Using Data & Diagnostics for Obesity Prediction

By Group 4L:

Daniel Jenkins, Teshinee Kukamjad,
John Sohn, Kwon Gyeong Min, Erika Yiu

Table of Contents

01 Background

02 Data Handling

03 Basic Models

04 Final Model

05 Takeaways

06 Conclusion

Background

Goal

- Predict if a person has obesity based on certain traits

Method

Simple Logistic Regression
K-Nearest Neighbors
Random Forest

Relevance

- Obesity is a dangerous level of weight classification
- Over 100 million Americans classified as such
- Linked with cardiovascular diseases like diabetes
- Decrease of life expectancy by 3 years

Data Cleaning

Missing Value Imputation

- About **8%** of all observations had missing values
- For 11 numeric variables: use **MEAN**
- For 19 categorical variables: factor & use **MODE**

Significant Predictors

FAF: Physical Activity Frequency	**CALC:** Caloric intake
CH2O: Daily Water Intake	**SCC:** Consumption of sweet drinks
FCVC: Frequency of Vegetable Consumption	**CAEC:** Consumption of food between meals
NCP: Number of main meals	**FAVC:** Frequent consumption of high-caloric food
Height	**Gender**
Age	**Family_history_with_overweight**

Simple Logistic Regression

Benefits

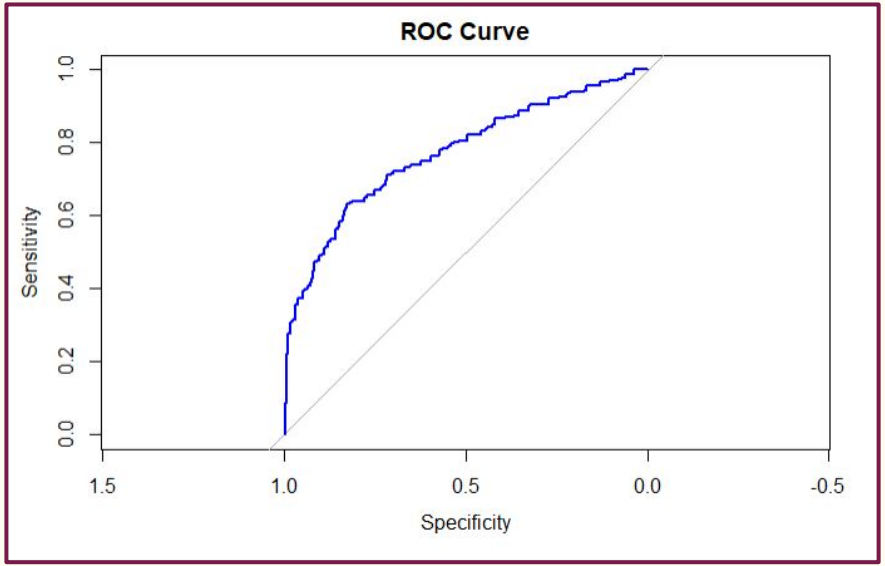
- Simple
- Quick
- Easily Interpretable

Method

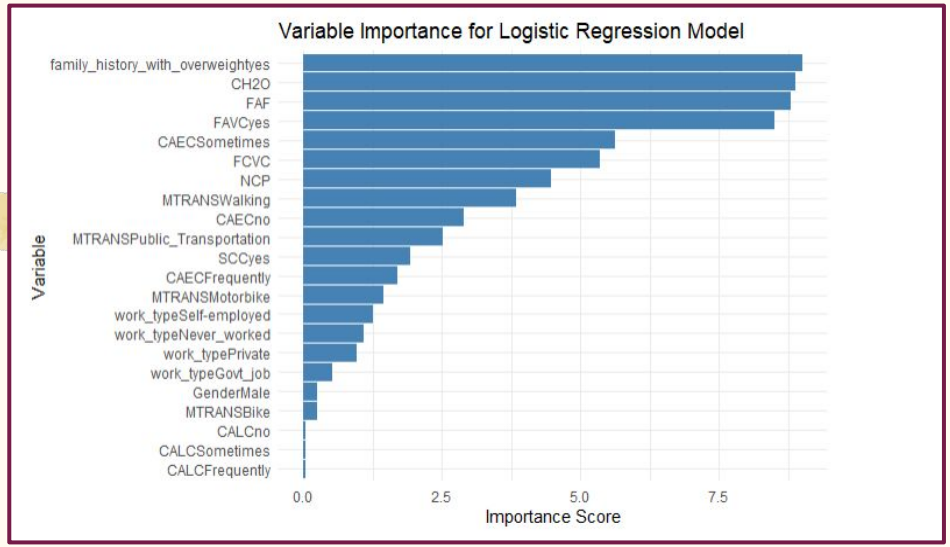
- Generalized linear model function
- Binomial family
- Response variable Obesity Status regressed on significant predictors

