Prediction of Early Onset Diabetes Using Classification Algorithms

By The Hardworking Procrastinators

Christopher Panlasigui[1], Jayashree Kapoor[2], Tia Whiteman[3], Wai Moon Chin[4]

[1,2,3,4]City University of N.Y. - Baruch College Zicklin School of Business

Data Mining for Business Analytics

Progress Summary

Exploratory data analysis has shown that our data has under-represented variables such as gender and class. Some of the findings are consistent with common risk factors with diabetes. We have conducted preliminary multiple logistic regression and Naive Bayes procedures using 80/20 re-sampling. The results we provide are sparse as we are considering making adjustments in our resampling methods.

Data Set

We obtained our dataset from UCI Machine Learning Repository. The data was originally collected using direct questionnaires from patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. It has 520 observations with 17 variables, which are all categorical except for Age. Our target variable is the response Class, which is categorical. There are no specific challenges in our dataset, nor does it contain any missing values. However, our dataset restricts our data mining procedures to classification. We intend to use all variables.

Exploratory Data Analysis

Class	Ave age	Median Age	n
(-)	46.1	45	200
(+)	49.1	48	320
		Total	**520**

Table 1. Breakdown of class distribution and their corresponding average and median ages.

Gender	Ave age	Median Age	n
Male	48.47	47	328
Female	47	48	192
		Total	**520**

Table 2. Age distribution by gender.

Class	Gender	Average age	Median age	Count	Total
(+)	M	51.4	53	147	320
(+)	F	47.1	47	173	
(-)	M	46.4	45	181	200
(-)	F	46.3	50	19	
				Total	**520**

Table 3. Breakdown of Class ((+): positive for diabetes, (-): negative for diabetes) by gender and their corresponding age distribution.

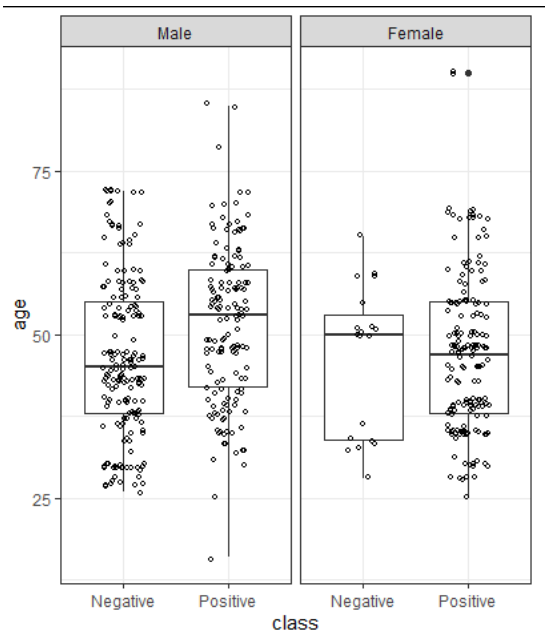
There are 320 positive cases for diabetes and 200 negative cases as shown in table 1. It is about 1.6:1 positive:negative ratio. There are more males than females in our data set. It is about a 1.7:1 male:female ratio as shown in table 2.

Among the positive, there are 147 males and 173 females. Among the negative, there are 181 males and 19 females.

		Class	
		(-)	(+)
Symptoms			
Polyuria (extreme urination)	Y	15	243
	N	185	77
Polydipsia (extreme thirst)	Y	8	225
	N	192	95
		Class	

		(-)	(+)
Weight loss	Y	29	188
	N	171	132
Polyphagia (extreme hunger)	Y	48	189
	N	152	131
Blurry vision	Y	58	175
	N	142	145
Fatigue (weakness)	Y	87	218
	N	113	102
Slow to heal	Y	86	153
	N	114	167
Obesity	Y	27	61
	N	173	259

Table 4. Common symptoms or risk factors associated with diabetes according to the CDC guidelines[1]



The box plot on the left shows the distribution of class by gender. The summary is in table 3.

Our exploratory data analysis provides hints and glimpse on what are the risk indicators. For example, polyuria or excessive urination is a common risk indicator. We see that almost half (258) have the symptom and of those with the symptom 243 is positive for diabetes. The conditional probability of being positive for diabetes given polyuria , P(+ | Y) ~ 0.454.

[1] https://www.cdc.gov/diabetes/basics/symptoms.html

Data Mining Task

We would like to determine whether someone is at risk (positive) or not at risk (negative) for the early onset of diabetes based on a set of symptoms. More specifically, what combination of symptoms is a good indicator?

