Coursework Coversheet

School of Geography

FACULTY OF ENVIRONMENT

UNIVERSITY OF LEEDS

Student ID	201788996
Module code	GEOG5404M
Module title	Analytics for Urban Policy
Assignment title	Analysing Childhood Obesity Rates Across London Boroughs Using Machine Learning Technologies
Marker	
Declared word count	1991

To improve your work from next time
1.
2.
3.

Justification of mark (using specific text from criteria)

Additional Comments

Analysing Childhood Obesity Rates Across London Boroughs Using Machine Learning Technologies

1. Introduction

The development and construction of healthy cities are important to improving global public health. Governments globally are implementing some policies to promote a better urban living environment. For instance, the United Kingdom (UK) government has implemented a sugar tax on sugary drinks to reduce sugar intake among children. The governments of the United States and Australia encourage schools to increase the number of physical education classes and promote after-school physical activities to ensure children get enough exercise. The Singaporean government supports community centres in organising healthy eating workshops and parent-child physical activities. However, Year 6 students in certain boroughs of London have childhood obesity rates that exceed the national average (Trust for London, 2023). This raises public concern and to address these challenges, the UK government is implementing some measures such as restricting fast-food restaurants near primary schools. It is essential to address physical and mental health during adolescence to reduce the lifelong risk of obesity (Triantafyllidis et al., 2020), which is linked to long-term health outcomes and affects sustainable urban development.

2. Literature Review

Recent advancements in applications of machine learning technologies have influenced health sector analytics. For example, Singh and Tawfik (2020) utilized different kinds of data, such as weight and Index of Multiple Deprivation (IMD), to predict the risk of obesity in adolescents by evaluating several machine learning (ML) algorithms; Zheng and Ruggiero (2017) created four enhanced ML models based on health-related behaviour data to predict obesity in high school students. Although the studies perform well and address data imbalance issues, they ignore other potential influencing factors such as income levels and urban environment, which could significantly impact the findings. Furthermore, Dunstan et al. (2020) created various nonlinear regression algorithms using food sales information to predict obesity rates. This research offers us a valuable understanding of predicting obesity rates at the national level, but the findings could not be generalized to other regions due to the exclusion of all income levels.

Zare et al. (2021) predicted obesity rates in fourth-grade students by calculating unhealthy food retailer environments, income, education levels and so on. Cheng et al. (2021) were the first to use physical activity (PA) data and ML techniques to find that PA is an important risk factor

for predicting obesity. These studies provided comprehensive insights into predictive features for further work, but the limitation is that these studies mainly focus on the obesity rates of specific children instead of total obesity rates in different areas of cities. Also, Lim et al. (2023) and Hammond et al. (2019) developed obesity prediction models based on different kinds of data, such as dietary habits and daily activities, but recent studies primarily concentrate on forecasting if there are any obesity issues instead of obesity rates. Additionally, Yuan and Hu (2023) utilised clustering algorithms to categorize different types of cities and studied patterns of urban disruption and adaptive resilience to COVID-19. Hager et al. (2021) conducted some clustering analyses on neighbourhoods to identify the most suitable stormwater management strategies. While clustering algorithms are widely used in each domain of urban planning, their application within health-focused urban studies remains underexplored.

Hence, based on these research gaps, further research can conduct different policies to address health challenges regarding specific neighbourhoods by creating cluster algorithms. In the meantime, to further explore the correlation in the dynamic urban environment, we need to collect more potential influencing factors related to obesity rates and focus on total rates in different areas of the same city.

3. Research Questions and Methodology

Based on the research gap and introduction, we can propose the following research questions (RQ):

- How do urban environment and socio-economic factors affect the level of physical activity participation and childhood obesity rates across different boroughs of London?
- Which machine learning model performs better in effectively predicting childhood obesity rates in this research?
- What substantial suggestions can we provide to lower childhood obesity rates?

Led by the above RQ, we need to determine what data we need. Research indicates that higher community exposure to fast food outlets is positively correlated with a higher risk of obesity (Burgoine et al., 2021), so we will collect information on fast food outlets and sports facilities, such as parks and gyms, across London due to their impact on children's lifestyle habits. Individuals in high-income groups in developed countries are less likely to be overweight (Oguoma et al., 2021), so low-income groups often lack access to healthy food options and median incomes across London boroughs will be collected. High crime rates and air pollution may deter children from engaging in outdoor activities, indirectly leading to higher obesity

rates, air pollution and crime rates also will be collected to analyse their relationship with obesity. Also, we will collect data on the proportion of physical activity participation to explore its impact on obesity rates. Specific data sources are referenced in Table 1.

