A Multimodal Neural Network for Binary Classification of Obesity Risk Based on Lifestyle Choices, Eating Habits, and Other Personal Attributes

MDSC 523 Fall 2023
Project Proposal
Abdullah Zubair
UCID: 30113730

Background

Obesity, a global epidemic, has seen its prevalence nearly triple since 1975, reflecting a disturbing trend in public health. In 2016, over 1.9 billion adults were overweight, with over 650 million among them classified as obese [1]. This issue is not limited to adults; in 2020, around 39 million children under five were overweight or obese [1]. In Canada, the magnitude of the problem is equally alarming. In 2018, 26.8% of Canadians aged 18 and older were obese, and 36.3% were overweight [2]. This brings the total population at increased health risk due to excess weight to 63.1%, a noticeable increase from 61.9% in 2015 [2].

The health implications of obesity are extensive and multifaceted[3]. It profoundly diminishes almost every aspect of health, from reproductive and respiratory function to memory and mood, and significantly increases the risk of diseases such as diabetes, heart disease, and certain cancers [3]. In terms of specific diseases, obesity is most strongly associated with type 2 diabetes, with a dramatic increase in risk correlated with higher body mass indices (BMIs) [3]. For instance, women with a BMI of 35 or higher had a 93 times higher risk of developing diabetes than those with a lower BMI [3]. The risk extends to cardiovascular diseases, as obesity correlates with an increase in blood pressure, LDL cholesterol, triglycerides, blood sugar, and inflammation, all of which are risk factors for coronary heart disease, stroke, and cardiovascular death [3]. Futhermore, obesity has been linked to various cancers,

including those of the esophagus, pancreas, colon and rectum, breast, endometrium, and kidney [3]. Additionally, obesity's association with mental health, particularly depression, and its effect on quality of life cannot be overlooked [3]. Obesity also complicates reproductive health, contributing to infertility, increased risks during pregnancy, and even congenital anomalies [3]. Obesity also impairs lung function, exacerbating conditions like asthma and obstructive sleep apnea, which affect a significant portion of the obese population [3]. Finally, obesity also has implications for cognitive health, increasing the risk of Alzheimer's disease and dementia, and places a mechanical and metabolic strain on the musculoskeletal system, leading to conditions like arthritis and osteoarthritis [3]. These health risks collectively underscore the urgency of addressing obesity through prevention and treatment strategies to improve public health and reduce healthcare costs [3].

Personal choices, such as lifestyle behaviors, play a pivotal role in influencing obesity risk, underscoring the importance of understanding the intricate relationship between obesity and lifestyle factors for effective prevention and management [4,5,6]. Lifestyle choices, including physical activity, dietary habits, and sedentary behaviors, have a direct impact on weight and general or abdominal obesity, with obese individuals often less active, particularly in terms of activity, and exhibiting less favorable dietary habits, such as lower intake of breakfast, fruits, and milk. Interestingly, lower intake of sugar-sweetened drinks and sweets/chocolates has also been associated with obesity [4,5,6]. Primary prevention of obesity through the

promotion of active lifestyles and healthy diets is emphasized as a public health priority, with research showing that factors like being male, consuming breakfast less than three times a week, and consuming sugar-sweetened drinks are significantly associated with overweight and obesity, while engaging in vigorous physical activity is associated with a reduced risk [4,5,6]. Having a tool that could take in the various factors and lifestyle choices an individual is making and then give the risk of various disease such as obesity could be useful in the field of preventative healthcare. The tool could then offer tailored recommendations for lifestyle modifications to effectively manage or prevent obesity. This approach could significantly contribute to the reduction of obesity prevalence and its associated health risks, offering a proactive solution in the fight against this global epidemic.

Objectives

The primary objective is to develop and validate a neural network model capable of accurately predicting obesity risk. This model will integrate multiple data types, including dietary habits, physical activity levels, and other relevant personal attributes, to facilitate a nuanced understanding of obesity risk factors. This model will be a binary classification model which will identify if an individual is at risk of developing obesity based on the individual's data. The creation of this neural network will involve preprocessing the data, splitting the data into training, validation and testing, optimizing the various parameters, and then observing the model's performance based

on metrics such as accuracy and precision with a specific focus on recall as for our use case having the least number of false negatives is the most important.

Methodology

The utilized dataset originates from the UCI Machine Learning Repository and is titled "Estimation of Obesity Levels Based on Eating Habits and Physical Condition". It comprises data from individuals in Mexico, Peru, and Colombia, focusing on obesity levels determined by their eating habits and physical conditions. The dataset is multivariate, encompassing health and medicine subject areas, and suitable for classification, regression, and clustering tasks. It contains 17 attributes across 2,111 records. These attributes include categorical, continuous, and binary features such as gender, age, height, weight, family history with overweight, frequency of high caloric food consumption, vegetable intake, number of main meals per day, eating habits between meals, and smoking behavior.

The model was developed using Python specifically a Jupyter notebook was used to run and code the model. Libraries such as Pandas, NumPy, and PyTorch, which are instrumental for data manipulation and neural network construction were used. The dataset is first loaded and processed, with the target variable 'NObeyesdad' being converted into binary form to distinguish between 'Obese' and 'Not Obese' classes. This binary classification is crucial for the model's goal to predict obesity risk. The categorical variables in the dataset are encoded using the LabelEncoder, and feature

selection is performed using Mutual Information (mi_selector), focusing on the top 10 features. This step is important for reducing dimensionality and highlighting the most relevant predictors of obesity. The continuous features are then standardized using StandardScaler, ensuring that the model is not biased towards variables with larger scales. The data is split into training, validation, and testing sets, maintaining a balance that allows for effective model training and evaluation without overfitting or underfitting. In constructing the neural network model, a simple architecture with three fully connected layers is employed. The first two layers use ReLU activation functions, while the output layer uses a sigmoid function, suitable for binary classification. The model dimensions are set based on the number of selected features (10 features) and the complexity of the task at hand.

The training and evaluation of the model are meticulously executed to ensure accuracy and effectiveness. The training loop incorporates both the training and validation phases, with the former focused on model learning and the latter on performance assessment. During training, the optimizer 'Adam' is employed with a learning rate of 0.001, which dictates the step size at each iteration while moving toward a minimum of the loss function. This learning rate is chosen for its balance between rapid convergence and stability. The model is trained over 30 epochs, where an epoch represents a complete pass through the entire training dataset. The training phase includes calculating loss using the Binary Cross-Entropy Loss (BCELoss) function, suitable for binary classification tasks, and backpropagation for updating the model's

weights. Accuracy metrics, such as training accuracy, are calculated by comparing the predicted outputs to actual labels. The model's performance is also evaluated in the validation phase, where it processes unseen data, and this phase's loss and accuracy are used to gauge the model's generalization capabilities. Early stopping is implemented with a patience of 5 epochs to prevent overfitting, where the model training halts if the validation loss does not improve for five consecutive epochs. This mechanism ensures that the model does not overlearn from the training data and maintains its ability to generalize to new data. Lastly, the accuracies from both training and validation phases are plotted against epochs, providing a visual representation of the model's learning curve and its ability to adapt and improve over time. This comprehensive training and evaluation process aims to optimize the model for accurate obesity risk prediction, ensuring its applicability and reliability in real-world settings.

Preliminary Analysis

Gender	Age	Height	Weight	family_histo	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
Female	21	1.62	64	yes	no	2	3	Sometimes	no	2	no	0	1	no	Public_Transportation	Normal_Weight
Female	21	1.52	56	yes	no	3	3	Sometimes	yes	3	yes	3	0	Sometimes	Public_Transportation	Normal_Weight
Male	23	1.8	77	yes	no	2	3	Sometimes	no	2	no	2	1	Frequently	Public_Transportation	Normal_Weight
Male	27	1.8	87	no	no	3	3	Sometimes	no	2	no	2	0	Frequently	Walking	Overweight_Level_I
Male	22	1.78	89.8	no	no	2	1	Sometimes	no	2	no	0	0	Sometimes	Public_Transportation	Overweight_Level_II
Male	29	1.62	53	no	yes	2	3	Sometimes	no	2	no	0	0	Sometimes	Automobile	Normal_Weight
Female	23	1.5	55	yes	yes	3	3	Sometimes	no	2	no	1	0	Sometimes	Motorbike	Normal_Weight
Male	22	1.64	53	no	no	2	3	Sometimes	no	2	no	3	0	Sometimes	Public_Transportation	Normal_Weight
Male	24	1.78	64	yes	yes	3	3	Sometimes	no	2	no	1	1	Frequently	Public_Transportation	Normal_Weight
Male	22	1.72	68	yes	yes	2	3	Sometimes	no	2	no	1	1	no	Public_Transportation	Normal_Weight
Male	26	1.85	105	yes	yes	3	3	Frequently	no	3	no	2	2	Sometimes	Public_Transportation	Obesity_Type_I
Female	21	1.72	80	yes	yes	2	3	Frequently	no	2	yes	2	1	Sometimes	Public_Transportation	Overweight_Level_II
Male	22	1.65	56	no	no	3	3	Sometimes	no	3	no	2	0	Sometimes	Public_Transportation	Normal_Weight
Male	41	1.8	99	no	yes	2	3	Sometimes	no	2	no	2	1	Frequently	Automobile	Obesity_Type_I
Male	23	1.77	60	yes	yes	3	1	Sometimes	no	1	no	1	1	Sometimes	Public_Transportation	Normal_Weight
Female	22	1.7	66	yes	no	3	3	Always	no	2	yes	2	1	Sometimes	Public_Transportation	Normal_Weight
Male	27	1.93	102	yes	yes	2	1	Sometimes	no	1	no	1	0	Sometimes	Public_Transportation	Overweight_Level_II
Female	29	1.53	78	no	yes	2	1	Sometimes	no	2	no	0	0	no	Automobile	Obesity_Type_I
Female	30	1.71	82	yes	yes	3	4	Frequently	yes	1	no	0	0	no	Automobile	Overweight_Level_II
Female	23	1.65	70	yes	no	2	1	Sometimes	no	2	no	0	0	Sometimes	Public_Transportation	Overweight_Level_I
Male	22	1.65	80	yes	no	2	3	Sometimes	no	2	no	3	2	no	Walking	Overweight_Level_II
Female	52	1.69	87	yes	yes	3	1	Sometimes	yes	2	no	0	0	no	Automobile	Obesity_Type_I
Female	22	1.65	60	yes	yes	3	3	Sometimes	no	2	no	1	0	Sometimes	Automobile	Normal_Weight
Female	22	1.6	82	yes	yes	1	1	Sometimes	no	2	no	0	2	Sometimes	Public_Transportation	Obesity_Type_I
Male	21	1.85	68	yes	yes	2	3	Sometimes	no	2	no	0	1	Sometimes	Public_Transportation	Normal_Weight
Male	20	1.6	50	yes	no	2	4	Frequently	yes	2	no	3	2	no	Public_Transportation	Normal_Weight
Male	21	1.7	65	yes	yes	2	1	Frequently	no	2	no	1	2	Always	Walking	Normal_Weight
Female	23	1.6	52	no	yes	2	4	Frequently	no	2	no	2	1	Sometimes	Automobile	Normal_Weight
Male	19	1.75	76	yes	yes	3	3	Sometimes	no	2	yes	3	1	Sometimes	Public_Transportation	Normal_Weight

Figure 1. Dataset obtained from UC Irvine Machine Learning Repository Containing 17 attributes and 2111 records.

