A
Major Project-II Report
On

DATA MINING USING SUPERVISED MACHINE LEARNING TECHNIQUES

Submitted in Partial Fulfilment of the Requirement for the Degree of

MASTER OF TECHNOLOGY
In
COMPUTER SCIENCE AND ENGINEERING

By
RITU
2K14/CSE/16

Under the Esteemed guidance of
Mr. MANOJ SETHI

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahabad Daulatpur, Main Bawana Road,
Delhi-110042.

JUNE, 2016

CERTIFICATE

This is to certify that Major Project-II Report entitled **“Data mining using supervised machine learning techniques”** submitted by **Ritu, Roll No. 2K14/CSE/16** for partial fulfilment of the requirement for the award of degree Master of Technology (Computer Science and Engineering) is a record of the candidate work carried out by her under my supervision.

Mr. Manoj Sethi
Department Of Computer Science & Engineering
Delhi Technological University

DECLARATION

I hereby declare that the major Project-II work entitled "**Data mining using supervised machine learning techniques**" which is being submitted to Delhi Technological University, in partial fulfilment of requirements for the award of degree of Master Of Technology (Computer Science and Engineering) is a bonafide report of Major Project-II carried out by me. The material contained in the report has not been submitted to any university or institution for the award of any degree.

Ritu
2K14/CSE/16

ACKNOWLEDGEMENT

First of all, I would like to express my deep sense of respect and gratitude to my project supervisor Mr. Manoj Sethi for providing the opportunity of carrying out this project and being the guiding force behind this work. I am deeply indebted to him for the support, advice and encouragement he provided without which the project could not have been a success.

Secondly, I am grateful to Dr. O.P.Verma, HOD, Computer Science & Engineering Department, DTU for his immense support. I would also like to acknowledge Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out. Last but not the least I would like to express sincere gratitude to my parents and friends for constantly encouraging me during the completion of work.

Ritu
University Roll no: 2K14/CSE/16
M.Tech (Computer Science & Engineering)
Department of Computer Science and Engineering
Delhi Technological University
Delhi – 110042

ABSTRACT

Every day human beings are generating vast data and this data comes from different sources, be it online or offline. It may be in the form of documents, may be in graphical formats, may be the video or may be the records (varying array). Since the data is available in different formats, appropriate action needs to be taken not only to analyze the data but also to fetch important information and patterns from it and maintain the data .The data should be made available as and when required by the clients. The data should be retrieved from the database to help them make better decision .This technique is actually what we call data mining.

Machine learning is a subfield of computer science which involves the study and construction of algorithms that can learn from and make predictions on data. First, a model is built from a training set of input observations so that the predictions can be data driven and then machine learning algorithms operates on test data for generating the predicted outcome. Here strict static program instructions are not followed. Instead historical data is considered for making the prediction.

Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

In this work, two popular machine learning algorithms: k-nearest neighbours and naïve bayes are studied and a new hybrid algorithm have been proposed using the best features of both these algorithms. In recent years, there has been a dramatic increase in the use of machine learning techniques within the healthcare systems to analyse, predict and classify clinical data. Therefore, we have selected 3 datasets containing the health related data. All three algorithms have been implemented in python and have been run against all the datasets.

Keyword: data mining, machine learning, knn, naïve bayes.

Table of Contents

List of Figures

CHAPTER 1- INTRODUCTION

1.1 DATA MINING

Data mining is the process of analyzing data from various perspectives from large database, extracting the hidden useful information from it and summarizing it so that it can be used to increase revenue, cut the costs, or both.

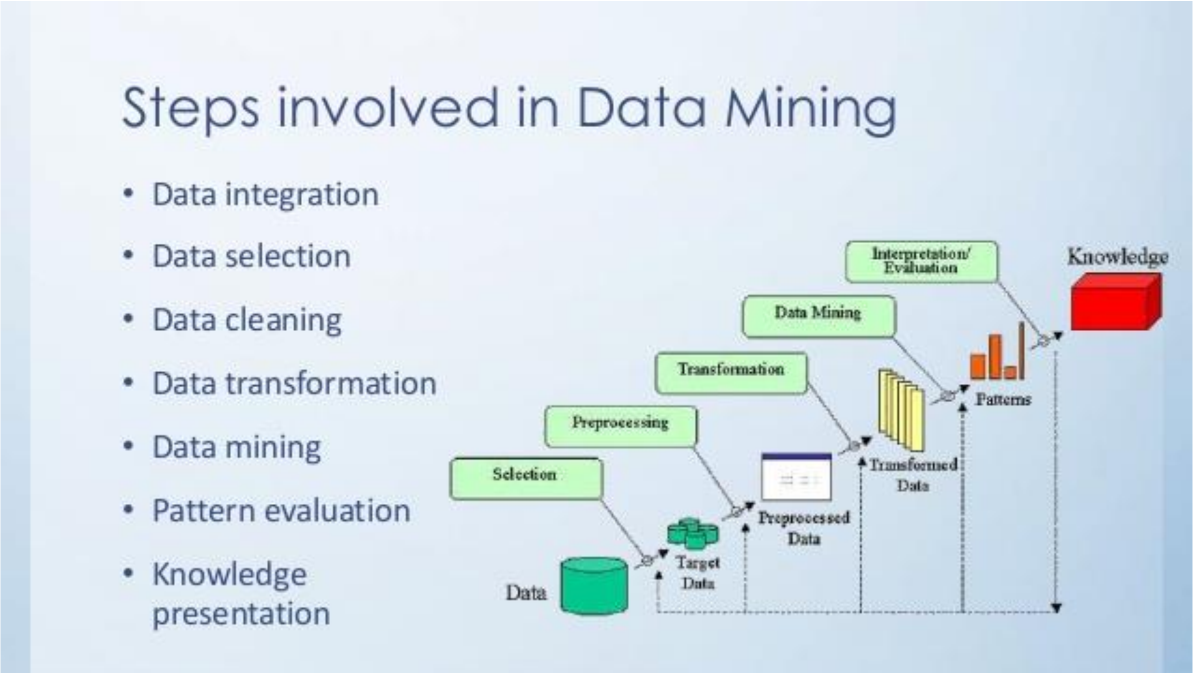
Data mining is a powerful new technology with great potential to help companies concentrate on the most important information in their data warehouses. There exists a difference between data in the databases and a data warehouse. The data stored in a database is always in the structured form whereas in the data warehouse, the data may or may not be present in the structured form. The structure of the data may be defined to make it compatible for processing. Therefore, in data mining, we also need to primarily focus on cleansing the data so as to make it feasible for further processing. The process of cleansing the data is also called noise elimination or noise reduction or feature elimination. The process of cleansing data can be either made by using tools such as ETL, tools available in the market or may be done by using various suitable techniques available

Various Data mining tools are available which predict future trends and behaviors and help the companies in making knowledge-driven decisions. Traditionally business questions were very time consuming but the data mining tools are able to resolve them comparatively easily with reduced time consumption. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.[6]

1.2 HOW DOES DATA MINING WORK

With the advancement of large-scale information technology, transaction and analytical systems have also been evolved separately and data mining provides the link between the two. Based on open-ended user queries, the relationships and patterns in the stored transactional data are analyzed using data mining softwares. Various types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to classify the new data into one of the predetermined classes. For example, customer purchase data from a restaurant chain can be mined to determine the time of customers visit and what they typically order. More customers can be attracted using this information by having daily specials.
- **Clusters**: Data items are grouped together in clusters taking logical relationships or consumer preferences into consideration. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Associations between data can be identifies using data mining. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, Based on a consumer's purchase of sleeping bags and hiking shoes, the likelihood of the purchase of a backpack can be predicted by an outdoor equipment retailer.



Steps involved in Data Mining

- Data integration
- Data selection
- Data cleaning
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation

Figure 1

A brief about the steps in the process of data mining are:

- **Data Integration:**

This is the very first step in data mining. Data is collected from different sources and stored at one place**.**

- **Data Selection:**

  All the data that we have collected in first step may not be useful to us. So in this step, from all the data, we select only that data that we actually want to work on.

- **Data Cleaning**:

  It is always the case that data we have collected is clean. There may present errors or missing values. The data may be noisy or inconsistent data. So, different techniques need to be applied to get rid of such anomalies.

- **Data Transformation**:

  The data that we get after cleaning may still not be ready for mining. So we need to transform it to forms that are appropriate for mining. There are various techniques for accomplishing this task like are smoothing, aggregation, normalization etc.

- **Data Mining**:

  Now the data is ready for applying data mining techniques on the data to discover the interesting patterns. There are various machine learning techniques like clustering , classification, association analysis etc.

- **Pattern Evaluation**:

  In this step, patterns are generated after the major step of data mining. Then data is visualized, transformed and redundant patterns if present are removed.

- **Knowledge Presentation**:

  After the application software has analyzed the data, this is the final step in which the data is presented in a useful format, like a graph or a table.

1.3 DATA MINING TASKS

There are usually six types of tasks that can be performed in data mining:

1. **Anomaly Detection**:

Anomaly detection is the process in which unusual data records are identified. These are the records that might be interesting or they might be data errors that require further investigation.

2. **Association Rule Mining**:

   Association rule learning is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness

3. **Clustering**:

   This is an unsupervised learning problem. In clustering, a set of inputs is to be divided into groups which are not known beforehand unlike in classification

4. **Classification**:

   In classification, input training set is fed to the algorithm which are classified into two or more classes, and based on the input provided , a model is produced which assigns a class label to the new data or multiple labels in case of multi-label classification. This is typically handled using supervised machine learning. Spam filtering is an example of classification, where email messages are inputs and they are classified in one of the classes: ‘spam’ and ‘not spam’.

