***Abstract*-Chronic disease endured for a long period of time. They are only be controlled but cannot be cured completely. Most of the people in the world are affected by chronic disease. Of all the diseases, Diabetes is the most hazardous disease. Diabetes means that blood glucose (blood sugar is too high. It is categorized into two divisions: Diabetes of category 1 and diabetes of category 2. In category 1, the human body does not make insulin, people with type1 need to take insulin every day. In type 2 the glucose level is of very high in the blood, it is one of the most common forms of diabetes. In type 2 diabetes, need to do physical activity and should have proper diet. The analysis on the data is performed using different machine learning algorithm like Naive Bayes algorithm and Support vector machine Algorithm and These two algorithm’s efficiency are compared to decide which is more efficient.**

**INTRODUCTION**

Diabetes mellitus is a complex, chronic illness requiring continuous medical care with multi factorial risk reduction strategies beyond glycemic control [1]. Ongoing patient self-management education and support are critical to preventing acute complications and reducing the risk of long-term complications. Significant evidence exists that supports a range of interventions to improve diabetes outcomes. As the term indicates it consists of large volume of raw data’s. The data’s consists of proper, improper and semi structured data’s.

**Machine learning** is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (i.e., progressively improve performance on a specific task) with [data](https://en.wikipedia.org/wiki/Data)[4], without being explicitly programmed. Machine learning algorithm broadly classified into 3 types: Supervised, Unsupervised and Reinforcement learning. There are many machine learning algorithms like Naive Bayes, SVM, Neural Net, Regressive, K-means algorithm….etc

**I. CLASSIFICATION AND DIAGNOSIS**

Diabetes can be classified into four clinical categories:

1. Type 1 diabetes (due to b-cell destruction, usually leading to absolute insulin deficiency)

2. Type 2 diabetes (due to a progressive insulin secretary defect on the back ground of insulin resistance)

3. Other specific types of diabetes due to other causes, e.g., genetic defects in b-cell

function, genetic defects in insulin action, diseases of the exocrine pancreas (such as cystic fibrosis), and drug- or chemical-induced

4. Gestational diabetes mellitus (GDM) (diabetes diagnosed during pregnancy that is not clearly overt diabetes)

Some patients cannot be clearly classified as type 1 or type 2 diabetic .Clinical presentation and disease progression vary considerably in both types of diabetes.

**OBJECTIVES**

* Detecting diseases at earlier stages can help to be treated more easily and effectively.
* This system leads to the improved focus on every individual patient health.
* The middle-income families can be with the high availability of medical facility at minimum cost.
* We compare prediction efficiency of different machine learning algorithms based on the results obtained.

**BACKGROUND STUDY**

**I. MACHINE LEARNING**

Machine learning is a paradigm that may refer to learning from past experience (which in this case is previous data) to improve future performance[4]. The sole focus of this field is automatic learning methods. Learning refers to modification or improvement of algorithm based on past “experiences” automatically without any external assistance from human.

Instead of designing an algorithm to address the problem directly, using Machine Learning, a researcher seek an approach through which the machine, i.e., the algorithm will come up with its own solution based on the example or training data set provided to it initially.

**A. MACHINE LEARNING: INTERSECTION OF STATISTICS AND COMPUTER SCIENCE**

Machine Learning was the phenomenal outcome when Computer Science and Statistics joined forces. Computer Science focuses on building machines that solve particular problems, and tries to identify if problems are solvable at all. The main approach that Statistics fundamentally employs is data inference, modeling hypothesizes and measuring reliability of the conclusions. The defining idea of Machine Learning is a little different but partially dependent on both nonetheless. Whereas Computer Science concentrate on manually programming computers, ML addresses the problem of getting computers to re-program themselves whenever exposed to new data based on some initial learning strategies provided [4]. On the other hand, Statistics focuses on data inference and probability, Machine Learning includes additional concerns about the feasibility and effectiveness of architectures and algorithms to process those data, compounding several learning tasks into a compact one and performance measures.

**B. MACHINE LEARNING AND HUMAN LEARNING**

A third research area closely related to Machine Learning is the study of human and animal brain in Neuroscience, Psychology, and related fields. The researchers proposed that how a machine could learn from experience most probably would not be significantly different than how animal or human minds learn with time and experience [8]. However, the research concentrated on solving machine learning problems using learning methods of human brain did not yield much promising result so far than the researches concerned with statistical - computational approach. This might be due to the fact that human or animal psychology remains not fully understandable to date. Regardless of these difficulties, collaboration between human learning and machine learning is increasing for machine learning is being used to explain several learning techniques seeing in human or animals. For example, machine learning method of temporal difference was proposed to explain neural signals in animal learning. It is fairly expected that this collaboration is to grow considerably in coming years.

