Risk Prediction of Diabetes at Early Stage using Logistic Regression and Naive Bayes Algorithms

Rajeev Agrawal

6/14/2021

1. Introduction

Diabetes is a metabolic disorder in which the patient experiences high blood sugar levels over a prolonged period. According to the World Health Organization (WHO), diabetes is one of the fastest growing chronic diseases that has affected hundreds of millions of people worldwide [1]. If left untreated, diabetes may cause serious long-term complications such as cardiovascular disease, stroke, kidney failure, foot ulcers, and blindness. Due to the presence of a relatively long asymptomatic phase, early detection of diabetes is very important for a clinically meaningful outcome and keeping people with diabetes healthy.

1.1 Research Question

The current study builds prediction models using logistic regression and naive Bayes algorithms to predict the risk of diabetes at an early stage.

1.2 Data

For this study, the data set containing sign and symptom data of newly diabetic or would be diabetic patients [2] collected by Sylhet Diabetes Hospital in Sylhet, Bangladesh, is used to build the prediction models. The data set contains total 520 records with the following 17 variables (16 of which are predictor variables and one outcome variable - *class*):

- **Categorical variables**
  - *Gender:* Male or Female.
  - *Polyuria:* (excessive or an abnormally large production or passage of urine) Yes or No.
  - *Polydipsia:* (excessive thirst or excess drinking) Yes or No.
  - *sudden_weight_loss:* Yes or No.
  - *weakness:* Yes or No.
  - *Polyphagia:* (abnormally strong sensation of hunger or desire to eat) Yes or No.
  - *Genital_thrush:* (a fungal infection caused by Candida yeasts) Yes or No.
  - *visual_blurring:* Yes or No.

- *Itching:* Yes or No.
- *Irritability:* Yes or No.
- *delayed_healing:* Yes or No.
- *partial_paresis:* (partial paralysis) Yes or No.
- *muscle_stiffness:* Yes or No.
- *Alopecia:* (hair loss) Yes or No.
- *Obesity:* Yes or No.
- *class:* (patient's diabetes diagnosis) 1 = Positive or 0 = Negative.

- **Quantitative variables**
  - *Age:* Patient's age.

1.3 Literature Review

The research article by M. M. Faniqul Islam et al. [2] makes predictions about the likelihood of diabetes at early stage using data mining classification methods such as naive Bayes, logistic regression, and random forest. The data set used in their work was collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. This is the same data set that is also used in the current study.