5. **Regression**:

   Regression can be explained as attempting to find a function which can model the data with the least error. It is also a supervised learning problem which deals with continuous outputs rather than discrete

6. **Summarization**:

   Summarization means to represent the dataset compactly. It includes visualization and report generation. It refers to the creation and study of the visual representation of data. The data can be multidimensional and the complex relationships between data is visually interpreted using graphics tools for the illustration of data relationships.[6]

1.4 MACHINE LEARNING

Machine learning is a subfield of computer science which involves the study and construction of algorithms that can learn from and make predictions on data. First, a model is built from a training set of input observations so that the predictions can be data driven and then machine learning algorithms operates on test data for generating the predicted outcome. Here strict static program instructions are not followed. Instead historical data is considered for making the prediction.

1.5 TYPES OF MACHINE LEARNING

Based on the nature of learning signal or feedback available to a learning system, machine learning is classified into three broad categories:

1. **Supervised learning**:

   The algorithm is presented with example inputs and their desired outputs, given by a "supervisor", and the goal is to learn a general rule that maps inputs to outputs.

2. **Unsupervised learning**:

   The algorithm does not need to provide with labels. It is left on its own for finding structure in its input. Unsupervised learning can be looked upon as a goal in itself which is to discover hidden patterns in data. This is the reason it is also referred to as feature learning.

3. **Reinforcement learning**:

   The algorithm communicates with a dynamic environment in which it must perform a certain goal such as driving a vehicle. This type of learning does not require a supervisor. It does not need to be told whether it is approaching its goal. Another example is learning to play a game by playing against an opponent.

There exists another type of machine learning between supervised and unsupervised learning, that is **semi-supervised learning,** where an incomplete training signal with some of the target outputs missing is provided by the supervisor.[11]

1.6 MACHINE LEARNING TECHNIQUES:

There are certain commonly used techniques in machine learning:

1. **Artificial neural networks**

   An artificial neural network learning algorithm can be explained as is a learning algorithm which resembles biological neurons in terms of the structure and functional aspects. Computations are carried out by first representing them in structured form as an interconnected group of artificial neurons. These interconnected neurons then process the information and follows a connectionist approach to perform the computation. Modern neural networks are constructed using non-linear statistical data modeling tools. They are usually helpful in constructing models that have complex relationships between inputs and outputs in order to find patterns in data.[13]

2. **Association Rule Learning**:

   Association rule learning is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness

3. **Decision Tree Learning**:

   It is a decision support tool that uses tree-shaped structures such as graph which represents the set of decisions and their possible outcomes. Rules are generated for the classification of a dataset through these decisions. It is one of the ways to represent an algorithm. Decision trees are specifically used in decision analysis where there is a need to identify a strategy which has a higher chances to reach a goal. For example, operations research. Machine learning is also a field where decision trees are very popular. Some of the decision tree methods are Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. Decision rules can be obtained from these techniques that can be applied to a new and unclassified dataset to predict which records will belong to which category

4. **Deep Learning**:

   In deep learning , the artificial neural network consists of multiple hidden layers. The concept of deep learning had developed a lot during the last few years because of the reduced hardware prices and the development of graphical user interface for personal use. This approach tries to model in a way similar to how the human brain processes light and sound into vision and hearing. Computer vision and speech recognition are a few successful application of deep learning.

5. **Inductive Logic Programming**

   Inductive logic programming (ILP) is an approach to rule learning using logic programming as a uniform representation for input examples, background knowledge, and hypotheses. Given an encoding of the known background knowledge and a set of examples represented as a logical database of facts, an ILP system will derive a hypothesized logic program that entails all positive and no negative examples. Inductive programming is a related field that considers any kind of programming languages for representing hypotheses (and not only logic programming), such as functional programs.

6. **Support Vector Machines:**

   Support vector machines are basically supervised learning methods which are mainly used for classification and regression. Given a set of training instances, each instance belongs to one of the two categories, an SVM training algorithm builds a model that predicts whether a new instance belongs to the one category or the other.

7. **Clustering:**

   Given a set of training data with each data instance having pre-designated in one of the clusters(subsets) ,clustering is the process of assigning new data instances one of the clusters based on the observations drawn from training data. Different clustering techniques make different assumptions on the structure of the data, often defined by some similarity metric and evaluated for example by internal compactness (similarity between members of the same cluster) and separation between different clusters. Other methods are based on estimated density and graph connectivity. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis.

8. **Baysian Networks**:

   A Bayesian network also known as belief network or directed acyclic graphical model is based on probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph (DAG). For example, a Bayesian network can be used to represent the probabilistic relationships between diseases and their symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Efficient algorithms exist that perform inference and learning.

9. **Genetic Algorithm**:

   It is a search heuristic that resembles the process of natural selection. It is mainly used to generate useful solutions to optimization and search problems whose design is based on the concepts of natural evolution.[7]

1.7 INTRODUCTION TO PYTHON

In this work, the high-level programming language, ‘python’ has been used for the implementation purpose. So, here is a brief about the language.

Python is a powerful, high-level programming language. Guido van Rossum created this language during 1985- 1990 at the National Research Institute for Mathematics and Computer Science in the Netherlands. It is derived for languages like ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages. Python is open source. Its source code is available under GNU General Public License.

Following are certain basic features of Python:

- **Python is Interpreted:**
  Like C and C++, Python need not to be compiled before execution. But it only has to be interpreted at runtime. This is similar to PERL and PHP.
- **Python is Interactive:**
  You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- **Python is Object-Oriented:**
  Python follows the concept of Object-Oriented programming that encapsulates code within objects.
- **Python is a Beginner's Language:**
  Python is very easy to learn as compared to other languages. If one has a basic understanding of programming terminologies, there is nothing much left to do for him and it would be a huge plus point.

- **Easy-to-learn**: Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.

- **Easy-to-read:** Python is designed to be highly readable. It makes use of English keywords frequently and the code is clearly defined.

- **Easy-to-maintain:** Python's source code is fairly easy-to-maintain.

- **A broad standard library:** Python has a portable standard library which is compatible on different platforms like UNIX, Windows, and Macintosh.

- **Portable:** Python can run on a wide variety of hardware platforms and has the same interface on all platforms.

- **Extendable:** New functions and data types implemented in C or C++ can be easily incorporated in the Python interpreter as extensions. Low-level modules can be added to the Python interpreter which enable programmers to customize their tools to be more efficient.

- **Databases:** Python provides interfaces to all major commercial databases.
- **GUI Programming:** Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable:** Python provides a better structure and support for large programs than shell scripting.

In addition to the above-mentioned features, Python has a long list of good features, few of them are listed below:

- Functional and structured programming methods as well as OOP are also supported by it.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It has an efficient set of data structures. High-level dynamic data types are provided by python and also it supports dynamic type checking.
- Automatic garbage collection is supported by it.
- Its integration with C, C++, COM, ActiveX, CORBA, and Java is also very easy.

1.8 MOTIVATION

There is an elementary problem that often arises in a variety of fields like image processing, pattern recognition, machine learning, and statistics, classification. Classification is an important part of exploratory data mining. Many algorithms exist to classify the data into categories. Two of them are k-nearest neighbours and naïve bayes. Both algorithms have nearly same accuracy levels. To get improved results we have combined their features to construct a hybrid classifier.

In this dissertation, we put emphasis on developing a hybrid classifier using k-nearest neighbours and naïve bayes classifier to minimize computational efforts of clustering.

1.9 RESEARCH OBJECTIVE

The objectives of this research work are as follows:

- To develop an algorithm for efficient classification of numerical data.
- To use the existing algorithms for constructing a new hybrid taking motivation from the previous work done in this area.
- To improve the algorithm developed so that it can be applied to low dimensional as well as high dimensional data.
- To find the medical application of the algorithm in the diagnosis of various diseases.

1.10 THESIS ORGANISATION

The dissertation starts with Chapter 1 that provides the Introduction to the work. Chapter 2 gives the literature survey of work done in parts. The works of scholars in fields of classification of datasets using the concerned algorithms using the medical data have been studied. Chapter 3 provides the research methodology used to reach the resultant hybrid algorithm. It emphasizes on the k-nearest neighbors and naïve bayes algorithms. Chapter 4 consists of the detailed explanation of the proposed algorithm. In the chapter 5, our proposed work has been evaluated on 5 medical related datasets. Finally, the conclusion and the future scope for this work is provide in Chapter 6

CHAPTER 2- LITERATURE REVIEW

2.1. HISTORY OF DATA MINING

Every day human beings are generating vast data and this data comes from different sources, be it online or offline. It may be in the form of documents, may be in graphical formats, may be the video or may be the records (varying array). Since the data is available in different formats, appropriate action needs to be taken not only to analyze the data but also to fetch important information and patterns from it and maintain the data .The data should be made available as and when required by the clients. The data should be retrieved from the database to help them make better decision .This technique is actually what we call data mining or simply KDD (Knowledge Discovery Process).

The most important reason that attracted a great deal of attention towards field of "Data mining" is due to the perception of *"we are data rich but information poor"*. Information technology required to discover useful information from large collections of data industry. There is huge volume of data but we were hardly able to convert them into useful information and knowledge for making managerial decision in business.

