Exploring Socioeconomic Factors and Accelerated Modeling in Understanding Global Obesity Trends

Manav Singh

Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala, India
singhmanav973@gmail.com

Arun Singh Pundir

Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala ,India
arun.pundir@thapar.edu

***Abstract*— Obesity is a significant worldwide health problem, with prevalence rates reaching epidemic levels. This study explores the intricate relationship between obesity and socioeconomic characteristics.. Furthermore, the application of GPU-accelerated training is used to speed up the process of model construction. By utilizing the parallel processing capabilities of GPUs, the time required for training is significantly decreased, allowing for the examination of computationally demanding models. The results emphasize the crucial significance of comprehending socioeconomic factors in addressing obesity and provide innovative perspectives on expediting predictive modelling in public health studies.**

I. Introduction

Obesity is an escalating worldwide health concern that has reached epidemic levels. In 2022, the World Health Organization (WHO)[1] reported that more than 2.5 billion persons aged 18 and older were overweight, with over 890 million of them classified as obese. This corresponds approximately to a ratio of 1 out of every 8 individuals worldwide. These figures emphasize the importance of understanding the patterns of obesity in different countries and income brackets An examination of the correlation between obesity and socioeconomic variables might provide significant insights into the socioeconomic aspects that contribute to this intricate health problem. Exploring the correlation between obesity and other causes poses numerous difficulties. A major obstacle is locating comprehensive data sets that cover all the relevant challenge. . Data can be spread out across several sources, which necessitates meticulous collection and harmonization. Efficiently reducing the training times for these models posed a substantial technical challenge. The conventional use of CPUs for training models led to excessively long processing times. In order to address this limitation, this paper investigated the potential of leveraging GPUs to enhance the speed of training. However, the process of obtaining and setting up dedicated GPUs can provide logistical difficulties. Thus, a software-based approach was implemented by examining and deploying NVIDIA's RAPIDS[2] suite of rapid data science libraries.

This study aims to tackle the issue of data accessibility by presenting a novel dataset. This dataset includes information on the prevalence of obesity, income class, health expenditure as a % of GDP, per capita GDP, and urban population size for 134 countries from 2000 to 2016. Furthermore, it explores the possibility of using an accelerated data science library such as RAPIDS.ai to construct predictive models for obesity rates.

II. Literature Review

Multiple studies have investigated the correlation between obesity and variables such as healthcare expenditures, GDP per capita, and the size of urban populations. For example the authors in [3] present an updated analysis of the current trends in obesity prevalence and investigate the potential reasons behind the observed deceleration in obesity rates in certain countries. Additionally, they stress the pressing necessity for effective measures to prevent the obesity epidemic. The authors in [4] provide an analytical framework to unravel the causes of obesity, with a specific focus on socioeconomic and intrapersonal aspects.

The writers in [5] examine the impact of urbanization on the worldwide obesity crisis. The authors in [6] examine the correlation between healthcare cost and the prevalence of obesity. The study examines the potential influence of variations in healthcare expenditure between nations on obesity rates and emphasizes the significance of healthcare policies in tackling the obesity pandemic. These publications, together with others, enhance our understanding of the worldwide obesity epidemic and the socioeconomic factors that influence this complex health problem.

