# Predictive Modeling for Diabetes Risk Using Machine Learning

Akwawo Ekpu and Aashi Vishnoi

BCH 339N

November 29th, 2023

# Overview

Background

Objective

Dataset

Risk Factors

Comparing Models

Limitations & Improvements

Conclusion

# Background Information

- Diabetes - chronic disease where individuals lose their ability to regulate levels of glucose. There are two types of diabete Type 1 and type 2.
  - Type 1 - the pancreas does not produce insulin
  - Type 2 - the pancreas makes less insulin and body becomes resistant
- Diabetes is a very complicated condition which may be caused by many given factors
- Additionally there is no cure.
- CDC (2018) - 34.2 million individuals in the USA have diabetes and 88 million have prediabetes
- Estimate 8 in 10 pre diabetics are unaware of their risk factors

# Background Information

- Behavioral Risk Factor Surveillance System
  - CDC
  - Health telephone survey
  - Data collection regarding health-related risk behaviors, chronic health conditions, and use of preventative services
  - Completes more than 400,000 adult interviews per year

# Dataset

- Kaggle - 2015

- Original - 441,455 individuals with 330 features
- Used - clean dataset with 253,680 individuals 21 features

- Variable examples
  - Blood pressure
  - Difficulty walking
  - BMI

# Objectives

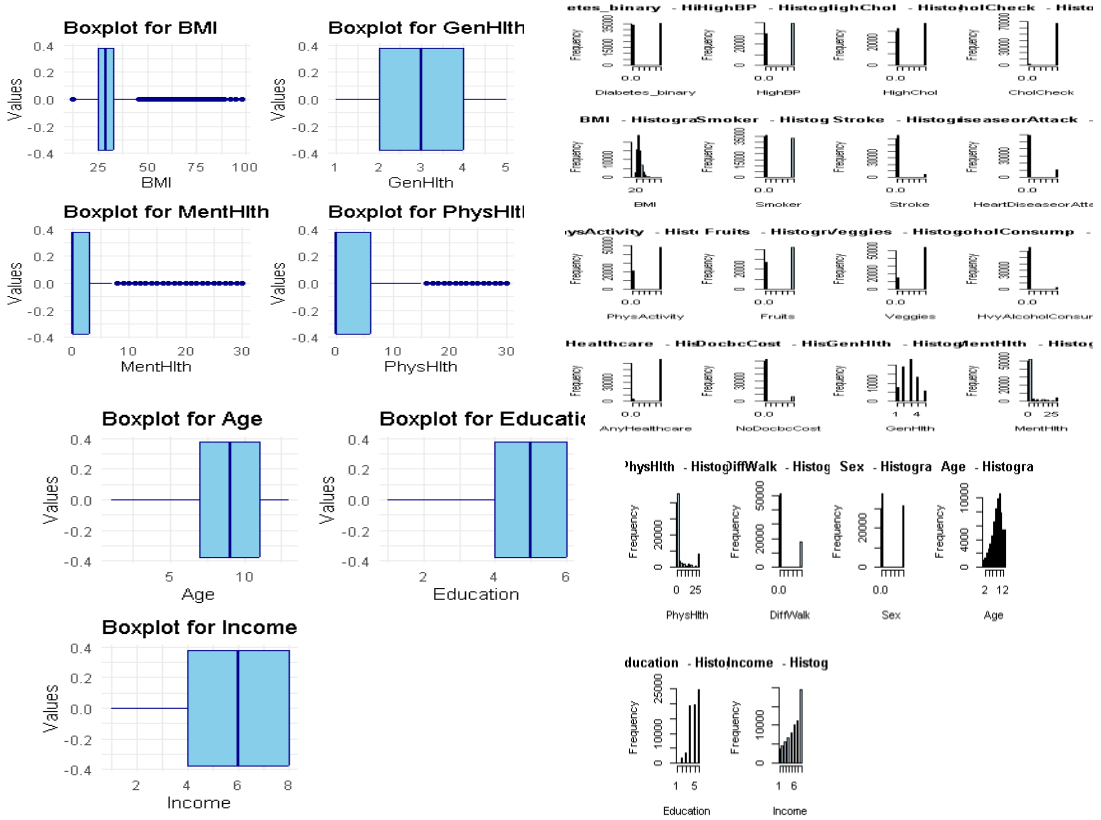
1. Determine which factors are more significant/important to diabetes diagnosis