Data Mining is among one of the areas which are gaining a lot of practical significance. It is progressing at an active pace with the advancements in new methods, methodologies and findings in different applications related to medicine, computer science, bioinformatics and stock market prediction, weather forecast, text, audio and video processing to mention a few of them. Data happens to be the key concern in data mining. With the huge online data generated from several sensors, Internet Relay Chats, Twitter, Face book, Online Bank or ATM Transactions, there have always been existed a need to manage it.

To take full advantage of data, the data retrieval is simply not enough, various tool are required for automatic summarization of data, extraction of the essential information from the data stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tools for the analysis and interpretation of such data and for the extraction of interesting knowledge that can help in decision-making. Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions.[7]

2.2 DATA MINING USING MACHINE LEARNING

Machine learning is a well-established research area of computer science. Early machine learning algorithms were perceptrons (later called neural networks Rumelhart & McClelland, 1986), decision tree learners like ID3 (Quinlan, 1979,1986) and CART (Breiman et al., 1984), and rule learners like AQ (Michalski,1969; Michalski et al., 1986) and INDUCE (Michalski, 1980). These early algorithms were typically used to induce classifiers from a relatively small set of training examples (up to a thousand) described by a small set of attributes (up to a 100). An overview of early work in machine learning can be found in (Michalski, Carbonell, & Mitchell, 1983, 1986).

Data mining and knowledge discovery in databases appeared as a recognizable research discipline in the early 1990s (Piatetsky-Shapiro & Frawley, 1991), with the advent of a series of data mining workshops. The birth of this area was triggered by a need in the database industry to deliver solutions enhancing the traditional database management systems and technologies. At that time, these systems were able to solve the basic data management issues like how to deal with the data in transactional processing systems. In online transactional processing (OLTP) most of the processing scenarios were predefined. The main emphasis was on the stability and safety of solutions. As the business emphasis changed from automation to decision support, limitations of OLTP systems in business support led to the development of the next-generation data management technology known as data warehousing. The motivation for data warehousing was to provide tools for supporting analytical operations for decision support that were not easily provided by the existing database query languages. Online analytical processing (OLAP) was introduced to enable inexpensive data access and insights which did not need to be defined in advance. However, the typical operations on data warehouses were similar to the ones from the traditional OLTP databases in that the user issued a query and received a data table as a result. The major difference between OLTP and OLAP is the average number of records accessed per typical operation. While a typical operation in OLTP affects only up to tens or hundreds of records in predefined scenarios, a typical operation in OLAP affects up to millions of records (sometimes all records) in the database in a non-predefined way. The role of data mining in the above framework can be explained as follows. While typical questions in OLTP and OLAP are of the form: '*What is the answer to the given query?*', data mining—in a somewhat simplified and provocative formulation—addresses the question '*What is the right question to ask about this data?*'. The following explanation can be given. Data warehousing/OLAP provides analytical tools enabling only user-guided analysis of the data, where the user needs to have enough advance knowledge about the data to be able to raise the right questions in order to get the appropriate answers. The problem arises in situations when the data is too complex to be appropriately understood and analyzed by a human. In such cases data mining can be used, providing completely different types of operations for handling the data, aimed at hypothesis construction, and providing answers to questions which—in most cases—cannot be formulated precisely.

2.3 MACHINE LEARNING CLASSIFIERS FOR MEDICAL APPLICATIONS

Machine-learning techniques have grown to be among the leading research topics within the health care systems and particularly for clinical decision support systems (CDSS), which are commonly used in helping physicians to make more accurate diagnosis**.** In recent years, there has been a dramatic increase in the use of machine learning techniques within the healthcare systems to analyze, predict and classify clinical data.

Applications of machine learning can improve the use of knowledge to support decision making, and therefore improving the quality of healthcare service being delivered to the patient [1, 2]. The concept of machine learning refers to a computer program that is able to learn and gain knowledge from past experiences and/or through identifying the important features of a given dataset in order to make predictions about other data that were not a part of the original training set [2, 3]. Data mining is applied in medical field since long back to predict disease like diseases of the heart, lungs and various tumors based on the past data collected from the patient.

This section reviews a number of studies that targeted to evaluate and compare the overall performance of supervised ML classifiers:[7]

The authors in [1], have compared two most popular machine learning classifiers, k-nearest neighbors and naïve bayes classifier with the aim of finding their predictive capabilities using Pima Indians diabetes datasets. According to the Indonesian Society of Endocrinology (PERKENI), people with diabetes is expected to increase because 50% of the patients with diabetes remain undiagnosed. Infact, only two thirds of patients exists who underwent treatment which is both pharmacological and non-pharmacological. However, most of them are still unable to diagnose the disease diabetes in a patient due to lack of experience. Therefore, the author felt the need to provide solutions to these problems by creating an application that can diagnose diabetes in order to help young physicians in establishing the diagnosis. The data mining tasks that need to be done are namely classification of data by tracking the historical medical records of diabetic patients to identify and classify the data prior diagnosis based on the properties that were identified previously. Data classification algorithms are applied in this study are two algorithms: Naive Bayes and K-Nearest Neighbor. From the results it can be concluded that the diagnosis algorithm Naive Bayes and K-Nearest Neighbor has the same level of accuracy.

The authors in paper [2] have compared naïve bayes classifier and KNN, two the most effective techniques for data classification (especially for medical diagnoses), implemented them using C language and using Weka tool respectively and classify the patient affected by tuberculosis into two categories (least probable and most probable). This algorithm extracts hidden patterns from available Tuberculosis database. Using data collected from various TB centers, they made an effort to fetch out hidden patterns and by learning this pattern through the collected data, tuberculosis can be diagnosed and the disease can be predicted. The authors have concluded that the efficiency of results using KNN can be further improved by increasing the number of data sets and for Naïve Bayesian classifier by increasing attributes or by selecting weighted features.

Another paper [4] have attempted to classify Diabetes Mellitus Using Machine Learning Techniques. Diabetes-Mellitus refers to the metabolic disorder that happens from misfunctioning in insulin secretion and action. They attempt to make an ensemble model by combining two techniques: Bayesian classification and Multilayer Perceptron for the accuracy, sensitivity and specificity measures of diagnosis of diabetes-mellitus. MLP is a development from the simple perceptron in which extra hidden layers (additional to the input and output layers, not connected externally) are added. In this process, More than one hidden layer can be used. The network topology is constrained to be feed forward, i.e., loop-free. Generally, connections are allowed from the input layer to the first (and only possible) hidden layer, from the first hidden layer to the second and so on, until the last hidden layer to the output layer. The presence of these layers allows an artificial neural network to approximate a variety of non-linear functions. The actual construction of network, as well as the determination of the number of hidden layers and determination of the overall number of units, is sometimes a trial-and-error process, determined by the nature of the problem at hand. The transfer function is generally a sigmoidal function. Multilayer Perceptron is a neural network that trains using back propagation learning. Bayesian Net is a statistical classifiers which can predict class membership probabilities, such as the probability that a given tuple belong to a particular class or not. It is concluded that the new hybrid model succeeded in the search of finding the better result in terms of Accuracy.

Automated diagnosis of skin cancer is also accomplished by the authors of paper [5]. However, classification models based on insufficiently labeled training data can badly influence the diagnosis process if there is no self-advising and semi supervising capability in the model. This paper presents a semi supervised, self-advised learning model for automated recognition of melanoma using dermoscopic images. The proposed model is tested on a collection of 100 dermoscopic images. The variation in classification error is analyzed with respect to the ratio of labeled and unlabeled data used in the training phase. The classification performance is compared with some popular classification methods and the proposed model using the deep neural processing outperforms most of the popular techniques including KNN, ANN, SVM and semi supervised algorithms like Expectation maximization and transductive SVM. It was observed that by increasing the number of labeled data in the training phase helps in reducing the classification error. This is inconsistent with a lot of finding in different other applications where combinations of labeled and unlabeled data sets are used.

The paper[6] proposes an intellectual classification system to recognize normal and abnormal MRI brain images. Nowadays, decision and treatment of brain tumors is based on symptoms and radiological appearance. Magnetic resonance imaging (MRI) is a most important controlled tool for the anatomical judgment of tumors in brain. In the present investigation, various techniques were used for the classification of brain cancer. Under these techniques, image preprocessing, image feature extraction and subsequent classification of brain cancer is successfully performed. When different machine learning techniques: Support Vector Machine (SVM), K- Nearest Neighbor (KNN) and Hybrid Classifier (SVM-KNN) is used to classify 50 images, it is observed from the results that the Hybrid classifier SVM-KNN demonstrated the highest classification accuracy rate of 98% among others. The main goal of this paper is to give an excellent outcome of MRI brain cancer classification rate using SVM-KNN.

In paper [3], gene samples have been classified according to their gene expression profiles such as to classify cancer versus normal samples or to classify different types or subtypes of cancer. One of the major tasks with gene expression data is to find co-regulated gene groups whose collective expression is strongly associated with sample categories. In this regard, a gene clustering algorithm is proposed to group genes from microarray data. It directly incorporates the information of sample categories in the grouping process for finding groups of co-regulated genes with strong association to the sample categories, yielding a supervised gene clustering algorithm. The average expression of the genes from each cluster acts as its representative. Some significant representatives are taken to form the reduced feature set to build the classifiers for cancer classification. The mutual information is used to compute both gene-gene redundancy and gene-class relevance. The performance of the proposed method, along with a comparison with existing methods, is studied on six cancer microarray data sets using the predictive accuracy of naive Bayes classifier, K-nearest neighbor rule, and support vector machine. An important finding is that the proposed algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability

CHAPTER 3- RESEARCH METHODOLOGY

3.1 K-NEAREST NEIGHBORS

K Nearest Neighbor (KNN) is one of the most popular machine learning algorithms that are very simple to understand and but performs incredibly well. KNN is a non-parametric method used for classification and regression. It means that it does not make any assumptions on the underlying data distribution. This is pretty useful because in the real world, most of the data does not depend on the typical theoretical assumptions made.