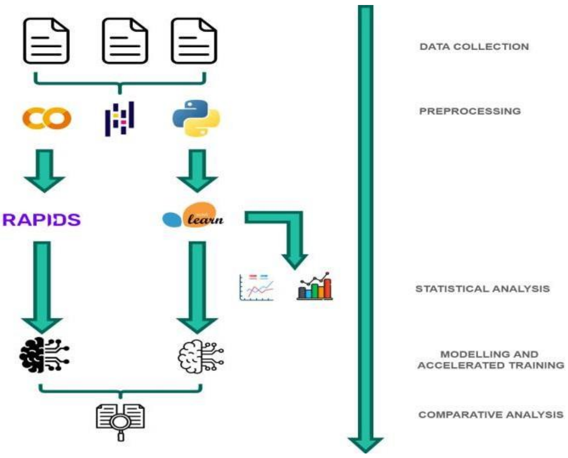

Fig. 1: Flow diagram for the study

III. METHODOLOGY

A. Data Collection:

The statistics on the prevalence of obesity among adults, namely persons with a BMI of 30 or higher, was collected from the World Health Organization's historical dataset titled "Prevalence of obesity among adults, BMI $\geqslant$ 30, age-standardized Estimates by country." The International Monetary Fund's (IMF) database included the annual percentage change in Real Gross Domestic Product (GDP) growth rates as economic indicators. The per capita GDP data was acquired from the World Bank database. The World Bank database also included historical statistics on the proportion of urban population sizes in relation to the total population. Data on healthcare expenditure and economic classification Economic Classifications were also acquired from the World Bank database.

B. Preprocessing:

The preprocessing stage employed popular Python libraries for efficient data manipulation and cleaning. The specific steps undertaken are as follows:

The obesity dataset originally contained distinct columns for the rates of obesity among males and females. Given that this study is specifically examining the overall prevalence of obesity regardless of gender, these columns were omitted. An analysis of missing values and imputation was conducted, carefully examining the fraction of missing values for each country in each dataset. Countries that have missing value ratios exceeding 10% were not included in the analysis. In these places, the process of collecting data may have been impeded. The process was consistently implemented across all five datasets. The final analysis was conducted by establishing a common timeframe of 2000- 2016 through the comparison of data availability across datasets. Missing values in each dataset were resolved using suitable methodologies. The average value of each column was used as the primary strategy for filling in missing data. However, in cases when columns had a substantial number of outliers, the median value was used as a more reliable method for filling in missing values.

Before combining the separate datasets, a thorough analysis of country names from all sources was performed. This step was essential to address potential variations in naming customs among organizations, such as the distinction between "Saint Lucia" and "St. Lucia" or "South Korea" and "Korea". Furthermore, nations that have undergone name changes in recent times, such as Turkey to Türkiye and Kyrgyz Republic to Kyrgyzstan, were recognized and dealt with to guarantee the merger process is accurate, using consistent country identifiers. The objective of this thorough strategy was to optimize the size of the dataset while ensuring the utmost accuracy and worldwide coverage.

C. Correlation and Trend Analysis:

This study involved the creation of line graphs to aid in the depiction of possible temporal patterns and connections between the incidence of obesity and socioeconomic determinants.

A heatmap and correlation matrix were created using the final dataset in order to thoroughly examine the connections between obesity prevalence and the other variables being investigated.

This study employed hypothesis testing to ascertain the presence of statistically significant associations between obesity rates and the other factors in the dataset. This analysis serves to validate or disprove any potential connections uncovered by exploratory techniques such as correlation analysis and trend visualization. Bar graphs were utilized in this study to visualize trends across different economic classes.

D. Modeling and Optimizing with Accelerated Data Science:

This study examined the capacity of different machine learning models to forecast obesity trends using the socioeconomic parameters present in the dataset. The choice of these models took into account the possibility of both linear and non-linear connections between obesity and socioeconomic factors, as well as their ability to handle large amounts of data and the intricacies within it. Utilizing GPUs for accelerated training: To speed up the training process for these models, the possibility of leveraging GPUs for accelerated training was investigated. Although specialized graphics processing units (GPUs) provide notable speed advantages, the process of obtaining and setting them up can be complex. Consequently, a solution that relies on software was pursued.