1. Build a predictive model to determine diabetes diagnosis using various factors

1. Compare models to determine which one is most efficient/effective for diagnosis prediction

# Data Visualization

- No missing values
- Few outliers



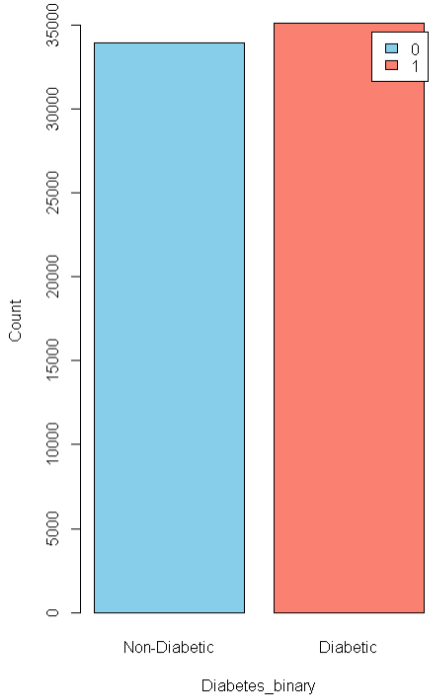
# Target Variable

Outr target variable was Diabetes _binary which basically describes if you have diabetes .

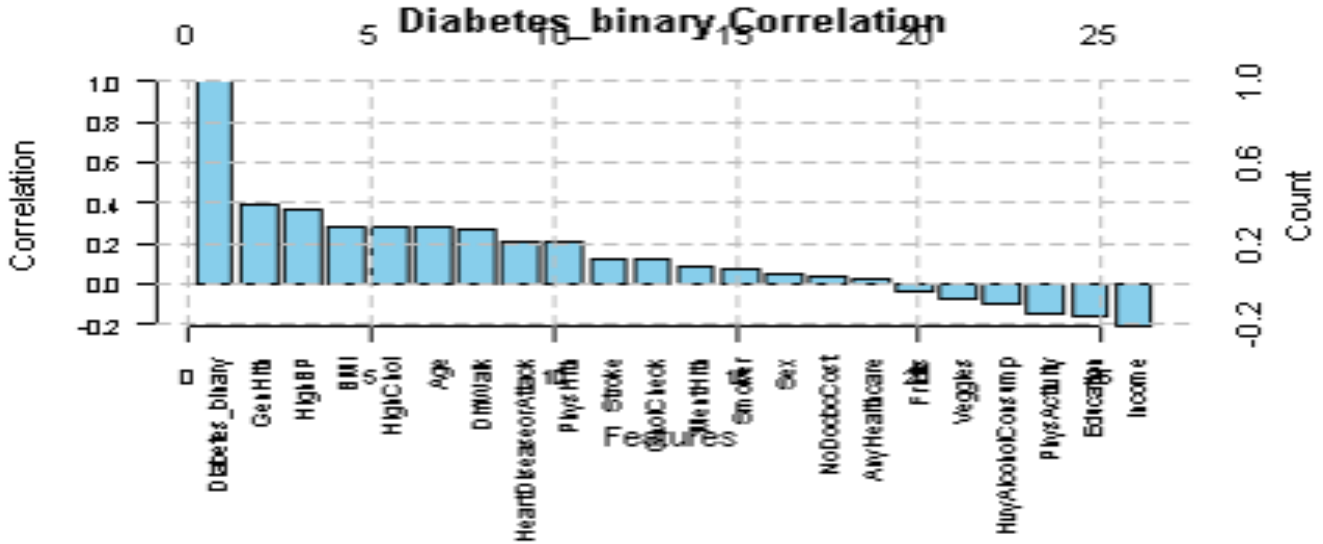
Yes= 1 , No = 0

We saved zeros in Non- Diabetic and the 1 in diabetic ]