From the dataset obtained from the UC Irvine repository there are 17 attributes given for each individual. These include: Gender: The gender of the individual, Age: The age of the individual, Height: The height of the individual in meters, Weight: The weight of the individual in kilograms, Family History with Overweight: Whether there is a family history of overweight, FAVC: Frequent consumption of high caloric food (yes/no), FCVC: Frequency of consumption of vegetables, NCP: Number of main meals, CAEC: Consumption of food between meals, SMOKE: Smoking status, CH2O: Consumption of water daily, SCC: Calories consumption monitoring, FAF: Physical activity frequency, TUE: Time using technology devices, CALC: Consumption of alcohol, MTRANS: Mode of Transportation, and NObeyesdad: Obesity level classification. Of the 2111 records, 77% was artificially created using the Weka tool and the SMOTE filter. Weka is a software for applying machine learning techniques, while SMOTE is a method for generating synthetic data. SMOTE specifically helps in balancing datasets by creating new samples for underrepresented classes. This approach enhances the dataset for more effective training of machine learning models.

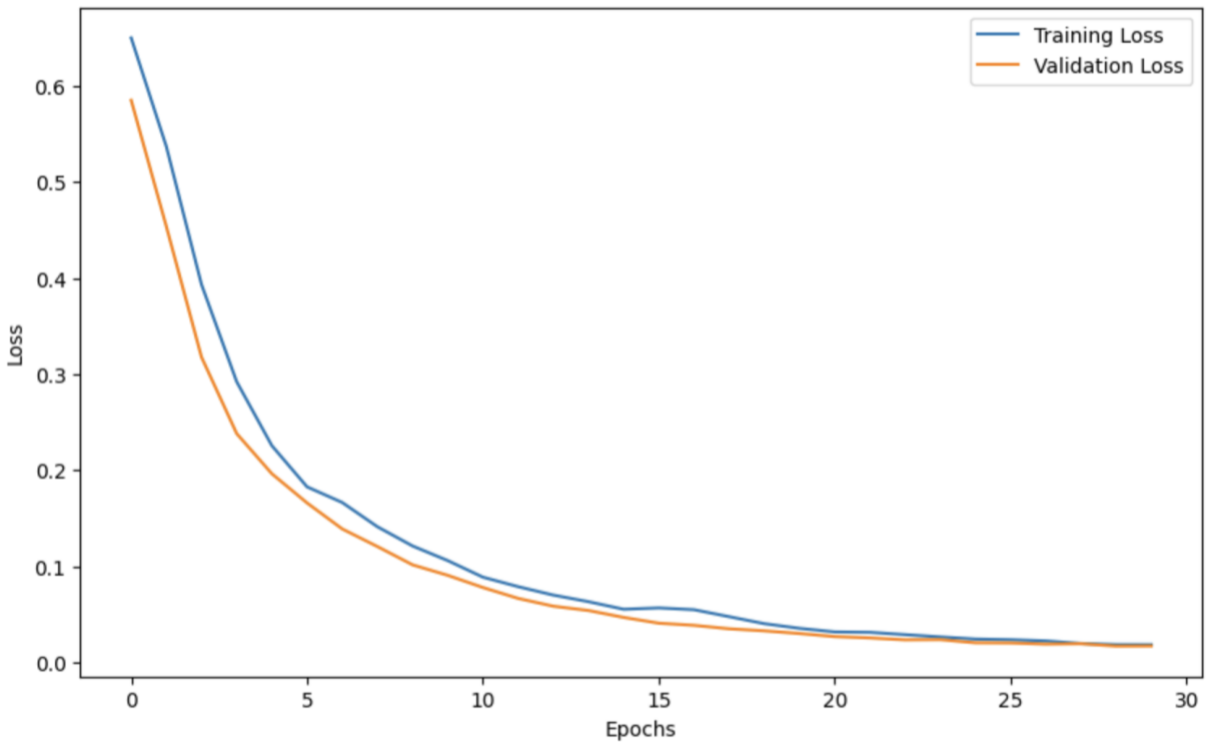

Figure 2. Training and Validation Loss per Epoch.

