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

1 Introduction

Diabetes is a major health problem afflicting 537M adults worldwide and is responsible for 6.7M deaths in 2021. It is also an economic problem costing 966B USD in health expenditure in the same year. Certain demographics have it worse than others as 3 out of 4 of adults with diabetes live in low-to-middle income countries. The outlook does not seem promising as the number of diabetes cases is projected to reach 634M by 2030[1].

A lot of contributions have been done to help mitigate this problem. Many of them are model predictions of diabetes using Deep Neural Network, Decision Tree[2], Support Vector Machines, k-Nearest Neighbors, Random Forests[3] just to name a few. The models involved complex data set from laboratory test results like plasma glucose concentration, blood pressure, skin fold thickness, serum insulin, cholesterol, high density lipoprotein (HDL), triglycerides, and state of pregnancy in female among others.

The complexities that made previous work great have also made them inaccessible, difficult to follow and implement. Some just cannot cannot afford expensive tests or are not able to manage laborious repeated monitoring and collection of data from participants. To help address the gap from these complex models, in this paper we will build several models from simpler data from survey questionnaires involving 16 predictors with mostly binary categorical data to classify whether or not someone is at risk for diabetes. The best model can be used to create a screening tool that is accessible by individuals, families and healthcare professionals to predict the disease in its early phase to prevent or delay health complications.

Since most of our variables are going to be categorical, our predictive data mining tasks are restricted to different classification methods. These methods involve multiple logistic (LR) regression to create models based on the log odds, Naive Bayes (NB) based on prior probabilities, and Random Forests (RF) based on ensemble of model prediction procedures such as decision trees, bagging, bootstrapping and aggregation. Additionally, we will investigate rules based on association procedure that might be used as a criterion to associate combination predictors that lead to the disease and whether these predictors match up as significant in our classification models.

2 Data

2.1 Descriptions

We obtained our diabetes data from the UCI Machine Learning Repository[4]. It was originally collected from patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. It comprises of `520` observations with `17` variables. Namely, `age`, `gender`, `polyuria`, `polydipsia`, `wtloss`, `weakness`, `polyphagia`, `genitalthrush`, `visualblur`, `itching`, `irritable`, `delayheal`, `partparesis`, `musclestiff`, `alopecia`, `obesity`, `class`.

As shown in Table 3, the numeric variable `age` has a minimum value of 16, mean value of 48.03, and maximum value of `90`; `gender` comprises 328 `Males` and `192` Females; `polyuria`, a term for excessive urination, has 258 Yes and 262 No; `polydipsia`, a term for excess thirst, has 233 Yes and 287 No; sudden weight loss, `wtloss` has 217 Yes and 303 No; `weakness` has 305 Yes and 215 No; `polyphagia`, a term for excessive eating has 237 Yes and 283 No; `genitalthrush`, a term for common yeast infection caused by fungus candida has 116 Yes and 404 No; visual blurring, `visualblur` has 233 Yes and 287 No; `itching` has 253 Yes and 267 No; `irritability` has 126 Yes and 394 No; `delayedheal` has 239 days of healing and a few days Yes and 281 No; partial paresis, `partparesis` or more commonly known as partial paralysis has 224 Yes and 296 No; `musclestiff` has 195 Yes and 325 No; `alopecia`, an immune system-induced hair loss has 179 Yes and 341 No; `obesity` has 88 Yes and 432 No; `class`, has 320 Positive and 200 Negative. Additional summary is provided in Table 1 for the Negative and Positive case for those that responded "yes" to questionnaires.

2.2 Distributions

The distributions of age by gender and class in Figure 1 (a) show that there are indeed more males than females in both classes. The boxplots in (b) show the mean and median of the age in the positive class is also higher in males. The means of both males and females are about the same in the negative class. However, the median age of female in the negative class is higher.

2.3 Central tendencies

The central tendencies for the age by class differ as shown in Table 2. The mean age for the positive class is `49.1`. For the negative class, the mean is `46.4` Both classes have the same standard deviation of `12.1`. Grouping by the response `class` reveals that there are about 46% positive case of male and 54% female. However, the proportion of male is higher than female overall. So there are about 45% of all males are positive and about 90% of all females are positive.

