PID: 389

A Presentation on

An Interpretable Machine Learning Approach for Identification of the Risk Factors of Early-Stage Overweight and Obesity

3rd International Conference on Advancement in Electrical and Electronic Engineering, 25-27 April, 2024

PID: 389

Authors Information

Priyanka Roy Amrita Das Tipu

Department of Computer Science and Engineering

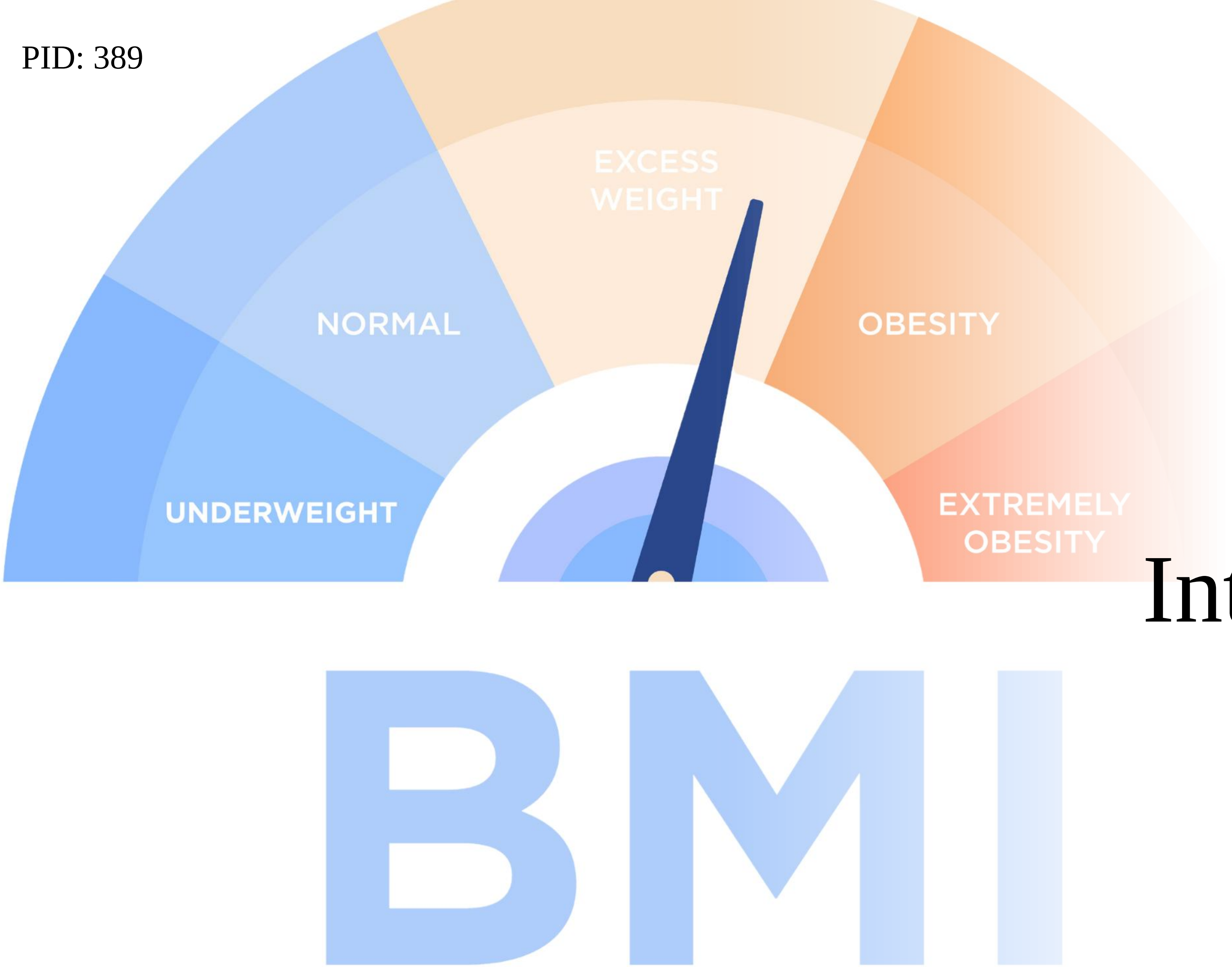
Presented By: Amrita Das Tipu

Hajee Mohammad Danesh Science and Technology University, Dinajpur-5200

Outline

- Introduction
- Motivation and Objectives
- Contributions
- Related Works
- Proposed Pipeline
- Result Analysis
- Conclusion and Future Work

