

Submitted by:

Rohit Gambhir

|  |  |  |
| --- | --- | --- |
| Section Number | Title | Subsections |
| 1 | Introduction | - Overview of the Obesity Prediction Project- Dataset Description and Objectives |
| 2 | Dataset Information | - Attributes Overview- Key Variables |
| 3 | Data Exploration and Preprocessing | - Data Loading and Initial Checks- Summary of Missing Values and Duplicates |
| 4 | Feature Analysis | - Univariate Analysis- Bivariate Analysis- Outlier Detection |
| 5 | Multicollinearity Analysis | - Correlation Matrix and Heatmap- Variance Inflation Factor (VIF) |
| 6 | Dimensionality Reduction | - Principal Component Analysis (PCA)- Clustering Insights |
| 7 | Predictive Modeling | - Logistic Regression- Random Forest- Support Vector Machines (SVM) |
| 8 | Prediction Insights | - Key Predictors Identified- Strengths and Weaknesses of Models |
| 9 | Conclusion and Recommendations | - Key Findings- Actionable Recommendations |
| 10 | Citations and Data Source | - Dataset Source |

**Table of Contents**

**Predicting Obesity Levels for Targeted Health Interventions**

**1. Introduction**

The primary objective of this analysis was to predict obesity levels in individuals from Mexico, Peru, and Colombia based on their eating habits and physical conditions. This analysis aims to address the growing obesity epidemic in these countries by identifying high-risk individuals and enabling targeted health interventions. The dataset used, ObesityDataSet\_raw\_and\_data\_sinthetic.csv, includes data for 2111 individuals, with 17 attributes related to eating habits and physical conditions.

This report strictly reflects the analyses conducted in the Python file, detailing steps, results, and their alignment with the business problem. Every insight provided is rooted in the findings from the data and model outcomes, ensuring relevance to actionable strategies.

**2. Dataset Information**

The dataset includes records from Mexico, Peru, and Colombia. The class variable NObesity classifies individuals into categories such as Insufficient Weight, Normal Weight, Overweight Levels I and II, and Obesity Types I, II, and III. Key dataset features include:

* **Numerical Variables**: Age, Height, Weight.
* **Categorical Variables**: Gender, family history of overweight, eating habits, and transportation mode.

The dataset ensures representation across obesity levels, making it suitable for building predictive models. By including attributes related to physical conditions and eating habits, it provides comprehensive coverage of factors influencing obesity, enabling informed decision-making for health interventions.

**3. Data Exploration and Preprocessing**

**3.1 Data Loading and Initial Checks**

The dataset was loaded and checked for structure, missing values, and duplicates. This step was critical to ensure the reliability of subsequent analyses.

print("First 5 records:")

print(df.head())

print("\nMissing entries per variable:")

print(df.isnull().sum())

print("\nNumber of duplicate records:")

print(df.duplicated().sum())

The dataset contained no missing values or duplicates, ensuring data quality and reliability for analysis.

Variable types were consistent with expectations, with numerical data for attributes like Weight and categorical data for variables like Gender and Obesity Level.

This clean dataset provided a strong foundation for exploratory and predictive analyses, reducing the risk of inaccuracies.

**3.2 Feature Analysis**

***Univariate Analysis***

To analyse distributions of individual variables (Age, Height, Weight) and understand their role in predicting obesity levels.

**Code Snippet:**

for var in numerical\_vars:

sns.histplot(df[var], kde=True, bins=30)

plt.title(f"Distribution of {var}")

plt.show()  
  
  
A graph of a number of age

Description automatically generated  
**Age:**

* Most individuals are between 18 and 50 years old, representing a working-age population. This aligns well with the target demographic for obesity-related interventions as this group is likely to benefit most from preventive and corrective measures.
* A smaller representation of younger and older individuals highlights potential gaps in addressing obesity for these age groups. Future studies may benefit from a more balanced age distribution to develop insights applicable across all demographics.  
    
  A diagram of a distribution of weight

  Description automatically generated

**Weight:**

* The weight distribution spans a wide range, with significant clusters in the healthy and overweight categories. Extreme weights, often associated with "Obesity Type II" and "Type III," underscore the urgency of focusing on individuals in these categories. These individuals represent the most critical group for intervention, including dietary and medical management.  
    
  A diagram of a distribution of height

  Description automatically generated

**Height:**

* Heights follow a near-normal distribution, with a few outliers. These outliers could represent rare cases, potential data entry errors, or unique physical conditions. Addressing these anomalies ensures that model performance remains robust and unbiased.