3 Methods

3.1 Re-Sampling

For our re-sampling procedures, we used 80/20 proportion where about 80% of the data will be used for training and the remainder for the testing of the models. We implemented stratification to our class variable for the proportion method to account for the imbalance in the data. We used 80/20 proportion split instead of other proportions based the paper published

by Gholami etal concluding that $\simeq$ 0.8 is empirically supported the best split.[5]. We also used cross-validation methods with k=10-folds.

3.2 Models and Tools

For the multiple logistic regression (LR), the `class` response variable has `(Negative, Positive)` levels that corresponds to the classifier:

$$\widehat{C}(x) = \begin{cases} 1\text{('Positive')} & \hat{p}(x) > 0.5 \\ 0\text{('Negative')} & \hat{p}(x) \leq 0.5 \end{cases} \tag{1}$$

that was used to obtain the predicted probabilities:

$$\hat{p}(x) = \hat{P}(Y = 1|X = x) \tag{2}$$

We used the function `glm()` from the `stats`[6] package to build full and reduced models to predict the response, `class`. The reduced model was based on the significant predictors from the logistic regression output of the full model. Significant predictors were based on `p-value <0.05`.

For the Naive Bayes (NB), we used the function `naiveBayes()` from the `e1071`[7] package to build a full NB models. Apart from removing continuous variable `age`, we did not perform feature subset selection.

For the Random Forests (RF), we used the function `randomForest()` from the `randomForest`[8] package to build a full model. Then we used the variable importance by mean decrease accuracy cut off value of greater than 20 to create a reduced model. We let the system provide default number of variables for classification, $\sqrt{p}$, to the `mtry` parameter and default out-of-bag value of 500 to the `ntree` parameter.

Since we addressed the class imbalance by utilizing stratified re-sampling, we used the prediction accuracy to measure model performance:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

where: TP: True Positive when an instance is positive and it is classified as positive;

FP: False Positive when an instance is negative and it is classified as positve. It is a Type I error;

TN: True Negative when an instance is negative and it is classified as negative;

FP: False Negative when instance is positive and it is classified as negative. It is a type II error.

We used the same measure performance for cross-validation re-sampling for consistency.

We used the `as()` , `inspect()` from the `arule`[9] package to determine the rules that are positvely associated for diabetes. We also used `arulesViz` package[10] to plot the rules.

Finally, we used a host of data wrangler and visualizer using the `tidyverse`[11] package.

4 Results

4.1 Logistic Regression

The significant predictors of the full LR model from the 80/20 split are `genderFemale`, `polyuriaYes`, `polydipsiaYes`, `genitalthrushYes`, `itchingYes`, `irritableYes` with corresponding odds ratio of `[66.056, 114.241, 117.491, 4.903, 0.043, 9.11]` as shown in Table 4.

The full model's performance in Table 5 show 5 false positive (FP) and 2 false negative (FN) from 104 test observations. Based on this result, the prediction accuracy is 93.27 %.

The full model probability estimate for our binary response, `class`:

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_{16}x_{16}}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_{16}x_{16}}} \tag{4}$$

where: $\beta_0, \beta_1, \cdots, \beta_{16}$ = `[-1.471, -0.044, 4.19, 4.738, 4.766, 0.167, 0.512, 1.124, 1.59, 1.298, -3.14, 2.209, -0.265, 0.821, -0.653, 0.39, 0.021]`.

$x_1, x_2, \cdots, x_{16}$ are all the predictors.

The reduced model from 80/20 data has odds ratio of `[0.088, 68.718, 84.281, 122.122, 2.43, 0.094, 8.683]` as shown in Table 6. The reduced model's performance in Table 7 show 8 false positives (FP) and 3 false negatives (FN) from 104 observations. Based on this result, the prediction accuracy is 89.42 %.

The prediction accuracy of the full LR model from cross-validation method was 92.68 %. While the reduced LR model 91.19 %.

4.2 Naive Bayes

The confusion matrix for the full NB from 80/20 data in Table 11 show that there are 8 false positives (FP) and 4 false negatives from 104 observations. This makes the prediction accuracy about 88.46 %. The corresponding table of prior probabilities with the symptoms are shown in Table 12 and with the gender are shown in Table 13.

