Analyzing Public Health Factors and BMI Levels

Shuyang Wu, Jacqueline Jia, Esther Davies

Introduction:

Body Mass Index (BMI) serves as a valuable estimate of body fat and an essential indicator of an individual's risk for various diseases associated with excess body fat, including heart disease and high blood pressure. With the understanding that a higher BMI might correlate with an increased risk of such health issues, our team tends to predict BMI levels based on a diverse set of public health factors. Our primary objective in this study is to empower individuals with insights into how these health factors influence their BMI, which can help promote awareness of these interconnected public health considerations.

Data Engineering Process:

Our group focuses on the dataset related to BMI, which contains 23,535 entries, comprising 14,896 females and 8,639 males. This imbalance across gender reflects the real-world Canadian population ratio of 63% to 37%. This study uses nine key public health factors to predict BMI levels: age, gender, calorie intake, sugar intake, physical activity, fast food frequency, screen time, height, and weight. In order to perform machine learning in the next step, we first checked if there were any NA values. We also comprehensively examined key statistics for each variable, including mean, standard deviation, and quantiles, to ensure data integrity.

Then, we categorized BMI values into four distinct levels: underweight, healthy, overweight, and obese. To gain insights into variable relationships, we created a correlation heatmap, identifying potential multicollinearities among features. This heatmap guided our feature selection and model building. Moreover, we optimized the "gender" variable by transforming it into a binary dummy variable (0 for male, 1 for female). This transformation enhanced data compatibility with our machine learning algorithms, ultimately improving model performance and interpretability.

Analysis:

Our analysis of predicting BMI levels based on public health factors utilizing the k-Nearest Neighbors (k-NN) algorithm with k=18. We selected the k-NN algorithm for this analysis since it is suitable for multiclass classification without making any assumptions about the data distribution. To evaluate the model's performance, we divided our dataset into a 65% training set and a 35% testing set, giving us the best performance. Since the KNN algorithm relies on the distance between data points, scaling the variables is essential. This ensured all predicted factors had equal weight in distance calculation for the algorithm and prevented any variables with larger weights from dominating the model decision. Optimizing the value of 'k' in the k-NN algorithm was an essential step since it will affect the prediction accuracy. We conducted a systematic exploration of different 'k' values, and we determined that 'k=18' struck an optimal balance between model complexity and predictive accuracy.

Findings:
The k-NN model with k = 18 achieved an overall accuracy of 0.8282. This means our k-NN model correctly classified 83% of individuals' BMI levels based on the given health factors.

Then, we took a deeper look into the classification report, which contains precision, recall, and F1 score. The precision measures the accuracy of positive predictions, with values ranging from 0.76 to 0.99 across BMI categories. "Underweight" has the highest precision at 0.99, followed by "obese" (0.94), "overweight" (0.79), and "healthy" (0.76). This suggests that the model excels in correctly identifying individuals as underweight or obese. Recall, ranging from 0.51 to 0.94, indicates the model's ability to capture true positive instances. "Healthy" boasts the highest recall at 94%, while "underweight" exhibits lower recall at 51%. F1-score, which balances precision and recall, ranges from 0.68 to 0.90, with "obese" achieving the highest score at 0.90.

Conclusion
The k-NN model in this study has demonstrated strong predictive capabilities to classify individuals into four BMI levels using individuals' public health factors. As observed in the finding, the model has effectively classified "healthy", "obese", and "overweight" individuals achieving high F1 scores. While most BMI categories can be correctly classified, there remain spaces for refinement, particularly in enhancing the underweight prediction. Our predictive model offers a reliable way to estimate individuals' BMI levels based on public health factors. This can assist healthcare practitioners in quickly assessing a patient's weight status and enabling more effective public health care.

Individual Contributions
Shuyang Wu:
- Spearheaded data cleaning and preprocessing, executed the k-NN model, and expertly visualized the k-NN results.
- Proofread the writing report
- Collaborated with designing the PowerPoint presentation.

Jacqueline Jia:
- Helped with data processing and meticulously fine-tuning the k-NN model
- Wrote the analysis report.
- Collaborated with designing the PowerPoint presentation.

Esther Davies:

Github link: https://github.com/JacquelineeJia/datathon1
Google Slide:
https://docs.google.com/presentation/d/1iAP0y6P1d66P76mNjJCkYgwcdPs0x4EhtUBtqsuLfBk/edit?usp=sharing