Master's Thesis

Data Mining and Analysis

of

Supervised & Unsupervised Learning Algorithms

Supervisor: Chong, Uipil

A thesis submitted to the Graduate Faculty, in partial fulfillment of the requirements for the Degree of
Master of Engineering, Graduate School of Industry

University of Ulsan, Republic of Korea

by

Yousaf Azeem

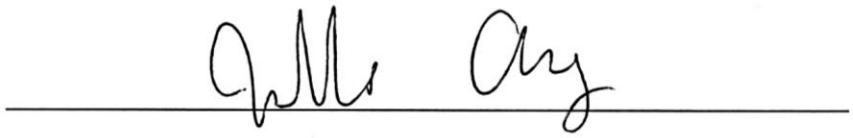
University of Ulsan, South Korea

May 2022

Data Mining through Supervised and

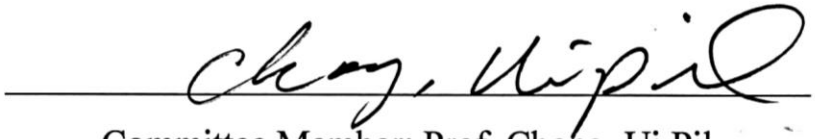
Unsupervised Learning Algorithms

This certifies that the dissertation of Yousaf Azeem is approved.



Committee Chair: Prof. Chung, Jin Ho

Committee Member: Prof. Park, Ju Chull



Committee Member: Prof. Chong, Ui Pil

Department of Global Smart IT Convergence

Graduate School of Industry

University of Ulsan,

Ulsan, Korea

July 2022

Abstract

Data mining is a process of investigating large pre-existing databases to gather new information. Data mining is a connection between computer science and statistics used to discover patterns in the data. The main objective of the data mining process is to mine the useful information from the data and formulate it into an understandable/logical structure for further use. The large data is sorted into sets to categorize patterns and create relationships to resolve problems through data analysis. Supervised (Classification) and unsupervised (Clustering) learning techniques are discussed in this research work.

Supervised machine learning is a method of machine learning. It involves allocating the specific data in such a way that a specific type of pattern or function can be extracted from that labeled data. Classification is defined as the function of learning in which provided data items are mapped into more than a few classes that are predefined.

Unsupervised techniques are essentially initiated from the sets of unlabeled data so, these are directly associated to figure out the unfamiliar properties in clusters. Clustering is a technique and process of unsupervised learning, used for the analysis of the statistical data exploited in several fields.

The analysis of Supervised (Classification) and Unsupervised (Clustering) learning techniques are based on accuracy and time studied in this research work. The Classification algorithms K Nearest Neighbor (KNN), Backpropagation (BP), Naïve Bayes, and Support Vector Machine (SVM) are compared by using different datasets through the testing tool Weka 3.8. The clustering algorithms K-means and Expectation-Maximization (EM) are also compared based upon accuracy and time by using Rapid miner and Weka 3.8 tools. The results show that the classification algorithm back-propagation performs with good accuracy as compared to the remaining classification algorithms. KNN performs timely executions as compared to other classification algorithms in supervised learning techniques. The clustering algorithm k-means shows good accuracy as compared to Expectation-Maximization (EM). K-means algorithm produces quality clusters as compared to Expectation-Maximization (EM).

ACKNOWLEDGMENTS

It is a pleasure to thank the people who made this thesis possible. It is my immense gratitude that I acknowledge the valuable guidance and encouragement of my supervisor, Prof. Ui-Pil Chong, whose kind encouragement, remarkable patience, timely advice, and gentle guidance made this possible. There aren't enough kind words to precise my appreciation for his involvement in my education and research.

Finally, I would like to thank my family and friends for the unbelievable amount of support and love shown me throughout my graduate school journey.

DECLARATION

I declare that I have written the Master's Thesis titled "**Data Mining and Analysis *of* Supervised & Unsupervised Learning Algorithms**", under the guidance of the advisor and other sources of data mentioned in the thesis and listed in the comprehensive appendix at the end of the thesis.

Table of Contents

List of Figures

List of Tables

List of Graphs

Chapter 1

Introduction & Background

1.1 Background

In the 1990s, the term "Data Mining" was introduced, early techniques for identifying patterns in data include the Bayes theorem (1700s), and the evolution of regression(1800s) [1]. The generation and growing power of computer science have boosted data collection, storage, and manipulation as data sets have wide in size and complexity level. Data mining is the process of applying computational methods to large amounts of data to reveal new important and relevant information. Data mining is used for finding patterns and exploring large data sets, building models that describe the important properties of data, and for making predictions based on the tested data [2]. The large data is sorted into sets to categorize patterns and create relationships to resolve problems through data analysis. Patterns are discovered in large datasets with the help of machine learning. The most important purpose and goal of data mining is information extraction from datasets to use this information for future use. Different approaches are used to explore data sets properties as Deep Learning from which Machine Learning is one. It is a data science sub-field that pays attention to designing algorithms that are discovered through the data and formulates calculations on it [3]. Furthermore, an advantage of the machine learning techniques is that it utilizes heuristic learning, arithmetical models, acquisitions of knowledge, and decision trees for making the decision. There is some misunderstanding on the meaning of data mining and its relation to machine learning and statistics. Data mining is a concept that places between information technology and statistics [4].

Supervised learning is a machine learning approach to labeled datasets. Supervise learning is designed to train or “supervise” algorithms to classify data or predict outcomes accurately. There are two types of supervised machine learning: classification and regression [5]:

- **Classification**: Classification is an algorithm to accurately assign specific categories of test data, such as splitting apples from oranges. In a real-world example, supervised learning algorithms can be used to categorize the spam folder in a separate folder from your inbox

folder. Linear classifiers, support vector machines (SVM), decision trees, and random forests are different types of classification algorithms.

- **Regression:** Regression is another supervised learning method that understands the relationship between dependent and independent variables. A regression algorithm is used for understanding the relationship. Regression models predict numerical values in different data points. Linear regression, logistic regression, and polynomial regression are the commonly used regression algorithms.

Unsupervised learning is a study to analyze and cluster unlabeled data sets. Unsupervised Learning algorithms find out hidden patterns in data without the need for human interference. That's why it is called Unsupervised learning [6].

For clustering, association, and dimensionality reduction Unsupervised learning models are used:

- **Clustering:** Clustering is a data mining method for categorizing unlabeled data based on their likenesses or differences. For example, K-means clustering algorithms assign the same data points into groups, whereas the value of K represents the size of the group and granularity. This method is useful for market segmentation, image compression, etc.
- **Association:** Association is another unsupervised learning technique that uses different rules to find relationships between variables in a given dataset. These methods are frequently used for market basket analysis and recommendation engines, along the lines of "Customers Who Bought This Item Also Bought" recommendations.
- **Dimensionality Reduction:** When the number of features (or dimensions) in a dataset is too high then we use the Dimensionality Reduction learning technique. It decreases the number of data inputs to a convenient size while also maintaining data integrity. Often, this technique is used in the preprocessing data stage, such as when auto encoders remove noise from visual data to improve picture quality.

1.2 Difference between Supervised and Unsupervised Learning

- Supervised machine learning uses labeled input and output data, while unsupervised learning uses unlabeled data and this is the main difference between the two machine learning techniques.
- In supervised learning algorithm "learns" from the training dataset by iteratively making predictions on the data and adjusting for the correct answer. While supervised learning models tend to be more accurate than unsupervised learning models, they require human interference to label the data appropriately. For example, a supervised learning model can predict how long your travel will be based on the time of day, weather conditions, and so on. But first, train it to know that rainy weather extends the driving time.
- While Unsupervised learning models work on their own to discover the essential structure of unlabeled data. But note that unsupervised learning still requires some human interference for authenticating output variables. For example, an unsupervised learning model can classify that online shopping customers often purchase specific products at the same time. However, a data expert would need to authorize that it makes sense to a recommendation of this specific product to the concerned buyer and group baby clothes with an order of diapers, apple sauce, and sippy cups.
- In supervised learning, the goal is to predict results for new data. You know the type of results to expect. With an unsupervised learning algorithm, the goal is to get an understanding of large volumes of new data. Machine learning itself defines what is different or interesting from the dataset.
- Supervised learning models are perfect for spam detection, sentiment analysis, weather forecasting, and pricing predictions, among other things. While on another hand unsupervised learning is a great fit for irregularity detection, recommendation engines, customer personas, and medical imaging.
- The supervised learning method is simple, typically calculated through the use of programming languages like R or Python. In unsupervised learning, you need powerful tools for working with large amounts of unlabeled data. Unsupervised learning models are computationally complex because they need a large training set to produce the required results.

- Supervised learning models may be a time-consuming process to train, and the labels for input and output variables require a skill. Meanwhile, unsupervised learning methods can have wildly incorrect results unless you have a human intervention to validate the output variables [7, 8].

1.3 Objective

The purpose of this work is to analyze the supervised and unsupervised learning algorithms by using the given datasets and analyzing the quality of data on the bases of Accuracy and Time. In supervised learning (classification) the analysis of Back Propagation, Naïve_Base, and KNN, and in unsupervised learning (clustering) analysis of K-means and EM algorithms are used for analysis in this study. For the algorithm analysis made in this study, a ready-made data set and its associated features are utilized.

1.4 Parameters of Analysis

Accuracy: The state or quality of being efficient, Efficiency performance level which consumes the minimum input quantity to generate the high quality of results. This study, it is tested which algorithm is the more accurate result producer in the given dataset.

Time: The rate at which someone or something moves/operates or can move/operate, in this study the Time of results-producing is monitored at given datasets.

1.5 Thesis Focus

The main thesis focus is to analyze the supervised (classification) and unsupervised (clustering) learning techniques by using their related algorithms upon Accuracy and Time to know which one is more accurate, efficient, and quickly results-producing among Backpropagation, Naïve Base, and KNN and the analysis of K means and EM.

1.6 Thesis Outline

Give a short overview of this thesis it consists of five chapters. The first chapter explains the background and introduction of data mining and its technologies.

The second chapter describes the subject of data mining which is the research area of computer science to extract the hidden information of data. For this purpose, many approaches such as deep learning and machine learning are used. In machine learning, supervised and unsupervised learning techniques are used. These techniques are further used in their algorithms to get the information from the given data. A brief knowledge about supervised learning technique algorithms like Support vector machine, backpropagation, Naïve Base, and KNN is also discussed in this chapter, and results are found based on accuracy, efficiency, and speed. The recent work on supervised and unsupervised learning techniques is also discussed in this chapter.

In the third chapter, the comparative analysis of Backpropagation, Naïve Base, KNN, and clustering algorithms k-means, EM at the given parameters Accuracy and Time. In this chapter, the details of the experiments are also discussed. The material regarding experiments is like datasets and algorithms. I have used some input for these algorithms and studied their results for analysis and result production which one is more accurate, efficient, and produces results.

In the fourth chapter, the experiment result is described and their analysis report is discussed.

And in the last fifth chapter, the conclusion and summary present the whole research work.

Chapter 2

Literature Review

2.1 Overview

Chapter two discusses machine learning and types of machine learning e.g. Supervised and Unsupervised learning. Classification of supervised and unsupervised learning techniques. Also, describe the different algorithms of supervised and unsupervised learning. The step involves supervised and unsupervised learning.

2.2 Machine Learning

In the field of computer science, a process named data detection plays a key role in data mining to get hidden knowledge from the dataset. In the field of computer science, machine learning algorithms have a big scope. In some cases, we cannot identify the information and hidden patterns from data after viewing it, in that case, machine learning helps us [9]. The purpose of machine learning is to solve daily life problems related to engineering. It allows understanding the problems without any explicit program. Machine learning is based on getting new ideas from the data. This process is frequently used in our daily lives. Machine learning is used in the decision-making process.

In the field of computing, Data Mining is an important field of research. In data mining, the data detection process involves getting hidden information and knowledge from the data by using the algorithms in machine learning [10]. Machine learning is a branch of artificial intelligence. Machine learning is about solving engineering problems in real life. In machine learning, all the information and knowledge are gained through the data by using different kinds of algorithms and machine learning tools. ML is the art to tell the machine how to handle the data more professionally and proficiently. Sometimes after inspecting the data we cannot interfere with the pattern or get statistics, facts, and information from the data, in such cases we apply ML. The purpose of ML is to learn from the data. This discipline is related to the Algorithm's coding, development, and design. These algorithms consent computers to modify performances that consist of experimental

information, as datasets. ML's main application of exploring is to mechanically find out the multifaceted patterns and produce quick results comprise of the data [11].

2.3 Types of Machine Learning

Machine learning extends the computer programs which is capable to modify the data. ML is generally classified into three divisions namely supervised learning (Classification), unsupervised learning (Clustering), and Reinforcement learning. Figure 2.1 [12] describes machine learning as divided into three approaches named supervised, unsupervised learning, and reinforcement learning techniques. But in this research work supervised learning, a classification technique (Backpropagation, Naïve Bayes, KNN, and Support Vector Machine), and an unsupervised learning clustering approach is used (Clustering algorithms, k-means, and EM) are going to compare the basis of accuracy and time [12].

There are different ways to train machine learning algorithms, each has advantages and disadvantages. To understand the pros and cons of each type of machine learning, we must first look at the type of data. In Machine Learning there are two kinds of data one is labeled data and other unlabeled data. Labeled data has both the input and output parameters in a completely machine-readable pattern, but requires a lot of human labor to label the data. Unlabeled data only have one or none of the parameters in a machine-readable form.

