Heaven's light is our guide

University Of Batna 2, Algeria

Faculty Of Mathematics And Computer Science
Department Of Computer Science
speciality : artificial intelligence and multimedia
module : seminar

Master 2 thesis report on
Predict Diabetes Using Machine Learning

Supervised by

Mohamed-Rida Abdessemed

Submitted by

Refis Fouad Chaouki

november 2023

Acknowledgment

all praises are for allah almighty the most gracious and most merciful who gave us the strength. then i would like to express my special thanks of gratitude to my teachers for their able guidance and to my parents for their support.

Abstract

Diabetes mellitus is a common chronic disease, the earlier we detected it much more the it will be possible to stop the progression of the disease and prevent further complications. So this work shows a model that can predict if a patient will develop the diabetes disease or not, based on significant attributes and the relationship between them using statistical methods and machine learning technics and data gathered from UCI repository which is named as ”pima indian“ diabetes Data-set we use different algorithm to build the model like (KNN) with training accuracy :0.79 and test accuracy: 0.78, decision tree it’s training accuracy 1 and test accuracy 0.714, for the (MLP) we get an accuracy test :0.823 and accuracy test: 0.802. so the Results shows that the model is capable of predicting the diabetes mellitus and the best model of them is the (MLP) model.

Key words: Diabetes, Machine-learning, Dataset, Prediction, KNN, MLP, Model, Health-care.

List of contents

List of Figures

Abbreviations and List of Symbols

KNN	K-Nearest Neighbours
DT	Decision Tree
MLP	multi-layer perceptron
WHO	world health organisation
GDM	Gestational diabetes mellitus
LR	Decision Tree
SVC	super vector classifier
RF	Random forest

Chapter 1

Introduction

More than 500 million people are living with diabetes worldwide, it's a chronic disease affecting people of all ages in every country, this number is expected to reach 1.3 billion people by 2030[1]. These terrifying numbers and the fact that half of people who have diabetes don't know it; "according to PAHO (pan American health organization) [2] in Americas 10 percent of the adult population -nearly 63 milion people – have diabetes, but an estimated 44 percent of them don't know it" leads us to think about ways to implement the new approaches of artificial intelligence and its sub-fields to build Health-Care apps and models to help both doctors and patients. Diagnosis of diabetes is considered a challenging problem for quantitative research. Based on the previous fact that nearly half of the people that have diabetes don't know that they actualy have it, and the other fact that Early diabetes detection can prevent serious complications [3] we must discover new approaches to help both doctors and patients to diagnose the appearance of diabetes earlier. So we can see that AI or precisely ML can or will provide a great tool to build predictive mosels to diagnose the appearance of diabetes in much more effective ways so :

1.1 Objectives

research objecives is answering those two questions :

- How might we apply machine learning techniques for the detection of diabetes mellitus?
- By what means might we compare the proposed machine learning model for detection diagnosis diabetes mellitus?

Chapter 2

Diabetes

according to the WHO (world health organization) Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves. [4]

2.1 Types Of Diabetes

diabetes mellitus classified into three categories:[5]

- 1. Type 1 diabetes mellitus is an autoimmune disease contributing to approximately 5 percent of diabetic cases with a high prevalence in adolescents. It is majorly caused due to destruction of pancreatic islet cells via humoral response and T-cell mediated inflammatory response. The presence of autoantibodies such as GAD65 glutamic acid decarboxylase, autoantibodies to insulin, IA2 and IA2β protein tyrosine phosphatase and ZnT8A zinc transporter protein against the pancreatic β cells is the diagnostic of this disease.
- 2. Type 2 diabetes, also termed as non-insulin-dependent condition that majorly affects the adults (20–79 yr. old), contributing to 95 percent of prevailing diabetes cases, is associated with obesity and insulin resistance that leads to decreased insulin production overtime.
- 3. Gestational diabetes mellitus is observed in 7 percent of all pregnant women worldwide. It often occurs due to hormonal imbalances that occur during pregnancy, leading to insulin resistance that subsides after pregnancy (American Diabetes Association 2015). Common risk factors for GDM are age, obesity, family history and susceptible race.