Result

- Moderately High AUC
- Sensitivity - Specificity Trade Off
- Decent job at classification



Logistic Regression Interpretation



Significant Variables

1. Family History with Overweight

2. CH2O

3. FAF

Diagnostics

Accuracy = 74.74%

Actual Class \ Predicted Class	Not Obese	Obese
Not Obese	15,902	3,498
Obese	4,591	8,034

K-Nearest Neighbors

Benefits

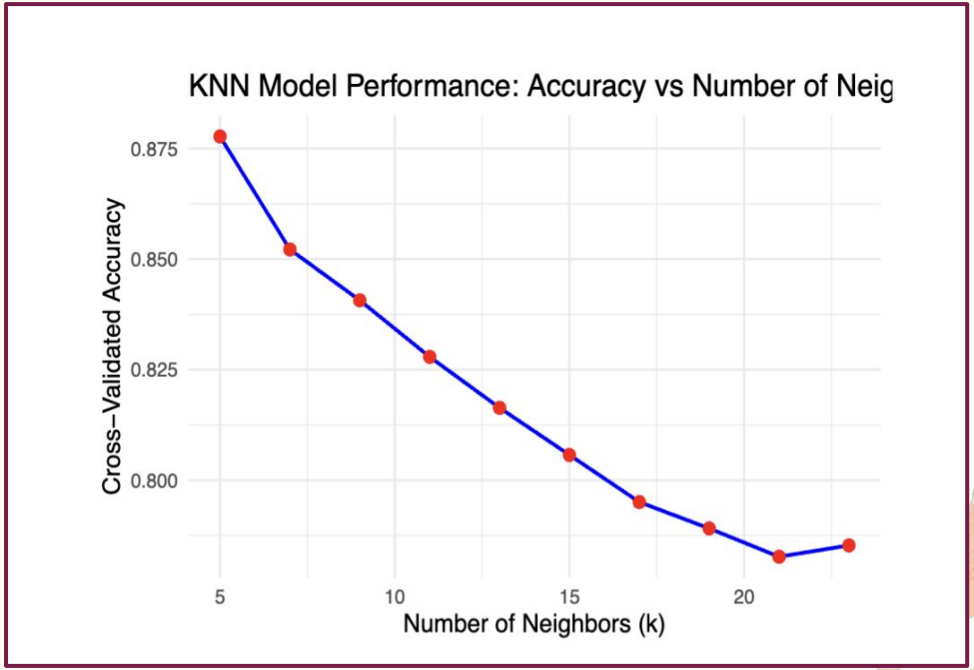
- Intuitive
- Non-Parametric
- Versatile

