Impact of Lifestyle on Obesity

Predicting Obesity Risk Based on Lifestyle & Demographic Factors Using Machine Learning

CS 412- Introduction to Machine Learning, Professor Zhaochen Gu

Srijita Banerjee, Blossom Egbuonu, Mohammad Nusairat, Muhammad Ali Khurram

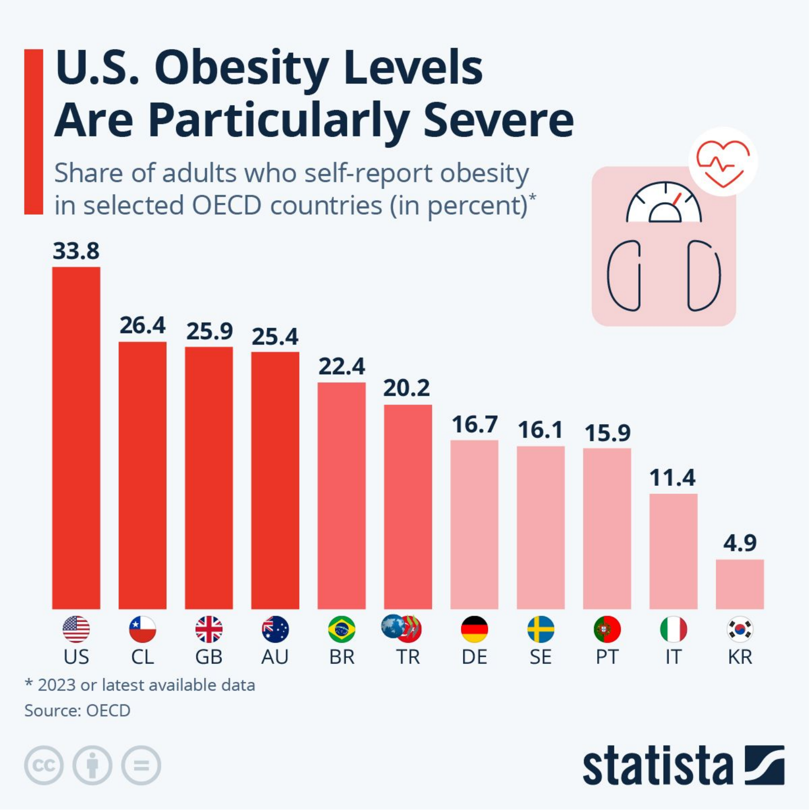
01

Introduction

Abstract

A Brief Introduction of Our Project

Obesity poses **serious health risks** globally; this project develops a machine learning model using demographic and lifestyle data to **predict obesity risk** and provide actionable insights for **prevention**.



Project Introduction

The problem we aim to solve and its relevance to machine learning

This project develops a predictive model to assess obesity risk based on **demographic** and **lifestyle** factors, such as **diet** and **physical activity**. Using **SVC** and **Naive Bayes**, our approach highlights key predictors and provides actionable recommendations to **reduce obesity risk**. This model offers practical insights to support personal and public health initiatives.

Index Terms: Obesity, Prediction, Lifestyle, Machine Learning, SVC, Naive Bayes, Health, Demographic, Preventive Health, Data-Driven Health Insights, Public Health Informatics, Risk

02

Dataset

Dataset Description and Relevant Pre-processing Conducted

Describing the dataset used, including its source and key characteristics, as well as data preprocessing steps

Dataset Source & Key Characteristics

- Source: UCI Machine Learning Repo
- Dataset has features: Gender, Age, Height, Weight, family_history_with_overweight, FAVC, FCVC, NCP, CAEC, SMOKE, CH2O, SCC, FAF, TUE, CALC, MTRANS, NObeyesdad, and BMI.
- Dimensionality: 2111 rows & 18 columns

Data Preprocessing Steps

- Encoded categorical variables
- Converted height from meters to inches
- Converted weight from kg to lbs
- Rounded numerical values
- Added `BMI` column
- Target column: NObeyesdad

03

Methodology

Methodology

Our chosen approaches to predictive modeling of the dataset

Naive Bayes

Effective for categorical data and is generally computationally efficient.

