**PROJECT ON DATA SCIENCE:**

**ABSTRACT:**

The remarkable advances in biotechnology and health sciences have led to a significant production of data, such as high throughout genetic data and clinical information, generated from large [Electronic Health Records](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/electronic-health-record) (EHRs). To this end, application of machine learning and [data mining](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/data-mining) methods in biosciences is presently, more than ever before, vital and indispensable .Diabetes mellitus (DM) is defined as a group of [metabolic disorders](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/metabolic-disorders) exerting significant pressure on human health worldwide. Extensive research in all aspects of diabetes (diagnosis, etiopathophysiology, therapy, etc.) has led to the generation of huge amounts of data. A wide range of machine learning algorithms were employed. In general, 85% of those used were characterized by supervised learning approaches and 15% by unsupervised ones, and more specifically, association rules.Concerning the type of data, clinical datasets were mainly used. The title applications in the selected articles project the usefulness of extracting valuable knowledge leading to new hypotheses targeting deeper understanding and further investigation in DM. The aim of the present study is to conduct a systematic review of the applications of machine learning, data mining techniques and tools in the field of diabetes research with respect to a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and e) Health Care and Management with the first category appearing to be the most popular.

**Keywords:**

* Machine learning
* Data mining
* Diabetes mellitus
* Diabetic complications
* Disease prediction models
* Biomarker(s) identification

1. [Introduction](https://www.sciencedirect.com/science/article/pii/S2001037016300733#s0005)
2. [Machine Learning and Knowledge Discovery](https://www.sciencedirect.com/science/article/pii/S2001037016300733#s0010)
3. [Methods](https://www.sciencedirect.com/science/article/pii/S2001037016300733#s0055)

4.Model building

5.[Conclusions](https://www.sciencedirect.com/science/article/pii/S2001037016300733" \l "s0115" \o "7. Conclusions)

**1.INTRODUCTION:**

Significant advances in biotechnology and more specifically [high-throughput sequencing](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/high-throughput-sequencing) result incessantly in an easy and inexpensive data production, thereby ushering the science of applied biology into the area of big data.

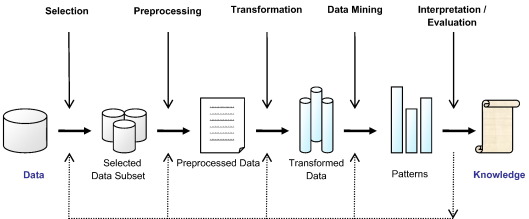
To date, besides high performance sequencing methods, there is a plethora of digital machines and sensors from various research fields generating data, including super-resolution digital microscopy, mass spectrometry, Magnetic Resonance Imagery (MRI), etc. Although these technologies produce a wealth of data, they do not provide any kind of analysis, interpretation or extraction of knowledge. To this end, the area of Biological [Data Mining](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/data-mining) or otherwise Knowledge Discovery in Biological Data, is more than ever necessary and important. The primary objective is to delve into the rapidly accruing body of biological data and set the basis potentiating answers to fundamental questions in biology and medicine.

Applying machine learning and data mining methods in DM research is a key approach to utilizing large volumes of available diabetes-related data for extracting knowledge. The severe social impact of the specific disease renders DM one of the main priorities in medical science research, which inevitably generates huge amounts of data. Undoubtedly, therefore, machine learning and data mining approaches in DM are of great concern when it comes to diagnosis, management and other related clinical administration aspects. Hence, in the framework of this study, efforts were made to review the current literature on machine learning and data mining approaches in diabetes research.

**2.MACHINE LEARNING AND KNOWLEDGE DISCOVERY:**

Machine learning is the scientific field dealing with the ways in which machines learn from experience. For many scientists, the term “machine learning” is identical to the term “artificial intelligence”, given that the possibility of learning is the main characteristic of an entity called intelligent .The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience . A more detailed and formal definition of machine learning is given by Mitchel : A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Knowledge discovery in databases (KDD) is a field encompassing theories, methods and techniques, trying to make sense of data and extract useful knowledge from them. It is considered to be a multistep process (selection, preprocess, transformation, [data mining](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/data-mining), interpretation-evaluation) .The most important step in the entire KDD process is data mining, exemplifying the application of machine learning algorithms in analyzing data. A complete definition of KDD is given by Fayyad et al. : KDD is the nontrivial process identifying valid, novel, potentially useful, and ultimately understandable patterns in data.



