# Study On Comparative Performance Analysis of Different Machine Learning Algorithms


## Submitted By:

**Syeda Nadia Islam**

**ID: 2015-3-50-003**

**Nusrat Jahan**

**2015-3-50-007**


## Submitted To

**Dr. Anup Kumar Paul**

**Assistant Professor**

**Department of ECE**

# Study On Comparative Performance Analysis of Different Machine Learning Algorithms

**Acknowledgement**

At first, all praises belong to the Almighty, the most graceful and most merciful.

In the development of this dissertation we are endowed with the guidance and the help of several individuals who assisted us in many ways towards the development of this project. We are greatly indebted to our respected supervisors Dr. Anup Kumar Paul for his guidance and valuable contribution to the development of this thesis. Hope his valuable advice will be helping us in future also.

We would also like to express our gratitude to the respected teachers of the Department of Electronics and Communications Engineering for their encouragement and constructive suggestions.

We are grateful to the Department for providing access to the Departmental Laboratory during the course of the project.

Last but not the least, we are thankful to our family members for supporting us by providing inspiration in stressful times.

**Abstract**

Diabetes mellitus is a group of metabolic disorders known as ‘diabetes’ ,it has affected hundreds of millions of individuals. Diabetes detection is of great significancewith regard to its serious complications. Many studies on diabetes prediction datasets have been conducted,where most of it are the studies on diabetes collected from individual,and it is also where the onset of diabetes dataset is high,studying the female in Pima Indian natives population during 1967. Most of the previous studies concentrated primarily on one or two specific complicated technique to test the data, while there is a lack of extensive research on popular technique . In this paper, we are conducting a thorough exploration of the most common technques like SVM (Support Vector Machine), (K Nearest Neighbors),etc.) used for identifying diabetes and other preprocessing method. Basically, we examine these techniques by precision of cross-validation on the dataset. We compare each classifier ‘s apects of analyzing and we modify the parameters to improve their accuracy. The best technique we find has 77.86% accuracy using 10-fold cross-validation.

**Overview**

## DESCRIPTION

Predict the onset of diabetes based on diagnostic measures.

## SUMMARY

This dataset originates from the Diabetes and Digestive and Kidney Diseases National Institut e.The goal is to predict whether a patient has diabetes based on diagnostic measurement.

Several limitations have been put on a bigger database choice of these cases.
In specific, all patients here are Pima Indian heritage women at back 21 years old.

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)

**Inspiration:**

1.Some values should not be regarded as missing values in the range where they are intended to be.

2.What technique is best to use to fill the missing value of this sort? How is the classification going to be like?

3. Is there a considerably higher likelihood of diabetes among subgroups?

# Chapter 1: Introduction

## Chapter 1 : Introduction

The Pima is a group of Arizona based indigenous Americans. A genetic predisposition has enables this group to survive for years in a diet that is poor in carbohydrates. In latest years they have developed the greatest incidence of type 2 diabetes due to a sudden change from traditional agricutural plants to processed foods, along with a decrease in physical activity, and this is the reason for which they have been a topic of many research.

Diabetes is a group of metabolic disorders with elevated concentrations of blood sugar over a n extended period of time.High blood sugar symptoms include frequent urination, higher thirs t, and hunger.It can cause many problems if left untreated. Diabetic ketoacidosis, hyperosmol ar hyperglycemic condition, or death may include acute complications. Serious complications in the long term include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, a nd eye harm. And
it has an instant elevated glucose flag, along with a few side effects including constant pee, i ncreased thirst, increased yearning, and weight loss.For the most portion, diabetes patients re quire continuous therapy, otherwise many dangerous complications may be prompted.
Diabetes is determined to have no less than 200mg / dL of the 2-
hour poststack plasma glucose[1 ], and the need to recognize diabetes conveniently results in various diabetes recognition tests.
Numerous previous study thoughts have been made on the identifiable evidence of machine l earning in diabetes. Research has been conducted focused on identifiable evidence of diabetes through various machine learning algorithms, and some rousing results have been achieved.
In contrast to past work, we are conducting a more comprehensive spot algorithms that suggesting to identify the best one among them in terms of accuracy.

Machine Learning (ML),an artificial intelligent sun field has been developed out of the need to teach computers on how to respond to a problem .

**Objective:**

We will try to build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not?

**Thesis Structure**

This paper is mainly divided into six chapters.

Chapter 1 introduces the topic of Defining the problem.

Chapter 2 presents the summarization of data.

Chapter 3 is data preprocessing.

Furthermore, Chapter 4 is about spot check of algorithm.

Chapter 5 is about Comparing the algorithms.

Chapter 6 is about predicting the model.