Encoded categorical variables and scaled numerical features.

SVC

Handles high-dimensional data well, with strong margins for classification.

Normalized numerical data and one-hot encoded categorical features.

Ensemble

Combines Random Forest, SVC, & Gradient Boosting for improved accuracy.

Used a ColumnTransformer for consistent feature preparation.

04

Results

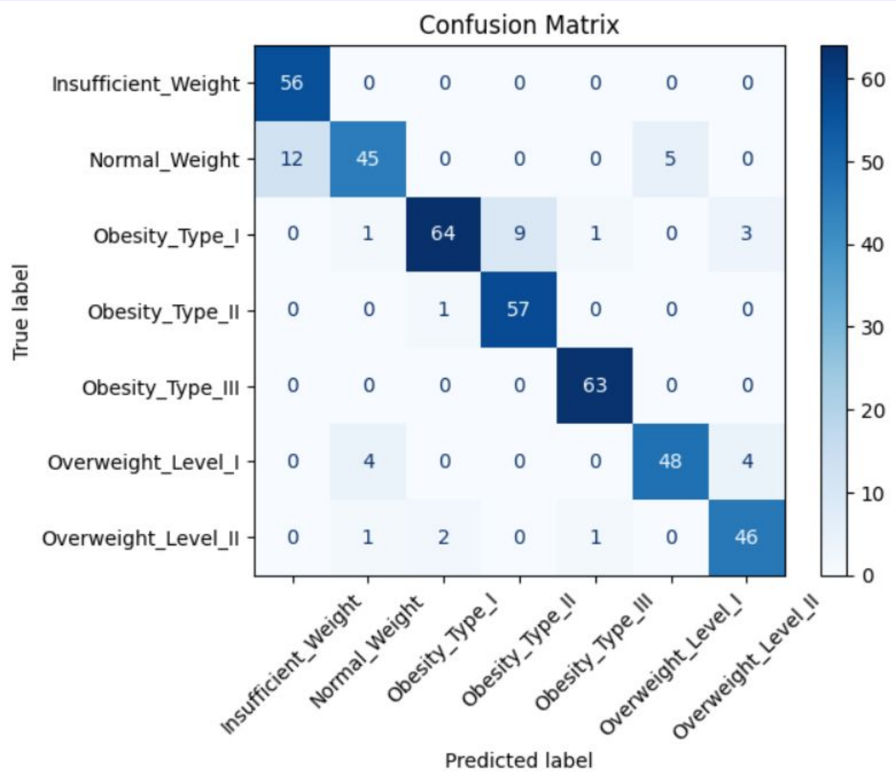
Results

Key metrics obtained from model implementations

Naive Bayes

Accuracy: 0.8959810874704491

	precision	recall	f1-score	support
Insufficient_Weight	0.82	1.00	0.90	56
Normal_Weight	0.88	0.73	0.80	62
Obesity_Type_I	0.96	0.82	0.88	78
Obesity_Type_II	0.86	0.98	0.92	58
Obesity_Type_III	0.97	1.00	0.98	63
Overweight_Level_I	0.91	0.86	0.88	56
Overweight_Level_II	0.87	0.92	0.89	50
accuracy			0.90	423
macro avg	0.90	0.90	0.89	423
weighted avg	0.90	0.90	0.89	423