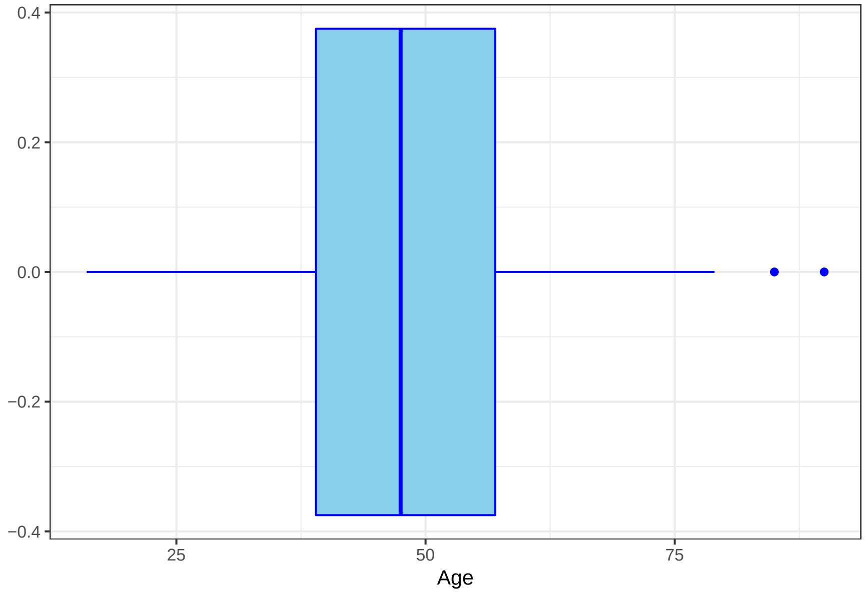
2. Exploratory Data Analysis

There are no missing values in the data set.

```
diabetes %>%
  summary()
```

```
##       Age           Gender    Polyuria  Polydipsia sudden_weight_loss weakness
##  Min.   :16.00   Male  :328   No :262   Yes:233    No :303            Yes:305
##  1st Qu.:39.00   Female:192   Yes:258   No :287    Yes:217            No :215
##  Median :47.50
##  Mean   :48.03
##  3rd Qu.:57.00
##  Max.   :90.00
##  Polyphagia Genital_thrush visual_blurring Itching   Irritability
##  No :283    No :404        No :287         Yes:253   No :394
##  Yes:237    Yes:116        Yes:233         No :267   Yes:126
##
##
##
##
##  delayed_healing partial_paresis muscle_stiffness Alopecia  Obesity   class
##  Yes:239         No :296         Yes:195          Yes:179   Yes: 88   0:200
##  No :281         Yes:224         No :325          No :341   No :432   1:320
##
##
##
##
```

```
diabetes %>%
  ggplot() + geom_boxplot(aes(Age), color = "blue", fill = "skyblue") +
  theme_bw()
```



3. Binomial Logistic Regression Model

Assumptions

1. Binary logistic regression requires the dependent variable to be binary.
2. The observations are independent of each other.
3. There is no severe multicollinearity among the explanatory variables.
4. There are no extreme outliers.
5. The independent variables are linearly related to the log odds.
6. The sample size of the dataset is large enough to draw valid conclusions from the fitted logistic regression model.

Out of the above 6 assumptions, the 3rd assumption about multicollinearity will be tested using variance inflation factor (VIF). We will remove the outliers from our data to take care of the 4th assumption. There is no evidence to suggest that the remaining 4 assumptions are violated.

Dealing with Outliers

We remove the 4 outliers for Age from our data.

```
d1 <- diabetes %>%
  filter(Age < 84)
```

Splitting the Data

We partition our data into training and test data sets using 75% to 25% split.

```
set.seed(1234)
sample_set <- sample(nrow(d1), round(nrow(d1)*.75), replace = FALSE)
d1_train <- d1[sample_set, ]
d1_test <- d1[-sample_set, ]
```

Dealing with Class Imbalance

We see that the class distributions across the three sets are similar. The test data should mirror the class distribution of the original data because a model's performance against the test data is a proxy for its generalizability against unseen data. However, any imbalance in the training data is balanced prior to the modeling process.

```
#Splitting the data
round(prop.table(table(select(d1, class))),4)*100
```

```
## 
##     0     1 
## 38.76 61.24
```

```
round(prop.table(table(select(d1_train, class))),4)*100
```

```
## 
##     0     1 
## 38.76 61.24
```

```
round(prop.table(table(select(d1_test, class))),4)*100
```

```
##
##     0     1
## 38.76 61.24
```

```
#Balance the training data
set.seed(1234)
d1_train <- smote(class ~ ., data.frame(d1_train), perc.over = 1, perc.under = 2)
round(prop.table(table(select(d1_train, class))),4)*100
```

```
##
##  0  1
## 50 50
```

Training and Evaluating the Model

We see that based on the p-values, some of the features in this full model are not significant.

```
d1_model <- glm(d1_train, family = binomial, formula = class ~.)
summary(d1_model)
```

```
##
## Call:
## glm(formula = class ~ ., family = binomial, data = d1_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.82404  -0.20314  -0.00681   0.02205   2.82427
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.34057    1.92537  -0.177 0.859597
## Age                   -0.07271    0.02994  -2.428 0.015164 *
## GenderFemale           4.20312    0.56767   7.404 1.32e-13 ***
## PolyuriaYes            5.53807    0.77098   7.183 6.81e-13 ***
## PolydipsiaNo          -4.94547    0.80725  -6.126 8.99e-10 ***
## sudden_weight_lossYes  0.38420    0.59926   0.641 0.521441
## weaknessNo            -0.90989    0.57378  -1.586 0.112787
## PolyphagiaYes          1.62260    0.59697   2.718 0.006567 **
## Genital_thrushYes      2.39582    0.55440   4.321 1.55e-05 ***
## visual_blurringYes     1.61049    0.70464   2.286 0.022281 *
## ItchingNo              2.31937    0.61093   3.796 0.000147 ***
## IrritabilityYes        2.74208    0.67809   4.044 5.26e-05 ***
## delayed_healingNo      0.17884    0.62443   0.286 0.774572
## partial_paresisYes     0.63444    0.55355   1.146 0.251735
## muscle_stiffnessNo     1.40970    0.57693   2.443 0.014547 *
## AlopeciaNo            0.74170    0.71696   1.035 0.300899
## ObesityNo             0.27755    0.61057   0.455 0.649409
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 831.78  on 599  degrees of freedom
## Residual deviance: 192.56  on 583  degrees of freedom
## AIC: 226.56
##
## Number of Fisher Scoring iterations: 8
```

Performing Stepwise Variable Selection

We perform stepwise variable selection based on full model and as can be seen, the variables - *sudden_weight_loss, delayed_healing, partial_paresis and Obesity* are dropped from the original full model. When comparing two models, the model with the lower AIC is preferred. We can see that the AIC of the new model (= 221.08) is slightly lower than the original full model’s AIC of 226.56.

```
#Step-wise Regression: AIC
step_model <- step(object = d1_model, trace = FALSE)
summary(step_model)
```

```
##
## Call:
## glm(formula = class ~ Age + Gender + Polyuria + Polydipsia +
##     weakness + Polyphagia + Genital_thrush + visual_blurring +
##     Itching + Irritability + muscle_stiffness + Alopecia, family = binomial,
##     data = d1_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.88253  -0.21497  -0.00664   0.01629   2.80031
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.38963    1.59255  -0.245  0.80672
## Age                -0.06658    0.02794  -2.383  0.01716 *
## GenderFemale        4.22924    0.56008   7.551 4.31e-14 ***
## PolyuriaYes         5.81127    0.73727   7.882 3.22e-15 ***
## PolydipsiaNo       -5.04003    0.77021  -6.544 6.00e-11 ***
## weaknessNo         -1.05795    0.45144  -2.344  0.01910 *
## PolyphagiaYes       1.62450    0.55207   2.943  0.00326 **
## Genital_thrushYes   2.34797    0.54082   4.341 1.42e-05 ***
## visual_blurringYes  1.99748    0.63489   3.146  0.00165 **
## ItchingNo           2.50619    0.55561   4.511 6.46e-06 ***
## IrritabilityYes     2.64448    0.62930   4.202 2.64e-05 ***
## muscle_stiffnessNo  1.48597    0.55387   2.683  0.00730 **
## AlopeciaNo          1.09919    0.59775   1.839  0.06593 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 831.78  on 599  degrees of freedom
## Residual deviance: 195.08  on 587  degrees of freedom
## AIC: 221.08
##
## Number of Fisher Scoring iterations: 8
```

Dealing with Multicollinearity

Multicollinearity is a problem because it makes it difficult to separate out the impact of individual predictors on response. A VIF of greater than 5 indicates the presence of multicollinearity and requires remediation. Our results show that none of the features have a VIF larger than 5.

```
vif(step_model)
```

```
##              Age           Gender         Polyuria       Polydipsia
##         2.997806         2.164186         2.783865         1.671043
##         weakness       Polyphagia   Genital_thrush  visual_blurring
##         1.438764         1.684581         1.593145         2.766585
##          Itching     Irritability muscle_stiffness         Alopecia
##         2.233669         1.446221         1.802139         2.389750
```

Choosing a Cutoff Value

```
d1_predl <- predict(step_model, d1_test, type = "response")

#Choosing a Cutoff Value
ideal_cutoff <- optimalCutoff(
  actuals = d1_test$class,
  predictedScores = d1_predl,
  optimiseFor = "Both")
ideal_cutoff
```

```
## [1] 0.4599999
```

Prediction Accuracy

Using the recommended cutoff value of 0.46, we transform our predictions and calculate our model predictive accuracy. Results show that logistic regression model's predictive accuracy is 93.8%.

```
d1_predl <- if_else(d1_predl >= ideal_cutoff, 1, 0)
d1_predl_table <- table(d1_test$class, d1_predl)
d1_predl_table
```

```
##    d1_predl
##       0  1
##   0 46  4
##   1  4 75
```

```
sum(diag(d1_predl_table))/nrow(d1_test)
```

```
## [1] 0.9379845
```

