

Supporting Information for

Early Obesity Risk Prediction via Non-dietary Lifestyle Factors using Machine Learning Approaches

Ker Ming Seaw,¹ Melvin Khee Shing Leow,^{1,2,3,4,5} Xinyan Bi,^{1}*

¹ Clinical Nutrition Research Centre (CNRC)

Singapore Institute of Food and Biotechnology Innovation (SIFBI)

Agency for Science, Technology and Research (A*STAR)

14 Medical Drive #07-02, MD 6 Building Yong Loo Lin School of Medicine, Singapore
117599, Singapore

² Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

³ Department of Endocrinology, Tan Tock Seng Hospital, Singapore

⁴ Human Potential Translational Research Programme (HPTRP), Yong Loo Lin School of Medicine, National University of Singapore (NUS)

⁵ Cardiovascular and Metabolic Program, Duke-NUS Medical School, Singapore

*Correspondence: bi_xinyan@sifbi.a-star.edu.sg

Table S1. Description of all the attributes in the data source collected from survey

Attribute Name	Type	Description	Data Categories
Id	Object	Serial number	1 to 2111
Gender	Object	Sexual identity	Male, female
Age	Int64	How young or old	Age in year
Height	Float64	How tall or short	Height in metres
Weight	Int64	Body mass	Weight in kilogram
Family history with overweight	Object	Presence or absence of overweight in family	Yes, no
FAVC	Object	Frequently consume high calorie food	Yes, no
FCVC	Object	Frequently consume vegetables	Never, sometimes, always
NCP	Int64	No of main meals (daily)	1, 2, 3, 4
CAEC	Object	Consume food between meals	No, sometimes, frequently, always
SMOKE	Object	Smoking status	Yes, no
CH2O	Object	Consume water amount (daily)	<1 L, between 1 and 2 L, > 2 L
SCC	Object	Calories consume monitoring	Yes, no
FAF	Object	Exercise frequency (days per week)	0, 1 to 2, 2 to 4, 4 to 5
TUE	Object	Time using technology devices (hours daily)	0 to 2, 3 to 5, > 5
CALC	Object	Alcohol consumption	No, sometimes, frequently, always
MTRANS	Object	Transportation	Automobile, motorbike, bike, public transportation, walking
NObeyesdad	Object	Obesity category	Insufficient, normal, overweight- I, overweight- II, obesity- I, obesity- II, obesity- III
BMI	Float	Body mass index	13.138 to 50.780

Note: Table S1 displays the information of the data collected in the original data source, including the column name, the data type as perceived in Python, the data description and the data value or categories in each column

Abbreviations: FAVC, Do you eat high caloric food frequently?; FCVC, Do you usually eat vegetables in your meals?; NCP, How many main meals do you have daily?; CAEC, Do you eat any food between meals?; CH2O, How much water do you drink daily?; SCC, Do you monitor the calories you eat daily?; FAF, How often do you have physical activity?; TUE, How much time do you use technological devices such as cell phone, videogames, television, computer and others?; CALC, How often do you drink alcohol?; MTRANS, Which transportation do you usually use?; NObeyesdad, Obesity Level; BMI, body mass index

Table S2. Final attributes used in this study to build the predictive models.

Data type	Final attributes	Perceived type of data
Input	Gender	Demographic
	Age	Demographic
	Height	Demographic
	Family history with overweight	Genetic
	Exercise frequency	Lifestyle
	Time using technology devices	Lifestyle
	Main transportation used	Lifestyle/Demographic
Output (target variable)	Obesity category	NIL

Note: Input reflected in table are features that were selected as the final attributes by the feature selection process

to build the predictive model. The output (target variable) of obesity category is the same as NObeyesdad, comprising from underweight to obese type III in their data value.