**II. CATEGORISATION OF ML ALGORITHMS**

An overwhelming number of ML algorithm have been designed and introduced over past years. Not everyone of them are widely known. Some of them did not satisfy or solve the problem, so another was introduced in its place. Here the algorithms are broadly grouped into two category and those two groups are further sub-divided.[8] This section try to name most popular ML algorithms and the next section compares three most widely used ML algorithms

.

**A. GROUP BY LEARNING STYLE**

1. **Supervised learning**

Input data or training data has a pre-determined label e.g. True/False, Positive/Negative, Spam /Not Spam etc. A function or a classifier is built and trained to predict the label of test data. The classifier is properly tuned (parameter values are adjusted)to achieve a suitable level of accuracy

.

1. **Unsupervised learning**

Input data or training data is not labeled. A classifier is designed by deducing existing patterns or cluster in the training datasets.

**3. Semi-supervised learning** --- Training dataset contains both labeled and unlabelled data. The classifiers train to learn the patterns to classify and label the data as well as to predict.

**4. Reinforcement learning** --- The algorithm is trained to map action to situation so that the reward or feedback signal is maximized. The classifier is not programmed directly to choose the action, but instead trained to find the most rewarding actions by trial and error.

**5. Transduction** --- Though it shares similar traits with supervise learning, but it does not develop a explicit classifier. It attempts to predict the output based on training data, training label, and test data.

**6. Learning to learn** --- The classifier is trained to learn from the bias it induced during previous stages.

**7.** It is necessary and efficient to organize the ML algorithms with respect to learning methods when one need to consider the significance of the training data and choose the classification rule that provide the greater level of accuracy.

**III. MEASURING AND COMPARING PERFORMANCES OF POPULAR ML ALGORITHMS**

Though various researchers have contributed to ML and numerous algorithms and techniques have been introduced as mentioned earlier, if it is closely studied most of the practical ML approach includes three main supervised algorithm or their variant. These three are namely, [4] Naive Bayes, Support Vector Machine and Decision Tree. Majority of researchers have utilized the concept of these three, be it directly or with a boosting algorithm to enhance the efficiency further. These three algorithms are discussed briefly in the following section.

1. **NAIVE BAYES CLASSIFIER**

It is a supervised classification method developed using Bayes’ Theorem of conditional probability with a ‘Naive’ assumption that every pair of feature is mutually independent [7]. That is, in simpler words, presence of a feature is not affected by presence of another by any means. Irrespective of this over-simplified assumption, NB classifiers performed quite well in many practical situations, like in text classification and spam detection. Only a small amount of training data is need to estimate certain parameters. Beside, NB classifiers have considerably outperformed even highly advanced classification techniques.

**B. SUPPORT VECTOR MACHINE SVM**

Supervised classification algorithm proposed by Vapnik in 1960s has recently attracted an major attention of researchers. The simple geometrical explanation of this approach involves determining an optimal separating plane or hyper plane that separates the two classes or clusters of data points justly and is equidistant from both of them. SVM was defined at first for linear distribution of data point[6]. Later, the kernel function was introduced to tackle nonlinear data as well.

**C. DECISION TREE**

A classification tree, popularly known as decision tree is one of the most successful supervised learning algorithm. It constructs a graph or tree that employs branching technique to demonstrate every probable result of a decision. In a decision tree representation, every internal node tests a feature, each branch corresponds to outcome of the parent node and every leaf finally assigns the class label. To classify an instance, a top-down approach is applied starting 99999at the root of the tree. For a certain feature or node, the branch concurring to the value of the data point for that attribute is considered till a leaf is reached or a label is decided. Now, the performances of these three were roughly compared using a set of tweets with labels positive, negative and neutral. The raw tweets were taken from Sentiment140 data set. Then those are pre-processed and labeled using a python program. Each of these classifier were exposed to same data. Same algorithm of feature selection, dimensionality reduction and k-fold validation were employed in each cases. The algorithms were compared based on the training time, prediction time and accuracy of the prediction.

**V. APPLICATIONS**

One clear sign of advancement in ML is its important real-life applications, some of which are briefly described here.It is to be noted that until 1985 there was no significant commercial applications of ML algorithms.

1. **SPEECH RECOGNITION**

All current speech recognition systems available in the market use machine learning approaches to train the system for better accuracy. In practice, most of such systems implement learning in two distinct phases: pre-shipping speaker independent training and post-shipping speaker-dependent training.