The prediction accuracy of full NB from cross-validation data is 87.09 %.

4.3 Random Forests

For the full RF training model from 80/20 data, there were 500 trees generated and the number of variables tried at each split was the default value, 4. Based on the false positive and false negative in the confusion matrix in Table 8 indicate an out-of-bag (OOB) error rate of 2.88 %. The plot of the OOB error rate in Figure 2 (black line) indicate high accuracy prediction. Variable importance parameter was used in the training set.

The confusion matrix of the full RF model in Table 9 show 1 false positive (FP) and 1 false negative (FN). Based on this the prediction accuracy of full RF model from 80/20 split data was 98.08 %. The mean decrease accuracy cut off greater than 20 was used for subset selection . As shown in Figure 3 there were 10 variables in the reduced model namely `polyuria`, `polydipsia`, `gender`, `age`, `delayheal`, `alopecia`, `irritable`, `partparesis`, `itching`, and `weightloss`.

The confusion matrix of the reduced RF model in Table 10 show 3 false positives (FP) and 2 false negatives (FN), based on this prediction accuracy was 95.19 %.

5 Association Rules

Two rules were produced from parameters `minlen=4`, `support=0.3`, `conf=0.5` `rhs='class=Positive'`. However only the first rule makes sense in the context of the disease. These rules are shown in Figure 4 where `rule 1`: `polyuria=Yes`, `polydipsia=Yes`, `alopecia=No` are likely to lead to `class=Positive` with support of 0.32, confidence of 1, lift of 1.63 and a count of 166.

6 Discussion

We have an interesting case in the full LR model about the significance of gender as high predictor. In general, middle-aged men tend to be more susceptible than women[12]. Our model reveals that being female carries a higher risk with an odds ratio of `66.056`. A possible

explanation that this be due to the disproportion of gender when grouped by `class`. For the positive cases female was 54% while male was 46%. In addition, the unequal proportion of positive cases (see Section 2.3) when comparing gender overall compounded this problem.

The full LR model from the 80/20 re-sampling had a prediction accuracy of 93.27%. As expected, it performed better when compared to the reduced LR from 80/20. It also performed better than both the full NB models using 80/20 and cross-validation. Although, it was not far off from the full LR 80/20, it was surprising that full LR from cross-validation under-performed with 92.68 % prediction accuracy. This may be because we did not implement stratified cross-validation. This could be done in future studies to improve the prediction accuracy of our current models.

The reduced RF model from 80/20, which had a 95.19 % prediction accuracy performed better than the highest performing full LR model. The best model is the full RF model from the 80/20 re-sampling at 98.08 % prediction accuracy as shown in Table 14.

In addition, the predictors `Polyuria=Yes`, `Polydipsia=Yes` and `Gender` always emerged very significant in different models as shown in Figure 5. These predictors combined make a compelling case for markers for the early phase of diabetes. These rules are easy to follow that they can be used as quick guideline for people who do not have an easy access to healthcare.

7 Conclusion

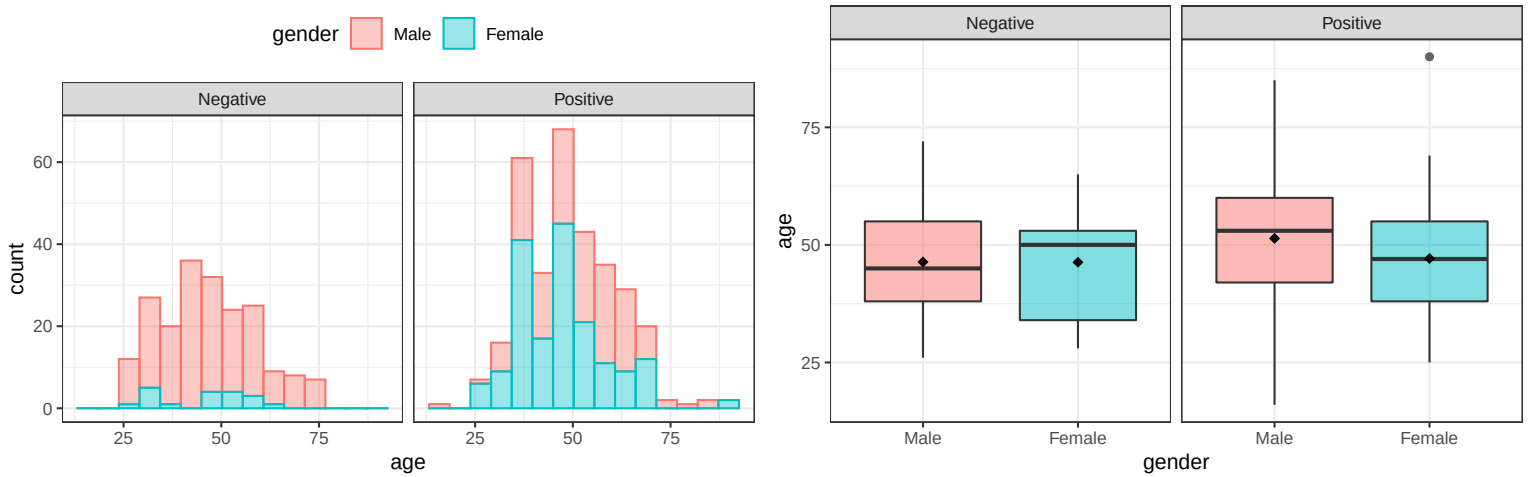
We have created a high performing random forest model with from 80/20 data with a prediction accuracy of 98.08 %. This model is very accessible as it only requires 16 non-invasive data such as age, gender, and certain symptoms and conditions for an input. It does not not even require a family medical history and the output is easy to understand and interpret. This model or an improved version can be used as part of a screening tool for prognosis of early onset of diabetes especially for those who are reluctant getting invasive diagnostic test and procedures. We see a great potential when it is deployed as an app for everyone to use. If there is no access to an app, the information in Figure 5 can be used a guideline. With this information and model in an app, we are one step closer in helping prevent or delay health complications from diabetes.

8 Figures and Tables

8.1 EDA

Table 1: Summary of Diabetes data acquired from UCI Machine Learning Repository. Characteristics in the Negative and Positive columns are those that responded 'Yes'.

Characteristic	**Overall**, N = 520	**Negative**, N = 200	**Positive**, N = 320
polyuria	258 (50%)	15 (7.5%)	243 (76%)
polydipsia	233 (45%)	8 (4.0%)	225 (70%)
wtloss	217 (42%)	29 (14%)	188 (59%)
weakness	305 (59%)	87 (44%)	218 (68%)
polyphagia	237 (46%)	48 (24%)	189 (59%)
genitalthrush	116 (22%)	33 (16%)	83 (26%)
visualblur	233 (45%)	58 (29%)	175 (55%)
itching	253 (49%)	99 (50%)	154 (48%)
irritable	126 (24%)	16 (8.0%)	110 (34%)
delayheal	239 (46%)	86 (43%)	153 (48%)
partparesis	224 (43%)	32 (16%)	192 (60%)
musclestiff	195 (38%)	60 (30%)	135 (42%)
alopecia	179 (34%)	101 (50%)	78 (24%)
obesity	88 (17%)	27 (14%)	61 (19%)



(a) Histogram

(b) Box plot

Figure 1: Distribution of Age

Table 2: Average age for each class and by class with gender

Characteristic	**Negative**, N = 200	**Positive**, N = 320	**p-value**
age	46 (12)	49 (12)	0.012
gender			<0.001
Male	181 (90%)	147 (46%)	
Female	19 (9.5%)	173 (54%)	

Table 3: Table of Responses

predictor	response	n
alopecia	No	341
alopecia	Yes	179
delayheal	No	281
delayheal	Yes	239
gender	Male	328
gender	Female	192
genitalthrush	No	404
genitalthrush	Yes	116
irritable	No	394
irritable	Yes	126
itching	No	267
itching	Yes	253
musclestiff	No	325
musclestiff	Yes	195
obesity	No	432
obesity	Yes	88
partparesis	No	296
partparesis	Yes	224
polydipsia	No	287
polydipsia	Yes	233
polyphagia	No	283
polyphagia	Yes	237
polyuria	No	262
polyuria	Yes	258
visualblur	No	287
visualblur	Yes	233
weakness	No	215
weakness	Yes	305
wtloss	No	303
wtloss	Yes	217

Table 4: Summary statistics of Full LR model using 80/20 data

term	oddsratio	estimate	p.value
(Intercept)	0.230	-1.471	0.189
age	0.957	-0.044	0.119
genderFemale	66.056	4.190	0.000
polyuriaYes	114.241	4.738	0.000
polydipsiaYes	117.491	4.766	0.000
wtlossYes	1.181	0.167	0.782
weaknessYes	1.669	0.512	0.396
polyphagiaYes	3.078	1.124	0.061
genitalthrushYes	4.903	1.590	0.011
visualblurYes	3.662	1.298	0.070
itchingYes	0.043	-3.140	0.000
irritableYes	9.110	2.209	0.001
delayhealYes	0.767	-0.265	0.660
partparesisYes	2.272	0.821	0.147
musclestiffYes	0.520	-0.653	0.297
alopeciaYes	1.477	0.390	0.557
obesityYes	1.021	0.021	0.972

Table 5: Confusion matrix of the full model LR using 80/20 data

	Predicted	
	Negative	Positive
Actual		
Negative	65	5
Positive	2	32

Table 6: summary statistics of reduced LR model using 80/20 data

term	oddsratio	statistic	p.value
(Intercept)	0.088	-5.892	0.000
genderFemale	68.718	6.819	0.000
polyuriaYes	84.281	6.834	0.000
polydipsiaYes	122.122	6.176	0.000
genitalthrushYes	2.430	1.641	0.101
itchingYes	0.094	-4.388	0.000
irritableYes	8.683	3.604	0.000

Table 7: Confusion matrix of the reduced LR model using 80/20 data

	Predicted	
	Negative	Positive
Actual		
Negative	62	8
Positive	3	31

Table 8: Confusion Matrix of the full RF training model

	Negative	Positive	class.error
Negative	122	8	0.0615385
Positive	4	282	0.0139860

8.2 Logistic Regression

8.3 Random Forests

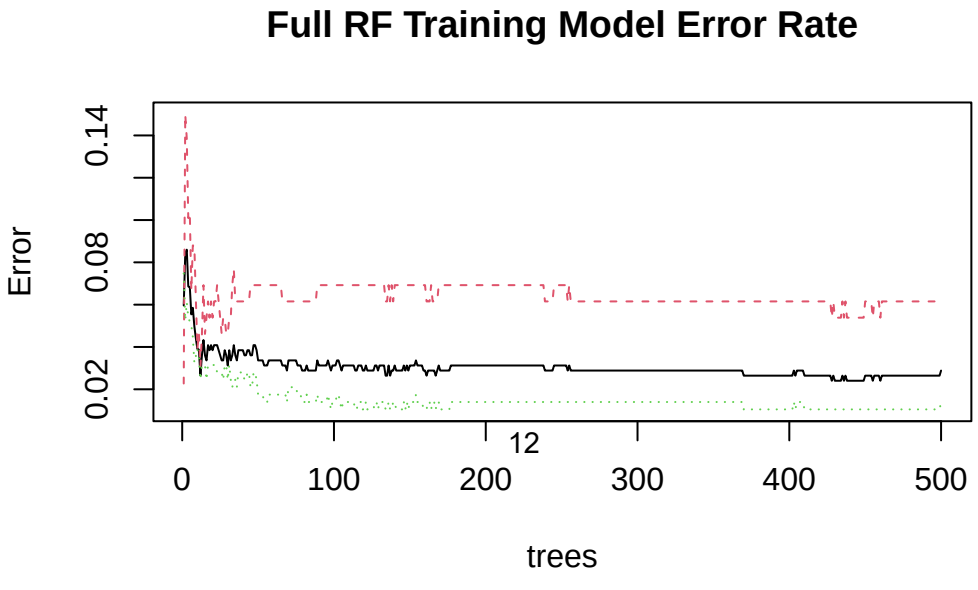


Figure 2: Plot of error rate of the full RF training model

Table 9: Confusion matrix of the full RF model using 80/20 data

	Predicted	
	Negative	Positive
Actual		
Negative	69	1
Positive	1	33

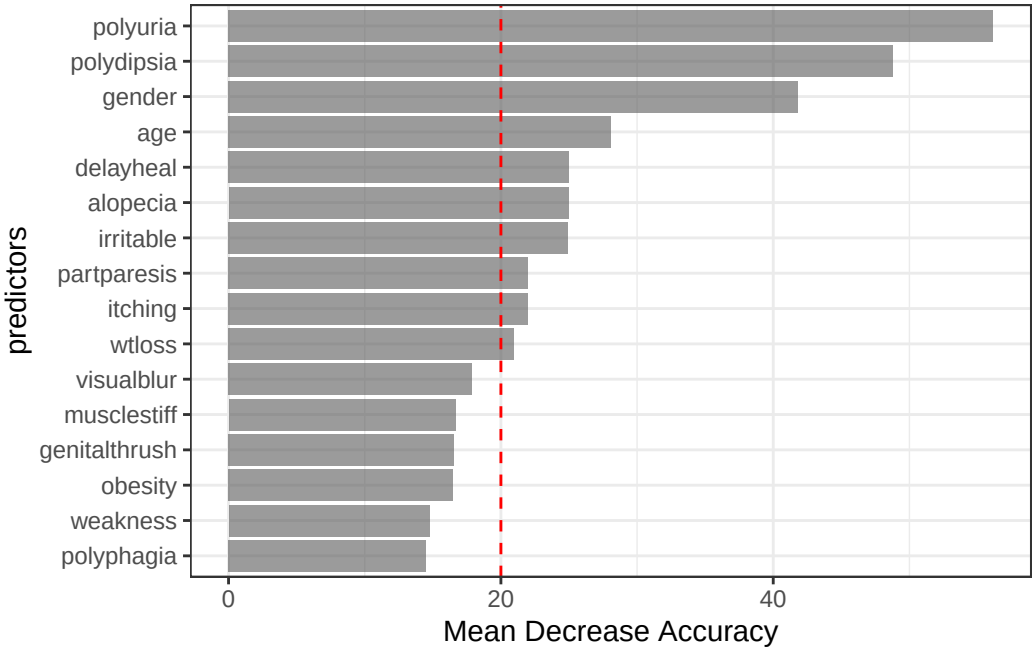


Figure 3: Variable Importance of the full RF model from 80/20 data

Table 10: Confusion matrix of the reduced RF model using 80/20 data

	Predicted	
	Negative	Positive
Actual		
Negative	67	3
Positive	2	32

8.4 Naive Bayes

Table 11: Confusion matrix of the full NB model using 80/20 data

	Predicted	
	Negative	Positive
Actual		
Negative	62	8
Positive	4	30

Table 12: Prior probability of class on symptoms

Prior Probabilities of class on symptoms
P(class | symptoms)

symptoms	Negative.No	Positive.No	Negative.Yes	Positve.Yes
polyuria	0.946	0.238	0.054	0.762
polydipsia	0.954	0.301	0.046	0.699
wtloss	0.854	0.416	0.146	0.584
weakness	0.546	0.318	0.454	0.682
polyphagia	0.792	0.409	0.208	0.591
genitalthrush	0.831	0.734	0.169	0.266
visualblur	0.723	0.441	0.277	0.559
itching	0.492	0.524	0.508	0.476
irritable	0.923	0.643	0.077	0.357
delayheal	0.585	0.514	0.415	0.486
partparesis	0.862	0.406	0.138	0.594
musclestiff	0.715	0.566	0.285	0.434
alopecia	0.477	0.745	0.523	0.255
obesity	0.854	0.801	0.146	0.199

Table 13: Prior probability of class on gender

Prior Probabilities
P(class | gender)

symptoms	Negative.No	Positive.No	Negative.Yes	Positve.Yes
gender	0.892	0.479	0.108	0.521

Table 14: Summary of the model performance

model	type	resampling	accuracy
LR	full	prop	93.27
LR	reduced	prop	89.42
LR	full	cv	92.68
LR	reduced	cv	91.19
NB	full	prop	88.46
NB	full	cv	87.09
RF	full	prop	98.08
RF	reduced	prop	95.19