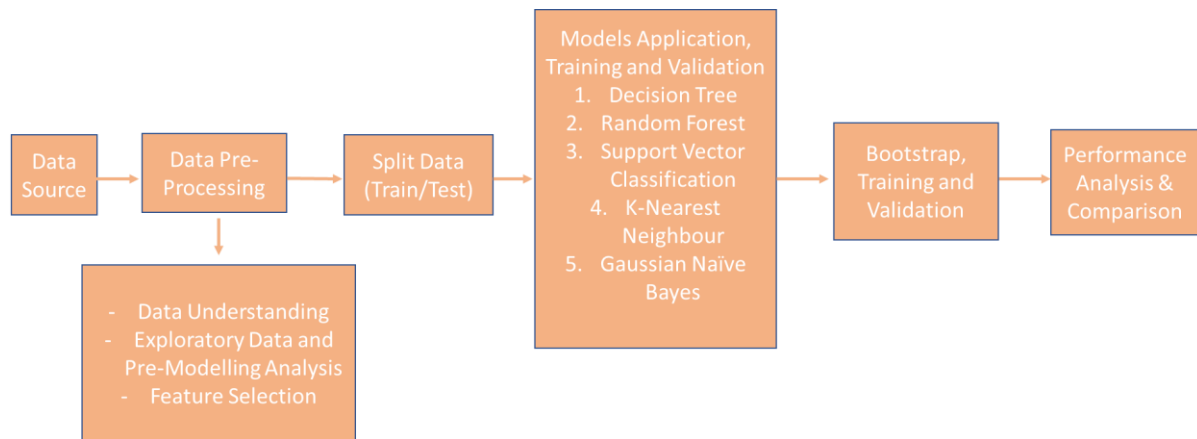


Figure S1. Workflow of the study.

Note: Figure S1 reflects the high-level overview of the study workflow and methodology

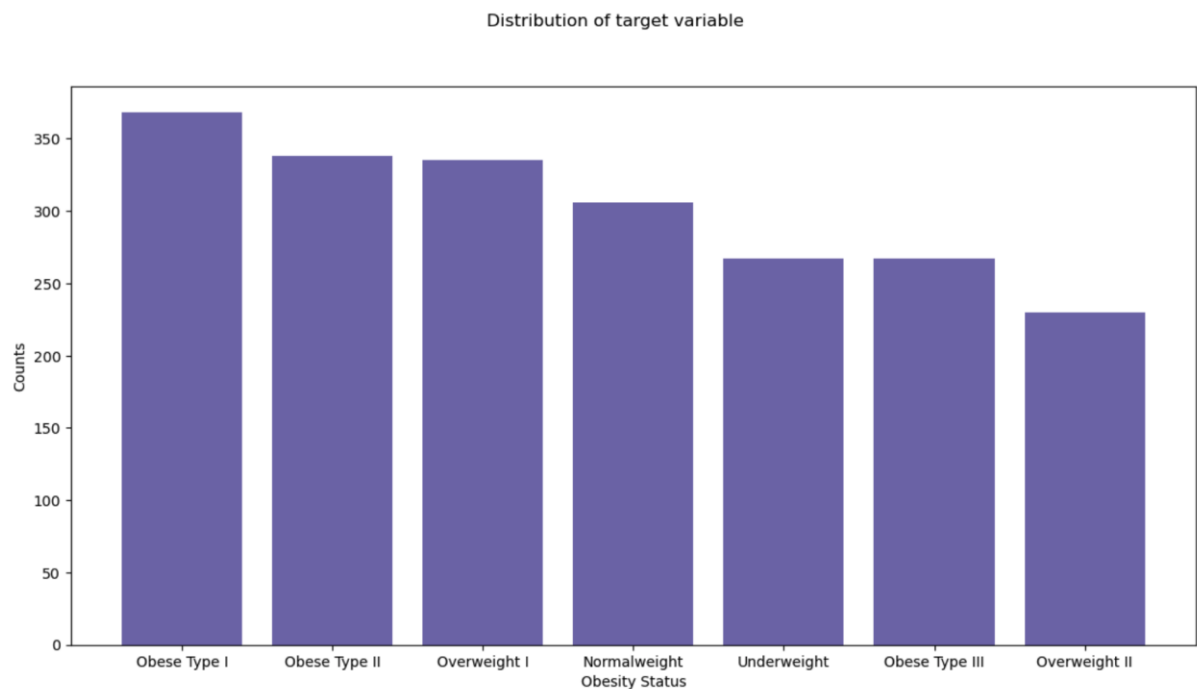


Figure S2. Counts of the different weight categories

Note: Bar graph displaying the distribution in numbers of different weight categories after data pre-processing as the target variable for prediction