DIABETES PREDICTION USING

MACHINE LEARNING TECHNIQUES

IMMARAJU SAMUEL samuelimmaraju@gmail.com

Abstract

Diabetes Mellitus 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 take necessary precautions and change his lifestyle to preserve the occurrence of disease or control of disease. The rapid development of machine learning, plays a significant role in healthcare industries. Using machine learning we can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. For this problem, we use Sylhet Diabetes Hospital in Sylhet, Bangladesh dataset. This dataset has 16 attributes. Age, Gender, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital Thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial Paresis, Muscle stiffness, Alopecia, Obesity. We train various machine learning classifiers such as logistic regression, decision tree, etc., and build models on this dataset and obtain the accuracy of each of the classifiers. If accuracy is low, we attempt to increase the accuracy. Next, we deploy the model with highest accuracy, and use it for Realtime prediction of Diabetes taking symptom inputs from the user.

TABLE OF CONTENTS

I. INTRODUCTION

One of the major health problems prevalent in these days irrespective of age is Diabetes Mellitus. It is the surfeit rise of sugar level in blood and this occurs when pancreas does not produce enough insulin, or the body cannot use the produced insulin effectively. And this is the root cause of many other health diseases like, Diabetic peripheral neuropathy, Diabetic retinopathy, Diabetic nephropathy, coronary heart diseases and so on. Diabetes is classified as,

- **Type-1 Insulin-Dependent Diabetes Mellitus (IDDM)** is the Inability of human's body to generate sufficient insulin for which insulin has to be injected into the patient.
- **Type-2 Non-Insulin-Dependent Diabetes Mellitus (NIDDM)** is when the body cells are not able to use the insulin produced properly.
- **Type-3 Gestational Diabetes** is due to insulin-blocking hormones produced during pregnancy.

Healthcare sectors have large volume databases. Such databases may contain structured, semi-structured or unstructured data. Data mining, Machine Learning is the process which analyses huge data sets and reveals hidden information, hidden patterns to discover knowledge from the given data. Machine Learning is a method that is used to train computers or machines explicitly. Machine learning algorithms are classified into three categories:

- **Supervised Learning or Predictive Models** is teaching or training the machine using data that is well labelled. Here some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples (data) so that the supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labelled data. Supervised learning is classified into two categories
    - Classification: A classification problem is when the output variable is a category, such as "disease" and "no disease" or "yes" and "no". E.g., Naive Bayes, Decision Trees, K-Nearest Neighbours, Support Vector Machine
    - Regression: A regression problem is when the output variable is a real value, such as "dollars" or "weight". E.g., Logistic Regression, Linear Regression
- **Unsupervised Learning or Descriptive Models** is training a machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. The main task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data. Unlike supervised learning, there is no teacher which means no training will be given to the machine. The machine has to find the hidden structure in unlabelled data by itself.

    - Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour. E.g., Hierarchical clustering, K-means clustering
    - Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.
- **Semi-supervised Learning** is a technique when we are dealing with a data which is a little bit labelled and rest large portion of it is unlabelled. We use unsupervised techniques to predict labels and then feed these labels to supervised techniques. This technique is mostly applicable in case of image data-sets where usually all images are not labelled.
- **Reinforcement Learning** is a technique where the model's performance keeps on increasing its performance using a Reward Feedback to learn the behaviour or pattern. These algorithms are specific to a particular problem e.g., Google Self Driving car, AlphaGo where a bot competes with human and even itself to getting better and better performer of Go Game. Each time we feed in data, they learn and add the data to its knowledge that is training data. So, more it learns the better it gets trained and hence experienced.

II. EXISTING SYSTEM

For the detection of Diabetes Mellitus traditional lab tests such as fasting blood glucose and post meals glucose, oral glucose tolerance etc., tests are conducted where the patient has to go to the diagnostic center. In the recent times, glucometers are being used at hospitals and even at home to check glucose levels at that particular time. But there are a large number of people who get tested for the first time just to know if they have diabetes or not, because of the symptoms that occur in them so that they adapt their lifestyle accordingly. Even for these types of people we require taking blood samples through the traditional way which is not actually needed as they might be yet in the initial state. For such ones to go to the hospital is time consuming and also not cost effective. Also, there are few machine learning models whose classification and prediction accuracy is not so high, also which are not user friendly as they ask complex inputs from the user in predicting this first time diabetes. So, the aim of this project is to build an efficient instant diabetes prediction system by taking symptoms a person has as inputs with high accuracy.