2.2 Diagnosis Of Diabetes Mellitus

2.2.1 signs and symptoms of diabetes

Anyone can develop diabetes. That's why it's important to get tested if you notice any signs or symptoms. These aren't always easy to spot. In fact type 2 diabetes is often diagnosed through blood tests for other conditions or health issues instead. The signs are [6] :

- Urinate (pee) a lot, often at night
- Are very thirsty.
- Lose weight without trying.
- Have blurry vision.
- Have numb or tingling hands or feet.
- Feel very tired.
- Have sores that heal slowly.
- Have more infections than usual.

2.2.2 Diabetes tests

- A1C Test : The A1C test measures your average blood sugar level over the past 2 or 3 months. An A1C below 5.7 percent is normal, between 5.7 and 6.4 percent indicates you have prediabetes, and 6.5 percent or higher indicates you have diabetes.
- Fasting Blood Sugar Test : This measures your blood sugar after an overnight fast (not eating). A fasting blood sugar level of 99 mg/dL or lower is normal, 100 to 125 mg/dL indicates you have prediabetes, and 126 mg/dL or higher indicates you have diabetes.
- Glucose Tolerance Test : This measures your blood sugar before and after you drink a liquid that contains glucose. You'll fast (not eat) overnight before the test and have your blood drawn to determine your fasting blood sugar level. Then you'll drink the liquid and have your blood sugar level checked 1 hour, 2 hours, and possibly 3 hours afterward. At 2 hours, a blood sugar level of 140 mg/dL or lower is considered normal, 140 to 199 mg/dL indicates you have prediabetes, and 200 mg/dL or higher indicates you have diabetes.
- Random Blood Sugar Test : This measures your blood sugar at the time you're tested. You can take this test at any time and don't need to fast (not eat) first. A blood sugar level of 200 mg/dL or higher indicates you have diabetes.[7]

Result*	A1C Test	Fasting Blood Sugar Test	Glucose Tolerance Test	Random Blood Sugar Test
Diabetes	6.5% or above	126 mg/dL or above	200 mg/dL or above	200 mg/dL or above
Prediabetes	5.7 – 6.4%	100 – 125 mg/dL	140 – 199 mg/dL	N/A
Normal	Below 5.7%	99 mg/dL or below	140 mg/dL or below	N/A

Figure 2.1: diabetes tests

2.2.3 Predictive Factors of diabetes

There are so many factors that we can predict the diabetes using it but we gonna choose just a few of them based on the dataset we have [8] :

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

Figure 2.2: diabetes attributes

2.3 prediction

aprediction is the outcome you would observe if your hypothesis were correct. Predictions are often written in the form of “if, and, then” statements, as in, “if my hypothesis is true, and I were to do this test, then this is what I will observe.”

We have some data X1,X2,,,,,Xp we call them preictors and Y is called a responce and we need to know the relationship between them, this relationship appears as a pattern that can be described with a mathematical function(s),

2.4 Predictive Model

Predictive modeling is a mathematical process used to predict future events or outcomes by analyzing patterns in a given set of input data. It is a crucial component ofpredictive analytics, a type of data analytics which uses current and historical data to forecast activity, behavior and trends.

The difference between statistics and machine learning is that the former involves testing hypotheses, whereas the latter involves the task of building knowledge and storing it in some form in the computer. The knowledge can be stored in the form of mathematical models, algorithms, or anything that can assist in determining patterns or predicting outcomes.

Chapter 3

AI and healthcare systems

3.1 The Role Of AI In Healthcare

AI plays a very important part in personalised medicine, drug discovery and development and gene editing therapies. It acts an interface between clinical image flow and archived image data, which does not need application-specific designing to utilise it. AI-based disease diagnostic systems expedite decision-making, reduce rate of false positives and therefore provide improved accuracy in the detection of diverse diseases. [9]

3.2 disease diagnostic AI model:

