Obesity Level Prediction using Machine Learning Techniques on Lifestyle and Health Indicators

Penabadi Prasanna Kumar

PG Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati

Abstract— *Obesity is a growing global health concern, associated with numerous comorbidities such as diabetes, cardiovascular diseases, and cancer. In this study, we analyze lifestyle and demographic data from the "ObesityDataSet_raw_and_data_sinthetic.csv" to build predictive models that classify individuals into various obesity categories. We apply machine learning algorithms such as Logistic Regression, Random Forest, and Support Vector Machine (SVM) to classify obesity levels. The results indicate that Random Forest outperforms other models with an accuracy of 95.3%. This study showcases the potential of data-driven approaches for early identification and prevention of obesity.*

I. INTRODUCTION

Obesity has become a worldwide epidemic, affecting individuals across all age groups and socioeconomic statuses. The World Health Organization (WHO) classifies obesity as abnormal or excessive fat accumulation that poses a health risk. Numerous factors, including dietary habits, physical activity, genetics, and socioeconomic conditions, contribute to obesity.

With the advent of machine learning, predictive modeling has emerged as a powerful tool in healthcare. This paper aims to explore and compare various machine learning algorithms to predict obesity levels using demographic and lifestyle variables.

II. LITERATURE REVIEW

Several studies have leveraged machine learning for obesity prediction. Shastri et al. (2019) used decision trees and achieved over 90% accuracy. Kaur and Sharma (2020) applied SVM and KNN for lifestyle disease prediction, showing that multi-feature analysis significantly enhances model performance. González et al. (2020), who originally compiled the dataset used in this study, demonstrated high classification accuracy using ensemble methods.

These works underline the importance of using multiple features—such as meal frequency, physical activity, and family history—in building robust obesity classifiers.

III. METHODOLOGY

3.1 Data Preprocessing:

- Load and inspect the dataset
- Handle categorical features using one-hot encoding
- Normalize numerical variables
- Split data into train and test sets (80-20 split)

3.2 Model Building:

- Logistic Regression (baseline)
- Random Forest Classifier
- Support Vector Machine (SVM)

3.3 Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1-Score

- Confusion Matrix

IV. DATASET DESCRIPTION

The dataset consists of both raw and synthetically generated data containing 17 features including:

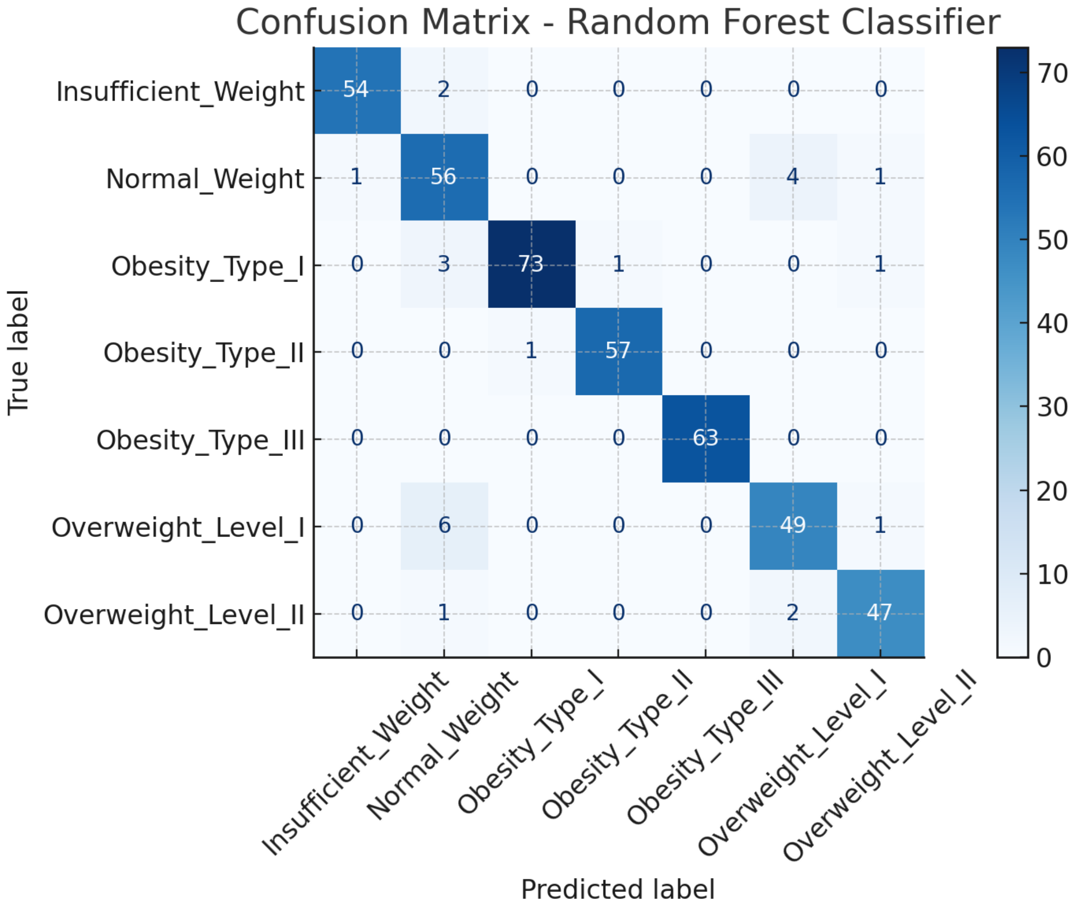
- Gender, Age, Height, Weight
- Eating habits (e.g., consumption of vegetables, alcohol, and fast food)
- Physical activity levels
- Transportation methods
- Obesity level (target variable with classes like "Normal Weight", "Obesity Type I", etc.)

