Prediction of Diabetes mellitus using Data Mining Techniques

Venkata Lakshmi Dasari
CS
Hood College
Frederick MD
vd4@hood.edu

Ahmed Alrefaei
CS
Hood College
Frederick MD
aaa15@hood.edu

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 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. The third one is gestational diabetes is happening some pregnant women. It is also like type 2 diabetes.

Methodology:

In this project, five algorithms will be used. 1. Support Vector Machine algorithm. 2.Naive Bayes classifier. 3. k-nearest neighbor algorithm. 4.Decision tree algorithm. 5.Logistic Regression algorithm.

1.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.

2. 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.

3. 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.

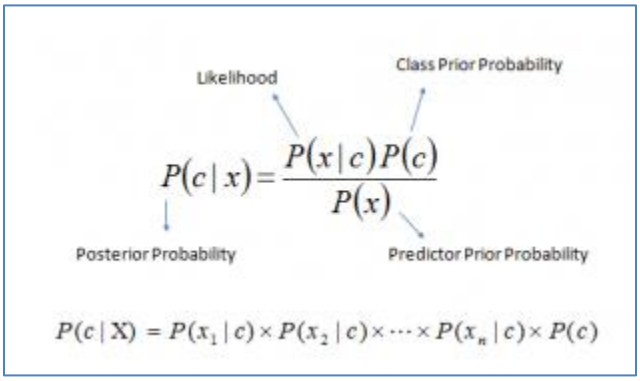

Figure 1: Naïve Bayes Equation

4. Decision Tree Algorithm:

It is one of the most popular machine learning algorithms. It belongs to the supervised learning algorithms family. It can be used for classification and regression purposes. It tries to solve the problem by using tree representation. Each internal node is taken from attribute, and each leaf node represents a class label. It starts from the root of the tree to predicating class label. We need to select proper attributes so that the tree start using that as root hence we need to select randomly and compares tries and modify these root attributes to get accurate results.

5. 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 2: Sigmoid function

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

Results

After applying SVM, NB, KNN, DT, and LR. We got the 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 3 : Result of the five algorithms and DT has the highest accuracy

Classification report

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

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

	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 4: Classification report

Data Management

Diabetes dataset is found in Kaggle. The dataset has different factors which affect diabetes. These factors are pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome. We will use this data and analyze then we will publish the report.

```
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 5: All the attributes and there are no missing data

