1. INTRODUCTION

Obesity is a significant global health concern, as it is associated with various chronic diseases such as diabetes, heart disease, and hypertension. Understanding the factors that contribute to obesity is crucial for developing effective prevention strategies and interventions. This study aims to predict obesity levels based on individuals' lifestyle choices and food consumption habits.

By identifying key contributors to obesity, this model can help in designing targeted preventive health programs. The insights gained from this research can be valuable to multiple stakeholders. Healthcare professionals can use the model for early detection and intervention, allowing them to provide tailored recommendations to patients at risk. Researchers can further explore the relationships between different lifestyle factors and obesity, contributing to a deeper understanding of the issue. Additionally, individuals can benefit by gaining awareness of how their daily habits influence their likelihood of obesity, encouraging healthier lifestyle choices.

2. DATA

The dataset used in this study contains information on obesity levels in individuals from Mexico, Peru, and Colombia, aged between 14 and 61. It was collected through an online survey, resulting in 2,111 records with 17 attributes. The dataset includes factors related to eating habits, physical activity, and lifestyle choices.

The attributes can be categorized into two main groups: lifestyle factors and food habits. The lifestyle factors include gender, age, smoking status, frequency of physical activity, time spent using technological devices, and the primary mode of transportation used. The food habits category consists of high-caloric food consumption, vegetable intake, the number of main meals per day, snack consumption, alcohol consumption, and daily water intake.

The target variable, NObeyesdad, categorizes individuals into different obesity levels, ranging from Insufficient Weight to Obesity Type III, based on the standards set by the World Health

Organization (WHO) and national health guidelines. Given its combination of categorical and numerical data, this dataset is well-suited for classification and predictive analysis.

3. EXPLORATORY DATA ANALYSIS

To better understand the dataset, we conducted an exploratory data analysis, visualizing key relationships between obesity levels and various features.

A **boxplot of BMI and obesity levels (Figure A)** revealed that Overweight Level I, Insufficient Weight, and Obesity Type III exhibit noticeable outliers. Overweight Level I contains individuals with extreme values in weight, height, or other physical factors. Insufficient Weight also has some individuals with unusually low weight values compared to the general trend.

A **stacked bar plot (Figure B)** comparing obesity levels by gender shows that Obesity Type II is predominantly observed in males, while Obesity Type III has a higher proportion of females.

Examining the relationship between family history of overweight and obesity levels, a **bar graph (Figure C)** indicates that individuals with a family history of overweight ("yes") have the highest counts in Obesity Type I and Obesity Type II, both exceeding 300. Overweight Levels I and II follow closely behind, with counts around 250-300. In contrast, individuals without a family history of overweight ("no") are most frequently found in the Normal Weight and Insufficient Weight categories, with much lower counts in the higher obesity levels. Notably, Obesity Type III is nearly negligible in this group.

Finally, a **bubble chart (Figure D)** visualizing height and weight, with obesity levels represented by color intensity, highlights a clear trend. The lighter-colored bubbles (higher obesity levels) are concentrated in the upper weight range (above ~100 kg), while the darker-colored bubbles (lower obesity levels) are found in the lower weight range (~40-70 kg). The gradient from low to high obesity levels is smooth, reinforcing that obesity increases progressively rather than in abrupt jumps. Obesity Type III, representing severe obesity, includes individuals with the highest weight values, often surpassing 100 kg, and correspondingly high BMI values.

4. METHODOLOGY

In this section, we describe the machine learning models used to predict obesity levels and the insights gained from each model. We began by experimenting with **Linear Regression** as a baseline model to establish a simple relationship between the features and the target variable (obesity levels). However, the model performed poorly, with an R^2 score of **0.12** and an Adjusted R^2 score of **0.10**, indicating that it explained only a small fraction of the variance in the data. The Mean Squared Error (MSE) was **3.24**, and the Root Mean Squared Error (RMSE) was **1.80**, further confirming the model's inability to capture the underlying patterns (see **Figure E** in the Appendix for detailed metrics). These results suggest that the relationship between the features and obesity levels is not linear, and a more sophisticated model is required.