PID: 389



Introduction

Introduction

- Overweight and obesity are **medical conditions** characterized by abnormal or excessive fat accumulation posing primary and secondary health risks (WHO)

- **Global Buzzword:**
  - Linked to several **NCDs and comorbidities**
  - Increase the risk of developing various **chronic diseases**, including cardiovascular diseases, diabetes, musculoskeletal disorders, and certain cancers.

- **Body Mass Index (BMI)** is commonly used indicator for classifying overweight and obesity.

$$\text{BMI} = \frac{weight}{height^2}$$

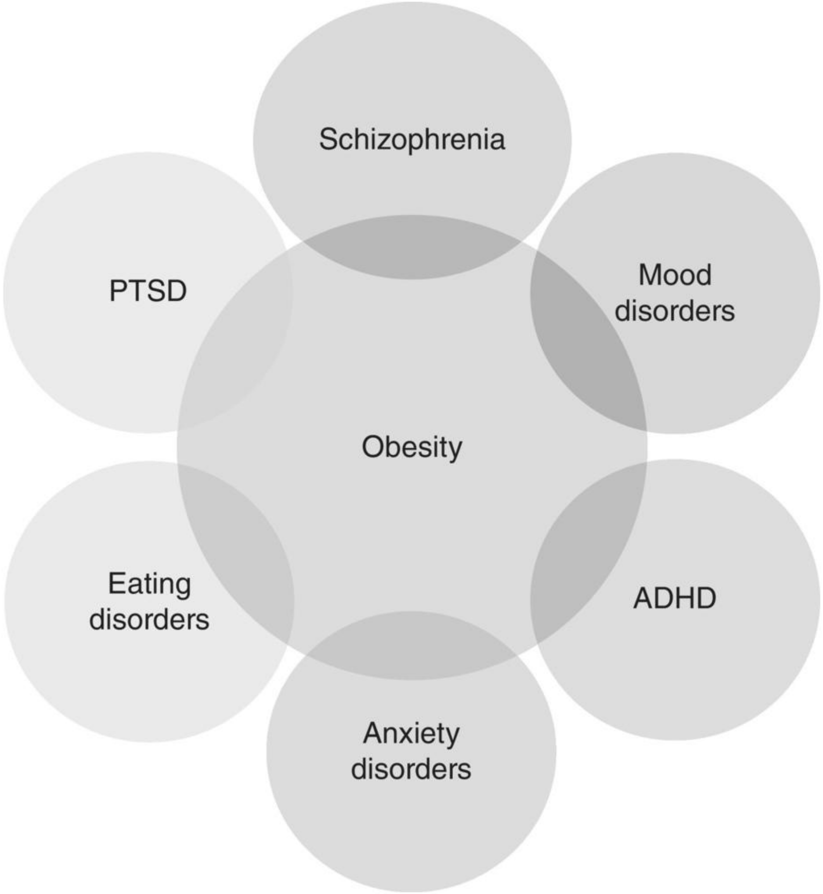


Fig. 1: Comorbidities

Motivation

- **Alarming Facts** from WHO**:**
    - ➢ Obesity rates tripled globally since 1975
    - ➢ 39 million children under 5 years of age and nearly **40% of adults** were overweight or obese in 2020
    - ➢ Prevalence of overweight and obesity will increase to **around 50% by 2030**

➢ In Bangladesh, over **2% children, 32% females, and 18% males** were obese in 2020 [1]

➢ Necessitates the **early-stage identification** of obesity and its **risk factors** as it allows for timely interventions, lifestyle modifications, and preventive measures