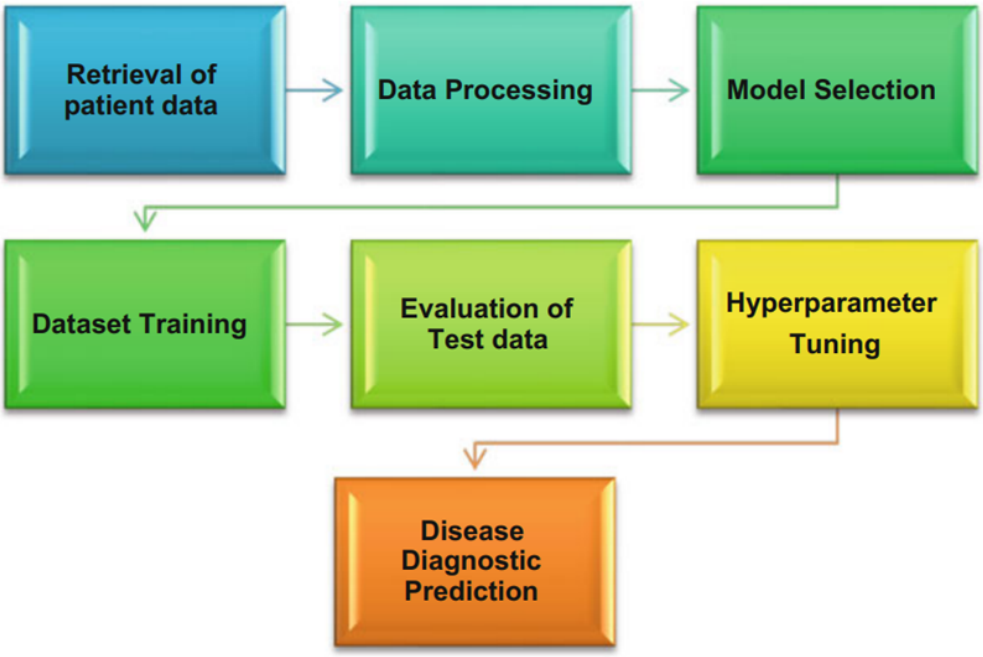

Figure 3.1: build a predictive model step by step

Chapter 4

Machine Learning

Machine learning is a branch of artificial intelligence. Using compuring, we design systems that can learn from data in a manner of beign trained. The systems might learn and improve with experience, and with time, refine a model that can be used to predict outcumes of equations based on previous learning.[10]

4.1 Types Of Machine Learning

4.1.1 Supervised learning

- The dataset is a pool of labelled examples.
- The goal of a supervised learning is to use the dataset to produce a model that takes a feature vector x as input and output information that allows deducing the label for this feature vector.
- It majorly solves classification and regression problems.
- Decision trees, random forest, k-nearest neighbours and logistic regression are the examples of supervised machine learning algorithms.

4.1.2 Unsupervised learning

- The dataset is a pool of labelled examples.
- The goal of a supervised learning is to use the dataset to produce a model that takes a feature vector x as input and output information that allows deducing the label for this feature vector.
- It majorly solves classification and regression problems.
- Decision trees, random forest, k-nearest neighbours and logistic regression are the examples of supervised machine learning algorithms.

4.1.3 Semi-Supervised Learning

- This combination will contain a very small amount of labelled data and a very large amount of unlabelled data.
- The goal of a semi-supervised learning algorithm is to improve supervised learning algorithm by using unlabelled data.
- It can solve problems of classification, regression, clustering and association.

4.1.4 Reinforcement Learning

- The machine is thriving in an environment where it recognises the state of that particular environment as feature vector in data.
- Each action brings different kind of rewards and can transfer the agent to another state (Sutton 1992).
- The goal of reinforcement learning is to make the system learn a policy.
- The policy is a function of the feature vector of a state that is considered as an input, and the outputs are an optimal action to implement in that state.
- If an action is ideal, it maximises the anticipated average reward.

4.2 Role of machine learning in diabetes melitus management

There are several applications of diabetes based on machine learning like : -Insuline controller : an automated artificial pancareatic system improves the efficiency of glucose monitoring.

- Detection of hypo/hyper-glycaemia : he identification of hypoglycaemia and hyperglycaemia is considered as a characteristic classification problem. For a given set of input factors, the model should identify the occurrence of a hypoglycaemic or hyperglycaemic condition.