***Bivariate Analysis***

To explore relationships between variables, particularly Height, Weight, and Obesity Levels, to uncover patterns that can guide predictive modelling and interventions.

**Code Snippet:**

sns.scatterplot(x='Height', y='Weight', hue='NObesity', data=df)

plt.title("Height vs Weight by Obesity Level")

plt.show()

A diagram of weight vs height by obesity

Description automatically generated

**Height vs. Weight:**

The scatterplot reveals distinct clustering of individuals by obesity levels. For instance:

* Individuals in "Obesity Type III" exhibit significantly higher weights across a range of heights, confirming weight as a dominant predictor.
* "Normal Weight" individuals cluster in a narrow range of both height and weight, reinforcing expected trends for healthy individuals.

Overlapping clusters in "Overweight Level I" and "Level II" categories suggest that other factors, such as eating habits and physical activity, are necessary to enhance classification accuracy.

These patterns validate the importance of weight as a critical feature for distinguishing obesity levels while emphasizing height's complementary role, particularly in borderline cases.

***How These Results Align with our problem:***

The univariate and bivariate analyses provide actionable insights into the physical characteristics driving obesity levels. By understanding these patterns:

* Intervention programs can focus on individuals with extreme weights, particularly in "Obesity Type II" and "Type III."
* Preventive strategies can target "Overweight Level I" and "Normal Weight" individuals to prevent progression to more severe categories.
* These insights validate the dataset’s utility for predictive modeling, ensuring reliable identification of high-risk groups and optimal resource allocation.

***Outlier Detection***

Outliers in numerical variables were detected using the Local Outlier Factor (LOF).

**Code Snippet:**

lof = LocalOutlierFactor()

outlier\_labels = lof.fit\_predict(X)

outliers = X[outlier\_labels == -1]

print(f"Number of outliers detected: {len(outliers)}")

Several outliers were identified and flagged. These observations are likely to skew model predictions if not addressed. Removing or re-evaluating these outliers ensures more robust and reliable modelling outcomes, leading to improved classification accuracy.  
  
  
***Multicollinearity Analysis***

Multicollinearity occurs when two or more predictor variables are highly correlated, leading to inflated variances in regression coefficients. This issue can significantly impact the stability of linear models, such as Logistic Regression, by making the estimated coefficients less reliable. Addressing multicollinearity ensures the interpretability and robustness of the predictive models.  
  
**Correlation Matrix**

A correlation matrix was computed to measure the strength of linear relationships between numerical variables. The matrix provided pairwise correlation coefficients for every combination of features, which were then visualized using a heatmap for clarity.

Variables with correlation coefficients above 0.8 (absolute value) were flagged as highly correlated, indicating potential redundancy.  
  
**Results:**

* The heatmap revealed certain strong correlations between numerical variables. For instance, Height and Weight exhibited a high correlation coefficient (>0.8), suggesting that these variables may share overlapping information.
* Other variables, such as Age, showed weaker correlations, indicating they might contribute unique, independent information to the models.

**Interpretation:**

* Features with high correlations often provide redundant information, leading to inefficiencies in the model. For example, if both Height and Weight are highly correlated, including both in the model may not improve predictive power but could increase complexity and instability.
* Identifying these relationships is crucial to simplify the model without losing significant predictive information. By addressing multicollinearity, we ensure that the features used are truly independent and meaningful.

**Variance Inflation Factor (VIF):**   
  
VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A higher VIF score indicates a stronger dependency among variables, making the affected coefficients less stable and interpretable.

Typically, features with VIF > 10 are considered problematic and are candidates for removal or transformation.  
  
**Results:**

* The VIF table revealed that certain features, such as Weight, had low VIF scores (< 5), indicating minimal multicollinearity. However, some features like Height had VIF values exceeding 10, highlighting potential redundancy.
* Features with high VIF scores were carefully analyzed to decide whether they should be removed or transformed.

**Interpretation:**

* Variables with high VIF scores contribute to instability in linear models. For instance, if Height and Weight both show high VIF values, one of these variables might be removed or combined with others to simplify the model.
* Addressing these issues ensures that the dataset is optimized for predictive modeling, particularly for models like Logistic Regression, which are sensitive to multicollinearity.