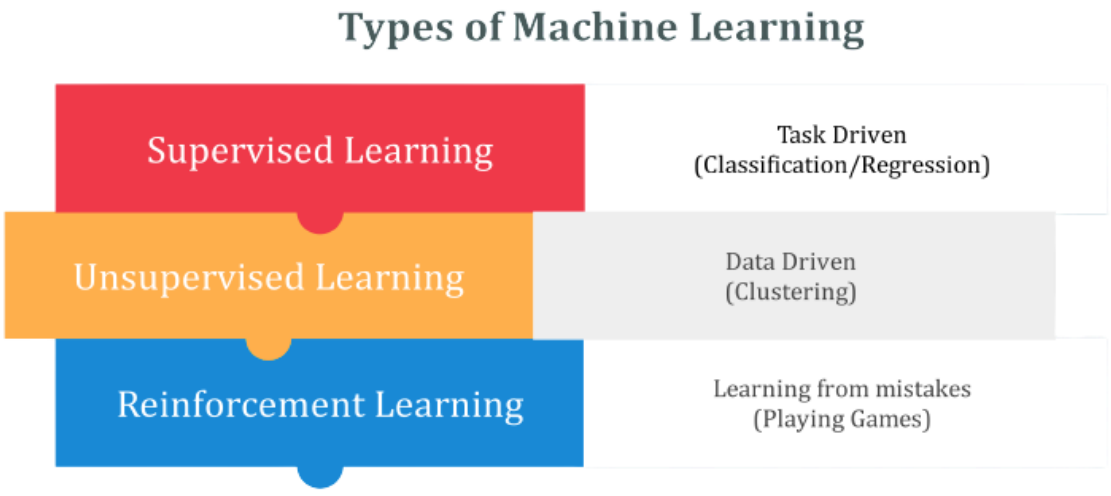


Fig 2.1: Types of Machine Learning

2.4 Types of Machine Learning Algorithms

The development of the machine learning process is comparable to data mining. Both machine learning and data mining are considered to extract information and hidden patterns from data. Fig: 2.2 shoes the Machine Learning algorithms [11]:

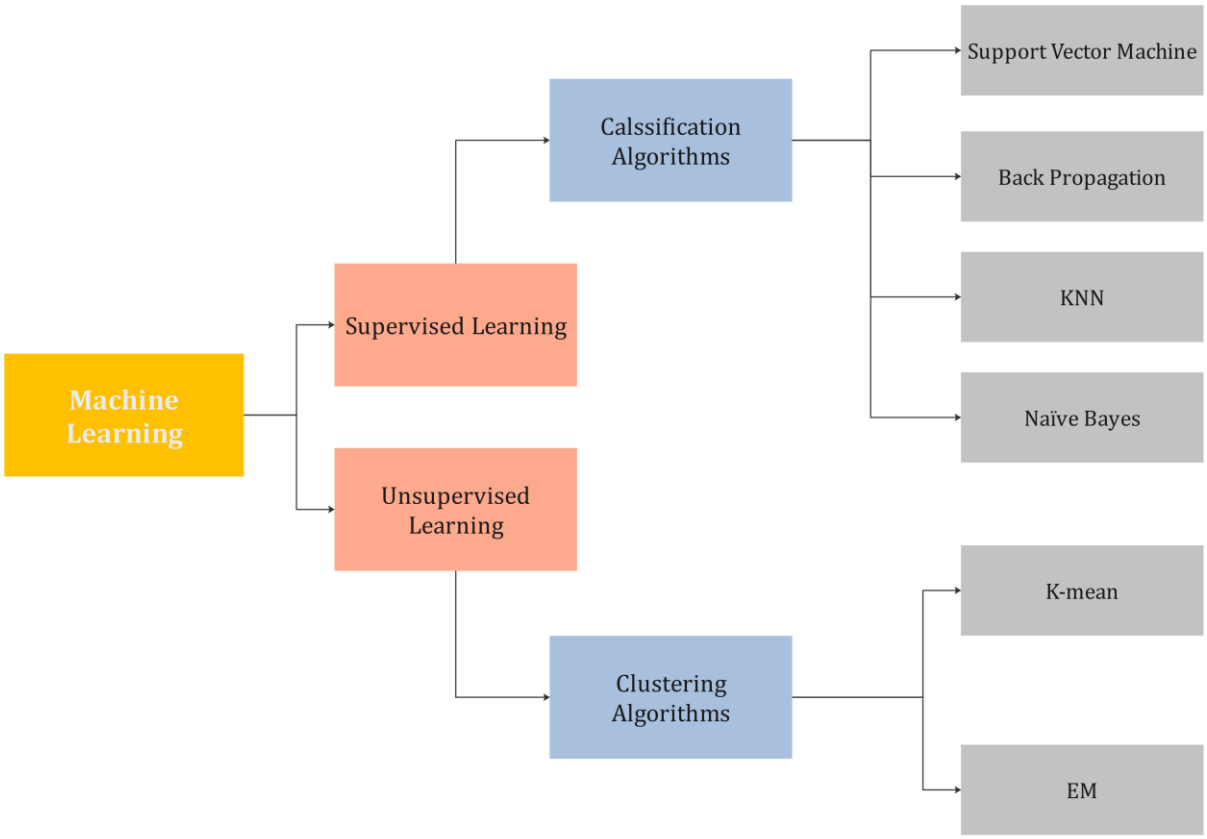


Fig 2.2: Types of Machine Learning Algorithms

Here we study only supervised (classification) and unsupervised learning (clustering) techniques and their analysis on Accuracy, speed, and Efficiency by applying datasets on related algorithms. Other techniques are beyond our research topic.

2.5 Supervised Machine Learning

Supervised machine learning involves allocating the specific data in such a way that a specific type of pattern or function can be extracted from that labeled data. It is an impotent feature of supervised machine learning that allocate the input as a vector and also gives the most desirable and important output value which is named the supervised signal. One of the important properties of supervised machine learning is that it classifies and labels the input data. In the datasets which are used by the machine learning algorithm, every instance of the dataset has some properties. It can be [13, 10]:

- Continuous
- Categorically
- Binary

If all the instances are described with specific labels then this type of learning is called supervised learning. Supervised machine learning is a process of searching algorithms from external instances to formulate a general hypothesis which then leads to making predictions for further knowledge. In simple words, the formation of a brief model of classification and distribution of specific labeled data in the sense of predictor properties is the main purpose of supervised machine learning. The supervised ML algorithms need outside assistance. The input dataset is divided into two parts (A) train and (B) test dataset. The training dataset has an output variable that needs to be predicted or classified. All algorithms read some kind of patterns from the training dataset and then apply these patterns to the test dataset for prediction or classification. The supervised machine learning classification algorithm's main purpose is to categorize data from prior information. Classification is carried out very frequently in data Science problems [14].

2.6 Classification

Classification is defined as the function of learning in which provided data items are mapped into more than a few classes that are predefined. It is a data investigation technique to formulate models related to significant data modules and foresees upcoming values. Classification techniques are used in data mining with ML, natural language processing, image processing, visualization, and statistical techniques to explore the knowledge in a format that is understandable. The latest

research on data mining has constructed such types of techniques that develop a robust and scalable model that can manage massive data. The classification has several applications like detection of fraud, performance prediction, target marketing, medical diagnosis, and manufacturing. The classification technique's performance is measured by speed, accuracy, robustness, comprehensibility, scalability, interpretability, and time. Classification and distribution techniques depend upon the inductive learning principle that observes and extracts different patterns from the database. As the nature of the environment is dynamic the model must be adaptive. It means it should have the ability to find and map the pattern effectively. Moreover, a model was introduced to classify labeled and unlabeled data more efficiently and appropriately.

For the theoretical learning of data, a framework was introduced in 2008, which replaced the Kernel Hilbert space. This model utilizes the non-negative specific function of the kernel. As compared to the previous framework it gives more efficient and better results. But there is a problem in choosing the specific kernel functions for a specific type of field. In 2008 Shilton and Palani Swami introduced an approach to support vector machines. This approach is firstly generated for binary classification and later on, it is extended to describe one-class classification and regression also. A framework that is based on binary classification for the two phases of kernel learning was introduced in 2012. One of the best advantages of this model is that it gives an easy way to perform research using binary classification and also develops the best algorithm based on the robust kernel function. In 2012 a robust kernel function is proposed which makes it best to optimize the previous and existing classification methods same as SVM. This model has many advantages such as distinct results related to theoretical research, enhancing the existing information about previous techniques, and describing the relation of different existing models [15].

2.7 Classification Process

The classification process required two types of data: Training Data and Test Data. A Training data set is a collection of all records which explain the past. A test data is a collection of all records which explain the present and for which we want to predict the future. These records do not have values for categorical fields. For example, we have a training data set of medical reports of all past patients and their relevant diseases. We could imagine that we have a test data set composed of

medical reports of current patients which are yet to be analyzed for diagnostic purposes. The training data must be accurate and complete. Authenticity refers to the correct, accurate, timely, and valid values. Completeness mentions the availability of all possible combinations. A training data set is provided to the classification algorithm for data mining purposes. A classification algorithm like Naïve Bayes, decision tree, KNN, etc. Analysis of the training data and applying statistical methods to determine hidden relationships among various features and outcomes. The output of the classification algorithm is a statistical model called a classifier. This phase is called the Training / Learning phase. Test data is fed to the classifier to predict the unknown class labels. Here the results will be decided based on an analysis of available features. The output is also called classification results with test data with labels. Here labels refer to the outcomes in the form of values assigned to one or more categorical unknown fields. This phase is called Testing Phase. Fig 2.3 shows the main mechanism of classification [16].

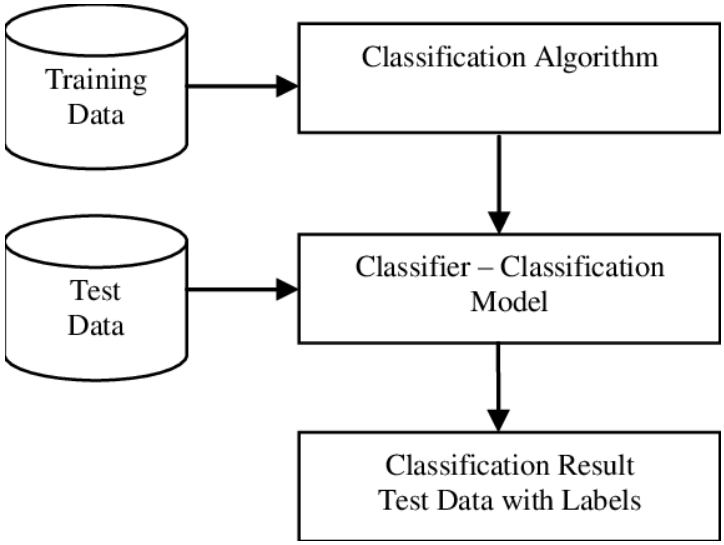


Fig 2.3: Classification Components

2.8 Classification Algorithms

Nowadays learning algorithms are useful in many fields and these are very appropriate and useful. A classifier makes many predictions to measure various trade-offs to perform many different activities. Supervised learning algorithms are shortly described below:

2.8.1 Logistic Regression

The algorithm describes the probability of recounting the possible outcomes of a different assessment modeled through a logistic function. It is a predictive analysis algorithm based on the idea of probability. We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complicated cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function. The hypothesis of logistic regression tends to bound the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression [17].

2.8.2 Naïve Bayes

Naïve Bayes algorithm based on Bayes theorem, one of the statistical classifications, and needs few amounts of training data to estimate the parameters, also known as probabilistic classifiers. It is considered to be the fastest classifier, highly scalable, and handles both discrete and continuous data. In addition, this algorithm is used to predict in real-time. This classifier works sound in various actual-world circumstances like spam filtering and classification of a document [18].

2.8.3 K-Nearest Neighbors

KNN is based on similar data, this classifier learns the patterns present within. It is a non-parametric and lazy learning algorithm. Non-parametric means that the assumption for underlying data distribution does not valid. In lazy loading, there is no condition for training data points for generating models. The training data is utilized in the testing phase causing the testing phase slower and costlier as compared with the training phase.

The Limitation of the algorithm is that the value of K Needs to be determined and the cost of computation is very high as it requests to the computer the space of every instance for all training samples [18] [19].

2.8.4 Decision Tree

Decision Trees facilitate prediction as well as classification. Using the decision trees, one can make decisions with a given set of input. Provided data along with all its classes, a series of rules is produced via a decision tree which can be utilized for the classification of the data. It involves small preparation of data and can manage both categorical and numerical data [18].

2.8.5 Random Forest

It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model. As the name defines, Random Forest is a classifier that holds several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. It is a complicated algorithm tricky to execute [18].

2.8.6 Support Vector Machine

SVMs are used mostly for classification. In SVM, we plot our data in an n-dimensional space. The value of each feature in SVM is the same as that of a specific coordinate. Then, we continue to find the ideal hyperplane differentiating between the two classes.

SVM is the demonstration of training data such as a point in the space that is divided into groups through an apparent gap that is possibly wide. Then the latest model is mapped to that identical space and predicted to fit in a group that is based on the side of the gap they decrease. It is efficient in the spaces of high dimensional and the function of the decision subset of the training points is used thus it is memory efficient as well. Probability estimation does not provide directly by the algorithm that is calculated by classy cross-validation [18] [19].

2.9 Unsupervised Machine Learning

Unsupervised learning may be a sort of algorithm that learns patterns from unlabeled data. In unsupervised learning, the training of the machine is done using information that is neither classified nor labeled. The machine learning algorithm work on information without guidance. It categorizes unsorted information according to similarities, patterns, and differences without any prior training or supervision [20].

There is no need to train the machine, the machine itself finds the hidden structure in unlabeled data and interprets it. The most frequent method of unsupervised learning is the analysis of clusters, which is exploited for investigative analysis of data to figure out the categories or secret patterns in the data.