-Data-Based Prediction Of Plasma Glucose Levels : data-based prediction of plasma glucose levels is categorised as a non-linear regression problem with input factors such as medications, dietary intake, physical activity, anxiety, etc.[11]

4.3 methodology for development of a prediction model

4.3.1 DataSet

The dataset used in this study has been originally obtained by the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to find whether a patient has

diabetes or not, given certain values for different parameters. All the patients considered in this dataset are females above 21 years old. There are 768 instances available in this dataset. The independent parameters for this dataset are number of times the patient was pregnant, plasma glucose concentration level, diastolic blood pressure, triceps skinfold thickness, serum insulin in 2 h, body mass index, diabetes pedigree and age of the patient discussed in Table.[12]

4.3.2 Data preprocessing

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.[12]

4.3.3 Model Construction

Five different classification models can bee created to see which model performs best. These classifiers are logistic regression (LR), support vector classifier (SVC), k-nearest neighbour (KNN) classifier, decision tree (DT) and random forest (RF). but here we use only three (KNN,DT,MLP) The parameters for all the models were declared For KNN, the best result was observed when the number of neighbours was set to 5. For DT when we use pre-pruning and max depth of 3 to avoid the overfitting, for MLP when we use 1000 iteration and augment the alpha rate and modifie the weights.

4.3.4 Results

For the analysis of the performance of our models, This means that the instances were randomly divided into ten parts, where one part would be treated as the testing data, while the remaining nine parts would be treated as the training data. This process would be repeated ten times where each partition experiences a chance to be the testing data. The average of all the metrics such as accuracy, sensitivity, specificity and F1 score is taken to showcase the performance of our models. k-nearest neighbours, decision tree and MLP were the 3 models employed in this classification studies for the automated prediction of diabetes mellitus based on the Pima Indians dataset.

4.4 a brief illustration to ML algorithms

4.4.1 KNN

K-nearest neighbors (KNN) is a type of supervised learning algorithm used for both regression and classification. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closet to the test data. The KNN algorithm calculates the probability of the test data belonging to the classes of ‘K’ training data and class holds the highest probability

will be selected. In the case of regression, the value is the mean of the 'K' selected training points.[13]

4.4.2 DT

decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.[14]

4.4.3 MLP

Multi layer perceptron (MLP) is a supplement of feed forward neural network. It consists of three types of layers—the input layer, output layer and hidden layer. The input layer receives the input signal to be processed. The required task such as prediction and classification is performed by the output layer. An arbitrary number of hidden layers that are placed in between the input and output layer are the true computational engine of the MLP. Similar to a feed forward network in a MLP the data flows in the forward direction from input to output layer. The neurons in the MLP are trained with the back propagation learning algorithm. MLPs are designed to approximate any continuous function and can solve problems which are not linearly separable. The major use cases of MLP are pattern classification, recognition, prediction and approximation.[15]

Chapter 5

predict diabetes using ML

5.1 steps of creating the model

5.1.1 dealing with data

first we import data and some of the important python libraries that we need and displaying data:

```
import pandas as pd      #a Python library used for data analysis.
import numpy as np       #a Python library used for working with arrays.
import matplotlib.pyplot as plt  #a library for creating static, animated, and interactive visualiza
import seaborn as sns
from sklearn import svm
%matplotlib inline
diabetes = pd.read_csv('Desktop/diabtese/diabetes.csv') #import diabetes data base.
print(diabetes.columns) #print the data base columns.
```

```
Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
      dtype='object')
```

```
diabetes.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 5.1: display data

```
print(diabetes.groupby('Outcome').size())
```

```
Outcome
0    500
1    268
dtype: int64
```

```
sns.countplot(data=diabetes, x="Outcome", hue="Outcome")
```

Figure 5.2: print the size of the data base and it's 'outcome' classification