Table 1: The used datasets for this research

Descriptions	Resources	Year
Childhood obesity rate in each borough of London	https://trustforlondon.org.uk/data/child-obesity/	2023
Locations of fast-food shops and sports facilities in each borough of London	OpenStreetMap API	2024
Area and population in each borough of London	https://data.london.gov.uk/dataset/land-area-and-population-density-ward-and-borough	2023
Median Income of each borough in London	https://data.london.gov.uk/dataset/earnings-workplace-borough	2023
The number of crimes in each borough of London	https://data.london.gov.uk/dataset/recorded_crime_summary	2023
The percentage of physically active children in each borough of London	https://fingertips.phe.org.uk/profile/physical-activity/data#page/1	2023
The percentage of population exposure in each borough of London (PM2.5 and NO2)	https://data.london.gov.uk/dataset/london-atmospheric-emissions-inventory--laei--2019	2019

Before conducting analysis, we should finish the stage of data wrangling and cleaning including dealing with missing values and converting formats – see the notebook. Also, we should ensure data alignment during combining multiple datasets into a single data frame.

After this, we first need to develop features related to the urban environment, such as fast-food restaurant density, physical activity facilities density, crime rate, and the proportion of physical activity participation across each borough of London. For RQ1, we can utilize exploratory data analysis and visualization techniques including heat maps to explore it, focusing on correlations between different features and childhood obesity rates.

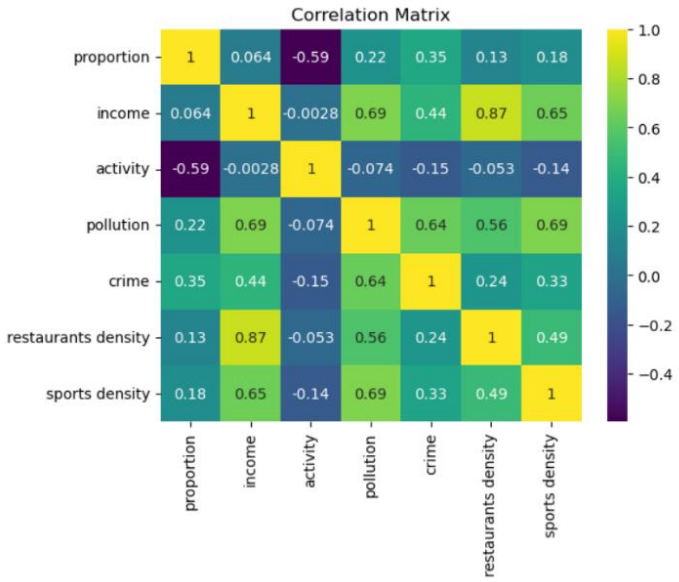
To answer RQ2, we will create some models including the Random Forest model, the Extreme Gradient Boosting model, and the Gradient Boosting model to predict childhood obesity rates. The selection of these models was based on their robustness in handling nonlinear relationships and previous studies. Initially, the dataset will be divided into training and test sets according to the 80/20 rule as the stage of validation. For model comparison, we will evaluate the metrics including the coefficient of determination and Mean Squared Error (MSE). Following this, we will optimize the models by manually tuning the parameters and utilizing grid search to identify the optimal parameters for each predictive model. The models' performance will be re-evaluated using the same metrics to ensure consistent assessment criteria.

To streamline the urban planning process for boroughs with similar characteristics, this research will utilise two clustering models - the K-means model and the hierarchical clustering model - to categorise the different types of boroughs in London. The performances of both clustering models will be assessed using silhouette scores. For RQ3, the model with better performance will be selected to further explore the characteristics of each cluster through visualizations such as radar charts. Finally, we will discuss policy recommendations based on our findings to answer RQ3.

4. Findings and Discussions

The urban environment directly or indirectly influences children's participation in physical activity participation and childhood obesity rates (see the notebook for other findings). Figure 1 shows the correlation between different factors related to the urban environment and childhood obesity rate.