Method

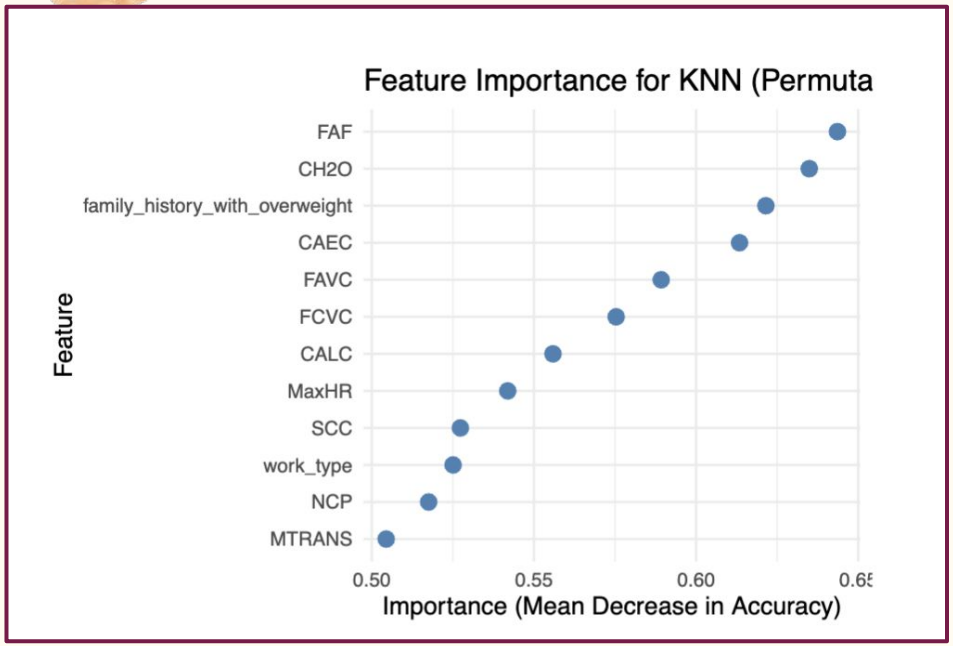
- Cross-validation from k = 5 to 23
- Accuracy primary metric

Result

- Highest accuracy at k = 5
- Accuracy declined with k increase
- Importance of optimizing k for bias and variance balance



KNN Interpretation



Significant Variables

1. FAF

2. CH2O

3. Family History with Overweight

Diagnostics

Training Accuracy = 92.84%

		Predicted Class: Not Obese	Predicted Class: Obese
Actual Class	Not Obese	18,101	1,009
Actual Class	Obese	1,282	11,622

Random Forest

Benefits

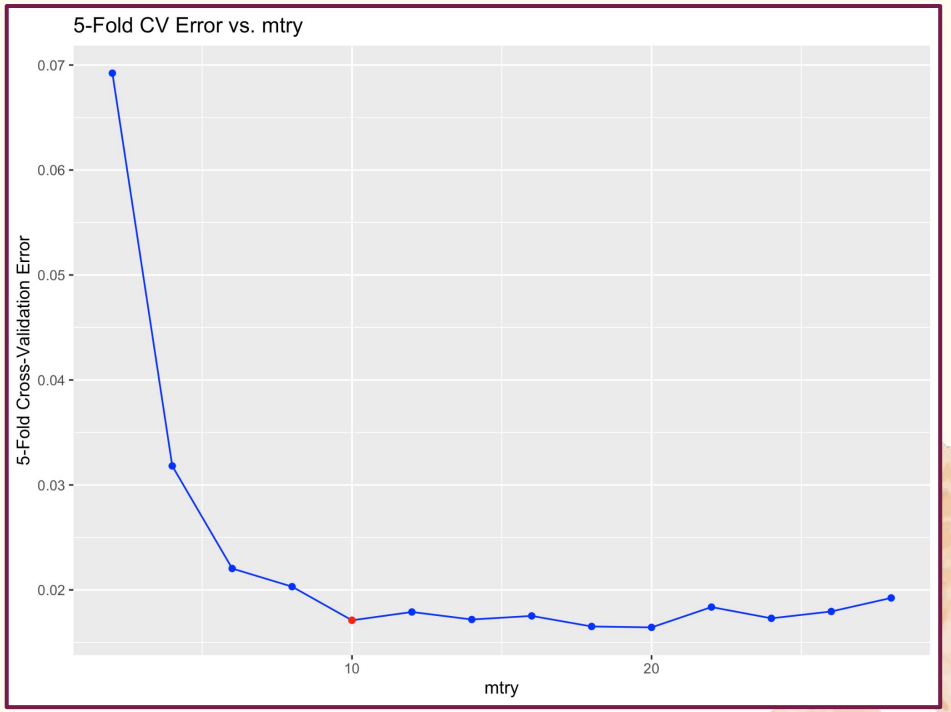
- Improved Accuracy
- Feature Importance