***The basic steps of the KDD process.***

### 2.1. Categories of Machine Learning Tasks

Machine learning tasks are typically classified into three broad categories .These are: a) supervised learning, in which the system infers a function from labeled training data, b) unsupervised learning, in which the learning system tries to infer the structure of unlabeled data, and c) reinforcement learning, in which the system interacts with a dynamic environment.

#### 2.1.1. Supervised Learning:

In supervised learning, the system must “learn” inductively a function called target function, which is an expression of a model describing the data. The objective function is used to predict the value of a variable, called dependent variable or output variable, from a set of variables, called independent variables or input variables or characteristics or features. The set of possible input values of the function, i.e. its domain, are called instances. Each case is described by a set of characteristics (attributes or features). A subset of all cases, for which the output variable value is known, is called training data or examples. In order to infer the best target function, the learning system, given a training set, takes into consideration alternative functions, called hypothesis and denoted by h. In supervised learning, there are two kinds of learning tasks: classification and regression. Classification models try to predict distinct classes, such as e.g. blood groups, while regression models predict numerical values. Some of the most common techniques are Decision Trees (DT), Rule Learning, and Instance Based Learning (IBL), such as k-Nearest Neighbors (k-NN), Genetic Algorithms (GA), Artificial Neural Networks (ANN), and Support Vector Machines (SVM).



#### 2.1.2. Unsupervised Learning:

In unsupervised learning, the system tries to discover the hidden structure of data or associations between variables. In that case, training data consists of instances without any corresponding labels.



#### 2.1.3. Reinforcement Learning:

The term Reinforcement Learning is a general term given to a family of techniques, in which the system attempts to learn through direct interaction with the environment so as to maximize some notion of cumulative reward .It is important to mention that the system has no prior knowledge about the behavior of the environment and the only way to find out is through trial and failure (trial and error). Reinforcement learning is mainly applied to autonomous systems, due to its independence in relation to its environment.

### 2.2. Feature Selection

Feature selection is one of the most important processes of the KDD's data transformation step. It is defined as the process of selecting a subset of features from the feature space, which is more relevant to and informative for the construction of a model. The advantages of feature selection are many and relate to different aspects of data analysis, such as better visualization and understanding of data, reduction of computational time and duration of analysis, and better prediction accuracy .

There are two main different approaches in the feature selection process. The first one is to make an independent assessment, based on general characteristics of data. Methods belonging to this approach are called filter methods, because the feature set is filtered out before model construction. The second approach is to use a machine learning algorithm to evaluate different subsets of features and finally select the one with the best performance on classification accuracy. The latter algorithm will be used in the end to build a predictive model. Methods in this category are called wrapper methods, because the arising algorithm wraps the whole feature selection process.

**3.METHODS**

Goal of the project is to investigate for model to predict diabetes with better accuracy, different classification algorithms to predict diabetes.

1. Dataset Description- The data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.

|  |  |
| --- | --- |
| S No. | Attributes |
| 1 | Pregnancy |
| 2 | Glucose |
| 3 | Blood Pressure |
| 4 | Skin thickness |
| 5 | Insulin |
| 6 | BMI(Body Mass Index) |
| 7 | Diabetes Pedigree Function |
| 8 | Age |

Table 1: Dataset Description

* The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.
* Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labeled as 0 means negative means no diabetes and 268 labeled as 1 means positive means diabetic.

1. Data Preprocessing- Data preprocessing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively ths process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform pre processing in two steps.
   1. Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and help to work faster.
   2. Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under same scale.
2. Apply Machine Learning- When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction.
3. The Techniques are follows
4. Support Vector Machine- Support Vector Machine also known as SVM which is a supervised machine learning algorithm. SVM is most popular classification technique. SVM creates a hyper plane that separate two classes. It can create a hyper plane or set of hyper plane in high dimensional space. This hyper plane can be used for classification or regression also. SVM differentiates instances in specific classes and can also classify the entities which are not sup- ported by data. Separation is done through hyper plane performs the separation to the closest training point of any class.

Algorithm:

* + Select the hyper plane which divides the class better.
  + To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.
  + If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to
  + Select the class which has the high margin.
  + Margin = distance to positive point + Distance to negative point.

1. K-Nearest Neighbor – KNN is also a supervised ma- chine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is lazy prediction technique.KNN assumes that similar things are near to each other. Many times data points which are similar are very near to each other.KNN helps to group new work based on similarity measure.KNN algorithm record all the records and classify them according to their similarity measure. For finding the distance between the points uses tree like structure. To make a prediction for a new data point, the algorithm finds the closest data points in the train- ing data set its nearest neighbors.

Here K= Number of nearby neighbors, its always a positive integer. Neighbor’s value is chosen from set of class. Closeness is mainly defined in terms of Euclidean distance.

The Euclidean distance between two points P and Q i.e. P (p1,p2, . ,pn) and Q (q1, q2,..qn) is defined by the following equation:-

Algorithm:

* + Take a sample dataset of columns and rows named as Pima Indian Diabetes data set.
  + Take a test dataset of attributes and rows.
  + Find the Euclidean distance by the help of formula.
  + Then, Decide a random value of K is the no. of nearest neighbors
  + Then with the help of these minimum distance and Euclidean distance find out the nth column of each.
  + Find out the same output values.

If the values are same, then the patient is diabetic, other- wise not.

1. Decision Tree

Decision tree is a basic classification method. It is supervised learning method. Decision tree is used when response variable is categorical. Decision tree has tree like structure based model which describes classification process based on input feature. Input variables are any types like graph, text, discrete, continuous etc.

Steps for Decision Tree are as follows:

Algorithm

* + Construct tree with nodes as input feature.
  + Select feature to predict the output from input fea- ture whose information gain is highest.
  + The highest information gain is calculated for each attribute in each node of tree.
  + Repeat step 2 to form a sub tree using the feature which is not used in above node.

1. Logistic Regression

Logistic regression is also a supervised learning classification algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories. It classifies the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or negative for diabetes. Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variable. Logistic regression is a based on Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and negative class.

Sigmoid function P = 1/1+e – (a+bx).

Here P = probability.

b = parameter of Model.

Ensembling:

Ensembling is a machine learning technique Ensemble means using multiple learning algorithms to- gather for some task. It provides better prediction than any other individual model that’s why it is used. The main cause of error is noise bias and variance, ensemble methods help to reduce or minimize these errors. There are two popular ensemble methods such as Bagging, Boosting, ada-boosting, Gradient boosting, voting, averaging etc. Here In these work we have used Bagging (Random forest) and Gradient boosting ensemble methods for predicting diabetes.

1. Random Forest

It is type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is grater then compared to other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest improve performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

Algorithm

* + The first step is to select the R features from the total features m where R<<M.
  + Among the R features, the node using the best split point.
  + Split the node into sub nodes using the best split.
  + Repeat a to c steps until l number of nodes has been reached.
  + Built forest by repeating steps a to d for a number of times to create n number of trees.

The random forest finds the best split using the Gin-Index Cost Function which is given by:

The first step is to need the take a glance at choices and use the foundations of each indiscriminately created decision tree to predict the result and stores the anticipated outcome at intervals the target place.

Secondly, calculate the votes for each predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula.

**4.MODEL BUILDING**

This is most important phase which includes model build- ing for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology:

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. K- Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

Step5: Build the classifier model for the mentioned ma- chine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned ma- chine learning algorithm based on test set.

Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

Step8: After analyzing based on various measures conclude the best performing algorithm.

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import confusion\_matrix

from sklearn.neighbors import KNeighborsClassifier

from sklearn.svm import SVC

from sklearn.naive\_bayes import GaussianNB

from sklearn.ensemble import RandomForestClassifier

df=pd.read\_csv('diabetes.csv')

print("Shape",df.shape)

df.head()

df.describe()

df['Outcome'].value\_counts()

df.groupby('Outcome').mean()

#seperating the data and labels

X = df.drop(columns ='Outcome',axis=1)

