DIABETES DETECTION PROJECT

Khyaati Khanna, Anish Bahl, Anand Sabnis, Eoin Donovan, Jacob Hegy

Diabetes is a common chronic ailment in the United States, impacting millions of individuals and placing a significant economic burden on the nation. From Type I diabetes which is insulin-dependent to Type II diabetes which is not insulin-dependent, many different factors contribute to whether or not you have an onset of diabetes at some point in your life.

NOTE→ The dataset was selected from Kaggle. The dataset has 70692 observations with 21 features.

By use of the Logistic Regression and Random Forests model, our goal is to accurately answer the following:

Which model best predicts whether a patient has diabetes or not?

Description of Variables:

- The response variable is the **diabetes_binary** variable. This variable will be 1 if a person does have diabetes and will be 0 if a person does not have diabetes.

The predictor variables are:

→ **HighBP :** 0 = no high BP 1 = high B
→ **HighChol :** 0 = no high cholesterol 1 = high cholesterol
→ **CholCheck:** 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
→ **BMI:** Body Mass Index
→ **Smoker:** Have you smoked at least 100 cigarettes in your entire life? (Note: 5 packs = 100 cigarettes) 0 = no 1 = yes
→ **Stroke:** you had a stroke. 0 = no 1 = yes
→ **HeartDiseaseorAttack:** coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
→ **PhysActivity:** physical activity in the past 30 days - not including job 0 = no 1 = yes
→ **Fruits:** Consume Fruit 1 or more times per day 0 = no 1 = yes
→ **Veggies:** Consume Vegetables 1 or more times per day 0 = no 1 = yes
→ **HvyAlcoholConsump:** (adult men >=14 drinks per week and adult women>=7 drinks per week) 0 = no 1 = yes
→ **AnyHealthcare:** Have any kind of health-care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
→ **NoDocbcCost:** Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes
→ **GenHlth:** Would you say that in general, your health is: a scale of 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
→ **MentHlth:** days of poor mental health scale 1-30 days

→ **PhysHlth:** physical illness or injury days in the past 30 days scale of 1-30
→ **DiffWalk:** Binary feature indicating if the individual exhibits an irregular walk (0 = false, 1 = true)
→ **Sex:** Binary feature indicating individual's sex (0 = female, 1 = male)
→ **Age:** A measurement of the individual's age on a 13-point scale where 1 = 18 - 24 years old and 13 = 80 years old or older
→ **Education:** A measurement of education level on a 6-point scale where 1 = No school/only kindergarten and 6 = 4 or more years of college
→ **Income:** A measurement of the individual's income on an 8-point scale where 1 = < $10000 and 8 = > $75000

Models

- The models that will be implemented for this research project include **logistic regression** and **random forest**.

Logistic Regression was chosen due the several advantages that it possesses in regards to classification problems. Some of logistic regression's advantages include its simplicity in implementation, its characteristic of providing a probabilistic output which is useful for this research's binary classification problem, and the interpretability of the output since it provides a means to measure how likely a specific outcome is. However, logistic regression does come with its disadvantages which include its limited expressiveness as seen in the fact that it assumes a linear relationship in the data and it can sometimes not perform as well with a large number of features in the dataset.

Random Forest was chosen due to its various advantages that come along with its classification abilities. Some of these advantages of random forest include its ability to aggregate the predictions of multiple trees and hence be less prone to overfitting and capture the most robust patterns in the data. Random forest is also better able to handle non-linear relationships unlike logistic regression. However, random forest does have some disadvantages which include its lack of interpretability as compared to logistic regression.

We decided to keep all predictor variables in the random forests model.

Logistic Regression Formula:

$$1 / 1 + e^{-(\beta_0 + \beta_1 \text{HighBP} + \beta_2 \text{HighChol} + \beta_3 \text{CholCheck} + \beta_4 \text{BMI} + \beta_5 \text{Stroke} + \beta_6 \text{HeartDiseaseorAttack} + \beta_7 \text{PhysActivity} + \beta_8 \text{Veggies} + \beta_9 \text{HvyAlcoholConsumption} + \beta_{10} \text{GenHlth} + \beta_{11} \text{MenHlth} + \beta_{12} \text{PhysHlth} + \beta_{13} \text{DiffWalk} + \beta_{14} \text{Sex} + \beta_{15} \text{Age} + \beta_{16} \text{Education} + \beta_{17} \text{Income})}$$

