Obesity Prediction through Machine Learning: Evaluating Model Performance on Hybrid Datasets

Abstract

Obesity is a growing global health concern, often associated with lifestyle and behavioural factors. This study explores the application of machine learning models to predict obesity levels based on a dataset from UCI containing individual habits such as physical activity, diet, and technology use (UCI, 2019). A key contribution of this work is the removal of weight as a feature due to its multicollinearity with body mass index (BMI) and the application of a stratified data splitting to preserve class distribution. Five models: Random Forest, Support Vector Machine (SVM), Logistic Regression, XGBoost, and a Voting Classifier were implemented and compared. The Random Forest model achieved the highest accuracy (86.44%) following XGBoost (85.17%) and the Voting Classifier (82.76%). The findings demonstrate that ensemble models, particularly Random Forest, are effective for obesity classification and support the role of machine learning in early public health interventions.

Introduction

Obesity is classified as a chronic disease that can harm health due to excessive fat deposits (WHO, 2024). It is a growing global public health crisis due to its strong links to chronic illnesses such as diabetes, cardiovascular diseases, and hypertension. Beyond individual health, obesity places a substantial burden on healthcare systems and has economic implications at a societal level.

The rise in obesity is closely tied to the increase in non-communicable diseases (NCDs) which are long-term conditions shaped by genetic, physiological, environmental, and behavioural factors. As NCDs are now a leading cause of death and disability worldwide, addressing obesity is a critical step toward improving population health and reducing healthcare system burdens (Hildebrand and Pfeifer, 2025).

Moreover, according to the WHO (2024), global obesity rates have approximately doubled, while adolescent obesity has quadrupled since the 1990s (WHO, 2024).

These alarming trends underscore the need for early detection and preventive strategies. With obesity influenced by various of lifestyle, dietary, and socio-economic variables, developing effective risk prediction models is essential.

Objectives:

This study explores the potential of machine learning (ML) algorithms to support early identification of obesity risk. The main objectives are to:

1. Highlight gaps in the existing literature concerning the use of ML models for obesity prediction and propose areas for future research.
2. Synthesize existing findings to uncover broader insights and trends in obesity-related ML research.
3. Develop practical recommendations for applying ML models in predictive healthcare, with a focus on behavioural and lifestyle data.

Literature review

A review of related literature provided valuable insights into existing research and helped shape the research objective. Moreover, numerous studies conducted suggest ICT-based solutions help physicians and public health agents make the best decisions particularly in information and communication technologies (Chatterjee *et al.*, 2023). Collectively, these studies demonstrate a consistent trend: ensemble methods and data preprocessing play a critical role in improving prediction accuracy.

I. Musa, Basaky and Osaghae (2022) applied five machine learning algorithms: Gradient Boosting Classifier, Random Forest Classifier, Decision Trees Classifier, K-Nearest Neighbour, and Support Vector Machines (SVM) to predict obesity status using a clinical dataset. Among these, it was founded that ensemble models like Gradient Boosting Classifier achieved the highest accuracy (99.05%). This was followed by Support Vector Machine (97.16%), Random Forest Classifier and Decision Tree Classifier (96.93%), and K-nearest neighbours (95.74%).

II. Dutta, Mukherjee and Chakraborty (2024) examined the use of machine learning models to assess obesity risk, with a focus on the importance of data preprocessing and feature extraction. Their study tested Support Vector

Machines, Random Forests, and Decision Trees, with Random Forest achieving the highest accuracy (96%). This research aligns with other studies on the effectiveness of ensemble methods in improving prediction accuracy for obesity-related health risks. They also emphasized the significance of data visualization in understanding the relationships between various obesity-related factors, contributing to the development of more accurate and efficient predictive models.

III. Dirik (2023) compared Multilayer Perceptron, SVM, Logistic Regression and Random Forest models, with ensemble methods slightly outperforming traditional classifiers like Logistic Regression. These results reinforce the advantage of using tree-based and ensemble algorithms in handling complex, non-linear relationships common in obesity-related data.

IV. DeGregory *et al.*, (2018) used machine learning techniques including logistic and linear repressors, artificial neural networks, deep learning, and decision tree analysis, to predict and categorize the degree of obesity from a large dataset obtained from sensors, smartphone app data, large insurance database, publicly available national health data, and electronic medical health records. Their work supports the long-term viability of machine learning in categorizing obesity risk, especially when leveraging diverse and high-volume data sources.

V. Wong *et al.* (2022) applied XGBoost, Random Forest, SVM, and Logistic Regression to predict obesity and overweight status among Malaysian working adults. Using Malaysia's Healthiest Workplace by AIA Vitality 2019 survey data, the study demonstrated that both advanced and traditional models can be effective in real-world, community-level applications. Notably, XGBoost slightly outperformed others, with key predictors identified as weight satisfaction, ethnicity, age, and gender.