➢ Machine Learning (ML) has the potential to transform this by ensuring the evidence-based decision-making

Objectives

Detecting	• Identifying early-stage overweight and obesity will pave the way to enhance quality of life
Analyzing	• Applying various data preprocessing techniques to analyze and improve the performance of the black-box ML methods
Reducing Complexity	• Utilizing feature selection algorithms to identify the most significant feature contributing to the model outcome
Balancing	• Balancing the imbalanced data and applying different XAI tools to add model explainability

Contributions

Generalized pipeline for early-stage overweight and obesity identification utilizing the power of ensemble ML models

Identified risk factors and variables causing overweight and obesity

Achieved 99.58% accuracy and 1.00 AUC score

Transforming opaque ML algorithms into transparent glass-box models by adding the flavor of model explainability

Related Works

Author	Dataset	Feature	Best Performing Algorithm	Accuracy (%)
Rodríguez et al. [2]	Private	14	Random Forest	77.69
Taghiyev et al. [3]	Custom	26	Hybrid (DT and LR)	91.42
Solomon et al. [4]	UCI [5]	16	Hybrid (GB, XGB and MLP)	97.16
De-La-Hoz-Correa et al. [6]	UCI	16	Decision tree (J48)	97.40 (Precision score)
Kaur et al. [7]	UCI	13	Gradient Boosting	98.11

Table 1: Recent Existing Works

Dataset Description

- *"Estimation of Obesity Levels Based On Eating Habits and Physical Condition"* from the UCI Machine Learning Repository [5]
- **2111 records** with no missing values
- Consists of **16 attributes** and target class ***NObesity***
- For binary classification
  - **0 (normal)** for BMI < 25
  - **1 (risk)** for BMI ≥ 25
- Fig. 2 represents the imbalanced nature of the dataset

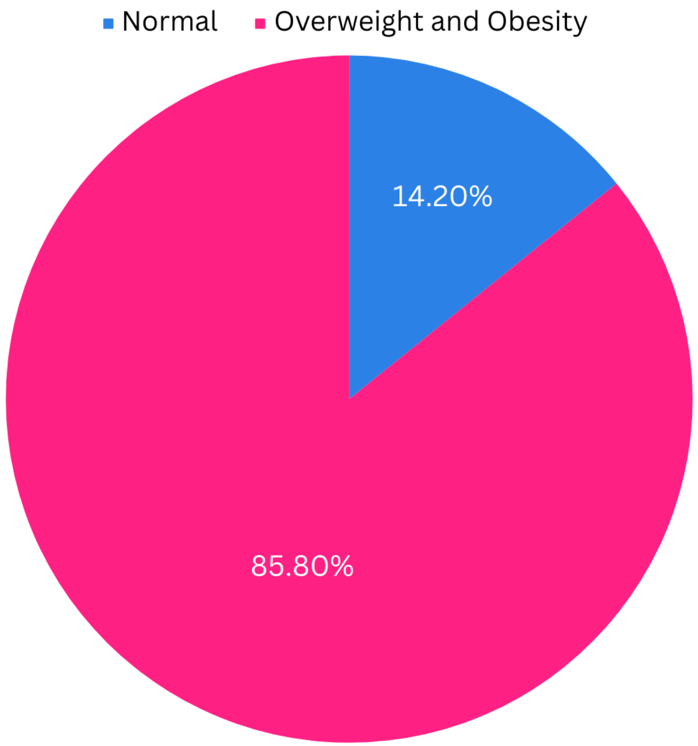


Fig. 2: Imbalanced Nature of the Dataset Used

PID: 389

Proposed Pipeline

Proposed Pipeline

- Data Collection
- Data Preprocessing
- Feature Selection
- Handling Imbalanced Data
- Training & Testing
- Performance Evaluation
- Result & Discussion
- Model Explainability

