**Predicting Obesity Through Physical Activity**

**UCI Obesity Dataset**

Cason Henderson, [chenderson2@bellarmine.edu](mailto:chenderson2@bellarmine.edu)

**ABSTRACT**

This study dives into the UCI Obesity Dataset to classify individuals as obese based on their physical activity levels. With 2111 rows and 17 columns, the dataset explores variables such as eating habits, physical conditions, and demographic information. The classification objective focuses on predicting obesity levels, ranging from insufficient weight to type III obesity, using physical activity as a key predictor. Multiple machine learning models were developed and tested through various setups, analyzing the effects of feature selection and other adjustments made. The exploratory analysis (EDA) showed notable correlations between physical activity patterns and obesity levels, emphasizing the significance of targeted features. The best-performing model demonstrated high accuracy which highlights the potential of machine learning techniques in predicting obesity. However, challenges such as imbalanced data and reliance of samples highlight some limitations. This project offers insights into the cross between health and technology, paving the way for further research and models in the future.

1. **INTRODUCTION**

The UCI Obesity Dataset contains data collected from individuals in Mexico, Peru, and Colombia, focusing on eating habits and physical activity. The classification target is to predict obesity levels based on physical activity attributes. This study aims to evaluate the feasibility of using physical activity data to classify individuals into obesity categories.

1. **BACKGROUND**

Obesity is a global health concern linked to various chronic diseases. The data set was created to study the relationship between eating habits, physical activity, and obesity levels. It includes both user-input generated data (77%), and real-world data (23%) collected via a web platform. The classification task involves predicting obesity levels categorized as insufficient weight, normal weight, overweight levels I and II, and obesity types I, II, and III.

1. **EXPLORATORY ANALYSIS**

The dataset contains 2111 rows with 17 columns.

Key variables relevant to this study include:​

* FAF: Frequency of physical activity
* Obesity Level: Categorical variable indicating obesity status.

Below is a summary of the data types:

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| Gender | Categorical |
| Age | Numerical |
| Height | Numerical |
| Weight | Numerical |
| Family History with Overweight | Categorical |
| FAVC (High Caloric Food) | Categorical |
| FCVC (Vegetable Consumption) | Categorical |
| NCP (Number of Meals) | Numerical |
| CAEC (Food Between Meals) | Categorical |
| Smoke | Categorical |
| CH20 (Daily Water Intake) | Numerical |
| SCC (Calorie Consumption) | Categorical |
| FAF (Physical Activity Frequency) | Categorical |
| TUE (Time Using Technology) | Numerical |
| CALC (Alcohol Consumption) | Categorical |
| MTRANS (Transportation) | Categorical |
| Obesity Level | Categorical |

Observations:

* No missing values were reported.
* The distribution of weight and height variables is skewed.
* Correlation analysis revealed strong relationships between physical activity and obesity levels.

Unusual Observations:

* Missing Values: Some entries have missing values, noticed in categorical variables like 'Family History' and 'MTRANS'.
* Skewed Distributions: Variables such as 'Age' and 'Weight' displayed skewed distributions, which may require editing to become normal.​

Visualizations:

* Correlation Matrix: A heatmap reveals strong correlations between 'Weight', 'Height', and 'Obesity Level'.
* Bar Charts: Distribution of 'FAF' and 'Obesity Level' categories.
* Scatterplots: To observe patterns between classes color coded by ‘Obesity Level’

1. **METHODS**
   1. *Data Preparation*

* Handling Missing Values: Imputed missing values using the mode for categorical variables and mean for numerical variables.
* Normalization: Applied Min-Max scaling to numerical variables to standardize the range.
* Feature Selection: Retained variables with significant correlation to 'Obesity Level', such as 'FAF' and 'Weight'.
  1. *Experimental Design*

The dataset was split into training and testing sets using a 70/30 ratio. Several machine learning models were trained and evaluated below:

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | Used all 17 features with an 80/20 train-test split. Random Forest classifier with default hyperparameters. |
| 2 | Focused on physical activity-related features only (e.g., CAEC, FCVC, NCP), using a Support Vector Machine (SVM) with a 70/30 train-test split. |
| 3 | Included demographic attributes (e.g., Gender, Age) and physical activity features. Applied Logistic Regression with cross-validation (k=5). |
| 4 | Dropped cells like CAEC and SMOKE due to low correlation with the target variable. Used a Gradient Boosting model with an 80/20 train-test split, and parameters (e.g., learning rate = 0.1, max\_depth = 5). |

* 1. *Tools Used*

The analysis was conducted using Python 3.8.5 within the Anaconda 4.9.2 environment. Key libraries included:​

* Pandas: Data manipulation and analysis.
* NumPy: Numerical computing.
* Matplotlib & Seaborn: Data visualization.
* Scikit-learn: Machine learning algorithms and evaluation metrics.​

I selected these tools for their ease of use, as well as their extensive and specific documentation.

1. **RESULTS**
   1. *Classification Measures/ Accuracy measure*

The models were evaluated using accuracy, precision, recall, and F1-score. The Random Forest model demonstrated the highest performance across these metrics.

* 1. *Discussion of Results*

The Random Forest model being the best performer can be caused by its ability to handle complex interactions between features and its strength against overfitting. In contrast, the much simpler Logistic Regression model showed lower performance due to its linear nature, which may not capture the complexities of the data.

* 1. *Problems Encountered*
* Imbalance of Data: The 'Obesity Level' variable had imbalanced classes, which could affect model performance.
* Feature Engineering: Determining the most relevant features for prediction required repeated testing.
  1. *Limitations of Implementation*

The models did not account for potential variables such as dietary habits and socioeconomic status, which could influence obesity levels. Additionally, the dataset's cross-sectional nature limits the ability to link causality.

* 1. *Improvements/Future Work*

Future studies could implement longitudinal data to examine causal relationships. They could include additional features like dietary intake and socioeconomic factors which could improve the model’s accuracy. Employing advanced techniques like combination methods could improve performance.

1. **CONCLUSION**

This study aimed to classify individuals as obese based on physical activity using the UCI Obesity Dataset. Among the models tested, the Random Forest classifier delivered the best performance, which highlighted the predictive value of physical activity frequency. The results suggest that machine learning can effectively support obesity risk assessment when it is combined with behavioral data. Future work could improve results by incorporating more features and testing advanced combinations or deep learning models.

**REFERENCES**

* UCI Machine Learning Repository
* Python libraries: Pandas, Scikit-learn, Seaborn, Matplotlib