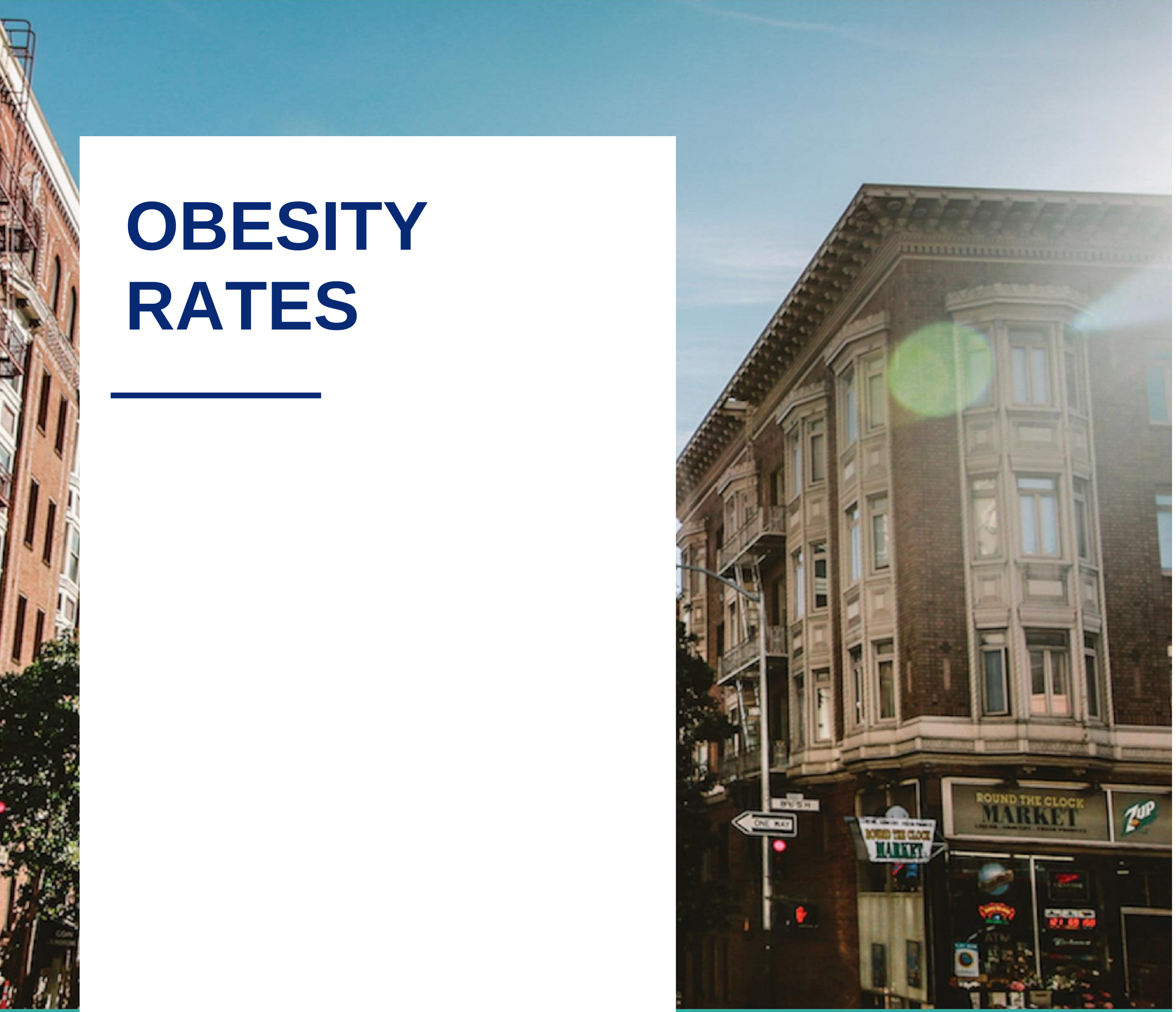


OBESITY RATES

NOVEMBER 4

Authored by: ASSIA ERGUIGUE

GitHub (code, visualizations, report):
[Obesity_rates_repository]

HUBBLEMIND
Innovate IT

Contents

Exploratory Data Analysis (EDA) 3

1. Summary Statistics 3
2. Distribution Analysis 3
3. Relationship Exploration 3
4. Additional Observations: 4

Preprocessing & Advanced visualizations: 4

Feature engineering & Modeling: 5

I. Feature engineering 5

II. Modeling 5

1. Logistic Regression 5
2. Random Forest Classifier 5
3. K-Nearest Neighbors (KNN) 5
4. Support Vector Machine (SVM) 6
5. XGBoost 6

