MSBD566 – Predictive Modeling and Analysis
Name: Megan Leak
Date: 12/2/2025
Project Title: Prediction of Obesity Levels

Project Description

The primary objective of this project is to develop a learning model that predicts an individual's obesity level from a set of physical measurements and self-reported lifestyle behaviors. Unlike binary classification problems that distinguish between simply "obese" and "non-obese," this dataset contains seven distinct outcome categories, which increases the complexity of the task. The project focuses on two main methodological components: reducing the dimensionality of the feature set using PCA and training a feedforward neural network to classify obesity levels. PCA is employed to extract dominant patterns from the data, helping reduce noise and redundancy before feeding the transformed features into the neural network. The feedforward network then learns to classify individuals based on these derived principal components. This approach builds on the idea that combining statistical techniques with nonlinear learning models can yield both interpretable and accurate predictions.

Data Description

The dataset used in this project is the Estimation of obesity levels (1), which contains 2,111 samples gathered from individuals across Latin American countries. The dataset includes a wide range of behavioral, nutritional, and demographic attributes such as age, height, weight, vegetable consumption frequency, water intake, family history of overweight, smoking habits, caloric food consumption, and time spent using technology. Additionally, categorical attributes such as gender, alcohol consumption habits, transportation modes, and physical activity levels were included. Because many of these features were non-numeric, one-hot encoding was applied to transform them into numerical representations suitable for machine learning models. This preprocessing step increased the dimensionality of the dataset from 17 original variables to 46 encoded features. The target variable represents the individual's obesity category and includes seven possible classes: Insufficient Weight, Normal Weight, Overweight Level 1, Overweight Level II, Obesity Type 1, Obesity Type II, and Obesity Type III. The dataset also includes synthetic sample produced using statistically informed data simulation techniques to protect privacy while preserving the distributional characteristics of the original data. The original dataset is publicly available via the UCI Machine Learning Repository, which provides further documentation on how the data were collected and validated.

Method and Analysis

Before applying PCA, all numerical features were standardized using a z-score transformation to ensure that variables with large scales did not dominate the principal components. PCA was chosen to address the high dimensionality resulting from one-hot encoding, to reduce noise, and to identify latent structures within the data. Ten principal components were extracted based on the cumulative explained variance, which retained a substantial proportion of the original information while transforming the dataset into a more compact and uncorrelated feature space. These components represent underlying lifestyle and demographic patterns, such as physical activity tendencies, dietary habits, and physiological characteristics.

Following dimensionality reduction, a feedforward neural network was trained on the PCA transformed features. The model was implemented using the scikit-learn MLP Classifier, configured with four hidden layers containing 64,32,16, and 8 neurons, respectively. The ReLU activation function was applied to each hidden layer to introduce nonlinearity, and the network was trained using the Adam optimization algorithm with categorical cross-entropy as the loss function. The dataset was split into training and test sets using an 80/20 ratio to evaluate generalization performance. The rationale for choosing a neural network model lies in its ability to capture complex, nonlinear interactions between the features and the obesity categories. Unlike linear models, neural networks can model subtle relationships across diverse behavioral and physiological inputs, which makes them well suited for multiclass classification tasks such as this one.

After training for up to 300 iterations, the model converged to a test accuracy of approximately 82%. This level of performance is strong given the complexity of predicting seven distinct obesity categories, each influenced by a combination of behavioral, dietary, and demographic factors. Despite the model's high accuracy, certain classes still exhibit overlapping lifestyle characteristics, particularly between adjacent categories such as Overweight Level 1 and Overweight Level II or Obesity Type 1 and Obesity Type II. These subtle similarities contribute to most misclassification errors and highlight the inherent challenge of distinguishing between closely related obesity levels.

Nevertheless, the neural network demonstrated a robust ability to identify meaningful patterns in lifestyle behaviors that differentiate obesity categories. Its capacity to model nonlinear relationships enhanced by PCA's dimensionality reduction allowed the network to capture key trends related to physical activity, eating habits, hydration, and other behavioral factors. The model's performance indicates that the combination of PCA and neural network-based classification is effective for uncovering complex health related patterns and provides a solid foundation for future refinements.

Evaluation

The results demonstrate that PCA served as an effective dimensionality reduction strategy, enabling the neural network to learn from a more compact and less redundant feature space. By reducing the original dataset to a smaller set of principal components, the model was able to focus on the most informative behavioral and demographic patterns without being affected by noise or multicollinearity. This contributed to both improved training stability and computational efficiency. The neural network achieved a final validation accuracy of approximately 82%, indicating that the selected principal components retained substantial discriminatory power for predicting obesity categories. This level of performance suggests that behavioral and demographic attributes such as physical activity, eating habits, hydration, and family history encode meaningful information related to obesity levels.