This algorithm stores the training data and classifies new data in one of the classes based on a similarity measure (e.g. distance functions). Classification is done from a majority vote of the k nearest neighbors of each data point. The query point is assigned that class label which has the most representatives within the nearest neighbors of the point.

The training dataset consists of vectors in a multidimensional feature space. Each member of training dataset has a class label. The training phase of the algorithm consists only of storing these feature vectors and class labels of the training samples.

In the classification phase, a user-defined constant, k is chosen, and the class label which is the most frequent among the k nearest training samples is assigned to the unlabeled vector, also known as query point or test point.

In other words, the algorithm for k nearest neighbors works as follows:

1. Calculate the distance of the test data from all training data using appropriate distance function.
2. After obtaining the distances, they are sorted from the smallest to largest.
3. After it, consider as much the number of data the value of k is to see the results
4. The class label which is the most frequent among the k nearest training samples is assigned to the test data element
5. Repeat steps 1 to 4 for each member of testing dataset.
6. Compare the actual and predicted result and find the accuracy of the algorithm.

Following are a few functions for calculating distance.

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Distance Functions

All three of these distance measures are only valid for continuous variables. Euclidean distance is the commonly used distance metric for continuous variables. In case of categorical variables the Hamming distance must be used.

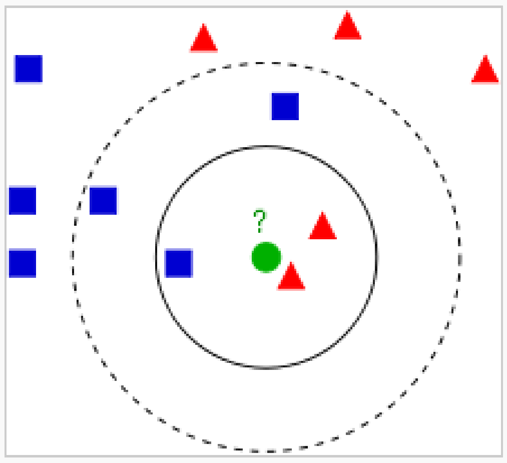


Figure 2 k-nn Classification Example

This Classification problem has 11 instances in its training dataset. The test sample i.e green circle has to be classified into one of the two classes : Blue or Red. If *k* = *3* (solid line circle) it is assigned to the Red class because there are 2 reds and only 1 blue inside the inner circle. If *k* =

5 (dashed line circle) it is assigned to the blue class since there are 3 blues and 2 reds inside the circle

3.2 NAÏVE BAYES

Naïve Bayes is a well known algorithm in the field of machine learning. It is particularly used for classification purpose which is the reason behind why it is called 'naïve bayes classifier'. As the name suggests, this algorithm is based on bayes' theorem , which describes the probability of an event depending on the conditions that might be related to that event.

Bayes' theorem can be represented in mathematical form as the following equation:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

where A and B are events and $P(B) \neq 0$.

- P(A) and P(B) are the probabilities of observing A and B independent of each other.
- P(A | B), a conditional probability, is the probability of observing event A given that B is true.
- P(B | A) is the probability of observing event *B* given that *A* is true.

It assumes that the features of a particular class are independent of each other. In other words, the presence of a particular feature in a class is not related to the presence of any other feature in that particular class. For example, a fruit, apple is considered to have certain features: it is red, round and about 3 inches in diameter. Even if these features are related to each other or depend upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is the reason it is known as 'Naive'.

Naïve Bayes Classifier is very easy to build and is particularly suited for the situations where input vector is of high dimensionality. Despite of its simplicity, this algorithm outperforms many sophisticated classification algorithms.

Probabilistic model

Naïve Bayes follows a conditional probability model. Given a problem instance, represented by vector x = (x_1, x_2, ….. ,x_n) having n features independent of each other , needs to be classified into k possible classes.

Using Bayes' theorem, the conditional probability of each class can be calculated according to the following formula:

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

In other words, it can be explained as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

When dealing with continuous data, it is typically assumed that the continuous values associated with each class are distributed according to a Gaussian distribution. Let there be an attribute, x in the training datset which is continuous in nature. According to the Gaussian distribution, dataset is first divided according to the classes and then mean and variance of x is computed. Let for all the values of x in some class c, mean is represented by μ_c and variance is σ_c^2. So the value of p(x_i | C_k) for a particular x = v that belongs to set x, can be calculated as:

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}}\, e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

Demonstration of Naïve Bayes with an example:

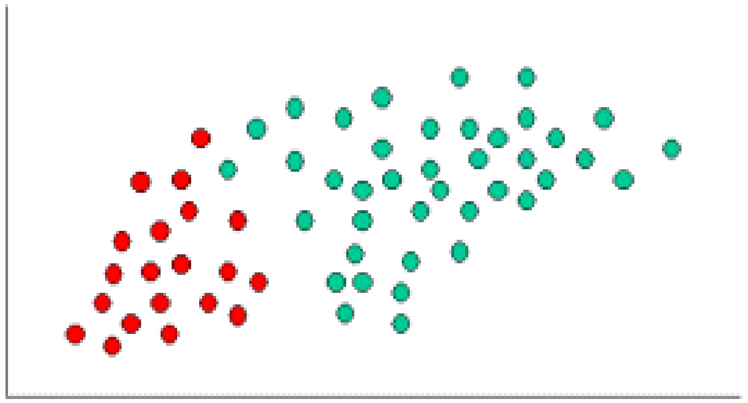


Figure 3

In this example, there are 60 object in total. They are classified into two classes : RED and GREEN. Our job is to categorize new objects, that is , to assign the class label to each arriving, based on the currently exiting objects.

Prior probability, which depends on the previous experience is calculated at first.

$$Prior\ Probability\ of\ GREEN\ \propto \frac{Number\ of\ GREEN\ objects}{Total\ number\ of\ objects}$$

$$Prior\ Probability\ of\ RED\ \propto \frac{Number\ of\ RED\ objects}{Total\ number\ of\ objects}$$

Out of 60, 40 are GREEN objects and 20 are RED objects. Therefore,

$$Prior\ Probabilty\ of\ GREEN = \frac{40}{60}$$

$$Prior\ Probability\ of\ RED\ = \frac{20}{60}$$

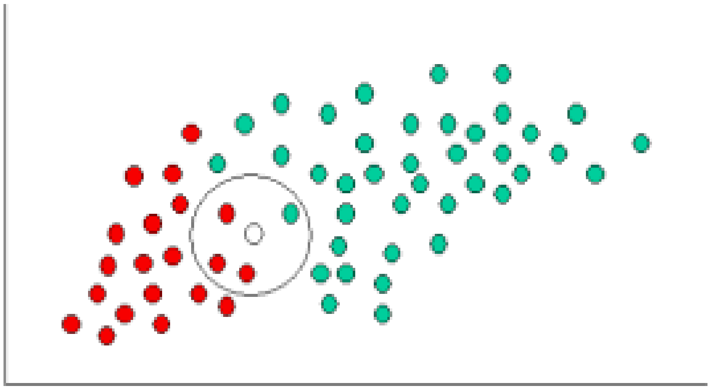


Figure 4

Now our task is to classify a new object represented as WHITE circle. It is quite clear from the above representation that the objects are well clustered, it would not be incorrect to assume that the more GREEN (or RED) objects in the vicinity of X, the more likely that the new object belong to that particular class. In order to measure this likelihood, a circle is drawn around X which covers a number of points irrespective of their class labels. The number of objects in the circle is to be calculated next and also the number of objects belonging to each class.

This likelihood can be calculated as follows:

$$Likelihood\ of\ X\ given\ GREEN\ \propto \frac{Number\ of\ GREEN\ objects\ in\ the\ vicinity\ of\ X}{Total\ number\ of\ GREEN\ objects}$$

$$Likelihood\ of\ X\ given\ RED\ \propto \frac{Number\ of\ RED\ objects\ in\ the\ vicinity\ of\ X}{Total\ number\ of\ RED\ objects}$$

$$Likelihood\ of\ X\ given\ GREEN\ \propto \frac{1}{40}$$

$$Likelihood\ of\ X\ given\ RED\ \propto \frac{3}{20}$$

According to the Bayes’ theorem, the final classification is obtained by combining both sources of information, that is, the prior probability and the likelihood. Together they are known to form a posterior probability. The posterior probability can be calculated as follows:

Posterior Probability of X being GREEN

= Prior Probability of GREEN X *Likelihood of GREEN given X*

=40/60 X *1/40*

=1/60

Posterior Probability of X being RED

= Prior Probability of RED X *Likelihood of RED given X*

=20/60 X *2/20*

=1/20

Hence, X is classified as RED because its class membership has the largest posterior probability.

The algorithm works according to the following steps:

1. The data is converted into a frequency table.
2. The Prior probability of each class is calculated.
3. The Likelihood table is created by finding the probabilities.

Now, Naive Bayesian equation is used to calculate the posterior probability of each class. The class with the highest posterior probability is the predicted outcome.