Overall, Figure 2 illustrates the model's loss on both the training and validation datasets. The loss is a measure of how well the model's predictions align with the actual data. Initially, both training and validation loss values are high, indicating significant error in the model's predictions. As the epochs progress, there is a sharp decline in loss, suggesting that the model is learning and improving its predictive accuracy. The curves converge and flatten out, which indicates that the model has reached a stable state where further training does not significantly improve performance. The closeness of the training and validation loss lines also suggests that the model is generalizing well and not overfitting to the training data.

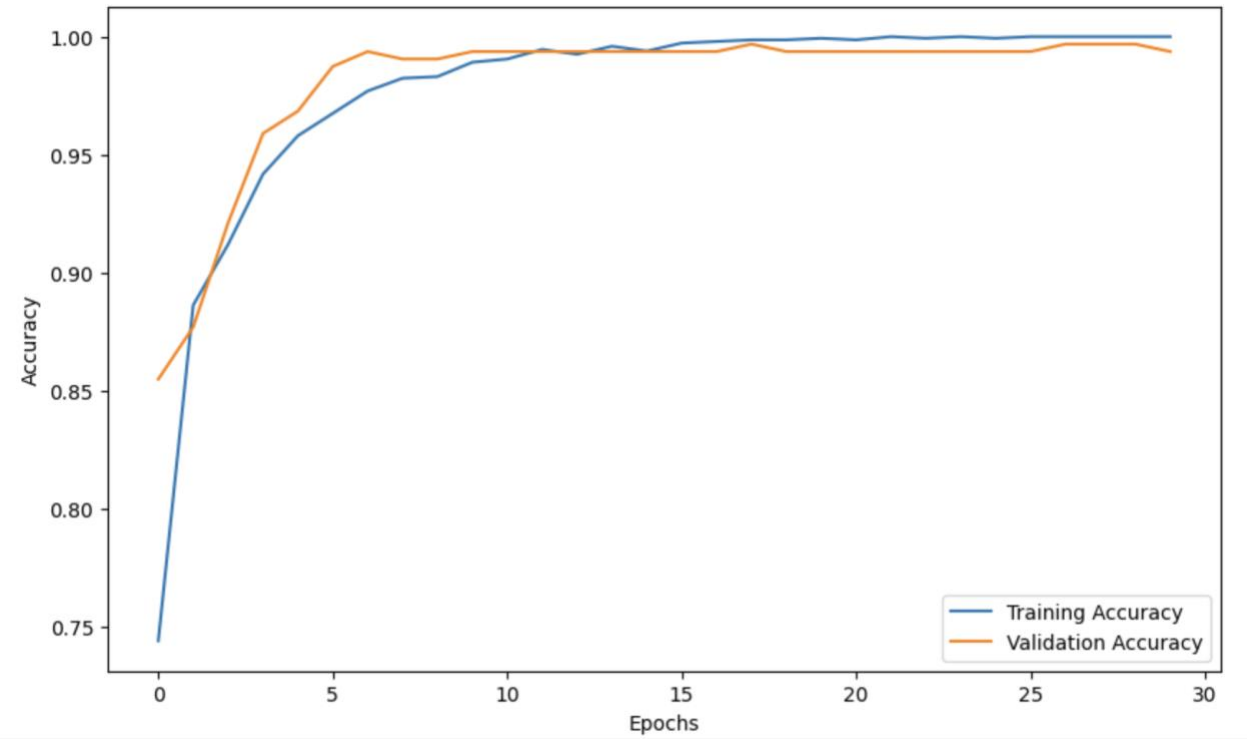

Figure 3. Training and Validation Accuracy per Epoch.