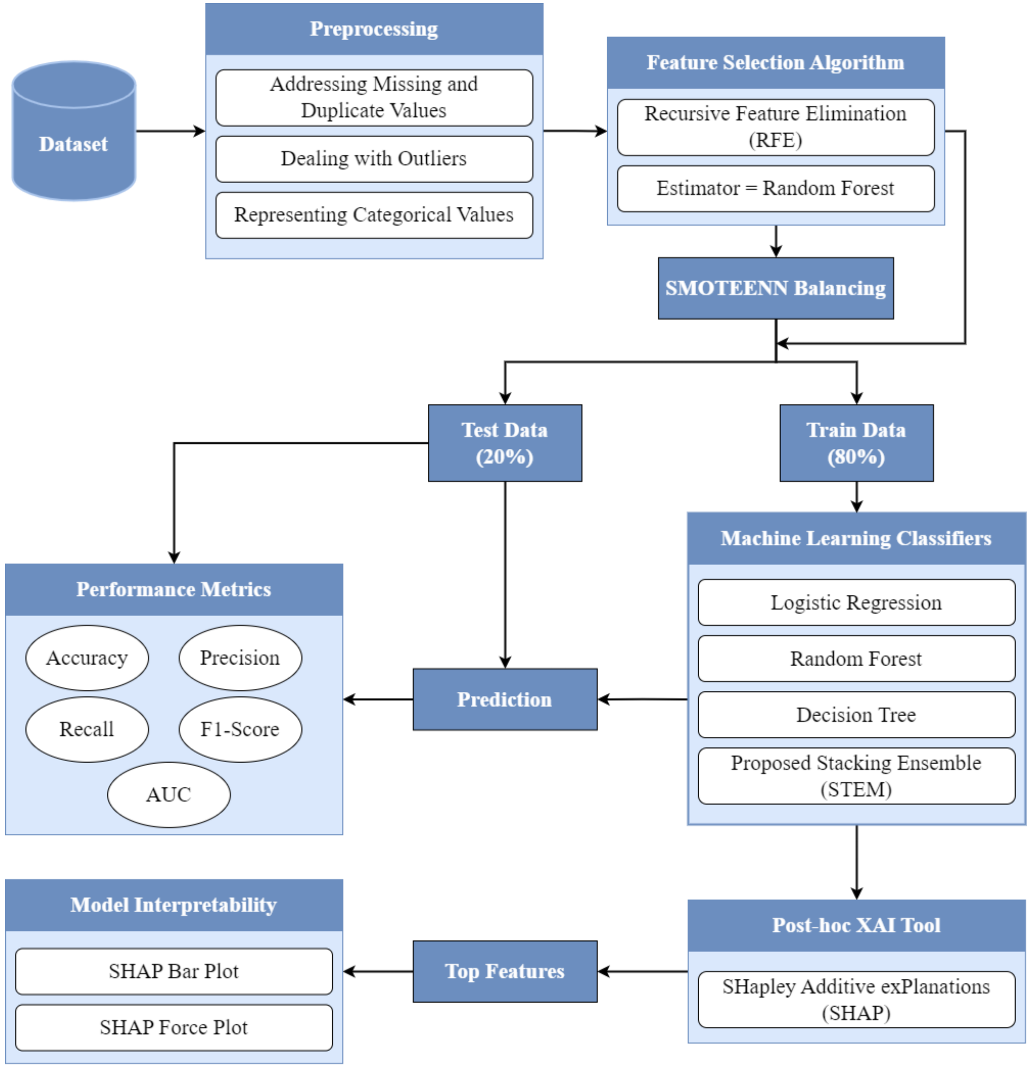


Fig. 3: Proposed Pipeline

ML Algorithms

- Linear Model
  - **Logistic Regression (LR)**: Comparatively simple and most effective for binary classification
- Tree Based Model
  - **Decision Tree (DT)**: Adaptable and intuitive data interpretation segmenting the feature space into different areas
  - **Random Forest (RF)**: Combines multiple DTs to improve generalization performance and is able to handle large datasets and reduce overfitting