Random Forest Formula:

diabetes_binary ~ HighBP + HighChol + CholCheck + BMI + Smoker + HeartDiseaseorAttack + PhysActivity + Fruits + Veggies + HvyAlcoholConsumption + AnyHealthCare + NoDocbcCost + GenHlth + MenHlth + PhysHlth + DiffWalk + Sex + Age + Education + Income

Logistic regression

To fit the model we looked at the summary of the original model with no exclusions. We then saw that 4 predictors had very high p values and we eliminated them. Once we eliminated them, we saw a minimal improvement, suggesting that the main predictors are much more dominant. We took out Smoker, Fruits, AnyHealthcare, NoDocbcCost.

```
lr_error_rates = list(10)
lr_mse = list(10)
for(i in 1:10) {
  set.seed(i)
  train <- sample(1:nrow(diabetes), (nrow(diabetes) * .8) + .5)
  test <- diabetes[-train,]
  print(train)
  print('\n')
  model2 <- glm(Diabetes_binary ~ . - Smoker - Fruits -
AnyHealthcare - NoDocbcCost, family = "binomial", data = train)
  pred <- predict(model, newdata = test, type = "response")
  yHat <- pred > 0.5
  yHat <- as.integer(as.logical(yHat))
  lr_error_rates[i] <- mean(yHat != test$Diabetes_binary)
  lr_mse[i] <- mean((yHat != test$Diabetes_binary))
}
lr_error_rates <- unlist(lr_error_rates)
print(lr_error_rates)
print(mean(lr_error_rates))
```

Random Forest

For random forest, we did not remove any variables. This way we are able to determine the most important variable in the problem of classifying whether a person has diabetes or does not have diabetes.

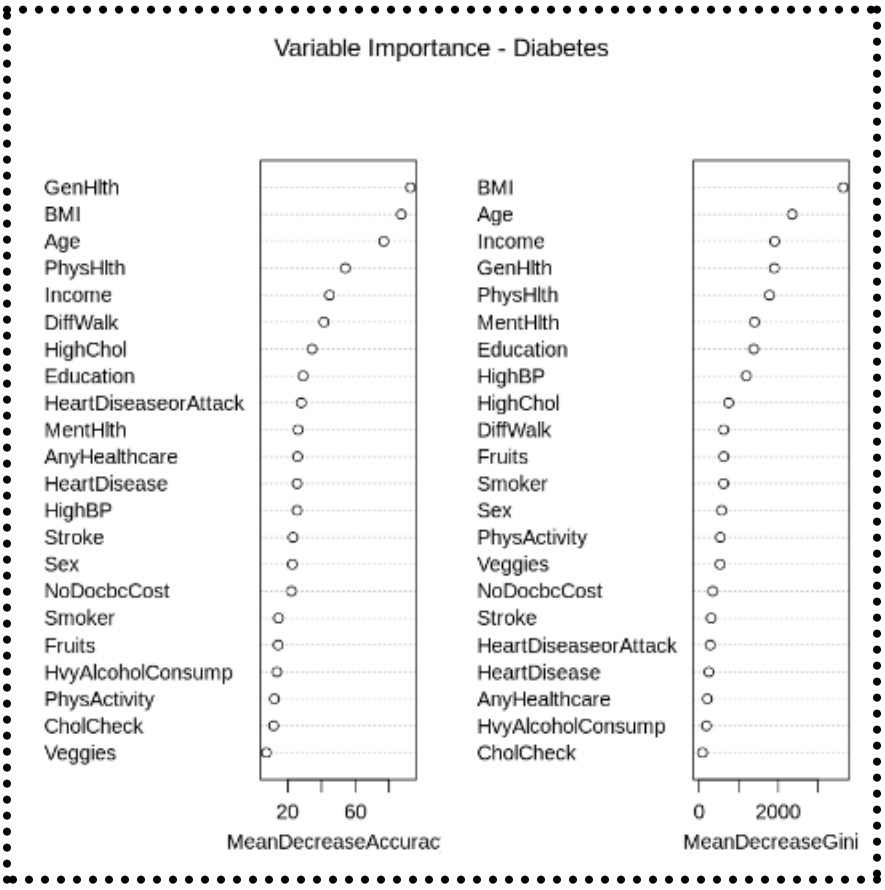
```
error_rates = list(10)
for(i in 1:10) {
  set.seed(i)
  train <- sample(1:nrow(diabetes), (nrow(diabetes) * .8) + .5)
  test <- diabetes[-train,]
  model = randomForest(Diabetes_binary ~ ., data = diabetes,
subset = train, ntree = 500, mtry = sqrt(ncol(diabetes) - 1),
importance = TRUE)
  dia.pred <- predict(model, newdata = test, type = "class")
  error_rates[i] <- mean(dia.pred != test$Diabetes_binary)
}
error_rates = unlist(error_rates)
print(mean(error_rates))
varImpPlot(model, sort = TRUE, main = "Variable Importance -
Diabetes")
```