III. METHODOLOGY

In this approach we divide the entire project into 5 modules. They are Data Collection, Data Pre-processing, Building models, Evaluating the models, Deploying the model with highest accuracy.

1. Dataset Collection

We acquire the dataset containing the sign and symptoms of newly diabetic or would be diabetic patients from UCI ML Repository. This dataset has been made by interviewing the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and after being approved by its doctor. This Diabetes dataset contains 520 records and 16 attributes. Data description is given below

Sno	Attributes	Input Type
1	Age	int
2	Gender	Yes/no
3	Polyuria	Yes/no
4	Polydipsia	Yes/no
5	Sudden weight loss	Yes/no
6	Weakness	Yes/no
7	Polyphagia	Yes/no
8	Genital Thrush	Yes/no
9	Visual blurring	Yes/no
10	Itching	Yes/no
11	Irritability	Yes/no
12	Delayed healing	Yes/no
13	Partial Paresis	Yes/no
14	Muscle stiffness	Yes/no
15	Alopecia	Yes/no
16	Obesity	Yes/no
-	Class Label	Positive/Negative

2. Data Pre-processing

Before Pre-processing the data, we first visualize the data as an intermediatory step. We try to get the insights of the data and its distribution. So, we obtain the class distribution for every feature i.e., symptoms.

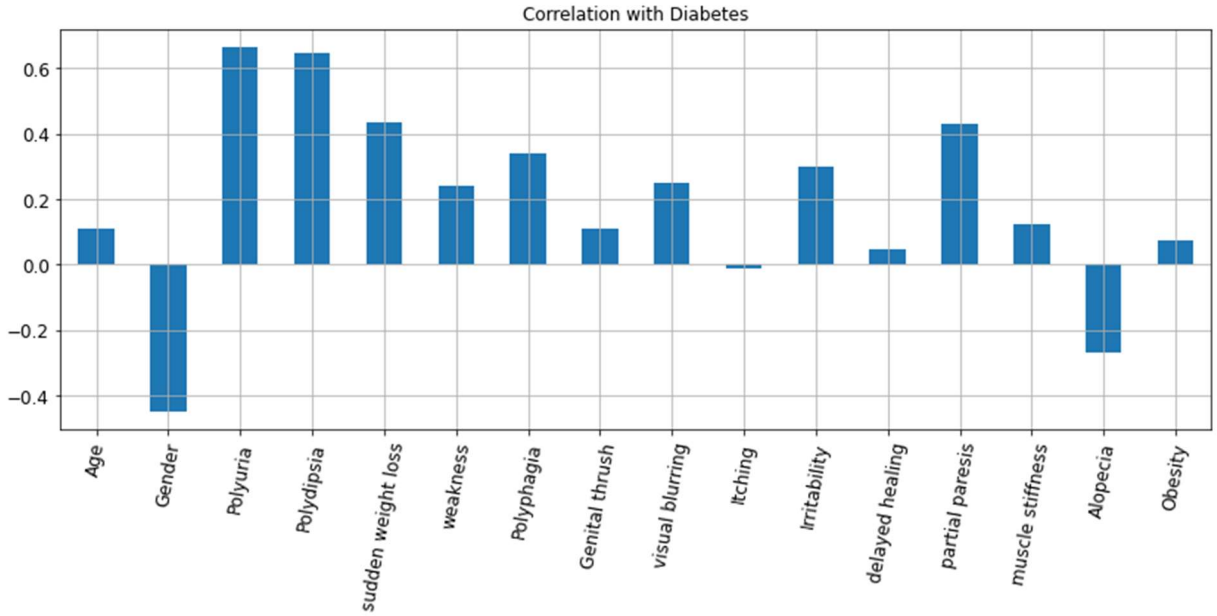


Fig:

fig: Correlation of features with diabetes

In the data pre-processing step, we deal with handling inconsistent data like missing values, null values etc., in order to get more accurate and precise results. And next it is checked for imbalanced data. The feature values consist of symptomatic values i.e. Yes/No, Male/Female and class label as Positive/negative. But many classifiers take only numeric inputs so, we need to convert all of them into numeric values. So, we encode these Yes, Male, Positive as 1 and No, Female, Negative as 0. As all the inputs are converted into 0&1 including the class label, we have to scale the age values also between 0&1. Hence, even though there are many Scaling techniques like Robust Scalar, Normalize etc., we use only the MinMax Scalar.