CHAPTER 3- PROPOSED WORK

3.1 PROBLEM STATEMENT

Classification is one of the most important data mining techniques used in statistics which involves the use of multiple variables. It is sometimes called the prediction problem, particularly in data mining. In statistics, classification is a process in which individual objects are categorized into groups based on the quantitative information about one or more characteristics of the objects (referred to as traits, attributes, variables, characters, etc) and based on a set of previously labeled objects in the training data. The problem can be stated as follows:

Given training data $\{(x_1 ,y_1), \ldots ,(x_n ,y_n)\}$ where $x \in X\,; the\ set\ of\ objects$ and $y \in Y; the\ set\ of\ class\ labels$ produce a classifier h :X→Y , which maps an object x∈X to its class label y∈Y .

3.2 PROPOSED SOLUTION

There are many classical approaches to deal with the classification problem. In this work, two most popular machine learning techniques ; k-nearest neighbors and naïve bayes have been incorporated to form a hybrid algorithm. This algorithm takes the best features of the base algorithms and combines them to produce a better solution.

The k-nearest neighbours algorithm selects a subset of training data based on the distance measured from them. There are many distance measurement schemes available for this purpose. Naïve bayes works by determine the conditional probability for each possible value of all attributes.

To classify a new object we first use the KNN algorithm to find the k Nearest Neighbor from the training dataset. After selecting the k nearest object, we shall then build a model using the Naïve Bayes algorithm

3.3 METHODOLOGY

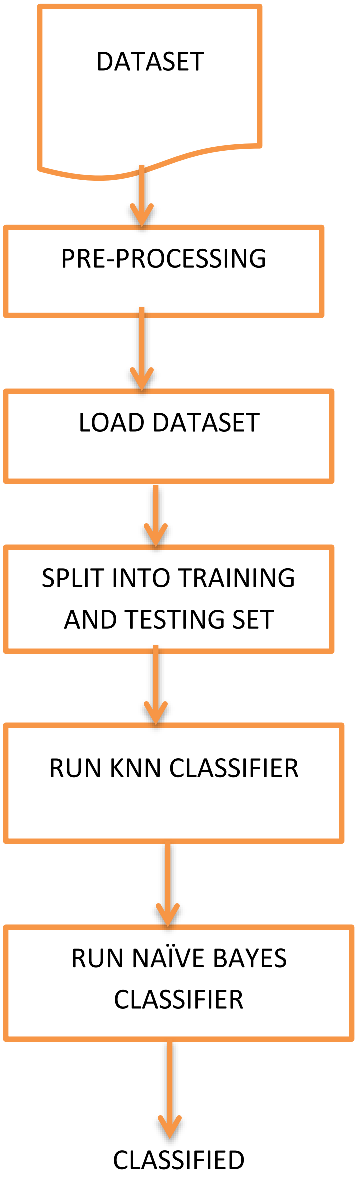

DATASET
PRE-PROCESSING
LOAD DATASET
SPLIT INTO TRAINING AND TESTING SET
RUN KNN CLASSIFIER
RUN NAÏVE BAYES CLASSIFIER
CLASSIFIED

Figure 5

3.4 PROPOSED ALGORITHM

1. Load Dataset.
2. Split the dataset into training set and testing set.
3. Apply KNN (k-nearest neighbors) to find out the k neighbors of each element of testing dataset.
4. Use naïve bayes to calculate the posterior probability of each neighborhood element
5. The class of the element having the highest probability will be the class of our testing data element.

Check Accuracy by comparing the actual class and the predicted class of each testing dataset element

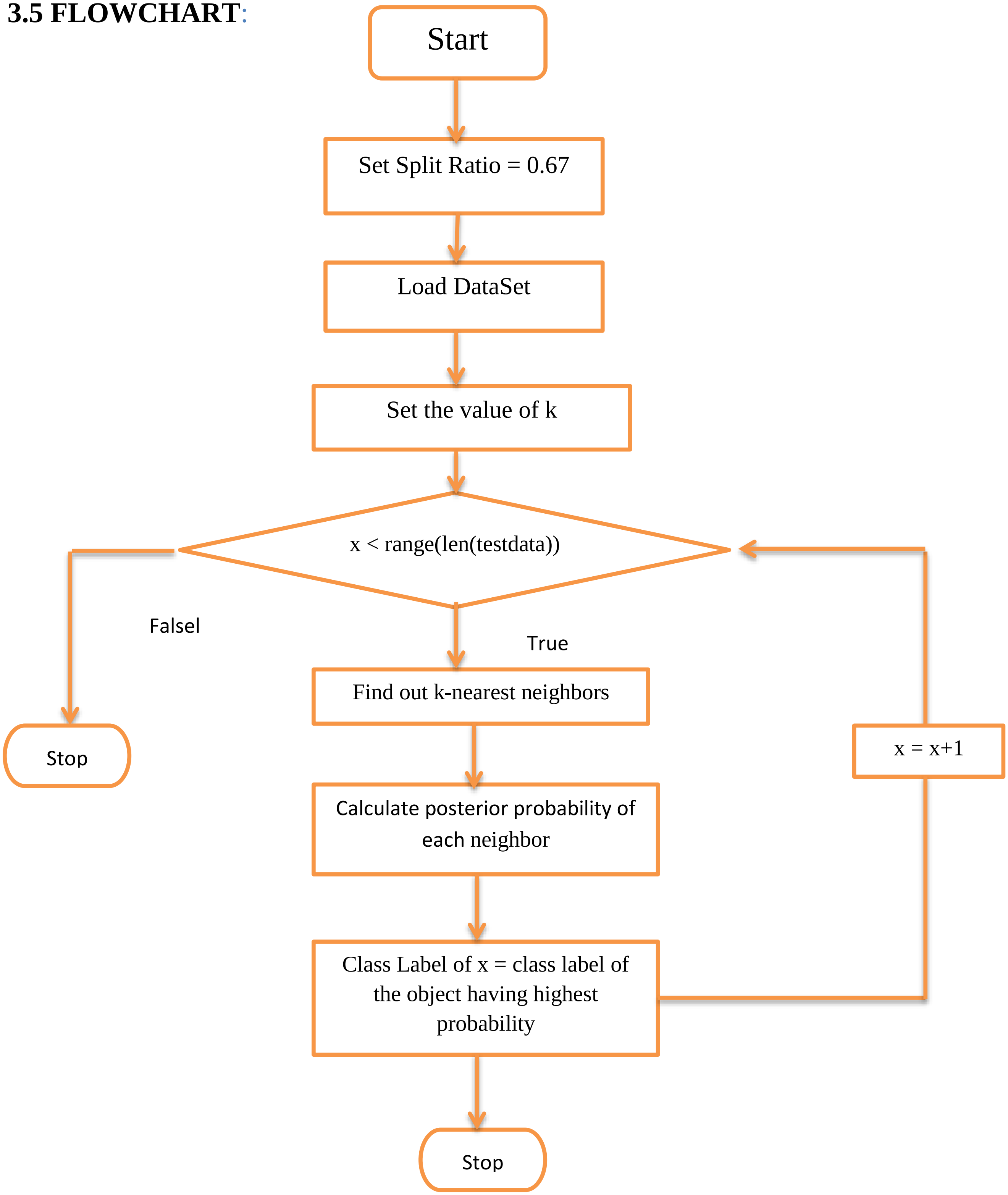
Firstly, we need to load dataset which is in CSV format. The file has no header lines or quotes . If present, they need to be removed. Next the dataset goes through pre-processing step in which any missing values are recovered by taking the average of the characteristic concerned. Also in this step, the string values are converted to float so that mathematical calculations can be run on them.

Next, we need to split the dataset randomly into training data and testing data using the provided split ratio. Usually 0.67 is considered ideal. So we have also set the value of split ratio as 0.67. After this, the process of finding k-nearest neighbors starts by setting the value of k. We have chosen to take different values of k for different datasets and by running the algorithm for different values for finding the one value that gives optimized results. The nearest neighbors can be simply collected by first calculating the Euclidean distance from all the member objects of training data and then sorting them into ascending order and keeping the first k neighbors since they are the nearest.

These neighbors are fed to the next phase of the hybrid classifier, i.e naïve bayes. The algorithm from now on works with only these neighbors. It will then collect summary of each neighbor. The summary of the neighbors collected involves the mean and the standard deviation for each attribute, by class value. For example, let there be two class labels/value and 10 numerical attributes, then we need to calculate mean and standard deviation for each attribute (10) and class value (2) combination, that is 20 attribute summaries. Once the summaries are collected, we are now ready to make predictions using the summaries prepared from the k nearest neighbors collected from the previous phase. Making predictions involves calculating the probability that a given data object belongs to each class, then selecting the class with the largest probability as the prediction. Gaussian function have been used estimate the probability of a given attribute value. While dealing with continuous data, it is a typical assumption that the continuous values associated with each class are distributed according to a Gaussian distribution . Next we need to combine the probabilities of all of the attribute values for a data object and come up with a probability of the entire data instance belonging to the class by multiplying them together. Since we have calculated the probability of a data object belonging to each class value, we can look for the largest probability and return the associated class.

The final step is calculating the accuracy of the algorithm by comparing the actual class label with the predicted one.

3.5 FLOWCHART:


Start
Set Split Ratio = 0.67
Load DataSet
Set the value of k
x < range(len(testdata))
Falsel
True
Stop
Find out k-nearest neighbors
x = x+1
Calculate posterior probability of each neighbor
Class Label of x = class label of the object having highest probability
Stop

CHAPTER 4-SIMULATION AND RESULTS

Machine-learning (ML) techniques have always been among the leading research topics for the health care systems and particularly for clinical decision support systems (CDSS), which are commonly used in helping physicians to make more accurate diagnosis. Therefore, we have chosen to simulate our k-nn naïve hybrid classifier for medical applications. Evaluating machine learning classifiers with only one sample of data appears to be unsatisfying, since it does not reflect the classifier's capabilities or their behavioral patterns under different circumstances.