VI. Thamrin *et al. (*2021) applied machine learning to predict adult obesity using data from the 2018 Indonesian Basic Health Research survey. Models used included Logistic Regression, Classification and Regression Trees (CART), and Naïve Bayes Classification. Logistic Regression achieved the highest accuracy, with key predictors including geographic location, marital status, age, education, dietary habits, physical activity, and health conditions. The study also addressed class imbalance using the Synthetic Minority Oversampling Technique (SMOTE) to enhance prediction accuracy.

In summary, the literature indicates that ensemble models such as Random Forest, Gradient Boosting, and XGBoost consistently perform well across different populations and data types. Studies also underscore the importance of data quality, preprocessing, and feature engineering.

Methodology

Dataset description:

The dataset consisted of 2,111 records with 17 features with our target variable representing obesity levels. Approximately 77% of the data was synthetically generated using the Weka tool and SMOTE, while the remaining 23% was collected directly via a web platform.

Feature	**Descriptions**
Gender	-
Age	-
Height	-
Weight	-
family_history_with_overweight	Has a family member suffered or suffers from being overweight?
FAVC (Frequent consumption of high caloric food)	Do you eat high caloric food frequently?
FCVC (Frequency of consumption of vegetables)	Do you usually eat vegetables in your meals?
NCP (Number of main meals)	How many main meals do you have daily?

CAEC (Consumption of food between meals)	Do you eat any food between meals?
SMOKE	Do you smoke?
CH2O (Consumption of water daily)	How much water do you drink daily?
SCC (Calories consumption monitoring)	Do you monitor the calories you eat daily?
FAF (Physical activity Frequency)	How often do you have physical activity?
TUE (Time using technology devices)	How much time do you use technological devices such as cell phones, video games, television, computer and others?
CALC (Consumption of alcohol)	How often do you drink alcohol?
MTRANS (Transportation use)	Which transportation do you usually use?

Target	Descriptions
Nobeyesdad	Obesity level

Fig 1. Table of features linked to eating habits, physical conditions, and other variables, with a target value predicting obesity level (Palechor and Manotas, 2019).

Methodology steps:

The data was analysed through using Google Colab, a hosted Jupyter Notebook service.

1. Data preprocessing

Data preprocessing converts raw data into a machine learning-ready format. It addresses noisy, unclear information, enhances data quality, improves analysis, and boosts model performance. The quality of the data directly impacts the performance of the model. Preprocessing ensures high-quality data, improves productivity, and simplifies data mining by preparing the original data for analysis.

Steps:

a) **Data cleaning:** involved checking for missing values. No missing values were found. 24 duplicates were removed to improve data reliability and accuracy, as identical rows across 17 variables were deemed highly unlikely.

b) **Data integration:** involved unifying the fields with the dataset, ensuring consistency. Categorical features were encoded using one-hot encoding for nominal variables and label encoding for ordinal ones. One-hot encoding was used for nominal categories, where each category is represented as a separate binary column, while label encoding was applied to ordinal categories where the order of the categories holds significance. This made them compatible with machine learning algorithms, which typically require numerical input, to process categorical data. (Garg and Pundir, 2021).

c) **Feature Selection:** checked for significance of values within the dataset, removing highly correlated variables with each other and not the target using logical reasoning. The weight feature was removed due to its high correlation with obesity, which is derived from BMI (BMI = weight / height^2). (Palechor and Manotas, 2019). Removing this variable addressed multicollinearity, improving model generalizability.

d) **Statistical Significance:** ANOVA and linear regression tests were conducted to identify significant predictors. Variables highlighted below showed P-values < 0.005, confirming their relevance to obesity prediction.

ANOVA Results for Categorical Predictors	
	P-value
gender	0.1449
family_history_with_overweight	0
favc	0
smoke	0.9278
scc	0
mtrans_automobile	0.1885
mtrans_motorbike	0.0888
mtrans_public_transport	0.0812
mtrans_walking	0

Linear Regression Results for Continuous Predictors	
	P-value
age	0
height	0
fcvc	0
ncp	0.6144
ch2o	0
faf	0
tue	0
ord_caec	0
ord_calc	0

Fig 2. p-value of variables

e) **Outlier Analysis:** several outliers were noted, particularly in age and meal frequency. Ages ranged from 14 to 61, with many in their 20s, and meal frequency varied between 1 and 4. Both features were deemed relevant due to real-world context and the dataset's synthetic augmentation.

f) **Feature Scaling:** Min-Max scaling was applied to normalize all features to a [0, 1] range, ensuring equal contribution to the analysis. This method was chosen due to its relevance in related works and its suitability for scale-sensitive models like SVM and Logistic Regression. It helped standardize variable influence during training (Dutta, Mukherjee and Chakraborty, 2024).

$$v' = \frac{v - \min \mathrm{A}}{\max \mathrm{A} - \min \mathrm{A}} (\mathrm{new\_max\,A} - \mathrm{new\_min\,A}) - \mathrm{new\_min\,A}$$

2.Exploratory Data Analysis (EDA)

EDA Enables researchers to analyse obesity dataset relationships, using statistics and visuals to identify patterns, trends, and correlations among factors like age, weight, height, and lifestyle.

