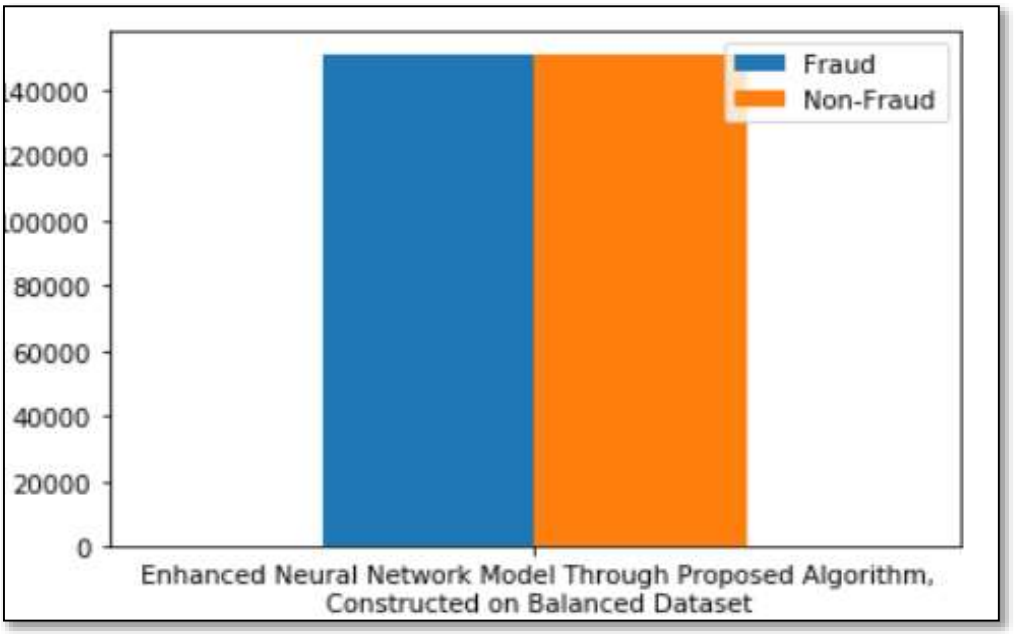


Figure 4.7: Balance Sample Dataset with the application of HPRT - Experiment II

4.2.1 Artificial Neural Network: An Introduction

The artificial neural network is an ML approach inspired by the brain processing system and helps in the giving out of a large amount of data. It acts just like human brains neural structure. The human brain is a collection of neurons where each neuron is connected to form a huge and complex set of interconnected neurons network. These interconnected neurons behave like switches and change their state to produce an output when they receive input signals of a defined threshold. In the complex architecture of neural network where many neurons are connected together input of one neuron can act as the output of other and again behaves as input for many other neurons and many neurons are activated again and again to learn different rules to produce output.

In 1943, Warren Mc Culoch and Walter Pits [107] published a research paper to narrate working principal of neurons for the first time. In the paper, they describe a neural network based model proposed by them. The model was self-possessed with an electronic circuit which can formulate meaning and determine patterns from complex dataset not possible for the human brain. ANN does not follow the traditional prototype to solve the given problem where a set of instruction is compulsory. ANN is smart enough to figure out output for the issue stated by learning given examples on its own through experiences. ANN algorithm given a set of examples can construct a model for prediction of future data based on the previous examples

but the challenge here is to choose the training examples carefully because model depends on the quality of examples provided as input for the algorithm.

Researcher's shows that human brain gathers information as patterns which can recognize some individual faces [108]. This technique is successfully incorporated by the field of computer technology and is known as Artificial Neural Network. ANN is a computing technique inspired by the biological process of the human brain and hence it is one of the advanced technologies of the modern era. The traditional computing system is not capable to identify complicated pattern easily and overall they cannot generalize a method for capturing pattern and use them after a period of time for detection of the future pattern.

The artificial neural network consists of many simple units which are connected together to form an immensely complex network helps in processing complicated big dataset. Each node in a network is a simple unit which either sends the signal or gets activated through input signal from other nodes in a network and hence works collectively to solve a specific problem. Nowadays this technology is used for solving many real-world complex problems such as speech recognition, image recognition, face recognition, etc. ANN can be used for both supervised and unsupervised training. In supervised learning input variable with their labels are used during ANN network training process, so that after training when the model is exposed to future data, the desired output is generated for a given input depending on examples. The learning in the ANN network occurs with weight adjustment. In the absence of a target variable that ANN is expected to learn on its own, through an unsupervised technique.

4.2.2 Artificial Neural Network: Architecture

ANN mimics biological neurons, therefore neurons present in ANN network works just as neurons of the brain [109]. ANN neurons get activated by the activation function, once an input received exceeds a certain value and change its state to produce the output, i.e. from 0 to 1 or 1 to 0. There are many activation functions used in a neural net. The most popular type activation function is Sigmoid, Tanh and Relu (Rectified Linear Unit).ANN architecture is connected with several neurons where the output of one neuron become input to other. In this hierarchical and complex architecture, each network can be represented as nodes. Node is a center which takes the weighted inputs, calculated them and sends to the activation function.

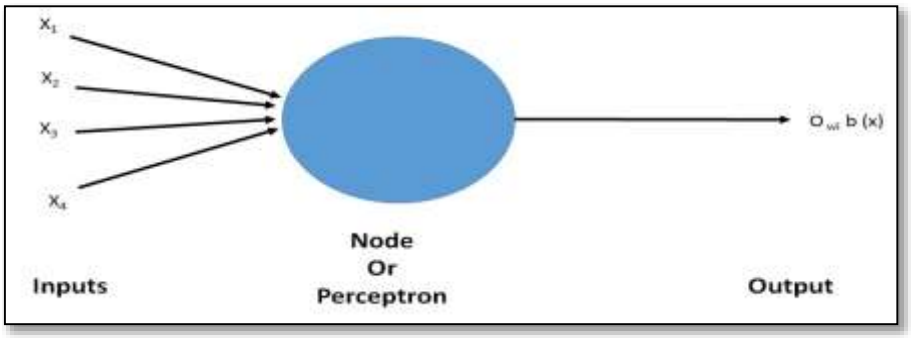


Figure 4.8: Node / Perceptron

In the figure 4.8, W_i is a weight in binary number multiplied by input and added up in the node as shown in eq. 4.3. B is a bias helps to change the state of the output.

$$X_1W_1 + X_2W_2 + X3W_3 + X4W_4 + b \quad \text{.....(4.3)}$$

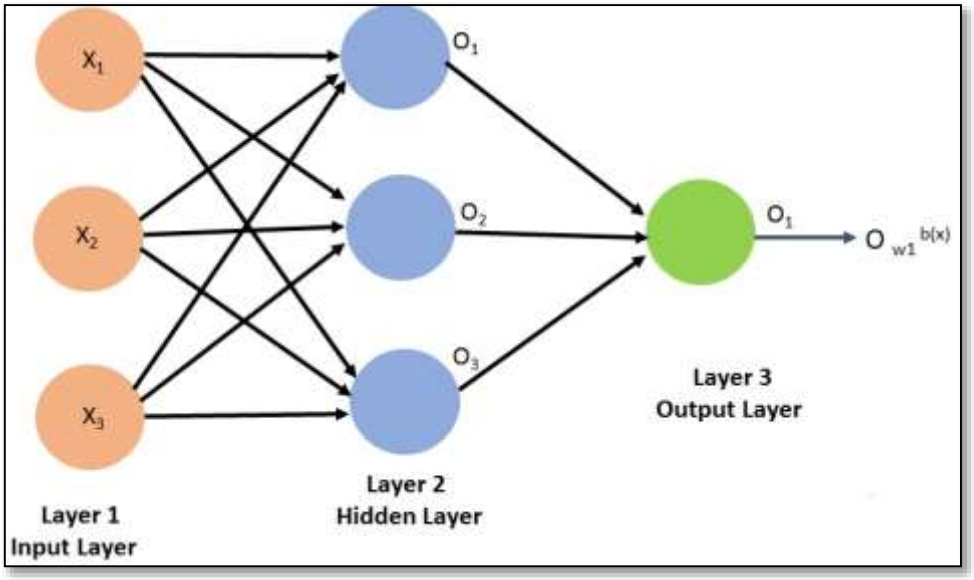


Figure 4.9: Neural Network Three-Layer Architecture

Figure 4.9 is a simple architecture of neural network containing an input layer, an output layer and a hidden layer [109].In layer 1 each node is connected with hidden layer nodes and the hidden layer is connected to the output layer. All the connection of ANN architecture has weights associated with it. A neural network receives input through the input layer, hidden layer process those input with the help of numerous neurons and then reflect the result to the output layer. This process is known as forwarding propagation. But every neuron contributes some error with an output which results in variation in desired output. To search for the seat of maximum error, neuron again travels back through the network and minimize the error by adjusting its weight. This is known as backward propagation. For minimize, error neural

network uses some optimization algorithm, such as gradient descent to complete the optimization task in an efficient way.

