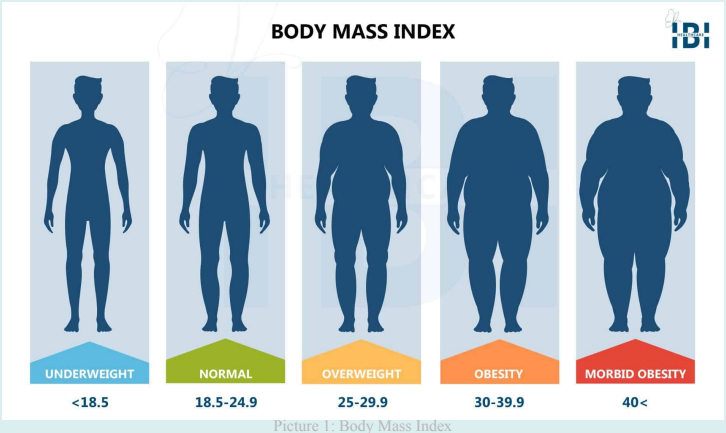
Predicting Obesity Status

12/02/2024

Hanyang Zhou, Jamie Tian,
Yiting Wang



Picture 1: Body Mass Index

Introduction

Obesity: A Major Public Health Concern

- Obesity is a medical condition of excess body fat, defined by a BMI over **30**.
- Obesity **significantly increases the risk of developing several serious health conditions** like heart disease, type 2 diabetes, stroke, high blood pressure, and osteoarthritis, leading to decreased quality of life and potential premature death.
- Obesity is a major public health concern due to its widespread impact and associated healthcare costs. According to the U.S. Center for Diseases Control and Prevention, obesity costs the US healthcare system almost **$173 billion a year**.

Picture 2: BMI

About the Study

Accurately predicting obesity allows for **early identification** of individuals at risk, enabling timely interventions and preventative measures to be taken, potentially reducing the **development of obesity-related health complications** later in life.

This study performs an exploratory data analysis on a dataset with a large sample size to classify the individuals' obesity status as "Obese" or "Not Obese." Statistics and machine learning techniques are used to perform a comprehensive investigation on the dataset.

About the Dataset

Number of Observations

- Training data: 32,014
- Testing data: 10,672
- Total: 42,686

Number of Variables

- Predictors: 29
- Response variable: 1

Categorical Predictors Examples

- Gender: Gender of the individual
- Family history of overweight: Whether there's a family history of overweight
- FAVC: Frequent consumption of high-calorie food

Numerical Predictors Examples

- Age: Age of the individual
- Height: Height of the individual in meters
- Cholesterol: Cholesterol Level
- MaxHR: Maximum Heart Rate
- avg_glucose_level: The individual average glucose level

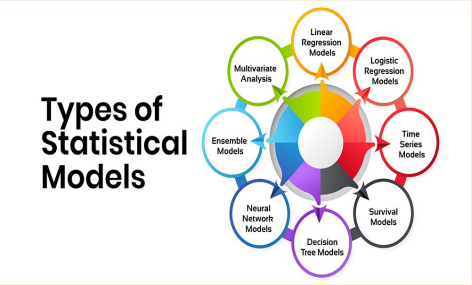
Methods

Methods Introduction

To predict the obesity status for each ID in the given test set and ensure consistency and accuracy, we performed data preparation and processing, model testing, and model selection. We tested a few different models and selected the one with the best accuracy.

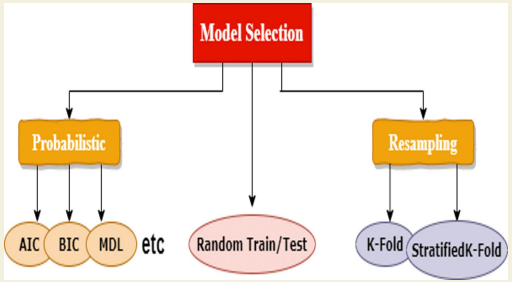
Picture 3: Data Cleansing

1. Data Cleaning



Picture 4: Types of Statistical Models

2. Model Testing



Picture 5: Model Selection

3. Model Comparison & Selection

Data Cleaning

Consistency

- We checked that the variables in our test and training datasets match by finding the intersection of their column names, and ensured that the datasets have the same structure.