Even distribution



# Correlation to Target Variable Diabetes_Binary

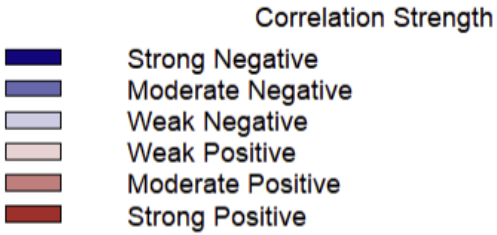


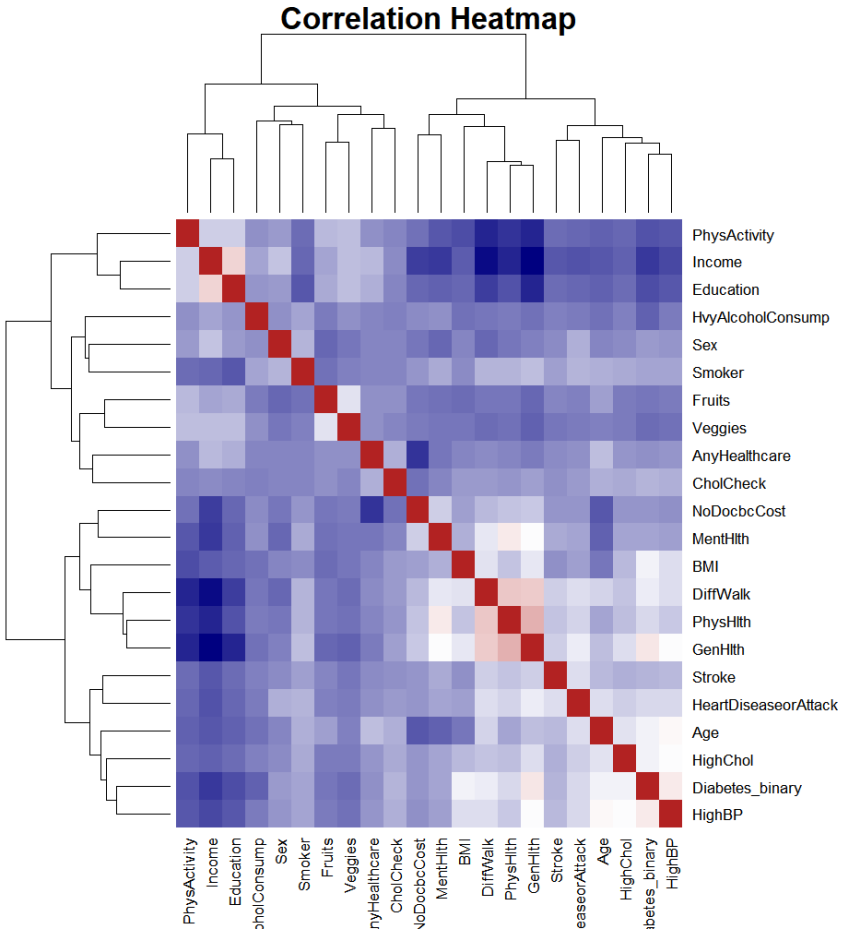
# Correlation Matrix

Partial Snippet

| | Diabetes_binary | HighBP | HighChol | CholCheck | BMI | Smoker | Stroke | HeartDiseaseorAttack | PhysActivity | Fruits | Veggies | HvyAlcoholConsump | AnyHealthcare | NoDocbcCost | GenHlth | MentHlth | PhysHlth | DiffWalk | Sex | Age | Education | In |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| iabetes_binary | 1 | 0.37204754 | 0.281399 | 0.118899983 | 0.28564292 | 0.075853409 | 0.122726777 | 0.20722856 | -0.15028119 | -0.044559815 | -0.07218106 | -0.098709111 | 0.027034114 | 0.036144992 | 0.39657084 | 0.080687506 | 0.206867582 | 0.26708166 | 0.042538131 | 0.274550155 | -0.15852221 | - |
| HighBP | 0.37204754 | 1 | 0.30898713 | 0.10659256 | 0.232372196 | 0.078122541 | 0.126869342 | 0.20677584 | -0.1283068 | -0.031818032 | -0.05982377 | -0.029764035 | 0.039658684 | 0.02180205 | 0.30845865 | 0.058133324 | 0.167820674 | 0.22963847 | 0.03782372 | 0.333720664 | -0.13003721 | |
| HighChol | 0.281399 | 0.30898713 | 1 | 0.088231116 | 0.123917284 | 0.086522244 | 0.098165616 | 0.17820724 | -0.08446875 | -0.040783475 | -0.03780063 | -0.027259301 | 0.034351608 | 0.029975959 | 0.22758797 | 0.079929484 | 0.138265933 | 0.15785903 | 0.013250379 | 0.235778962 | -0.07536401 | - |
| CholCheck | 0.11889998 | 0.10659256 | 0.08823112 | 1 | 0.047778545 | -0.002853773 | 0.023368238 | 0.04479458 | -0.01007231 | 0.015852566 | -0.00103992 | -0.026850049 | 0.106549251 | -0.061974765 | 0.06311625 | -0.009364526 | 0.036441764 | 0.04642052 | -0.008116308 | 0.103413775 | -0.01126586 | |
| BMI | 0.28564292 | 0.2323722 | 0.12391728 | 0.047778545 | 1 | 0.002760648 | 0.019503148 | 0.0553446 | -0.16417877 | -0.076932708 | -0.05016261 | -0.060795053 | -0.010527003 | 0.061860787 | 0.25664174 | 0.09928612 | 0.15566129 | 0.24066721 | -0.002821766 | -0.045129813 | -0.08911209 | - |
| Smoker | 0.07585341 | 0.07812254 | 0.08652224 | -0.002853773 | 0.002760648 | 1 | 0.061957309 | 0.12045681 | -0.0724006 | -0.068191646 | -0.02375952 | 0.076393576 | -0.010227542 | 0.031895633 | 0.14065842 | 0.086354285 | 0.114730414 | 0.11371344 | 0.113422187 | 0.099699128 | -0.13078973 | - |