# Chapter 2: Descriptive and Summary Statistics

## Chapter 2: Descriptive and Summary Statistics

The first step in any ML analysis is to examine the data using summary statistics. These statistics address two important issues that can impact the quality of the analysis:

1. **Data quality issues:** e.g., outliers and missing values can change the relationship between model outputs and inputs
2. **Distributions of inputs:** ML algorithms are often more predictive when the inputs are standardized.

   The describe() function on the pandas dataframe lists 8 statistical properties of each attributes,they are:

   a.count-it gives the number of how many times the attributes has appeared.

   b.mean-it gives us the mean of each attributes

   c.standard deviation-

   d.minimum value-it gives us the minimum value of each attribute

   e.25$^{th}$ percentile

   f.50$^{th}$ percentile

   g.75$^{th}$ percentile

   h.maximum value-it gives us the maximum of each attributes

   code:

   #load the csv file using read_csv function of pandas library

   from pandas import read_csv

   from pandas import set_option

   filename = 'pima-indians-diabetes.csv'

```
#url = 'https://myfilecsv.com/test.csv'

names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']

data = read_csv(filename, names=names)


set_option('display.width', 200)

set_option('display.max_columns', 10)

set_option('precision', 3)


description = data.describe(), // gives us the decription of each attributes

print(description)
```

**output:**

```
          preg     plas     pres     skin     test     mass     pedi      age
class
count  768.000  768.000  768.000  768.000  768.000  768.000  768.000  768.000
768.000
mean     3.845  120.895   69.105   20.536   79.799   31.993    0.472   33.241
0.349
std      3.370   31.973   19.356   15.952  115.244    7.884    0.331   11.760
0.477
min      0.000    0.000    0.000    0.000    0.000    0.000    0.078   21.000
0.000
25%      1.000   99.000   62.000    0.000    0.000   27.300    0.244   24.000
0.000
50%      3.000  117.000   72.000   23.000   30.500   32.000    0.372   29.000
0.000
75%      6.000  140.250   80.000   32.000  127.250   36.600    0.626   41.000
1.000
max     17.000  199.000  122.000   99.000  846.000   67.100    2.420   81.000
1.000
```

Here, the describe() function give us the count, mean,maximum, minimum and percile values of each of 8 attributes.

### 2.1 Data Visualization:

Since, we have already analysd the data set by using statistic to get the meaning of each attribute. But now, inorder to better understand them, we use data visualization because:

Data visualization is the act of capturing and putting information (data) in a visual context, su ch as a map or graph.

Data visualizations facilitate the understanding of large and small data for the human brain, a nd visualization also facilitates the detection of patterns, trends and outliers

in data groups.Also,good visualizations of information should position meaning in complex d atasets to make their message clear and concise.

Inorder to better understand visually, we use **univariate plots** ,it will us to understand each attribute in a data set individually. We ill be using 3 visualization techniques to understand our datasets:

1)Histogram

2)Density Plots

3)Box and whisker plots

#### 2.1.1 Histogram

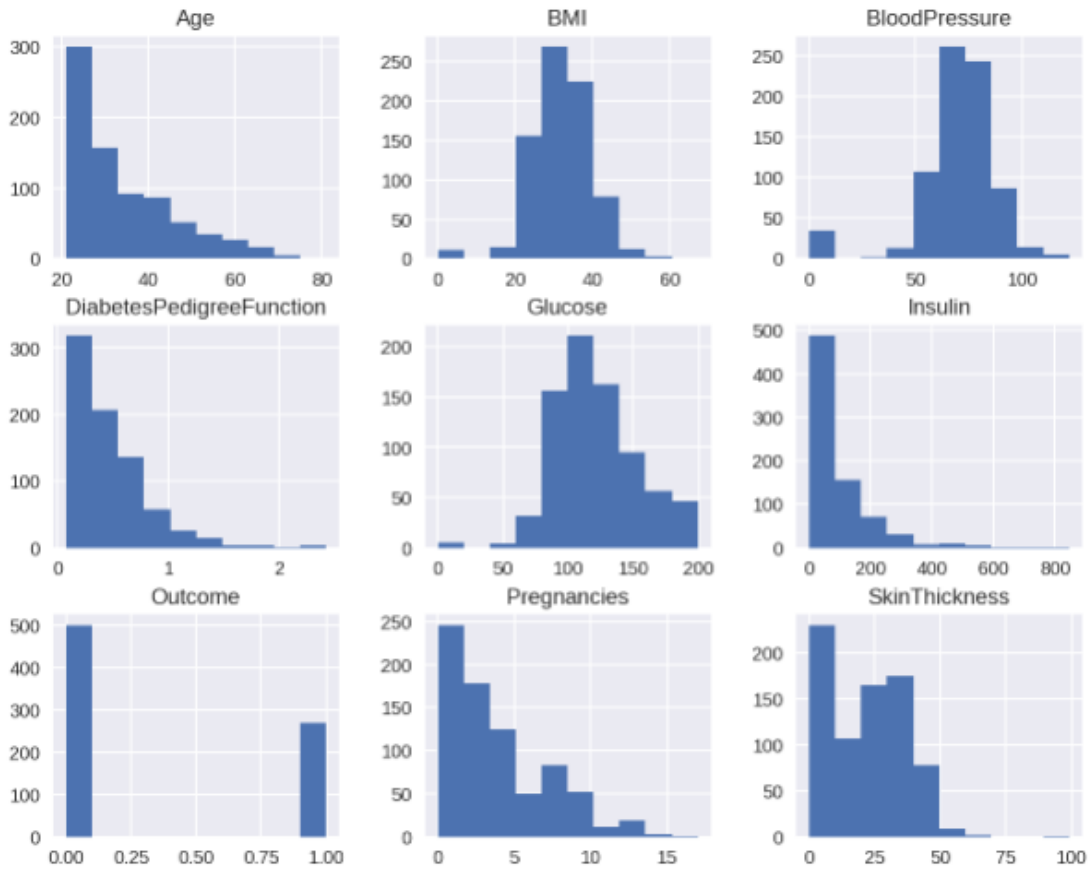
Histogram groups data into bins and count of observations into each bins.