Method

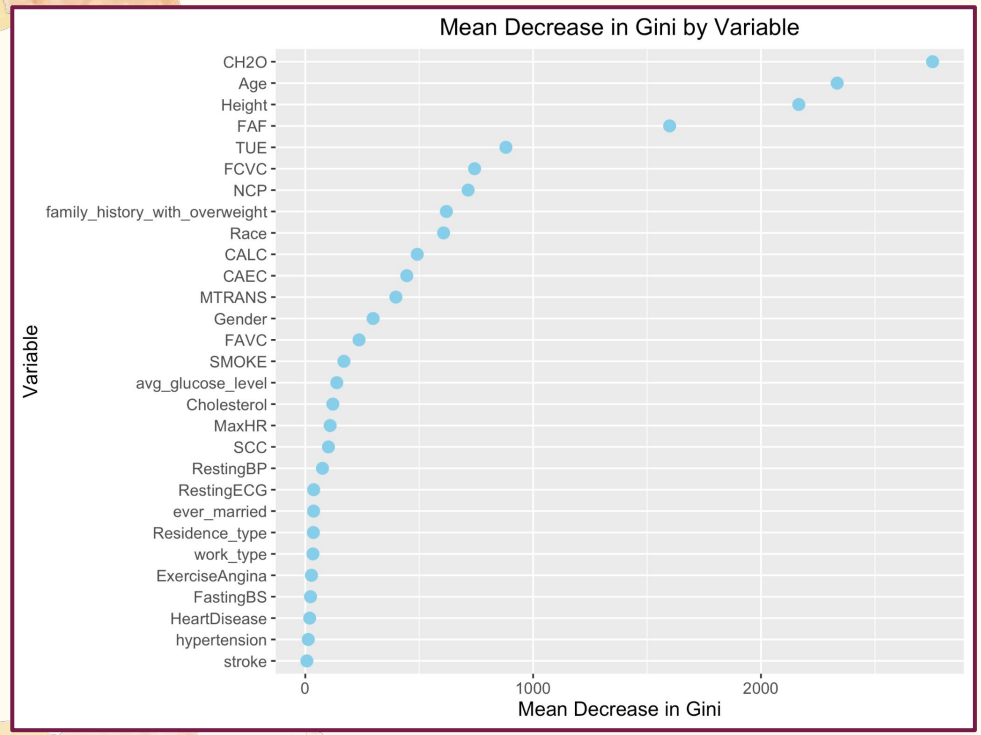
- Bootstrap sampling
- Mtry: Range of number of variables considered at each split from 2 to 29
- Fixed at 5 trees

Result

- Validation Error decreased with mtry increase
- Slowed at mtry = 10, used to balance complexity and performance



Random Forest Interpretation



Significant Variables

1. CH20

2. Age

3. Height

Diagnostics

Training Accuracy = 99.9%

Actual Class \ Predicted Class	Not Obese	Obese
Not Obese	19,524	30
Obese	7	12,453

Improved Model

Best Predictors

- CH2O, FAF, Family History with Overweight, Age, Height

Best Model

- Random Forest (mtry = 4, trees = 5)

Diagnostics

- **Training Accuracy =** 98.2%
- **Testing Accuracy =** 93.4%

Actual Class \ Predicted Class	Not Obese	Obese
Not Obese	19,331	388
Obese	200	12,095

Takeaways

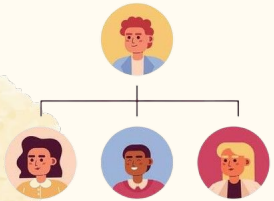
Model Comparison

- **Logistic Regression:** Fast and easy, low accuracy
- **KNN:** Moderate speed and setup, improved accuracy
- **Random Forest:** Slowest and most complicated, high accuracy

Predictor Interpretation

- **CH2O:** Appetite controller, calorie replacement, metabolism boost
- **FAF:** Reduces fat accumulation, calorie burn
- **Family History:** Genetic predisposition, shared lifestyle
- **Age:** Metabolic slowdown, decreased activity level, increase fat mass
- **Height:** Shortness increases BMI





Conclusion

Goal

- Able to predict obesity of person using age, height, family history with overweight, water intake, and physical activity

Possible Improvements

- **Feature Transformation** (Interaction)
- **Advanced Models** (XGBoost)

Future Questions

- How can generative AI models improve obesity prediction?
- Can this model be adapted for real-life healthcare settings?

Thank You!