However, obesity classification remains challenging due to the natural overlap in lifestyle characteristics among adjacent categories. These subtle distinctions can reduce class separability, highlighting the complexity of capturing fine grained health related patterns using tabular features alone. While the neural network's nonlinear modeling capabilities helped address this complexity, further improvements may require richer datasets, additional engineered features, or domain specific modeling approaches. PCA also enhanced interpretability by reducing the dimensionality of the input space, making it easier to explore how broad behavioral components relate to obesity outcomes. Overall, the combination of PCA and a neural network provided a strong balance between computational efficiency and predictive performance.

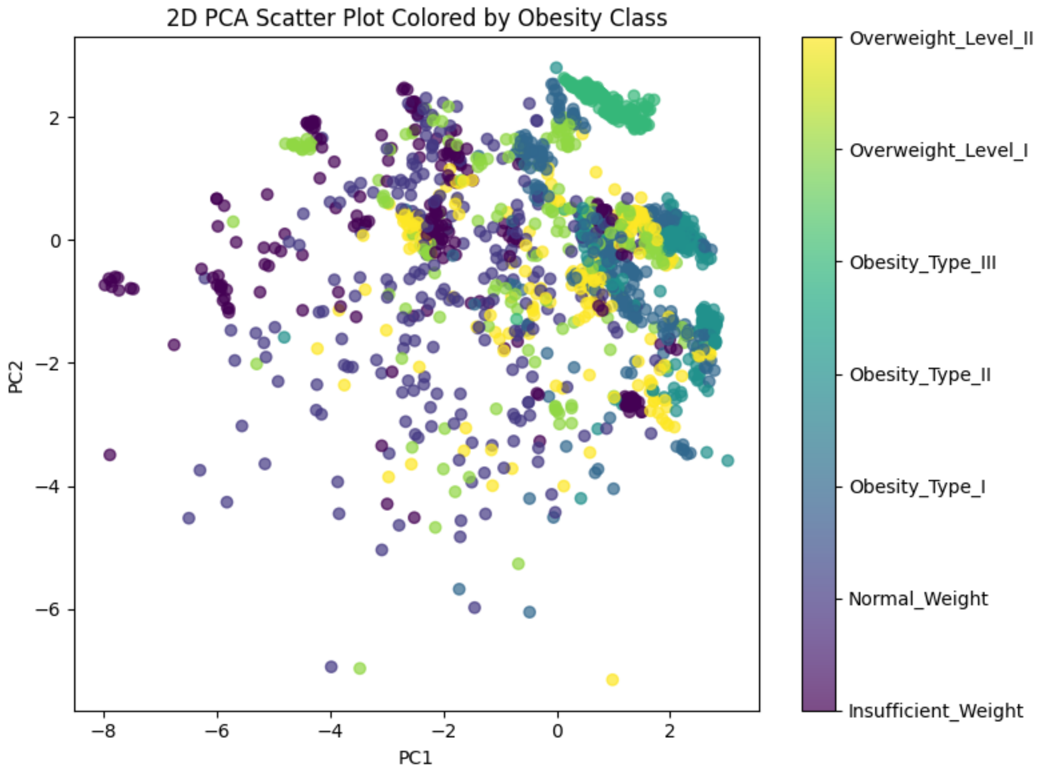

Figure 1: 2D PCA Scatter Plot

The scatter plot in figure 1 shows a two-dimensional PCA representation of many health and lifestyle variables, with each point representing one person and the color showing their obesity category. Overall, the points from different weight classes overlap a lot, meaning that when all the original variables are reduced to just two PCA components, people from different obesity levels do not form clearly separated groups. Some obesity categories, especially the more severe ones, for small clusters, suggesting those individuals share more similar patterns in their data. In contrast, normal-weight and underweight individuals appear more spread out and mixed with other classes, indicating greater variability in their characteristics. Although there are hints of structure, the plot mainly shows that obesity categories blend rather than forming distinct boundaries in this simplified view, highlighting how complex and overlapping the underlying factors really are.