Code:

```
from matplotlib import pyplot
#load the csv file using read_csv function of pandas library
from pandas import read_csv
filename = 'pima-indians-diabetes.csv'
#url = 'https://myfilecsv.com/test.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)

data.hist()
pyplot.show()
```

output:




We can't comprehend them better, but we can get from the bins form whether the attribute is gauussian, skewed or even exponential.

So to get the shape of each distribution for each attributes , we use density plots.

### 2.1.2 Density Plots

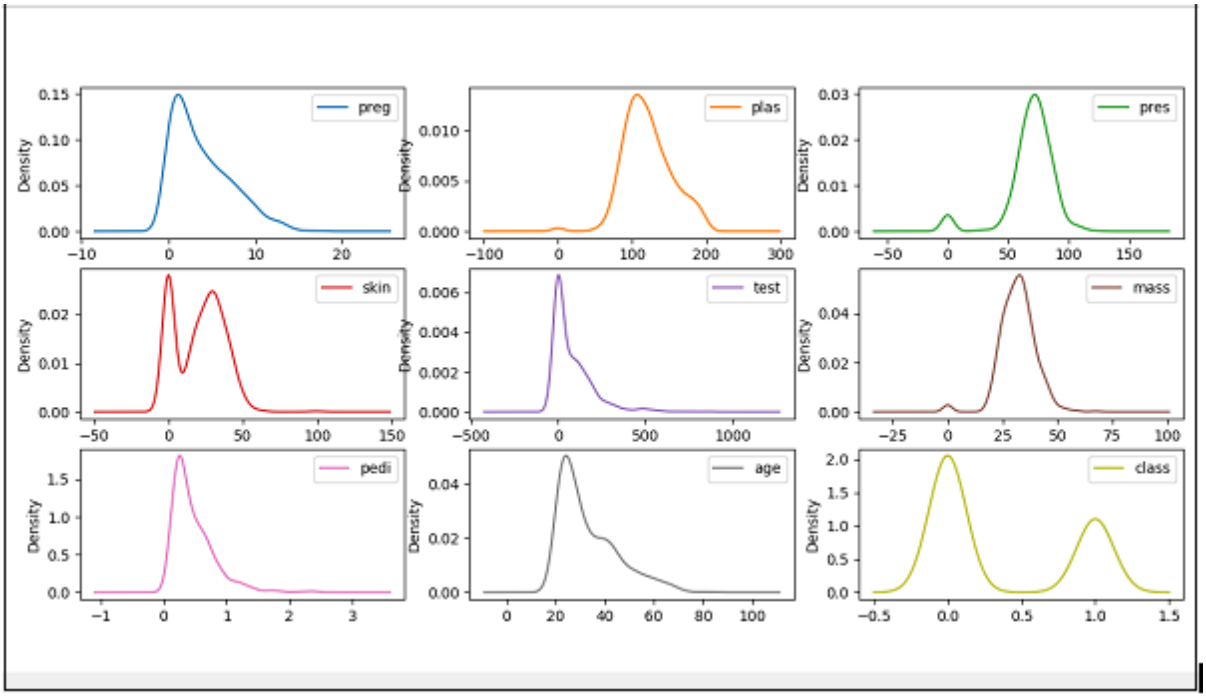
Density plot is an abstracted histogram that is curved smoothly through the top of each bin.

