A PROJECT REPORT ON

**“DIABETES PREDICTION**

**USING**

**DATA SCIENCE”**

SUBMITTED BY :- Sonal Sanjay Shitole

INTERNSHIP PROJECT :- Exposys Data Labs

PROBLEM STATEMENT:-

Diabetes is a type of chronic disease which is more common among the people of all age groups. Predicting this disease at an early stage can help a person to take the necessary precautions and change his/her lifestyle accordingly to either prevent the occurrence of this disease or control the disease (For people who already have the disease).

Task:-

1. Prepare the data-set using several methods to train the model.

2. Build a model which can give high accuracy of predicting the disease.

Abstract:-

The field of medical diagnosis with the help of machine learning is increasing gradually. Medical data sets that contain irrelevant data (noise) are used to train and evaluate the machine learning algorithms.

Diabetes leads to serious complications or even premature death. However, to diagnose diabetes, several time consuming tests and analyzing critical factors are done. Nowadays, machine learning algorithms are used to classify and diagnose diseases, in order to eliminate the problem and reduce the required cost. Besides that, using the machine learning algorithm leads to meaningful and accurate decisions. In this project, a support vector machine [SVM] is applied in diabetes prediction. The performance of the SVM algorithm is analyzed for different available kernels. The best kernel is selected and used for prediction. The proposed approach is implemented in python programming language and its performance is as good as other algorithms.

TABLE OF CONTENTS:-

| Contents | Page no. |
| --- | --- |
| Introduction | 1 |
| Existing Methodology | 1 |
| Proposed Method With Architecture | 2 |
| Methodology | 3 |
| Implementation | 4 |
| Conclusion | 6 |

Introduction:-

Diabetes is becoming a very common disease in India. The annual report of the World Health Association, added up to the number of individuals experiencing diabetes is 422 million the year . Consistently, there is a significant increment in the number of individuals experiencing diabetes in different healing centers. The World Health Organization (WHO) reports on “Diabetes Care 2018” by American Diabetes Association and Standards for Medical Care in Diabetes, a study for correlation between diverse races and their pay.

Machine learning algorithms are mathematical techniques, which are very useful in analyzing large amounts of data and suggesting some actions on the basis of that data. These algorithms are useful in analyzing a data set and predict the values for a new entry. Many researchers are applying ML algorithms for prediction and control of various diseases. The results of machine learning algorithms are found very well in the prediction of different diseases. There is a need to apply machine learning algorithms to explore their power for diabetes prediction so that necessary precautions can be taken to avoid diabetes.

EXISTING METHODOLOGY:-

Health care professionals most often use the fasting plasma glucose (FPG) test or the A1C test to diagnose diabetes. In some cases, they may use a random plasma glucose (RPG) test.

The FPG blood test measures your blood glucose level at a single point in time. For the most reliable results, it is best to have this test in the morning, after you fast for at least 8 hours. Fasting means having nothing to eat or drink except sips of water.

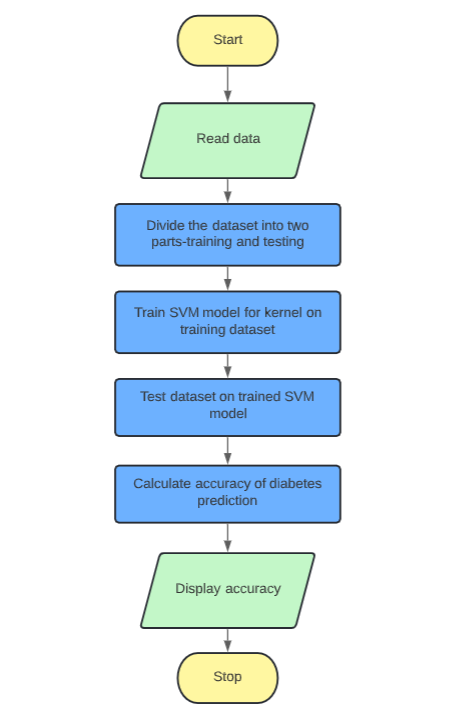
The [A1C test](https://www.niddk.nih.gov/health-information/diagnostic-tests/a1c-test) is a blood test that provides your average levels of blood glucose over the past 3 months. Other names for the A1C test are hemoglobin A1C, HbA1C, glycated hemoglobin, and glycosylated hemoglobin test.

Sometimes healthcare professionals use the RPG test to diagnose diabetes when diabetes symptoms are present and they do not want to wait until you have fasted. You do not need to fast overnight for the RPG test. You may have this blood test at any time.

Proposed Method With Architecture:-

We have applied an SVM machine learning algorithm for diabetes prediction. The SVM algorithm is implemented in the python programming language and tested on a data set. The SVM model is created by using python programming language. The dataset is divided into training parts and testing parts. Then, the SVM model is trained accordingly. The novelty in this work is that the performance of SVM kernels on medical data set for disease prediction is proposed. We compare the performance of SVM kernels.0020 The model is trained on one kernel available for SVM and its prediction accuracies are calculated by testing set. The SVM is tested on one kernels which are Linear. The best SVM kernel is selected and used for diabetes prediction. The proposed model is implemented in Python programming language and tested on a data set of 768 patients. The data set is freely available on Kaggle with the name Pima Indians Diabetes Database for research . The data set is available in the form of a CSV file which is best suited for python programming language. The prediction accuracy of the SVM ML algorithm depends on the selection of its model.

Architecture:-



Methodology:-

The SVM models for classification have been developed for the classification of diabetes dataset. The experiments are conducted on Matlab R2010a. The datasets are stored in MS Excel documents and read directly from Matlab. The diagnostic performance of the developed models is evaluated using the Receiver Operating Characteristic (ROC) curve. In the ROC curve the true positive rate (sensitivity) is plotted in function of the false positive rate for different cut-off points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold.

The Pima Indian diabetes dataset, donated by Vincent Sigillito, is a collection of medical diagnostic reports from 768 records of female patients at least 21 years old of Pima Indian heritage, a population living near Phoenix, Arizona, USA [19]. The binary target variable takes the values “0” or “1” while “1” means a positive test for diabetes, “0” means a negative test. There are 268 cases in class “1” and 500 cases in class “0”.

The significance of the automatically selected set of variables was further manually evaluated by fine tuning parameters. The variables included in the final selection were those with the best discriminative performance. There are eight numeric variables: (1) Number of times pregnant, (2) Plasma glucose concentration a 2h in an oral glucose tolerance test (3) Diastolic blood pressure (mm Hg) (4) Triceps.

Implementation:-

import numpy as np

import pandas as pd

from sklearn.preprocessing import StandardScaler

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

from sklearn import svm

from sklearn.metrics import accuracy\_score

import matplotlib.pyplot as plt

%matplotlib inline

import seaborn as sns

sns.set()

sns.set(style="darkgrid")

diabetes\_dataset = pd.read\_csv('/content/sample\_data/diabetes.csv')

diabetes\_dataset.describe()

diabetes\_dataset['Outcome'].value\_counts()

diabetes\_dataset.groupby('Outcome').mean()

X = diabetes\_dataset.drop(columns = 'Outcome', axis=1)

Y = diabetes\_dataset['Outcome']

print(X)

print(Y)

scaler = StandardScaler()

scaler.fit(X)

standardized\_data = scaler.transform(X)

print(standardized\_data)

X = standardized\_data

Y = diabetes\_dataset['Outcome']

print(X)

print(Y)

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X,Y, test\_size = 0.3, stratify=Y, random\_state=2)

print(X.shape, X\_train.shape, X\_test.shape)

classifier = svm.SVC(kernel='linear')

classifier.fit(X\_train, Y\_train)