| Stroke | 0.12272678 | 0.12686934 | 0.09816562 | 0.023368238 | 0.019503148 | 0.061957309 | 1 | 0.22206208 | -0.07677079 | -0.005810653 | -0.04486947 | -0.024496003 | 0.007801144 | 0.034304777 | 0.18653747 | 0.084799695 | 0.161823506 | 0.18971436 | 0.004149474 | 0.123344059 | -0.06960165 | - |
| iseaseorAttack | 0.20722856 | 0.20677584 | 0.17820724 | 0.044794583 | 0.055344602 | 0.120456813 | 0.222062075 | 1 | -0.09385845 | -0.014931129 | -0.03232705 | -0.038745479 | 0.017603433 | 0.033397166 | 0.2715016 | 0.071530424 | 0.194963254 | 0.22918819 | 0.099019737 | 0.220789811 | -0.09040325 | - |
| PhysActivity | -0.15028119 | -0.1283068 | -0.08446875 | -0.010072315 | -0.164178766 | -0.072400598 | -0.076770793 | -0.09385845 | 1 | 0.127578055 | 0.14339159 | 0.021623808 | 0.02416824 | -0.059078733 | -0.26414228 | -0.1245348 | -0.228329375 | -0.27098753 | 0.052068899 | -0.097456433 | 0.18015934 | |
| Fruits | -0.04455981 | -0.03181803 | -0.04078347 | 0.015852566 | -0.076932708 | -0.068191646 | -0.005810653 | -0.01493113 | 0.12757806 | 1 | 0.2345047 | -0.031517781 | 0.026963877 | -0.042214604 | -0.08683569 | -0.057019114 | -0.041835956 | -0.04401689 | -0.088016697 | 0.066047646 | 0.089283 | |
| Veggies | -0.07218106 | -0.05982377 | -0.03780063 | -0.00103992 | -0.050162611 | -0.023759524 | -0.04486947 | -0.03232705 | 0.14339159 | 0.234504703 | 1 | 0.024104098 | 0.02683206 | -0.033643343 | -0.10613637 | -0.047458146 | -0.060874884 | -0.07809761 | -0.053422437 | -0.015761849 | 0.14475546 | |
| :oholConsump | -0.09870911 | -0.02976403 | -0.0272593 | -0.026850049 | -0.060795053 | 0.076393576 | -0.024496003 | -0.03874548 | 0.02162381 | -0.031517781 | 0.0241041 | 1 | -0.012691446 | 0.008452984 | -0.06370548 | 0.013913711 | -0.038739073 | -0.05189367 | 0.015436713 | -0.059553272 | 0.03997057 | |
| lnyHealthcare | 0.02703411 | 0.03965868 | 0.03435161 | 0.106549251 | -0.010527003 | -0.010227542 | 0.007801144 | 0.01760343 | 0.02416824 | 0.026963877 | 0.02683206 | -0.012691446 | 1 | -0.22045118 | -0.02847723 | -0.047688859 | -0.000416106 | 0.01106594 | -0.006804284 | 0.139837853 | 0.10356353 | |
| NoDocbcCost | 0.03614499 | 0.02180205 | 0.02997596 | -0.061974765 | 0.061860787 | 0.031895633 | 0.034304777 | 0.03339717 | -0.05907873 | -0.042214604 | -0.03364334 | 0.008452984 | -0.22045118 | 1 | 0.16475786 | 0.191107641 | 0.153951534 | 0.12341486 | -0.04846917 | -0.133526667 | -0.09167795 | |