Results

Key metrics obtained from model implementations

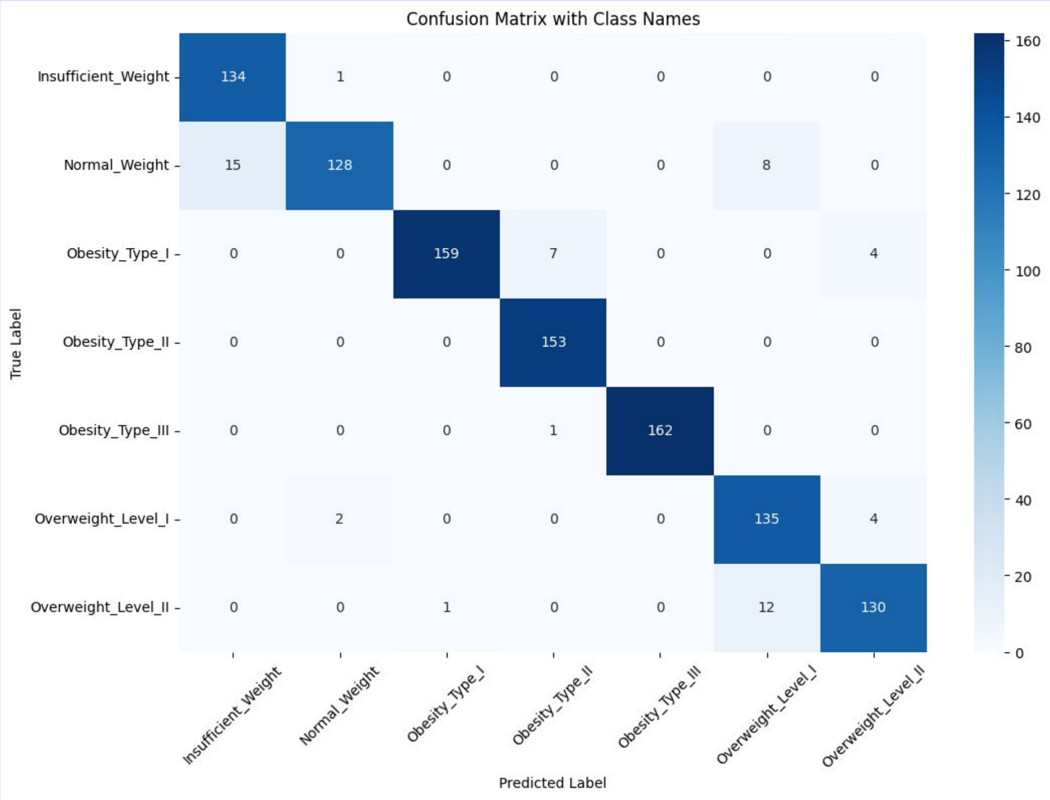
Support Vector Classifier (SVC)

SVC

```
SVC(kernel='linear', random_state=0)
```

```
Accuracy: 0.9479166666666666
```

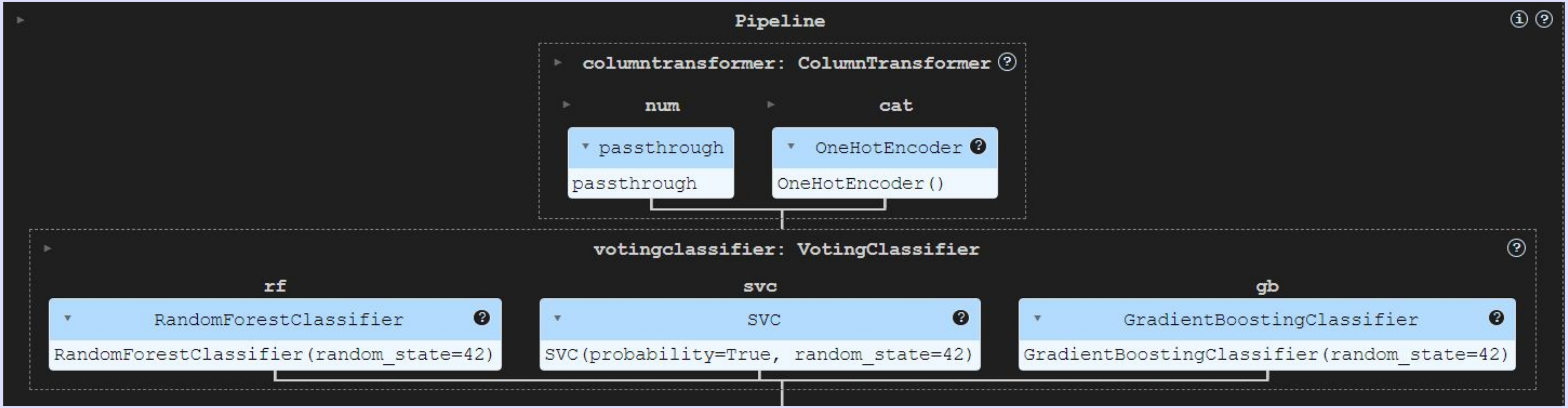
	precision	recall	f1-score	support
0	0.90	0.99	0.94	135
1	0.98	0.85	0.91	151
2	0.99	0.94	0.96	170
3	0.95	1.00	0.97	153
4	1.00	0.99	1.00	163
5	0.87	0.96	0.91	141
6	0.94	0.91	0.93	143
accuracy			0.95	1056
macro avg	0.95	0.95	0.95	1056
weighted avg	0.95	0.95	0.95	1056