Results

Logistic regression summary after removing unnecessary values

```
Call:
glm(formula = Diabetes_binary ~ . - Smoker - Fruits - AnyHealthcare -
    NoDocbcCost, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3721  -0.8002  -0.1356   0.8365   2.9787

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error z value             Pr(>|z|)
(Intercept)          -6.874521   0.133632 -51.444 < 0.0000000000000002 ***
HighBP1               0.744242   0.022089  33.693 < 0.0000000000000002 ***
HighChol1             0.605191   0.021105  28.676 < 0.0000000000000002 ***
CholCheck1            1.368989   0.091335  14.989 < 0.0000000000000002 ***
BMI                   0.077007   0.001767  43.576 < 0.0000000000000002 ***
Stroke1               0.152724   0.045829   3.333              0.000861 ***
HeartDiseaseorAttack  0.244152   0.031863   7.663   0.0000000000000182 ***
PhysActivity1        -0.051662   0.023752  -2.175              0.029622 *
Veggies1             -0.084097   0.025617  -3.283              0.001028 **
HvyAlcoholConsump1   -0.752646   0.054726 -13.753 < 0.0000000000000002 ***
GenHlth               0.587604   0.012847  45.739 < 0.0000000000000002 ***
MentHlth             -0.004889   0.001424  -3.433              0.000597 ***
PhysHlth             -0.007454   0.001333  -5.592   0.0000000224143728 ***
DiffWalk1             0.097914   0.028957   3.381              0.000721 ***
Sex1                  0.268617   0.021190  12.677 < 0.0000000000000002 ***
Age                   0.152823   0.004275  35.750 < 0.0000000000000002 ***
Education            -0.041484   0.011390  -3.642              0.000270 ***
Income               -0.054216   0.005759  -9.415 < 0.0000000000000002 ***
HeartDisease1               NA         NA      NA                    NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 78466  on 56600  degrees of freedom
Residual deviance: 57712  on 56583  degrees of freedom
AIC: 57748

Number of Fisher Scoring iterations: 5
```

Random Forest importance graph



The plot above shows that some of the important variables include: **GenHlth, BMI, Age, PhysHlth,** and **Income.**

What variable(s) have the greatest effect on the prediction of diabetes?

- For random forest the greatest effect comes from GenHelth, BMI, Income, Age, and Physhealth.
- For Logistic regression the greatest effect comes from all of them except stroke1, phys health, veggies1, MentHealth, diffWalk, education.

What is the accuracy of the models?

- Test Error Rate of the Logistic Regression(s):

```
> print(lr_error_rates_2)
 [1] 0.1371807 0.1365697 0.1349732 0.1363726 0.1365894 0.1355645 0.1375158 0.1351506 0.1345199 0.1371610
> print(mean(lr_error_rates_2))
[1] 0.1361597
```

- Test Error Rate of the Random Forest(s)

```
> print(error_rates)
 [1] 0.1367865 0.1355448 0.1346973 0.1360572 0.1348155 0.1340665 0.1367865 0.1347564 0.1343228 0.1357222
> mean(error_rates)
[1] 0.1353556
```

Conclusion

When comparing the two models the Random forest takes a slight lead with a lower test error rate. The slight lead is not worth the computational trade off. Logistic regression used all variables besides 4, which are : Smoker, Fruits, AnyHealthcare, NoDocbcCost. Random Forest did not remove any variables. What we have learned from this project is that patients should monitor their BMI and their Gen health the most-this will help them determine if they have diabetes or not. However, it is worth noting that GenHlth is a self-reported variable, making it susceptible to bias and skew based on the perception of the individual participants as well as failing to be quantitative. Thus, despite its importance, it is not necessarily the ideal predictor for people to track within their personal life.

Bibliography

Teboul, Alex. “Diabetes Health Indicators Dataset.” *Kaggle*, 8 Nov. 2021, www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.