B. **COMPUTER VISION**

Majority of recent vision systems, e.g., facial recognition software’s, systems capable of automatic classification microscopic images of cells, employ machine learning approaches for better accuracy. For example, the US Post Office uses a computer vision system with a handwriting analyzer thus trained to sort letters with handwritten addresses automatically with an accuracy level as high as 85%.

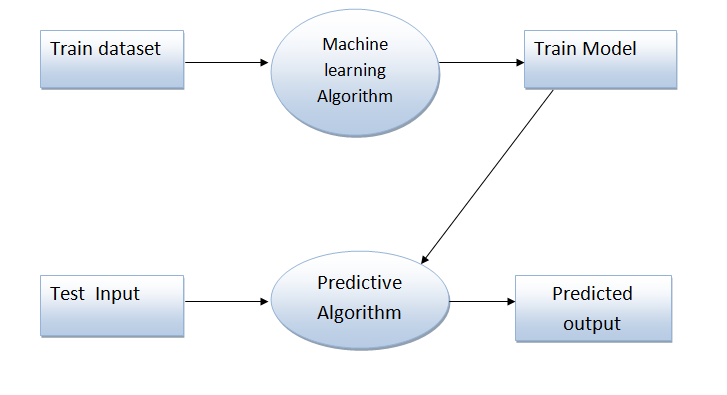
1. **BIO-SURVEILLANCE**

Several government initiatives to track probable outbreaks of diseases use ML algorithms. Consider the RODS project in western Pennsylvania. This project collects admissions reports to emergency rooms in the hospitals there, and the an ML software system is trained using the profiles of admitted patients in order to detect aberrant symptoms, their patterns and areal distribution. Research is ongoing to incorporate some additional data in the system, like over-the counter medicines’ purchase history to provide more training data. Complexity of this kind of complex and dynamic data sets can be handled efficiently using automated learning methods only.

1. **ROBOT OR AUTOMATION CONTROL**

ML methods are largely used in robot and automated systems. For example, consider the use of ML to obtain control tactics for stable flight and aerobatics of helicopter [9]. The self driving car developed by Google uses ML to train from collected train data. A large group data-intensive science disciplines use ML methods in several of it researches. For example, ML is being implemented in genetics, to identify unusual celestial objects in astronomy, and in Neuroscience and psychological analysis. The other small scale yet important application of ML involves spam filtering, fraud detection, topic identification and predictive analytics (e.g., weather forecast, stock market prediction, market survey etc.).

**METHODOLOGY**

****

**Training Datasets**

A training dataset is a dataset of examples used for learning, which is to fit the parameters (e.g., weights) of, for example, a classifier. Most approaches that search through training data for empirical relationships tend to over fit the data, meaning that they can identify apparent relationships in the training data that do not hold in general.

**Naive Bayes Algorithm**

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature [6]. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or 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 why it is known as ‘Naive’. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

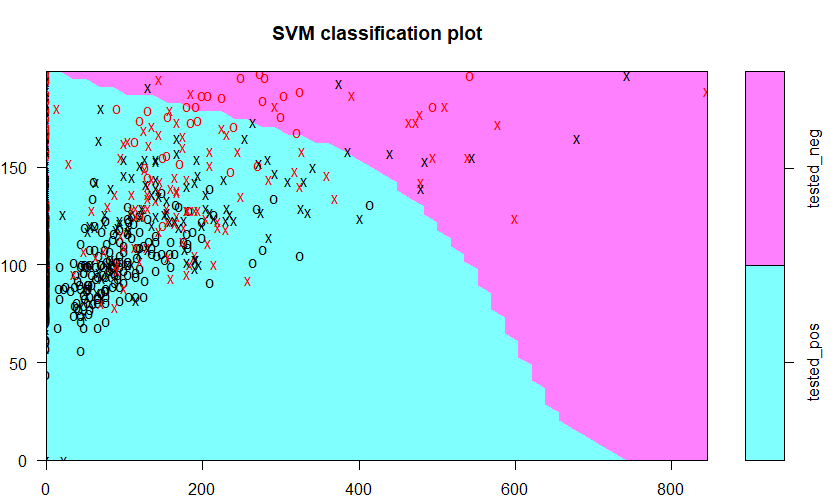


Above,

* *P*(*c|x*) is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
* *P*(*c*) is the prior probability of *class*.
* *P*(*x|c*) is the likelihood which is the probability of *predictor* given *class*.
* *P*(*x*) is the prior probability of *predictor*.