Figure 1: correlation matrix of the study

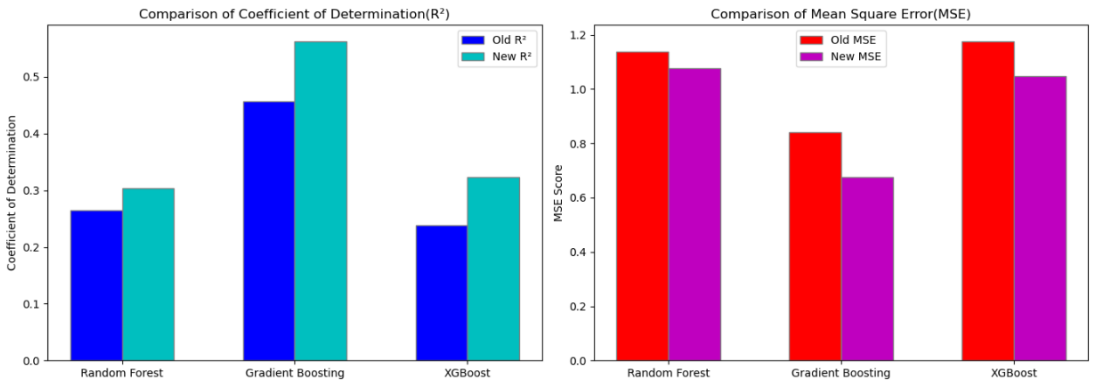


The childhood obesity rate shows a strong negative correlation (-0.59) with the percentage of physical activity participation, showing that boroughs with higher percentages of physical activity participation have lower childhood obesity rates. The positive correlation (0.35) with crime rate indicates that in boroughs with higher crime rates, childhood obesity rates also tend to be higher. These results are similar to previous studies in that crime rates are significantly related to insufficient physical activity and obesity in adults (Singleton et al., 2023), and physical activity are significant health risk factor (Tainio et al., 2021). The correlation between childhood obesity rate and sports facilities density is weakly positive (0.18), these findings could not support the study that people living in areas with low availability of sports facilities are more prone to obesity (Cereijo et al., 2022).

Hence, an unsafe environment may limit children's physical activity c and thus reduce the amount of physical activity and increasing physical activity can reduce obesity issues. Children's physical activity participation in London may not directly negatively correlate with income levels and sports facilities density. The correlation between children's physical activity participation and both income levels and sports facility density appears not directly negative, which may be because the correlation between obesity rates and socioeconomic indices is complex (Oguoma et al., 2021). Furthermore, the plot shows that poor air quality in some boroughs correlates with higher childhood obesity rates, likely discouraging physical activity participation. Conversely, better air quality supports higher percentages of physical activity participation among children, potentially lowering obesity rates.

We compared the performance of different models in predicting child obesity rates from Figure 2, by assessing the coefficient of determination and mean squared error (MSE) of these models.

Figure 2: the performances of each model



It can be found that the Gradient Boosting model has the best performance in both the coefficient of determination and MSE, although this model still has lots of improved spaces.

Although the performances of both the Random Forest model and the XGBoost model have improved after optimizing models, scores of the coefficient of determination still belove 0.5. Therefore, in this research question, it can be found that the Gradient Boosting model is the best fit among the three models to predict childhood obesity rates across London boroughs.

Through clustering analysis, both models have divided the boroughs into three clusters, and the performances of the models are the same. Therefore, we consider providing some suggestions for reducing childhood obesity rates based on calculating the average value of each feature in each cluster. The differences in the average values of various features across the three cluster groups can be observed in Figure 3.

Figure 3: the performances of each model

Some boroughs belonging to Cluster 0, such as Lambeth and Southwark, have high childhood rates, lower sports facilities density and the percentage of physical activity participation. Governments could construct additional indoor sports facilities with security monitoring in these boroughs due to high crime rates. Additionally, they should implement effective policing strategies and environmental protection measures to reduce crime rates and improve air quality, thereby improving children's participation in physical activity and ultimately lowering childhood obesity rates.

For Cluster 1, such as Barking and Dagenham, these boroughs show lower childhood obesity rates and higher percentages of physical activity participation. Governments may conduct some education workshops related to healthy diets due to lower socioeconomic conditions, and help low-income families get access to healthier foods and sports facilities.

The City of London is the only borough belonging to Cluster 2. Although sports facilities density is higher and crime is lower, the percentage of physical activity participation is lower and childhood obesity rates are higher, this may be due to higher pollution influencing the children's motivation for outdoor activities. Governments can conduct environmental policies to improve air quality and make outdoor environments more suitable for activities. For example, enhancing green spaces and blue spaces near the neighbourhood and limiting the increasing number of fast-food restaurants.