Missing Values Handling

- **Numerical Variables**: We replaced the missing values ‘NA’ in each variable with the **mean of that variable**.
- **Categorical Variables**: We replaced the missing values ‘NA’ in each variable with the **most frequent category in that variable (mode)**.

Choice of Models

Logistic Regression

Pros:

- **Simplicity**: Easy to implement and computationally efficient
- **Interpretable**: Provides insights for important contributing factors in obesity predictions.

Cons:

- **Not flexible:** Struggles with non-linear relationships
- **Sensitive to Multicollinearity**: Correlated predictors can distort coefficient estimates

Random Forest

Pros:

- **Versatility**: Handles a mix of numerical and categorical data well
- **Accuracy**: High model performance with large dataset

Cons:

- Hard to interpret individual predictions.
- Computationally Complex

Specifically, to leverage the simplicity of the Logistic Regression and the accuracy of the Random Forest method,
We first apply Logistic Regression to the data, aiming at exploring the relationship between variables.
We then apply the Random Forest method for a more accurate prediction of individual obesity status.

Why NOT…

- For a binary classification problem, some other conventional methods are **LDA, QDA and KNN**. Theoretically, these models can also be used in modeling mixed data with both numerical and categorical variables.

- However, it worth noting that we have more categorical predictors (18) than numerical ones (11), and a lot of our categorical predictors in the obesity data set are binary.

- This data structure makes it difficult to
  - satisfy the normality assumption for LDA and QDA
  - find linear discriminants with LDA
  - separate classes in the data space using distance metrics with KNN

Multicollinearity

- **Multicollinearity:** several independent variables in a model are correlated.
- Multicollinearity among independent variables will result in **less reliable statistical inferences**.
- **Variance Inflation Factor (VIF)** is a statistical metric used to measure the degree of multicollinearity among the independent variables in a regression model.

We used our logistic regression model with all 29 variables to check for multicollinearity.

- **VIF = 1:** No correlation (ideal case).
- **1 < VIF ≤ 5:** Moderate correlation; generally acceptable.
- **VIF > 5:** High correlation; multicollinearity might be problematic and therefore the variable will be excluded from model.

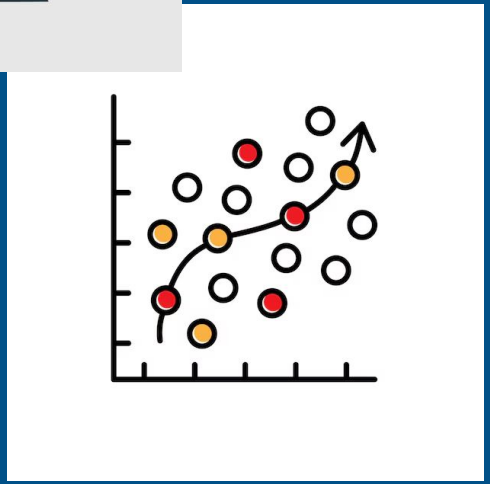
We found that all of the predictors have a VIF value between 1 and 2, indicating that there are no multicollinearity problem in the training data.

Model Testing

Model 1.1

Logistic Regression - Full Model

Picture 6: Machine Learning



Picture 7: Statistical Model

Model 1.1

Method: We created a **Logistic Regression** model using **ALL 29 predictors**. This model was used to predict the probability that an individual is **obese** or **not obese** and categorize them based on the more likely option.

Significant Predictors (17): Height, FCVC, NCP, CH2O, FAF, Cholesterol, avg_glucose_level, Gender, family_history_with_overweight, FAVC, CAEC, SCC, MTRANS, Race, FastingBS, ExerciseAngina, stroke

Results: This model had a **testing error rate** of **25.37%** which is about **14%** better than the maximum error rate allowed. This error rate is good, but not great. Additionally, this model has **high dimensionality** and therefore **harder interpretability**.

TRAINING ERROR RATE: 25.36%

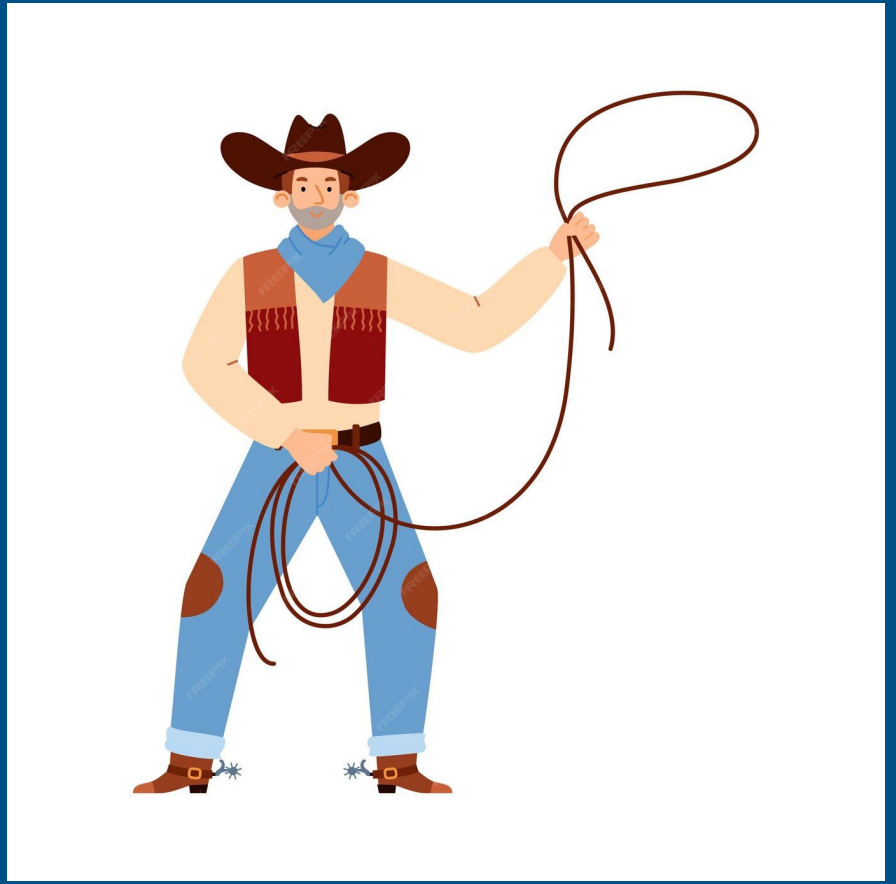
	Not Obese	Obese
Not Obese	16526	5114
Obese	3005	7369

Figure 1: Confusion Matrix of Model 1.1

Model 1.2

Logistic Regression - Partial Model

(Manual Stepwise Lasso Regression)



Picture 8

Model 1.2

Method: We used **Lasso Regression** as a feature selection tool. We started from a full lasso regression model and removed the least significant predictor(s) step by step and recreated a new partial lasso regression model with the remaining predictors each time. We found that the most balanced **Lasso Logistic Regression** model used **ONLY 5 predictors**. This model was used to predict the probability that an individual is **obese** or **not obese** and categorize them based on the more likely option.

Significant Predictors (5): CH2O (+), CAEC (Frequently/Sometime +, No -), FAF (-), FAVC (+), MTRANS (Motorbike/Public_Transportation +, Bike/Walking -)

Results: This model had a **testing error rate** of **25.89%** which is about **13%** better than the maximum error rate allowed and only about **0.5%** worse than Model 1.1. However, this model has significantly **lower dimensionality** and thus much **more interpretable**.

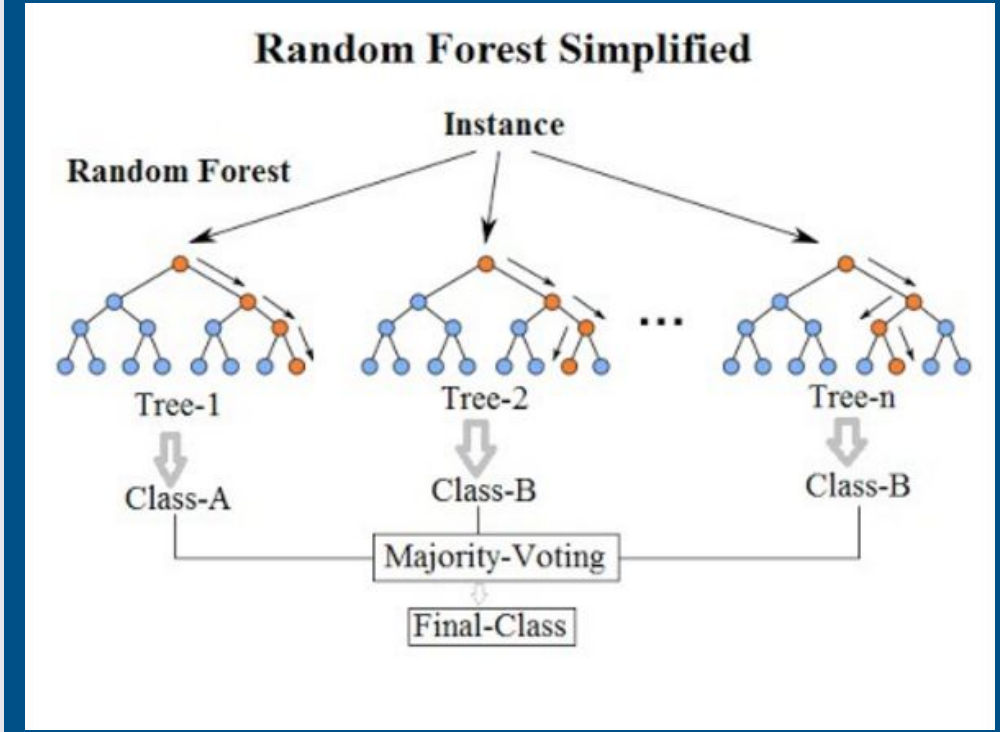
TRAINING ERROR RATE: 26.11%

	Not Obese	Obese
Not Obese	16984	5813
Obese	2547	6670

Figure 2: Confusion Matrix of Model 1.2

Model 2

Random Forest - Model with Full Predictors



Picture 9: Random Forest

Model 2

Method: We created a **Random Forest** model using **ALL 29 predictors**. The model had 1000 trees. A subset of predictors was randomly sampled at each split. Its size was equal to the square root of the total number of predictors.

Results: We found a training error rate almost equal to zero. The class error is about 0.0001 for "Not Obese" and 0.0140 for "Obese."

TRAINING ERROR RATE: 0%

	Not Obese	Obese
Not Obese	19529	2
Obese	175	12308

Figure 3: Confusion Matrix of Model 2

Results
Final Model

Our best model is Random Forest with all predictors, and 1000 trees. With this model we reached a low testing error rate of 0.00638%. However, with a testing sample size of 10672, there are still about 70 people whose obesity status were not accurately predicted. In future studies, researchers can attempt other missing-data handling techniques, advanced modeling techniques or consider including informative new predictors in order to reach a better prediction accuracy of obesity status for the general population. Some potential future directions are discussed in the following sections.

Limitations

Data Cleaning

The approach to handling missing data by replacing missing values with the mean or the most frequent category might introduce bias, especially when missing values not similar to the mean or mode.

Limited Methods

Only logistic regression and random forest were tried. Other classification methods, like Support Vector Machines (SVM) or Gradient Boosting Machines (GBM), were not used, and they could potentially enhance the performance of our models.

Limitation of the Dataset

While the dataset comprises 42,686 observations, it is uncertain that whether the sample accurately represents the broader demographic characteristics like sex, age, socioeconomic status, and geographic diversity. If the dataset is not representative, the model's ability to predict obesity status in wider population may be limited.

Future Improvements

Dataset Expansion

Enhanced Demographic Details:

- Add age categories (e.g., children, adolescents, adults, elderly) to capture **age-specific** obesity risks.
- Expand the "Race" feature to represent diverse ethnic groups, providing insights into **cultural dietary habits** and genetic predispositions.

Temporal and Longitudinal Data

- Record **variables** like BMI, caloric intake (CALC), and physical activity (FAF) over **months or years** to study obesity trends..
- Add variables to track **seasonal variations**, such as increased activity in summer or higher caloric intake during holidays.

Real Word Application

Wearable Integration:

- Leverage **fitness trackers** (e.g., Fitbit, Apple Watch) to monitor continuous health data like heart rate, activity, and sleep.
- Deliver **personalized alerts** for prolonged inactivity or **high-risk BMI trends.**

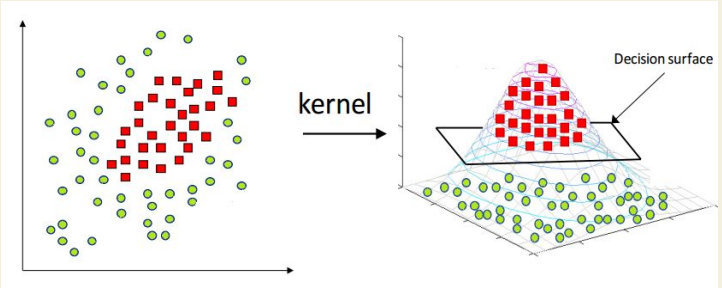
Healthcare Support:

- Deploy predictive models in clinics to identify high-risk patients during **routine check-ups.**
- Prioritize early interventions like **nutritional counseling** or preventive care programs.

Further Improvements - Advanced Methodologies

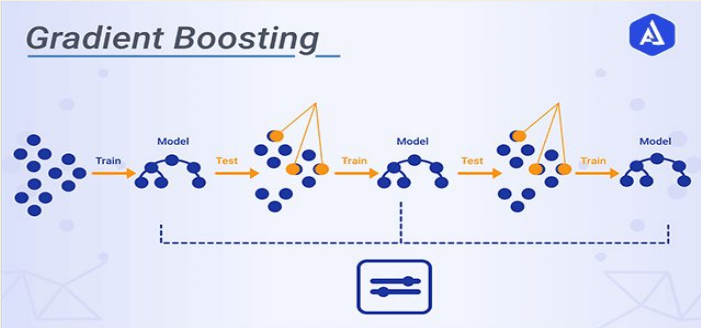
Support Vector Machines (SVM):

- **Strengths**:
  - Effective for high-dimensional datasets.
  - Non-linear relationships handled through kernel tricks.
- **Applications**:
  - Ideal for distinguishing complex obesity patterns.
  - Example: Classify obesity status using mixed feature types (numerical and categorical).



Gradient Boosting Machines (GBM):

- **Strengths**:
  - Reduces bias and variance.
  - Handles missing data naturally during training.
- **Applications**:
  - Boost predictive accuracy with iterative model improvements.
  - Example: Use GBM for ranking the most critical obesity predictors.



Ensemble Methods:

- **Strength**:
  - Combines predictions from multiple models (e.g., Random Forest and Logistic Regression).
  - more robust generalization to unseen data, particularly for complex datasets with imbalanced classes.

Summary

- We started with **Logistic Regression** due to its simplicity and interpretability, achieving an error rate of approximately 25%.
- Our **Random Forest** model with 1,000 trees excelled by capturing complex, non-linear interactions and handling mixed data types effectively.
- Despite its strengths, the Random Forest model accuracy still has space for improvement, and the model complexity made interpretation challenging.
- Future work should explore advanced imputation techniques and alternative models, such as Support Vector Machines (SVM) and Gradient Boosting Machines (GBM), to further enhance accuracy and applicability.

References

Bhaskar, G. (2021, February 18). Data Cleaning. Data Preprocessing. https://medium.com/data-preprocessing/data-cleaning-788b1cdd151f

Centers for Disease Control and Prevention. (2024, May 16). About Obesity. Obesity. https://www.cdc.gov/obesity/php/about/index.html

FACS, A. C. I. M. (2022, November 11). Morbid Obesity: What BMI Really Reveals? IBI Healthcare Institute. https://www.ibihealthcare.com/bariatric/bmi-of-morbid-obesity/

Koehrsen, W. (2020). Random Forest Simple Explanation. Medium. https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d

MarkovML. (2024, January 16). Machine Learning Model Selection and Parameter Tuning: A Guide. https://www.markovml.com/blog/machine-learning-model-selection

What is Statistical Modeling in Data Science? (2024). Dasca.org. https://www.dasca.org/world-of-data-science/article/what-is-statistical-modeling-in-data-science