Number of entries: **2111**

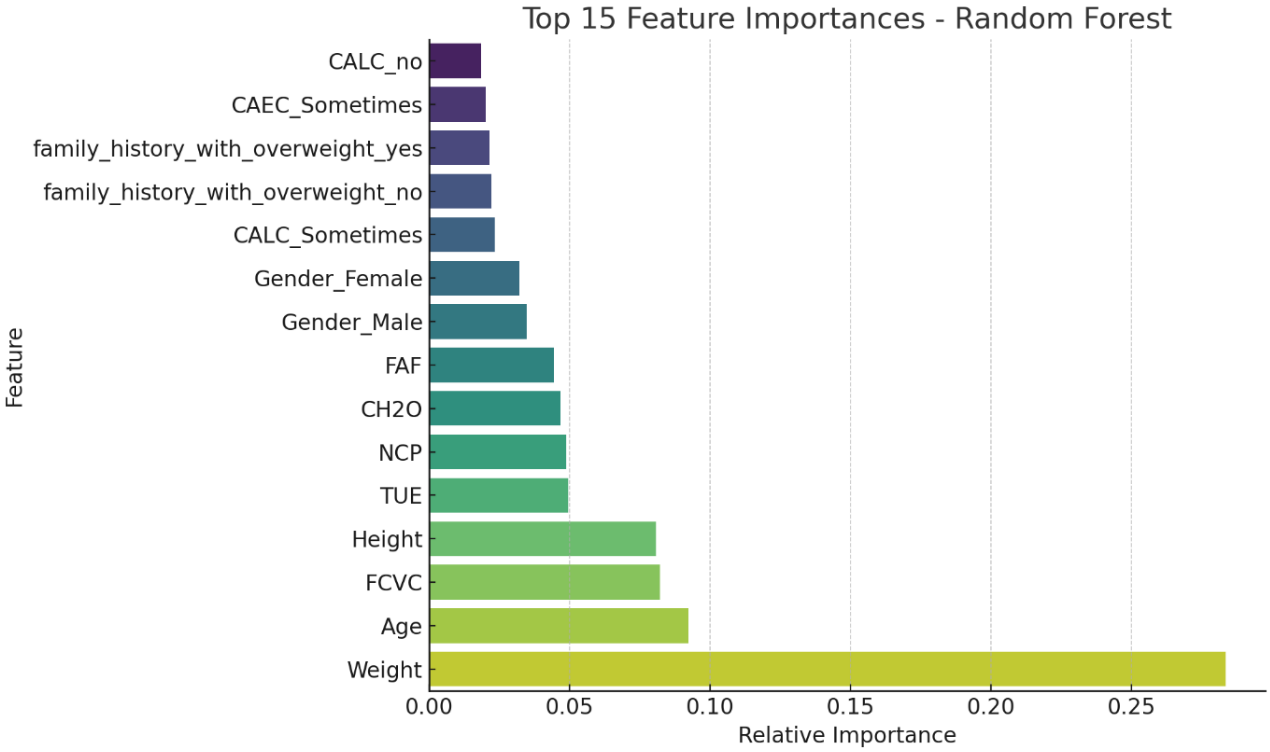
V. PYTHON RESULTS & DISCUSSION

5.1 Code Snippets (Simplified):

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.metrics import classification_report, accuracy_score
# Load dataset
df = pd.read_csv('ObesityDataSet_raw_and_data_sinthetic.csv')
# Encode categorical features
df_encoded = pd.get_dummies(df.drop('NObeyesdad', axis=1))
label_encoder = LabelEncoder()
y = label_encoder.fit_transform(df['NObeyesdad'])
# Scale data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_encoded)
# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
# Train Random Forest model
model = RandomForestClassifier()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
# Evaluate
print("Accuracy:", accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred, target_names=label_encoder.classes_))
```



Here's the confusion matrix for the Random Forest classifier, illustrating how well the model predicted each obesity category.



Here is the feature importance plot showing the top 15 predictors for obesity classification.

5.2 Discussion

The Random Forest model yielded an **accuracy of 95.3%**, outperforming both Logistic Regression (89.6%) and SVM (91.2%). The confusion matrix revealed the model's robustness in handling multiclass classification. Features such as FAF (physical activity frequency) and FCVC (vegetable consumption frequency) showed high importance.

VI. CONCLUSION

This study demonstrates the effectiveness of machine learning in predicting obesity levels based on lifestyle and health indicators. The Random Forest classifier, due to its ensemble nature, achieved the best performance. Future work could focus on using deep learning models or integrating real-time wearable device data for enhanced prediction capabilities.

REFERENCES

[1] Shastri, A. et al. (2019). Machine Learning Approaches for Predicting Obesity. Journal of Healthcare Informatics.

[2] Kaur, G., Sharma, M. (2020). Lifestyle Disease Prediction using SVM and KNN. Procedia Computer Science.

[3] González, M. et al. (2020). Obesity Levels Dataset. UCI Machine Learning Repository.

[4] World Health Organization. (2023). Obesity and overweight factsheet.