Although this study fills in some research gaps, some limitations still need to be further explored. First, while various variables were considered, there may still be unobserved factors potentially influencing child obesity rates, such as demographic characteristics and cultural backgrounds. Future studies should incorporate more data to provide a more comprehensive understanding of the factors related to obesity rates. Secondly, there might be some inherent biases due to the limited available dataset and temporal alignment. We only accessed socioeconomic data in 2023, with the facility information related to this in 2024. If new facilities were opened in 2024, it might impact the children's behaviours, but this impact would not be reflected in the data from 2023. Furthermore, since we collect obesity rate data from Year 6 students, the models may not fully represent all age groups. We should consider gathering current-year data on obesity rates across different ages and ensure the same temporal for all datasets.

References

Burgoine, T. *et al.* (2021) "Independent and combined associations between fast-food outlet exposure and genetic risk for obesity: a population-based, cross-sectional study in the UK," *BMC medicine*, 19(1), p. 49. doi: 10.1186/s12916-021-01902-z.

Cereijo, L. *et al.* (2022) "Exercise facilities and the prevalence of obesity and type 2 diabetes in the city of Madrid," *Diabetologia*, 65(1), pp. 150–158. doi: 10.1007/s00125-021-05582-5.

Cheng, X. *et al.* (2021) "Does physical activity predict obesity-A machine learning and statistical method-based analysis," *International journal of environmental research and public health*, 18(8), p. 3966. doi: 10.3390/ijerph18083966.

Childhood obesity, London health issues (no date) *Trust for London*. Available at: https://trustforlondon.org.uk/data/child-obesity/ (Accessed: May 14, 2024).

Dunstan, J. *et al.* (2020) "Predicting nationwide obesity from food sales using machine learning," *Health informatics journal*, 26(1), pp. 652–663. doi: 10.1177/1460458219845959.

Hager, J. K. *et al.* (2021) "Integrated planning framework for urban stormwater management: one water approach," *Sustainable and resilient infrastructure*, pp. 1–22. doi: 10.1080/23789689.2020.1871542.

Hammond, R. *et al.* (2019) "Predicting childhood obesity using electronic health records and publicly available data," *PloS one*, 14(4), p. e0215571. doi: 10.1371/journal.pone.0215571.

Lim, H., Lee, H. and Kim, J. (2023) "A prediction model for childhood obesity risk using the machine learning method: a panel study on Korean children," *Scientific reports*, 13(1), p. 10122. doi: 10.1038/s41598-023-37171-4.

Oguoma, V. M. *et al.* (2021) "Prevalence of overweight and obesity, and associations with socio-demographic factors in Kuwait," *BMC public health*, 21(1), p. 667. doi: 10.1186/s12889-021-10692-1.

Pang, X. *et al.* (2021) "Prediction of early childhood obesity with machine learning and electronic health record data," *International journal of medical informatics*, 150(104454), p. 104454. doi: 10.1016/j.ijmedinf.2021.104454.

Singh, B. and Tawfik, H. (2020) "Machine learning approach for the early prediction of the risk of overweight and obesity in young people," in *Lecture Notes in Computer Science*. Cham: Springer International Publishing, pp. 523–535.

Singleton, C. R. *et al.* (2023) "Violent crime, physical inactivity, and obesity: Examining spatial relationships by racial/ethnic composition of community residents," *Journal of urban health: bulletin of the New York Academy of Medicine*, 100(2), pp. 279–289. doi: 10.1007/s11524-023-00716-z.

Tainio, M. *et al.* (2021) "Air pollution, physical activity and health: A mapping review of the evidence," *Environment international*, 147(105954), p. 105954. doi: 10.1016/j.envint.2020.105954.

Triantafyllidis, A. *et al.* (2020) "Computerized decision support and machine learning applications for the prevention and treatment of childhood obesity: A systematic review of the literature," *Artificial intelligence in medicine*, 104(101844), p. 101844. doi: 10.1016/j.artmed.2020.101844.

Yuan, Z. and Hu, W. (2023) "Urban resilience to socioeconomic disruptions during the COVID-19 pandemic: Evidence from China," *International journal of disaster risk reduction: IJDRR*, 91(103670), p. 103670. doi: 10.1016/j.ijdrr.2023.103670.

Zare, S. *et al.* (2021) "Use of machine learning to determine the information value of a BMI screening program," *American journal of preventive medicine*, 60(3), pp. 425–433. doi: 10.1016/j.amepre.2020.10.016.

Zheng, Z. and Ruggiero, K. (2017) "Using machine learning to predict obesity in high school students," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.