4.1 SIMULATION SETUP

To simulate the three algorithms – k-nearest neighbors, naïve bayes and the proposed hybrid classifier, the following configurations of systems have been used:

1. AMD A8
2. 64-bit configuration
3. Windows 8
4. 4 GB RAM

The parameter used for simulation is:

Split ratio = 0.67 for dividing the dataset into training set and testing set since this is the standard ratio.

4.2 DATASET INFORMATION

Source of dataset – UCI machine learning Repository,

http://archive.ics.uci.edu/ml/

FERTILITY DIAGNOSIS DATASET

This dataset is collected from 100 volunteers who provide a semen sample analyzed according to the WHO 2010 criteria. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits.

The dataset comprises of following 10 attributes:

1.Season in which the analysis was performed. 1) winter, 2) spring, 3) Summer, 4) fall. (-1, -0.33, 0.33, 1)
2.Age at the time of analysis. 18-36 (0, 1)
3.Childish diseases (i.e, chicken pox, measles, mumps, polio) 1) yes, 2) no. (0, 1)
4.Accident or serious trauma 1) yes, 2) no. (0, 1)
5.Surgical intervention 1) yes, 2) no. (0, 1)
6.High fevers in the last year 1) less than three months ago, 2) more than three months ago, 3) no. (-1, 0, 1)
7.Frequency of alcohol consumption 1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never (0, 1)
8.Smoking habit 1) never, 2) occasional 3) daily. (-1, 0, 1)
9.Number of hours spent sitting per day ene-16 (0, 1)
10.Output: Diagnosis normal (N), altered (O)

HEART DISEASE DATASET

This dataset was the Cleveland Clinic Foundation heart disease dataset, which is available. It contains 76 attributes, but all published experiments referred have used only 14 of them. The "goal" attribute refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0). It comprises of 270 instances and without missing values. Each instance contains 13 attributes in addition to the output class.

1. age
2. sex
3. Chest_pain_type
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic

4. resting_blood_pressure (in mm Hg on admission to the hospital)
5. Serum_cholestoral (mg/dl)
6. Fasting_blood_sugar > 120 mg/dl (1 = true; 0 = false)
7. Resting_electrocardiographic_results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

8. thalach: maximum heart rate achieved
9. Exercise_induced_angina (1 = yes; 0 = no)
10.oldpeak = ST depression induced by exercise relative to rest
11 slope: the slope of the peak exercise ST segment
-- Value 1: upsloping
-- Value 2: flat
-- Value 3: downsloping
12. ca: number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. the predicted attribute, num: diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing
-- Value 1: > 50% diameter narrowing
(in any major vessel: attributes 59 through 68 are vessels)

BREAST CANCER DATASET

This dataset is donated by University of Wisconsin Hospitals Madison, Wisconsin, USA. It comprises of 699 instances and 11 attributes.

1. Sample_code_number: id number
2. Clump_Thickness: 1 - 10
3. Uniformity_of_Cell_Size: 1 - 10
4. Uniformity_of_Cell_Shape: 1 - 10
5. Marginal_Adhesion: 1 - 10
6. Single_Epithelial_Cell_Size: 1 - 10
7. Bare_Nuclei: 1 - 10
8. Bland_Chromatin: 1 - 10
9. Normal_Nucleoli: 1 - 10
10. Mitoses: 1 - 10
11. Class: (2 for benign, 4 for malignant)

PIMA INDIANS DIABETES DATASET

This dataset is in accordance with the criteria of the World Health Organization (WHO) about the number of women whose age is more than 21 years, from Pima India heritage who lives near Phoenix, Arizona. There are 768 objects in this dataset and 9 attributes in total. The original owner is National Institute of Diabetes and Digestive and Kidney Diseases.

The dataset comprises of following attributes:

1. Number_of_times_pregnant
2. Plasma_glucose_concentration
3. Diastolic_blood_pressure (mm Hg)
4. Triceps_skin_fold_thickness (mm)
5. 2-Hour_serum_insulin (mu U/ml)
6. Body_mass_index (weight in kg/(height in m)^2)
7. Diabetes_pedigree_function
8. Age (years)
9. Class_label (0 or 1)

BLOOD TRANSFUSION DATASET

This is the data taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiwan. The center passes their blood transfusion service bus to one university in Hsin-Chu City to gather blood donated about every three months. They selected 748 donors at random from the donor database.

The database comprises of following attributes:

1.Recency - months since last donation,
2. Frequency - total number of donation,
3.Monetary - total blood donated in c.c.,
4.Time - months since first donation, and
5.a binary variable representing whether he/she donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood)

S.No	**Name of Dataset**	**Number of instances**	**Number of attributes**
1	Fertility Diagnosis	100	10
2	Heart Disease Dataset	270	13
3	Breast Cancer Dataset	699	11
4.	Pima Indians Diabetes Dataset	768	9
5	Blood Transfusion Dataset	748	5

Table 1 Summary of Datasets used

4.3 RESULTS AND GRAPHS

Dataset is split randomly into training set and testing set. The split ratio is set as 0.67. Therefore, we have carried out the execution of each algorithm 10 times and have taken the average of these 10 iterations in order to make the comparisons on the basis of accuracy (in percentage). Accuracy is calculated by comparing the actual result with the predicted result.

FERTILITY DIAGNOSIS DATASET

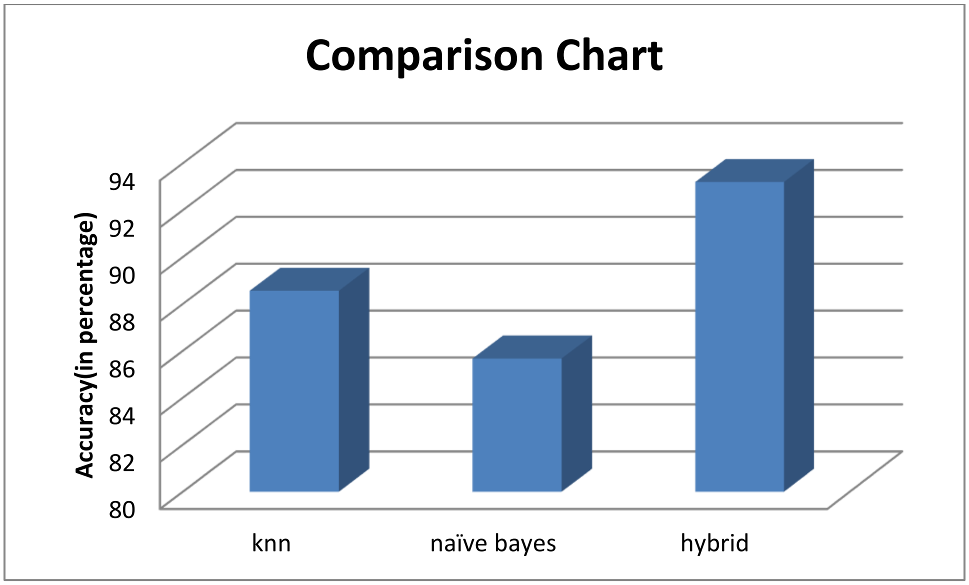


Figure 7 Results for k=50

The results shows that the hybrid classifier outperforms the other two algorithm with accuracy level equal to 93.2% while the accuracy of k-nn is 88.56 and that of naïve bayes is 85.67.

This following line-graph shows the variation of the average accuracy on varying the value of k. The value of k must be large enough that noise in the data is minimized and small enough so the samples of the other classes are not included. This has been achieved by hit and trial.We have chosen that value of k which gives the most efficient results for simulating the hybrid algorithm

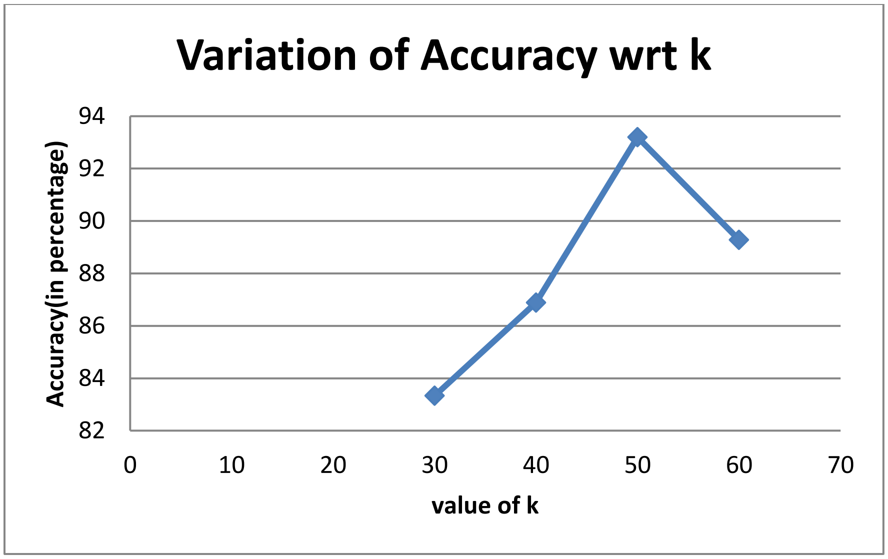


Figure 8

HEART DISEASE DATASET

.