Results

Key metrics obtained from model implementations

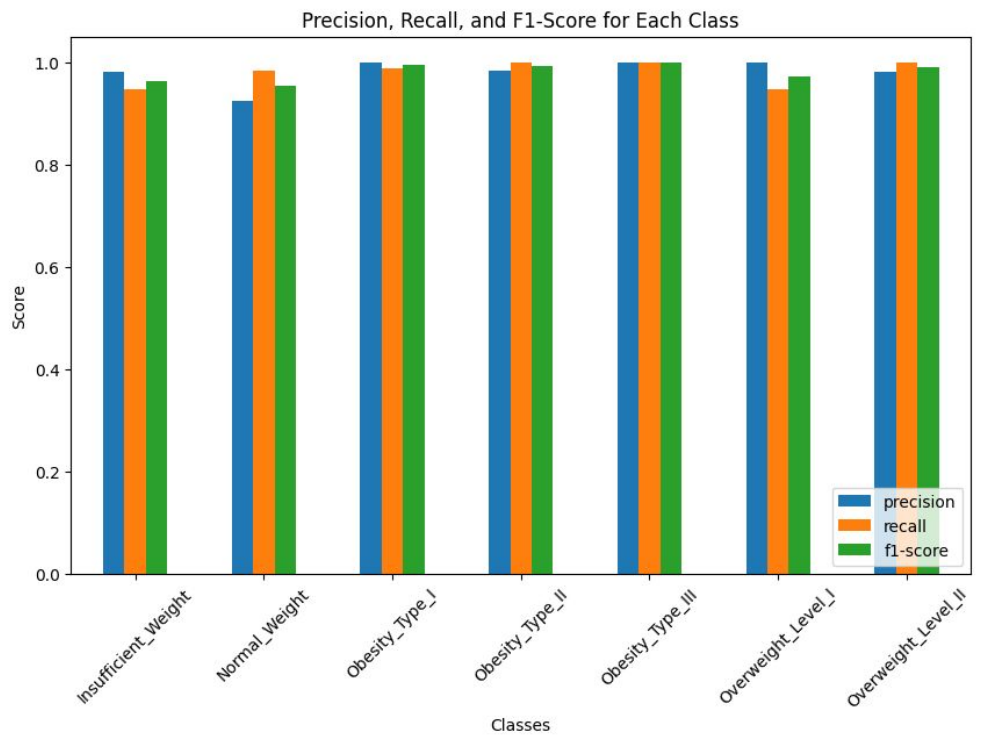
Ensemble with Support Vector, Random Forest and Gradient Boosting Classifier



