Prediction of Diabetes mellitus using Data Mining Techniques

Venkata Lakshmi Dasari CS Hood College vd4@hood.edu

Ahmed Alrefaei CS Hood College aaa15@hood.edu

Dr. Liu Xinlian

Abstract

Data Mining now a day's plays an important role in the prediction of diabetes mellitus. Data Mining is the process of selecting, exploring, and modeling large quantities of records to discover unknown styles or relationships useful to the fact's analyst. Various data mining techniques help diabetes Data Mining now a day's plays an important role in the prediction of diabetes mellitus. Data Mining is the process of selecting, exploring, and modeling large quantities of records to discover unknown styles or relationships useful to the fact's analyst. Various data mining techniques help diabetes studies and in the long run improve the quality of health care for diabetes patients. In this project, we will take the diabetes dataset and provide the results using algorithms. Studies and in the long run improve the quality of health care for diabetes patients. In this project, we will take the diabetes dataset and provide the results using algorithms.

Introduction

Diabetes is a group of metabolic disorders it also referred to as diabetes mellitus. Diabetes is due to not producing enough insulin in the pancreas or the cells of the body not responding proper way to the insulin produced. Diabetes can cause many complications like damage to the eyes, kidney diseases, stroke, cardiovascular disease. These all are serious long-term complications.
There are three types of diabetes. They are type1, type2, gestational diabetes. The type1 diabetes is from pancreas failure to produce enough insulin because of the loss of beta cells. The type2 diabetes is different from type1 diabetes. Starts with insulin resistance, in this type2, the person makes insulin, but the insulin doesn't work in a person's body, or they don't make enough insulin to process the glucose. This type 2 diabetes happens in older persons most of the time who is overweight. The third one is gestational diabetes is happening some pregnant women. It is also like type 2 diabetes. See figure1.

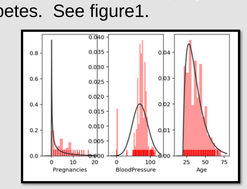

figure1

Some attributes such as, pregnancies, blood pressure, and age which impact both type of diabetes.

Dataset Management:

Diabetes dataset is found in Kaggle. The dataset has different factors which effect diabetes. These factors are pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome which can be either 0 or 1 (has diabetes or not). We will use this data and analyze then we will publish the report. See the attributes in figure 2.

```
Data columns (total 9 columns):
 Pregnancies                 2000 non-null int64
 Glucose                     2000 non-null int64
 BloodPressure               2000 non-null int64
 SkinThickness               2000 non-null int64
 Insulin                     2000 non-null int64
 BMI                         2000 non-null float64
 DiabetesPedigreeFunction    2000 non-null float64
 Age                         2000 non-null int64
 Outcome                     2000 non-null int64
dtypes: float64(2), int64(7)
memory usage: 140.7 KB
```

figure 2

Methodology:

Support Vector Machine (SVM) algorithm:
A Support Vector Machine (SVM) is a machine learning algorithm. SVM is usually used to classification problems. It is also used for regression purposes. SVM uses the concept of finding a hyperplane which uses to divide the dataset into two classes.
We are using SVM to classify our dataset into two different classes which will help us to determine the classes for patients with diabetes and patients with no – diabetes. We are using SVC model from sklearn.svm in our project. We are providing this model with training, and test data and model will provide us with predictions. We are using this prediction to find out the accuracy.
K-nearest neighbors' algorithm:
This algorithm uses the non-parametric method to classify and regression the data. Based on the majority of votes by its neighbor object is classified. It does not use training data to make any generalization. We will be providing K values to choose some neighbors we are going to use to classify our data. We are then using the accurate method to get accuracy for our data classification.
Naïve Bayes Classification:
This algorithm classifies data by assuming the value of the particular feature is independent of the value of any other feature, given the class variable. This algorithm considers each of the features to contribute independently to the probability of the result. This algorithm is easy to build and particularly for large datasets. This algorithm converts dataset into frequency then it creates a likelihood table with less probability and then it uses a Naïve Bayes equation to calculate posterior probability for each class. The formula given below:

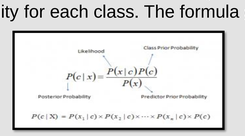

figure 3

Logistic regression algorithm:
It is a statistical method for analyzing a data set. This algorithm is used when there are one or more independent variables which might determine the results. This algorithm uses the sigmoid function.

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

figure 4

It is predictive analysis which used to predict a binary outcome.