```
sns.countplot(data=diabetes, x="Outcome", hue="Outcome")
```

```
<Axes: xlabel='Outcome', ylabel='count'>
```

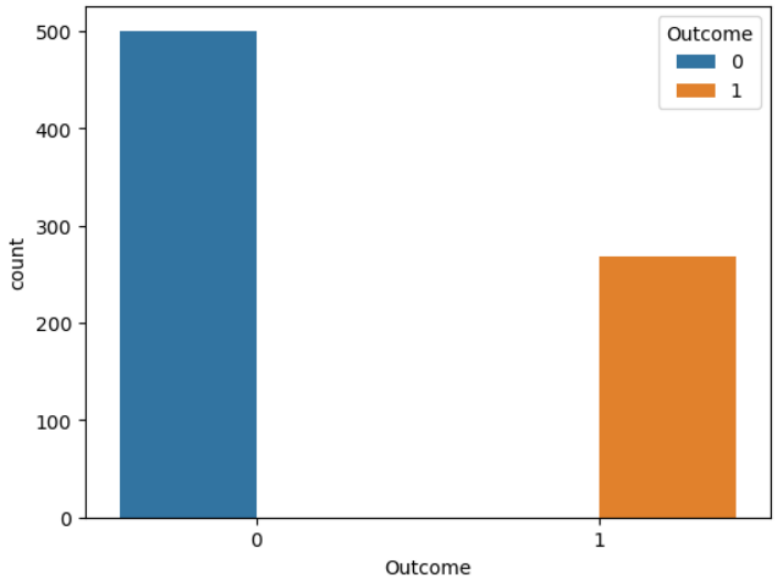

Figure 5.3: then we use bar charts to represent the data :

```
diabetes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Figure 5.4: get informations of the Data-frame

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(diabetes.loc[:, diabetes.columns != 'Outcome'],
from sklearn.neighbors import KNeighborsClassifier
training_accuracy = []
test_accuracy = []
```

Figure 5.5: train the model whith KNN algorithm

```
# try n neighbors from 1 to 10
neighbors_settings = range(1, 11)
for n_neighbors in neighbors_settings:
    # build the model
    knn = KNeighborsClassifier(n_neighbors=n_neighbors)
    knn.fit(X_train, y_train)
    # record training set accuracy
    training_accuracy.append(knn.score(X_train, y_train))
    # record test set accuracy
    test_accuracy.append(knn.score(X_test, y_test))
plt.plot(neighbors_settings, training_accuracy, label="training
accuracy")
plt.plot(neighbors_settings, test_accuracy, label="test accuracy")
plt.ylabel("Accuracy")
plt.xlabel("n_neighbors")
plt.legend()
```

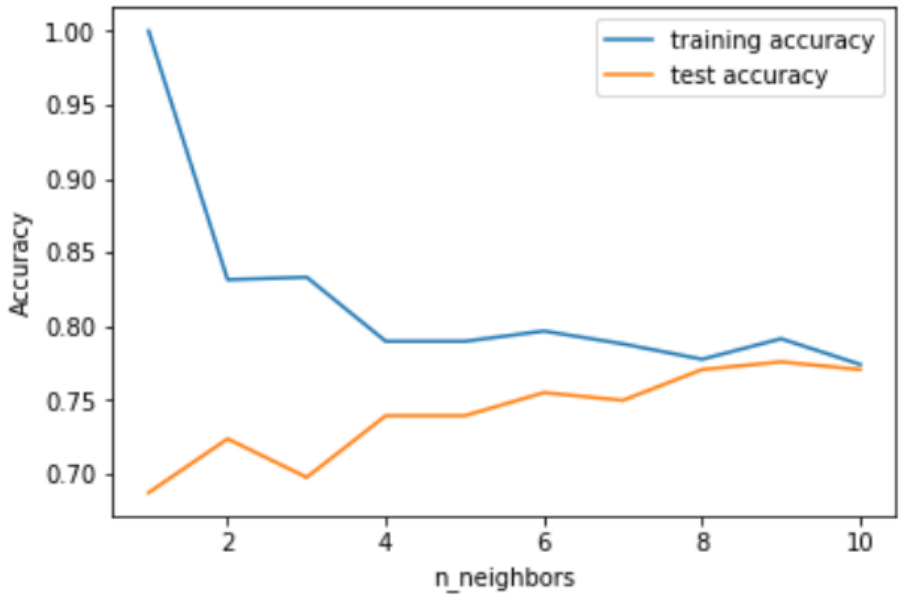

Figure 5.6: train and test accuracy of the model

