**To:** Professor Yu Du

**From:** Paul Draper,Hung Nguyen, Ramya Meduri

**Date:** 5/4/2024

Diabetes is one of the most common chronic diseases in the US, impacting millions. It results in severe complications such as heart disease, vision loss, lower-limb amputation, and kidney disease. The CDC estimates that 1 in 5 diabetics, and roughly 8 in 10 pre-diabetics, are unaware of their risk. Diabetes also places a massive burden on the economy, with undiagnosed diabetes approaching $400 billion annually. This project aims to predict whether an individual has diabetes or is at high risk of developing it. The accurate identification of individuals at risk is crucial for patient risk assessment, insurance underwriting, and pricing. Additionally, these predictions contribute to the development of effective long-term public health policy strategies for diabetes prevention and management.

**EXECUTIVE SUMMARY**

**Major Findings**

1. Initial EDA showed a positive correlation between being diabetic and BMI, stroke, heart disease, high blood pressure, high cholesterol and age. These proved to be among the top predictors in our model (Appendix a). In contrast, non-health related factors such as Income and Education were poor predictors of Diabetes.
2. The final model included 13 predictors to classify an individual as being diabetic or not. The variables selected were high blood pressure, high cholesterol, cholesterol check in the last 5 years, BMI, history of stroke, history of heart disease, daily vegetable consumption, heavy drinker, self-given general health score (1-5), days of physical activity over the last 30 days, serious difficulty walking, sex, and age (Appendix b).
3. The overall model accuracy at correctly identifying those with and without diabetes was 74.3% (Appendix c). Precision and Recall were 73.2% and 76.8% respectively indicating the model produced fewer false negatives than false positives. Subsequent models using other advanced machine learning algorithms, Random Forest, K-Nearest Neighbor, and Lasso regression, yielded similar results (Appendix d).
4. Additional factors not in the dataset should be evaluated to more accurately identify Diabetes. Additional survey questions on family history and diet would be recommended and research shows that implementing more clinical predictors like triglycerides and glucose levels could also potentially improve Diabetes predictions.

**Analytical Overview**

The dataset used for our analysis, found [here,](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv) is from Kaggle and was originally sourced by the Centers for Disease Control and Prevention (CDC). It is generated via the Behavioral Risk Factor Surveillance System which surveys respondents over the phone and is collected annually. The data, collected in 2015, contains 70,692 survey responses with 21 attributes and is a 50/50 split between those with and without a diabetes diagnosis.

We started the project with data cleaning by looking for missing values, duplicates, and outliers that could be the result of input errors. Exploratory data analysis was then conducted to evaluate the distribution of the data and identify any apparent relationships with the variables to get an overall better understanding of the dataset and begin hypothesizing about the relationships and most important predictors.

Before beginning model development, the response variable and the binary and multi-category variables needed to be converted to factors. The initial model was built with all predictors included using the caret package with cross-validation to give an idea of what the most important predictors are. The data was then split into training and testing data sets (80/20 split) before further model development. To refine our model and aid in predictor selection, we employed the use of several R functions such as regsubsets(), bestglm(), stepAIC, and varIMP() (appendix e), which yielded our final 13 predictor variables to include in the model.

Model evaluation/validation was performed by ensuring the assumptions of logistic regression were met, deviance test, cross-validation, ROC curve analysis, comparing AIC scores and using several metrics to evaluate performance (appendix f). The final step in our analysis was to utilize other advanced machine learning algorithms to compare results and conclude our selected model was the best performing.

**Conclusion**

This project represents a comprehensive analytical approach to diabetes diagnosis. The logistic regression model offers a baseline interpretation of the relationship between health features and the target variable. Other machine learning algorithms build upon that and enhance prediction accuracy and robustness. The findings from the logistic regression model, Lasso Regression, Random Forest, K-Nearest Neighbor underline the importance of health and lifestyle attributes in assessing diabetes risk. Factors such as BMI, Age, High Blood Pressure, High Cholesterol, and General Health come up across different models as consistent predictors, highlighting their significance in determining diabetes risk. Although we had challenges in refining the model’s accuracy markedly, we believe our work yields valuable interpretations and insights that can be applied in future diabetes-related projects.