Visualizations such as histograms, box plots, and correlation heatmaps were created to clearly depict the distribution of variables and highlight any potential outliers or unusual patterns.

3.Model Selection

Based on the literature, objectives and dataset analysis, the following models were selected for this research to enable further analysis and comparisons, guided by the findings in the literature review.

- **Random Forest Classifier:** a versatile ML algorithm that performs well on both classification and regression tasks by constructing an ensemble of decision trees trained through the bagging method, which combines learning models to improve the model's overall performance (Donges, 2021).
- **Logistic Regression:** a supervised learning algorithm for classification tasks, class probabilities using the sigmoid function. The sigmoid function maps outputs to a range between 0 and 1. Its simplicity and effectiveness make it widely used across fields like healthcare and finance (Bisong, 2019).
- **Support Vector Machine:** an essential and adaptable ML algorithm that excels in regression, outlier detection, and linear and nonlinear classification. Support Vector Machine is commonly used for classification and regression due to its accuracy and low consumption requirements but can be sensitive to unclean data, outliers, and mislabelled observations (Debruyne, 2009).
- **XGBoost:** a highly efficient ML algorithm that implements a gradient boosting framework designed to optimize performance and speed. It builds an ensemble of decision trees in a sequential manner, where each tree attempts to correct the errors of the previous ones. It prevents overfitting with regularization, enables efficient handling of large datasets with parallel computation, and has built-in mechanisms to manage sparse data and missing values (Chen and Guestrin, 2016).

- **Voting Classifier:** an ensemble method that combines multiple models' predictions to boost accuracy, using majority voting for classification or averaging for regression. It reduces overfitting by leveraging diverse model strengths (Cornelio *et al.*, 2021). Soft voting is used in this study, which aggregates probability estimates to better capture model confidence, often improving accuracy and robustness. It was preferred over hard voting as it considers the confidence of each model's prediction, rather than relying solely on the majority class.

4.Model Optimization

Stratification, uncommon in related works, was applied when splitting data (70% training, 30% test) to maintain obesity category distribution, reduce biases, and ensure a fair model evaluation.

Additionally, Grid Search with 5-fold cross-validation was implemented to optimize hyperparameters, tailored for multiclass classifiers to handle multiple prediction classes effectively across models, addressing the need for multi-class classification of obesity. This balanced bias-variance trade-offs and improved generalization, despite being computationally intensive. It also ensured the model achieved the best performance as grid search exhaustively tested combinations of values within a predefined grid.

With only 2,087 rows remaining after removal of duplicates, grid search was crucial for optimizing model parameters. However, it is notable that grid search along with cross validation can be computationally expensive compared to other hyperparameter tuning techniques (Rimal, Sharma and Alsadoon, 2024).

5.Model Evaluation

Each model's performance was assessed using standard classification metrics.

$$Accuracy = \frac{no.\ of\ correct\ classifications}{total\ no.\ of\ classifications} \times 100$$

- **Accuracy:** measures the proportion of correctly classified instances out of the total instances, providing an overall performance metric (Miller *et al.*, 2024).

$$p = \frac{TP}{TP + FP}$$

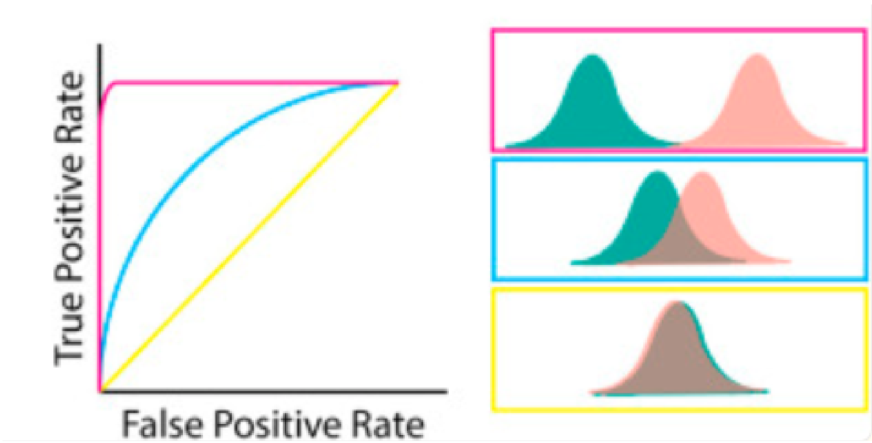
- **Precision:** evaluates the ratio of true positive predictions to the total predicted positives, reflecting the model's ability to avoid false positives.

$$r = \frac{TP}{TP + FN}$$

- **Recall:** assesses the ratio of true positive predictions to the total actual positives, indicating the model's sensitivity (Goutte and Gaussier, 2005).

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R}.$$

- **F1-score:** harmonic mean of precision and recall balancing the trade-off between the two (Hand, Christen and Kirielle, 2021).



- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** measures binary classifier performance by assessing its ability to differentiate classes across thresholds. It calculates the area under the curve of true positive rate (recall) versus false positive rate. The example above shows yellow (total crossover), blue (partial separation), and pink (total separation) distribution patterns (Miller et al., 2024).

Moreover, in the case of a confusion matrices for the obesity dataset, true positives and true negatives are key, as we want to predict instances of obesity rather than false predictions which can harm patient trust (El-Shahat *et al.*, 2024). Additionally, precision ensures most positive predictions are true obesity cases, while recall identifies most actual cases. The F1-score balances the trade-off between precision and recall.

(See Appendix 1 for Confusion matrices)

Discussions

	Precision	Recall	F1-score	Accuracy	AUC-ROC
SVM	74.50%	74.96%	74.37%	74.96%	93.69%
XGBoost	86.12%	85.17%	85.26%	85.17%	97.50%
Random Forest	**87.07%**	**86.44%**	**86.45%**	**86.44%**	**98.06%**
Logistic Regression	56.05%	57.10%	55.11%	57.10%	86.63%
Voting Classifier	83.00%	82.78%	82.63%	82.78%	96.85%

Fig 3. Classification metrics comparison table

The results in Fig 3 offer valuable insights into the effectiveness of machine learning algorithms for predicting obesity levels using lifestyle and behavioural data Among the models tested, ensemble methods consistently outperformed traditional classifiers, with Random Forest achieving the highest scores across most metrics. XGBoost followed closely, further supporting that ensemble-based algorithms tend to outperform traditional classifiers in complex classification tasks like obesity prediction.

Although the Voting Classifier, which combined multiple models, showed robust but not superior performance, highlighting that ensemble effectiveness depends on the individual classifiers and their complementary strengths (Alhamid, 2025). In this case, the ensemble may have lacked sufficient diversity beyond the capabilities already offered by Random Forest and XGBoost.

SVM demonstrated moderate performance, consistent with prior findings that highlight its sensitivity to data scaling and noise. While effective with smaller, well-pre-processed datasets, SVM may underperform on highly imbalanced data or features nonlinear relationships that aren't easily separable with kernel tricks (Singla and Shukla, 2020).

Logistic Regression, often considered a baseline model, achieved the lowest performance. This aligns with expectations, given the complexity of obesity as a multifactorial condition influenced by various behavioural, dietary, and physiological factors. Linear models like Logistic Regression are limited in capturing such interactions unless explicitly engineered through feature interactions, which may not always be feasible or interpretable (Levy and O'Malley, 2020).

Overall, the results highlight the strengths of ensemble methods and tree-based models in obesity detection, suggesting their potential for further development and application in healthcare settings. Moreover, comparisons of various machine learning models in related works shows that the random forest was mostly the best classifier.

In comparison to similar works such from Dutta, Mukherjee and Chakraborty (2024), where the Voting Classifier and Logistic Regression were highlighted for further exploration, our findings provide a nuanced contrast. In our case, both models underperformed, which may be attributed to our methodology choices, notably, the exclusion of the weight variable (due to multicollinearity with obesity) and the use of stratified data splitting to maintain class balance. While these steps improved model fairness and generalizability, they may have impacted certain models that performed better with simpler, less-refined datasets, as seen in other studies.

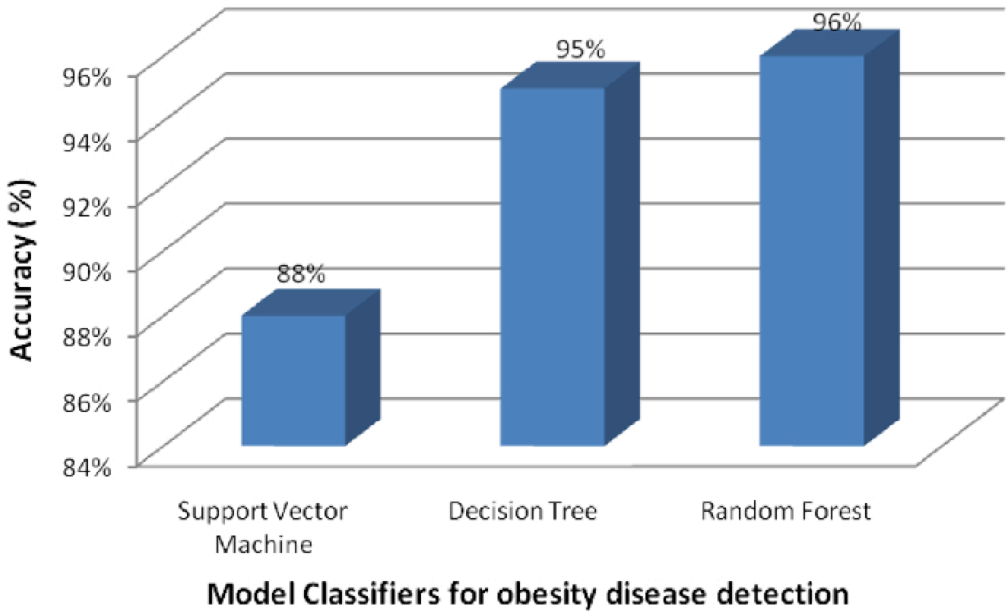


Fig 4. Comparative accuracy of different models from Dutta, Mukherjee and Chakraborty (2024).