Next, we applied **Polynomial Regression** with degree 2 to capture non-linear relationships. While the R^2 score improved slightly to **0.17**, the Adjusted R^2 score dropped to **0.07**, indicating that the model was overfitting the data. The MSE and RMSE values were **3.08** and **1.75**, respectively, which were only marginally better than Linear Regression (see **Figure F** in the Appendix for detailed metrics). These results reinforced the need for a model that could handle the complexity of the data without overfitting.

We then turned to **Decision Tree**, which was applied to the **Food Habits** dataset. The model achieved an accuracy of **59.34%**, with a macro average F1-score of **0.61** and a weighted average F1-score of **0.60**. While the model showed high precision for "Severe Obesity" (0.92), it struggled with recall for "Insufficient Weight" (0.52) and "Overweight" (0.44) (see **Figure G** in the Appendix for the full classification report). This indicated that the Decision Tree model, while interpretable, was not sufficiently robust for this classification task.

To improve performance, we implemented **Random Forest**, which was applied to both the **Food Habits** and **Lifestyle Habits** datasets separately. For the **Food Habits Dataset**, the model achieved an accuracy of **69.98%**, with a macro average F1-score of **0.70** and a weighted average F1-score of **0.70**. The model showed high precision for "Severe Obesity" (0.87) and recall for "Obesity" (0.78), indicating a good ability to identify severe obesity cases correctly (see **Figure H** in the Appendix for the full classification report). For the **Lifestyle Habits Dataset**, the model

achieved an accuracy of **61.23%**, with a macro average F1-score of **0.64** and a weighted average F1-score of **0.62**. The model performed well for "Healthy Weight" (recall: 0.72) and "Severe Obesity" (precision: 0.88) but struggled with predicting "Overweight" (precision: 0.42, recall: 0.45) (see **Figure I** in the Appendix for the full classification report). Overall, the **Food Habits Model** outperformed the **Lifestyle Habits Model**, with higher accuracy and improved precision and recall across most classes.

To further optimize the Random Forest model, we performed hyperparameter tuning on the combined dataset. The key hyperparameters we tuned included **n_estimators**, which determines the number of trees in the forest, **max_depth**, which controls the maximum depth of each tree, and **min_samples_split**, which specifies the minimum number of samples required to split an internal node. We tested values of [50, 100, 200, 300] for **n_estimators**, [None, 10, 20, 30] for **max_depth**, and [2, 5, 10] for **min_samples_split**. Our approach involved looping through different combinations of these hyperparameters, training the Random Forest model for each combination, and evaluating the model's accuracy on the test set. The best-performing model had the following hyperparameters: **n_estimators** set to 100, **max_depth** set to None, and **min_samples_split** set to 2. This model achieved an accuracy of **84.16%**, balancing complexity and generalization. Increasing the number of trees improved stability, while limiting splits prevented overfitting. The unrestricted depth allowed the model to capture complex patterns effectively.

We also experimented with **Logistic Regression** as an alternative classification model. The results showed an accuracy of **86%** (see **Figure K**). While the model achieved high accuracy, it produced a significant number of false positives for class 0, making it less reliable for our specific use case. As a result, we decided not to proceed with Logistic Regression for further analysis.

Finally, we combined the **Food Habits** and **Lifestyle Factors** datasets and applied the tuned Random Forest model. This combined approach significantly improved the model's performance, achieving an accuracy of **84.16%**. The classification report (see **Figure J** in the Appendix) shows high precision, recall, and F1-scores across all classes. For example, "Severe Obesity" achieved a precision of **0.98** and a recall of **0.89**, while "Obesity" achieved a precision

of **0.90** and a recall of **0.89**. The model also performed well for "Overweight" (precision: 0.82, recall: 0.81) and "Healthy Weight" (precision: 0.61, recall: 0.72). These results indicate that both food habits and lifestyle factors contribute valuable information for predicting obesity levels, and their combination leads to a more robust and accurate model.

5. CONCLUSION

The goal of this project was to predict obesity levels using machine learning models based on lifestyle and dietary factors. We experimented with several models, including Linear Regression, Polynomial Regression, Decision Tree, and Random Forest. While Linear and Polynomial Regression performed poorly due to their inability to capture non-linear relationships, Decision Tree provided interpretable results but was outperformed by Random Forest in terms of accuracy and class-wise performance.

The **Random Forest** model, when applied to the combined dataset of **Food Habits** and **Lifestyle Factors**, achieved the highest accuracy of **84.16%**. This significant improvement in performance highlights the importance of considering both food habits and lifestyle factors in obesity prediction. The model's high precision, recall, and F1-scores across all classes demonstrate its robustness and suitability for real-world applications.

Key insights from the analysis include:

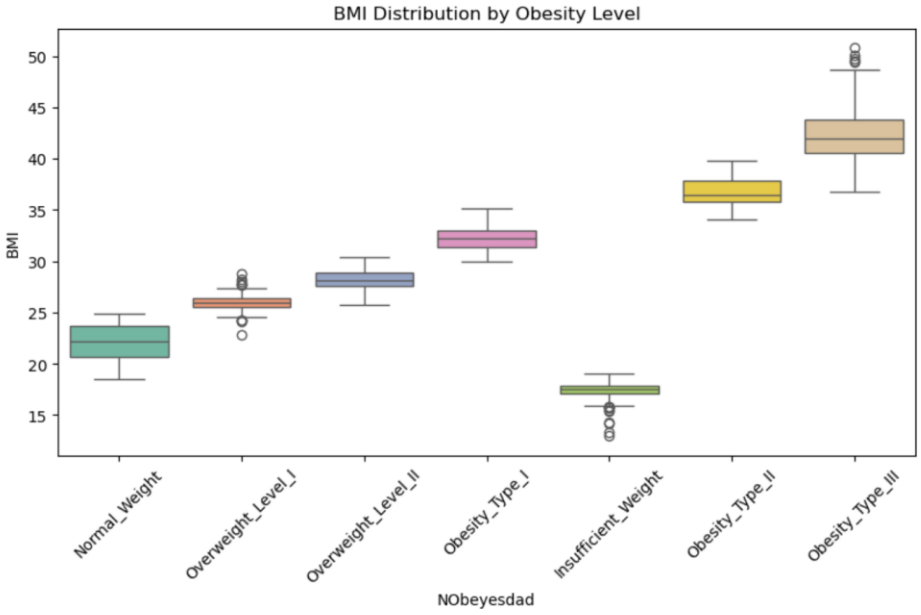
- **Food Habits** are a stronger predictor of obesity levels compared to **Lifestyle Factors**, as evidenced by the higher accuracy of the Food Habits Model (69.98%) compared to the Lifestyle Habits Model (61.23%).
- Combining both datasets significantly improves model performance, indicating that both factors contribute valuable information for predicting obesity levels.
- The model performs exceptionally well in identifying severe obesity cases, with high precision and recall for the "Severe Obesity" class.

This combined model can be used by healthcare professionals, nutritionists, and individuals to assess obesity risk and take preventive measures. Future work could involve collecting more