Proposed Ensemble Classifier

- Stacking Ensemble Model (STEM)
- **Weak-learners:** LR, KNN, DT are chosen because they can handle high-dimensional data and learn intrinsic patterns
- Predictions of weak-learners are feed into the **meta-learner RF**
- Output of meta-learner is the final prediction

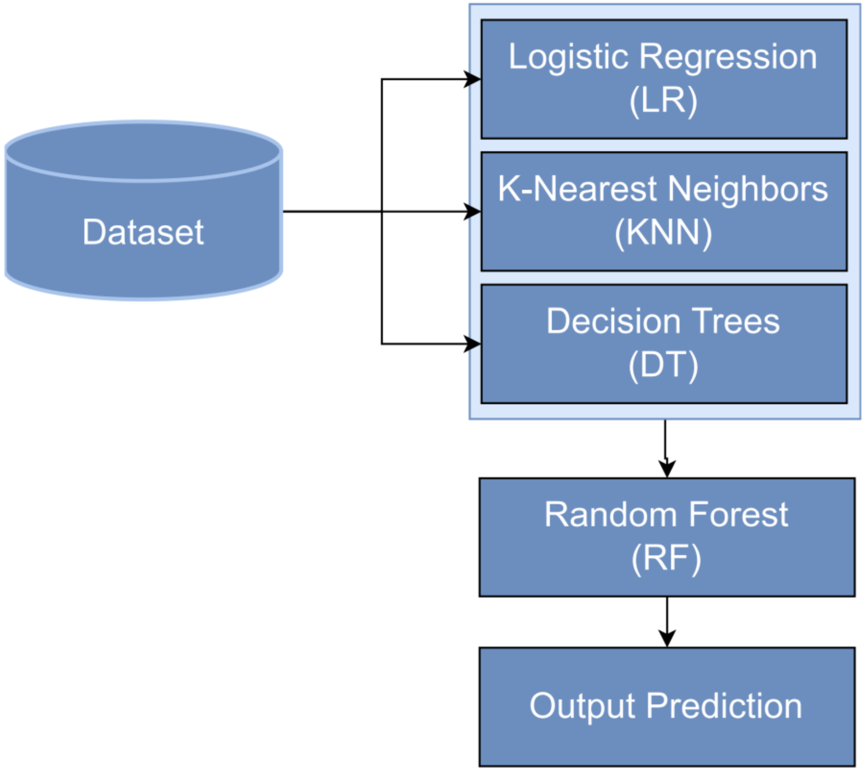


Fig. 4: Proposed STEM model

PID: 389

Result Analysis

Algorithm	Parameters
LR	C = 1.0, max_iter = 100, random_state = 42
RF	n_estimators = 100, bootstrap = True, max_depth = None
DT	Criterion = ’gini’, splitter = ’best’, random_state = 42
STEM	**Weak-learner:** LR, KNN (n_neighbors = 7, p = 2, leaf_size = 30, metric = ‘minkowski’), DT **Meta-learner:** RF

Table 2: Hyperparameter Settings for the Algorithms

Algorithm	Class	Precision	Recall	F1-Score	Avg. Acc. (%)
LR	0 1	0.39 0.85	0.10 0.97	0.16 0.91	83.22
RF	0 1	**0.90** **0.97**	0.82 **0.98**	0.86 0.97	95.74
DT	0 1	0.84 **0.97**	0.85 0.97	0.84 0.97	95.04
STEM	0 1	0.88 **0.97**	**0.87** **0.98**	**0.87** **0.98**	**95.98**

Table 3: Performance on Imbalanced Dataset

On Imbalanced Dataset

- Tuned parameters for each classifier
- Although the accuracy surpassed 90% except LR, the scores from other metrices clearly denotes the fluctuating poor performance in terms of stability for both majority and minority classes
- **STEM** achieved the **highest accuracy** of 95.98% even at the **imbalanced dataset** however Table 3 indicates scope for further improvements

Feature Selection

- Fig. 5 represents the final **10 significant features** extracted by the Recursive Feature Elimination (RFE) algorithm
- Ranked in descending order depending on individual **importance score**
- Reduced the dataset dimensionality and, therefore, related model complexity