Conclusion and Model Comparison 7

Model evaluation: 7

Exploratory Data Analysis (EDA)

1. Summary Statistics

- **Objective**: Obtain an initial understanding of each feature's range, mean, median, and variance.
- **Findings**:
  - Continuous variables, like age, caloric intake, and physical activity levels, were distributed across a wide range, necessitating data scaling before modeling.
  - Categorical variables such as Gender, Family history of being overweight, and Frequent consumption of high-caloric food were also analyzed to understand their distribution within each obesity level.
  - Obesity levels, the target variable, had a balanced representation across the dataset, which supports unbiased model training.

2. Distribution Analysis

- **Objective**: Analyze the distribution of obesity levels across different demographic and lifestyle factors.
- **Visualizations Used**: Histograms, bar charts, and box plots.
- **Insights**:
  - **Age and Obesity Levels**: The dataset showed an age gradient, with certain obesity levels skewed towards older age groups.
  - **Gender**: Male and female distribution did not significantly differ in obesity prevalence, but lifestyle factors contributing to obesity showed gender-related tendencies.
  - **High-Caloric Food Consumption**: Frequent consumption of high-calorie foods had a noticeable correlation with higher obesity levels.
  - **Physical Activity Levels**: Lower activity levels correlated strongly with higher obesity categories, especially in individuals classified as Overweight Level II and Obesity Type III.

3. Relationship Exploration

- **Objective**: Identify relationships between key features to detect multicollinearity or feature importance for obesity prediction.

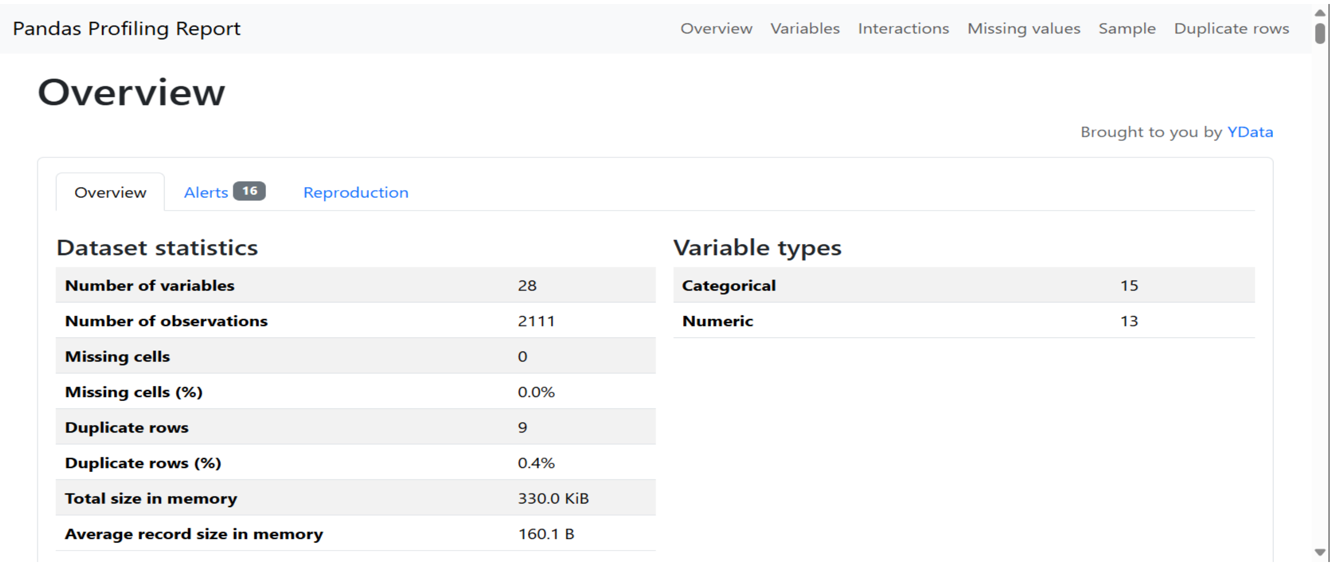
- **Visualizations Used**: Pair plots and correlation heatmaps.
- **Findings**:
  - **Correlations**: Features like daily caloric intake, frequency of physical activity, and consumption of vegetables showed notable correlations with obesity levels.
  - **Multicollinearity**: The correlation matrix indicated some multicollinearity between dietary habits and obesity levels, suggesting that combined dietary factors could be powerful predictors in the modeling phase.