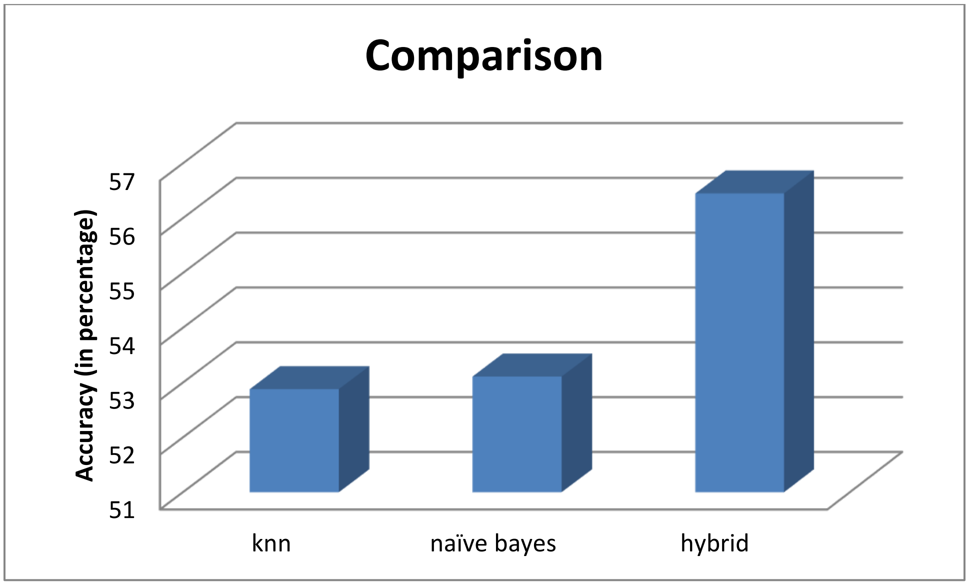


Figure 9 Results for k = 50

The results shows that the hybrid classifier outperforms the other two algorithm with accuracy level equal to 56.45% while the accuracy of k-nn is 52.88 and that of naïve bayes is 53.11.

This following line-graph shows the variation of the average accuracy on varying the value of k. We have chosen that value of k which gives the most efficient results for simulating the hybrid algorithm.

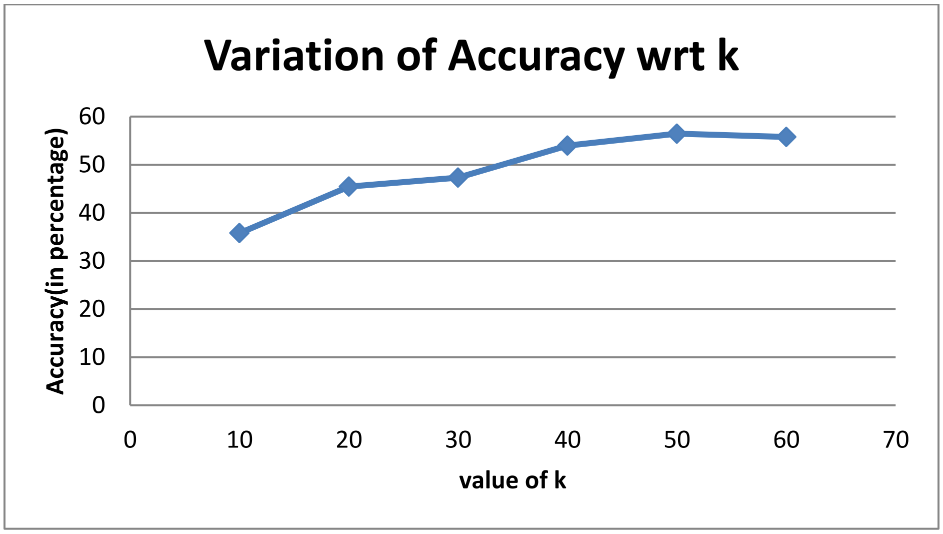


Figure 10

BREAST CANCER DATASET

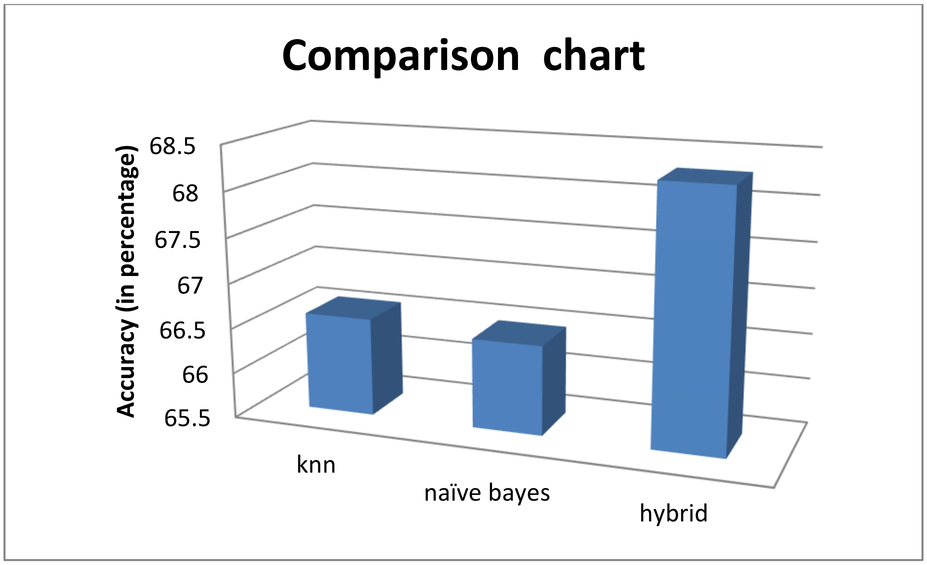


Figure 11 Results for k=300

The results shows that the hybrid classifier outperforms the other two algorithm with accuracy level as high as 68.30% while the accuracy of k-nn is 66.58 and that of naïve bayes is 66.48

The following line-graph shows the variation of the average accuracy for Breast cancer dataset on varying the value of k.

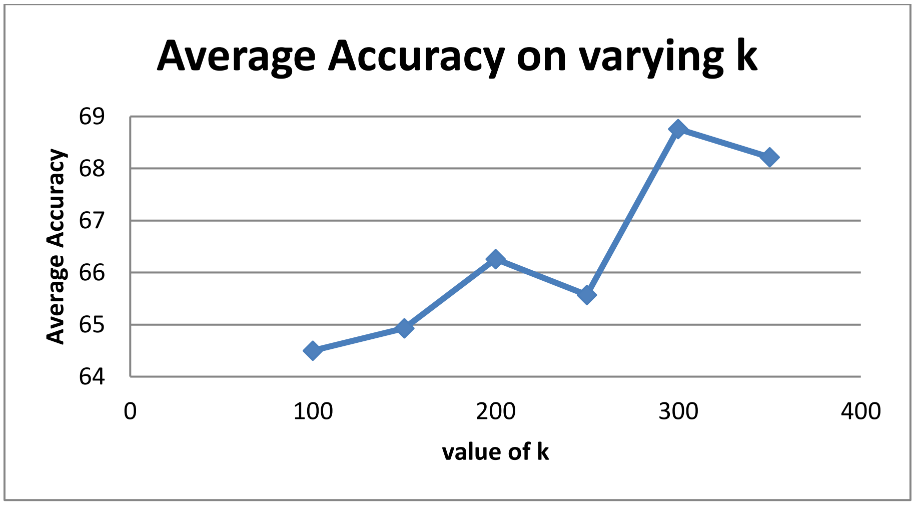


Figure 12

PIMA INDIANS DIABETES DATASET

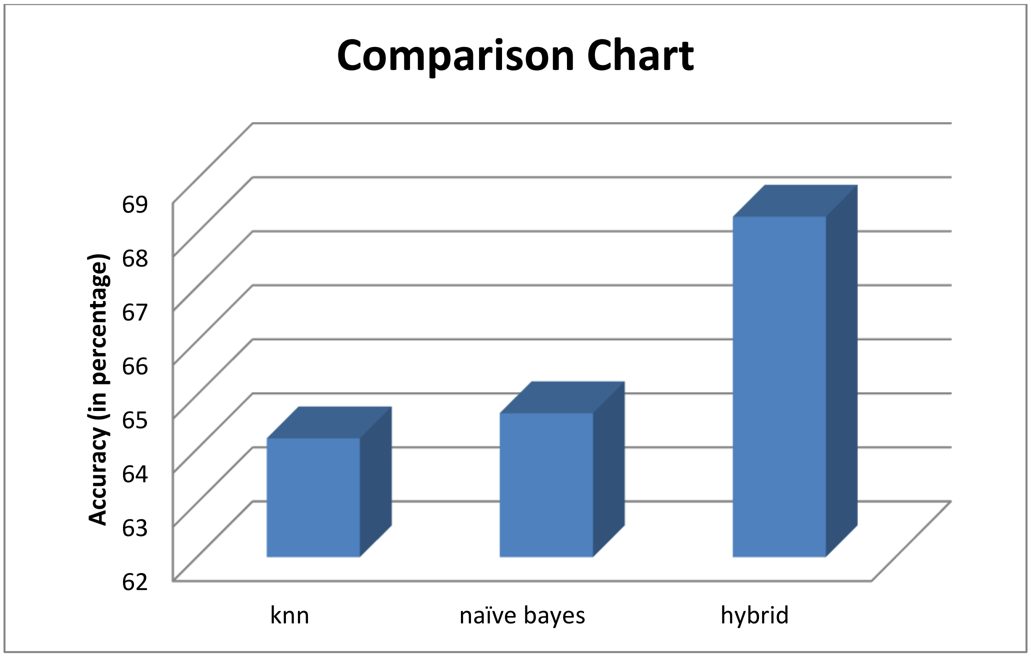


Figure 13 Results for k = 300

The results shows that hybrid algorithm gives better results as compared to the other two with accuracy level as high as 68.30% while the accuracy of k-nn is 64.2 and that of naïve bayes is 64.66.

