AOL Report Data Mining

Comparative Analysis of Machine Learning Algorithms for Predicting Obesity Levels Based on Eating Habits and Physical Activity

Disusun Oleh:

2602124316 - Dionisius Hendi Krisnanto

2602152553 - Bryan Jo

2602123572 - Geovanka Thersia Kurniawan

1. Introduction

A. Background of the Study

Obesity, which leads to many physical and mental health problems, is a major global health issue with serious consequences. The number of people affected by obesity is steadily increasing, making it essential to conduct new research to understand the factors that contribute to this condition. Such research should explore the various elements that influence obesity, including lifestyle choices and genetics. By studying these factors, we can learn how they affect the development of obesity. Additionally, it is important to create methods to predict who is likely to become obese based on these factors. These predictive tools can help design specific strategies and preventive measures to reduce the growing number of obesity cases and lessen its health impacts.

Eating habits, including dietary intake and meal patterns, are crucial in the development of obesity. Consuming high amounts of calorie-dense, nutrient-poor foods and having irregular meal patterns are often linked to weight gain. On the other hand, a balanced diet rich in fruits, vegetables, and whole grains can help maintain a healthy weight.

Physical activity is also a vital factor in obesity. Regular physical activity helps with energy expenditure, weight management, and overall health. Inactive lifestyles characterized by prolonged periods of inactivity are strongly associated with weight gain and obesity.

In recent years, machine learning techniques such as Naïve Bayes, Support Vector Machines (SVM), Random Forest, and Decision Trees have become increasingly prominent in predicting health outcomes and tailoring medical interventions. Naïve Bayes is valued for its simplicity and efficiency, SVM for its robustness in high-dimensional spaces, Random Forest for its accuracy and ability to handle large datasets, and Decision Trees for their intuitive and interpretable nature. By processing diverse inputs like eating habits and physical activity levels, these algorithms can effectively forecast obesity levels, allowing for better predictive tools and preventive measures.

B. Problem Definition

- How can we address the escalation in the number of obese individuals?
- What is the primary factor that has the greatest impact on obesity level?

C. Solution to the Problem

- Making a good model that is able to predict people's obesity level based on their lifestyle and eating habits enabling timely lifestyle adjustments to prevent obesity.
- Analyze the correlation between every feature in each class and the type of obesity level.

D. Other Studies

The study titled "Obesity and Severe Obesity Forecasts Through 2030" projects future obesity trends and estimates potential savings from successful prevention efforts. It utilizes nonlinear regression models based on data from the Behavioral Risk Factor Surveillance System (BRFSS) spanning 1990 to 2008. These models consider various factors at both individual and state levels, such as gender, age, race/ethnicity, education, marital status, annual household income, annual unemployment rates, and prices for items like alcohol, gas, fast food, groceries, healthier foods, Internet access, and fast-food availability. By comprehensively analyzing these factors, the study aims to predict obesity and severe obesity prevalence up to 2030, taking into account a broad spectrum of socio-economic and environmental influences.
https://www.ajpmonline.org/article/S0749-3797(12)00146-8/fulltext

The study, titled "Obesity disease risk prediction using machine learning," gathers data on obesity and preps it for machine learning. They clean and organize the data and visualize it. Then, they use different machine learning methods like support vector machines, random forests, and decision trees to classify obesity risk. Results show that the random forest model works best with 96% accuracy. They also test a clinical support system for kids' health, finding it's 85% accurate

with the ID3 model. Another study by Wang et al. explores using genetic data to predict obesity risk.
https://www.researchgate.net/publication/377155765_Obesity_disease_risk_prediction_using_machine_learning

In contrast, our study takes a different approach by focusing on immediate classification of obesity status using current data. We analyze variables such as lifestyle choices, age, and weight to determine whether an individual is obese. Unlike the broader scope of the other study, which aims for long-term forecasts and financial estimates, our research centers on the here and now. We investigate how specific lifestyle choices impact obesity status in the present moment, rather than projecting future trends. While the other study provides extensive insights into long-term outcomes using a wide range of variables, our study offers a concentrated examination of current individual factors related to lifestyle, age, and weight to assess obesity status.

2. Data

A. Dataset Used

In this study, the dataset is collected from Kaggle (https://www.kaggle.com/datasets/fatemehmehrparvar/obesity-levels), where it has 17 columns and 2111 rows. The 23% of the dataset was collected directly from people through a web platform, and the remaining 77% data was generated synthetically using Weka Tool and SMOTE filter.

Attributes	Type	Description
Gender	Categorical	The user's gender
Age	Continuous	The user's age
Height	Continuous	The user's height
Weight	Continuous	The user's weight
family_history_	Binary	Has a family member suffered or

with_overweight		suffers from being overweight?
FAVC	Binary	Do you eat high caloric food frequently?
FCVC	Integer	Do you usually eat vegetables in your meals?
NCP	Continuous	How many main meals do you have daily?
CAEC	Categorical	Do you eat any food between meals?
SMOKE	Binary	Do you smoke?
CH2O	Continuous	How much water do you drink daily?
SCC	Binary	Do you monitor the calories you eat daily?
FAF	Continuous	How often do you have physical activity?
TUE	Integer	How much time do you use technological devices such as cell phones, video games, television, computers and others?
CALC	Categorical	How often do you drink alcohol?
MTRANS	Categorical	Which transportation do you usually use?
NObeyesdad	Categorical	Obesity level

B. Pre-processing

The data preprocessing was conducted to enhance and simplify our dataset. We start by rounding down the age feature to convert it from continuous data to integers, as age is typically represented as an integer rather than a continuous value.