In this, research work, we only study clustering and clustering algorithms k means and EM other techniques are beyond our research work [21].

2.10 Clustering

The grouping of similar data objects is called clustering. Clustering is the most common technique in unsupervised learning where data is grouped based on the similarity of the data points. Clustering has many real-life applications where it can be used in different situations. The basic principle behind cluster is to assign a given set of observations into sub-groups or clusters such that observations present in the same cluster possess a degree of similarity. It is the implementation of the human cognitive ability to differentiate objects based on their nature [22]. For example, when you go out grocery shopping, you easily distinguish between apples and oranges in a given set containing both of them. We can differentiate these two objects based on their color, texture, and other sensory information that is processed by your brain. Clustering is a simulation of this process so that machines can differentiate between different objects.

It is a technique of unsupervised learning since there is no external label attached to the object. The machine has to learn the features and patterns by itself without any given input-output mapping or human intervention. The algorithm can extract inferences from the nature of data objects and then create distinct classes to group them appropriately. The applications of clustering

frequently deal with huge datasets and data having many features. Investigation of such data is a focus of data mining. In clustering data is grouped and similar objects are kept in the same group [23].

2.10.1 Clustering Workflow

To cluster the data, we follow these steps:

Prepare Data: This process involves selecting the suitable attribute, performing the proper extraction of the feature, and preprocessing the data items to estimate the values of the set of the feature that is chosen. It will frequently be appropriate to select a subset of the entire available features, to overcome the problem of space dimensionality [24].

Create Similarity Metric: In the clustering process, similarity measures perform a significant part in the procedure of clustering where many clusters are formed by the set of objects, hence the same objects should be in an identical cluster. In this function, two data sets are taken as input and proceed output among them as the similarity measure.

Run Clustering algorithm: A clustering algorithm uses the similarity metric to cluster data. The clustering algorithm tries to discover ordinary component groups (or data) that depend on the number of similarities.

Interpret Result: If the result does not make any sense we can improve the result requires iteratively experimenting with the previous steps to see how they affect the clustering. Check the quality of your clustering output is iterative and exploratory because clustering lacks “truth” that can verify the output. You verify the result against expectations at the cluster level and the sample level [25].

2.11 Clustering Algorithms

2.11.1 K-means Clustering Algorithm

In the field of data mining, k-means clustering refers to cluster analysis. It directs data into partition and observations into k clusters in which every observation belongs to a cluster with the nearest mean. The algorithm uses an iterative improvement technique, that's why it is called a k-means algorithm. It is also called Lloyd's algorithm, particularly in the community of computer science, K-means is a widely used technique. In scientific and industrial fields, it is the most common and popular used algorithm [26].

The basics of the algorithm are very simple [27].

1. Choose K points as primary points
2. Repeat
3. Now Form the K cluster by assigning each point to its neighboring closest points
4. Recomputed the central point of each cluster until the central point does not change

2.11.2: Expectation-Maximization (EM) Algorithm

EM stands for Expectation-Maximization Algorithm. In data mining, the EM algorithm is also significant. We can discover patterns and correlations in large preexisting datasets. It is a way to discover new meaning in data. When we are satisfied with the results of K-Means algorithms, we can use EM algorithms. It is an iterative process for finding the uttermost likeliness or MAP (Maximum A Posterior) approximation of parameters in statistical models. Hence, models depend on unobserved latent variables. Maximum likelihood becomes intractable if there are variables that interact with those in the dataset but were hidden or not observed as so-called latent variables [28].

An expectation-maximization algorithm is an approach for performing maximum likelihood estimation in the presence of latent variables. It does this by first approximating the values for the latent variables, then optimizing the model, then iterating these two steps till convergence. It is an effective and general approach and is most commonly used for density estimation with missing data, such as clustering algorithms [29].

The EM iteration alternates between an E (Execution) step and M (Maximization) step. E-step computes the expectations of the likelihood, while M-step computes parameters maximizing the expected log-likeliness found on E-step. Then we can use these parameter estimates to determine the distribution of the latent variables in the next E- step. In a Band-Name class index, the results of the cluster analysis will be written. In this band, the values indicate the class directories. Here a value (0) refers to the first-class cluster, while value (1) refers to the second-class cluster, and so on. The main use of the EM algorithm is to locate the maximum likelihood parameters of a statistical model in the case where the equations cannot be solved at once. These models also involve observations of known data. That's why, either missing values exist among data or the model can be formulated in such a way that it is more simply by accepting without verification, the existence of further unobserved data points. As an example, we can describe a mixture model, more simply by assuming that each observed data point has similar unobserved data points or latent variables. These latent variables specify the mixture component to which each data points belong. Typically, for finding a maximum likelihood solution, we need to take the derivatives of the likelihood function concerning all the ambiguous values, parameters, and latent variables. Then simultaneously solving the resulting equations. This is normally impossible in statistical models with latent variables. Alternatively, the results are typically a set of interlocking equations [28] [29] [18].

2.12 Summary

In this chapter, the basics of machine learning, machine learning types are discussed and how computing power has changed over time, and advanced machine learning algorithms. Also, discuss different types of machine learning algorithms and further took a brief look at some of the popular ML algorithms. This chapter explains supervised machine learning, unsupervised machine learning, and several algorithms that are part of this machine learning. In the above, learned about the various algorithms that are used for machine learning classification. This chapter went through clustering and how clustering has brought advanced data analysis techniques to unlabeled datasets. Overview of the clustering algorithms.

Chapter 3

Experimental Study

3.1 Overview

Two approaches to machine learning are tested in this experimental study. The first experiment is done between the classification algorithms named KNN, Naïve Bayes, and Back Propagation. The features are Accuracy and Time. And in the second part the unsupervised learning technique (Clustering) algorithms K-means and EM are tested upon two datasets and the same features.

3.2 Datasets

In this research work, I am retrieving the data from one data source e.g WEKA. The datasets for classification and clustering techniques are obtained from the weka website. The received data is unlike in nature. This useful data can be applied directly to the data mining tools and predict future results.

Sr. No	Datasets	No of Classes	No. of Attributes	No. of Instances
1	Breast Cancer	2	10	286
2	Ionosphere	2	35	351
3	Glass	7	10	214
4	Unbalanced	2	33	856
5	Soybean	19	36	683
6	Labor	2	17	57
7	Credit	2	21	1000

Table 3.1: Datasets

3.3 Methodology

It is the computational process of unfolding patterns in the large raw data set which will support making the right decisions and designing strategies for further usage.

Raw Data in the data mining process could be anything as below:

1. CSV files (comma separated values)
2. Data warehouse
3. CRM
4. Transactional Data
5. Text, fact files, etc.

The data mining process is mainly applied to a large amount of. This process identifies the relationship between input data, analyzes patterns, and extracts information that gets transformed into a user-understandable format like the dashboard, tables, charts, reports, etc. The resultant data generated is called “information”, thus it can be concluded that knowledge discovery is the main aim of the data mining process.

I take datasets from the repositories on WEKA. The WEKA applied versatile classification algorithms and then predict a helpful result that is very useful for further use and new users.

3.4 What is Weka?

Weka is a data mining visualization tool that contains a collection of machine learning algorithms for data mining tasks. It is an open-source software issued under the GNU General Public License. It provides result information in the form of a chart, tree, table, etc.

Weka work on a file that is in Attribute-Relation File Format (ARFF) file. So, first, we have to convert any file into ARFF before we start mining with it in Weka [30].

3.4.1 Features of Weka

Data Preprocessing: It is the cleaning of data while the data gathering and selection phase. It removes/adds default value to missing fields and resolves conflicts.

Data Classification and Prediction: It classifies data based on the relationship between things and predicts data labels. For example, a Bank based on available data on loans classifies and predicts customer labels 'risky' or 'safe'.

Clustering: Group of related data into the cluster, used to discover a distinct group. For example, we have data on weather, and the data based on that we want to decide whether to play outside or not, in such a case, using the Weka tool we can visualize overall data and can make a decision according to the charts [30].

3.5 Experimental Study of Classification Algorithms

Algorithms are compared on the bases of Accuracy and Time. Accuracy refers to the understanding of a measured value to a standard or known value. In this study, it is tested which algorithm produces a more accurate result in the given dataset. The state or quality of being efficient, efficiency is a level of performance that uses the lowest amount of inputs to create the greatest amount of outputs. In this study, it is also tested which algorithm is more efficient at given practices. The followings are the classification algorithms:

1. KNN (K Nearest Neighbor)
2. Naïve Bayes
3. Back Propagation

3.6 Data Mining Process

The process involves different steps and the followings are the basic steps:

1. The first step is selecting the dataset.
2. In the second step, the data preprocessing is done and data is prepared to analyze the hidden patterns. In some cases, the missing values and outliers disturb the accuracy and time of

the algorithms, so first, deal with those missing values and outliers (The values which are outside, the interquartile range are known as outliers and should be removed). This is called data preprocessing.

3. In the next step, the features are engineered. This step is further divided into two parts (a) Feature Extraction and (b) Feature Selection. The removal of irrelevant features from the dataset is called feature extraction.
4. Select the tool to enhance your productivity. Using the tools, build the model and assess initial results.
5. Select appropriate algorithms for the required task and necessary parameters.
6. Evaluate preliminary results and test the model on different sample data sets and review the results [31].

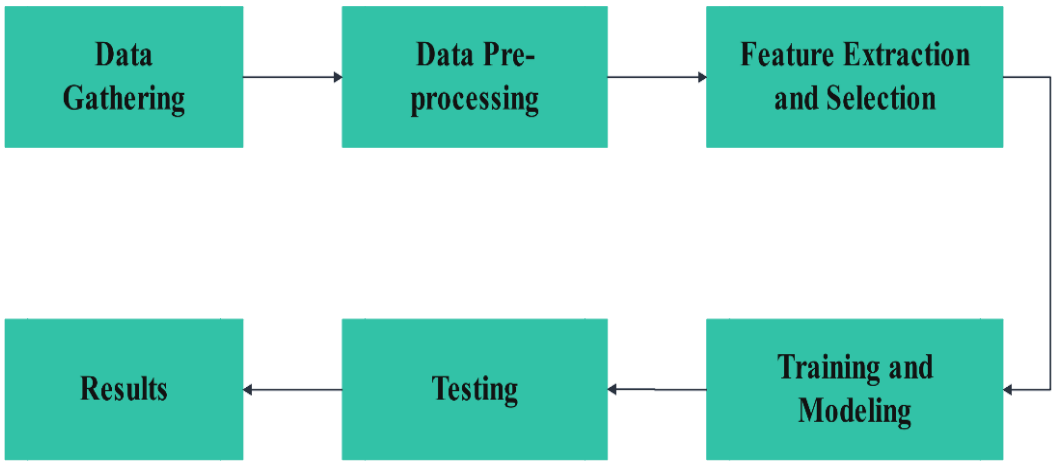


Fig 3.1: Data Mining Process

3.7 Experimental Analysis of Classification (Supervised Learning) Algorithms

3.7.1 Performance of KNN Algorithm

The datasets are tested by the weka tool applying classification algorithms. The findings are given below in Table. The testing time, training time, testing accuracy, and training accuracy are described for the different values of K at different datasets with the setting of cross-validation being 10.

Dataset Name	No of K's	Training Time (Sec)	Testing Time (Sec)	Testing Accuracy (%)	Training Accuracy (%)
Breast Cancer	3	0	0.01	73.7762	80.0699
	6	0	0.01	73.0769	76.2238
	9	0	0.01	73.4266	75.8741
	12	0	0.01	72.3776	74.8252
	15	0	0.01	73.0769	74.1259
	18	0	0.01	73.0769	74.4755
	21	0	0.01	72.7273	74.4755
	24	0	0.01	72.3776	74.4755
	27	0	0.02	72.3776	73.7762
	30	0	0.01	72.028	73.7762
Ionosphere	3	0	0.03	86.6097	91.1681
	6	0	0.03	86.0399	88.3191
	9	0	0.03	84.6154	85.1852
	12	0	0.03	85.4701	85.1852
	15	0	0.04	84.0456	84.9003
	18	0	0.04	84.9003	85.755
	21	0	0.04	81.7664	84.9003
	24	0	0.03	82.906	84.6157
	27	0	0.05	79.2023	82.3362
	30	0	0.03	79.2023	81.4851
Glass	3	0	0.01	71.9626	80.8411
	6	0	0.01	66.8224	74.7664
	9	0	0.01	63.0841	69.1589
	12	0	0.01	63.0841	71.4953