| GenHlth | 0.39657084 | 0.30845865 | 0.22758797 | 0.063116249 | 0.256641742 | 0.14065819 | 0.186537468 | 0.2715016 | -0.26414228 | -0.086835685 | -0.10613637 | -0.063705482 | -0.028477229 | 0.164757862 | 1 | 0.310093325 | 0.550137724 | 0.472338 | -0.016880197 | 0.149215545 | -0.27103728 | - |
| MentHlth | 0.08068751 | 0.05813332 | 0.07992948 | -0.009364526 | 0.09928612 | 0.086354285 | 0.084799695 | 0.07153042 | -0.1245348 | -0.057019114 | -0.04745815 | 0.013913711 | -0.04768886 | 0.191107641 | 0.31009333 | 1 | 0.37662536 | 0.2469476 | -0.08992649 | -0.106157606 | -0.09923656 | - |
| PhysHlth | 0.20686758 | 0.16782067 | 0.13826593 | 0.036441764 | 0.15566129 | 0.114730414 | 0.161823506 | 0.19496325 | -0.22832938 | -0.041835956 | -0.06087488 | -0.038739072 | -0.000416106 | 0.153951534 | 0.55013772 | 0.37662536 | 1 | 0.48409231 | -0.045928755 | 0.08184705 | -0.15010464 | - |
| DiffWalk | 0.26708166 | 0.22963847 | 0.15785903 | 0.046420522 | 0.240667214 | 0.113713445 | 0.189714359 | 0.22918819 | -0.27098753 | -0.044016887 | -0.07809761 | -0.051893666 | 0.011065935 | 0.123414864 | 0.472338 | 0.246947596 | 0.484092309 | 1 | -0.08285825 | 0.193668605 | -0.19386946 | - |
| Sex | 0.04253813 | 0.03782372 | 0.01325038 | -0.008116308 | -0.002821766 | 0.113422187 | 0.004149474 | 0.09901974 | 0.0520689 | -0.088016697 | -0.05342244 | 0.015436713 | -0.006804284 | -0.04846917 | -0.0168802 | -0.08992649 | -0.045928755 | -0.08285825 | 1 | -0.004754655 | 0.04296438 | |
| Age | 0.27455015 | 0.33372066 | 0.23577896 | 0.103413775 | -0.045129813 | 0.099699128 | 0.123344059 | 0.22078981 | -0.09745643 | 0.066047646 | -0.01576185 | -0.059553272 | 0.139837853 | -0.133526667 | 0.14921554 | -0.106157606 | 0.081847051 | 0.1936686 | -0.004754655 | 1 | -0.10177365 | - |
| Education | -0.15852221 | -0.13003721 | -0.07536401 | -0.011265862 | -0.089112087 | -0.130789729 | -0.069601651 | -0.09040325 | 0.18015934 | 0.089283002 | 0.14475546 | 0.03997057 | 0.103563528 | -0.091677947 | -0.27103728 | -0.099236558 | -0.150104636 | -0.19386946 | 0.042964381 | -0.101773649 | 1 | |
| Income | -0.2128456 | -0.17636028 | -0.09871182 | 0.005067423 | -0.113706233 | -0.093897423 | -0.132637329 | -0.14070159 | 0.18605247 | 0.068602795 | 0.14644161 | 0.068505365 | 0.127422128 | -0.193669498 | -0.3702036 | -0.212170036 | -0.271148247 | -0.33581877 | 0.161565945 | -0.124261452 | 0.45037598 | |