Conclusion

This study explored the potential of machine learning algorithms in predicting obesity levels using behavioural and lifestyle data. Rather than focusing on clinical or medical variables, this research emphasized non-invasive assessment tools such as eating habits, physical activity, and technology usage. A key methodological contribution was the exclusion of the weight variable due to its overlap with BMI and the use of stratified data splitting to maintain class balance across training and test sets. These steps helped refine the dataset and improve model generalisation.

Five models were evaluated: Support Vector Machine, XGBoost, Random Forest, Logistic Regression, and a Voting Classifier. Among these, ensemble methods particularly Random Forest and XGBoost demonstrated superior performance across all metrics. This reinforces the value of tree-based approaches in handling complex, nonlinear patterns within health-related data.

Key strengths of this study were the use of a high-quality dataset, with no missing values, collected from individuals across Mexico, Peru, and Colombia. The data included a wide range of lifestyle and dietary habits, which significantly contributed to the performance of the models.

However, limitations exist. Notably, only 23% of the data was derived from real-world observations, while the remaining 77% was synthetically generated through SMOTE and Weka. While this allowed for class balancing and increased sample size, it also introduces concerns about generalizability and model reliability when applied to broader populations (Badawy, Ramadan and Hefny, 2024). Furthermore, the age distribution within the dataset ranging from 14 to 61 years and predominantly concentrated among individuals in their 20s does not fully represent the demographic diversity of the populations studied. Additionally, given that the dataset was added to the UCI Machine Learning Repository in 2019, it may not reflect more recent shifts in dietary behaviours and lifestyle trends, collection of more recent data with various distributions would be optimal (Dutta, Mukherjee and Chakraborty, 2024)

These limitations underscore the importance of future research focused on improving the generalisability and accuracy of obesity prediction models. Incorporating more recent and representative real-world data, particularly from underrepresented age groups and regions, will enhance the robustness of such models. Further research could also explore the integration of additional lifestyle variables such as mental health, motivation, or duration of exercise routines, which may influence obesity but are often overlooked. Finally, advancing model performance through more diverse ensemble techniques and robust hyperparameter tuning can further enhance prediction accuracy. Validating these models across diverse cultural and socioeconomic settings will be critical to ensuring their broader real-world applicability and impact.

Bibliography

1. Alhamid, M. (2025) *Ensemble Models: What Are They and When Should You Use Them?*, *Built In*. Available at: https://builtin.com/machine-learning/ensemble-model (Accessed: 12 April 2025).

2. Badawy, M., Ramadan, N. and Hefny, H.A. (2024) ‘Big data analytics in healthcare: data sources, tools, challenges, and opportunities’, *Journal of Electrical Systems and Information Technology*, 11(1), p. 63. Available at: https://doi.org/10.1186/s43067-024-00190-w.
3. Bisong, E. (2019) ‘Logistic Regression’, in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. Berkeley, CA: Apress, pp. 243–250. Available at: https://doi.org/10.1007/978-1-4842-4470-8_20.
4. Chatterjee, A. *et al.* (2023) ‘A systematic review and knowledge mapping on ICT-based remote and automatic COVID-19 patient monitoring and care’, *BMC Health Services Research*, 23(1), p. 1047. Available at: https://doi.org/10.1186/s12913-023-10047-z.
5. Chen, T. and Guestrin, C. (2016) ‘XGBoost: A Scalable Tree Boosting System’, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery (KDD ’16), pp. 785–794. Available at: https://doi.org/10.1145/2939672.2939785.
6. Cornelio, C. *et al.* (2021) ‘Voting with random classifiers (VORACE): theoretical and experimental analysis’, *Autonomous Agents and Multi-Agent Systems*, 35(2), p. 22. Available at: https://doi.org/10.1007/s10458-021-09504-y.
7. Farayola, M. (2025) *Data-Science-ML-Projects/Bank_Note_Analysis.ipynb at main · mmfara/Data-Science-ML-Projects*, *GitHub*. Available at: https://github.com/mmfara/Data-Science-ML-Projects/blob/main/Bank_Note_Analysis.ipynb (Accessed: 16 April 2025).
8. Debruyne, M. (2009) ‘AN OUTLIER MAP FOR SUPPORT VECTOR MACHINE CLASSIFICATION’, *The Annals of Applied Statistics*, 3(4), pp. 1566–1580.
9. DeGregory, K.W. *et al.* (2018) ‘A review of machine learning in obesity’, *Obesity Reviews*, 19(5), pp. 668–685. Available at: https://doi.org/10.1111/obr.12667.