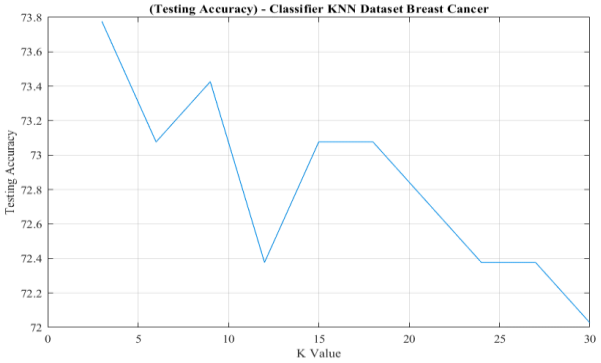
	15	0	0.02	62.6168	66.8224
	18	0	0.01	63.0841	66.3551
	21	0	0.01	65.4206	64.486
	24	0	0.01	63.0841	66.3551
	27	0	0.01	61.6822	65.8879
	30	0	0.01	60.7477	64.9533
Unbalanced	3	0	0.07	98.3645	98.715
	6	0	0.08	98.5981	98.5981
	9	0	0.09	98.5981	98.5981
	12	0	0.11	98.5981	98.5981
	15	0	0.17	98.5981	98.5981
	18	0	0.01	98.5981	98.5981
	21	0	0.19	98.5981	98.5981
	24	0	0.13	98.5981	98.5981
	27	0	0.13	98.5981	98.5981
	30	0	0.02	98.5981	98.5981
Soybean	3	0	0.07	91.3616	94.1435
	6	0	0.09	89.6074	91.8009
	9	0	0.09	88.5798	89.7511
	12	0	0.08	86.6764	88.5798
	15	0	0.08	84.9195	87.2621
	18	0	0.08	83.6018	85.5051
	21	0	0.12	80.9663	84.3338
	24	0	0.13	77.7452	81.2592
	27	0	0.11	73.7921	78.7701
	30	0.01	0.1	70.7174	75.8419
Labor	3	0	0	91.2281	100
	6	0	0	82.4561	92.9825
	9	0	0	91.2281	91.2281
	12	0	0	91.2281	92.9825
	15	0	0	80.7018	94.7361
	18	0	0	80.7018	91.2281
	21	0	0	80.7018	82.4561
	24	0	0	75.4386	80.7018
	27	0	0	75.4386	77.793
	30	0	0	73.6842	77.193
Credit	3	0	0.1	73.3	86
	6	0	0.1	74.3	79.7
	9	0	0.14	74.3	79.8

	12	0	0.17	74.4	77.9
	15	0	0.13	73.5	77.6
	18	0	0.13	73.2	75.2
	21	0	0.14	73.6	75.5
	24	0	0.16	73.7	75.2
	27	0	0.13	73.5	76
	30	0	0.17	73	74.7

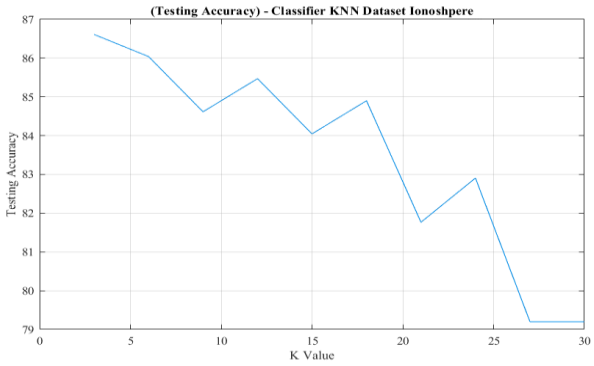
Table 3.2: KNN Algorithm Results

3.7.2 Graphical Representation of Testing Accuracy (KNN)

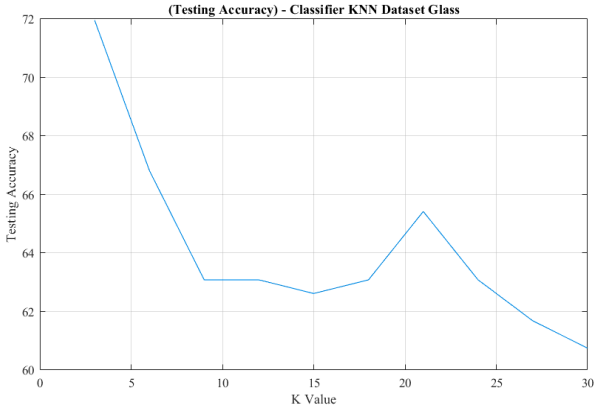
Testing accuracy of all datasets by using the KNN algorithm is graphically represented below. The graph is drawn using Matlab Software. The X-axis shows the no of K's values and Y-axis shows the testing accuracy.



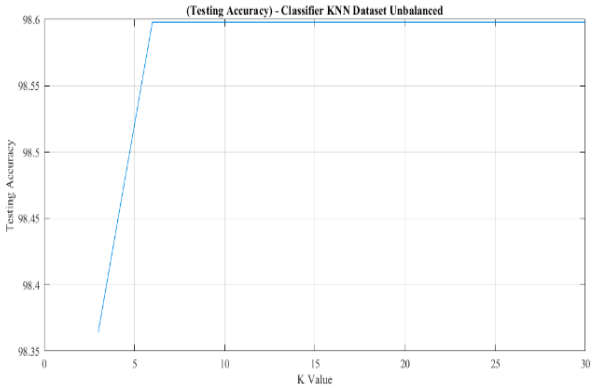
Graph 3.1: KNN Testing Accuracy of Breast Cancer



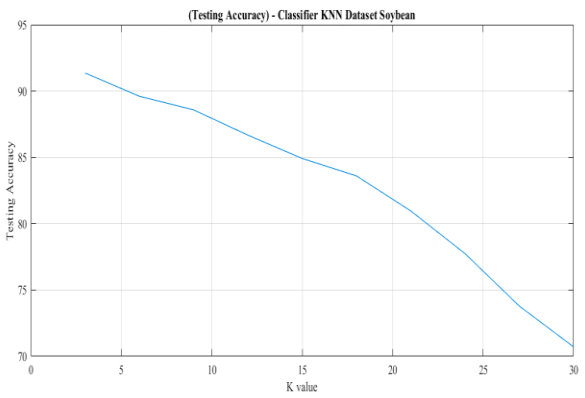
Graph 3.2: KNN Testing Accuracy of Ionosphere



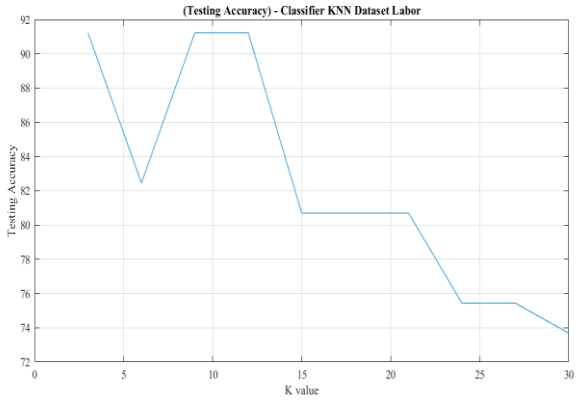
Graph 3.3: KNN Testing Accuracy of Glass



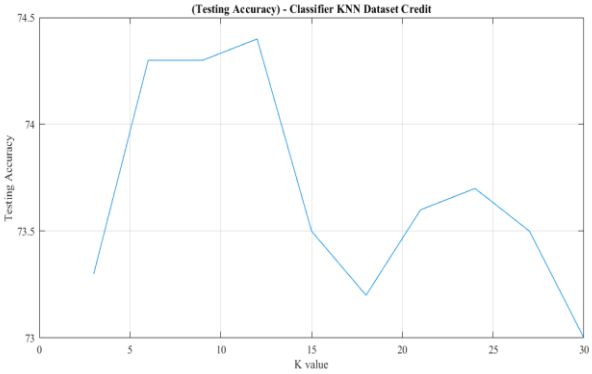
Graph 3.4: KNN Testing Accuracy of Unbalanced



Graph 3.5: KNN Testing Accuracy of Soybean



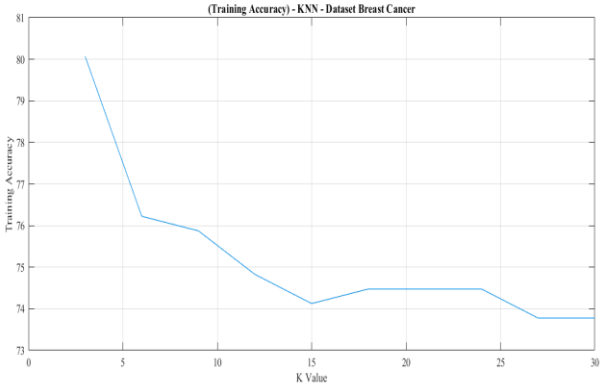
Graph 3.6: KNN Testing Accuracy of Labor



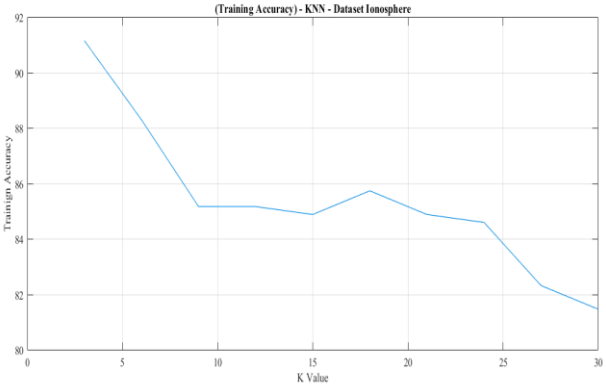
Graph 3.7: KNN Testing Accuracy of Credit

3.7.3 Graphical Representation of Training Accuracy (KNN)

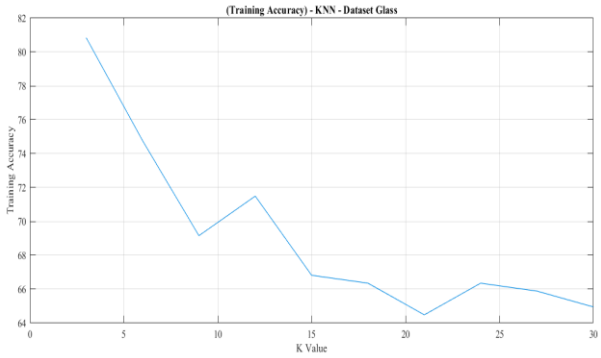
The Training accuracy of all datasets by using the KNN algorithm is graphically represented below. The graph is drawn using Matlab Software. The X-axis shows the no of K's values and Y-axis shows the training accuracy.



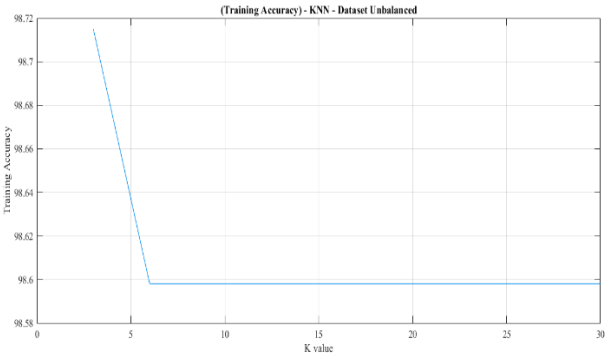
Graph 3.8: KNN Training Accuracy of Breast Cancer



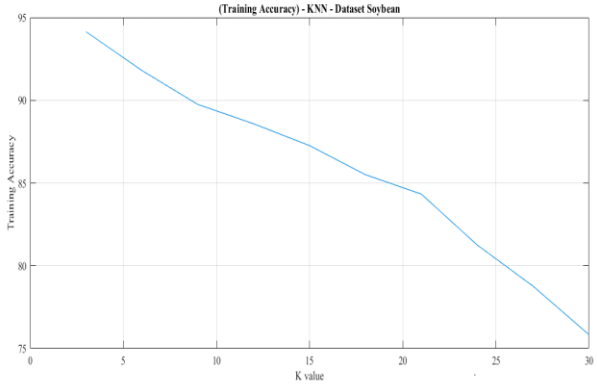
Graph 3.9: KNN Training Accuracy of Ionosphere



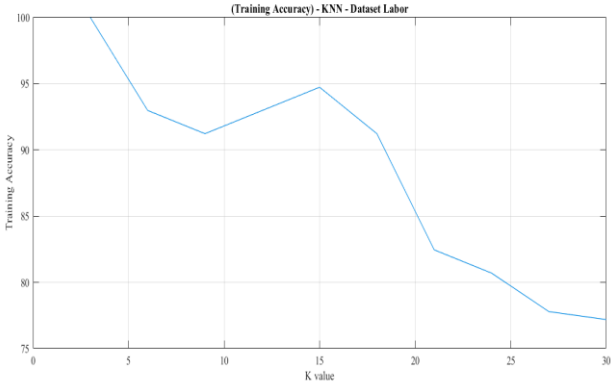
Graph 3.10: KNN Training Accuracy of Glass



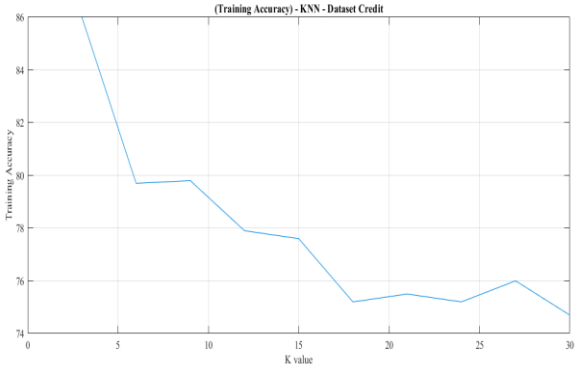
Graph 3.11: KNN Training Accuracy of Unbalanced



Graph 3.12: KNN Training Accuracy of Soybean



Graph 3.13: KNN Training Accuracy of Labor



Graph 3.14: KNN Training Accuracy of Credit

3.7.4 Performance of Backpropagation Algorithm

The datasets are tested by the Weka tool applying Classification Algorithms. The findings of the backpropagation algorithm are given below in Table 3.3. The Testing Time, Training Time, Testing Accuracy, and Training Accuracy are described for different values of hidden layers 10 to 100 with the multiple of 10 at different datasets with the setting of cross-validation being 10.