4. Naive Bayes Classifier

Building and Evaluating the Model

Results show that naive Bayes classifier's predictive accuracy is 83.7%.

```
#Building the naive Bayes model
d2_model <- naiveBayes(class ~., data = d1_train, laplace = 1)

#Evaluating the model
d2_predl <- predict(d2_model, d1_test, type = "class")
d2_predl_table <- table(d1_test$class, d2_predl)
d2_predl_table
```

```
##    d2_predl
##       0  1
##   0 43  7
##   1 14 65
```

```
sum(diag(d2_predl_table))/nrow(d1_test)
```

```
## [1] 0.8372093
```

ROC curves for the prediction models

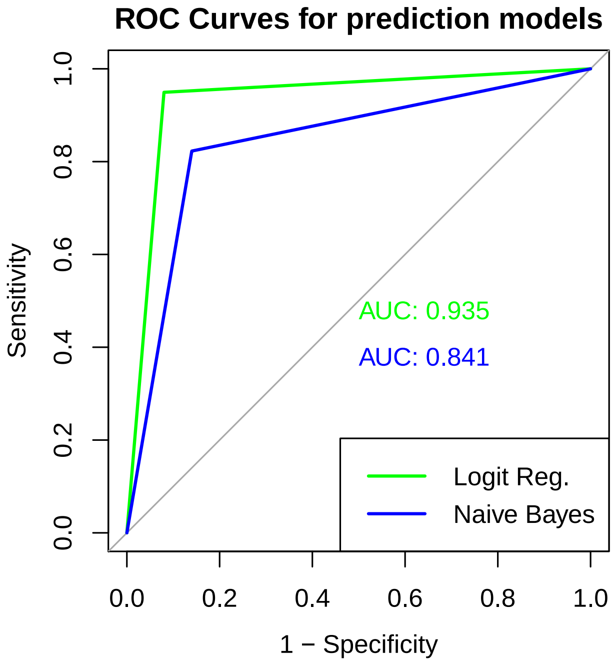
ROC curve is commonly used to visually represent the relationship between a model's true positive rate and false positive rate for all possible cutoff values. ROC curve is summarized into a single quantity known as area under the curve (AUC), which measures the entire two-dimensional area underneath the ROC curve from (0,0) to (1,1). The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. As can be seen, the AUC for naive Bayes model is 0.841 while that for logistic regression model is 0.935.

```r
nB_predl <- d2_predl %>%
  as.vector() %>%
  as.numeric()

par(pty="s")
ROC1 <- roc(d1_test$class ~ d1_predl, plot=TRUE, print.auc=TRUE,
               col="green", lwd =2, legacy.axes=TRUE,
               main="ROC Curves for prediction models")

ROC2 <- roc(d1_test$class ~ nB_predl, plot=TRUE, print.auc=TRUE,
               col="blue", lwd = 2, print.auc.y=0.4, legacy.axes=TRUE, add = TRUE)

legend("bottomright", legend=c("Logit Reg.","Naive Bayes"),
       col=c("green","blue"), lwd=2)
```



5. Conclusion

The capability to predict diabetes early assumes a vital role for the patient's appropriate treatment procedure. Machine learning methods are valuable in this early diagnosis of diabetes. In the current study, two machine learning techniques were applied on a training data set and validated against a test data set; both of these data sets were based on the data collected from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. The results of our model implementations show that based on both the measures of future performance - prediction accuracy and the AUC, the logistic regression classifier outperforms the naive Bayes classifier. One limitation of the current study is that it may only be valid on a similar data set as was used for this study, which was sourced from a very specific location. Further research is needed to check if similar results are seen for data collected elsewhere.

6. References

1. World Health Organization https://www.who.int/news-room/fact-sheets/detail/diabetes [accessed in June 2021]
2. Islam M.M.F., Ferdousi R., Rahman S., Bushra H.Y. (2020) Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. In: Gupta M., Konar D., Bhattacharyya S., Biswas S. (eds) Computer Vision and Machine Intelligence in Medical Image Analysis. Advances in Intelligent Systems and Computing, vol 992. Springer, Singapore. https://doi.org/10.1007/978-981-13-8798-2_12