**Code Snippet:**

# Calculate correlation matrix

correlation\_matrix = df[numerical\_vars].corr()

sns.heatmap(correlation\_matrix, annot=True, cmap='bwr')

plt.title("Correlation Matrix")

plt.show()

# Identify pairs of highly correlated variables

high\_correlations = np.where(np.abs(correlation\_matrix) > 0.8)

high\_correlations = [(correlation\_matrix.index[x], correlation\_matrix.columns[y], correlation\_matrix.iloc[x, y])

for x, y in zip(\*high\_correlations) if x != y and x < y]

print("\nHigh correlations (>0.8):")

for var1, var2, corr in high\_correlations:

print(f"{var1} and {var2}: {corr:.2f}")

# Calculate VIF for each feature

vif\_data = pd.DataFrame()

vif\_data["Feature"] = X.columns

vif\_data["VIF"] = [variance\_inflation\_factor(X.values, i) for i in range(X.shape[1])]

print(vif\_data)  
  
A red and blue squares

Description automatically generated  
  
 **Action Taken:**

* Features with high VIF scores were either transformed (e.g., through scaling or creating derived features) or considered for removal to improve model stability and interpretability.

**4. Dimensionality Reduction**

Dimensionality reduction was performed to simplify the dataset while retaining the most informative features. Principal Component Analysis (PCA) was used to condense the high-dimensional data into a smaller number of principal components, preserving most of the variance.

PCA was applied to the standardized numerical features to extract the first two principal components.

The first two components were visualized in a scatterplot, with data points coloured by their obesity levels to observe clustering patterns.

**Code Snippet:**

pca = PCA(n\_components=2)

df\_pca = pca.fit\_transform(df[numerical\_vars])

plt.scatter(df\_pca[:, 0], df\_pca[:, 1], c=df['NObesity'].factorize()[0])

plt.title("PCA Result")

plt.show()  
  
A screen shot of a computer screen

Description automatically generated

PCA revealed distinct clusters corresponding to different obesity levels. These clusters highlight the potential for targeted interventions tailored to specific groups.

For example, clusters representing "Obesity Type II" and "Type III" can guide resource prioritization for high-risk groups, while tightly grouped clusters for "Normal Weight" suggest effective preventive measures.

Dimensionality reduction through PCA ensures model efficiency while maintaining the interpretability of key obesity-related patterns.  
  
**Results:**

* The first two components explained a significant proportion of the variance in the dataset.
* Distinct clusters emerged in the scatterplot:
  + Individuals classified as Normal Weight and Insufficient Weight formed compact, distinct groups.
  + Higher obesity levels, such as Obesity Type II and Type III, were more spread out but formed separate clusters.

**Interpretation:**

* PCA revealed that the dataset contains clear separable patterns, indicating that obesity levels can be distinguished using the available features.
* This step reduced the dataset's complexity, making predictive models more efficient without sacrificing interpretability. The clustering patterns also validate that the dataset is well-suited for classification tasks.

**5. Predictive Modelling**

***Logistic Regression***

To create a baseline classification model and evaluate its performance.

**Code Snippet:**

lr = LogisticRegression()

lr.fit(X\_train, y\_train)

y\_pred\_lr = lr.predict(X\_test)

print(classification\_report(y\_test, y\_pred\_lr))

Logistic regression provided moderate accuracy, reflecting its limitations in capturing non-linear relationships present in the data.

As a baseline model, logistic regression offers a starting point for comparison against more complex models like Random Forest and SVM.

***Random Forest***

To improve prediction accuracy and identify critical features influencing obesity levels.

**Code Snippet:**

rf = RandomForestClassifier()

rf.fit(X\_train, y\_train)

print(classification\_report(y\_test, rf.predict(X\_test)))

Random forest outperformed logistic regression, demonstrating its ability to capture non-linear relationships and handle complex datasets.

Feature importance scores revealed that Weight and Physical Activity are the most significant predictors, aligning with insights from exploratory analysis.

This model provides a robust framework for identifying high-risk individuals and tailoring interventions accordingly.

***Support Vector Machines (SVM)***

To evaluate the performance of non-linear models and compare them against Random Forest.

**Code Snippet:**

svc = SVC()

svc.fit(X\_train, y\_train)

print(classification\_report(y\_test, svc.predict(X\_test)))

SVM provided competitive accuracy, particularly for separating overlapping categories such as "Overweight Level I" and "Level II."

The flexibility of kernel functions allows SVM to adapt to complex decision boundaries, making it a valuable tool for refining classifications.