3. Model Building

This is most important phase which includes model building for prediction of diabetes. In this we implement various machine learning classifiers for diabetes prediction. These algorithms include Logistic Regression, K-Nearest Neighbour, Gaussian Naive Bayes, Decision Tree Classifier. The Pre-processed data is split into Training and Training data using Stratify approach which is very import for even distribution of classes. We pass the training dataset to the classifier to train the model and next we pass the testing dataset to evaluate the model.

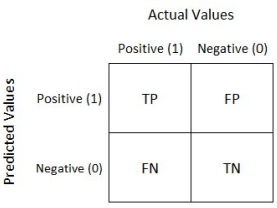
4. Evaluation

We evaluate the models using various evaluation metrics like classification accuracy, confusion matrix, recall, precision and f1-score,

Classification Accuracy - It is the ratio of number of correct predictions to the total number of input samples. It is given as

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Confusion Matrix - It gives us a matrix as output and describes the complete performance of the model. Accuracy of the confusion matrix is given by taking the average of diagonal values.



$$\text{Accuracy} = \frac{(TP+F\ )}{N}$$

TP-true positive, TN-true negative, FP-False negative, FP-false positive, N total number of samples

Precision - It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall - It is the number of correct positive results divided by the number of all relevant samples.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1 score - It is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. F1 Score tries to find the balance between precision and recall.

$$\text{F1 score} = 2 * \frac{1}{(\frac{1}{\text{precision}})+(\frac{1}{\text{recall}})}$$

This table below compares the results of various evaluation metrics on the classifiers used.

	MODEL	ACCURACY	CROSS VALIDATION ACCURACY	PRECISION	RECALL	F1 SCORE
1	Random Forest	0.990385	0.978223	1.00000	0.984375	0.992126
2	Logistic Regression	0.971154	0.918118	0.984127	0.968750	0.976378
3	Decision Tree	0.971154	0.954181	1.000000	0.953125	0.976000
4	K-Nearest Neighbour	0.942308	0.915854	1.000000	0.906250	0.950820
5	Gaussian Naive Bayes	0.913462	0.889431	0.936508	0.921875	0.929134

Table: Evaluation of classifiers used

5. Deployment

As we have got logistic regression with highest accuracy and cross validation accuracy so we use it to implement our model. Here we predict diabetes for a random given user. The user answers for the symptoms asked which will be taken as input. This input values are encoded and given to the model which predicts diabetes accordingly.

IV. IMPLEMENTATION

Language and Libraries Used:

- Python – 3.8.9
- Scikit learn – 0.24.2
- Pandas – 1.2.4
- Seaborn – 0.11.1
- Matplotlib – 3.3.4

IV. CONCLUSION

In this project, we built a diabetics prediction model with 99% accuracy which can take symptoms of the user as inputs and predict if he has diabetes or not. Thus, help him adapt his life and diet accordingly

OUTPUTS:

```
#RUN THIS CELL TO PREDICT THE DIABETES OF A PERSON.
diapred()
```

```
ENTER ALL VALUES
What's your age
55
What's your Gender (M/F)
F
Do you have Polyuria (Y/N)
Y
Do you have Polydipsia (Y/N)
N
Do you have sudden weight loss (Y/N)
N
Do you have weakness (Y/N)
Y
Do you have Polyphagia (Y/N)
Y
Do you have Genital thrush (Y/N)
N
Do you have visual blurring (Y/N)
Y
Do you have Itching (Y/N)
N
Do you have Irritability (Y/N)
N
Do you have delayed healing (Y/N)
N
Do you have partial paresis (Y/N)
N
Do you have muscle stiffness (Y/N)
Y
Do you have Alopecia (Y/N)
N
Do you have Obesity (Y/N)
N
Aww!! You might have Diabetes so start taking Precautions as soon as possible
```

```
#RUN THIS CELL TO PREDICT THE DIABETES OF A PERSON.
diapred()
```

```
ENTER ALL VALUES
What's your age
24
What's your Gender (M/F)
M
Do you have Polyuria (Y/N)
N
Do you have Polydipsia (Y/N)
N
Do you have sudden weight loss (Y/N)
N
Do you have weakness (Y/N)
Y
Do you have Polyphagia (Y/N)
N
Do you have Genital thrush (Y/N)
N
Do you have visual blurring (Y/N)
N
Do you have Itching (Y/N)
Y
Do you have Irritability (Y/N)
Y
Do you have delayed healing (Y/N)
N
Do you have partial paresis (Y/N)
N
Do you have muscle stiffness (Y/N)
N
Do you have Alopecia (Y/N)
N
Do you have Obesity (Y/N)
N
Your Safe!! You don't have diabetes
```