By observing Figure 3, the accuracy of the model on both the training and validation datasets can be seen. Accuracy is the proportion of true results among the total number of cases examined. Figure 3 reveals a rapid increase in accuracy during the initial epochs, demonstrating the model's swift learning curve. As training progresses, both training and validation accuracies plateau, indicating that the model has reached its optimal performance level. Notably, the validation accuracy closely tracks the training accuracy, which is a positive indicator of the model's generalization capabilities. The model does not appear to overfit, as the validation accuracy does not significantly diverge from the training accuracy. Overall, these figures demonstrate a successful training process, with the model achieving high accuracy and low loss,

suggesting it is well-tuned for the task of classifying obesity risk based on the given dataset.

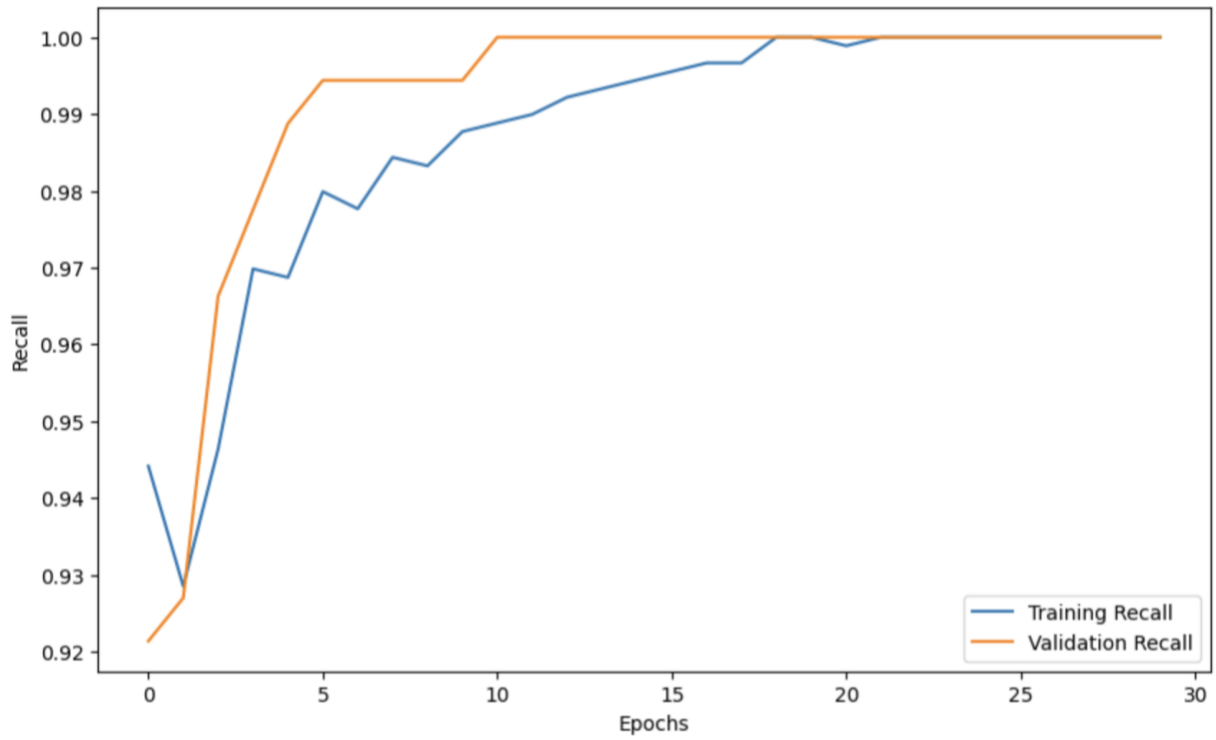

Figure 4. Training and Validation Recall per Epoch.