This following line-graph shows the variation of the average accuracy on varying the value of k.

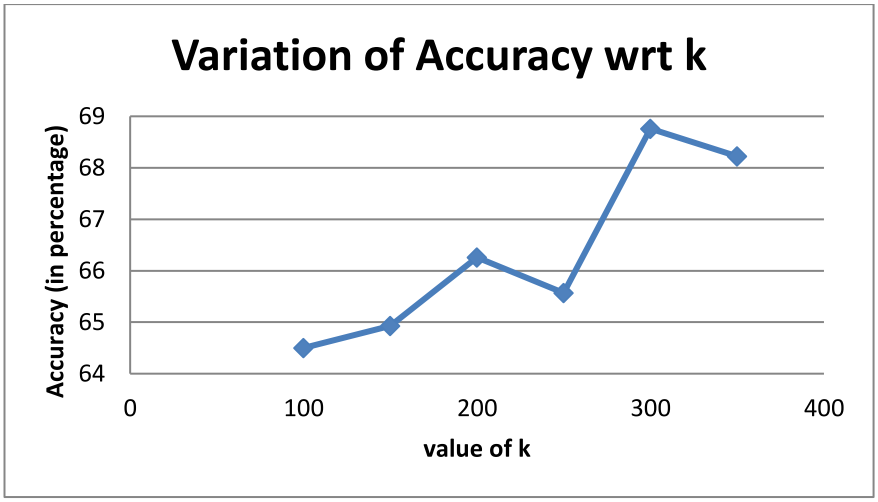


Figure 14

BLOOD TRANSFUSION DATASET

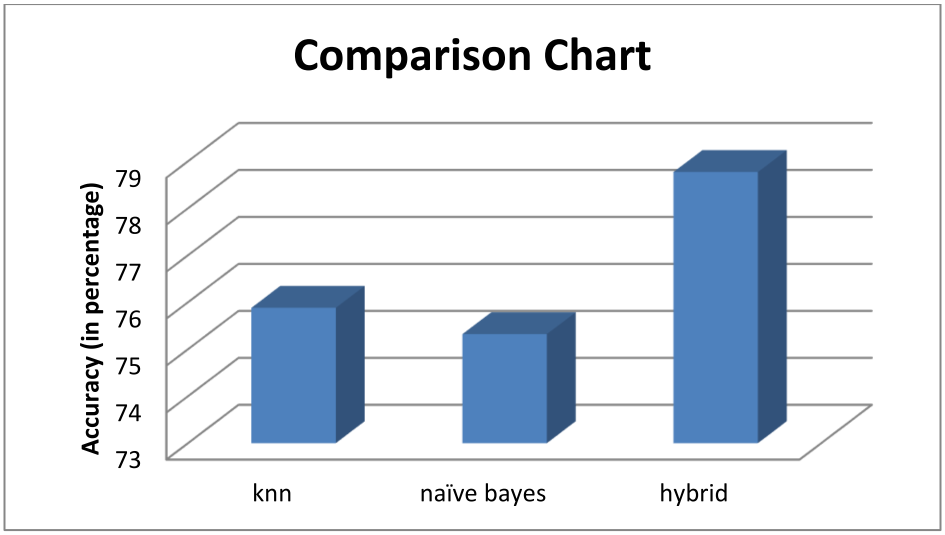


Figure 15 Results for k=350

The results shows that hybrid algorithm gives better results as compared to the other two with accuracy level as high as 78.77% while the accuracy of k-nn is 75.88 and that of naïve bayes is 75.32.

This following line-graph shows the variation of the average accuracy on varying the value of k

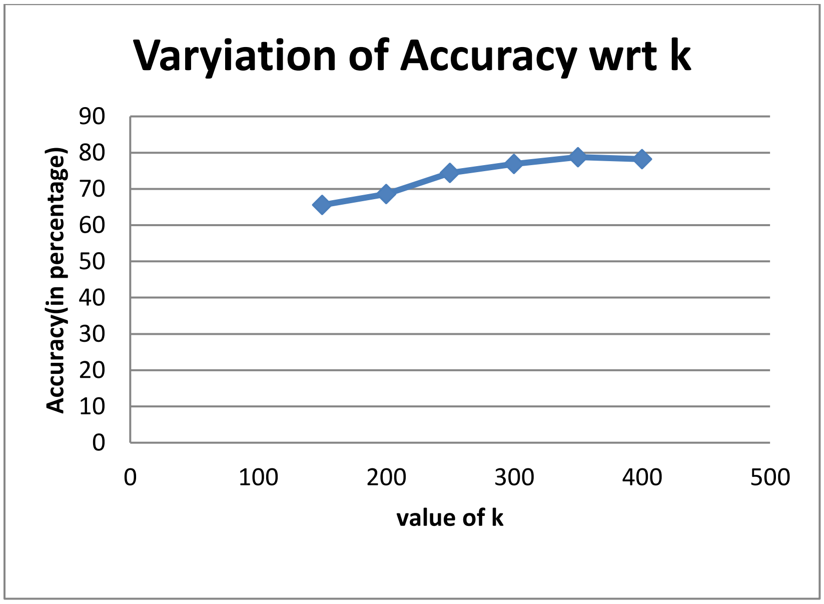


Figure 16

CHAPTER 6-CONCLUSION AND FUTURE SCOPE

The main objective of this work is to build a hybrid classifier which has the best features of two algorithms k-nearest neighbors and naïve bayes which are very popular machine learning techniques for data classification The performance of the proposed hybrid classifier is compared with them by executing all three algorithms against 5 datasets of varying sizes. The implementation has been done in python. It is concluded that the traditional k-nearest neighbors and naïve bayes have almost similar levels of accuracy and the hybrid classifier outperforms both algorithms. It has always been attempted by the researchers to contribute to the health care field since most of the times, the disease is not diagnosed which leads to the suffering of the patient. This hybrid algorithm have been applied to the 5 medical datasets and will help physicians make more accurate diagnosis which can cure diseases.

There is a shortcoming of the proposed hybrid algorithm that it does not perform well if we consider less number of neighbors for prediction. This short-coming has to be attempted to remove in future. The hybrid classifier can also be made user-interactive so that it can classify any dataset without making explicit changes in the algorithm.

REFERENCES

1. 2011Mohammed J. Islam, Q. M. Jonathan Wu, Majid Ahmadi, Maher A. Sid-Ahmed, *Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers*, International Conference on Convergence Information Technology, 2007

2. P. Viswanath and T. Hitendra Sarma, *An Improvement to k-Nearest Neighbor Classifier*, IEEE ,2011.

3. Pradipta Maji and Chandra Das, *Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification* , IEEE Transactions on nanobioscience, VOL. 11, NO. 2, JUNE 2012

4. Amit kumar Dewangan, Pragati Agrawal, *Classification of Diabetes Mellitus Using Machine Learning Techniques,* International Journal of Engineering and Applied Sciences (IJEAS***)*** Volume-2, Issue-5, May 2015

5. Ammara Masood, Adel Al- Jumaily, Khairul Anam, *Self-Supervised Learning Model for Skin Cancer Diagnosis,* 7th Annual International IEEE EMBS Conference on Neural Engineering, April, 2015

6. Ketan Machhale, Hari Babu Nandpuru, Vivek Kapur, Laxmi Kosta, *MRI Brain Cancer Classification Using Hybrid Classifier (SVM-KNN),* International Conference on Industrial Instrumentation and Control (ICIC), 2015

7. M.S.B. PhridviRaja, C.V. GuruRaob, *Data mining – past, present and future – a typical survey on data streams*, The 7th International Conference Interdisciplinarity in Engineering , 2013

8. Shu-Hsien Liao , Pei-Hui Chu, Pei-Yuan Hsiao, *Data mining techniques and applications – A decade review from 2000 to 2011*, ScienceDirect, 2013

9. Hsiao, H. C. W., Chen, S., Chang, J. P., & Tsai, J. J. P., *Predicting Subcellular Locations of Eukaryotic Proteins Using Bayesian and K-Nearest Neighbor Classifiers,* Journal of Information Science and Engineering, 2008

10. *Application of k-Nearest Neighbour Classification in Medical Data Mining,* ResearchGate , April 2014, https://www.researchgate.net/publication/270163293

11. Sagar S. Badhiye, Nilesh U. Sambhe, P. N. Chatur, *KNN Technique for Analysis and Prediction of Temperature and Humidity Data,* International Journal of Computer Applications, Volume 61– No.14, January 2013

12. [Online] https://en.wikipedia.org/wiki/Machine_learning.

13. **Dell.** naive-bayes-classifier. *www.statsoft.com.* [Online] http://www.statsoft.com/textbook/naive-bayes-classifier.

14. Artificial_neural_network. *en.wikipedia.org.* [Online]https://en.wikipedia.org/wiki/Artificial_neural_network

15. naive_bayes_classifier. *en.wikipedia.org.* [Online] https://en.wikipedia.org/wiki/Naive_Bayes_classifier.

16. K-nearest_neighbors_algorithm. *en.wikipedia.org.* [Online] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.

17. Nurhayati, Arif Nur Rahman , *Implementation of Naive Bayes and K-Nearest Neighbor Algorithm for Diagnosis of Diabetes Mellitus* , Applied Computational Science