X\_train\_prediction = classifier.predict(X\_train)

training\_data\_accuracy = accuracy\_score(X\_train\_prediction, Y\_train)

print('Accuracy score of the training data : ', training\_data\_accuracy)

X\_test\_prediction = classifier.predict(X\_test)

test\_data\_accuracy = accuracy\_score(X\_test\_prediction, Y\_test)

print('Accuracy score of the test data : ', test\_data\_accuracy)

input\_data = (5,166,72,19,175,25.8,0.587,51)

# changing the input\_data to numpy array

input\_data\_as\_numpy\_array = np.asarray(input\_data)

# reshape the array as we are predicting for one instance

input\_data\_reshaped = input\_data\_as\_numpy\_array.reshape(1,-1)

# standardize the input data

std\_data = scaler.transform(input\_data\_reshaped)

print(std\_data)

prediction = classifier.predict(std\_data)

print(prediction)

if (prediction[0] == 0):

print('The person is not diabetic')

else:

print('The person is diabetic')

# Distribution of Each Feature: Histogram

colnames=diabetes\_dataset.columns

fig, ax=plt.subplots(nrows=2, ncols=4, figsize=(15,12), sharey=True)

for i in range(4):

x = colnames[i]

ax[0,i].hist(diabetes\_dataset[str(x)], color="blue", bins=20)

ax[0,i].set\_ylabel(str(x))

ax[0,i].set\_title(str(x) + "\nHistogram")

for i in range(4,8):

x = colnames[i]

ax[1,i-4].hist(diabetes\_dataset[str(x)], color="green", bins=20)

ax[1,i-4].set\_ylabel(str(x))

ax[1,i-4].set\_title(str(x) + "\nHistogram")

plt.subplots\_adjust(left=0.1,

bottom=0.1,

right=0.9,

top=0.9,

wspace=0.3,

hspace=0.4)

Output:-

Accuracy score of the training data : 0.7821229050279329

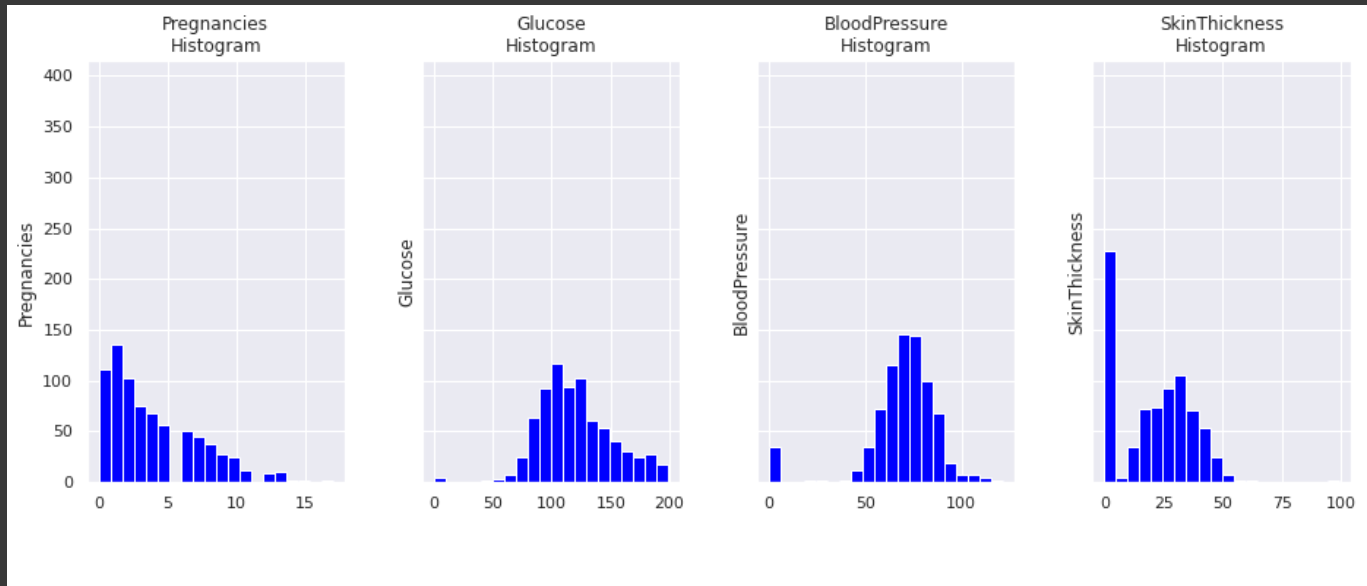
Accuracy score of the test data : 0.7748917748917749