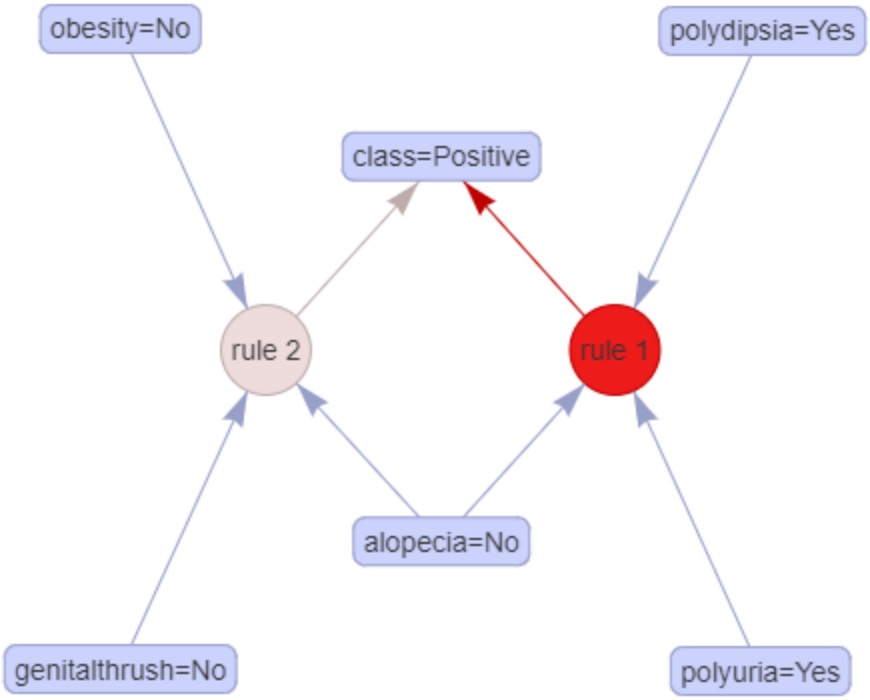


Figure 4: diabetes rules

ASSOCIATION	LOGISTIC REG reduced	RANDOM FOREST	
Polyuria Y	Polyuria Y, 84x	Polyuria Y	Weight loss
Polydipsia Y	Polydipsia Y, 122x	Polydipsia Y	
Alopecia N		Alopecia Y	
	Gender F, 68x	Gender	
	Irritable Y, 8x	Irritable Y	
	Itching Y	Itching Y	
	Thrush, 2x		
		Age	
		Delay heal Y	
		Part paresis	

Figure 5: common predictors

References

1. *Global diabetes data report 2000 — 2045*. (n.d.). Retrieved November 30, 2022, from https://diabetesatlas.org/data/en/world/
2. Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Computer Methods and Programs in Biomedicine*, *152*, 23–34. https://doi.org/https://doi.org/10.1016/j.cmpb.2017.09.004
3. Muhammad, L. J., Algehyne, E. A., & Usman, S. S. (2020). Predictive Supervised Machine Learning Models for Diabetes Mellitus. *SN COMPUT. SCI.*, *1*(5), 240. https://doi.org/10.1007/s42979-020-00250-8
4. *Early stage diabetes risk prediction*. (2020). [Web Page]. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/machine-learning-databases/00529/
5. Gholamy, A., Kreinovich, V., & Kosheleva, O. (n.d.). *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation*. 7.
6. R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
7. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2022). *e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), TU wien*. https://CRAN.R-project.org/package=e1071
8. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. https://CRAN.R-project.org/doc/Rnews/
9. Hahsler, M., Buchta, C., Gruen, B., & Hornik, K. (2022). *Arules: Mining association rules and frequent itemsets*. https://CRAN.R-project.org/package=arules
10. Hahsler, M. (2021). *arulesViz: Visualizing association rules and frequent itemsets*. https://CRAN.R-project.org/package=arulesViz
11. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686
12. Gale, E. A., & Gillespie, K. M. (2001). Diabetes and gender. *Diabetologia*, *44*(1), 3–15. https://doi.org/10.1007/s001250051573