```
Outcome
0    1316
1     684
dtype: int64
```

Figure 6 : Diabetic and none diabetic

As shown in figure 6, the outcome is represented as a binary value which explain the number of diabetic patients as 1 and the number of non-diabetic patients as 0.

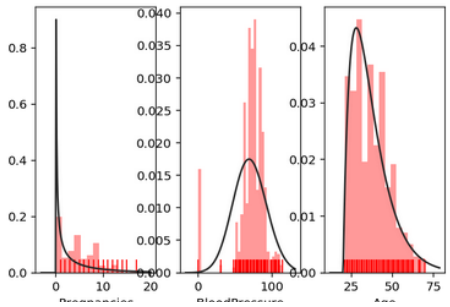

Figure 7: Pregnancies, blood pressure, and age compared with outcome while its positive (1)

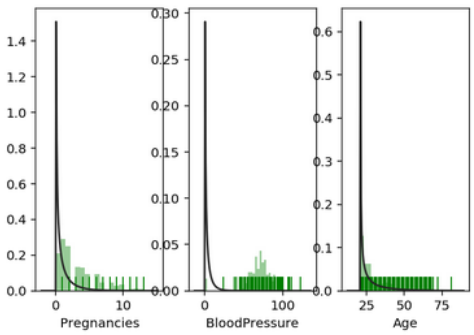

Figure 8: Pregnancies, blood pressure, and age compared with outcome while its negative (0)

Finding correlation between attributes using heat map:

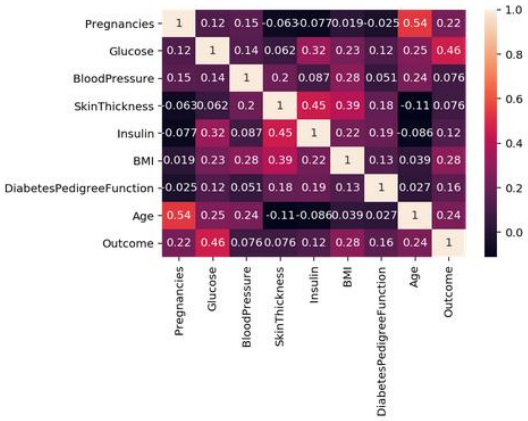

Figure 9 : Heat map

The software will be used

- Jupiter Notebook
- Python

Conclusion

Based on algorithms that we used, and diabetes dataset from Kaggle, we will be finding out the number of people who have 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.

ACKNOWLEDGMENTS

Thanks to Dr. Liu for giving us this opportunity to apply our knowledge in real life issue which will definitely help our community in the future. At the level of programming skills, we have learned a lot of the machine learning libraries, techniques and methods in order to apply them in this project. This might be the first step, but we are very happy to take it.

REFERENCES

[1] Rakesh Motka, Viral Parmar, Balbindra Kumar, A. R. Verma, "Diabetes Mellitus Forecast Using Different Data Mining Techniques," International conference on computer and Communication Technology

[2] Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar, "Diagnosis of Diabetes Mellitus based on Risk Factors", International Journal of Computer Applications, Vol.10, Issue No.4, November.2010

[3] Expert Committee on the Diagnosis and Classification of Diabetes Mellitus Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Diabetes Care 1997; 20: 1183– 1197 [PubMed]

[4] Geiss LS, Pan L, Cadwell B, Gregg EW, Benjamin SM, Engelgau MM: Changes in incidence of diabetes in U.S. adults, 1997–2003. Am J Prev Med 2006; 30: 371– 377 [PubMed]

[5] Pradhan AD, Rifai N, Buring JE, Ridker PM: Hemoglobin A1c predicts diabetes but not cardiovascular disease in nondiabetic women. Am J Med 2007; 120: 720– 727 [PMC free article] [PubMed]

[6] Sherwin R, Jastreboff AM. Year in diabetes 2012: The diabetes tsunami. J Clin Endocrinol Metab. 2012;97:4293–4301. [PMC free article] [PubMed]

[7] Sacks DB. Diagnosis of gestational diabetes mellitus: it is time for international consensus. Clin Chem. 2014;60:141–143. [PubMed]

[8] Mbanya JC. The burden of type 2 diabetes mellitus in the African diaspora. Available at www.medscape.com/view article/560718_2.

[9] Cameron CG, Bennett HA. Cost-effectiveness of insulin analogues for diabetes mellitus. CMAJ 2009 Feb;180(4):400-407.

[10] Fujioka K. Pathophysiology of type 2 diabetes and the role of incretin hormones and beta-cell dysfunction. JAAPA 2007; suppl 3-8

[11] Ahmed AM. History of diabetes mellitus. Saudi Med J 2002 Apr;23(4):373- 378

[12] Anand A. Chaudhari, Prof.S.P.Akarte, " Fuzzy and Data Mining based Disease Prediction using K-NN Algorithm," International Journal of Innovations in Engineering and Technology, Vol. 3, Issue No. 4, April 2014

[13] Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar, " Diagnosis of Diabetes Mellitus based on Risk Factors," International Journal of Computer Applications, Vol. 10, Issue No. 4, November 2010

[14] O'Sullivan JB, Mahan CM: Criteria for the oral glucose tolerance test in pregnancy. Diabetes 1964; 13: 278. [PubMed]

[15] Carpenter MW, Coustan DR: Criteria for screening tests for gestational diabetes. Am J Obstet Gynecol1982; 144: 768– 773 [PubMed]

.