4. Additional Observations:

- The combination of sedentary lifestyle indicators (low physical activity and high-caloric intake) was found to be a strong predictor of higher obesity levels.
- Overall, age and gender had less direct influence on obesity compared to lifestyle-related factors, reinforcing the focus on dietary and activity modifications for obesity prevention strategies.

Preprocessing & Advanced visualizations:

After the preprocessing part, we use the library **ydata_profiling**; this leads to an HTML page containing all information on the new data, including the relations and insights where some are discussed in the previous part. This file is attached along-side this project.For instance, here’s the overview of the encoded data:



Overview

Dataset statistics

Number of variables	28
Number of observations	2111
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	9
Duplicate rows (%)	0.4%
Total size in memory	330.0 KiB
Average record size in memory	160.1 B

Variable types

Categorical	15
Numeric	13

Feature engineering & Modeling:

I. Feature engineering

We created three new columns that are heavily related to the precision of obesity rates in general, presented in the following table:

BMI	Body mass index is a value derived from the mass and height of a person. The BMI is defined as the body mass divided by the square of the body height
DIET SCORE	To express the type of food consumed by people and how healthy are they.
LIFESTYLE SCORE	Calculate Lifestyle Score (FAF, TUE, and MTRANS related to activity/sedentary level)

II. Modeling

1. Logistic Regression

- **Performance**:
    - **Accuracy**: Moderate, with better classification for extreme categories.
    - **Precision and Recall**: Moderate for most classes, with notable misclassifications between adjacent obesity levels.
- **Confusion Matrix Analysis**: Highlighted some misclassification between adjacent obesity levels, suggesting the need for more nuanced features.

2. Random Forest Classifier

- **Performance**:
    - **Accuracy**: Outperformed Logistic Regression, especially for Overweight and Obesity Type III classes.
    - **Feature Importance**: Key predictors included daily caloric intake, physical activity, and high-caloric food consumption.

3. K-Nearest Neighbors (KNN)

- **Model Choice**: KNN was chosen for its simplicity in capturing proximity-based patterns.

- **Performance**:
    - **Accuracy**: Performed reasonably well for Normal Weight and Obesity Type III but struggled with intermediate obesity levels, likely due to feature scaling sensitivity.
    - **Precision and Recall**: Moderate recall for extreme classes; however, precision varied, showing inconsistency for borderline obesity levels.
- **Limitations**: KNN's performance was sensitive to feature scaling and required tuning of the number of neighbors (k) for optimal results.

4. Support Vector Machine (SVM)

- **Model Choice**: SVM was selected for its robustness in high-dimensional spaces and its ability to handle complex decision boundaries.
- **Performance**:
    - **Accuracy**: Achieved strong accuracy across most classes, with improvements in classifying borderline categories (e.g., Overweight Level I vs. Normal Weight).
    - **Precision and Recall**: High precision for Obesity Type III and Overweight categories, though slightly lower recall in predicting Normal Weight.
- **Kernel Choice**: A non-linear kernel (such as RBF) was effective, capturing complex patterns in the data, particularly for lifestyle-related obesity classes.

5. XGBoost

- **Model Choice**: XGBoost was applied for its powerful gradient-boosting approach and capability to capture complex feature interactions.
- **Performance**:
    - **Accuracy**: XGBoost outperformed other models, especially for higher obesity levels.
    - **Precision and Recall**: Highest precision and recall across all classes, particularly effective in distinguishing adjacent obesity levels.
- **Feature Importance**: XGBoost's feature importance analysis confirmed that daily caloric intake, activity level, and caloric food consumption were crucial, but it also revealed interactions not seen in other models, underscoring its nuanced predictive power.
- **Confusion Matrix Analysis**: Minimal misclassification, indicating robust generalization and sensitivity to the underlying data structure.

Conclusion and Model Comparison

- **Best Performing Model**: XGBoost achieved the highest overall accuracy and precision-recall balance, demonstrating strong potential for obesity classification based on lifestyle factors.
- **Secondary Models**: SVM and Random Forest also showed strong performance and are viable alternatives depending on computational resources and interpretability requirements.

Model evaluation:

Models	Evaluation metrics			
	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.69	0.67	0.69	0.67
Random Forest Classifier	0.95	0.95	0.95	0.95
XGBOOST	0.95	0.95	0.95	0.95
SVM	0.95	0.95	0.95	0.95
KNN	0.95	0.95	0.95	0.95