Perceptron is a basic central unit of the neural network responsible for the processing of numerous inputs and generates signal output [110]. An output of a neuron is generated by calculating weighted input of all the neuron of a layer which is performed in a network through activation function.

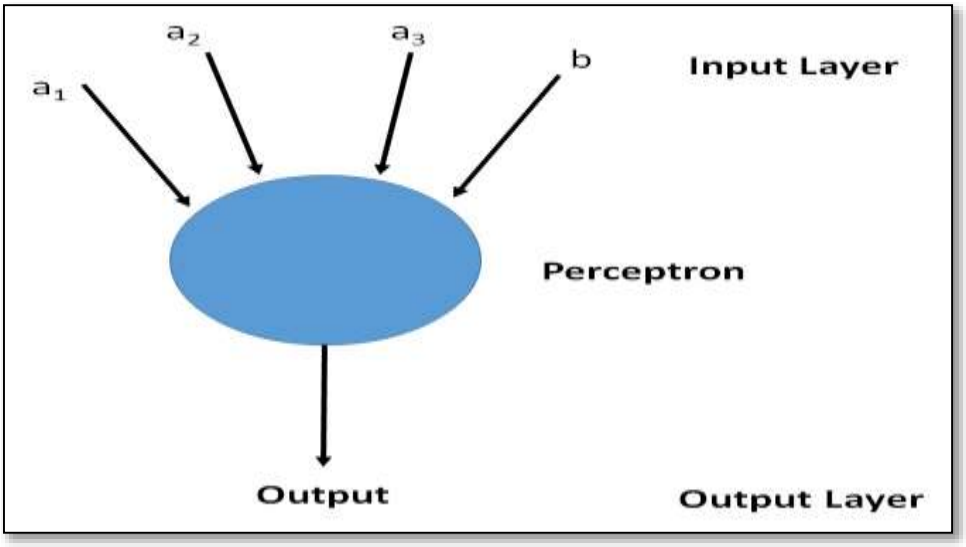


Figure 4.10: Perceptron with Bias

In the figure 4.10, a simple preceptor with three input and bias is shown Threshold is set in order to calculate an output. In a case where the threshold is 0, activation function will generate an output by calculating input values as shown in eq. 4.4

$$w_1 * a_1 + w_2 * a_2 + w_3 * a_3 + b \quad \text{.....(4.4)}$$

Where w_1, w_2, w_3 is a weight of input a_1, a_2, a_3 and b is a bias, can be written as eq. 4.5

$$f(x) = \int \sum_{i=0}^{n} w_i a_i + b \quad \text{.....(4.5)}$$

In eq. 4.5 b is a bias which is very important co-efficient for providing flexibility to perceptron. It provides flexibility to the calculated value to moves up and down, without b prediction will fit poorer line as the line will pass through the origin always.

We already discussed simple neural network consist of input and output layer and a single hidden layer but for most of the real world complex problem, the simple neural network

is not sufficient. In such cases, we have to look forward to MLP, which is comprised of a number of hidden layers in between input and output layer depending upon the complexity of the problem.

4.2.3 Stochastic Gradient Descent: An Optimizer

We use Stochastic Gradient Descent as an optimizer [111] to reduce the error of the predicted outcome of the neural network model using forward propagation. Error or loss in an output is determined by traveling from output to back into the neural network through backward propagation. This step is continued several times to minimize the error depending upon the error size. A complex round of forwarding and backward propagation is termed an epoch. Stochastic gradient descent algorithms help optimization of error in neural network model built for the huge dataset. Let us consider objective function as eq. 4.6.

$$f(h) = \frac{1}{N}\sum_{i=1}^{N} f_i\,(h) \qquad \text{.....(4.6)}$$

In eq. 4.6 , $f_i(h)$ is an error or loss function for a training set of N size, where N is too large for which computation cost can too high. Stochastic gradient descent provides a solution for such huge dataset, that rather than using the whole gradient, it uses randomly created I number of small samples and calculate $\nabla f_i(h)$ as in eq. 4.7.

$$\nabla f_i(h) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i h \qquad \text{.....(4.7)}$$

To update h if λ is a learning rate where computation cost is o (|β|)…(4.8)

$$h \coloneqq x - \lambda\, \nabla \mathrm{f_i(h)} \qquad \text{.....(4.8)}$$

For large training samples, SGD is a powerful enough to find solution in a few iteration. SGD minimizes the error in each iteration, where each of the iteration is consist of three steps:

1. Calculate the function closer to the output value as shown in eq. 4.9

$$f\,(h_{i)=}\ \textstyle\sum_{i=0}^{N} \Omega_j f\left(h_j\right) \qquad \text{.....(4.9)}$$

2. Calculate error rate from difference between actual value and f(h) as shown in eq. 4.10

$$\Omega_i = y_i - f(h)_i \quad \text{.....(4.10)}$$

3. Error is minimized in each iteration by adjusting the co-efficient of variable as shown in eq. 4.11

$$\Omega_i = \Omega_j + \delta(\in_i)h_{ij} \quad \text{.....(4.11)}$$

δ Is a learning rate here and it should be very small for large dataset. Batch methods such as BFGS use entire data to calculate the next update which can be very slow and costly for a large dataset in the single machine but stochastic gradient descent overcome this challenge and can provide fast convergence.

We construct an MLP model for prediction of frauds and non-frauds of credit card transaction dataset. It is a large imbalance dataset consisting of approx. 3 billion records. Here MLP is consist of 1 input layer containing all the features of the dataset as node and bias, 1 hidden layer of 32 nodes and 1 output layer with 2 nodes, each for 1 possible output. Before neural network architecture is constructed proposed algorithm is applied in a dataset for reducing its complexity. Then MLP is used on a balanced dataset to train the model. The neural network is used here for building model is to predict both frauds and non-frauds accurately.

Neural network model training is started by initializing learning rate 0.01. Rectified Linear Units (RELU) activation function is used as a unit of processing for the neural net as a type of activation function matters a lot for accurate training. Now a day's most of the neural network and deep learning network are using RELU because it calculates the output in a simpler manner and it required less time to train large network.

Adam optimizer is used to calculate and then minimize the error [112]. Adaptive Moment Estimation optimization (ADAM) is a variation of stochastic gradient algorithm, requires less memory and first-order gradient only. Adams calculate learning rate individually for each parameter from estimated moments of 1st and 2nd gradients. We use 5 epochs during optimization.

Keras is software uses for construction of neural net architecture, for calculating output, for minimizing errors and for evaluation of test prediction. Keras is a high-level software application written in python. Keras is composed of deep learning library and API to be used. Tenserflow, Theone, CNTK act as a base for Keras. It gives a facility to convert an idea into

result in very short time by focusing on fast experimentation. It is an efficient computational library for performing the numeric calculation. It helps to construct and train neural network model with very less line of code.