10. Dirik, M. (2023) ‘Application of machine learning techniques for obesity prediction: a comparative study’, *Journal of Complexity in Health Sciences*, 6(2), pp. 16–34. Available at: https://doi.org/10.21595/chs.2023.23193.
11. Donges, N. (2021) *Random Forest: A Complete Guide for Machine Learning*, *Built In*. Available at: https://builtin.com/data-science/random-forest-algorithm (Accessed: 12 April 2025).
12. Dutta, R.R., Mukherjee, I. and Chakraborty, C. (2024) ‘Obesity disease risk prediction using machine learning’, *International Journal of Data Science and Analytics* [Preprint]. Available at: https://doi.org/10.1007/s41060-023-00491-9.
13. El-Shahat, D. *et al.* (2024) ‘Machine learning and deep learning models based grid search cross validation for short-term solar irradiance forecasting’, *Journal of Big Data*, 11(1), p. 134. Available at: https://doi.org/10.1186/s40537-024-00991-w.
14. Galloway, J. (2021) classification_of_obesity_levels/Classification_of_Obesity_Levels.ipynb at main · jgalloway42/classification_of_obesity_levels · GitHub. Available at: https://github.com/jgalloway42/classification_of_obesity_levels/blob/main/Classification_of_Obesity_Levels.ipynb (Accessed: 16 April 2025).
15. Garg, S. and Pundir, P. (2021) *MOFit: A Framework to reduce Obesity using Machine learning and IoT*, p. 1740. Available at: https://doi.org/10.23919/MIPRO52101.2021.9596673.
16. Goutte, C. and Gaussier, E. (2005) ‘A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation’, in D.E. Losada and J.M. Fernández-Luna (eds) *Advances in Information Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 345–359.
17. Levy, J.J. and O’Malley, A.J. (2020) ‘Don’t dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning’, *BMC Medical Research Methodology*, 20(1), p. 171. Available at: https://doi.org/10.1186/s12874-020-01046-3.
18. Hildebrand, S. and Pfeifer, A. (2025) ‘The obesity pandemic and its impact on non-communicable disease burden’, *Pflügers Archiv - European Journal of*

Physiology [Preprint]. Available at: https://doi.org/10.1007/s00424-025-03066-8.

19. Miller, C. *et al.* (2024) 'A review of model evaluation metrics for machine learning in genetics and genomics', *Frontiers in Bioinformatics*, 4, p. 1457619. Available at: https://doi.org/10.3389/fbinf.2024.1457619.
20. Mistral AI (2025) *Le Chat*. Available at: https://www.mistral.ai (Accessed: 14 April 2025).
21. Musa, F., Basaky, F. and Osaghae, E. (2022) 'Obesity prediction using machine learning techniques', *Journal of Applied Artificial Intelligence*, 3, pp. 24–33. Available at: https://doi.org/10.48185/jaai.v3i1.470.
22. Palechor, F.M. and Manotas, A. de la H. (2019) 'Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico', *Data in Brief*, 25, p. 104344. Available at: https://doi.org/10.1016/j.dib.2019.104344.
23. Rimal, Y., Sharma, N. and Alsadoon, A. (2024) 'The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms', *Multimedia Tools and Applications*, 83(30), pp. 74349–74364. Available at: https://doi.org/10.1007/s11042-024-18426-2.
24. Scikit-learn (2025) *MinMaxScaler*, *scikit-learn*. Available at: https://scikit-learn/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html (Accessed: 14 April 2025).
25. Scikit-learn (2025) *train_test_split*, *scikit-learn*. Available at: https://scikit-learn/stable/modules/generated/sklearn.model_selection.train_test_split.html (Accessed: 14 April 2025).
26. Shen, S. (2022) 'ITP 449 Exploratory Data Analysis Project: Obesity Levels Based on Eating Habits and Physical…', *Medium*, 30 May. Available at: https://stacyy.medium.com/itp-449-
27. Singla, M. and Shukla, K.K. (2020) 'Robust statistics-based support vector machine and its variants: a survey', *Neural Computing and Applications*, 32(15), pp. 11173–11194. Available at: https://doi.org/10.1007/s00521-019-04627-6.

28. Thamrin, S.A. *et al.* (2021) ‘Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018’, *Frontiers in Nutrition*, 8. Available at: https://doi.org/10.3389/fnut.2021.669155.