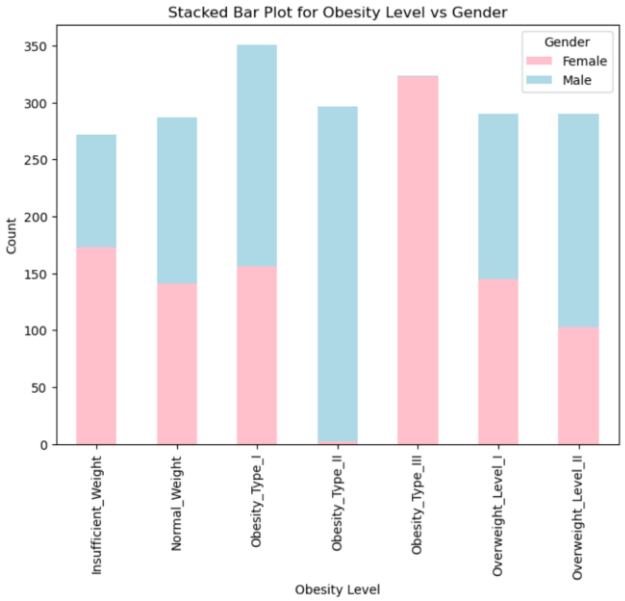
data, experimenting with additional algorithms, and exploring feature engineering techniques to further improve model performance.