Dataset Name	No of Hidden Layers	Training Time (Sec)	Testing Time (Sec)	Testing Accuracy (%)	Training Accuracy (%)
Breast Cancer	10	0.91	0	69.5804	96.1538
	20	1.87	0	67.1329	97.9021
	30	2.91	0	67.4825	97.9021
	40	3.57	0	66.0839	97.2028
	50	4.51	0	67.1329	97.9021
	60	5.45	0	67.4825	97.9021
	70	6.22	0.01	67.4825	97.9021
	80	7.63	0.01	69.5804	97.2028
	90	8.14	0.01	67.1329	97.9021
	100	9.11	0	67.1329	97.5524
Ionosphere	10	0.86	0	91.453	99.7151
	20	1.7	0	90.3134	99.7151
	30	2.54	0	91.453	99.7151
	40	3.36	0	90.3134	99.7151
	50	4.26	0	89.1738	100
	60	5.23	0	90.3134	99.7151
	70	5.97	0.01	89.7436	98.8604
	80	6.92	0.01	89.7436	99.7151
	90	7.66	0.01	89.1738	98.8604
	100	8.58	0.01	90.0285	98.8604
Glass	10	0.3	0	66.3551	82.243
	20	0.53	0	69.1589	85.9813
	30	0.79	0	70.0935	85.514
	40	1.01	0	70.0935	88.3178
	50	1.25	0	69.6262	86.9159
	60	1.48	0	70.0935	84.5794
	70	1.74	0	70.935	87.8505

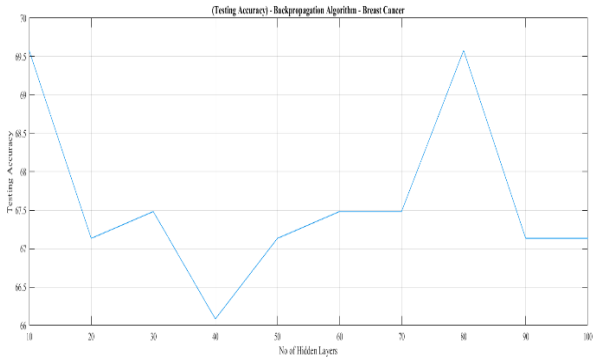
	80	1.97	0	71.4953	86.9159
	90	2.2	0	68.6916	87.3832
	100	2.5	0	67.2897	87.3832
Unbalanced	10	2	0	98.1308	99.4159
	20	3.97	0	97.8972	99.1822
	30	5.9	0.01	98.1308	99.0654
	40	7.94	0.01	97.8972	99.1822
	50	9.79	0.01	98.1308	99.1822
	60	11.88	0.01	97.8972	99.2991
	70	13.93	0.01	97.6636	99.1822
	80	15.89	0.01	98.014	99.2991
	90	17.62	0.01	98.014	99.0654
	100	20.51	0.02	98.014	98.5981
Soybean	10	4.73	0.01	91.3616	99.8536
	20	9.33	0.01	89.6074	99.8536
	30	13.65	0.01	88.5798	99.8536
	40	18.16	0.02	86.6764	99.8536
	50	23.39	0.03	84.9195	99.8536
	60	27.15	0.02	83.6018	99.8536
	70	32.64	0.03	80.9663	99.8536
	80	36.41	0.03	77.7452	99.8536
	90	40.78	0.04	73.7921	99.8536
	100	45.32	0.06	70.7174	99.8536
Labor	10	0.12	0	87.7193	100
	20	0.23	0	85.0659	100
	30	0.33	0	85.9549	100
	40	0.44	0	85.9649	100
	50	0.55	0	85.9649	100
	60	0.69	0	87.7193	100
	70	0.77	0	85.9649	100
	80	0.88	0	87.7193	100
	90	0.99	0	87.7193	100
	100	1.16	0	89.4737	100
Credit	10	4.11	0	71.8	97.8
	20	8.11	0.01	70.8	98.8
	30	13.85	0.02	73	99
	40	19.85	0.02	71.7	99.3

	50	25.39	0.03	72.7	99
	60	38.89	0.04	71.8	99.3
	70	38.79	0.03	72.2	98.5
	80	46.64	0.03	73.4	98.5
	90	36.78	73.5	0.04	99
	100	52.63	71.8	0.04	98.6

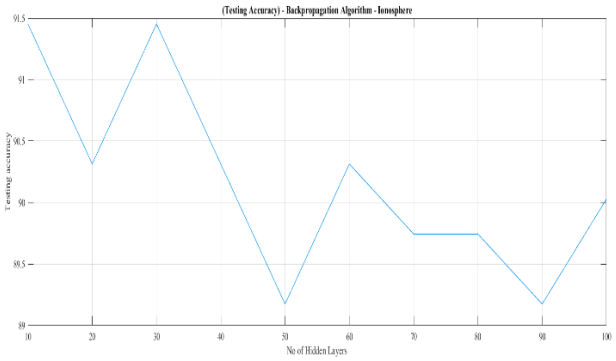
Table 3.3: BP Algorithm Results

3.7.5 Graphical Representation of Testing Accuracy (BP)

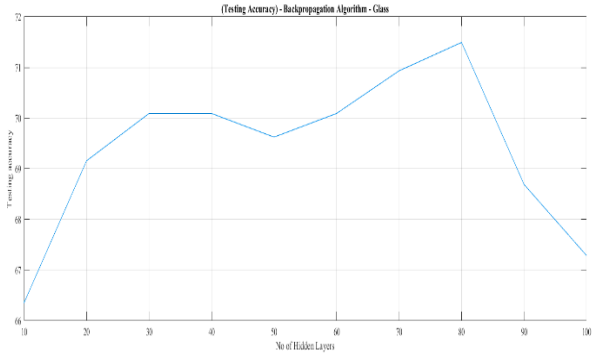
The Testing accuracy of all datasets by using the backpropagation algorithm is graphically represented below. The graph is drawn using Matlab Software. The X-axis shows the no of hidden layers values and Y-axis shows the testing accuracy with the cross-validation 10.



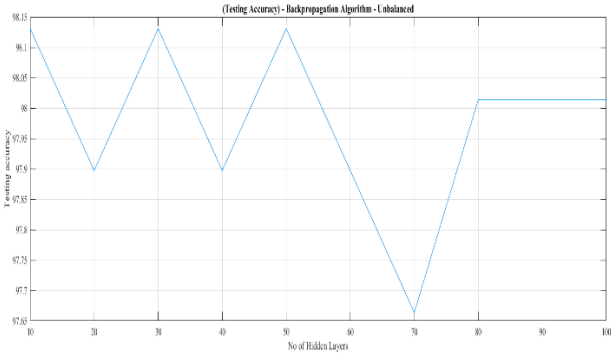
Graph 3.15: BP Testing Accuracy of Breast Cancer



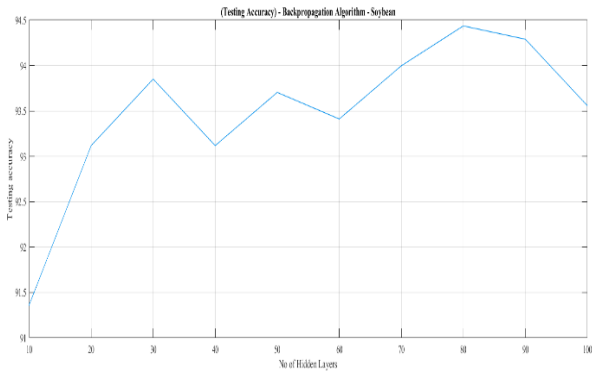
Graph 3.16: BP Testing Accuracy of Ionosphere



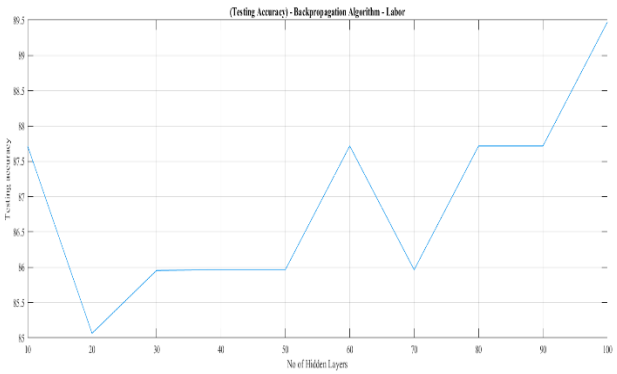
Graph 3.17: BP Testing Accuracy of Glass



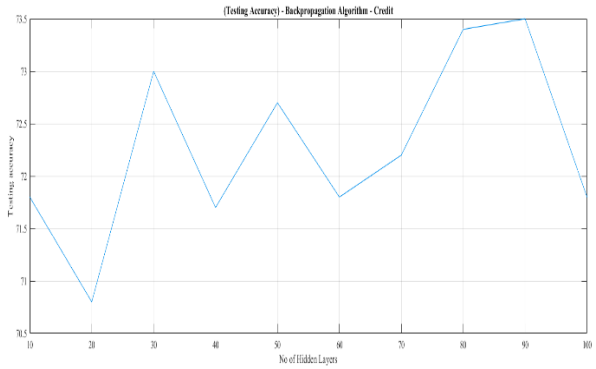
Graph 3.18: BP Testing Accuracy of Unbalanced



Graph 3.19: BP Testing Accuracy of Soybean



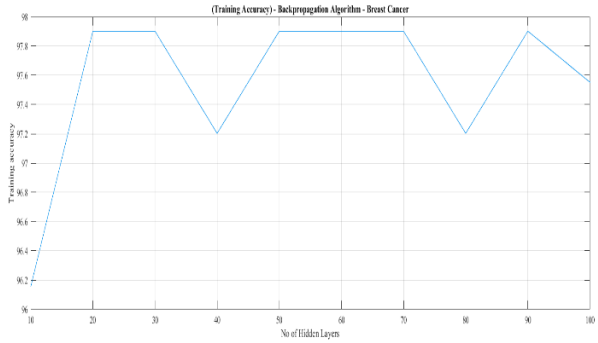
Graph 3.20: BP Testing Accuracy of Labor



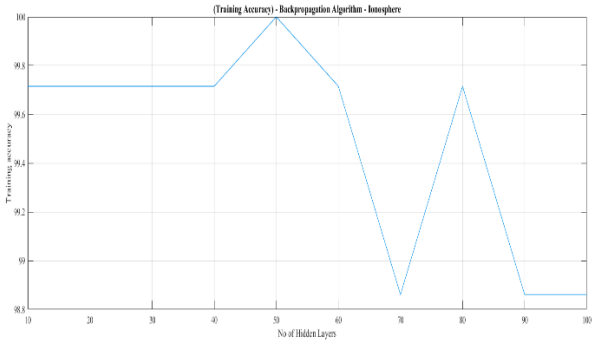
Graph 3.21: BP Testing Accuracy of Credit

3.7.6 Graphical Representation of Training Accuracy (BP)

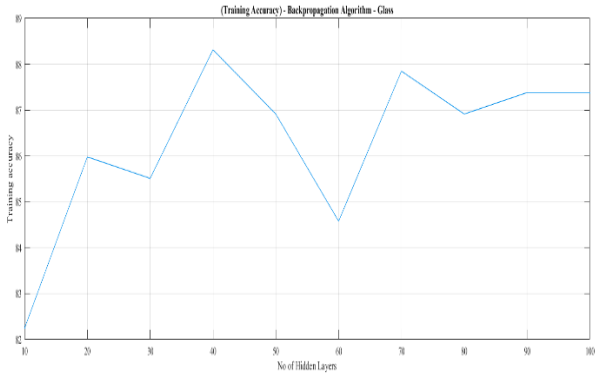
The Training accuracy of all datasets by using the backpropagation algorithm is graphically represented below. The graph is drawn using Matlab Software. The X-axis shows the no of hidden layers values and Y-axis shows the training accuracy with the cross-validation 10.



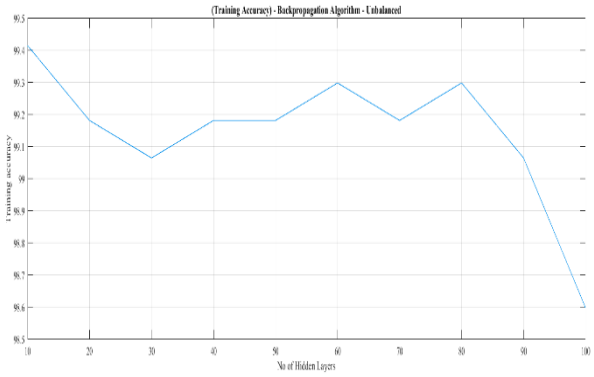
Graph 3.22: BP Training Accuracy of Breast Cancer



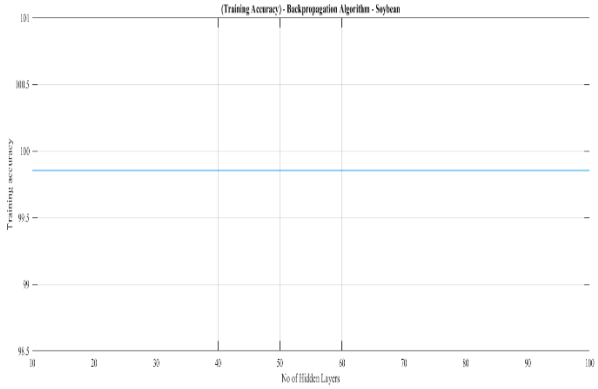
Graph 3.23: BP Training Accuracy of Ionosphere



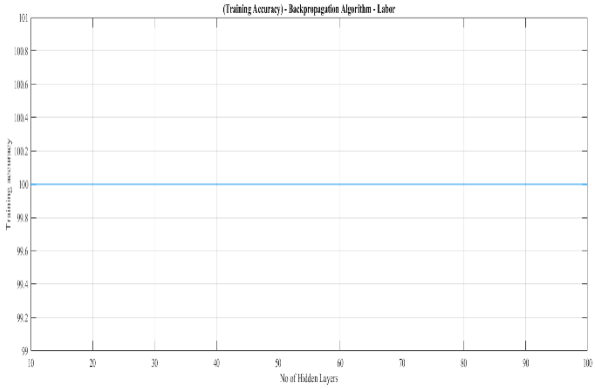
Graph 3.24: BP Training Accuracy of Glass