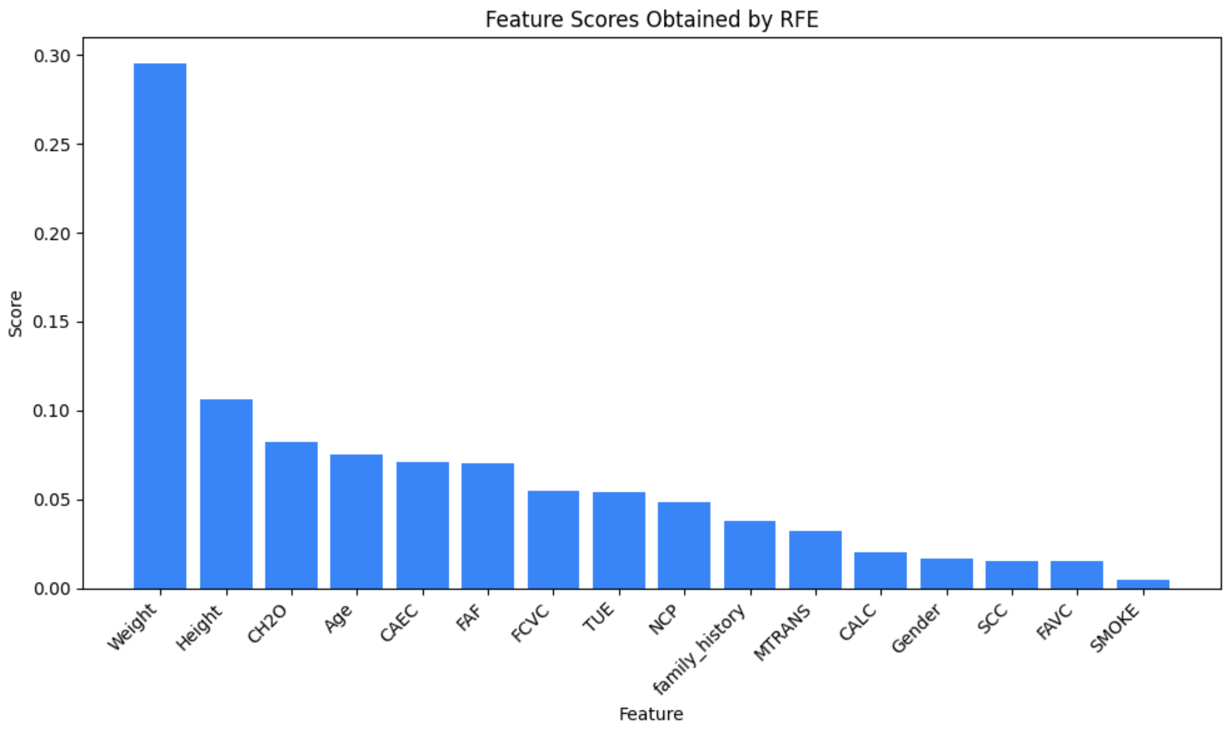


Fig. 5: Feature Ranking According to Importance Score

After Applying Proposed Pipeline

- The proposed **STEM surpassed** other existing ML models significantly by securing **99.58%** accuracy
- The Precision, Recall, and F1-scores demonstrates the **reliable performance** for both the majority and minority classes with **nearly a perfect score** for each metric
- Fig. 6 further illustrates the robustness and stability of the model **(AUC = 1.00)**

Algorithm	Class	Precision	Recall	F1-Score	Avg. Acc. (%)
LR	0 1	0.83 0.87	0.89 0.80	0.86 0.83	84.76
RF	0 1	0.97 **1.00**	**1.00** 0.96	0.98 0.98	98.18
DT	0 1	0.97 0.98	0.98 0.96	0.97 0.97	97.20
STEM	0 1	**1.00** 0.99	0.99 **1.00**	**1.00** **1.00**	**99.58**

