Data Description and Exploratory Analysis

The data set is sourced from the UC Irvine Machine Learning Repository, named Obesity Levels Based on Eating Habits and Physical Condition. It collected individuals' information about their eating habits, physical condition, and relative obesity level from Mexico, Peru, and Colombia in 2019,[1].

This data set contains 2,111 observations, with 16 features and one target variable. Of the 16 features, 8 are categorical and 8 are numerical. The target variable, NObeyesdad, is categorical and includes 7 levels: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III.

Variable	Type	Description	Variable	Type	Description
Gender	chr	Gender	CAEC	chr	Eat any food between meals?
Age	num	Age	SMOKE	chr	Smoke?
Height	num	Height	CH2O	num	Drink how much water daily?
Weight	num	Weight	SCC	chr	Monitor the calories intake daily?
Family History	chr	Has a family history of overweight?	FAF	num	How often do physical activities?
FAVC	chr	Eat high caloric food frequently?	TUE	num	How much time spent on smart devices?
FCVC	num	Eat vegetables frequently?	CALC	chr	How often drink alcohol?
NCP	num	Number of meals daily?	MTRANS	chr	Daily transportation?

Table 1: Description of Variables in the Dataset

From the graph of numerical variables (Figure 1) shows that height and weight have the highest correlation (0.46) compared to other variables. Additionally, weight provides the clearest separation among the obesity categories. The graph of categorical variables (Figure 2) indicates that the dataset is nearly balanced across seven classes, with only obesity_type_I having slightly higher counts compared to the other types. The gender distribution is also balanced, with almost equal numbers of females and males. Furthermore, most people in this dataset have a family history of being overweight, consume more high-calorie foods, do not smoke, occasionally drink alcohol, and use public transportation.

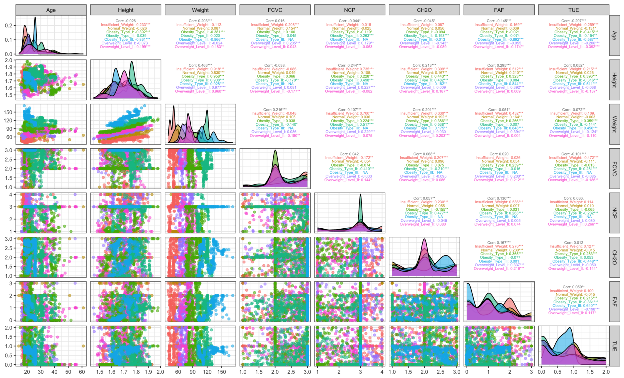

Figure 1: Numerical Variables

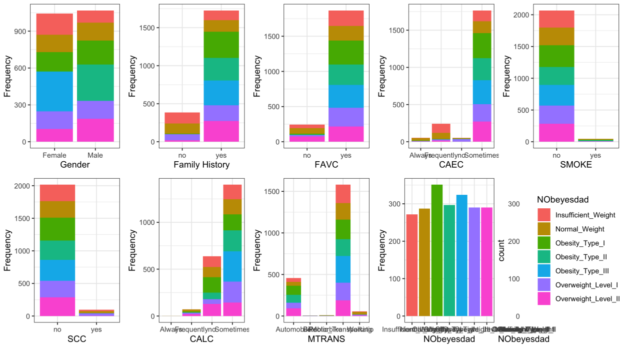

Figure 2: Categorical Variables

Problem Statement

According to the latest report from the World Health Organization (2024), by 2022 the obesity rate among adults had reached 1 in 8, twice the rate in 1990, and adolescent obesity had increased fourfold. The report also revealed that millions of deaths each year are caused by noncommunicable diseases linked to overweight and obesity, with higher BMI levels than optimal,[2].

This study aims to predict obesity levels using a neural network model. It seeks to help individuals prevent and manage obesity by adjusting their eating habits and increasing physical activity.

Method

A neural network is a deep learning model composed of node layers, including a input layer, one or more hiddenlayers, and an output layer. It mimics the way how human brain processes information, enabling computers to recognize patterns and solve problems. Each input is processed through input layer nodes and passed to the next layers. The neural network's output, $F(x)$, is calculated as a weighted sum of the activations $h_k(x)$ of the hidden layer nodes:

$$F(x) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k(x),$$

where $h_k(x)$ represents the activations of the k-th hidden node. Each activation is computed as:

$$h_k(x) = g(w_{k0} + \sum_{j=1}^{p} w_{kj} x_j),$$

where w_{kj} are the weights, x_j are the inputs, and $g(z)$is a non-linear activation function. During training, the network adjusts its weights and biases using optimization methods like gradient descent to minimize the error and achieve the best accuracy on the trainging data.

I performed stratified splitting to divide the dataset into 75% training and 25% test data. Since the nnet function requires numerical input, I removed all categorical variables from the dataset, leaving 8 numerical predictors (Age, Height, Weight, FCVC, NCP, CH2O, FAF, TUE) and one target variable, NObeyesdad, which was converted to a factor. I built three neural network models: the first was a baseline model using the original data without parameter tuning. For the second model, I optimized the parameters using 5-fold cross-validation. For the third model, I scaled all numerical predictors in the training and test datasets separately using the same scaling parameters to avoid data leakage. The optimal parameters for this model were identified using 10-fold cross-validation.

Results and Discussion

The results showed a significant improvement in model performance through parameter tuning and data scaling. The base model achieved an accuracy of 69.0% with an AIR of 0.507, showing limited performance without parameter optimization or data scaling. Optimal Model 1, which used parameter tuning but no data scaling, improved accuracy to 84.6% and AIR to 0.739. Optimal Model 2, incorporating both parameter tuning and scaled predictors, achieved the best performance with 94.7% accuracy and an AIR of 0.885. These findings highlight the importance of scaling numerical predictors and tuning parameters, as they substantially enhance the predictive capability of neural network models.

Model	**Accuracy**	**AIR**
Base Model (no scale, no parameter tuning)	0.690	0.507
Optimal Model 1 (no scale, tuning parameter)	0.846	0.739
Optimal Model 2 (scale data, tuning parameter)	0.947	0.885

Table 2: Model Performance Comparison

Model	**Size**	**Decay**	**CrossV**	**Opt-Size**	**Opt-Decay**
Optimal Model 1	1:20	0:5	5	19	0
Optimal Model 2	1:20	0:8	10	8	0

Table 3: Tuning Parameters for Optimal Models

Conclusion

This study successfully demonstrated how a neural network can be used to predict obesity levels based on numerical predictors. The results showed that parameter tuning and scaling significantly improve model accuracy and performance. The base model provided limited results, but the optimized models, especially the one with scaled predictors, achieved high accuracy and AIR

References

[1] *Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]*. 2019. DOI: 10.24432/C5H31Z. URL: https://doi.org/10.24432/C5H31Z.

[2] World Health Organization. *Obesity and Overweight*. Accessed: 2025-01-24. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight#:~:text=In%202022%2C%201%20in%208,who%20were%20living%20with%20obesity.

[3] F. H. Yagin et al. "Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique". In: *Applied Sciences* 13.6 (2023), p. 3875. DOI: 10.3390/app13063875. URL: https://doi.org/10.3390/app13063875.