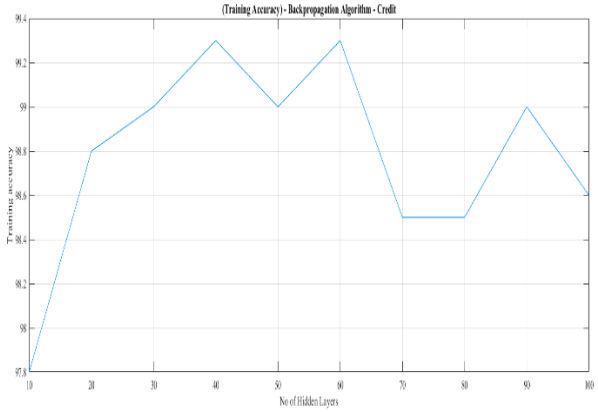
Graph 3.25: BP Training Accuracy of Unbalance



Graph 3.26: BP Training Accuracy of Soybean



Graph 3.27: BP Training Accuracy of Labor



Graph 3.28: BP Training Accuracy of Credit

3.7.7 Performance of Naïve Bayes Algorithm

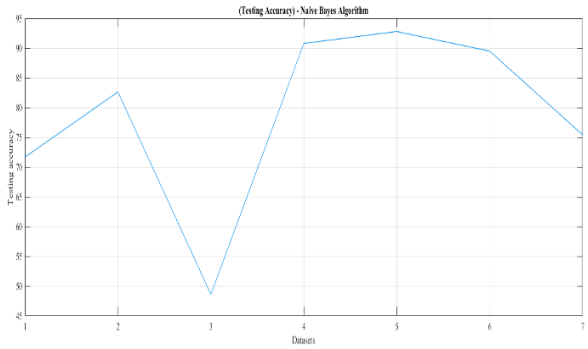
All the datasets are tested by the Naïve Bayes algorithm through the Weka tool. The findings of the Naïve Bayes algorithm are given below in table 3.4. The testing time, training time, testing accuracy, and training accuracy are described with the setting of cross-validation being 10.

Dataset Name	Training Time (Sec)	Testing Time (Sec)	Testing Accuracy (%)	Training Accuracy (%)
Breast Cancer	0	0	71.6783	75.1748
Ionosphere	0.01	0.01	82.6211	82.906
Glass	0	0	48.5981	55.6075
Unbalanced	0.01	0.01	90.771	93.3411
Soybean	0	0.02	92.7722	93.7042
Labor	0	0	89.4737	98.2456
Credit	0	0.01	75.4	77.2

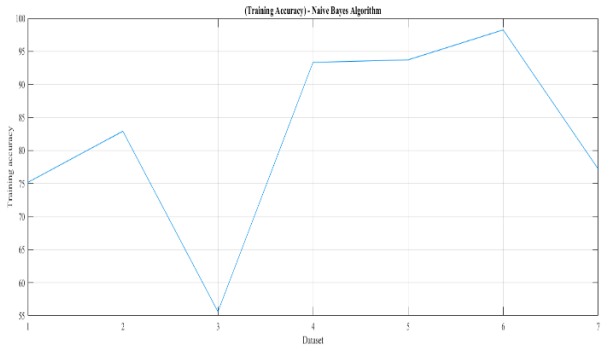
Table 3.4: Naïve Bayes Algorithm Results

3.7.8 Graphical Representation of Naïve Bayes Algorithm

The Training accuracy and Testing accuracy of all datasets by using the Naïve Bayes algorithm is graphically represented below. The graph is drawn using Matlab Software. The X-axis shows the datasets and Y-axis shows the training accuracy and testing accuracy.



Graph 3.29: Naïve Bayes Testing Accuracy



Graph 3.30: Naïve Bayes Training Accuracy

3.7.9 Highest Testing and Training Accuracy of Classification Algorithms

The highest testing and training accuracy is given in tables 3.5 and 3.6. This analysis shows the clear difference between the performances of different algorithms. And the graphical representation of these highest testing and training accuracies.

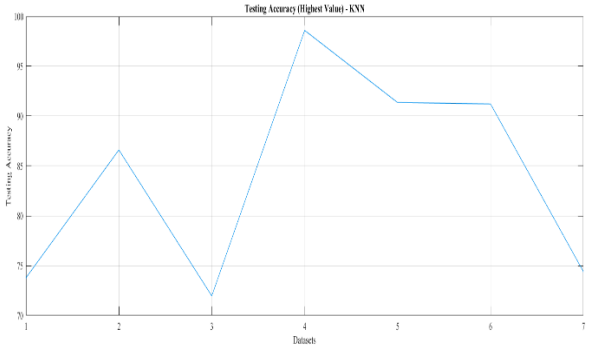
Highest Testing Accuracy of Classification Algorithms								
Datasets	**KNN**			**BP**			**Naïve Bayes**	
	Time (Sec)	**Testing (%)**	**K**	**Testing (%)**	**Hidden Layers**	**Time (Sec)**	**Testing (%)**	**Time (Sec)**
Breast Cancer	0	73.7762	3	69.5804	10	0	71.6783	0
Ionosphere	0	86.6097	3	91.453	10	0	82.6211	0.01
Glass	0	71.9626	3	71.4953	80	0	48.5981	0
Unbalanced	0	98.5981	6	98.1308	10	0	90.771	0.01
Soybean	0	91.3616	3	94.4363	80	0.03	92.7722	0.02
Labor	0	91.2281	3	89.4737	100	0	89.4737	0
Credit	0	74.4	12	73.5	90	0.04	75.4	0.01

Table 3.5: Highest Testing Accuracy of Classification Algorithms

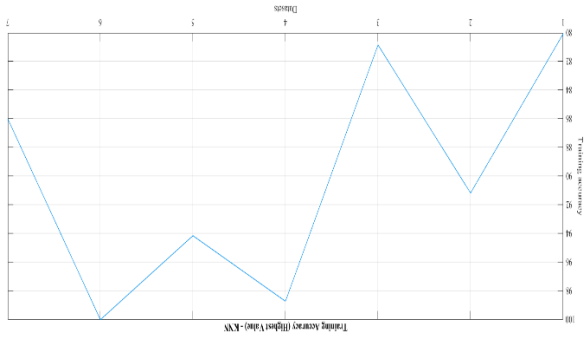
Highest Training Accuracy of Classification Algorithms								
Datasets	**KNN**			**BP**			**Naïve Bayes**	
	Time (Sec)	**Training (%)**	**K**	**Training (%)**	**Hidden Layers**	**Time (Sec)**	**Training (%)**	**Time (Sec)**
Breast Cancer	0.02	80.0699	3	97.9021	20	0.91	75.1748	0
Ionosphere	0.03	91.1681	3	100	50	4.26	82.906	0.01
Glass	0.01	80.8411	3	88.3178	40	1.01	55.6075	0
Unbalanced	0.07	98.715	3	99.4159	10	2	93.3411	0.01
Soybean	0.07	94.1435	3	99.8536	10	4.73	93.7042	0
Labor	0	100	3	100	10	0.12	98.2456	0
Credit	0.1	86	3	99.8	20	8.11	77.2	0

Table 3.6: Highest Training Accuracy

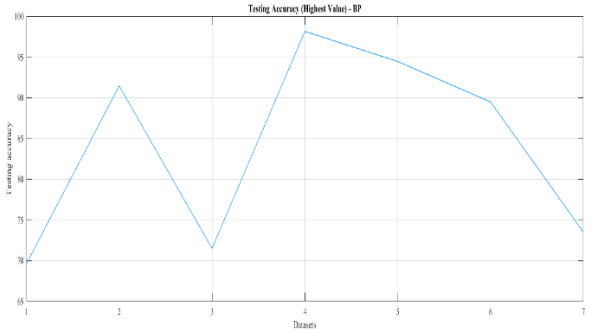
3.7.10 Graphical Representation of Highest Testing & Training Accuracy



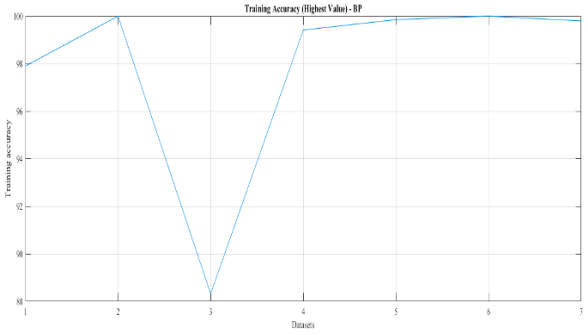
Graph 3.31: Highest Testing Accuracy – KNN



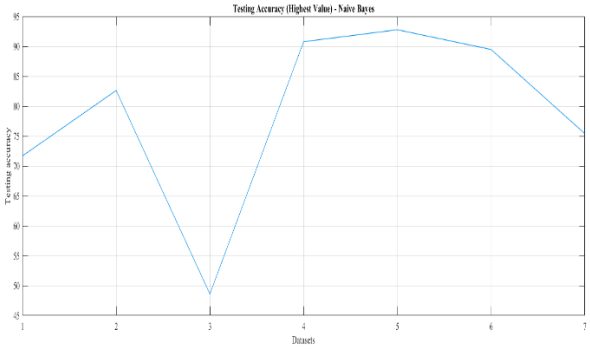
Graph 3.32: Highest Training Accuracy – KNN



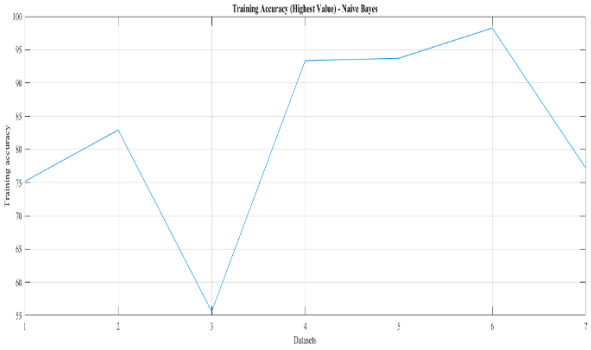
Graph 3.33: Highest Testing Accuracy – BP



Graph 3.34: Highest Training Accuracy – BP



Graph 3.35: Highest Testing Accuracy – Naïve Bayes



Graph 3.36: Highest Training Accuracy – Naïve Bayes

3.8 Experimental Analysis of Clustering (Unsupervised Learning) Algorithms

The experimental study of clustering algorithms and analysis between K- Means and EM is discussed in this portion. The datasets are clustered with the value of 2, 3, and 4 clusters.

3.8.1 K-Means

The idea of K-Means belongs to an earlier time in 1957. However, in 1967, K-Means was first introduced and used by James MacQueen. In 1957, Stuart Lloyds suggested that the standard algorithm of K-Means can be very helpful. This standard form of the algorithm was proposed as a method for pulse code modulation but till 1982 this proposal was not publicized. Industries are using K-Means to a great degree as a partitioned clustering method. The reason for its wide use is that it can be easily implemented and most efficient. Efficient in terms of execution time in the field of data mining. K-Means clustering refers to cluster analysis. It directs data into partition and observations. Observations are direct into K-Clusters in which every observation is in the possession of a cluster with the nearest mean. An iterative technique is used by the algorithm, that's why it is known as the K-Means algorithm. In the community of computer science, K-Means are also known as Lloyd's algorithm. In scientific research and industrial fields, it is a widely used technique. It processes the largest data sets efficiently. It can work also on numeric values. In data mining clustering refers to the method of cluster analysis in which observations are partitioned into K-Clusters. In K-Clusters each observation is in the possession of the cluster with the nearest mean. One of the simplest unsupervised learning algorithms is K-Means because it solves the frequently experienced clustering problems. The algorithm works as the following:

1. First of all, put K-points into the space symbolized by the objects that are being clustered. These k-points represent initial group centroids
2. In the second step, give each object to the group that has the nearest centroid
3. In the third step, after assigning all objects to the group, calculate a new position of the K-Centroids

4. In the fourth and final step, repeat step 2 as well as step 3 until the centroids no more move. This repetition creates a detachment of objects into groups. In this way, we can calculate the metric to be minimized.

3.8.2 Expectation-Maximization (EM) Algorithm

EM stands for Expectation-Maximization Algorithm. When we are satisfied with the results of K-Means algorithms, we can use EM algorithms. The EM algorithm is an iterative method that involves expectation (E-step) and maximization (M-step); to find the local maximum likelihood from the data. Commonly, EM is used on several distributions or statistical models, where there are one or more unknown variables. Therefore, it is called missing or "*latent*" variables.

In other words, we can use the EM algorithm if we have incomplete datasets (for example, it doesn't have groups or labels in the data), but we need to predict its group or label. The group or label has never been observed or recorded but is very important in explaining the data. It is called a "*latent*" variable, meaning missing, hidden, or invisible.

This is normally impossible in statistical models with latent variables. Alternatively, the results are typically a set of interlocking equations. In these equations, the solution to the parameters demands the values of the latent variables and contrariwise. But interchanging one set of equations into the other produces unsolvable equations. The EM algorithm goes forward from the observations in such a way that there is a way to solve these two sets of equations numerically. We can pick arbitrary values, from one of the two sets of unknowns. Then these arbitrary values are used to estimate the second set. Then we use these new values to find a better estimate of the first set. Then keep replacing the two until the resulting values, both meet a fixed point. This is not obvious that this will work. But it can be proved that in this circumstance it may work. It is also proven that the derivatives of the likelihood are zero or close to zero at that point. Successively, it means that the point is either a maximum or a saddle point in this context. Multiple maxima may occur, with no guarantee that the global maximum will be obtained. Some likelihood also has singularities in them such as strange maxima. For example one of the solutions that may be found by the EM algorithm in a mixture model involves setting one of the components to have zero

discrepancies. While the mean parameters for the same component are equal to one of the data points.