Table 4: Performance after Applying Proposed Pipeline

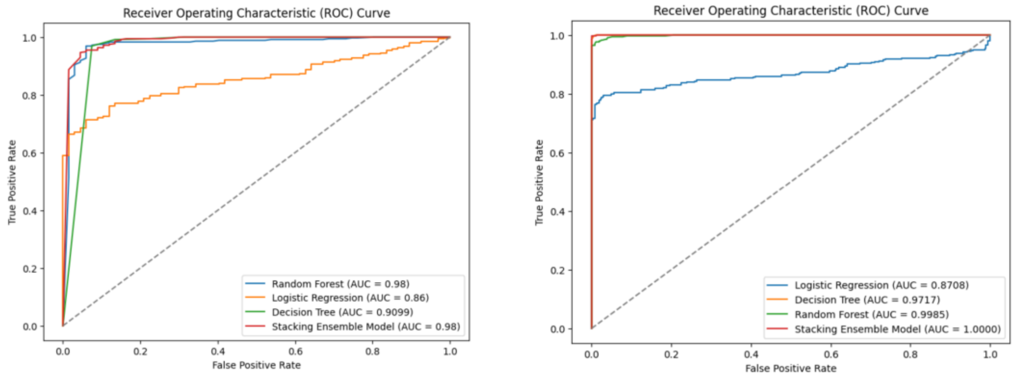
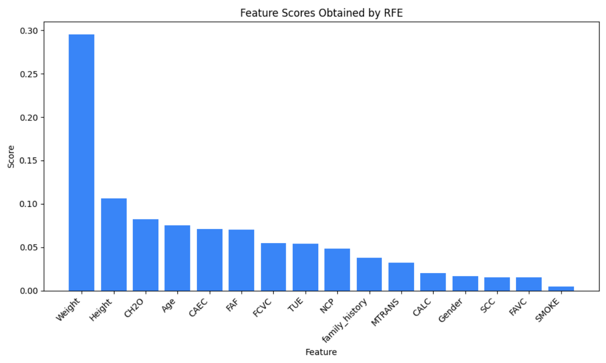


Fig. 6: Comparison of AUC-ROC Curves (before and after applying proposed pipeline)

Applying XAI

- Almost identical features indicating **Weight, Height, CAEC, NCP, CH2O, Age, and FAF** most significant contributors
- The SHAP force plot illustrates the individual feature contributions to the model outcome.
- Fig. 8 denotes that a young person of 25 years with 110.9 kgs and 166.5 cm height has a **100% chance to suffer from obesity**



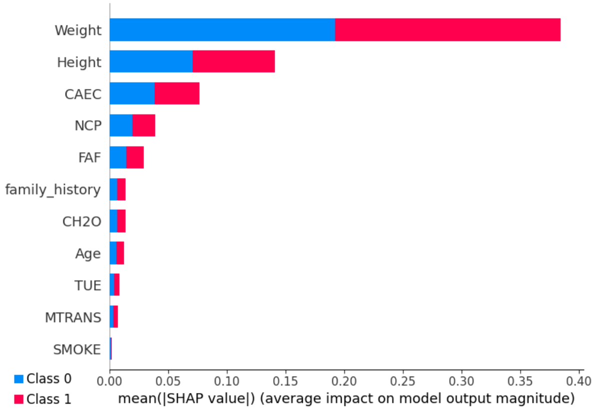


Fig. 7: Comparison between Extracted Features (RFE and SHAP)