Code:

```
from matplotlib import pyplot
#load the csv file using read_csv function of pandas library
from pandas import read_csv
filename = 'pima-indians-diabetes.csv'
#url = 'https://myfilecsv.com/test.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)

data.plot(kind='density', subplots=True, layout=(3,3), sharex=False)
pyplot.show()
```

Output:



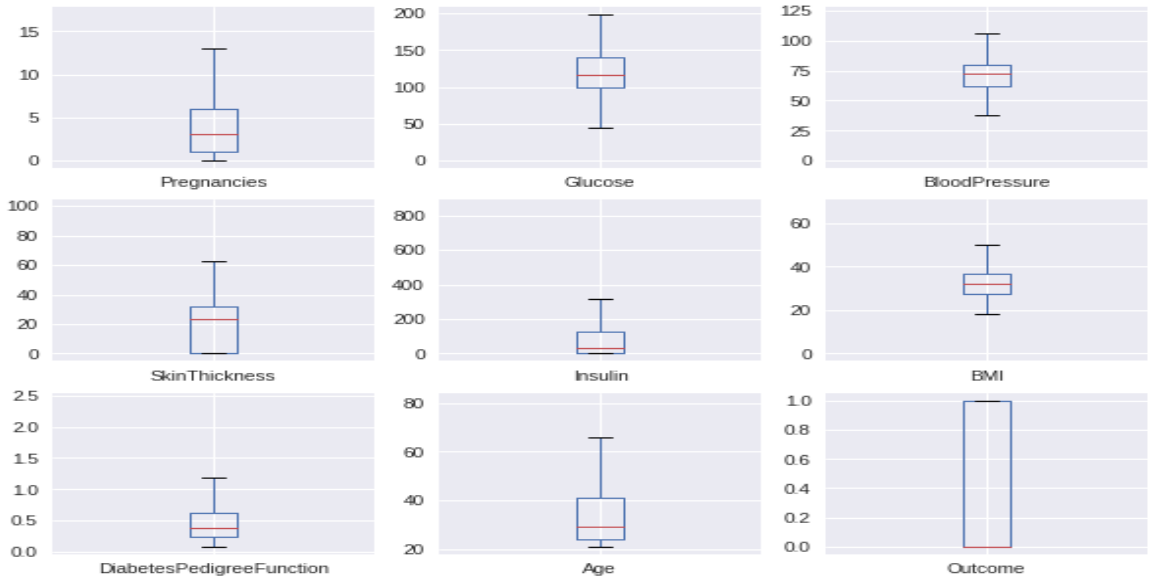

Now, we can better understand the histogram, that is from the output, it seems that attributes such as age,pedi and test have an exponential distribution.

Also , the mass, plasma and pres attribution have an gaussian or nearly gaussian distribution.

### 2.1.3 Box and Whisker Plots

Boxplots summarize each attribute's distribution, drawing on a median (middle) row, and a b ox around 25th and 75th percentile. Whiskers give an impression of disseminating informatio n outside the whiskers.

Output:




Here, from the box-plot output , it seems that,the spread of each attributes is quite different. Some like ages,test and skin appear quite skewed toward smaller value.

# Chapter 3: Data Preprocessing

## Chapter 3:Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format [1]- https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa.

Realworld information in certain behaviors or trends is often incomplete, inconsistent and/or missing, and is likely to contain many mistakes.

Real world informatiom in certain behaviours or trends is often incomplete, inconsistent or missing and is likely to contain more mistakes.

So, in our Diabetes datasets, we find from the histogram that, in the past analysis,we have noticed that for some of the characteristics of the patients are missing. When data is missing, machine learning algorithms don′t perform well. Here , are the following missing datas of attributes:-

1)**Blood Pressure**: By observing, we notice 0 blood pressure values. So it is obvious that the dataset appear to be incorrect because a living individual is unable to have zero diastolic blood pressure.We can see 35 counts where the value is 0 from the visualization graph

2**) Plasma glucose levels:** It would nt be as small as zero even after fasting,therefore zero is invalid to read.

3) **Skin Fold Thickness :** Skin folding thickness cannot be less than 10 mm for individual .

4) **BMI** :Unless the individual is really underweight that could be life threathening, it should not be zero or near to zero.

so we need to discover a solution to "clean" the data we have. That is there are 4 ways of data preprocessing:-

1) Data Rescale.

2) Data Standardization

3) Data Normalization

4) Data Binarization

So, for our data preprocessing, we use data standardization.

### 3.1 Standardize data

Standardize data is comparing data with different units using Standardize Deviation is called standardize.
Using the following formula, we compare individual information values with their mean relat ive to their standard deviation:

$$Z=\frac{\square-\square'}{\square}$$

The standardized value is denoted by z .