# Heatmap

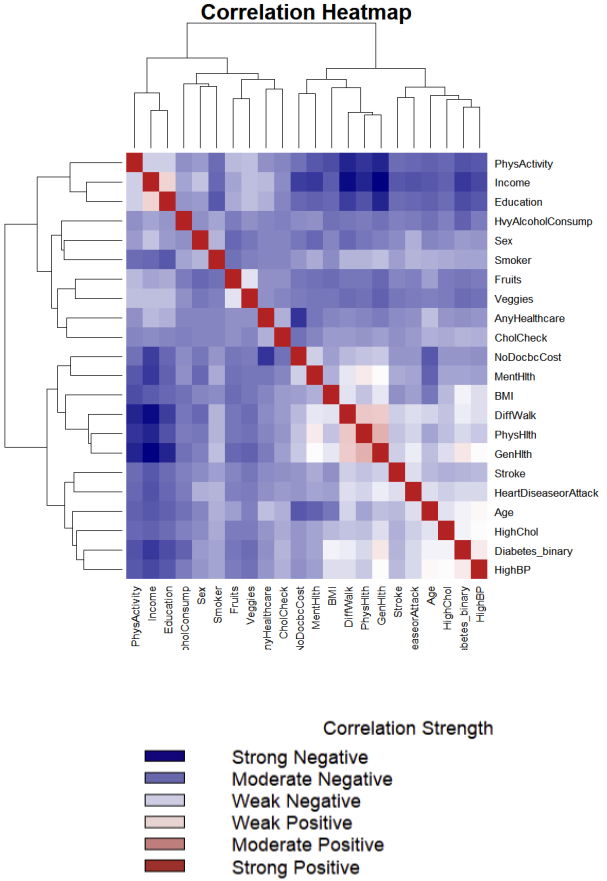

Correlation Strength
Strong Negative
Moderate Negative
Weak Negative
Weak Positive
Moderate Positive
Strong Positive


Correlation Heatmap
PhysActivity
Income
Education
HvyAlcoholConsump
Sex
Smoker
Fruits
Veggies
AnyHealthcare
CholCheck
NoDocbcCost
MentHlth
BMI
DiffWalk
PhysHlth
GenHlth
Stroke
HeartDiseaseorAttack
Age
HighChol
Diabetes_binary
HighBP

# Heat Map Interpretation

- Positive Correlation
  - Stroke
  - Heart Disease or Attack
  - Difficulty Walking
  - Sex
  - Age

- Negative Correlation
  - Income and general health



# Logistic Regression Model - Greatest Risk Factors

- We used the logistic regression model because our outcome is binary in nature
- More likely to be associated with the likelihood of diabetes
- p<0.05 = significant
    - BMI, GenHlth, High Blood Pressure , Heart Disease or Attack, Age
- Insignificant predictors in model
    - Smoker, Physical Activity

```
Coefficients:
                        Estimate Std. Error z value
(Intercept)           -6.739557   0.123869 -54.409
HighBP                 0.723653   0.019812  36.527
HighChol               0.577544   0.018937  30.499
CholCheck              1.360819   0.081034  16.793
BMI                    0.074834   0.001572  47.601
Smoker                -0.013908   0.018922  -0.735
Stroke                 0.157097   0.040809   3.850
HeartDiseaseorAttack   0.251768   0.028423   8.858
PhysActivity          -0.021339   0.021219  -1.006
Fruits                -0.028009   0.019609  -1.428
Veggies               -0.058871   0.023270  -2.530
HvyAlcoholConsump     -0.750961   0.048604 -15.451
AnyHealthcare          0.055314   0.046930   1.179
NoDocbcCost            0.008320   0.033944   0.245
GenHlth                0.545746   0.010773  50.661
MentHlth              -0.006343   0.001249  -5.079
DiffWalk               0.078005   0.024909   3.132
Sex                    0.265308   0.019232  13.795
Age                    0.150042   0.003909  38.386
Education             -0.035572   0.010208  -3.485
Income                -0.056434   0.005184 -10.887
```