**Appendix**

Appendix a:

A graph of bmi

Description automatically generatedA graph of cholesterol response

Description automatically generatedA diagram of diabetes

Description automatically generated

A graph with a green rectangle

Description automatically generated

A graph of disease

Description automatically generated

A graph of a number of people with diabetes

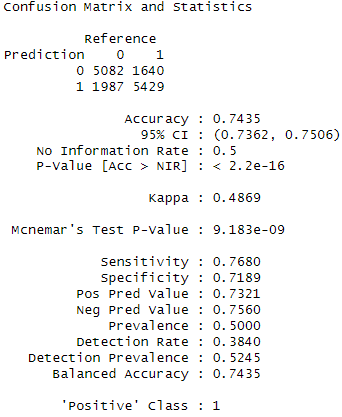
Description automatically generated

Appendix b.

A screenshot of a computer code

Description automatically generated

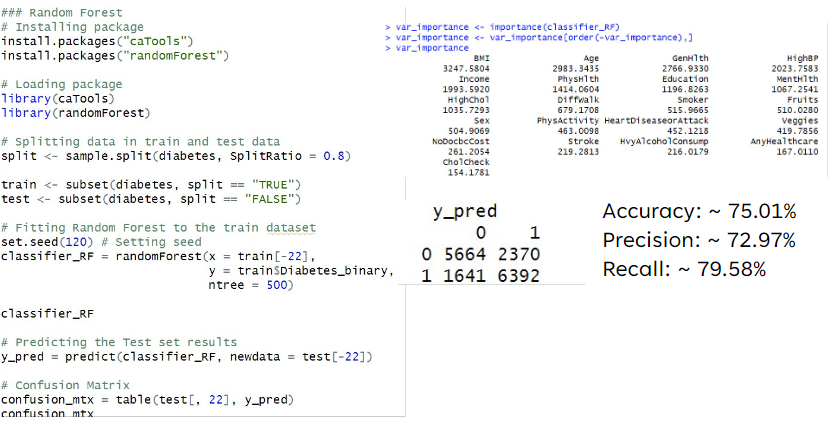
Appendix c.

A diagram of a pie chart

Description automatically generated

Appendix d.

**Random Forest:**



**K-Nearest Neighbor:**

A computer code with black text

Description automatically generatedA number on a white background

Description automatically generated

**Lasso Regression:**

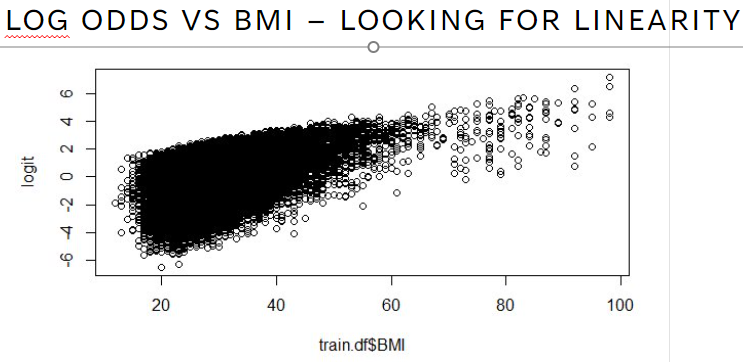
A computer screen shot of a computer code