Here we convert the characteristics with a Gaussian distribution and different means and standard deviation into a Gaussian normal distribution with a mean of 0 and a standard deviation of 1. It is best suited for methods such as logistical regression and linear discriminating assessment which assume a Gussian distribution in the input variable and operate better with rescaled data. [2]-

Output

```
[[ 0.64   0.848  0.15   0.907 -0.693  0.204  0.468  1.426]
 [-0.845 -1.123 -0.161  0.531 -0.693 -0.684 -0.365 -0.191]
 [ 1.234  1.944 -0.264 -1.288 -0.693 -1.103  0.604 -0.106]
 [-0.845 -0.998 -0.161  0.155  0.123 -0.494 -0.921 -1.042]
 [-1.142  0.504 -1.505  0.907  0.766  1.41   5.485 -0.02 ]]
```

here, we see in the output that all the attributes data are between the gaussian distribution with a mean 0 and standard deviation of 1.

# Chapter 4: Spot-Check Algorithm

## Chapter 4: Spotcheck Algorithms

Spotchecking algorithms is an applied machine learning method s aim to deliver a first set of results rapidly and pointedly on a fresh predictive modeling issue. It is is how you find the best algorithm for your dataset. The reason we are using spot checking algorithm is that,
Unlike grid search and other algorithm tuning for the optimal configuration of the algorithm, The spot check is intended to evaluate a range of algorithms rapidly and provide a rough unde rstanding of the problemoutcome.
This first cut result can be used to get an understanding of the predictability of a problem or p roblem representation and, if so, the types of algorithms that can be used.

The importance of spot checking agorithm is that it is used for classification and regression

on a predictive modelling analysis to create a normal structure that is :-

1)Spotchecking allows you to rapidly uncover the kinds of algorithms that work well with yo ur predictive modeling issue.

2)How generic information loading framework can be developed, models defined, models ev aluated and findings summarized?

3) How to apply classification and regression for the problem framework?

The answer to this is,

Spotchecking is a way to address this issue. It includes testing a wide range of different machi ne learning algorithms on a issue quickly to find out what algorithms might work and where t o concentrate attention. The reasons are as follows:-

- **It's quick :** it bypasses the days or weeks of preparing and analyzing and playing with algorit hms that may never result
- **It's objective**: enabling you to find out what might work well for an issue, rather than what y ou used last time.
- **Quick Results :** That is it is going to effectively fit models,create projections and understand the issue whether it can be predicted and what it might looks like. It involves working with small samples of datasets to rapidly turnout the results.

Spot check is for both classification and regression algorithms.

In this project, trial and error method is implemented to detect number of algorithms that work well with the problem and tune it down further .

For Spot-checking , We will be using 6 machine learning models:-

1)Logistic Regression

2)Linear Discriminant Analysis

3)K-Nearest Neighbour

4)Classification And Regression Trees(CART)

5)Support Vector Machines

6)Naives Bayes

We will be simply checking the accuracy for each model and nothing more.

The following parameters are used in this research as input pregnancies, glucose, blood press ure, thickness of the skin, insulin, BMI, pedigree function of diabetes, and age. A number of machine learning and statistical techniques can be used to forecast illnesses of diabetes.Based on the scope of the literature, we settled on using six of the most well-known machine learning algorithms (Logistic Regression, KNN, Naïve Bayes, Linear Discri minant Analysis, CART, Support Vector Machine) classification algorithms and/or ensemble s in one using base learner. These methods of classification and their distinctive specifications used in this project are described in the following chapter.

### 4.1 Logistic Regression

Logistic regression is another technique borrowed by machine learning from the field of statistics.

In statistics Logistic regression is a regression model where the dependent variable is categori cal, i.e. binary dependent variablethat is, where only two values, "0" and "1," represent results such as pass / fail, win / lose, alive / dead or healthy / sick, can be taken.
The method can also be used in engineering, in particular to predict a process, system or prod uct's likelihood of failure. It is also used in marketing apps such as predicting the propensity o f a customer to buy a product or stopping a subscription

It is the go-to method for binary classification problems (problems with two class values). In this post you will discover the logistic regression algorithm for machine learning. [3]-

We can construct a logistic regression model using the logistic regression class .