Methods and Models

Since all but one of the variables are categorical, our predictive data mining task is restricted to classification methods. Thus, we intend to implement multiple logistic regression, Naïve Bayes classifier, and random forest methods. We may perform binning data transformation on the Age variable to turn it into ordinal categorical data to improve model performance. Finally, categorical values will be re-coded.

Due to under-representation in both gender and class, we might consider stratified hold-out in our resampling methods.

Multiple Logistic Regression

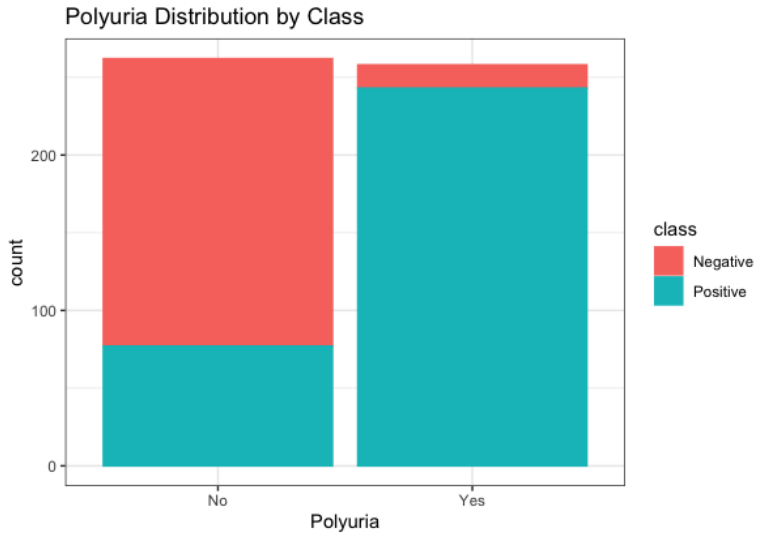
For the multiple logistic regression model, the R function glm() will be used to perform numerous univariate analyses, that is identify the relation of the response with each predictor one at a time and investigate any association to scientific plausibility. We will compare a full model (with all variables) with 2 reduced models obtained from the aggregate of predictors from univariate analysis and from the backward stepwise selection. Since the predictors are categorical, we cannot think of the model outcome as a multiplicative model of odds. As an example, we are not going to be able to talk about the association of response increase with a single unit increase in X_1, holding others (X_2, … X_{16}) constant, because the results depend on the actual (numerical) value of X_1. Rather, the presence or absence of individual predictors in our model will provide the odds of being negative/positive for the disease.

Our preliminary results show that gender, polyuria, polydipsia, polyphagia, genital thrush, itching, irritability and delay healing are all significant based on p-value less than 0.05.

Naïve Bayes

The naïve Bayes classifier is computationally efficient and able to handle categorical variables directly, so we do not need to re-code our categorical variables. It is oftentimes cited to provide good performance when the goal is classification or ranking. The function naiveBayes from the CRAN package e1071 will be used to perform the naïve base classifier. Since it is based on conditional probabilities assuming independence, our model can provide the likelihood of being positive/negative for the disease based on the existence or absence of different symptoms. Since it is best suited for the categorical classifier, we hope that it would perform better than Logistic Regression.

For our exploratory data analysis we perform various visualizations to take a better look at several variables in our data such as Obesity, Polyuria and Polyphagia, which are common risk factors associated with diabetes.



Naive Bayes Classifier for Discrete Predictors

NB_Predictions	Negative	Positive
Negative	180	43
Positive	20	277

-

After we fit the model for the Naive Bayes classifier to process, we were able to classify 180 out of 200 “Negative” cases correctly and 277 out of 320 “Positive” cases correctly. This means the ability of Naive Bayes algorithm to predict “Negative” cases is about 90% but it falls down to only 86.6% of the “Positive” cases resulting in an overall accuracy of 87.9%.

Random Forest

Since we want to predict whether a person is positive for early onset of diabetes based on a combination of symptoms. This prediction will be based on a decision tree where the root node and the internodes represent the symptoms, and the terminal/leaf nodes represent the output or test of whether someone is positive or negative for the disease. Random Forest is a machine learning model that uses an ensemble of decision trees to make its prediction. It randomly samples data and builds an ongoing series of decision trees on a subset. Since it is random, it reduces overfitting and bias.