Data Analytics :

- Description of the dataset is represented below:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	3.703500	121.182500	69.145500	20.935000	80.254000	32.193000	0.470930	33.090500	0.342000
std	3.306063	32.068636	19.188315	16.103243	111.180534	8.149901	0.323553	11.786423	0.474498
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	63.500000	0.000000	0.000000	27.375000	0.244000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	40.000000	32.300000	0.376000	29.000000	0.000000
75%	6.000000	141.000000	80.000000	32.000000	130.000000	36.800000	0.624000	40.000000	1.000000
max	17.000000	199.000000	122.000000	110.000000	744.000000	80.600000	2.420000	81.000000	1.000000

figure 6

- Finding correlation between attributes using heatmap:

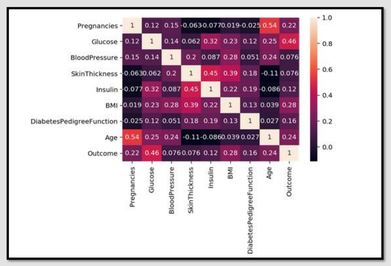

figure 5

- Finding Correlation between attributes as tables:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.120405	0.149672	-0.063375	-0.076600	0.019475	-0.025453	0.539457	0.224437
Glucose	0.120405	1.000000	0.138044	0.062368	0.320371	0.226864	0.123243	0.254496	0.458421
BloodPressure	0.149672	0.138044	1.000000	0.198800	0.087384	0.281545	0.051331	0.238375	0.075958
SkinThickness	-0.063375	0.062368	0.198800	1.000000	0.448859	0.393760	0.178299	-0.111034	0.076040
Insulin	-0.076600	0.320371	0.087384	0.448859	1.000000	0.223012	0.192719	-0.085879	0.120924
BMI	0.019475	0.226864	0.281545	0.393760	0.223012	1.000000	0.125719	0.038987	0.276726
DiabetesPedigreeFunction	-0.025453	0.123243	0.051331	0.178299	0.192719	0.125719	1.000000	0.026569	0.155459
Age	0.539457	0.254496	0.238375	-0.111034	-0.085879	0.038987	0.026569	1.000000	0.236509
Outcome	0.224437	0.458421	0.075958	0.076040	0.120924	0.276726	0.155459	0.236509	1.000000

figure 6

As we can see from the above heat map and the table, the percentage of all attributes are given compared to each other and the outcomes.
This will give us a clue on which features are the most important. After we identify the priority of each features based on the outcomes. We use **SelectK Best where K=4** and **Chi2** to choose the highest four

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

X = data_func.iloc[:,0:8]
Y = data_func.iloc[:,8]
select_top_4 = SelectKBest(score_func=chi2, k = 4)
fit = select_top_4.fit(X,Y)
features = fit.transform(X)
```

```
features[0:5]
```

```
array([[138. ,   0. ,  33.6,  47. ],
       [ 84. , 125. ,  38.2,  23. ],
       [145. ,   0. ,  44.2,  31. ],
       [135. , 250. ,  42.3,  24. ],
       [139. , 480. ,  40.7,  21. ]])
```

figure 7
The results of highest attributes are explained in array in figure 7.

Results:

After applying SVM, NB, KNN, DT, and LR, we got accuracy of each one individually :

```
accurcy of  SVM : 0.7985714285714285
accurcy of  NB : 0.7642857142857142
accurcy of  KNN : 0.7828571428571429
accurcy of  DT : 0.9428571428571427
accurcy of  LR : 0.7785714285714285
```

figure 8

Our aim is computing the F1 score, which is balanced F-measure. The F1 score can be explained as a weighted average of the precision and recall, F1 will have the best value at 1 and worst value at 0. Both contribution of precision and recall to F1 score are equal. The formula is given by

F1 = 2 * (precision * recall) / (precision + recall)

Classification report:

	precision	recall	f1-score	support
0	0.80	0.88	0.83	443
1	0.68	0.54	0.60	217
avg / total	0.76	0.77	0.76	660

figure 9

Conclusion:

Based on algorithms that we used, and diabetes dataset from Kaggle, we will be finding out the number of people who has diabetes and how many people who do not have diabetes. We will show the accuracy of the result using these algorithms. The importance of our project is to help medical industries not only to have a high prediction but also to reduce the cost of the treatment if doctors have an idea on which attributes are impacted the most.

Acknowledgements

Thanks to Dr. Liu for give us this opportunity to apply our knowledge in real life issue which will definitely help our community to have better life. This might be a first step, but we are very happy to take it.