Y= df['Outcome']

print(X)

print(Y)

df.info()

df.corr()

fig = plt.figure(figsize = (19,15))

plt.matshow(df.corr()\*\*2,fignum = fig.number)

plt.xticks(range(df.shape[1]),df.columns,fontsize = 14, rotation = 45)

plt.yticks(range(df.shape[1]),df.columns,fontsize = 14)

cb = plt.colorbar()

cb.ax.tick\_params(labelsize = 14)

plt.title("correlation matrix of squared correlations \n\n\n\n\n",fontsize =20)

#1-->diabetic

#0-->healthy

print(df["Outcome"].value\_counts())

fig = plt.figure(figsize = (10, 6))

#glucose for diabetic

fig = plt.figure(figsize =(16,6))

sns.distplot(df["Glucose"][df["Outcome"] == 1])

plt.xticks([i for i in range(0,201,15)],rotation = 45)

plt.ylabel("Glucose count")

plt.title("Glucose",fontsize = 20)

#insulin for diabetic

fig = plt.figure(figsize = (16,6))

sns.distplot(df["Insulin"][df["Outcome"]==1])

plt.xticks()

plt.title("Insulin",fontsize = 20)

#BMI for diabetic

fig = plt.figure(figsize =(16,6))

sns.distplot(df["BMI"][df["Outcome"]==1])

plt.xticks()

plt.title("BMI",fontsize = 20)

#diabeticpedigreefunction for diabetic

fig = plt.figure(figsize = (16,5))

sns.distplot(df["DiabetesPedigreeFunction"][df["Outcome"] == 1])

plt.xticks([i\*0.15 for i in range(1,12)])

plt.title("diabetespedigreefunction")

#Age for diabetic

fig = plt.figure(figsize = (16,6))

sns.distplot(df["Age"][df["Outcome"] == 1])

plt.xticks([i\*0.15 for i in range(1,12)])

plt.title("Age")

#seperating dependent and independent columns

X = df.drop(["Pregnancies","BloodPressure","SkinThickness"],axis = 1)

y = df.iloc[:,-1]

X\_train,X\_test,y\_train,y\_test = train\_test\_split(X,y,test\_size = 0.2,random\_state=0)

print("X\_train size:", X\_train.shape)

print("y\_train size: ",y\_train.shape,"\n")

print("X\_test size:", X\_test.shape)

print("y\_test size:",y\_test.shape)

#standard scaling

sc= StandardScaler()

X\_train = sc.fit\_transform(X\_train)

X\_test = sc.fit\_transform(X\_test)

print(sc)

print(X\_train)

print(X\_test)

kn\_classifier = KNeighborsClassifier(n\_neighbors =5,metric = 'minkowski',p = 2)

kn\_classifier.fit(X\_train,y\_train)

kn\_y\_pred = kn\_classifier.predict(X\_test)

cm\_kn = confusion\_matrix(y\_test, kn\_y\_pred)

print(cm\_kn)

print("Correct:",sum(kn\_y\_pred == y\_test))

print("Incorrect : ",sum(kn\_y\_pred != y\_test))

print("Accuracy:",sum(kn\_y\_pred ==y\_test)/len(kn\_y\_pred))

svc\_classifier = SVC(kernel ="linear",random\_state = 0)

svc\_classifier.fit(X\_train,y\_train)

svc\_y\_pred = svc\_classifier.predict(X\_test)

svc\_cm = confusion\_matrix(y\_test,svc\_y\_pred)

print(svc\_cm)

print("Correct:",sum(svc\_y\_pred == y\_test))

print("Incorrect : ",sum(svc\_y\_pred != y\_test))

print("Accuracy:",sum(svc\_y\_pred ==y\_test)/len(kn\_y\_pred))

nb\_classifier = GaussianNB()

nb\_classifier.fit(X\_train,y\_train)

nb\_y\_pred =nb\_classifier.predict(X\_test)

nb\_cm = confusion\_matrix(nb\_y\_pred,y\_test)

print(nb\_cm)

print("Correct:",sum(nb\_y\_pred == y\_test))

print("Incorrect : ",sum(nb\_y\_pred != y\_test))

print("Accuracy:",sum(nb\_y\_pred ==y\_test)/len(nb\_y\_pred))

**5.CONCLUSION:**

* The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully.

I **Ms. HARSHINI SA** of St.Joseph’s Institute of technology has completed my project in exposys data labs on data science. I thank exposys data labs for giving this wonderful opportunity.