Description automatically generated**73% Prediction Accuracy**

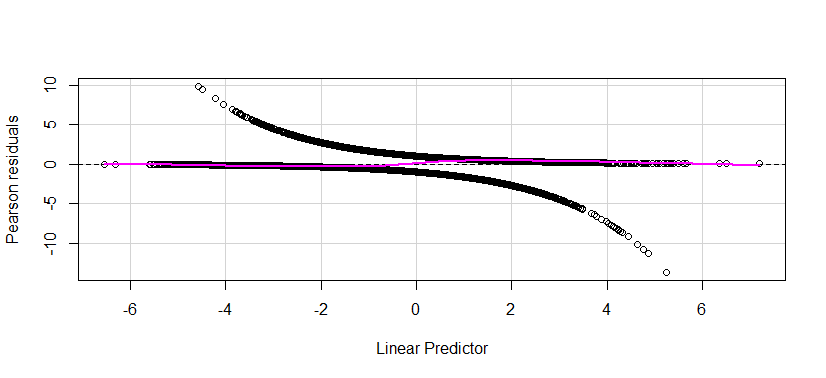
Appendix e.



Appendix f.



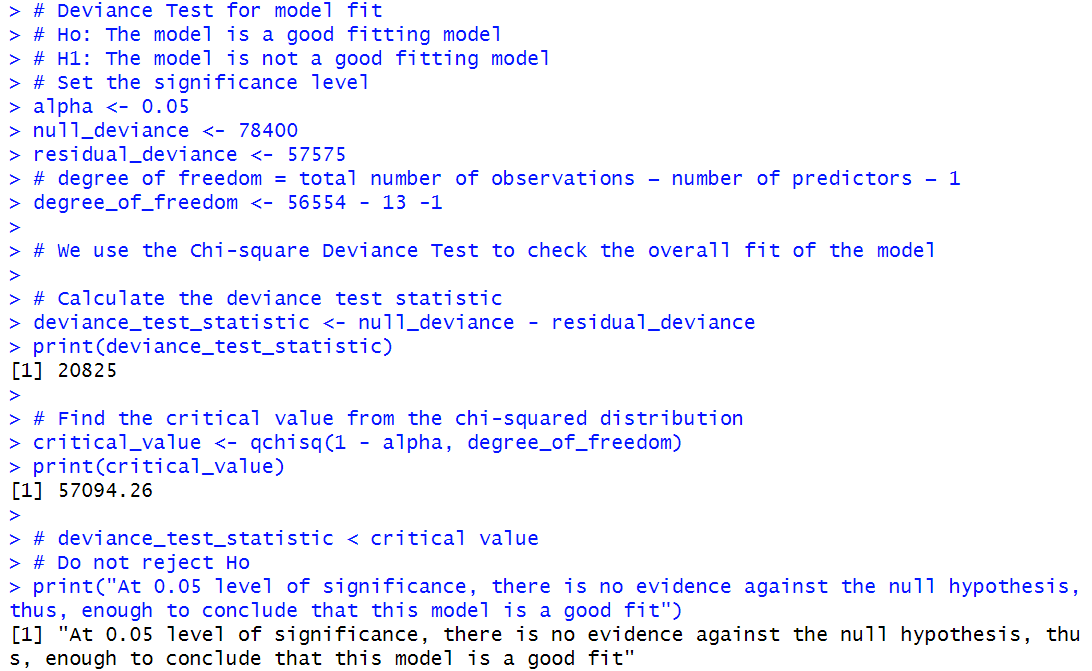
A screenshot of a graph

Description automatically generated

A graph with a blue line

Description automatically generated

**Deviance Test**



**5-Fold Cross Validation**

With the initial full model

A close up of numbers

Description automatically generated

With the final model post feature selection

A close-up of a text

Description automatically generated The smallest prediction error

With the final model - ‘Age’



**References**

<http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/>

<https://www.datacamp.com/tutorial/k-nearest-neighbors-knn-classification-with-r-tutorial>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4885900/#:~:text=Summary%20statistics%20based%20on%20Pearson,helpful%20in%20identifying%20non%2Dlinearity>.

<https://zief0002.github.io/bespectacled-antelope/notes/11-logistic-regression.html>

<https://www.statology.org/partial-residual-plot-in-r/>

<https://builtin.com/data-science/precision-and-recall>

ChatGPT