```
from sklearn.tree import DecisionTreeClassifier
tree = DecisionTreeClassifier(random_state=0)
tree.fit(X_train, y_train)
print("Accuracy on training set: {:.3f}".format(tree.score(X_train, y_train)))
print("Accuracy on test set: {:.3f}".format(tree.score(X_test, y_test)))
```

```
Accuracy on training set: 1.000
Accuracy on test set: 0.714
```

```
tree = DecisionTreeClassifier(max_depth=3, random_state=0)
tree.fit(X_train, y_train)
print("Accuracy on training set: {:.3f}".format(tree.score(X_train, y_train)))
print("Accuracy on test set: {:.3f}".format(tree.score(X_test, y_test)))
```

```
Accuracy on training set: 0.773
Accuracy on test set: 0.740
```

Figure 5.7: when we train the model with DT algorithm we get into an overfiting problem so we need to apply the pre-pruning on the tree

```
print("Feature importances:\n{}".format(tree.feature_importances_))
```

```
Feature importances: [ 0.04554275 0.6830362 0. 0. 0. 0.27142106 0. 0. ]
```

```
def plot_feature_importances_diabetes(model):
    plt.figure(figsize=(8,6))
    n_features = 8
    plt.barh(range(n_features), model.feature_importances_,
align='center')
    plt.yticks(np.arange(n_features), diabetes_features)
    plt.xlabel("Feature importance")
    plt.ylabel("Feature")
    plt.ylim(-1, n_features)
plot_feature_importances_diabetes(tree)
```

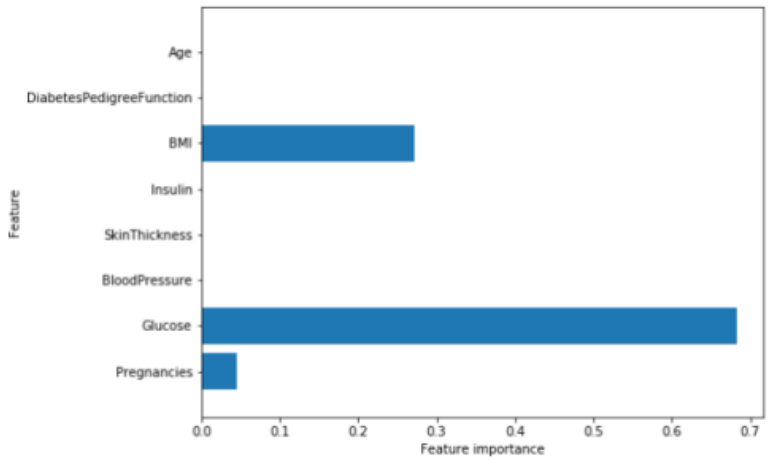

Figure 5.8: features imprtance

```
from sklearn.neural_network import MLPClassifier
mlp = MLPClassifier(random_state=42)
mlp.fit(X_train, y_train)
print("Accuracy on training set: {:.2f}".format(mlp.score(X_train, y_train)))
print("Accuracy on test set: {:.2f}".format(mlp.score(X_test, y_test)))
```

```
Accuracy on training set: 0.73
Accuracy on test set: 0.72
```

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.fit_transform(X_test)
mlp = MLPClassifier(random_state=0)
mlp.fit(X_train_scaled, y_train)
print("Accuracy on training set: {:.3f}".format(mlp.score(X_train_scaled, y_train)))
print("Accuracy on test set: {:.3f}".format(mlp.score(X_test_scaled, y_test)))
```

```
Accuracy on training set: 0.823
Accuracy on test set: 0.802
```

Figure 5.9: train the model using MLP leads first to a weak model than we optimized it We increase the number of repetitions and the weights

```
mlp = MLPClassifier(max_iter=1000, alpha=1, random_state=0)
mlp.fit(X_train_scaled, y_train)
print("Accuracy on training set: {:.3f}".format(mlp.score(X_train_scaled, y_train)))
print("Accuracy on test set: {:.3f}".format(mlp.score(X_test_scaled, y_test)))
```

```
Accuracy on training set: 0.806
Accuracy on test set: 0.797
```

```
plt.figure(figsize=(20, 5))
plt.imshow(mlp.coefs_[0], interpolation='none', cmap='viridis')
plt.yticks(range(8))
plt.xlabel("Columns in weight matrix")
plt.ylabel("Input feature")
plt.colorbar()
```

```
<matplotlib.colorbar.Colorbar at 0x1c921cb6f50>
```