Code:

```
"""
#load the csv file using read_csv function of pandas library
from pandas import read_csv
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression

filename = 'pima-indians-diabetes.csv'
#url = 'https://myfilecsv.com/test.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
dataframe = read_csv(filename, names=names)

array = dataframe.values

#splitting the array to input and output
X = array[:,0:8]
Y = array[:,8]

num_folds = 10
seed = 7

kfold = KFold(n_splits = num_folds, random_state = seed)
model = LogisticRegression(solver='liblinear')

results = cross_val_score(model, X, Y, cv=kfold)
print("Mean Estimated Accuracy Logistic Regression: %f " % (results.mean()))
```

Output:

```
Mean Estimated Accuracy Logistic Regression: 0.769515
```

In logistic regression , the accuracy is about 77 %.

### 4.2 Linear Discriminant analysis

If you have more than two classes then Linear Discriminant Analysis is the preferred linear classification technique. It is used when there is more than one clasees.

**Limitations of Logistic Regression**

Logistic regression is a linear classification algorithm, that is convinient and most preferred in majority of the classification but it also has some constraints that indicate the requirement for alternate linear classification algorithms.

- Logistic Regression is designed for two classes,that is binary classification. It can also be expanded in multiple class classification but it is never used in majority of the cases.
- If the classes are well separated. Logistic Regression will not be able to classify .
- Logistic regression can fluctuate when there are less variance parameters to be estimated.

Linear Discriminant Analysis addresses each of these points and is the go-to linear method for multi-class classification to solve the issues. Even with binary-classification for solution of each problem, attempting both logistic regression and linear discriminant analysis is a good idea

Here in our datasets, we have only one class that is the result, whether it is ‘1’ or ‘0’. Linear Discriminant Analysis is done to project the features in higher dimension space onto a lower dimensional space. [4]

Output:

```
Mean Estimated Accuracy Linear Discriminant: 0.773462
```

### 4.3 K-Nearest Neighbour

In machine learning, KNN is conceptually simple and shown to be a popular non-parametric technique for classification . [5]

It has been extensively studied and widely exploited in many fields such as data mining, image processing and statistical pattern recognition for many years owing to its superiorities.[5]-, [6] [7] [8]. Also, The k-NN algorithm is probably the easiest machine learning algorithm. The model building comprises only of storing the data set for training. The algorithm only searches for data point that is its nearest data,it is called 'nearest neighbour'.

It is sort of tiring where we have use a function called K classifier in KNN, the function willingly approximates the points in vicinity until it is ready for classification.

Best part of the KNN system is that its neighbour when found classifies an entity. K is positive integer all the time. The right classification is known as neighbours from a set of neighbours.Knearest neighbor algorithm is a straightforward method that stores all accessible instances and classifies fresh similarity-based instances.

Output:

```
Mean Estimated Accuracy KNeighbors: 0.726555
```

Here, the accuracy of training set in KNN is about 73%.

### 4.4 Support Vector Machine (SVM)

SVM is a collection of associated supervised learning methods used for classification and re gression in medical diagnosis. At the same time, SVM minimizes the error of empirical classif ication and maximizes the geometric margin. SVM is therefore called Maximum Margin Clas sifiers.SVM is a particular algorithm based on statistical learning theory's guaranteed risk bo undaries, known as the concept of structural risk minimization.[9]-

Support Vector Machine (SVM) is based on the notion of judgment planes that describe the consequence limitations. A judgment notion is one that separates a set of objects with a class membership truthfulness.

SVM is an appropriate method for binary classification assignments, so we choose SVM for diabetes prediction.SVM is a technique suitable for binary classification tasks, so we choose SVM to predict the diabetes.

Using what is called the kernel trick,
SVMs can perform nonlinear classification efficiently, mapping their inputs into high-dimensional feature spaces. The kernel trick allows the classifier to be constructed without the functio n's room explicit knowledge.

The regular SVM uses collection of input data and predicts which input consists of two feasib le classes. An SVM representation of the examples as space points, mapped as a clear divid e as wide as possible to divide the cases of the distinct categories.

Code:

```
#load the csv file using read_csv function of pandas library
from pandas import read_csv
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.svm import SVC

filename = 'pima-indians-diabetes.csv'
#url = 'https://myfilecsv.com/test.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
dataframe = read_csv(filename, names=names)

array = dataframe.values

#splitting the array to input and output
X = array[:,0:8]
Y = array[:,8]

num_folds = 10
seed = 7

kfold = KFold(n_splits = num_folds, random_state = seed)
model = SVC(gamma='auto')

results = cross_val_score(model, X, Y, cv=kfold)
print("Mean Estimated Accuracy SVM: %f " % (results.mean()))
```

Output:

```
Mean Estimated Accuracy SVM: 0.651025
```

### 4.5 CART

Together Leo Breiman, Jerome Friedman, Richard Olsen and Charles Stone developed an algorithm called the Classification and Regression Tree (CART) and developed a periodic scheme for the development of easy-to-face arithmetic models. CART is important while dealing with non-complete information, established data and contribution features.

The method will study some of the instances related to the information description will tip the information minimization and will continue until some stop criteria have been reached. [10] Decision tree constructs models of regression or classification as a tree structure. It breaks do wn a dataset into smaller and smaller subsets while increasingly developing an related decisio n tree. The end outcome is a tree with nodes of choice and nodes of leaf.

We construct a CART model using the Decision Tree Classifier Class.

Code:

```
#load the csv file using read_csv function of pandas library
from pandas import read_csv
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.tree import DecisionTreeClassifier

filename = 'pima-indians-diabetes.csv'
#url = 'https://myfilecsv.com/test.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class
dataframe = read_csv(filename, names=names)

array = dataframe.values

#splitting the array to input and output
X = array[:,0:8]
Y = array[:,8]

num_folds = 10
seed = 7

kfold = KFold(n_splits = num_folds, random_state = seed)
model = DecisionTreeClassifier()

results = cross_val_score(model, X, Y, cv=kfold)
print("Mean Estimated Accuracy CART: %f " % (results.mean()))
```

Output:

```
Mean Estimated Accuracy CART: 0.684774
```

In this output,the accuracy is about 68 % in CART algorithm.

### 4.5 Naïve-Bayes

The classification of Bayesian is based on the theorem of Bayes. Naive Bayesian classifiers suppose that the impact on a specified class of an attribute value is independent of the other attribute values. This hypothesis is called independence conditional class. Simplifying the computation concerned is done and is deemed "naive" in this sense. Let X={ x1, x2,. xn} is a sample with parts representing values created on a set of n characteristics. X is regarded "proof" in Bayesian terms. Let H be some hypothesis, such as that information X belongs to a particular class C. P(H), the probability that the H hypothesis holds given the "proof," (i.e. the data sample X observed) must be determined.

The probability we want to calculate P(H) can be expressed in terms of probabilities P(H), P( X), and P(X) as P(H) = P(X) P(H)/P(X), according to Bayes ' theorem. [11]

Classifiers of Naive Bayes suppose that characteristics have separate distributions. It is regarded quick and effective in space. It also provides simple approach, with clear semantics, representing and learning probabilistic knowledge.It is known as Naive because it relies on two important simplifying assumptions. The predictive attributes are conditionally independent .[12It is quick to train and effective to classify.

Code:

```
"""
#load the csv file using read_csv function of pandas library
from pandas import read_csv
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.naive_bayes import GaussianNB

filename = 'pima-indians-diabetes.csv'
#url = 'https://myfilecsv.com/test.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
dataframe = read_csv(filename, names=names)

array = dataframe.values

#splitting the array to input and output
X = array[:,0:8]
Y = array[:,8]

num_folds = 10
seed = 7

kfold = KFold(n_splits = num_folds, random_state = seed)
model = GaussianNB()

results = cross_val_score(model, X, Y, cv=kfold)
print("Mean Estimated Accuracy Naive Bayes: %f " % (results.mean()))
```

Output:

```
Mean Estimated Accuracy Naive Bayes: 0.755178
```

# Chapter 5: Choosing the best machine learning model

**Chapter 5: Choosing among the six machine learning model**

Here, in this, six different classification algorithms compared on a single datasets.

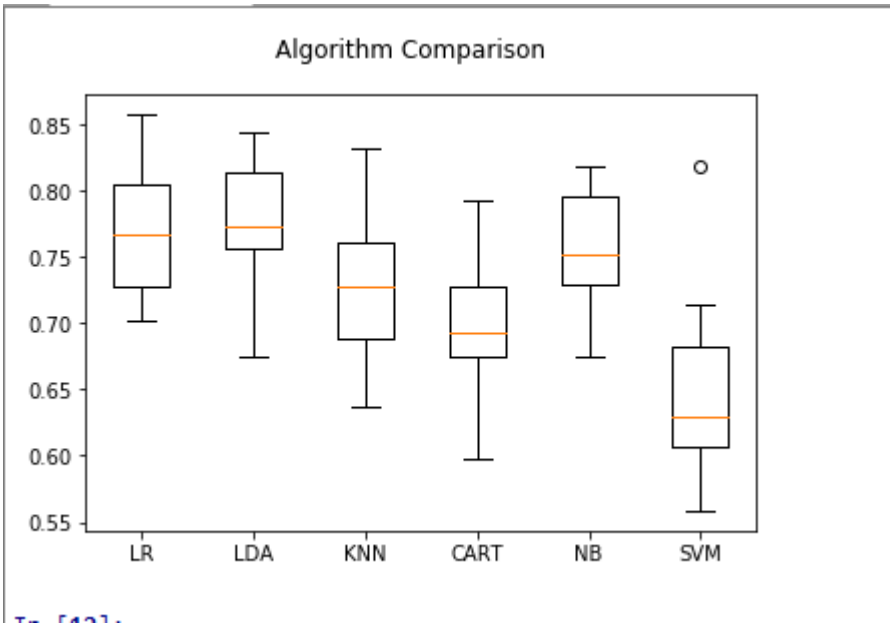
Each machine learning model has different performance characteristics. So,here, we use resampling methods like, we can obtain an estimate of how precise an unseen data for each model.