**Prediction Insights with Metrics**

The classification models were evaluated using metrics like **precision**, **recall**, **F1-score**, and **support**, which offer critical insights into each model’s performance and suitability for predicting obesity levels across categories.

***Logistic Regression   
  
A screenshot of a computer

Description automatically generated***

**Overall Performance:**

Accuracy: 87%

Macro Average (F1-Score): 86%

Weighted Average (F1-Score): 87%

**Class-Wise Performance:**

For Insufficient Weight, the model achieved **100% recall**, meaning it identified all individuals in this category correctly, though its precision was slightly lower at **86%**.

For Normal Weight, the recall was **63%**, indicating significant misclassification of individuals in this category.

For overlapping categories like Overweight Level I, recall was **73%**, showing difficulty in distinguishing it from related classes.

**Strengths:**

Logistic Regression provides interpretable and consistent results for classes like Insufficient Weight and Obesity Type III.

It is computationally inexpensive and serves as a useful baseline for comparing more advanced models.

**Weaknesses:**

The low recall for Normal Weight highlights its struggle to differentiate between closely related obesity levels.

Sensitivity to multicollinearity and linearity assumptions further limit its effectiveness for this dataset.

***Random Forest  
  
A screenshot of a computer

Description automatically generated***

**Overall Performance:**

Accuracy: 94%

Macro Average (F1-Score): 94%

Weighted Average (F1-Score): 94%

**Class-Wise Performance:**

Perfect **100% precision and recall** for critical categories like Obesity Type III and Obesity Type II, demonstrating its reliability in identifying high-risk individuals.

Even for challenging categories like Overweight Level I, it achieved a precision of **91%** and recall of **86%**.  
  
**Strengths:**

Random Forest performed consistently across all classes, with minimal trade-offs between precision and recall.

The **support metric**, which measures the number of test samples per class, shows that Random Forest handled imbalanced classes like Insufficient Weight (support = 56) and Obesity Type III (support = 53) without compromising accuracy.

Feature importance analysis identifies actionable predictors like Weight, Physical Activity, and Caloric Intake, aligning with public health knowledge.

**Weaknesses:**

The model is computationally intensive, making it less suitable for real-time predictions.

***Support Vector Machines (SVM)   
  
A screenshot of a computer

Description automatically generated***

**Overall Performance:**

Accuracy: 98%

Macro Average (F1-Score): 98%

Weighted Average (F1-Score): 98%

**Class-Wise Performance:**

Achieved near-perfect classification for almost all categories, with **100% precision and recall** for Obesity Type II and Obesity Type III.

For Normal Weight, it achieved **92% precision** and **95% recall**, showing improvement compared to Logistic Regression and Random Forest.

**Strengths:**

SVM excels in separating overlapping categories like Overweight Level I and Overweight Level II, as reflected in its high precision and recall for these classes.

The ability to adapt complex decision boundaries using kernel functions makes it particularly effective for this dataset.

**Weaknesses:**

High computational cost and sensitivity to hyperparameter tuning limit its scalability for larger datasets or real-time applications.

**Comparative Insights Across Metrics**

1. **Precision and Recall Across Models:**
   * **Logistic Regression:** Struggled with recall for overlapping categories like Overweight Level I, resulting in a lower overall F1-score.
   * **Random Forest:** Delivered high precision and recall for most categories, balancing performance across both dominant and minority classes.
   * **SVM:** Provided near-perfect precision and recall, making it the most accurate model for distinguishing closely related obesity levels.
2. **Support and Class Balance:**
   * The **support metric** highlights the distribution of test samples per class:
     + Classes like Insufficient Weight (support = 56) and Obesity Type III (support = 53) were handled effectively by Random Forest and SVM.
     + Logistic Regression struggled with smaller classes, leading to lower recall.
3. **Actionable Predictors:**
   * Across all models, Weight, Physical Activity, and Caloric Intake consistently emerged as key predictors. These features provide actionable insights for targeting high-risk groups and designing public health interventions.
4. **Overall Accuracy:**
   * Logistic Regression: 87% – A strong baseline model for simpler tasks but limited in complex datasets.
   * Random Forest: 94% – The best balance of accuracy, interpretability, and robustness.
   * SVM: 98% – The most accurate model, particularly for overlapping classes, but with higher computational requirements.