Clustering Algorithm Result											
Datasets	Cluster	K-Means					Expectation-Maximization (EM)				
		Cluster Instances		Time (Sec)	Square Error	Iterations	Cluster Instances		Time (Sec)	Square Error	Iterations
Segment	2	0	612 (41%)	0.02	1765.1	2	0	706 (47%)	0.19	0	25
		1	888 (59%)				1	794 (53%)			
	3	0	1027 (68%)	0.01	1363.89	3	0	218 (15%)	0.13	0	3
		1	240 (16%)				1	492 (33%)			
		2	233 (16%)				2	790 (53%)			
	4	0	581 (39%)	0.02	993.22	4	0	584 (39%)	0.17	0	5
		1	237 (16%)				1	207 (14%)			
		2	220 (15%)				2	219 (15%)			
		3	462 (31%)				3	490 (33%)			
Diabetes	2	0	500 (65%)	0.01	149.51	2	0	432 (56%)	0.09	0	46
		1	268 (35%)				1	336 (44%)			
	3	0	132 (17%)	0.01	127.72	3	0	228 (30%)	0.05	0	9
		1	268 (35%)				1	203 (26%)			
		2	368 (48%)				2	337 (44%)			
	4	0	166 (22%)	0.01	119.18	4	0	133 (17%)	0.06	0	9
		1	268 (35%)				1	176 (23%)			
		2	220 (29%)				2	354 (46%)			
		3	114 (15%)				3	105 (14%)			

Table 3.7: Clustering Algorithm Results

3.9 Summary

In this chapter different algorithms are discussed and the experimental implementation of classification algorithms KNN, Backpropagation, and Naïve Bayes. And shows the results in tables and graphical representations of all results. I also discussed the Clustering algorithms K-Mean and Exception Maximization (EM). The results are shown in the table.

Chapter 4

Results and Discussion

4.1 Overview

Classification analysis is an important tool for studying labeled data in data mining. Classification analysis consists of several steps like preprocessing, algorithm development, validity, and evaluation. The findings are described in chapter 3 in detail in tables and graphically represented by using Matlab. In this chapter, results are discussed, as also discussed the factors that can affect the results.

4.2 Supervised Learning Algorithms

4.2.1 KNN Results Analysis

This experimental analysis shows that the accuracy affects while increasing the number of K. in all 7 datasets the accuracy increases while increasing the no of K values. In every dataset the different value of K gives the highest accuracy and, in some datasets, it remains the same. Mostly when the value of K is 3, 6, and 12 the accuracy is high. The training time, testing time, testing accuracy, and Training Accuracy of all datasets at all K values are shown in Table 3.2. The KNN gives the highest Testing accuracy value of 98.5981% for the dataset named Unbalanced and the lowest accuracy of 60.7470% for the dataset named Glass. The KNN gives the highest Training accuracy value of 100% for the dataset named Labor and the lowest accuracy of 64.486% for the dataset named Glass.

Breast cancer is showing the highest testing accuracy of 73.7762 % and training accuracy of 80.0699%, dataset ionosphere showing the testing accuracy of 86.6097% and training accuracy of 91.1681%, dataset Glass shows the highest testing accuracy of 71.9626 % and training accuracy 80.8411%, dataset Unbalanced shows the testing accuracy of 98.5981% and training accuracy of 98.715%, dataset Soybean shows the testing accuracy of 91.3616% and training accuracy of 94.1435%%, dataset Labor shows the Testing accuracy of 91.2281% and training accuracy 100% and dataset Credit shows the highest testing accuracy of 74.4% and training accuracy 86.0%.

4.2.2 Back Propagation Results Analysis

While using the multilayer perception (backpropagation algorithm) with the setting of different parameters in weka especially changing the hidden nodes from 10 to 100 the training time increases while increasing the no of hidden nodes and accuracy is also affected. The backpropagation takes more time as compared to the other three algorithms. The backpropagation gives the highest testing accuracy of 98.1308% for the dataset named Unbalanced and the highest training accuracy of 100% for the datasets labor and Ionosphere and the lowest testing accuracy value of 66.039% for the dataset named Breast Cancer and the lowest training accuracy of 82.243% for the dataset named Glass.

The highest testing and training accuracy of dataset Breast Cancer is 69.5804%, 97.021%, dataset ionosphere is 91.453%, 100%, dataset Glass is 71.4953%, 88.3137%, dataset Unbalanced is 98.1308%, 99.4159%, dataset Soybean 94.4363%, 99.8536, dataset Labor is 89.4737 %, 100% and dataset Credit 72.5%, 99.8% respectively.

4.2.3 Naïve Bayes Results Analysis

Naïve Bayes is one of the simplest classifiers with less computational complexity in handling new instances. The experiment is performed on all seven datasets using the weka tool to measure the performance of Naïve Bayes and the time to build the model. The table shows the Testing and training accuracy of the dataset for breast cancer is 71.6783%, 75.1748%, dataset ionosphere is 82.6211%, 82.906%, dataset Glass shows 48.5981%, 55.6075%, dataset Unbalanced is 90.771%, 93.341%, dataset Soybean is 92.7722%, 93.7042%, dataset Labor shows 89.4737%, 98.2456%, and dataset Credit shows 77.2%, 70% respectively. The highest accuracy of Naïve Bayes is 92.7722% in the Soybean dataset and the highest training accuracy is 98.2456% for the Labor dataset. The lowest Testing Accuracy is 48.5981% and Training accuracy is 55.6075% for dataset Glass.

4.3 Unsupervised Learning Algorithms

4.3.1 Result Analysis of Clustering Algorithms

Cluster analysis is an important tool for studying unlabeled data in data mining. Cluster analysis consists of several steps like preprocessing, algorithm development, validity, and evaluation. In this research work, I focus on the two algorithms of unsupervised machine learning technique (clustering) for the analysis of performance by using two datasets and use the weka tool for cluster analysis.

4.3.2 K-means Analysis

The K-means algorithm classifies the *k* number of centroids and then assigns every data point to the nearest cluster while keeping the centroids as small as possible. The K-means algorithm in data mining starts with the first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. The following Fig 4.1, 4.2, and 4.3 show the result of the K-Mean algorithm on the bases of cluster sizes 2, 3, and 4. The centroid of each cluster is shown in the result window, along with measurements of the number and percent of instances allocated to each cluster. Each cluster centroid is represented by a mean vector. This cluster can be used to describe a cluster. Graph 4.1 shows the graphical representation of the K-Mean algorithm result.

Table 3.7 shows the cluster instances on the bases of cluster number. One column describes the number of iterations. There are 13 iterations when the cluster size is 2 and 32.20% of data is incorrectly clustered. There are 8 iterations on 3 cluster size and 48.30% of data is incorrectly clustered. And 6 iterations on 4 clusters with 48.04% data incorrectly clustered.

```
                        Cluster#
Attribute    Full Data          0          1
                (768.0)    (515.0)    (253.0)
============================================
preg              3.8451     2.0835     7.4308
plas            120.8945   115.3282   132.2253
pres             69.1055    65.9903    75.4466
skin             20.5365    21.8194    17.9249
insu             79.7995    85.0194    69.1739
mass             31.9926    31.7751    32.4352
pedi              0.4719     0.4708     0.4741
age              33.2409    26.7728    46.4071



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      515 ( 67%)
1      253 ( 33%)


Class attribute: class
Classes to Clusters:

   0   1  <-- assigned to cluster
 380 120 | tested_negative
 135 133 | tested_positive

Cluster 0 <-- tested_negative
Cluster 1 <-- tested_positive

Incorrectly clustered instances :       255.0     33.2031 %
```

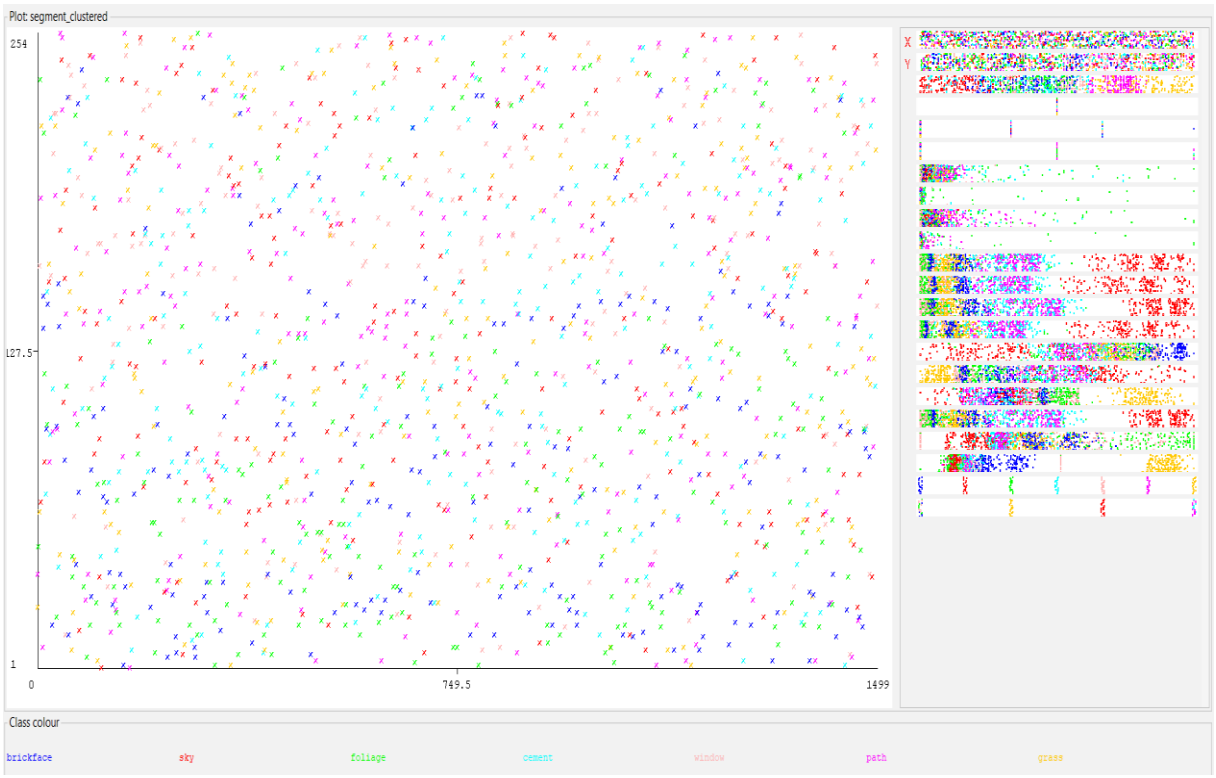
Fig 4.1: K-Mean Result – No of Cluster 2

```
Attribute    Full Data          0          1          2
                (768.0)    (208.0)    (229.0)    (331.0)
=========================================================
preg              3.8451     2.0144     7.7031     2.3263
plas            120.8945   137.0913   130.2969   104.2115
pres             69.1055    73.226     76.4716    61.4199
skin             20.5365    34.7452    16.9389    14.0967
insu             79.7995   175.1058    58.8297    34.4169
mass             31.9926    36.6567    32.3948    28.7834
pedi              0.4719     0.5931     0.4713     0.3961
age              33.2409    29.2933    46.8035    26.3384



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      208 ( 27%)
1      229 ( 30%)
2      331 ( 43%)


Class attribute: class
Classes to Clusters:

   0   1   2  <-- assigned to cluster
 112 110 278 | tested_negative
  96 119  53 | tested_positive

Cluster 0 <-- No class
Cluster 1 <-- tested_positive
Cluster 2 <-- tested_negative

Incorrectly clustered instances :       371.0     48.3073 %
```

Fig 4.2: K-Mean Result – No of Cluster 3

```
Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3
                (768.0)    (173.0)    (222.0)     (36.0)    (337.0)
================================================================
preg              3.8451     2.1214     7.7297     3.5556     2.2018
plas            120.8945   143.9191   130.0495        117   103.4599
pres             69.1055     73.104    77.4865     0.6667    68.8427
skin             20.5365    35.0289    16.8964          2    17.4748
insu             79.7995   194.6879    59.8063     0.6944    42.4421
mass             31.9926    36.9064    32.3122    25.7639    29.9249
pedi              0.4719     0.6211      0.465     0.3932     0.4082
age              33.2409    29.8613    46.8874    30.4444    26.2849



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      173 ( 23%)
1      222 ( 29%)
2       36 (  5%)
3      337 ( 44%)

Class attribute: class
Classes to Clusters:

   0   1   2   3  <-- assigned to cluster
  87 108  20 285 | tested_negative
  86 114  16  52 | tested_positive

Cluster 0 <-- No class
Cluster 1 <-- tested_positive
Cluster 2 <-- No class
Cluster 3 <-- tested_negative

Incorrectly clustered instances :       369.0     48.0469 %
```

Fig 4.3: K-Mean Result – No of Cluster 4



Graph 4.1: Graphical Representation of K-Mean Algorithm Results

4.3.3 Expectation-Maximization (EM) Analysis

The following Fig 4.4, 4.5, and 4.6 show the result of the K-Mean algorithm on the bases of cluster sizes 2, 3, and 4. The centroid of each cluster is shown in the result window, along with measurements of the number and percent of instances allocated to each cluster. Each cluster centroid is represented by a mean vector. Graph 4.2 shows the graphical representation of the EM algorithm result.