6. APPENDIX

A) BMI vs Obesity levels

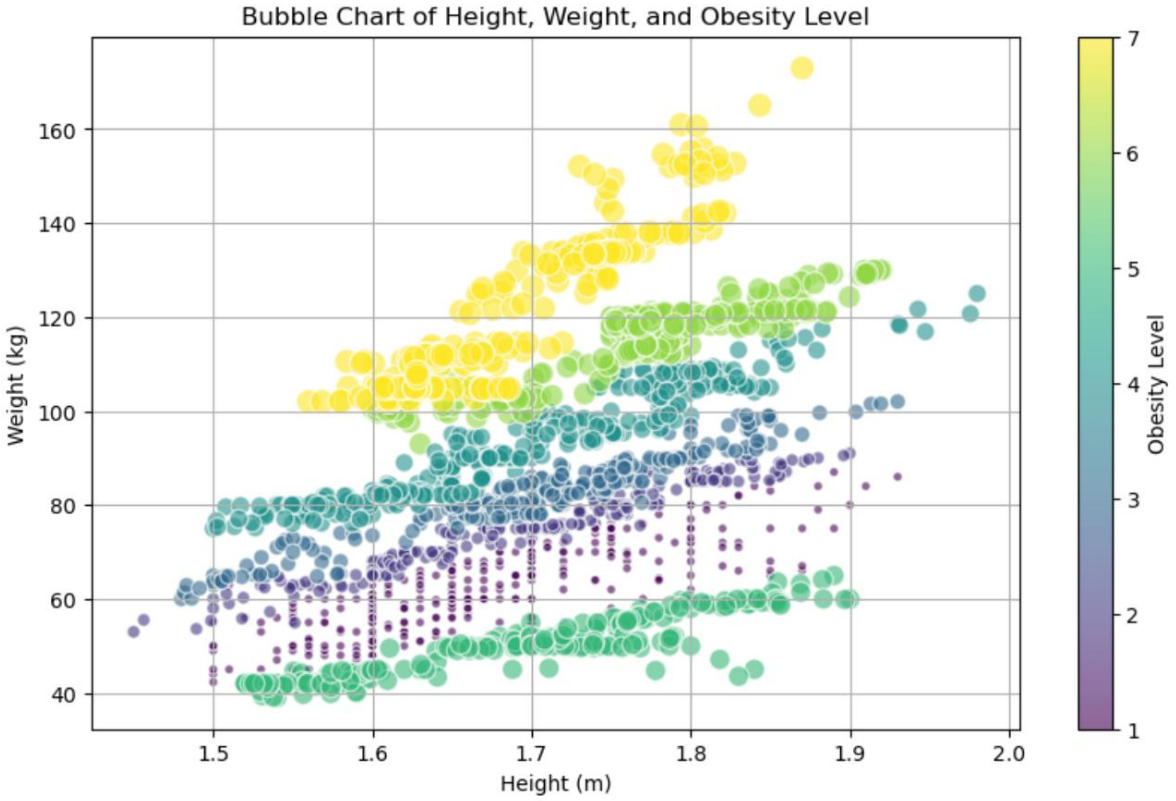


B) Obesity level vs Gender



C) Family History with Overweight vs Obesity Level

D) Correlation between Height, Weight and Obesity Level



E) Linear Regression Result

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.69e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
Linear Regression with Numerical and Categorical Features using Statsmodels
Mean Squared Error: 3.2366911006593817
R2 Score: 0.12360187714903625
Root Mean Squared Error: 1.7990806265032653
Adjusted R2 Score: 0.10233007805071181
```

F) Polynomial Regression Result

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.34e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
Polynomial Regression (Degree 2) with Numerical and Categorical Features using Statsmodels
Mean Squared Error: 3.076178502811044
R2 Score: 0.16706383724141471
Root Mean Squared Error: 1.7539037894967455
Adjusted R2 Score: 0.06516207264860907
```

G) Random Forest model on Food Habits Result

```
Accuracy: 0.6122931442080378

Classification Report:
                     precision    recall  f1-score   support

     Healthy Weight       0.55      0.72      0.62        67
Insufficient Weight       0.84      0.57      0.68        54
            Obesity       0.64      0.64      0.64       143
         Overweight       0.42      0.45      0.43       101
     Severe Obesity       0.88      0.76      0.81        58

           accuracy                           0.61       423
          macro avg       0.67      0.63      0.64       423
       weighted avg       0.63      0.61      0.62       423
```

H) Decision Tree Classifier on Food Habits Result

```
Accuracy: 0.5933806146572104

Classification Report:
                     precision    recall  f1-score   support

      Healthy Weight       0.51      0.73      0.60        67
Insufficient Weight       0.68      0.52      0.59        54
             Obesity       0.63      0.60      0.61       143
         Overweight       0.44      0.44      0.44       101
      Severe Obesity       0.92      0.76      0.83        58

            accuracy                           0.59       423
           macro avg       0.63      0.61      0.61       423
        weighted avg       0.61      0.59      0.60       423
```

I) Random Forest model on Lifestyle Factors Result

```
Random Forest Accuracy: 0.6998

Classification Report:
                     precision    recall  f1-score   support

      Healthy Weight       0.51      0.60      0.55        60
Insufficient Weight       0.70      0.61      0.65        54
             Obesity       0.78      0.78      0.78       142
         Overweight       0.63      0.61      0.62       113
      Severe Obesity       0.87      0.87      0.87        54

            accuracy                           0.70       423
           macro avg       0.70      0.69      0.70       423
        weighted avg       0.70      0.70      0.70       423
```

J) Random Forest model on Combined features Result

```
Random Forest Accuracy: 0.8416

Classification Report:
                     precision    recall  f1-score   support

      Healthy Weight       0.61      0.72      0.66        60
Insufficient Weight       0.92      0.87      0.90        54
             Obesity       0.90      0.89      0.90       142
         Overweight       0.82      0.81      0.81       113
      Severe Obesity       0.98      0.89      0.93        54

            accuracy                           0.84       423
           macro avg       0.85      0.84      0.84       423
        weighted avg       0.85      0.84      0.84       423
```

K) Logistic Regression Result

```
Logistic Regression Accuracy: 0.8629

Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.62      0.71       114
           1       0.87      0.95      0.91       309

    accuracy                           0.86       423
   macro avg       0.85      0.79      0.81       423
weighted avg       0.86      0.86      0.86       423
```

REFERENCES

1. **Dataset Source**:
   Palechor, F. M., & de la Hoz Manotas, A. (2019). *Estimation of obesity levels based on eating habits and physical condition*. UCI Machine Learning Repository.
   Available at:
   https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

2. **Research Paper**:
   Palechor, F. M., & de la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition. *Data in Brief, 25*, 104344.
   Available at: https://www.sciencedirect.com/science/article/pii/S2352340919306985