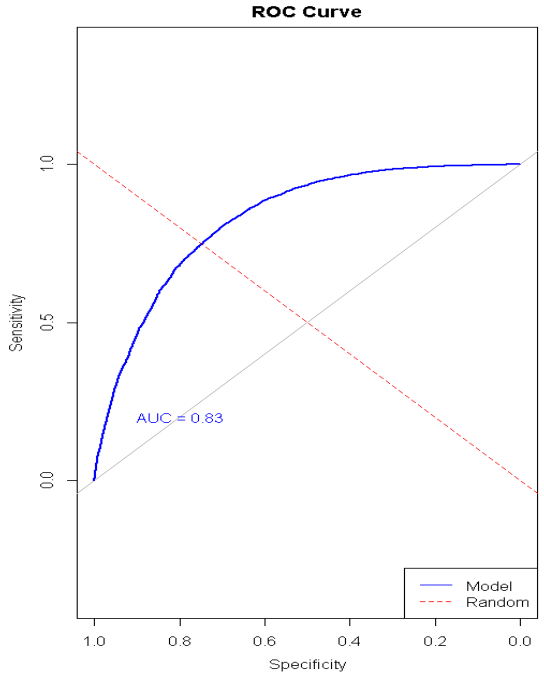
# ROC/AUC Plot

Visualization of the performance of  our logistic regression model

Given that :

1. The curve rises steeply towards  hugs the upper left corner, indicating high sensitivity and low false positive rate.
2. 1>AUC > 0.5
3. The accuracy of our model came out to be about 0.83

We concluded that the accuracy was good



# Random Forest

This learning method that is used for both classification and regression tasks. We choice this method:

- Can be used for both classification and regression task
- It design for predicting categorical outcomes
- Less prone to overfitting

Accuracy = 0.7471742

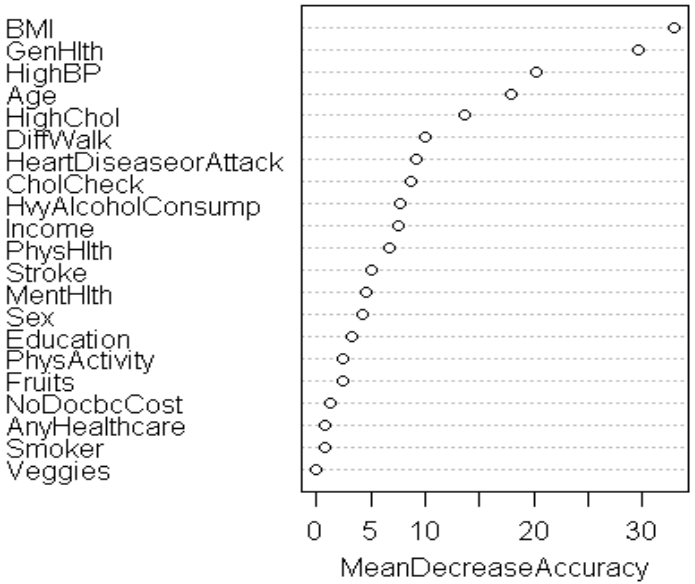
```
Call:
 randomForest(formula = Diabetes_binary ~ ., data =
training,      mtry = 4, ntree = 501, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 501
No. of variables tried at each split: 4

        OOB estimate of  error rate: 25.23%
Confusion matrix:
      0     1 class.error
0 20218  7989   0.2832276
1  6281 22065   0.2215833
```