Table 3.7 shows the cluster instances on the bases of cluster number. One column describes the number of iterations. There are 4 iterations when the cluster size is 2 and 33.98% of data is incorrectly clustered. There are 13 iterations on 3 cluster size and 57.01% of data is incorrectly clustered. And 12 iterations on 4 clusters with 45.83% data incorrectly clustered.

```
mass
  mean          30.2275  33.4198
  std. dev.      8.3431   7.1719

pedi
  mean           0.3802    0.546
  std. dev.      0.2031   0.3907

age
  mean          24.9083  39.9785
  std. dev.       3.516   11.759


Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       353 ( 46%)
1       415 ( 54%)


Log likelihood: -29.12836


Class attribute: class
Classes to Clusters:

   0   1  <-- assigned to cluster
 296 204 | tested_negative
  57 211 | tested_positive

Cluster 0 <-- tested_negative
Cluster 1 <-- tested_positive

Incorrectly clustered instances :       261.0     33.9844 %
```

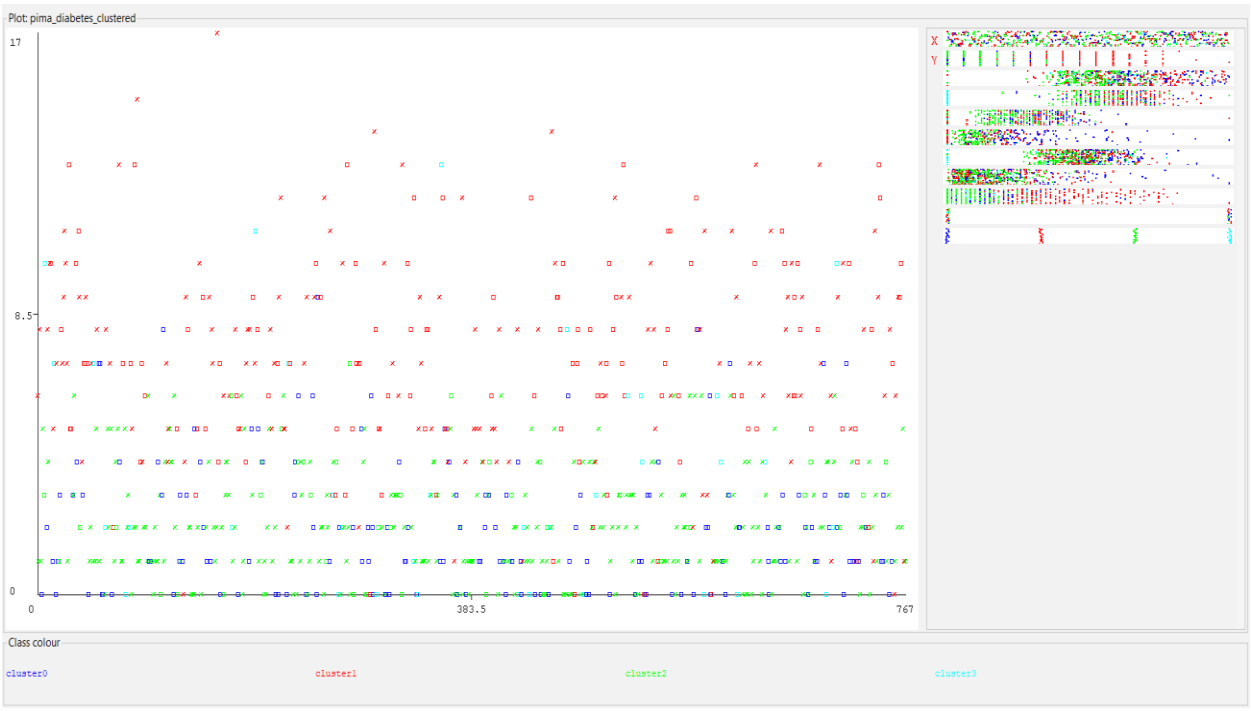
Fig 4.4: EM Result – No of Cluster 2

```
  std. dev.      8.9628   7.8967   5.8762

pedi
  mean           0.3491   0.5238   0.5432
  std. dev.      0.2085   0.3844   0.3345

age
  mean          32.3749  24.5802  44.2618
  std. dev.      9.7378   3.0226  11.1922


Time taken to build model (full training data) : 0.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       369 ( 48%)
1       250 ( 33%)
2       149 ( 19%)


Log likelihood: -23.10594


Class attribute: class
Classes to Clusters:

   0   1   2  <-- assigned to cluster
 234 195  71 | tested_negative
 135  55  78 | tested_positive

Cluster 0 <-- tested_positive
Cluster 1 <-- tested_negative
Cluster 2 <-- No class

Incorrectly clustered instances :       438.0     57.0313 %
```

Fig 4.5: EM Result – No of Cluster 3

```
pedi
  mean           0.646    0.463   0.4018   0.3884
  std. dev.     0.4646   0.3016   0.2363   0.2506

age
  mean          29.5771  45.9571  25.7186  30.7143
  std. dev.      7.8184  10.1528   4.6182   9.6793


Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       152 ( 20%)
1       263 ( 34%)
2       318 ( 41%)
3        35 (  5%)


Log likelihood: -26.93384


Class attribute: class
Classes to Clusters:

   0   1   2   3  <-- assigned to cluster
  74 127 280  19 | tested_negative
  78 136  38  16 | tested_positive

Cluster 0 <-- No class
Cluster 1 <-- tested_positive
Cluster 2 <-- tested_negative
Cluster 3 <-- No class

Incorrectly clustered instances :       352.0     45.8333 %
```

Fig 4.4: EM Result – No of Cluster 4



Graph 4.2: Graphical Representation of EM Algorithm Results

Chapter 5

Conclusion

5.1 Conclusion

The above research work took a look at the different types of machine learning paradigms like supervised and unsupervised learning. It is an application of artificial intelligence. Also, it allows software applications to become accurate in predicting outcomes.

The primary aim is to allow computers to learn automatically without human intervention. As humans become more addicted to machines, we're witness to a new revolution that's taking over the world and that is going to be the future of Machine Learning. Also learned about the various algorithms that are used for machine learning classification. These algorithms are used for a variety of tasks in classification. And described clustering and how clustering has brought advanced data analysis techniques to unlabeled datasets. I give an overview of various types of clustering algorithms.

An experimental analysis of three supervised learning algorithms is done, measuring the Testing time, training time, testing accuracy, and training accuracy upon 7 different datasets. The final results are shown in Table 3.5, Table 3.6, Table 3.8, and Table 3.9. The final analysis shows that the Backpropagation shows constant high accuracy upon a maximum number of datasets. Backpropagation shows the highest accuracy upon 4 datasets out of 7. And the last is Naïve Bayes. There is not as such constant ratio inaccuracy. When comparing the performance of algorithms, It is found that Backpropagation is a better algorithm in most of the datasets. Different algorithms show the different results upon different datasets on the base of accuracy. After this experiment analysis, we can predict on the bases of the given results, every algorithm gives a different accuracy on each dataset. If one algorithm gives the highest accuracy on one dataset, it's not mean that it provides the highest accuracy on all other datasets. The parameters also play a vital role in the performance of algorithms. The change in parameters affects the performance of algorithms. In the backpropagation increase in the number of hidden layers increases the accuracy but also increases the time. Backpropagation gives the highest accuracy ratio but it takes a lot of time to

perform tasks. But the performance of the Backpropagation algorithm is constant on datasets as compared to other algorithms. The performance wise first comes from backpropagation, KNN, and in the last Naïve Bayes. According to the other analysis factor Time, the KNN shows a fast speed and less time to execute the results as compared to the other algorithms.

The unsupervised learning techniques (clustering) of the algorithms K-means and EM are tested upon two datasets. No clustering and classification algorithm can be universally used to solve all problems. Usually, algorithms are designed with certain assumptions and some type of bias. In this sense, it is not accurate to say that "best" in the context of classification and clustering algorithms, although some results upon some analysis are possible as we did in this research work. These analyses are mostly based on some specific applications, under certain conditions and the results may differ, and quite different if the conditions change.

The building blocks of data mining are the evolution of a field that includes database management systems(DBMS), Statistics, Artificial Intelligence(AI), and Machine Learning(ML). Data mining algorithms work best for numerical data, and various data mining techniques have evolved in this process. The field of data mining has been greatly influenced by the development of fourth-generation programming languages and various related computing techniques. In, the early days of data mining most of the algorithms only deals with statistical techniques. Ans later on, it advanced with various computing techniques like AI, ML, and Pattern Reorganization.

Various data mining applications have been successfully implemented in several domains like health care, finance, retail, telecommunication, fraud detection, risk analysis, etc. The increasing complications in various fields have posed new challenges to data mining; the various challenges include different data formats, data locations, computation and networking resources, research and scientific fields, business challenges, etc.

References

[1] E. Alpaydın, in *Introduction to Machine Learning*, The MIT Press, 2010, pp. 180-181.

[2] C. Stedman, "TechTarget," 9 2021. [Online]. Available: https://www.techtarget.com/searchbusinessanalytics/definition/data-mining.

[3] D. Y. Li Deng, "Deep Learning: Methods and Applications," *Foundation and Trends in Signal Processing*, vol. 7, no. 3-4, pp. 197-387, 2013.

[4] T. J. (. D. M. I. D. W. W. O. C. K. Y. H. Kallio A., "Data Mining," in *Encyclopedia of Systems Biology*, New York, 2013, pp. 525-528.

[5] C. P. N. S.-M. R. M.-G. Jesús de-Prado-Gil, "To predict the compressive strength of self-compacting concrete with recycled aggregates utilizing ensemble machine learning models," *Case Studies in Construction Materials*, vol. 16, no. e01046, pp. 2214-5095, 9 2022.

[6] B. a. Y. R. Chowdary, "A Survey on Applications of Data Mining Techniques," *International Journal of Applied Engineering Research*, vol. 13, no. 7, pp. 5384-539, 2018.

[7] J. Delua, "IBM," IBM Analytics, 12 03 2021. [Online]. Available: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.

[8] A. A. R. Sathya, "Analysis of Supervised and Unsupervised Learning Algorithms for Pattern Classification," *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 2, pp. 34-38, 2013.

[9] A. Dey, "Machine Learning Algorithms: A Review," vol. 7, pp. 1174-1179, 2016.

[10] S. L. Z. a. P. P. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Journal of Intelligent Learning Systems and Applications*, vol. 7, no. 2, pp. 3-24, 2015.

[11] S. L. Z. a. P. P. Kotsiantis, "Machine learning: A review of classification and combining techniques," *Artificial Intelligence Review,* vol. 26, no. 3, pp. 159-190, 2006.

[12] M. P. a. V. J. Jaiganesh, "A Literature Review on Supervised Machine Learning Algorithms and Boosting Process," *International Journal of Computer Applications,* vol. 169, no. 8, pp. 32-35, 2017.

[13] C. Donalek, "Supervised, and Unsupervised learning," Astronomy Colioquia, USA, 2011.

[14] A. E. Mohamed, "Comparative Study of Four Supervised Machine Learning Techniques for Classification," *International Journal of Applied Science and Technology ,* vol. 7, no. 2, pp. 5-18, 2017.

[15] A. a. J. M. Jain, "Text classification by combining text classifiers to improve the efficiency of classification," *International Journal of Computer Application,* vol. 6, no. 2, pp. 2250-1797, 2016.

[16] D. A. D. Punjani, A Comprehensive Study of Various Classification Techniques in Medical Application using Data Mining, 2018.

[17] A. Pant, "Towards Data Science," 22 01 2019. [Online]. Available: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148.

[18] A. N. T. a. A. S. Singh, "A review of supervised machine learning algorithms, in Computing for Sustainable Global Development," in *3rd International Conference on. 2016. IEEE*, India, 2016.

[19] T. a. V. B. Siriteerakul, "Support Vector Machine accuracy improvement with k-means clustering," in *Computer Science and Engineering Conference (ICSEC).*, 2013.

[20] T. R. T. a. J. F. Hastie, "Unsupervised learning, in The elements of statistical learning," Springer, 2009, pp. 485-585.

[21] T. C. R. a. K. N. Sajana, "A survey on clustering techniques for big data mining," *Indian Journal of Science and Technology*, vol. 9, no. 3, pp. 1-12, 2016.

[22] T. Mitchell, Machine learning (McGraw-hill international editions computer science series), 1997.

[23] C.-H. a. H.-C. Y. Lee, "Implementation of Unsupervised and Supervised Learning Systems for Multilingual Text Categorization. in Information Technology," in *Fourth International Conference, IEEE*, 2007.

[24] A. e. a. Fahad, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE transactions on emerging topics in computing*, vol. 2, no. 3, pp. 267-279, 2014.

[25] R. a. A. A. Sathya, "Analysis of supervised and unsupervised learning algorithms for pattern classification," *Advanced Research in Artificial Intelligence*, vol. 2, no. 2, pp. 34-38, 2013.

[26] M.-C. a. C.-H. C. Su, "A modified version of the K-means algorithm with distance-based," *Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 674-680, 2001.

[27] J. a. M. W. Hartigan, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society - Series C(Applied Statistics)*, vol. 28, no. 1, pp. 100-108, 1979.

[28] O. A. Abbas, "Analysiss Between Data Clustering Algorithms," *Int. Arab J. Inf. Technol.*, vol. 5, pp. 320-325, 2008.

[29] U. F. a. C. R. P.S. Bradley, "Scaling Clustering Algorithms to Large Databases," KDD-98 Proceedings, USA, 1998.

[30] D. M. W. WEKA, "TataSoft," 12 11 2021. [Online]. Available: https://www.tatvasoft.com/blog/data-mining-with-weka/.

[31] B. D. Vijay Kotu, "Predictive Analytics and Data Mining," in *Data Mining Process*, Vijay Kotu, Bala Deshpande, 2015, pp. 17-36.