IV.RESULT AND ANALYSIS:

A. Statistical Analysis :

A visual representation of obesity rates across the analyzed timeframe (2000-2016) shows a worrisome increase over time. This continuous rise corresponds with the conclusions of the authors in [7], where they emphasize the increasing occurrence of obesity as a major worldwide public health challenge. While obesity rates consistently increase, GDP growth rates display a more erratic pattern. This volatility corresponds to the intrinsic intricacies of economic systems.[8] A hypothesis test indicated a statistically insignificant correlation between the two variables. This discovery is consistent with the increasing amount of data indicating that economic growth by itself may not be enough to prevent the increase in obesity rates.[9] As such there is no reason to include factor in further analysis. In contrast to the consistent upward trajectory of obesity rates, healthcare expenditure has a cyclical pattern, marked by periods of substantial growth followed by modest declines. The global "zig-zag" pattern may be attributed to an intricate interplay of various causes, which include: Demographic shifts, Technological advancements, and Alterations in government policies and laws concerning healthcare. Although the increase in healthcare cost does not follow a straight line like the growth in obesity rates, the hypothesis test showed a statistically significant positive association between the two variables.

The rising prevalence of obesity results in an increased need for healthcare services to address obesity-related health conditions such as diabetes and heart disease. Consequently, healthcare spending also increases.[10]

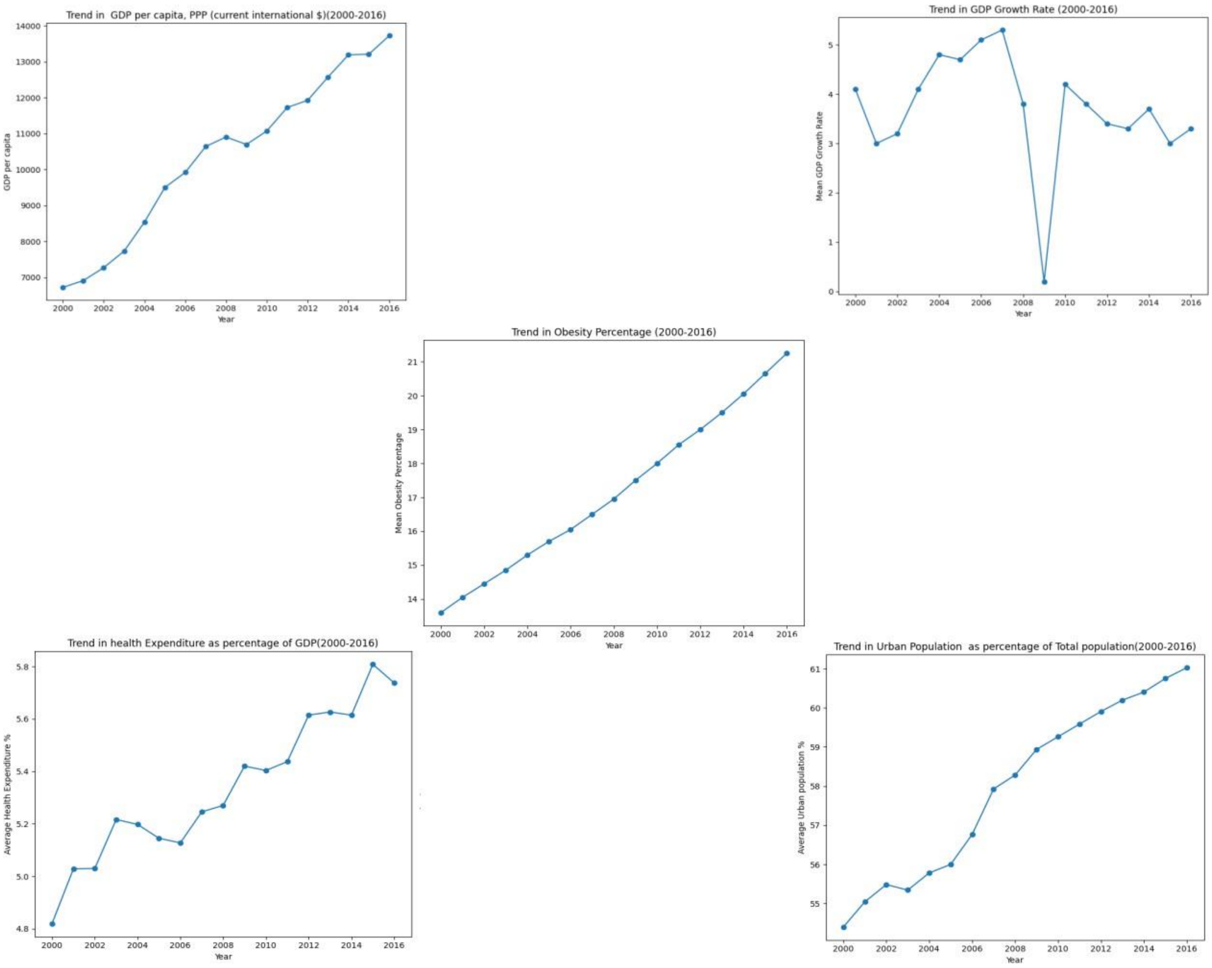

Fig 2. Comparative Trends in Socioeconomic and Health Indicators Across Countries (2000-2016)

The per capita GDP also exhibited an upward trend but there is a significant decline in per capita GDP around 2008-2009. The decrease can be ascribed to the extensively documented worldwide financial crisis during that time. It is noteworthy that even during this period, the trends of obesity remained consistently stable. The hypothesis test demonstrated a statistically significant and positive relationship between per capita GDP and obesity rates. This analysis is consistent with the concept of the "nutrition transition" described in [7]. The nutrition transition concept posits that economic progress is frequently followed by changes in eating habits towards the consumption of processed foods that are rich in calories and harmful fats. This shift in dietary patterns might contribute to the increasing prevalence of obesity. Continuing the investigation the line graph of urban populations was analysed .It showed progressive increase .This increase in general corresponds to the extensively documented occurrence of global urbanization, as elucidated by World Urbanization Prospects: The 2018 Revision [11]. The hypothesis test showed a statistically significant positive link between urban population size and obesity rates, similar to the analysis of per capita GDP. This finding aligns with research conducted in [7]. Urbanization results in dietary changes towards processed foods that are rich in calories potentially leading to a higher prevalence of obesity. Also , urban surroundings may provide limited possibilities for physical activity due to dependence on transportation and a decrease in physically demanding occupations.

The diverse levels of correlation between the obesity rate and the analyzed factors as indicated by the heatmap indicate complex connections that require further examination. The size of the urban population shows the strongest link with the proportion of obesity. The strong positive correlation observed in this study highlights the significant impact of urbanization on the prevalence of obesity.

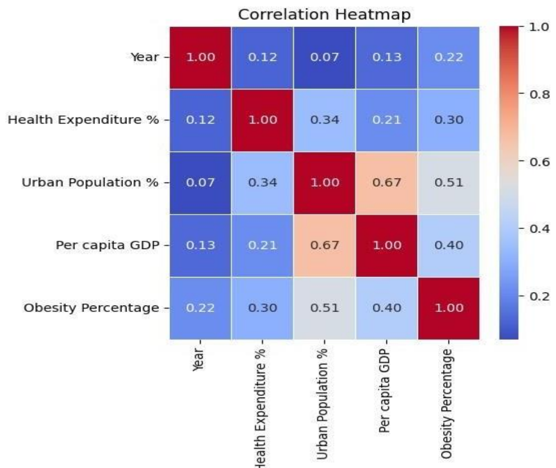

Fig 3. This heatmap depicts the correlation coefficients between all the features of the final in dataset

On the other hand, the moderately positive correlations between the percentage of obesity and both health expenditure and per capita GDP indicate a more complex link. This decrease in intensity may arise from various sources. The complex structure of healthcare systems and how they distribute. Although more healthcare spending may indicate improved treatment strategies for obesity-related illnesses, its effectiveness in reducing obesity itself is influenced by broader socioeconomic variables and lifestyle choices. The association between per capita GDP and obesity is also complex. Wealth can provide individuals with more options for both healthy and bad lifestyle choices, which might affect the strength of the correlation seen. To clarify the complex connection between economic classes and the prevalence of obesity a series of bar graphs for each year over this period was created, allowing for a comparative analysis of average obesity rates among different economic groups. Interestingly, the results consistently showed that wealthier countries had higher rates of obesity. This supports previous research that has found a positive relationship between economic prosperity and the prevalence of obesity [12]. Moreover, the difference between obesity rates among high-income and upper-middle-income countries was relatively small, it expanded considerably between upper-middle-income and lower-middle-income nations, and even more so between low-income and lower-middle-income countries. The increase in obesity rates based on economic status highlights the worsening impact of economic Inequalities on health outcomes, a well-documented phenomena in research [13]. Following bar graphs illustrating the trend of average obesity rates for each economic class during the study period confirmed these findings, showing a consistent and rapid rate of increase in high-income countries, followed by upper-middle-income countries, and then low-middle- income and low-income countries. This phenomena aligns with previous studies that emphasize the impact of swift urbanization, inactive lifestyles, and dietary shifts linked to economic progress on the rise of obesity [7]. Nevertheless, during the period of 2007 to 2010, there was a significant decrease in obesity rates in low and lower- middle-income nations, which contrasts with the overall trend of increasing obesity rates. The occurrence of this historical anomaly might be ascribed to different contextual variables, such as worldwide economic recessions or fluctuations in food costs. These factors might have affected dietary patterns, food security, and access to healthcare, thereby influencing the prevalence of obesity [14].

Furthermore, it is possible that socio-political upheavals, environmental disasters, or public health measures occurring during this time could have caused temporary changes in obesity rates. This suggests the need for more research to understand these oscillations. To summarize, the thorough analysis of the connection between economic class and the prevalence of obesity highlights the complex interaction of socioeconomic factors on health outcomes. This emphasizes the importance of implementing specific interventions that address economic inequalities in order to reduce the growing global obesity crisis.

B. *Modelling and Accelerated Training:*

To achieve accurate predictive modelling for the dataset being analysed, a careful method was used to choose and evaluate four machine learning models that are well-known for their effectiveness in regression tasks. Particular attention was given to integrating models that utilize both bagging and boosting strategies, as both methods have a tendency to improve forecast accuracy by combining many models and iteratively refining them. The decision to use bagging and boosting techniques in predictive modelling tasks is based on the understanding that these strategies are beneficial in reducing overfitting and enhancing the ability to generalize, as acknowledged in the literature.[15]

Fig 4. *Average Obesity Rates by Economic Class (2000-2016)*

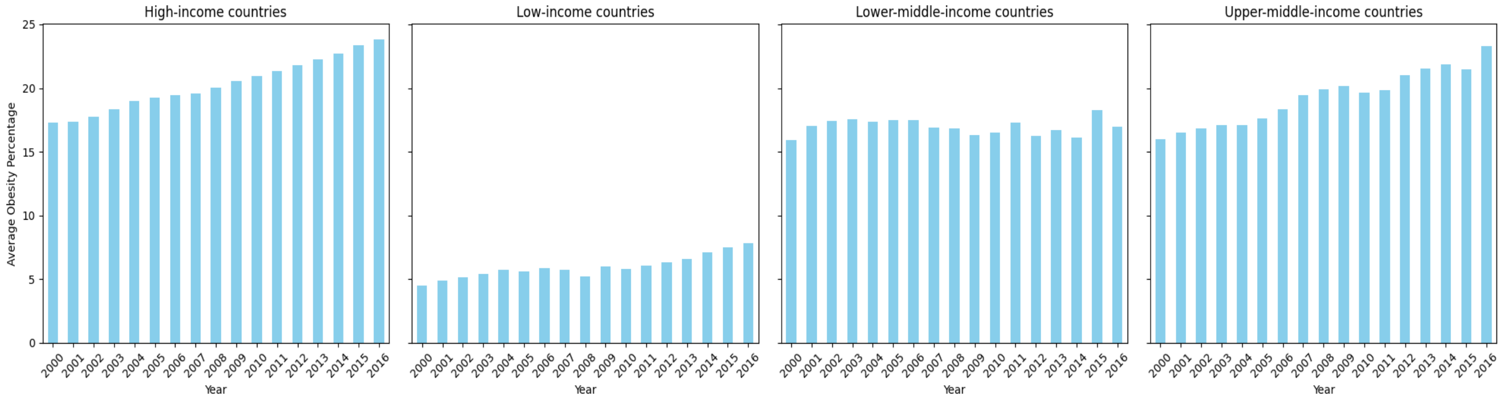

Fig 5. *Bar chart showing the average obesity percentages in four income classes from 2000 to 2016*

The selected models included a wide range of techniques, such as linear regression, decision trees, random forest, and XGBoost. Each of these models is well-known for its unique abilities to capture different levels of complexity and nonlinearity in the data. Before training the model, a decision was taken to use 10- fold cross-validation, a generally recommended technique that is crucial for evaluating the model's performance and estimating its effectiveness [16]. This methodological approach guarantees an impartial evaluation of the model's capacity to generalize by repeatedly dividing the dataset into training and validation subsets. This process produces dependable estimations of the model's predictive performance. After instantiating the models, a process of feature selection followed, with the goal of determining the best mix of predictor variables to use in the modelling process. The inclusion of numerical features such as per capita GDP, urban population size, and healthcare expenditure was considered essential.

However, incorporating categorical features required careful research to determine their effect on predictive accuracy. The initial modelling iterations mostly concentrated on numerical predictors. Subsequent iterations gradually integrated categorical data, such as country and year. The results for some of these iterations are indicated in a tabular form. Notably, the gradual addition of categorical characteristics resulted in subtle enhancements in predicted accuracy across models, but this came at the cost of increasing computational burden. In the end, adding categorical characteristics, specifically country and year, improved the predictive performance by the largest amount. The linear regression model was shown to be the most effective. The values shown in the tables are rounded up to 2 decimal places.

However, due to the limited amount of time available for training, made worse by the large size of the dataset and the process of selecting features in an iterative manner, to optimize training time it was necessary to use different training methods. Considering these limitations, an innovative strategy utilizing GPU-accelerated training was implemented, resulting in the adoption of the Rapids AI platform[2]. Rapids AI is a new approach to predictive modelling that utilizes the powerful parallel processing capabilities of GPUs to speed up model training and inference. For the purposes of this implementation, the Google colab platform was chosen and the powerful T4 GPU was used. The use of GPU-accelerated training resulted in significant enhancements in computational efficiency across all models. The results of which are summarized in a table below. Crucially, these substantial improvements in the speed of training were accomplished without sacrificing the ability to accurately forecast, as seen by the consistent RMSE values in both CPU and GPU implementations. The introduction of GPU-accelerated training in the predictive modelling of health data, particularly obesity rates is a novel approach. It provides researchers with exceptional computational efficiency and maintains high predictive performance. This enables faster model construction and experimentation.

V. CONCLUSION :

In this paper, the intricate patterns of the worldwide obesity crisis have been illuminated, and the techniques used to forecast future trends in public health studies have been enhanced. A comprehensive dataset has been created by carefully collecting and organizing data from 134 nations over a period of 17 years. This dataset provides valuable insights into the connections between obesity rates and important socioeconomic factors. Strong relationships have been discovered with parameters such as healthcare spending, GDP per person, and the size of metropolitan populations. Efforts to speed up model building and solve computational hurdles using GPU-accelerated training have shown substantial decreases in training times. By employing NVIDIA's RAPIDS suite, this methodological innovation speeds up research progress and creates opportunities for further exploration of obesity and other public health concerns.

The significance of interdisciplinary collaboration and methodological innovation in tackling global health concerns has been emphasized. Through the utilization of data-driven methodologies and state-of-the-art technologies, we can make significant advancements towards a future that is both healthier and more equal for everyone.

TABLE I.

Model performance when 'Country' and 'Year' are included.

Model Name	Mean Root Mean Squared Error	Time Taken (seconds)
Linear Regression	1.14	0.26
Decision Tree	1.98	0.65
Random Forest	1.32	37.28
XGBoost	1.37	1.34
Total Time	-	39.53

TABLE II.

Model performance when all categorical features are included.

Model Name	Mean Root Mean Squared Error	Time Taken (seconds)
Linear Regression	2.00	0.34
Decision Tree	1.49	0.91
Random Forest	1.19	37.94
XGBoost	1.29	4.21
Total Time	-	43.4

TABLE III.

Model training times when CPU trained vs when GPU trained.

Model Name	CPU Time (seconds)	GPU Time (seconds)
Linear Regression	0.26	0.06
Decision Tree	0.65	0.14
Random Forest	37.28	1.87
XGBoost	1.34	0.27
Total Time	39.53	2.34

VI. REFERENCES :

[1] World Health Organization (WHO). "Obesity and overweight", who.int. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (accessed April 11, 2024).

[2] Hricik, T., Bader, D., & Green, O. (2020, September). Using RAPIDS AI to accelerate graph data science workflows. In 2020 IEEE High Performance Extreme Computing Conference (HPEC) (pp. 1-4). IEEE.

[3] Koliaki, C., Dalamaga, M., & Liatis, S. (2023). Update on the obesity epidemic: after the sudden rise, is the upward trajectory beginning to flatten?. Current Obesity Reports, 12(4), 514-527.

[4] Dogbe, W., Salazar-Ordóñez, M., & Gil, J. M. (2021). Disentangling the drivers of obesity: an analytical framework based on socioeconomic and intrapersonal factors. Frontiers in Nutrition, 8, 585318.

[5] Kirchengast, S., & Hagmann, D. (2021). " Obesity in the City"–urbanization, health risks and rising obesity rates from the viewpoint of human biology and public health. Human Biology and Public Health, 2.

[6] Okunogbe, A., Nugent, R., Spencer, G., Ralston, J., & Wilding, J. (2021). Economic impacts of overweight and obesity: current and future estimates for eight countries. BMJ global health, 6(10), e006351.

[7] Popkin, B. M., Adair, L. S., & Ng, S. W. (2012). Global nutrition transition and the pandemic of obesity in developing countries. Nutrition reviews, 70(1), 3-21.

[8] Hnatkovska, V. (2004). Volatility and growth (Vol. 3184). World Bank Publications.

[9] Malik, V. S., Willet, W. C., & Hu, F. B. (2020). Nearly a decade on—trends, risk factors and policy implications in global obesity. *Nature Reviews Endocrinology*, *16*(11), 615-616.

[10] van den Broek-Altenburg, E., Atherly, A., & Holladay, E. (2022). Changes in healthcare spending attributable to obesity and overweight: payer- and service-specific estimates. *BMC Public Health*, *22*(1), 962.

[11] World Health Organization. (208). WUP2018. WHO

[12] Swinburn, B. A., Sacks, G., Hall, K. D., McPherson, K., Finegood, D. T., Moodie, M. L., & Gortmaker, S. L. (2011). The global obesity pandemic: shaped by global drivers and local environments. *The lancet*, *378*(9793), 804-814.

[13] Marmot, M., Friel, S., Bell, R., Houweling, T. A., & Taylor, S. (2008). Closing the gap in a generation: health equity through action on the social determinants of health. *The lancet*, *372*(9650), 1661-1669.

[14] Hawkes, C., Smith, T. G., Jewell, J., Wardle, J., Hammond, R. A., Friel, S., ... & Kain, J. (2015). Smart food policies for obesity prevention. *The lancet*, *385*(9985), 2410-2421.

[15] Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123-140.

[16] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145)..