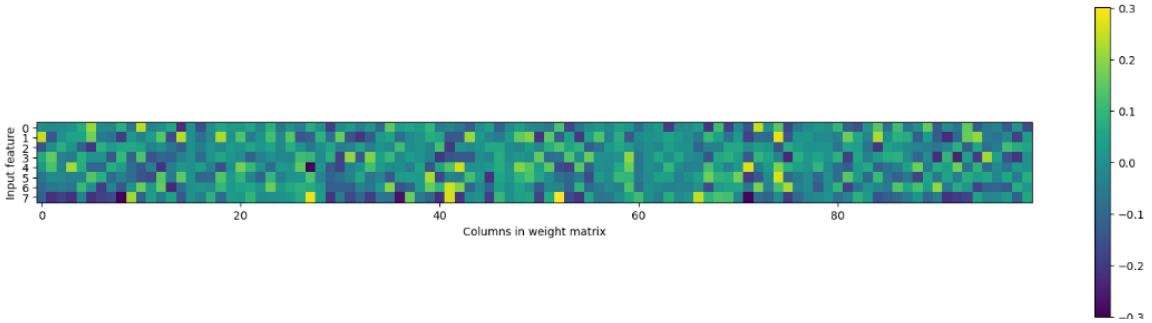

Figure 5.10: build a heat map

```
input_data = (5,166,72,19,175,25.8,0.587,51)
input_data_as_numpy_array = np.asarray(input_data)
input_data_reshaped=input_data_as_numpy_array.reshape(1,-1)
std_data = scaler.transform(input_data_reshaped)
print(std_data)
prediction= mlp.predict(std_data)
print(prediction)

if (prediction[0] == 0):
   print ("no diabetes")
else:
   print("diabetes")
```

```
[[ 0.25771615  1.27714848  0.12678478 -0.12734202  0.64299205 -0.93399952
   0.46135351  1.48820037]]
[1]
diabetes
```

Figure 5.11: test the model with a new input

Chapter 6

Conclusion

This papers succeed to answer the questions of the research showing in the introduction, it shows the details of how we can apply machine learning to help both doctors and patients to predict the diabetes mellitus. we built 3 predictive models using KNN, DT, MLP and the best results appears according to train and test accuracy that have been calculated in the code source is the MLP or the neural network model.

References

[1] K. L. Ong, L. K. Stafford, S. A. McLaughlin, E. J. Boyko, S. E. Vollset, A. E. Smith, B. E. Dalton, J. Duprey, J. A. Cruz, H. Hagins, *et al.*, "Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the global burden of disease study 2021," *The Lancet*, 2023.

[2] pan american health organisation, "Half of people who have diabetes don't know it," 2013.

[3] M. H. Care, "Early diabetes detection can prevent serious complications."

[4] W. H. Organisation, "Diabetes,"

[5] A. Saxena and S. Chandra, *Artificial intelligence and machine learning in healthcare*. Springer, 2021.

[6] C. for Disease Control and Prevention, "Diabetes symptoms."

[7] C. for Disease Control and Prevention, "Diabetes tests."

[8] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima indians diabetes mellitus classification based on machine learning (ml) algorithms," *Neural Computing and Applications*, vol. 35, no. 22, pp. 16157–16173, 2023.

[9] A. Saxena and S. Chandra, *Artificial intelligence and machine learning in healthcare*. Springer, 2021.

[10] J. Bell, *Machine learning: hands-on for developers and technical professionals*. John Wiley & Sons, 2020.

[11] A. Saxena and S. Chandra, *Artificial intelligence and machine learning in healthcare*. Springer, 2021.

[12] A. Saxena and S. Chandra, *Artificial intelligence and machine learning in healthcare*. Springer, 2021.

[13] Medium, "k-nearest neighbours."

[14] IBM, "decision trees."

[15] S. Direct, “multilayer perceptron.”