Subsequently, features with weak correlations to the target variable, such as Gender and MTRANS (the mode of transportation typically used by respondents) are eliminated, as they have minimal impact on the label.

Following this, we removed duplicate rows from our dataset to avoid bias, redundant information, and misleading model interpretations, ensuring that no single row is given excessive importance during training.

Following that, we addressed outliers by removing features with a low number of outliers, such as height and weight. For features with numerous outliers, we replaced outlier values with the mean of the non-outlier data.

After managing the outliers, we normalized our dataset using the `StandardScaler` from scikit-learn. This step ensured that all features were on a similar scale, enhancing the training speed and reducing bias towards features with larger scales.

As a final step, we addressed data imbalance using SMOTE (Synthetic Minority Over-sampling Technique) or upsampling methods. This balanced the dataset, reducing bias towards the majority class and improving overall model performance.

3. Dataset Problem Solving

A. Exploratory of the Dataset

We start by importing our dataset into Python using the pandas library which we then print the dataset for inspection. We check for duplicate entries, review general information including data types, and generate descriptive statistics. We also use pandas' built-in functions to identify any null values in the dataset. Our initial exploration reveals that the dataset contains some duplicate rows, no missing values, and that the age variable is continuous.

Afterward, we assess the correlation between the features and the target label by creating a correlation matrix. This analysis shows that Gender and MTRANS (the mode of transportation typically used by respondents) have weak correlations with the label, with correlation coefficients of 0.025 and -0.046, respectively.

Eventually, we identify outliers in each feature using the Interquartile Range (IQR) method. We find that height, weight, NCP (number of daily meals),

and age all contain outliers. Specifically, NCP has 577 outliers, height and weight each have one outlier, and age has 147 outliers.

4. Method

A. Technique Used for the Dataset

We utilized our dataset to employ five distinct models: Gaussian Naive Bayes, Bernoulli Naive Bayes, Support Vector Machine, Decision Tree, and Random Forest. Following the execution of these models, we conducted a thorough comparison to determine the most suitable model for predicting obesity levels. For models featuring parameters, we employed grid search techniques to identify the optimal parameters. Additionally, we utilized K-Fold Cross Validation to evaluate the performance of each model.

5. Result & Analysis

TRAINING	Accuracy	Precision	Recall	F1-score
SVM	0.98	0.98	0.98	0.98
GaussianNB	0.57	0.58	0.57	0.51
BernoulliNB	0.57	0.56	0.57	0.55
Decision Tree	1	1	1	1
Random Forest	1	1	1	1

The table shows percentages for various evaluation metrics (accuracy, precision, recall, and F1-score) for different models using the training data. The SVM model consistently scores 0.98 across all metrics. In contrast, the GaussianNB model scores 0.57 in accuracy and recall, 0.58 in precision, and 0.51 in F1-score. Similarly, the BernoulliNB model scores 0.57 in accuracy and recall, 0.56 in precision, and 0.55 in F1-score. Both the Decision Tree and Random Forest models perform strongly and consistently across all metrics, each achieving a perfect score of 1.

TESTING	Accuracy	Precision	Recall	F1-score
SVM	0.96	0.96	0.96	0.96
GaussianNB	0.59	0.59	0.59	0.53
BernoulliNB	0.55	0.55	0.55	0.54
Decision Tree	0.94	0.94	0.94	0.94
Random Forest	0.95	0.95	0.95	0.95

The table above presents the percentages of various evaluation metrics, including accuracy, precision, recall, and F1-score, for various models from the testing data. The SVM model demonstrates consistent performance across all evaluation metrics with a score of 0.94. In contrast, the GaussianNB model shows 0.59 for accuracy, precision, and recall, and 0.53 for F1-score. Similarly, the BernoulliNB model reports 0.55 for accuracy, precision, and recall, and 0.54 for F1-score. Both the Decision Tree and Random Forest models exhibit consistent and strong performance across all metrics, each achieving a score of 0.94 and 0.95.

6. Evaluation

Based on our model's outcomes, GaussianNB and BernoulliNB, which are Naive Bayes variants, show poor performance, recording the lowest scores across all metrics. Conversely, the Decision Tree and Random Forest models exhibit robust performance, obtaining a score of 0.94 and 0.95. Yet, the SVM (Support Vector Machine) model shines with exceptional performance, securing a top score of 0.96, thus establishing itself as the superior model among those assessed.

Support Vector Machines (SVM) demonstrate high accuracy on test data, making them very effective for predicting new, unseen data. They also exhibit the least overfitting, as shown by the small gap between training and testing accuracy when compared to Decision Trees and Random Forests. This is because SVM evaluates different hyperplanes and chooses the one with the maximum margin. SVM aims to create a decision boundary that not only effectively separates the classes in the training data but also generalizes well to unseen data.

7. Conclusions

The result indicates that SVM stands out as the optimal model for predicting individuals' obesity levels to prevent a rise in obesity cases. Utilizing this model enables us to categorize and identify an individual's obesity level accurately. Moreover, we can determine the key features most strongly associated with obesity levels, including Weight, Family History, and CAEC.

8. Implications

Based on our study findings, we can proactively prevent obesity by identifying individuals at risk of nearing obesity and encouraging lifestyle changes. Additionally, we have identified key features strongly correlated with obesity levels, such as Weight, Family History, and CAEC (How often do the respondents eat food between meals), enabling individuals to target specific lifestyle adjustments for better health.