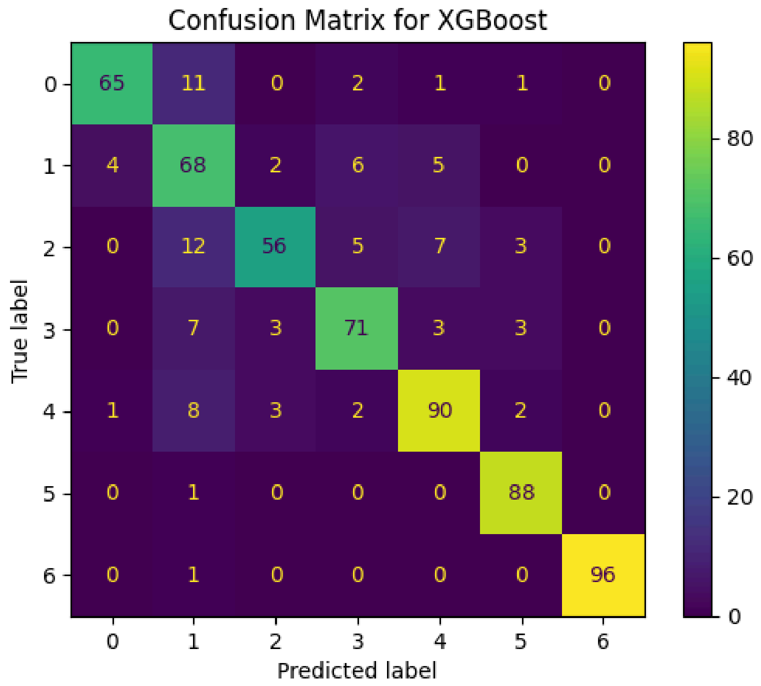
29. UCI (2019) ‘Estimation of Obesity Levels Based On Eating Habits and Physical Condition’. UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition (Accessed: 18 April 2025).

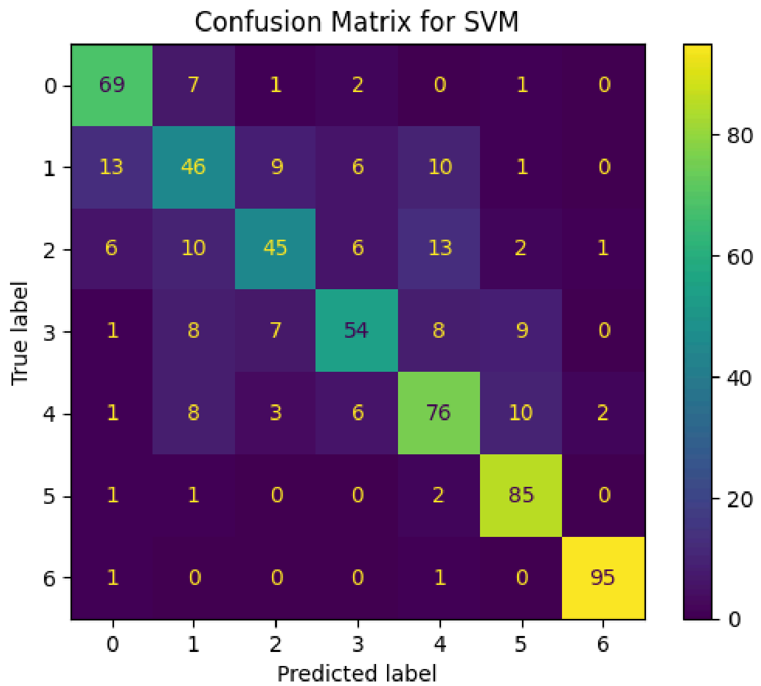
30. WHO (2024) Obesity and overweight. Available at: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (Accessed: 2 April 2025).

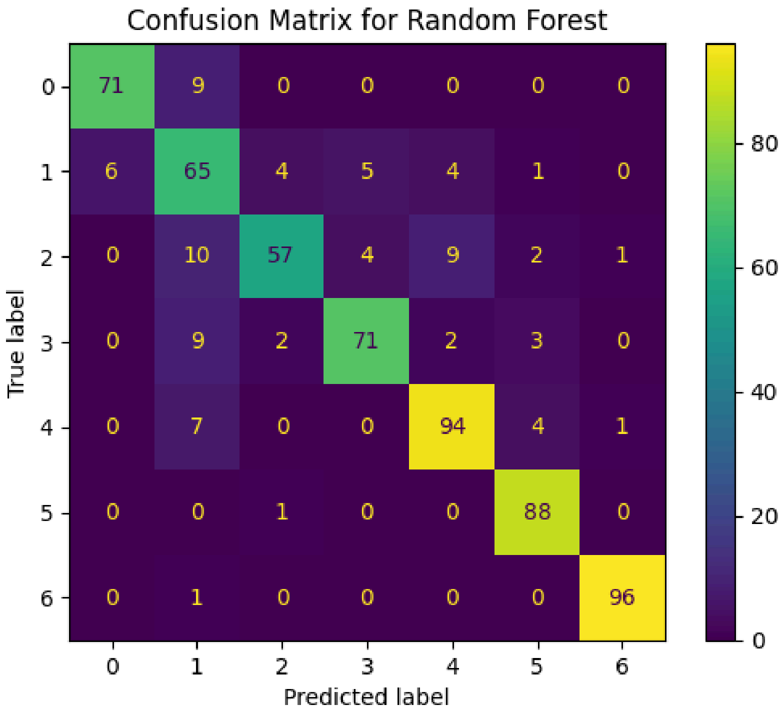
31. Wong, J.E. *et al.* (2022) ‘Predicting Overweight and Obesity Status Among Malaysian Working Adults With Machine Learning or Logistic Regression: Retrospective Comparison Study’, *JMIR Form Res*, 6(12), p. e40404. Available at: https://doi.org/10.2196/40404.