To define the problem here, we are first

1) The dataset is the onset of pima indian diabetes problem.
2) There are two classes and 8 differing input of varying scale.
3) Each algorithm is evaluated using 10 Fold Classification,has been configured with the same random seed.

Output:

```
LR: 0.769515 (0.048411)
LDA: 0.773462 (0.051592)
KNN: 0.726555 (0.061821)
CART: 0.706989 (0.060623)
NB: 0.755178 (0.042766)
SVM: 0.651025 (0.072141)
```




In this output , Logistic regression and linear discriminant analysis are close to 77 %

after comparing to five other distinct macgine learning algorithm.

**Predicting the model**

Since, we have already analysed six different machine learning model, we are going to create a model based on logistic regression to predict whether we are going to have diabetes or not.

**Code:**

```
from pickle import load

#Name of the saved model from another computer
filename = 'final_pima_indian.sav'

#load the file
loaded_model = load(open(filename, 'rb'))


# define one new data instance for prediction
Xnew = [[0,132,90,33,167,40.1,3.288,30]]

# make a prediction
ynew = loaded_model.predict(Xnew)
print("Input =%s, Predicted =%s" % (Xnew[0], ynew[0]))
```

**Output:**

```
Input =[0, 132, 90, 33, 167, 40.1, 3.288, 30], Predicted =1.0
```

After giving my inputs for each attributes, the machine using logistic regression in the back end predict that I will be having diabetes . Since , it is giving ‘1’ as an output.

**Conclusion:**

Various data mining techniques and its application were studied or reviewed .application of machine learning algorithm were applied in different medical data sets Machine learning methods have different power in different data set. Single algorithm provided less accuracy than ensemble one. Diabetes mellitus is a condition that can cause a lot of complications. It is worth studying on how to predict and diagnose this disese accurately using machine learning.

As performance of kNN algorithm has minimum accuracy. Based on the parameters taken for analysis, the perfor-mances of the six algorithms are analyzed. The results show that the performance of Linear discriminat Analysis technique is significantly superior to the other five techniques for the classification of diabetes data. To improve the overall accuracy, it is necessary to use more data set with large number of attributes and use hybridisation model in the future.

**Future work**

In other words, today's computers can use patient information from various sources to diagnose disease, inform therapy choices, and predict results, including genomic sequencing and sensors.

It's my goal to introduce diabetes to the AI revolution. Machine Learning and Artificial Intelligence can obtain data from different machines to generate customized progrogrammes that gives importance to medical adherance and management of glucose level.

## References


[1]- machinelearningmastery.com

[2]- www.i-scholar.in

[3]-www.irjet.net

[4]-briangriner.github.io

[5]-www.pgtfb.com

[6]-dr.ntu.edu.sg

[7]- Submitted to Gergia Institute of Technology Main Campus

[8]- Submitted to Univerrsity of Central England in Birmingham

[9]-Mahuri Panwar,Amit Aacharya, Rishad A Shafiq,Dwaipayan Biswas."K-nearest neighbour based methology for accurate diagnosis of diabetes mellitus",2016 Sixth International Symposium on Embedded Computing and System Design (ISED),2016

[10]-eprint.ncl.ac.uk

[11]-Submitted to Chittagong University of Engineering And Technology

[12]-pdfs.semanticshilar.org

[13]- Mangesh J. Shinde .. "COMPARATIVE STUDY OF DECISION TREE ALGORITHM AND NAIVE BAYES CLASSIFIER FOR SWINE FLU PREDICTION", International Journal of Research in Engineering and Technology, 2015

[14]- Submitted to University of Northumbria at Newcastle

[15]-www.tandfondline.com

[16]-baadalsg.inflibnet.acin

[17]-Submitted to La Sagesse University

[18]-Submitted to Middle East Technical University

[19]-onlinedatasciencecourses.com

[20]-m.scirp.org

[21-]Janvier Omar Sinayobye, Swaib Kaawaase Kyanda, N. Fred Kiwanuka, Richard Musabe. "Hybrid Model of Correlation Based Filter Feature Selection and Machine Learning Classifiers Applied on Smart Meter Data Set", 2019 IEEE/ACM Symposium on Software Engineering in Africa (SEiA), 2019