4.2.4 Neural Network Application for Big and Imbalanced Data

Big data is a multidimensional huge volume dataset presenting a complex structure which carries many difficulties while storing, analyzing and processing it for futuristic decision making. In such a situation where traditional approaches fail to process data, machine learning and artificial intelligence provide a way of training machines for processing big data and making it ready for prediction and classification. Lots of machine learning techniques are used for extracting information from big data, among them neural network is most common because of its robustness. Here training and classification of a dataset using different current technology and the use of trained data for diverse areas, such as healthcare, financial and different business has been discussed [88]. Further, several neural network technologies for the detection and prediction of cancer in the early stages through inferring pattern between the input and output of a dataset has been described.

The replacement of huge information processor with that of the neural network [60] has been marked. Big data is enormous to not preferable handling with normal tools; therefore the neural system is used here for processing of massive dataset and for the identification of thyroid patient through back propagation. Here, NN architecture is trained through feed forward using gradient descent optimizer and produces a good result. Therefore neural network has been recommended as upcoming tools for big data analytics.

With an escalating pace as data is changing with an equivalent tempo business is also changing at current circumstances and the businesses are bound to choose latest technologies, for there rising market value as well as profit [92]. In such cases using big data and data analytics can help in extracting hidden facts and upbringing automation in the current system. An artificial neural network is recommended here for processing large scale big business data and converting manual services into automated tools, thus securing business for a long period of time. An artificial neural network is just like neurons inside our body. Artificial neural network when applied can fulfill many aims. This paper articulated challenges caused while searching for reliable and fast machine learning-based tools and techniques for big data analytics [91]. Latest developed tools and techniques has also discussed for the said purpose. Authors suggested that with the growth of complexity and volume in a data, the requirements of intelligent computational techniques are required.

4.2.5 Experiment II: HPRT based Neural Network Model

Credit card transaction dataset used during collaboration of machine learning group is considered here in this research. Dataset consists 2 days credit card transactions of European card holder (2013). It is highly imbalanced dataset having 0.17% frauds only and 30 features. Much information about the dataset is not provided due to the security reason. The intention here is to balance the dataset and then built machine learning model to catch the fraudulent so that increasing rate of frauds can be minimized by extracting necessary pattern used during the fraud.

We download dataset and apply our proposed preprocessing enhanced algorithm for resampling and balancing the target class of dataset. Our enhance algorithm is a machine learning technique and applicable for any imbalance dataset as it automatically reduces the complexity of an imbalanced dataset. Our algorithm is designed to overcome the drawback of random oversampling and SMOTE oversampling. Both the process is efficient having some drawbacks but our algorithm prevails over these challenges. An algorithm is an enhanced resampling technique, which works stepwise; In the first step it reduces features of majority class by eliminating redundancy and outliers and in the second step increases minority class by adding synthetic feature using SMOTE, such that both minority and majority classes form normal distribution balance dataset automatically. This algorithm can be used be for large as well as big dataset as we have included PCA for dimensionality reduction and K-Means algorithm for reducing comparison. Divide and Conquer is used for reducing redundancy and Tukey's rule is applied for detecting outliers and dropping outliers.

MLP constructs a model for a balance credit card transaction dataset. We use Keras for training neural network architecture. Nodes and layers are added in a Keras through the creation of the sequential model. We use three layers fully connected neural network architecture. A number of layers can vary from problem to problem. Feed forward and back propagation technique is used for training purpose. Learning rate (0.1) is kept as learning rate because for high learning rate time consumed for training increases while accuracy decreases. In a model architecture number of nodes is the same as the number of features in a dataset. For the hidden layer we increase the node by 2, for providing flexibility to a model but it can vary from analyst to analyst. In this study, we get a satisfactory result by adding one hidden layer with 32 nodes, which can again vary depending upon the problem and quality of its solution. Initially, we assigned weights to all the input nodes along with bias by using a random number generator. The training set is fed as input to the input layer of the neural network. The output is passed

with its weight from the input to hidden and then to the output layer. At the output, layer error is calculated by Adam Stochastic Gradient Descent algorithm and back propagated again into the input layer in case of error arises. The step is continued until the error is minimized. This step can be explicitly provided to the architecture by setting a number of epochs. Epoch is a number of iteration for which the training process will continue, here we use 5 epochs i.e. 5 cycles of feed forward and back propagation will happen to search and minimize the error thereafter reflects final result. Rectifier (RELU) is used as an activation function of input and hidden layer. Softmax calculates probability which ranges from 0 to 1, therefore, we use softmax as an activation function of output layer because it produces result ranges only from 0 to 1, which is easy in mapping with target class 0 or class 1 depending upon the default threshold which is 0.5 for this experiment.

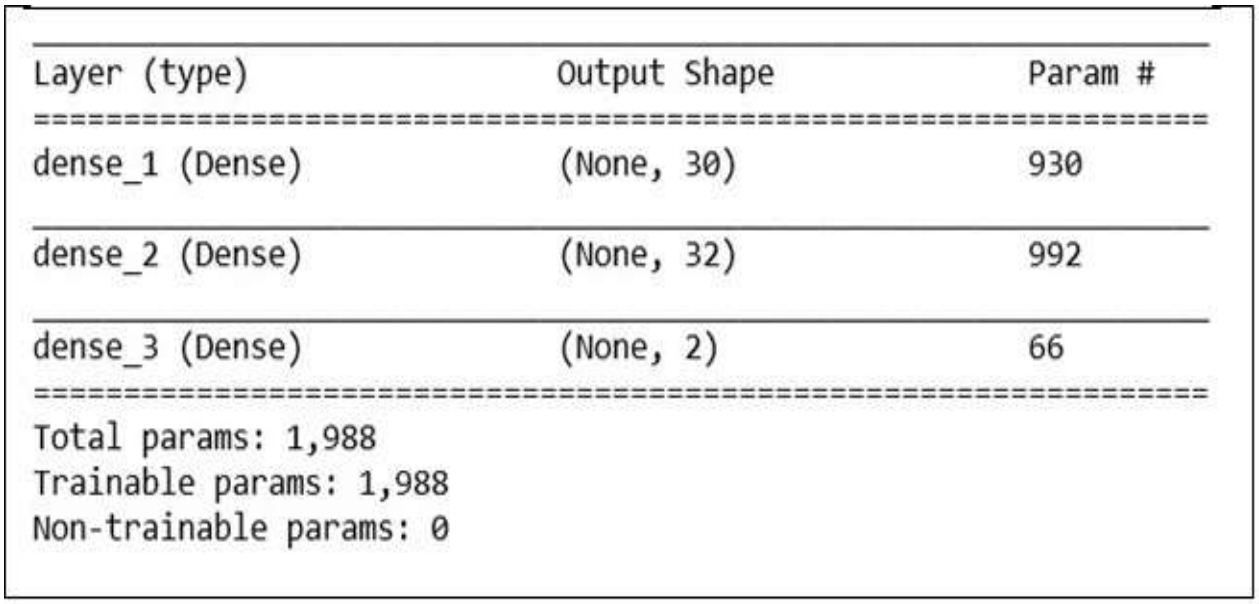


Figure 4.11: Designing of Input, Hidden and Output Layer for HPRT based Neural Network

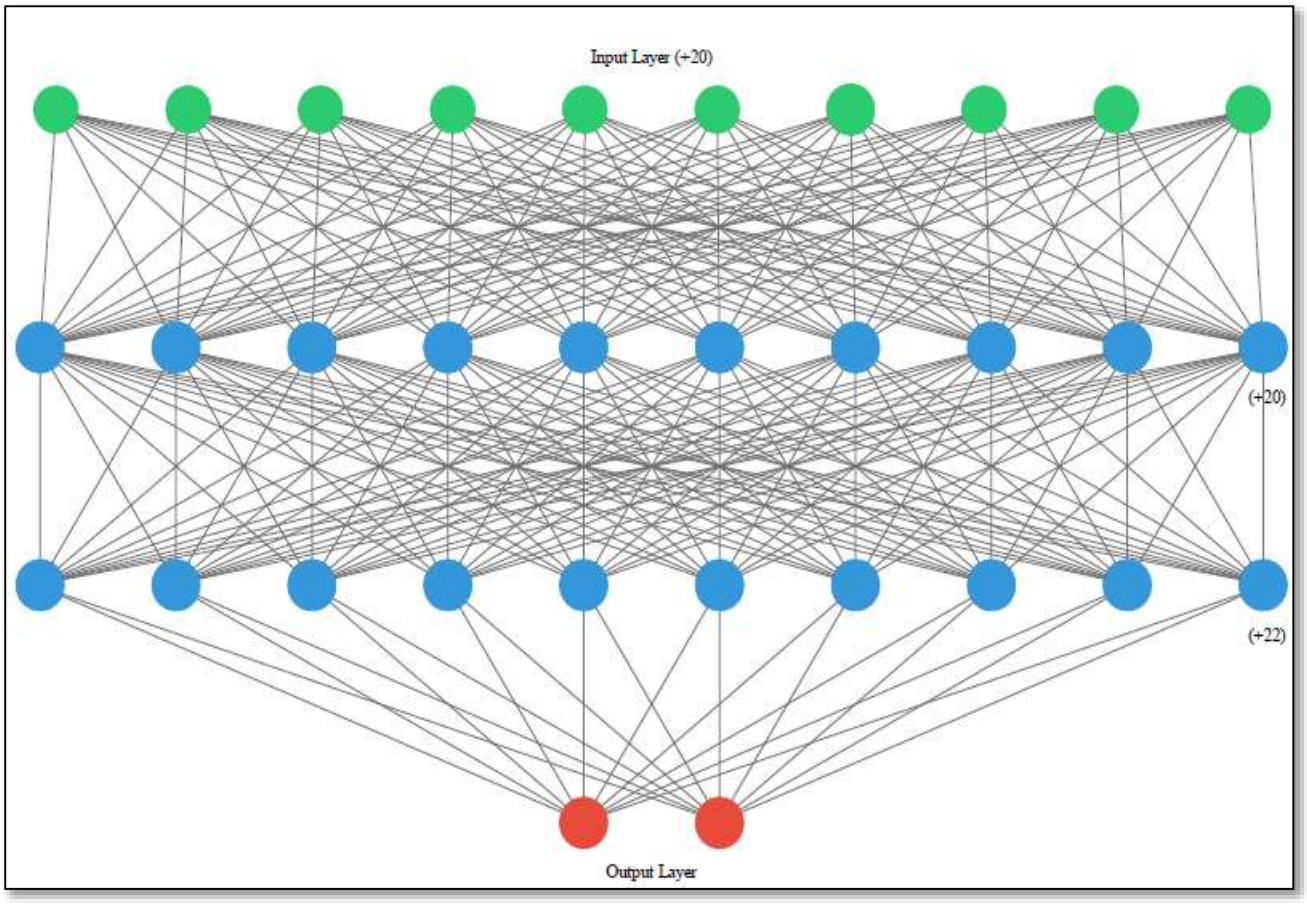


Figure 4.12: HPRT based Neural Network Model

After processing of neurons, we must compile the model to calculate the loss. During compilation, we use sparse categorical cross-entropy for loss calculation and Adam optimizer for loss minimization. After compilation, we will check the model on the same dataset to determine the quality of the model. Here we will use epochs (no. of iteration) for improving the quality of our model by reducing errors. After training the neural network model, the performance of the model is evaluated by testing it with test data. This will provide us an overview that how precisely our model captures the training rule. Further in future if real life unseen data is provided to the model, will predict output based on the training provided earlier.

```
Train on 243980 samples, validate on 60996 samples
Epoch 1/5
 - 4s - loss: 0.0192 - acc: 0.9928 - val_loss: 0.0036 - val_acc: 1.0000
Epoch 2/5
 - 3s - loss: 0.0036 - acc: 0.9989 - val_loss: 6.0133e-04 - val_acc: 1.0000
Epoch 3/5
 - 3s - loss: 0.0032 - acc: 0.9990 - val_loss: 0.0122 - val_acc: 0.9979
Epoch 4/5
 - 4s - loss: 0.0031 - acc: 0.9990 - val_loss: 0.0017 - val_acc: 0.9997
Epoch 5/5
 - 3s - loss: 0.0020 - acc: 0.9994 - val_loss: 6.8450e-04 - val_acc: 1.0000
```

Figure 4.13: Loss and Error Calculation during each epoch

4.2.6 HPRT based Neural Network Algorithm

Input: HPRT applied dataset C

Learning Rate 0.01

Output: Neural Network Model

Method: Initialization of weights and biases for nodes of the input layer using random numbers.

While condition for termination not get satisfied

{

For each Feature T in C

{

Feed the training features to the input layer

Note: Feed Forward Propagation

For each node of the input layer I unit x propagation

{

$I_x = T_x$

For each node of hidden layer H at unit x

{

H_x = sum ($W_{Yx}I_x$ + αx)

Note: α_x is a bias at H and W_{Yx}isa weighted input at layer H_x to previous layer y.

}

For each node of output layer O at unit x

{

O_x = Computed output from hidden layer H_x

}

For each node of output layer O at unit x

E_x = A_x - O_x

Note: A_x is an actual value from the training set

For each node from the output layer towards the input layer

{

Checking for the neuron causing a maximum error during back propagation

}

For each weight in network

{

ΔW_{yx}= Adjust weight

Note: ΔW_{yx} is a weight of higher connection to previous in network

W_{Yx} = ΔWyx

Note: weight update

}

For each bias in network

{

$\Delta\alpha_{yx}$ = Adjust bias

Note: α_{yx} is a bias of higher connection to previous in network

$\alpha_{yx} = \Delta\alpha_{yx}$

Note: bias update

}

}

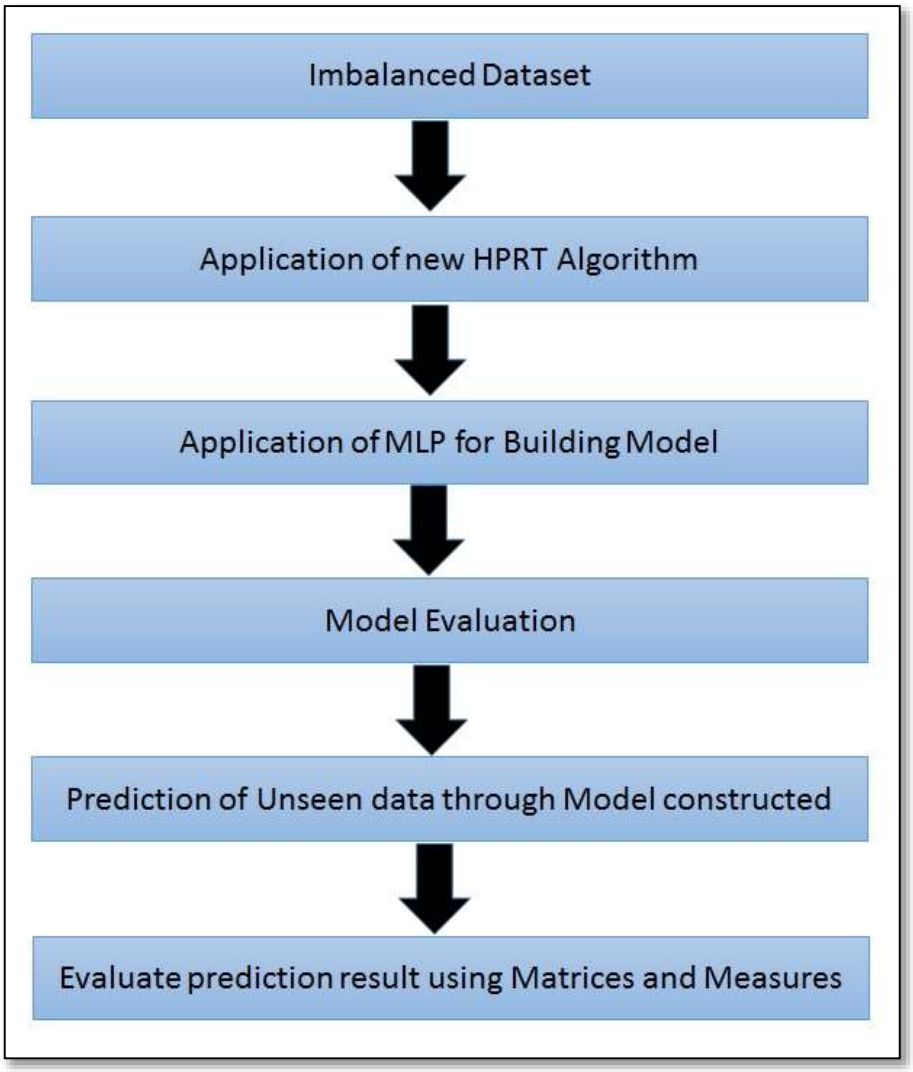


Figure 4.14: Steps for HPRT based Neural Network Model-Experiment II

4.2.7. Experiment II: Result

MLP model prediction for fraud transactions has been compared with the test set sample, which is a part of the original dataset, extracted from it and kept unseen at the beginning of the experiment. It produces a very good result for various matrices and measures. Accuracy (99%), precision (100%), Recall (85%), F1-Score (89%) results have been observed. According to confusion matrix out of 38,221 test samples, the model predicted, 38, 108-time

true negative, 87 times true positive, 15 times false positive and only 11 times false negative. Therefore we can conclude the model is precisely predicting frauds reflecting too small misclassification and can be neglected.

Table 4.2: Results of Neural Network Models during Experiment II

Matrices & Measures (%)	Traditional Neural Network Model	HPRT Based Neural Network Model
Accuracy	99%	99%
Precision	75%	100%
Recall	78%	85%
F1 – Score	39%	89%

Confusion Metrics		
Total Test Samples 38221	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

38112	11
60	38

38108	15
11	87

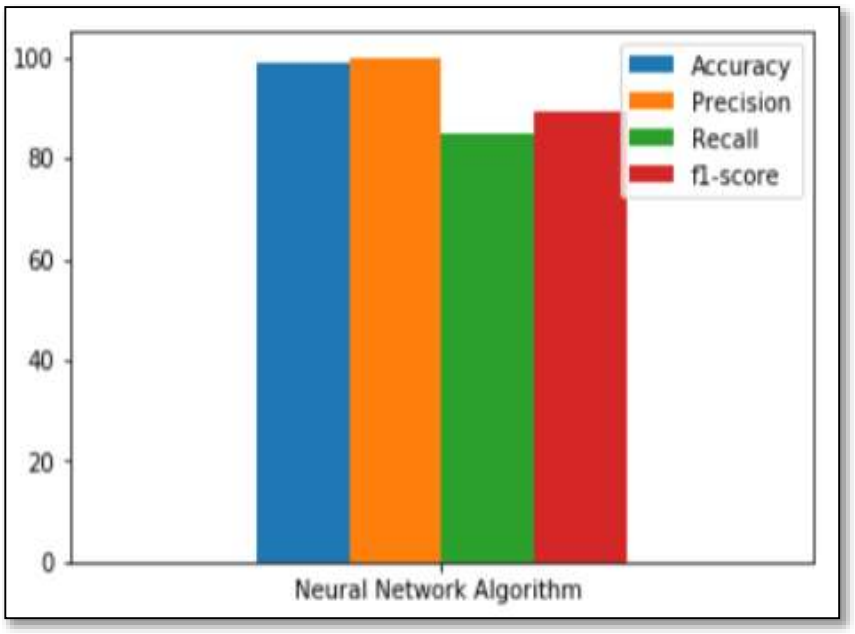


Figure 4.15: Performance of HPRT based Neural Network Model

4.2.8 Summary

An objective of this work is to apply and check the performance of the neural network to a highly imbalanced dataset. The neural network has been considered as the most robust and efficient machine learning algorithm. Its architecture can process a huge amount of big data in a very less time. Lots of research on neural network determined performance and accuracy provided by the neural network are very excellent as it is able to calculate and minimize the error during the time of processing itself. The neural network works well when the data provided to it is balanced containing target class of normal distribution. But for an imbalance, dataset neural network result is quite confusing, although accuracy (99%) is observed confusion matrix, exhibit lots of misclassification with very low F1-Score (39%). Therefore the traditional neural network algorithm does not fit for an imbalanced dataset. In the second experiment, we applied our proposed resampling algorithm on the dataset for reducing its complexity, then the performances of neural network algorithm also boost up bringing forth precise and accurate result, so we can conclude that our proposed algorithm act as a pre-processing tool, for a noisy and imbalance dataset, thus intensify the performance of machine learning classification algorithms. Finance industry could be benefited through this model by recognizing the fraudulent pattern and can alert in future to minimize losses happen through fraud transactions. This effort can save millions of dollars every year from bustling to the wrong person during fraud transactions.

The chapter starts with the introduction of neural networks, their use in different area, it extensive literature and description of MLP architecture, which is the most common and popular kind of neural net. Then chapter moves towards an experiment to show the result of an application of proposed resampling algorithm, on an imbalanced dataset. Result illustrates that application of the HPRT to dataset before the developments of the neural network model intensify its accuracy and decreases misclassification.

HPRT can utilize to any imbalance dataset having a lot of redundancy and outliers but the algorithm can be extended further for removing other noises from the dataset so that the algorithm turns into more generalized and fit for any dataset. Our enhanced HPRT based neural network model can utilize for prediction of credit card frauds for future data provided the dimensions of the dataset is same as that of the dataset used here. But the study can be elaborated in a direction to enhance the algorithm as well as a model so that it is widely accepted to any domains in a different area.

CHAPTER-5

Findings, Conclusion and Future Work

5.1 Findings

Present data management tools and techniques are either expensive or not suitable to accept challenges caused by the complex Bigdata. Therefore, data mining and machine learning is emerging technique, capable to extract decision making information from Bigdata and helps various industries using it to gain profit from those extracted ideas. During mining data, for revealing the intuitive power of prediction, huge amount of data example is required by machine learning algorithm, for which Bigdata is making sense. In this way, Bigdata and machine learning are helping each other.

One of the contribution of this study is to investigate the challenges of machine learning algorithm in extracting rules during classification problem. During the study lots of literature has been scrutinized and examined a number of factors affecting the performance of machine learning due to complex and messy dataset. Out of many such complexities, data imbalance is a focus area of this research. It has been observed in an experiment during this study, that if the dataset is an imbalance, machine learning techniques could not fabricate accurate result. Therefore, before feeding dataset to the machine learning algorithm pre-processing step should be applied to it, to reduce its complexity. Various disadvantages of current popular approaches of data sampling such as Under-Sampling and Over-Sampling has been premeditated and observed that in order to get benefits from Bigdata and machine learning algorithm those issue should be resolved. Therefore, HPRT, a new Hybridization Preprocessing and Resampling Technique, has been develop during this research, which solves several issues of messy data and automatically balances the dataset, thus reducing its complexity and enhancing the performance of ML classifier. Several outcomes of this research are as follow:

1. A new Hybridization Preprocessing and Resampling Technique (HPRT), act as a perfect single data management solution having a capability of cleaning the data simultaneously during optimizing complexity of an imbalanced dataset.

2. HRPT, acts as preprocessing algorithm, able to enhance the performances of several traditional machine learning algorithms, which has been proved through the result of several experiments performed during study.
3. HPRT based Neural Network model for detection of credit card fraud is an important outcome of this study which is capable to classify frauds and non-fraud transaction precisely and hence can rescue the financial industry from mammoth losses.
4. HRPT tested with several sample real time datasets such as credit card transaction dataset and breast cancer prediction dataset to achieve good result.
5. HRPT is designed in such a way that it can be integrated with several ML algorithms such as to construct a powerful model for solving several problem.
6. HPRT can utilize to any imbalance dataset having a lot of redundancy and outliers and the algorithm can be extended further for removing other noises from the dataset so that it turns into more generalized technique and fit for any dataset.
7. HPRT based Neural Network model can be a perfect solution for every domain specific problem having dataset imbalance and messy data challenge.

5.2 Conclusion

Machine learning is an amalgamation of several open-handed technologies that can adapt for the various problem-solving goals. Statistics analysis, mathematics, data mining, data analysis, deep learning, and artificial intelligence are some of the technology. The arrival of Bigdata helped the industry to expand the usage of machine learning techniques because old and traditional algorithms are not proficient to Bigdata. Vast and extravagant techniques at the back of machine learning are attracting good number of researchers to emphasize much on it to extracts new hidden facts from Bigdata, which can be beneficial for today's economy.

Big data has gained huge prominence in present market place. Data analysis becomes stiff during this Bigdata age because of its complex nature. As of today's scenario, data analytics becomes a major area for research. Data analytics processes to uncover the hidden pattern from the massive data useful for decision making. Now a day's machine learning is extensively used for extracting vital facts and figures from Bigdata almost every area to achieve diverse objectives. The growth of Bigdata also affected the finance domain, with its increases financial frauds. Credit cards or other financial cards fraud persuade magnificent hazard to the financial industry. Present techniques and tools are either expensive or unsuitable to accept challenges caused by the complex and big financial data. Data mining and machine learning is an emerging technique, capable to classify frauds and non-fraud transaction and hence can

rescue the financial industry from mammoth losses. During data mining for revealing the intuitive power of prediction huge amount of data example is required by machine learning algorithm, for which Bigdata is making sense. In this way, Big data and machine learning are helping each other. Machine learning and data mining usage are being utilized in every domain with the advent of Bigdata.

The major contribution of this study is to investigate the challenges of machine learning algorithm in extracting rules during classification problem. During the study lot of literature has been scrutinized to conclude a number of factors affecting the performance of machine learning, if a complex dataset is provided to it. Out of many such complexities, data imbalance is a focus area of this research. It has been observed in an experiment during the study, that if the dataset is an imbalance, machine learning techniques could not fabricate accurate result. In spite of showing high accuracy confusion matrix shows lots of misclassification. Data preparation is another activity which need to be perform in order to clean messy dataset before feeding it to the ml algorithms because unclean data is again a threat for ml techniques. Therefore, before constructing rule out of ml algorithm pre-processing step should be applied to it, to reduce its complexity.

HPRT, a machine learning based, Hybridization Preprocessing and Resampling Technique, is design and implemented during this research, to automatically balance the dataset while cleaning it and thus reducing its complexity in several steps as a result performance of ML classifier enhances. HPRT receives unbalanced and messy dataset and in the first steps original dataset is divided into two separate dataset where one contains only positive (minority) class and other contains only negative (majority) class. PCA is applied to reduce the dimensionality of both the dataset, as a result of which irrelevant features are discarded from it. Then, In next step K-Means algorithm is applied to a class consisting majority sample for redundancy removal by using divide and conquer rule and outliers is removed by using tukey's methods. At this stage dataset is free from redundant and outliers record thus discarding unwanted sample and cleaning the dataset while reducing majority sample. Then at next step SMOTE oversampling is applied to minority class to balance the dataset. Several experiment using HPRT with number of machine learning algorithm is tested on imbalance European credit card dataset selected as a case study to detect frauds. Experiment result observed positive impact of the HPRT on different machine learning classification algorithms like KNN, SVM and LG.

Lastly, our enhanced MLP model using three layers neural architecture combined with HPRT algorithm applied on imbalanced credit card transaction dataset, achieve promising

result when compares traditional neural network model. Therefore we can conclude that the purposed algorithm enhances the performance of the machine learning algorithm by reducing dataset complexity.

The solution provided here will provide a lot of benefits to credit card domain in detecting future frauds beforehand and thus will reduce the losses through fraud. A single solution is easy to understand and use and therefore can be proved as easily accepted tools among the users. HPRT implemented with big amount of data will solves numbers of issues related to it. The algorithm with little modification can be applied to various domain starting from health care to education. Thus a single solution corroborate to be a perfect data management tool to reduce lot of complexity of a dataset.

5.3 Future work

Research is a never ending process. Scope of improvement is always there. Although HPRT algorithm experimented several time keeping different prospective in mind ,which produce very good result, but still there is a chance of improvement where other cleaning parameter can also be consider.it can be applied to any imbalanced dataset of different domains with some limitations. Our algorithm is designed for reducing complexity by removing redundancy and dropping outliers, but other cleaning factor such as treating missing value, null values, etc. can also be considered which is missed during this study. Our MLP model shows very good result for detection of fraud using credit card dataset and can be used in future for fraud detection. This model can be extended with a little change in the hidden layer, input layer and output layer or inactivation function so that it can be used by any domain which is missed in this study. The number of nodes on each layer can be increased or decreased depending upon the problem complexity. This model can generalize to be used by any domain with no parameter left in my post doctorate which is left here as future work.

References

[1] Douy Lancy, 3d Data Management: Controlling Data Volume, Velocity and Variety, META Group, February 2001.

[2] Dan Vesset, et al. , Worldwide Big Data Technology and Services 2012 – 2015 Forecast, IDC Report, July 2014.

[3] S.Vijayarani and S.Sharmila, Research In Big Data – An Overview, Informatics Engineering, an International Journal (IEIJ), Vol. 4 (3), September 2016.

[4] Emmanuel Letouzé, Big Data for Development: Challenges & Opportunities, UN Global Pulse, May 2012.

[5] Live Vault, 2.5 Quintillion Bytes of Data Are Created Every Day, October 2015, Accessed: https://www.livevault.com/2-5-quintillion-bytes-of-data-are-created-every-day/.

[6] Roshani K. Chaudhari and Prof. D. M. Dakhane, Contribution of Hadoop to Big Data , International Journal of Advanced Research in Computer Science and Software Engineering, Problems, Vol. 5(4), April 2015.

[7] Mohammad Saeid Mahdavinejad, et al., Machine learning for internet of things data analysis: a survey, Digital Communication and Networks, Vol. 4 (3), August 2018.

[8] Pushpa Suri and Meenakshi Sharma, Maintenance Modification Algorithms and its Implementation on object oriented data warehouse. Global Journal of Computer Science and Technology, Vol. 11(12), July 2011.

[9] Xindong Wu, Data mining with big data, IEEE Transactions on Knowledge and Data Engineering, Vol. 26(1), January 2014.

[10] Mariette Awad and Rahul Khanna, Efficient Learning Machines, Springer, Vol. 2(1), April 2015.

[11] Hossin, M. and Sulaiman, A Review on Evaluation Metrics for Data Classification Evaluation, International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol. 5(2), March 2015.

[12] Akinsola et al. , Supervised Machine Learning Algorithms: Classification and Comparison, International Journal of Computer Trends and Technology (IJCTT), Vol. 8, June 2017.

[13] Vidushi Sharma et al., A Comprehensive Study of Artificial Neural Networks, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2(10), October 2012.

[14] Xinjian Guo et al., On the Class Imbalance Problem, Fourth International Conference on Natural Computation, October 2008.

[15] Ricardo Barandela et al., The Imbalanced Training Sample Problem: Under or over Sampling, The International Association for Pattern Recognition, Vol. 3138, August 2004.

[16] Jingke Xi, Outlier detection algorithm in data mining, 2008 Second International Symposium on Intelligent Information Technology Application, December 2008.

[17] Prity Vijay, Bright Keshwani, Emergence of Big Data with Hadoop: A Review, IOSR Journal of Engineering (IOSRJEN), Vol. 6(3), March 2016.

[18] S.B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Emerging Artificial Intelligence Applications in Computer Engineering, Vol. 2(7), 2007.

[19] Fabian Pedregosa et al., Fabian Pedregosa Scikit-learn: Machine Learning in Python Journal of Machine Learning Research 12, January 2012.

[20] Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, IEEE, August 2011.

[21] Brian McKenna, Doug Cutting, 'father' of Hadoop, talks about big data tech evolution, June, 2018, Accessed: https://www.computerweekly.com/news/450420002/Doug-Cutting-father-of-Hadoop-talks-about-big-data-tech-evolution.

[22] Abdelrahman Elsayed et al., MapReduce: State-of-the-Art and Research Directions, International Journal of Computer and Electrical Engineering, Vol. 6(1), February 2014.

[23] Vidyasagar S. D, A Study on "Role of Hadoop in Information Technology era, Vol. 2(2), February 2013.

[24] Shilpa and Kaur M, Big Data and Methodology-A review, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3(10), October 2013.

[25] Jens Dittrich and Jorge Arnulfo Quiane-Ruiz, Efficient Big Data Processing in Hadoop MapReduce, Information Systems Group Saarland University, October 2014.

[26] B. Arputhamary and L. Arockiam, A Review on Big Data Integration. International Journal of Computer Applications, Advanced Computing and Communication Techniques for High Performance Applications, Vol. 2(2), March 2014.

[27] Ashish Thusoo et al. , Facebook Data Infrastructure Team, Hive–A Petabyte Scale Data Warehouse Using Hadoop, 2009.

[28] Alan F. Gates et al., Building a High-Level Dataflow System on top of Map-Reduce: The Pig Experience, Yahoo Team, VLDB, 2009.

[29] Clark Bradley et al., Data Modeling Considerations in Hadoop and Hive, Accessed: https://support.sas.com/resources/papers/data-modeling-hadoop.pdf.

[30] Bo Li, Survey of Recent Research Progress and Issues in Big Data, 2013, Accessed: http://www.cse.wustl.edu/~jain/cse570-13/ftp/bigdata2.pdf.

[31] Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, Vol. 8, February 2006.

[32] S. B. Kotsiantis, et al., Data Preprocessing for Supervised Leaning, International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol. 1(12), March 2007.

[33] Zena M. Hira and Duncan F. Gillies, A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Hindawi Publishing Corporation Advances in Bioinformatics, Vol.2, Aprail 2015.

[34] Prity Vijay, Bright Keshwani, A Study on Big Data Analytics through R, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4(8), August 2016.

[35] Nick F Ryman-Tubb and Artur d'Avila Garcez, SOAR — Sparse Oracle-based Adaptive Rule extraction: Knowledge extraction from large-scale datasets to detect credit card fraud, The 2010 International Joint Conference on Neural Networks (IJCNN), October 2010.

[36] Saravanan Sagadevan et al., Credit Card Fraud Detection Using Machine Learning As Data Mining Technique, Journal of Telecommunication and Computer Engineering (JTEC), Vol. 10, August 2018.

[37] Wei Feng et al., Class Imbalance Ensemble Learning Based on the Margin Theory, Appl. Sci. 2018, 8(5), May 2018.

[38] Satyam Maheshwari, et al., A Review on Class Imbalance Problem: Analysis and Potential Solutions, IJCSI International Journal of Computer Science Issues, Vol. 14 (6), November 2017.

[39] Zhuoyuan Zheng et al., Oversampling Method for Imbalanced Classification, Computing and Informatics, Vol. 34, October 2015.

[40] Satyam Maheshwari et. al, A New approach for Classification of Highly Imbalanced Datasets using Evolutionary Algorithms, International Journal of Scientific & Engineering Research, Vol. 2 (7), July 2011.

[41] Mikel Galar, et al., A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting and Hybrid-Based Approaches, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 42 (4), July 2012.

[42] N. V. Chawla et al., SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, Vol. 16, June 2002.

[43] Rheza Harliman and Kaoru Uchida, Data- and Algorithm-Hybrid Approach for Imbalanced Data Problems in Deep Neural Network, International Journal of Machine Learning and Computing, Vol. 8 (3), June 2018.

[44] M. Mostafizur Rahman and D. N. Davis, Cluster Based Under-Sampling for Unbalanced Cardiovascular Data, Proceedings of the World Congress on Engineering 2013 Vol.3, July 2013.

[45] Tasadduq Imam et al., z-SVM: An SVM for Improved Classification of Imbalanced Data, Springer, June 2006.

[46] Yuxuan Li and Xiuzhen Zhang, Improving k Nearest Neighbor with Exemplar Generalization for Imbalanced Classification, Springer, Vol 4(5), August 2011.

[47] Siyang Zhang, Fangjun Kuang, A Novel SVM Model with PSO on Power Transformer Fault Diagnosis, Journal of Computational Information Systems, Vol. 8 (14), July 2012.

[48] Vaishali Ganganwar, An overview of classification algorithms for imbalanced datasets, International Journal of Emerging Technology and Advanced Engineering, Vol. 2 (4), April 2012.

[49] Keshav Dahal, et al., GA-based learning for rule identification in fuzzy neural networks, Elsevier , Vol. 35, October 2015.

[50] Radha R and Murlidhara S., Removal Of Redundant And Irrelevant Data From Training Datasets Using Speedy Feature Selection Method, International Journal of Computer Science and Mobile Computing (IJCSMC),Vol. 5(7), July 2016.

[51] Shuchu Han et al., Automatically Redundant Features Removal for Unsupervised Feature Selection via Sparse Feature Graph, Cornell University Library, Vol 8(5), May 2017.

[52] Qinbao Song, et al., A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data, IEEE, Vol. 25(1), Jan 2013.

[53] Annalisa Appice, et al., Redundant Feature Elimination for Multi-Class Problems, Machine Learning, Proceedings of the Twenty-first International Conference,Vol 7(6) January 2004.

[54] Bharati Kamble and Kanchan Doke, Outlier Detection Approaches in Data Mining, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Vol. 04 (3), March 2017.

[55] Tze Siong Lau et al., A Binning Approach to Quickest Change Detection with Unknown Post-Change Distribution, Cornell University Library, Vol 2, January 2018.

[56] Manju Kaushik and Bhawana Mathur, Comparative Study of K-Means and Hierarchical Clustering Techniques, International Journal of Software & Hardware Research in Engineering, Vol. 2(6), June 2014.

[57] Rishikesh Suryawanshi and Shubha Puthran, Novel Approach for Data Clustering using Improved K-means Algorithm, International Journal of Computer Applications, Vol. 142(12), May 2016.

[58] SK Ahammad Fahad and Md. Mahbub Alam, A Modified K-Means Algorithm for Big Data Clustering. IJCSET, Vol. 6 (4), April 2016.

[59] Ankita Sinha and Prasanta K. Jana, A Novel K-Means based Clustering Algorithm for Big Data, International Conference on Advances in Computing, Communications, IEEE, Vol 4(5), September 2016.

[60] S.Sapna Fusion of Big Data and Neural Networks for Predicting Thyroid. 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques, IEEE, June 2017.

[61] Aman Gulati et al., Credit card fraud detection using neural network and geolocation. 14th ICSET, Vol 7, 2017.

[62] Raghavendra Patidar and Lokesh Sharma, Credit Card Fraud Detection Using Neural Network, International Journal of Soft Computing and Engineering (IJSCE), Vol. 1, June 2011.

[63] Massimiliano Zanin, et al., Credit Card Fraud Detection through Parenclitic Network Analysis, Vol. 3(4) , January 2018

[64] Vikash Sharma, et al., Importance of Big Data in financial fraud detection. Int. Automation and Logistics, Vol. 2, June 2016.

[65] Navanshu Khare and Saad Yunus Sait , Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models, International Journal of Pure and Applied Mathematics, Vol. 118, September 2016.

[66] WorldPay, Global payments report preview: your definitive guide to the world of online payments, September 2016.

[67] Furnkranz et al, Foundations of Rule Learning, Springer, Vol 6, 2012.

[68] Hossin and Sulaiman, A Review on Evaluation Metrics for Data Classification Evaluations, International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol. 5, March 2015.

[69] S.Deviranjitham and S.Thamilarasan, A Study on Usage and Satisfaction of Credit Cards by Customers in Krishnagiri District , International Journal of Business and Administration Research Review, Vol. 2(4), March 2014.

[70] Mike Finlay, Using Indicators and Internal Data to Forecast Fraud, Association of Certified Fraud Examiners (ACFE), 2012.

[71] K.R. Seeja and Masoumeh Zareapoor, FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining, The Scientific World Journal, Vol. 2014, September 2015.

[72] Maria R. Lepoivre et al., Credit Card Fraud Detection with Unsupervised Algorithms, Journal of Advances in Information Technology, Vol. 7(1), February 2016.

[73] Prity Vijay, Bright Keswani, Support Vector Machine (SVM) Kernels based approach for detection of Breast Cancer,CASS, Vol. 2(2), November 2018.

[74] Pedro Domingos, A Few Useful Things to Know About Machine Learning, Communications of the ACM, Vol. 55(10), October 2012.

[75] Nitish Srivastava, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, Vol. 4(5),September 2014.

[76] Ong Shu Yee,Credit Card Fraud Detection Using Machine Learning ss Data Mining Technique, Journal of Telecommunication, Electronic and Computer Engineering (JTEC), Vol. 10,February 2016.

[77] Xiaolong Jin et. al, Significance and Challenges of Big Data Research, Elsevier, Vol. 2(2), June 2015.

[78] Dajana Barbic et al., Logistic regression analysis of financial literacy implications for retirement planning in Croatia, CRORR Journal, Vol. 7, December 2016.

[79] Durgesh K. Srivastava et al., Data Classification Using Support Vector Machine, Journal of Theoretical and Applied Information Technology, 2009.

[80] Xiaopeng Yu and Xiaogaoyu, The Research on an Adaptive k-Nearest Neighbors Classifier, IEEE International Conference on Cognitive Informatics, Vol. 5(8), July 2006.

[81] Viashalu Ganganwar, An overview of classification algorithms for imbalanced datasets, International Journal of Emerging Technology and Advanced Engineering, Vol. 2(4), April 2012.

[82] K.Madasamy and M.Ramaswami, Data Imbalance and Classifiers: Impact and Solutions from a Big Data Perspective. International Journal of Computational Intelligence Research, Vol. 13, September 2017.

[83] Akila Somasundaram and U. Srinivasulu Reddy, Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data, International Conference on Research in Engineering, Computers and Technology (ICRECT 2016), January 2016.

[84] Nitesh V. Chawala, Data Mining for Imbalanced Datasets: An Overview, Springer, July 2017.

[85] Alexandra L'Heureux et al. , Machine Learning with Big Data: Challenges and Approaches, IEEE, Vol. 5, April 2017.

[86] Dunren Che et al., From Big Data to Big Data Mining: Challenges, Issues, and Opportunities, Springer, Vol 6(3), July 2013.

[87] Alberto Fernández, et al., An insight into imbalanced Big Data classification: outcomes and challenges, Springer, Vol. 3 (2), March 2017.

[88] L.Álvarez Menéndez, et al., Big data analytics using a neural network for earlier cancer detection, Vol. 52(8), Elsevier, October 2010.

[89] M. Mostafizur Rahman and D. N. Davis, Addressing the class imbalance problem in medical datasets, International Journal of Machine Learning and Computing, Vol. 3(2), April 2013.

[90] Arpit Singh and Anuradha Purohit, A Survey on Methods for Solving Data Imbalance Problem for Classification, International Journal of Computer Applications, Vol. 127 (15), October 2015.

[91] Fuchun Sun, et al., Efficient and rapid machine learning algorithms for big data and dynamic varying system. IEEE, Vol. 47 (10), October 2017.

[92] Ronald Davis, BIG DATA and Neural Networks & Deep Learning: 2 Manuscripts Paperback, April 2017.

[93] Hui Han1,et al., Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data

[94] Bhaskar N. Patel et al., Efficient Classification of Data Using Decision Tree, International Journal of Data Mining, Vol. 2, March 2012.

[95] B.S.Mounika Yadav and SeshaBhargavi Velagaleti, Challenges in Handling Imbalanced Big Data: A Survey, International Journal Of Current Engineering And Scientific Research (IJCESR), Vol. 5 (3), 2018.

[96] Piyasak Jeatrakul et al., Data Cleaning for Classification Using Misclassification Analysis, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 14(3), 2010.

[97] Jeannie Fitzgerald and Conor Ryan, A Hybrid Approach to the Problem of Class Imbalance, International Conference on Soft Computing, Vol. 12(1), June 2013.

[98] Jianqing Fan, et al., Principal component analysis for big data, January 2018, Accessed: https://arxiv.org/pdf/1801.01602.pdf.

[99] Mugdha Jain, ChakradharVerma, Adapting k-means for Clustering in Big Data, International Journal of Computer Applications, Vol. 101, September 2014.

[100] Olga Kurasova et al., Strategies for big data clustering. IEEE, Vol. 5(7), January 2014.

[101] Ali Seyed Shirkhorshidi, et al., Big data clustering: Algorithm and challenges. Computational science and its applications - ICCSA 2014: 14th international conference Guimarães, Vol. 6(7), June 2014.

[102] D. Saidulu and R. Sasikala Data, Machine Learning and Statistical Approaches for Big Data: Issues, Challenges and Research Directions, International Journal of Applied Engineering Research, Vol. 12(21), 2017.

[103] ULB Machine Learning Group, Accessed: http://mlg.ulb.ac.be.

[104] Ruchi Sharma and Nidhi Gulati, Improving the Accuracy and Reducing the Redundancy in Data Mining, International Journal of Engineering Science and Computing, Vol. 6, May 2016.

[105] Xu Chu et al., Data Cleaning: Overview and Emerging Challenges, SIGMOD 16, June Vol. 3(4), October 2016.

[106] Yi l. Murphey, hong guo, lee a. Feldkamp, Neural Learning from Unbalanced Data. Kluwer Academic Publishers, Manufactured in the United States, Vol. 6(6), May 2004.

[107] Warren S. McCulloch and Walter Pitts, A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, Bulletin of Mathematical Biophysics, Vol. 5 (4), December 1943.

[108] S.B. Maind, P. Wankar, Research Paper on Basic of Artificial Neural Network. January 2014.

[109] Axay J Mehta, et al., A Multi-layer Artificial Neural Network Architecture Design for Load Forecasting in Power Systems. World Academy of Science, Engineering and Technology International Journal of Electrical and Computer Engineering, Vol. 5(2), July 2011.

[110] Perceptron, Wikipedia. Accessed: https://en.wikipedia.org/wiki/Perceptron.

[111] L´eonBottou, Large-Scale Machine Learning with Stochastic Gradient Descent. NEC Labs America, Princeton, Vol. 5 (4), December 2016.

[112] Diederik P. Kingma, Jimmy Lei Ba, Adam: A Method for Stochastic Optimization. Published as a conference paper at ICLR, Vol. 5(2), June 2015.