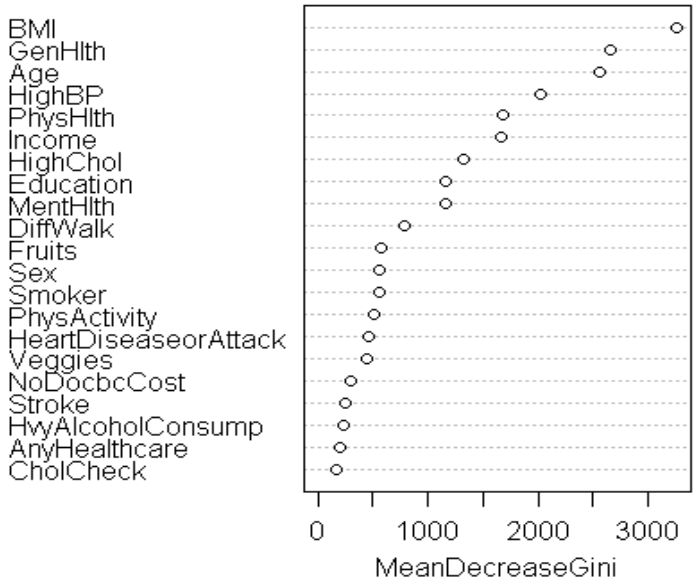
# Random Forest

| | 0 | 1 | MeanDecreaseAccuracy |
|---|---|---|---|
| HighBP | 14.7625264 | 17.34720036 | 20.1942976 |
| HighChol | 10.0031377 | 9.96206268 | 13.5642978 |
| CholCheck | 1.5888433 | 6.36690262 | 8.6319233 |
| BMI | 15.5602050 | 14.24871052 | 32.9674506 |
| Smoker | 2.6375022 | -1.63476957 | 0.7495181 |
| Stroke | 5.9162228 | -0.01656464 | 5.1002642 |
| HeartDiseaseorAttack | 7.3264700 | 2.68403257 | 9.2348282 |
| PhysActivity | 4.5625771 | 0.12150768 | 2.4340969 |
| Fruits | 3.9622258 | -0.95959377 | 2.4229055 |
| Veggies | 0.8587002 | -1.01316123 | -0.1156270 |
| HvyAlcoholConsump | 3.2069940 | 6.79626300 | 7.6781600 |
| AnyHealthcare | 2.0932391 | -0.79772483 | 0.7578497 |
| NoDocbcCost | 2.3505565 | -0.26410599 | 1.2828409 |
| GenHlth | 18.8792728 | 21.57448406 | 29.5607726 |
| MentHlth | 1.4719074 | 0.97060179 | 4.4996397 |
| PhysHlth | 8.7540357 | 1.24212953 | 6.6573000 |
| DiffWalk | 8.4854327 | 2.78752949 | 10.0402167 |
| Sex | 1.6939199 | 3.01587853 | 4.2600519 |
| Age | 10.9509357 | 21.53901797 | 17.8502709 |
| Education | 2.7835636 | 2.63897772 | 3.2595897 |
| Income | 5.8377532 | 1.99028658 | 7.4804251 |

| | MeanDecreaseGini |
|---|---|
| HighBP | 2023.6853 |
| HighChol | 1323.3360 |
| CholCheck | 159.9370 |
| BMI | 3265.0010 |
| Smoker | 565.5910 |
| Stroke | 244.6540 |
| HeartDiseaseorAttack | 463.1931 |
| PhysActivity | 510.6787 |
| Fruits | 572.1854 |
| Veggies | 451.5815 |
| HvyAlcoholConsump | 233.3037 |
| AnyHealthcare | 193.9032 |
| NoDocbcCost | 297.7510 |
| GenHlth | 2655.2716 |
| MentHlth | 1159.4606 |
| PhysHlth | 1688.5708 |
| DiffWalk | 784.1230 |
| Sex | 565.6392 |
| Age | 2568.4840 |
| Education | 1167.0194 |
| Income | 1663.5491 |

# Random Forest : Feature Importance



# Comparing Models

| Models | Accuracy Score |
| --- | --- |
| Logistic Regression | 0.83 |
| Random Forest | 0.7471742 |

Best performing model was Logistic Regression with a score of 0.83

# Limitations & Improvements

Limitations

- Selection bias - telephone interviews
- Self reported data by individuals
- Class imbalances

Improvements for Future Iterations

- Increasing the number of trees
- More representative data collection
- Better handling of class imbalances
- Introduce weighted classes - support underrepresented groups
- Test scalability across years

# Conclusion

Significant factors that influence diabetes in the models:

- BMI
- General Health
- High Blood Pressure
- Age

Best performing/highest scoring model was Logistic Regression with a score of 0.83

# References


1. Centers for Disease Control and Prevention. (2023, September 5). What Is Diabetes? Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/diabetes.html
2. “Introduction to Random Forest in R.” Simplilearn.com, www.simplilearn.com/tutorials/data-science-tutorial/random-forest-in-r.
3. UVA Health. (n.d.). Types of Diabetes. Retrieved from https://uvahealth.com/services/diabetes-care/types
4. Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021, December 20). *Machine learning and deep learning predictive models for type 2 diabetes: A systematic review - diabetology & metabolic syndrome*. BioMed Central. https://dmsjournal.biomedcentral.com/articles/10.1186/s13098-021-00767-9
5. Nayak, L. (2022, March 22). *Predicting diabetes with random forest classifier*. Medium. https://towardsdatascience.com/predicting-diabetes-with-random-forest-classifier-c62f2e319c6e
6. *What is logistic regression?*. IBM. (n.d.). https://www.ibm.com/topics/logistic-regression