## Support Vector Machine

Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges [5]. However,  it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.



**Train Model**

The process of training an ML model involves providing an ML algorithm (that is, the *learning algorithm*) with training data to learn from. The term *ML model* refers to the model artifact that is created by the training process.

The training data must contain the correct answer, which is known as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.

**Input**

Test data is data which has been specifically identified for use in tests, typically of a computer program. Some data may be used in a confirmatory way, typically to verify that a given set of input to a given function produces some expected result. Other data may be used in order to challenge the ability of the program to respond to unusual, extreme, exceptional, or unexpected input.

**Predicted Output**

After the analysis of large diabetic data went, the final results are distributed over various server. By employing proper electronic communication technology to exchange the information of individual patients among health care centres will leads to get proper treatment at right time in remote locations at low cost.

**IMPLEMENTATION**

**Naive Bayes Algorithm**

require(e1071) #Holds the Naive Bayes Classifier

Train <- read.csv(file.choose())

Test <- read.csv(file.choose())

#Make sure the target variable is of a two-class classification problem only

levels(Train$Item\_Fat\_Content)

model <- naiveBayes(Item\_Fat\_Content~., data = Train)

class(model)

pred <- predict(model,Test)

**SVM Algorithm**

#Import Library

require(e1071) #Contains the SVM

Train <- read.csv(file.choose())

Test <- read.csv(file.choose())

# there are various options associated with SVM training; like changing kernel, gamma and C value.

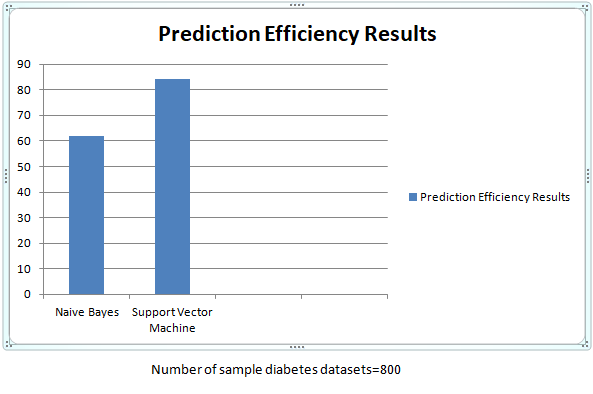
# create model

model <- svm(Target~Predictor1+Predictor2+Predictor3,data=Train,kernel='linear',gamma=0.2,cost=100)

#Predict Output

preds <- predict(model,Test)

**RESULT**

****

As per the above implementation of machine learning algorithms on diabetes sample datasets(Number of datasets 800), we came across the result that Support Vector Machine learning algorithm is more efficient in predicting compare to Naive Bayes algorithm.

**CONCLUSION**

Data sets are created from health reports of various diabetic patients from different geographical regions. These Train Datasets are given as input to the Naïve Bayes Algorithm.This Algorithm will produces Train Module from the given Train Datasets. Prediction algorithm will predict the diabetes rate over the region as well as error rate. Finally we compare the efficiency of prediction of different machine learning algorithm like Naive Bayes and SVM for given dataset.

**REFERENCES**

**[1**] Raghupathi W: Data Mining in Health Care. Healthcare Informatics: Improving Efficiency and Productivity. Edited by: Kudyba S. 2010, Taylor & Francis, 211-223

**[2]** Harrell, Frank E., Kerry L. Lee, and Daniel B. Mark. "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors." *Statistics in medicine* 15.4 (1996): 361-387.

**[3]** Schneeweiss, Sebastian. "Learning from big health care data." *New England Journal of Medicine* 370.23 (2014): 2161-2163.

**[4]** International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Website: www.ijircce.com Vol. 5, Issue 2, February 2017 Copyright to IJIRCCE DOI: 10.15680/IJIRCCE.2017. 0502001 1301 A Survey on Machine Learning: Concept, Algorithms and Applications Kajaree Das1 , Rabi Narayan Behera 2 B.Tech, Dept. of I.T., Institute of Engineering and Management, Kolkata, India1 Asst.

**[5]** N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, 2000.

**[6]** K.P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012.

**[7]** K.P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012.

**[8]**  E. Alpaydın, Introduction to Machine Learning, 3rd ed., MIT Press, 2014.

**[9]** Autonomous Mobile Robot Navigation using Machine Learning Xiyang Song, Huangwei Fang, Xiong Jiao, and Ying Wang School of Engineering, Southern Polytechnic State University, Marietta, GA, 30060, USA.