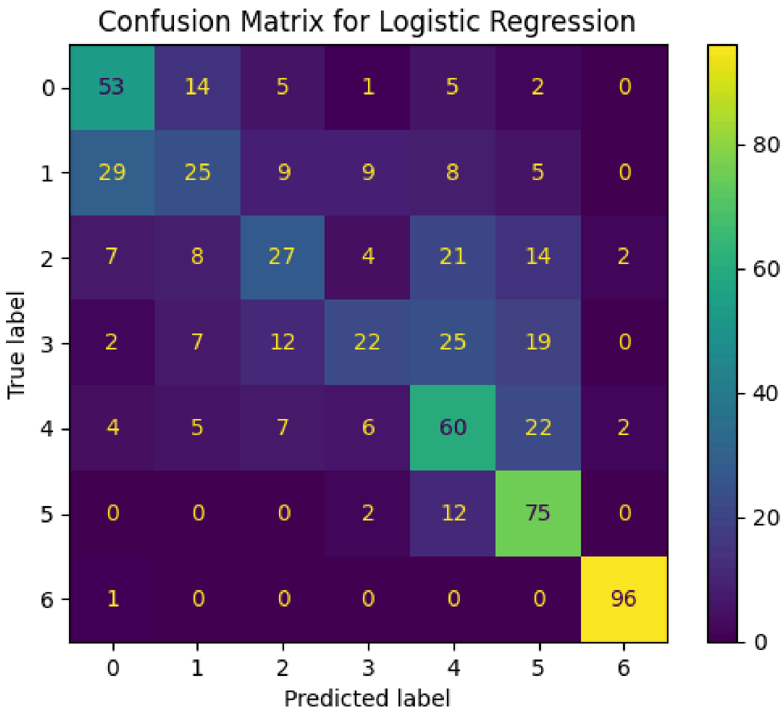
Appendix

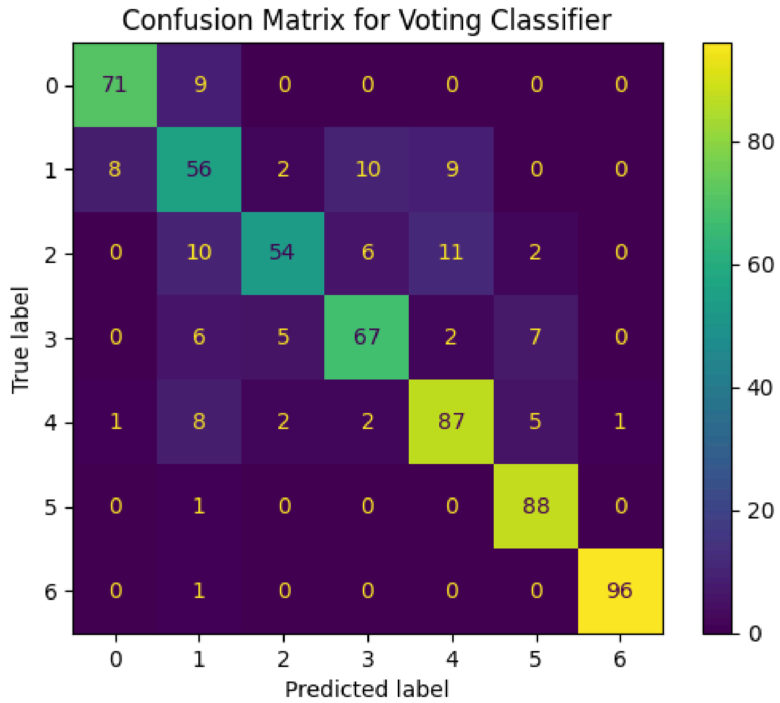
Appendix 1 – Confusion Matrices






Confusion Matrix for Random Forest
71 9 0 0 0 0 0
6 65 4 5 4 1 0
0 10 57 4 9 2 1
0 9 2 71 2 3 0
0 7 0 0 94 4 1
0 0 1 0 0 88 0
0 1 0 0 0 0 96
True label
Predicted label
0 1 2 3 4 5 6
0 20 40 60 80


Confusion Matrix for Logistic Regression
53 14 5 1 5 2 0
29 25 9 9 8 5 0
7 8 27 4 21 14 2
2 7 12 22 25 19 0
4 5 7 6 60 22 2
0 0 0 2 12 75 0
1 0 0 0 0 0 96
True label
Predicted label
0 1 2 3 4 5 6
0 20 40 60 80


Confusion Matrix for Voting Classifier
True label
Predicted label
0 1 2 3 4 5 6
71 9 0 0 0 0 0
8 56 2 10 9 0 0
0 10 54 6 11 2 0
0 6 5 67 2 7 0
1 8 2 2 87 5 1
0 1 0 0 0 88 0
0 1 0 0 0 0 96
0
20
40
60
80