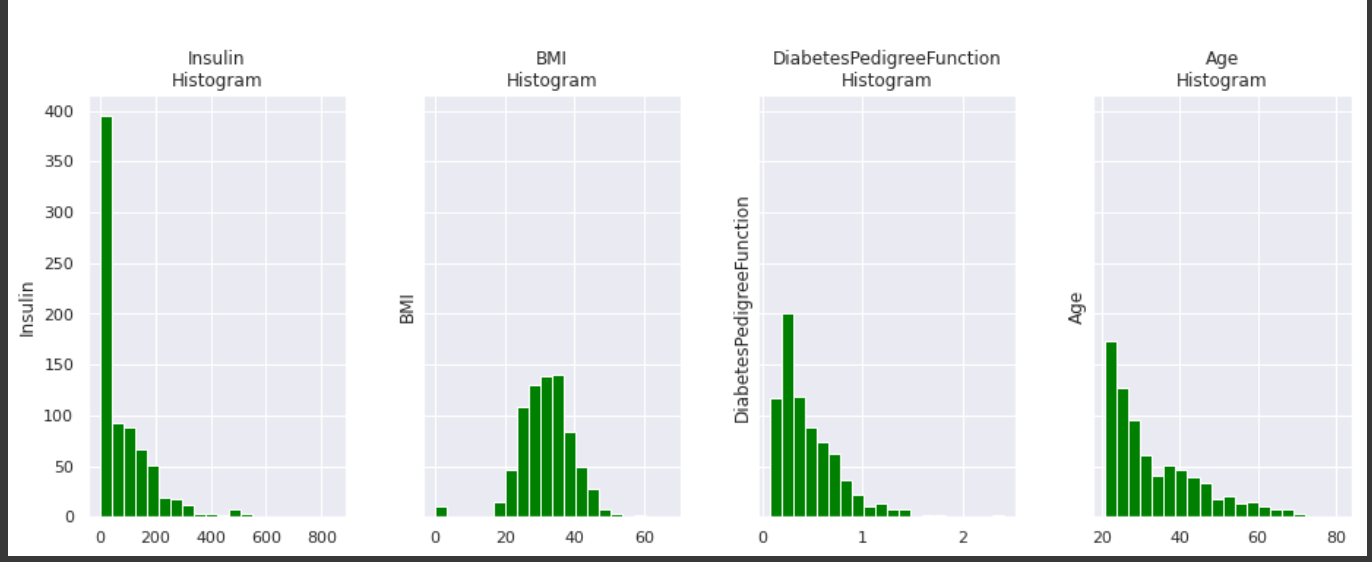
[[ 0.3429808 1.41167241 0.14964075 -0.09637905 0.82661621 -0.78595734

0.34768723 1.51108316]]

[1]

The person is diabetic





Conclusion:-

In this project, kernels of the SVM are used for prediction and their prediction accuracy is measured. It is found that the kernel best performs for the diabetes prediction of Indian patients as its prediction accuracy is found. In future the SVM kernel can be tested in prediction of other diseases such as Cancer, Thyroid etc. Further the work can be extended to compare the performance of SVD on other performance metrics such as precision, recall and F-measure and can be compared with other existing techniques. In future other machine learning and Deep learning algorithms can also be applied for diabetes prediction and their performance can be compared with SVD.