Results

Key metrics obtained from model implementations

Ensemble with Support Vector, Random Forest and Gradient Boosting Classifier



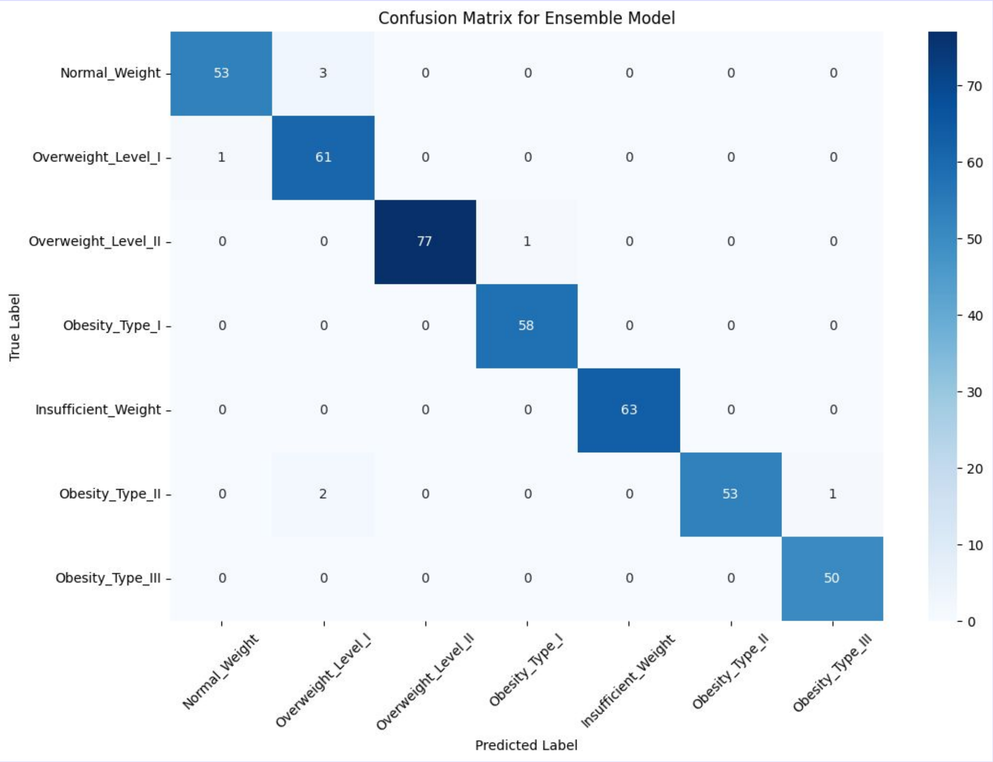
Results

Key metrics obtained from model implementations

Ensemble with Support Vector, Random Forest and Gradient Boosting Classifier

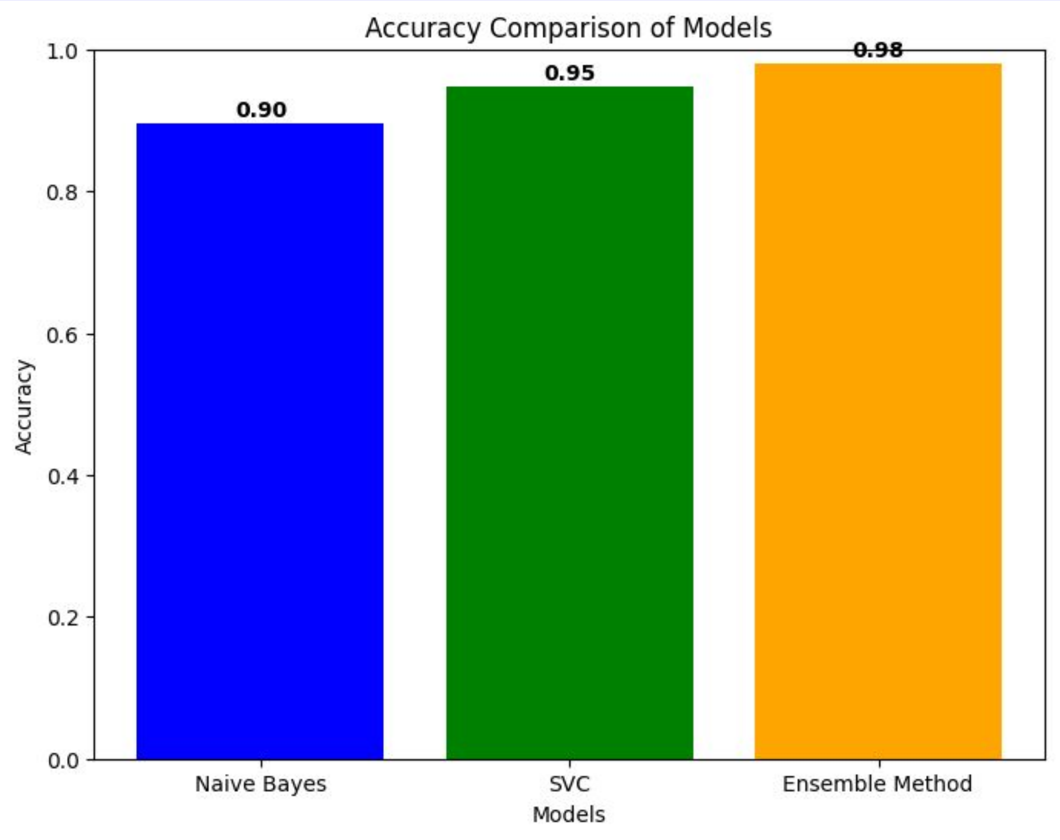
```
Overall Accuracy: 0.98
```

	precision	recall	f1-score	support
Insufficient_Weight	0.98	0.95	0.96	56
Normal_Weight	0.92	0.98	0.95	62
Obesity_Type_I	1.00	0.99	0.99	78
Obesity_Type_II	0.98	1.00	0.99	58
Obesity_Type_III	1.00	1.00	1.00	63
Overweight_Level_I	1.00	0.95	0.97	56
Overweight_Level_II	0.98	1.00	0.99	50
accuracy			0.98	423
macro avg	0.98	0.98	0.98	423
weighted avg	0.98	0.98	0.98	423



Results

Key metrics obtained from model implementations



05

Challenges and Adaptations

Challenges and Adaptations

Problems faced and methods used to tackle them

Data Imbalance:

- Challenge: Some obesity classes were underrepresented (ie. had fewer data points compared to other classes, which could lead to **model bias** and **poor generalization**).
- Adaptation: Used **soft voting** (handles imbalanced datasets better) in the ensemble model and evaluated performance using **F1-scores** (balances precision and recall for each class) to handle imbalance effectively.

Preprocessing Complexity:

- Challenge: Categorical and numerical features required different preprocessing steps.
- Adaptation: Implemented a **ColumnTransformer** for **streamlined feature preparation**.

06

Conclusion

Conclusion

Evaluation and review of the results obtained

Findings and Implications:

- The **ensemble model** achieved the **highest accuracy (98%)**, outperforming individual models like Naive Bayes and SVC.
- This project highlighted **key predictors of obesity**, such as **diet** and **activity patterns**, which is useful as actionable insights to be used by healthcare workers for **obesity risk reduction**.
- These results can support **public health efforts** by allowing personalized obesity risk assessments and hopefully curb the severe rise in obesity globally

Future Work and Improvements:

- Enhancements:
  - Incorporate additional datasets to **increase feature diversity**
  - experiment with **deep learning models** for more **complex** feature interactions.
- Could develop a user-friendly app for real-time obesity risk predictions and personalized health recommendations
- Could integrate the model into public health systems.

Questions?

Comments?

THANK YOU

Srijita Banerjee, Blossom Egbuonu, Mohammad Nusairat, Muhammad Ali Khurram