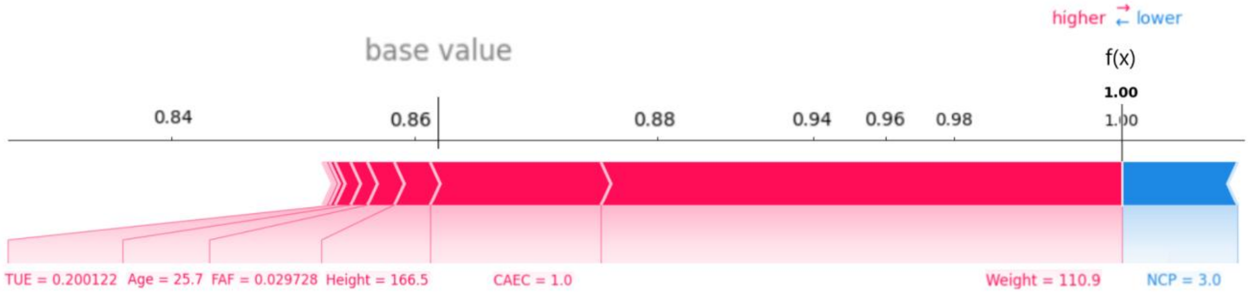
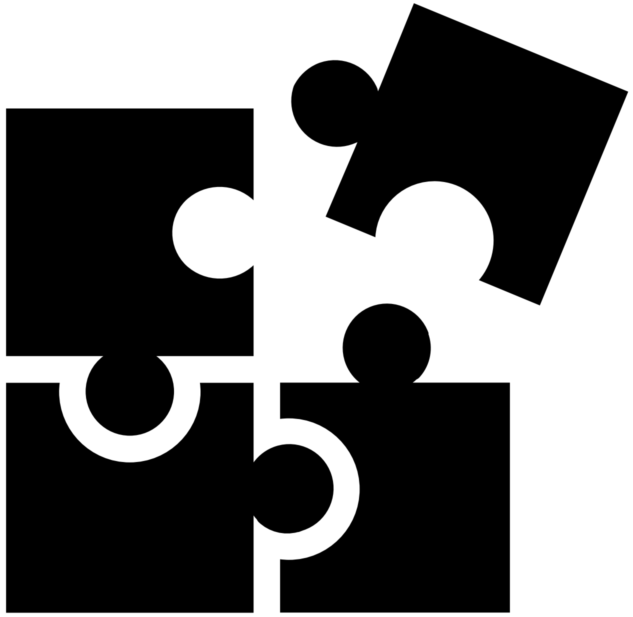


Fig. 8: SHAP Force Plot

Conclusion and Future Works

- Proposed an **effective generalized pipeline** including feature selection, dataset balancing and proposed STEM model
- Achieved the best **accuracy score of 99.58%** and AUC score of 1.0
- Identified the root causes behind overweight and obesity
- **Future directions?**
  - Federated machine learning
  - More Diversified Datasets



PID: 389

References

1. National Institute of Population Research and Training (NIPORT), “Bangladesh demographic and health survey 2017–18,” *The DHS Program*, 2020.
2. E. Rodríguez, E. Rodríguez, L. Nascimento, A. F. da Silva, and F. A. S. Marins, “Machine learning techniques to predict overweight or obesity,” in *IDDM*, 2021, pp. 190–204.
3. A. Taghiyev, A. A. Altun, and S. Caglar, “A hybrid approach based on machine learning to identify the causes of obesity,” *Journal of Control Engineering and Applied Informatics*, vol. 22, no. 2, pp. 56–66, 2020.
4. D. D. Solomon, S. Khan, S. Garg, G. Gupta, A. Almjally, B. I. Alabduallah, H. S. Alsagri, M. M. Ibrahim, and A. M. A. Abdallah, “Hybrid majority voting: Prediction and classification model for obesity,” *Diagnostics*, vol. 13, no. 15, 2023.
5. F. M. Palechor and A. de la Hoz Manotas, “Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico,” *Data in Brief*, vol. 25, p. 104344, 2019.
6. E. De-La-Hoz-Correa, F. E. Mendoza-Palechor, A. De-La-Hoz-Manotas, R. C. Morales-Ortega, and S. H. Beatriz Adriana, “Obesity level estimation software based on decision trees,” *Journal of Computer Science*, vol. 15, no. 1, pp. 67–77, Jan 2019.
7. R. Kaur, R. Kumar, and M. Gupta, “Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence,” *Endocrine*, vol. 78, no. 3, pp. 458–469, 2022.

PID: 389

Thank you

Any Question?