Figure 4 shows the recall metric trajectory for the training and validation datasets over a span of 30 epochs. Recall is an important metric in binary classification models, especially in medical or health-related contexts, is pivotal as it measures the model's capability to correctly identify all actual positive cases. The importance of recall in this obesity risk prediction model stems from the grave implications associated with false negatives; failing to identify an individual at risk for obesity could lead to missed opportunities for preventive health interventions. The y-axis quantifies the recall value, indicating that the model demonstrates a remarkable ability to identify true positives, as

the recall values are close to the ideal score of 1.00. The initial phase shows a steep ascent in recall for both datasets, suggesting a rapid learning rate where the model swiftly enhances its predictive accuracy for positive cases. As training progresses, the recall for both datasets achieves and sustains high performance, with the training recall slightly outpacing the validation recall. This convergence of recall values illustrates the model's adeptness at generalizing well to unseen data while maintaining robust sensitivity. Figure 4 not only confirms the model's effectiveness and reliability in detecting positive instances, crucial for a condition like obesity where early identification is essential, but also underscores why recall is an important metric for this model, as the cost of not recognizing at-risk individuals is significantly high.

Conclusions and Future Directions

The preliminary analysis of the model demonstrates the neural network model's ability to accurately predict obesity risk, as evidenced by high accuracy and recall metrics alongside low training loss. The accuracy metric highlights the model's overall correctness in classification, while the high recall specifically indicates the model's strength in identifying individuals at high risk for obesity, a critical aspect of the binary classification task at hand. The low training loss corroborates the model's efficiency in learning from the dataset and its capability to generalize well to unseen data, minimizing the error between predicted and actual values. The convergence of training and validation recall scores assures that the model is sensitive and reliable, avoiding the critical issue of false negatives, which is of utmost importance in health-related

predictions. The high performance in these key metrics positions the model as a valuable tool for healthcare professionals and policymakers aiming to mitigate obesity risk through early identification and intervention. Looking ahead, the expansion of the dataset to include more diverse populations from different countries presents an exciting avenue for enhancing the model's applicability and robustness. As lifestyle and environmental factors contributing to obesity may vary significantly across regions, incorporating data that reflect these regional differences could unveil nuanced insights into the relative impact of various metrics on obesity. For instance, the method of transit to work might hold varying levels of significance in different locations, depending on the baseline physical activity levels inherent to those regions. To add on to this, the model has the potential to be extended into a multi-class classification model instead of just a binary classification model. Since the dataset encompasses multiple levels of obesity, the model could be refined to not only predict the risk of obesity but also determine the severity of risk. The development of such a model could be very useful and provide one tool that combats the obesity epidemic.

Appendix

The code for the model which is a Jupyter notebook file as well as the dataset in CSV format are attached to this document.

References

1. World Health Organization. Obesity and overweight [Internet]. Geneva: World Health Organization; 2020 [cited 2023 Dec 16]. Available from: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
2. Statistics Canada. Overweight and obese adults, 2018 [Internet]. Ottawa: Statistics Canada; 2019 [cited 2023 Dec 16]. Available from: https://www150.statcan.gc.ca/n1/pub/82-625-x/2019001/article/00005-eng.htm#
3. Harvard T.H. Chan School of Public Health. Health Risks | Obesity Prevention Source [Internet]. Boston: Harvard T.H. Chan School of Public Health; [cited 2024 Jan 01]. Available from: https://www.hsph.harvard.edu/obesity-prevention-source/obesity-consequences/health-effects/
4. Kerkadi A, Sadig AH, Bawadi H, Al Thani AA, Al Chetachi W, Akram H, Al-Hazzaa HM, Musaiger AO. The relationship between lifestyle factors and obesity indices among adolescents in Qatar. International journal of environmental research and public health. 2019 Nov;16(22):4428.
5. Rassy N, Van Straaten A, Carette C, Hamer M, Rives-Lange C, Czernichow S. Association of Healthy Lifestyle Factors and Obesity-Related Diseases in Adults in the UK. JAMA Network Open. 2023 May 1;6(5):e2314741-.
6. Cha E, Akazawa MK, Kim KH, Dawkins CR, Lerner HM, Umpierrez G, Dunbar SB. Lifestyle habits and obesity progression in overweight and obese American young adults: Lessons for promoting cardiometabolic health. Nursing & health sciences. 2015 Dec;17(4):467-75.