**Extended Analysis of Model Performance**

The additional evaluation metrics provide deeper insights into the performance of each model, including training vs. testing accuracy, cross-validation scores, and an improved logistic regression model using cross-validation. These results further refine our understanding of how well the models predict obesity levels and their implications for the business problem.

**Training vs Testing Scores**

The training and testing accuracy scores for each model reveal how well they generalize to unseen data:

***Logistic Regression:***

**Training Accuracy:** 98.58%

**Testing Accuracy:** 95.74%

Logistic Regression achieves excellent performance on both training and testing datasets, showing minimal overfitting. Improvements made using LogisticRegressionCV optimized hyperparameters and enhanced generalization.

***Random Forest:***

**Training Accuracy:** 99.94%

**Testing Accuracy:** 94.09%

Random Forest has near-perfect training accuracy but slightly lower testing accuracy, indicating mild overfitting. Despite this, its testing performance remains high and reliable for unseen data.

***Support Vector Machines (SVM):***

**Training Accuracy:** 98.16%

**Testing Accuracy:** 97.87%

SVM demonstrates the highest testing accuracy, reflecting its ability to generalize well to unseen data. This model excels in handling complex patterns and separating overlapping categories.

* Logistic Regression offers a balance between accuracy and simplicity, making it suitable for scalable applications.
* Random Forest provides robust predictions and insights into key predictors, ideal for guiding interventions.
* SVM’s precise classification makes it valuable for targeted campaigns or clinical settings.

**Cross-Validation Results**

Cross-validation evaluates model stability by splitting the training data into multiple folds. The mean cross-validation scores provide a robust measure of model performance:

1. **Logistic Regression:**
   * **CV Scores:** [96.15%, 93.49%, 94.08%, 94.65%, 97.32%]
   * **Mean CV Score:** 95.14%
   * **Interpretation:** Logistic Regression exhibits consistent performance across folds, indicating reliability and robustness.
2. **Random Forest:**
   * **CV Scores:** [93.19%, 93.19%, 95.26%, 92.58%, 96.14%]
   * **Mean CV Score:** 94.07%
   * **Interpretation:** Random Forest delivers high accuracy across folds but shows slightly more variation compared to Logistic Regression.
3. **SVM:**
   * **CV Scores:** [96.45%, 94.97%, 94.67%, 95.25%, 97.92%]
   * **Mean CV Score:** 95.85%
   * **Interpretation:** SVM achieves the highest mean CV score, reinforcing its consistency and strength in predicting obesity levels.

Cross-validation confirms the robustness of all models, with SVM and Logistic Regression being the most consistent across folds.

Random Forest remains reliable for handling non-linear patterns but may require further tuning to reduce variability.

**Improved Logistic Regression Performance**

Using LogisticRegressionCV significantly enhanced the performance of Logistic Regression:

**Overall Accuracy:** 96%

**Class-Wise Performance:**

* Precision and recall scores were high across all categories, with F1-scores ranging from 92% to 100%.
* Perfect classification was achieved for Obesity Type III and strong performance was observed for overlapping categories like Normal Weight (Precision: 96%, Recall: 89%).

**Improvements:**

* The optimized Logistic Regression model now handles overlapping categories effectively.
* Cross-validation ensures better generalization, reducing the risk of overfitting.

**6. Conclusion**

Weight and Physical Activity are the most significant predictors of obesity levels, underscoring their role in targeted interventions.

Random forest emerged as the best-performing model, offering high accuracy and actionable insights through feature importance.

Logistic regression serves as a baseline model but is limited in handling the complexity of the data.

SVM demonstrated strong performance in managing overlapping categories and refining predictions.

1. **Recommendations**
2. **High-Risk Identification:** Focus interventions on individuals in "Obesity Type II" and "Type III" categories, prioritizing medical and dietary support for these groups.
3. **Awareness Campaigns:** Use insights from feature importance to design targeted campaigns emphasizing weight management and increased physical activity.
4. **Resource Allocation:** Leverage clustering insights from PCA to allocate resources efficiently, ensuring high-risk areas receive adequate support.
5. **Model Optimization:** Expand and validate the dataset to enhance model generalizability, incorporating additional lifestyle factors for improved accuracy.  
     
     
     
     
     
   ***Citations and Data Source:*** *Dua, D., & Graff, C. (2019). Estimation of obesity levels based on eating habits and physical condition. UCI Machine Learning Repository. Retrieved from* [*https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition*](https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition)*.*