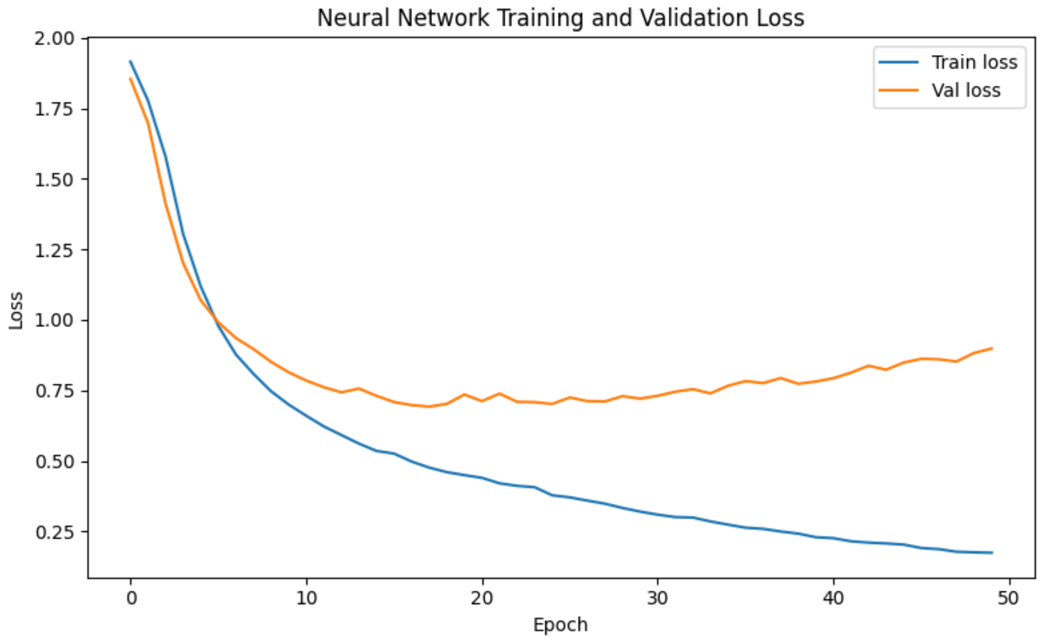

Figure 2: Neural Network Training and Validation Loss

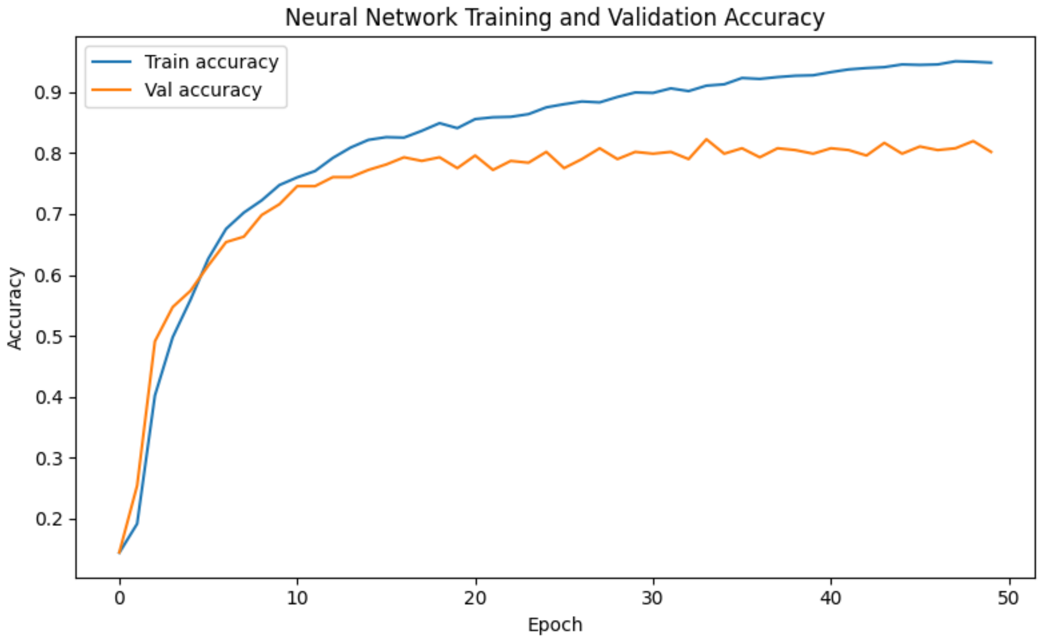

Figure 3: Neural Network Training and Validation Accuracy

These two graphs show how well the neural network learned over time by tracking its loss and accuracy during training and validation. The loss curves in figure 2 show that both training and validation loss drop quickly as first, meaning the model is learning to make better predictions. After around the midpoint of training, the training loss continues to decrease smoothly, but the validation loss levels off and then slowly begins to rise. This pattern indicates overfitting, the model keeps getting better at the specific data it was trained on, but its performance on new, unseen data stops improving and slightly worsens.

The accuracy curves in figure 3 shows a similar story. Both training and validation accuracy increase steadily in the early epochs, meaning the model is learning useful patterns. Eventually the training accuracy continues to climb toward a high value, while the validation accuracy plateaus around the low 80% range and fluctuates slightly. This again suggests that after a certain point, the model is learning details that only fits the training data instead of general patterns that apply to new data. Overall, the results show that the model learns effectively at first, but begins to overfit later, meaning further training past that point doesn’t improve real-world performance.

References

1. UCI machine learning repository [Internet]. [cited 2025 Dec 10]. Available from: https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition