Mini Project 2 - Estimation of obesity levels based on eating habits and physical condition

- This dataset includes data for the estimation of obesity levels in individuals from the countries of **Mexico, Peru and Colombia**, based on their eating habits and physical condition.
- The data contains **17 attributes** and **2111 records**
- The records are labelled with the class variable 'Nobesity' (Obesity Level), that allows classification of the data using the values of **Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.**
- 77% of the data was generated synthetically using the Weka tool and the SMOTE filter
- 23% of the data was collected directly from users through a web platform.

Mini Project 2 - Estimation of obesity levels based on eating habits and physical condition

The attributes related with eating habits are:

- Frequent consumption of high caloric food (FAVC)
- Frequency of consumption of vegetables (FCVC)
- Number of main meals (NCP)
- Consumption of food between meals (CAEC)
- Consumption of water daily (CH20)
- Consumption of alcohol (CALC)

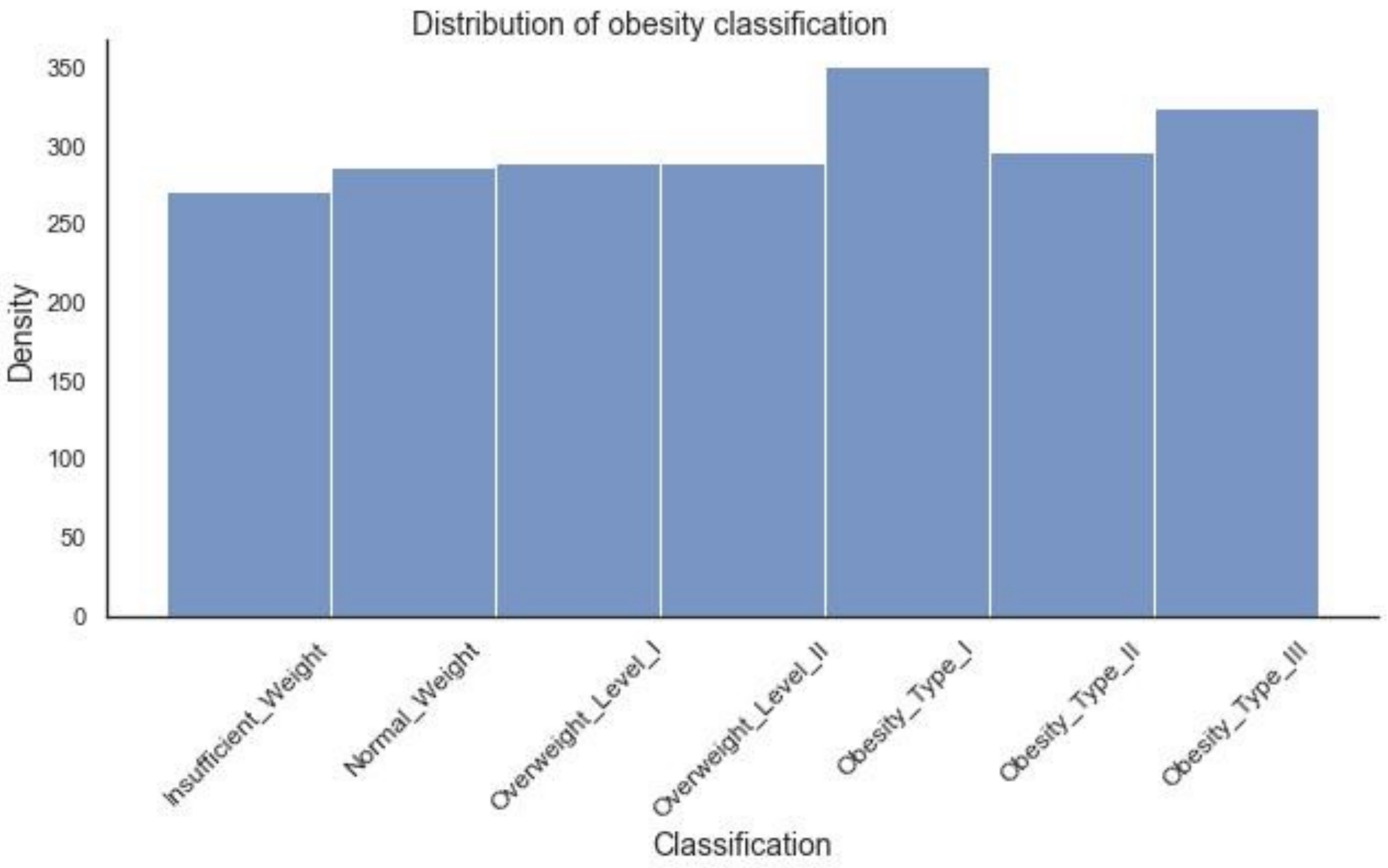
The attributes related with the physical condition are:

- Calories consumption monitoring (SCC)
- Physical activity frequency (FAF)
- Time using technology devices (TUE)
- Transportation used (MTRANS)

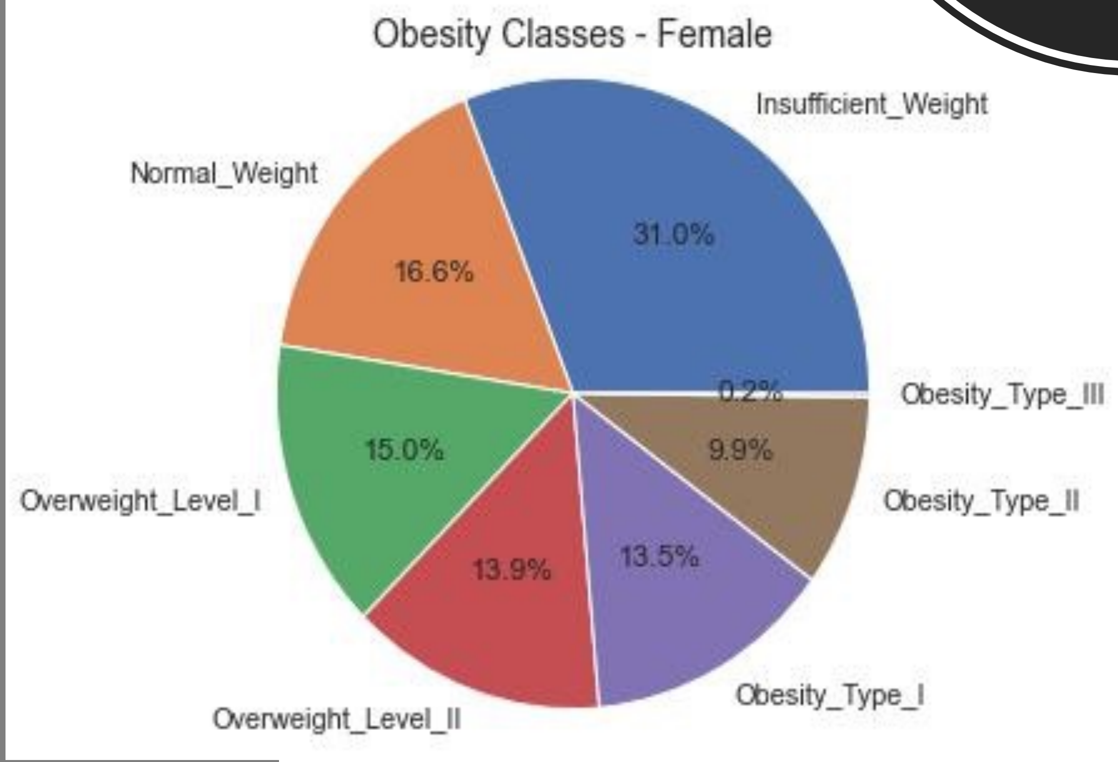
Other variables obtained were:

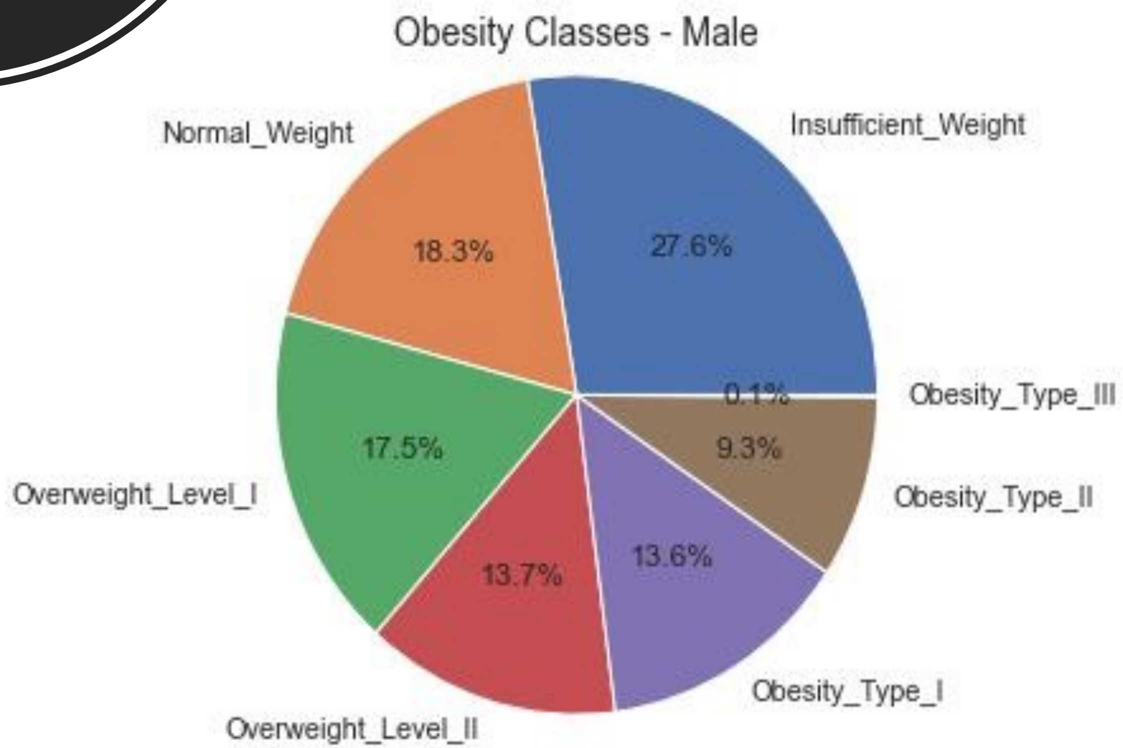
- Gender
- Age
- Height
- Weight

There are seven ‘obesity level’ classes, with roughly equal frequencies

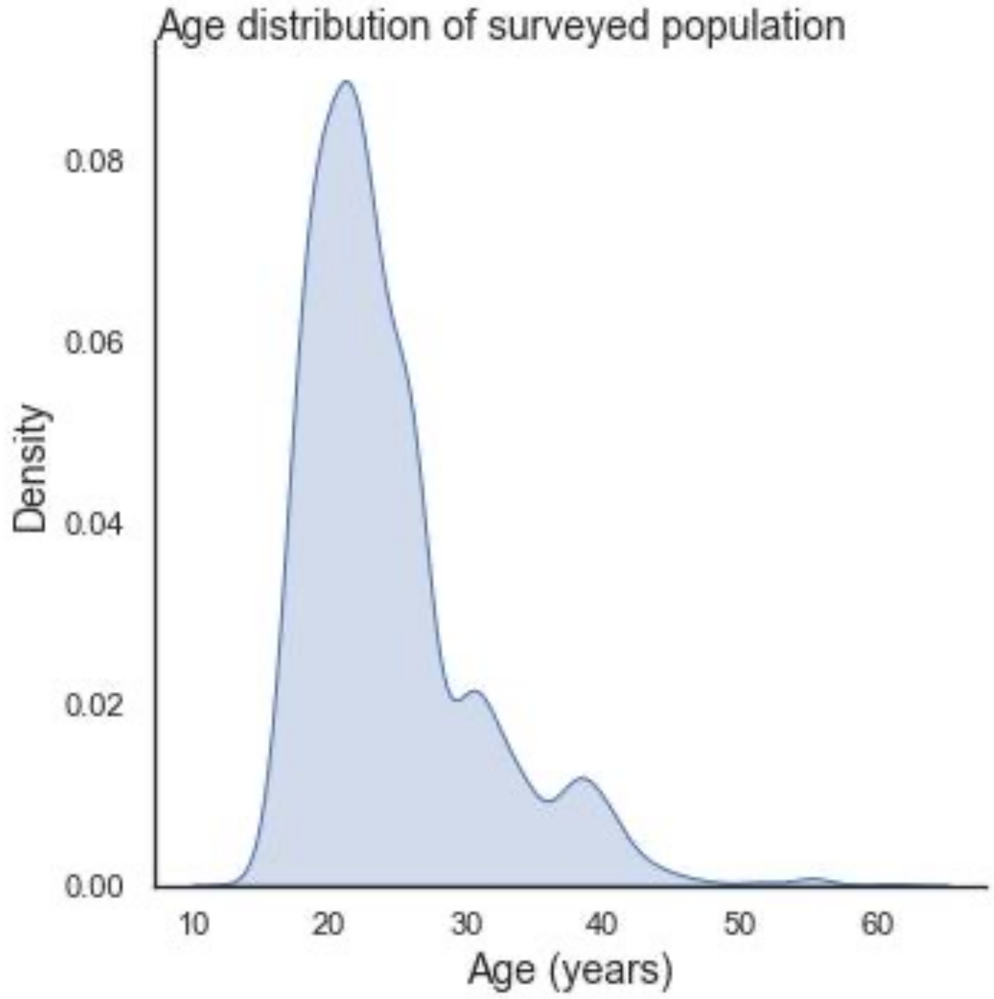


There is no difference in obesity levels between the sexes

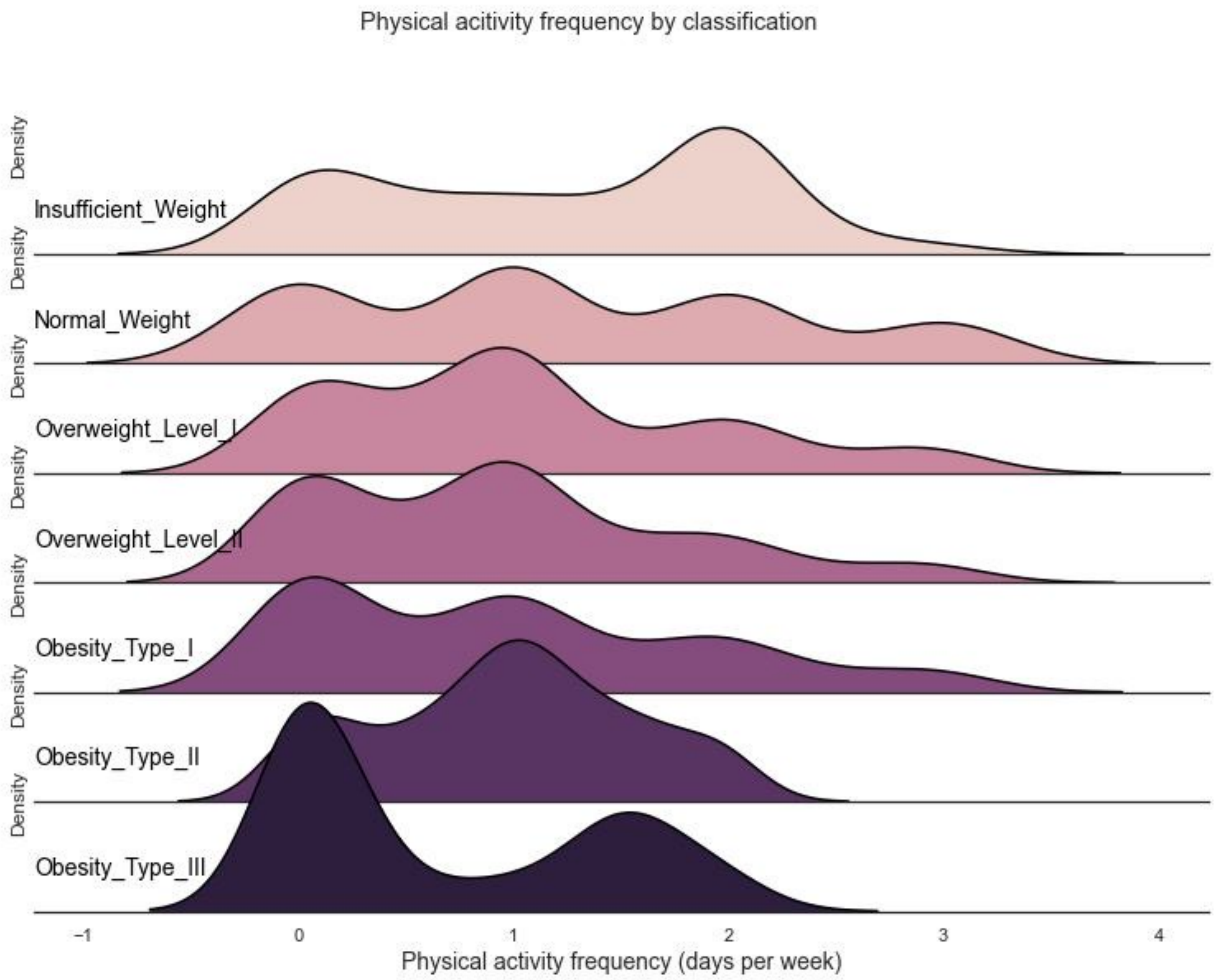




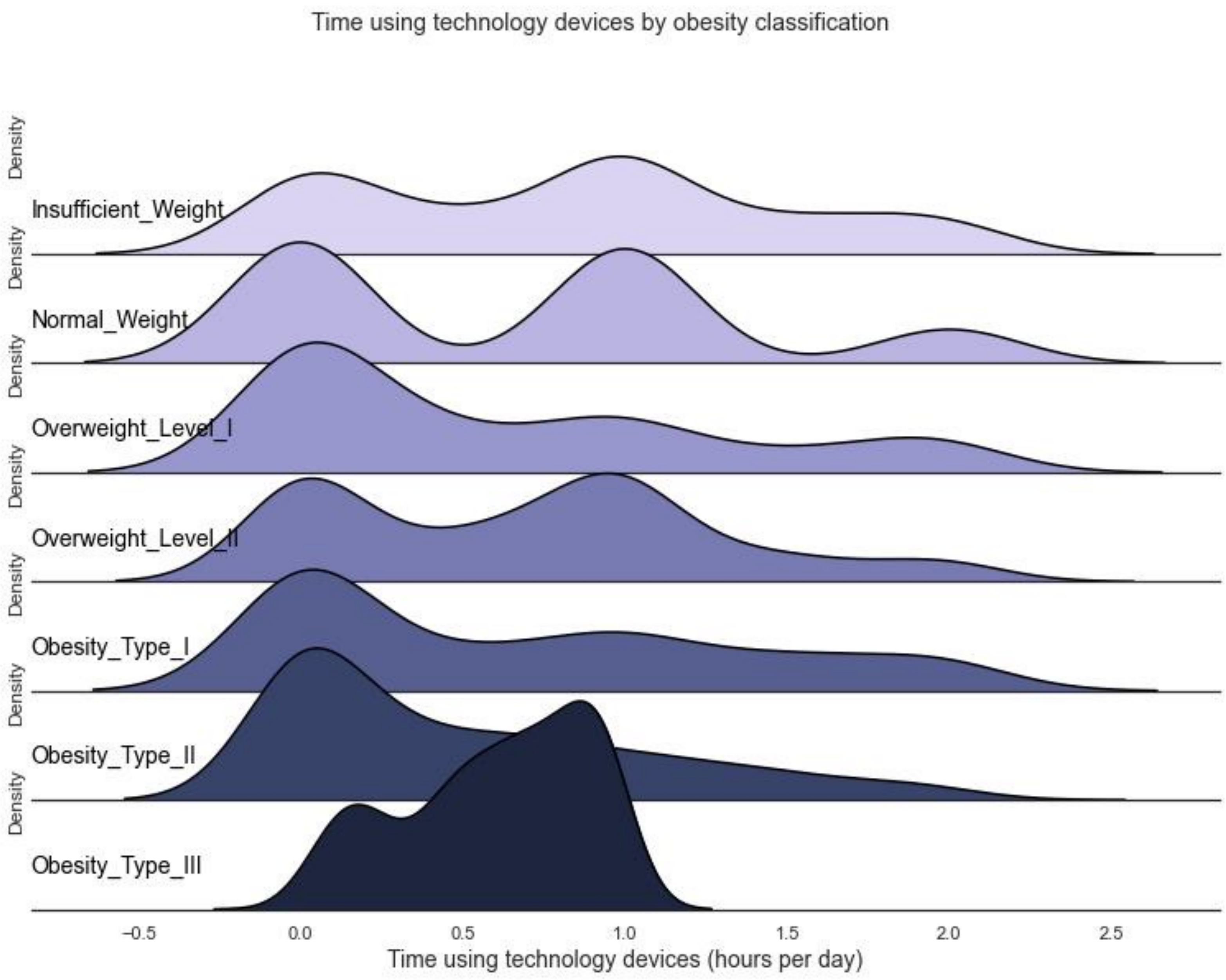
Age distribution of the surveyed population is narrow, with most aged 18-30



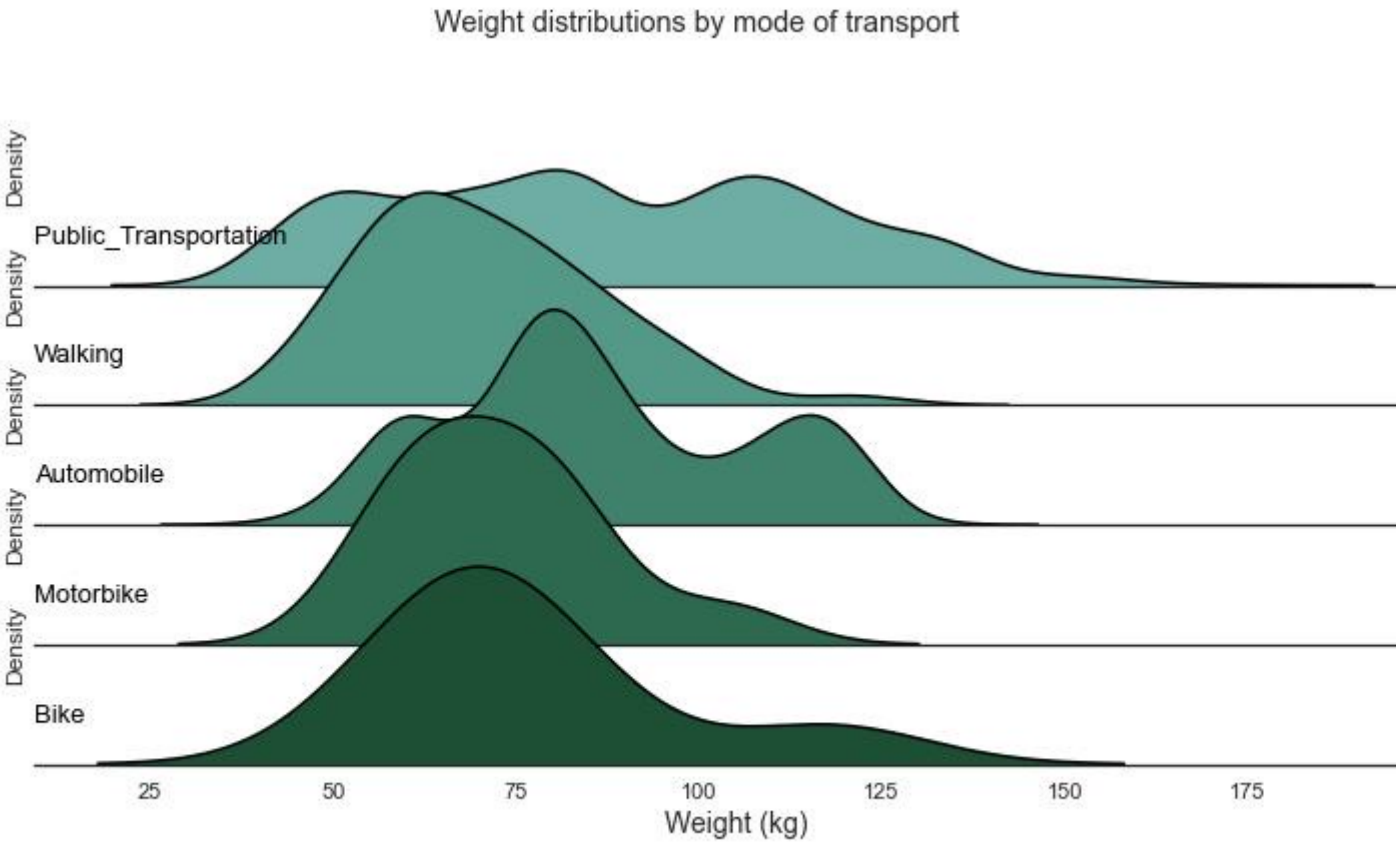
Highest levels of obesity tend to be associated with lower levels of physical activity



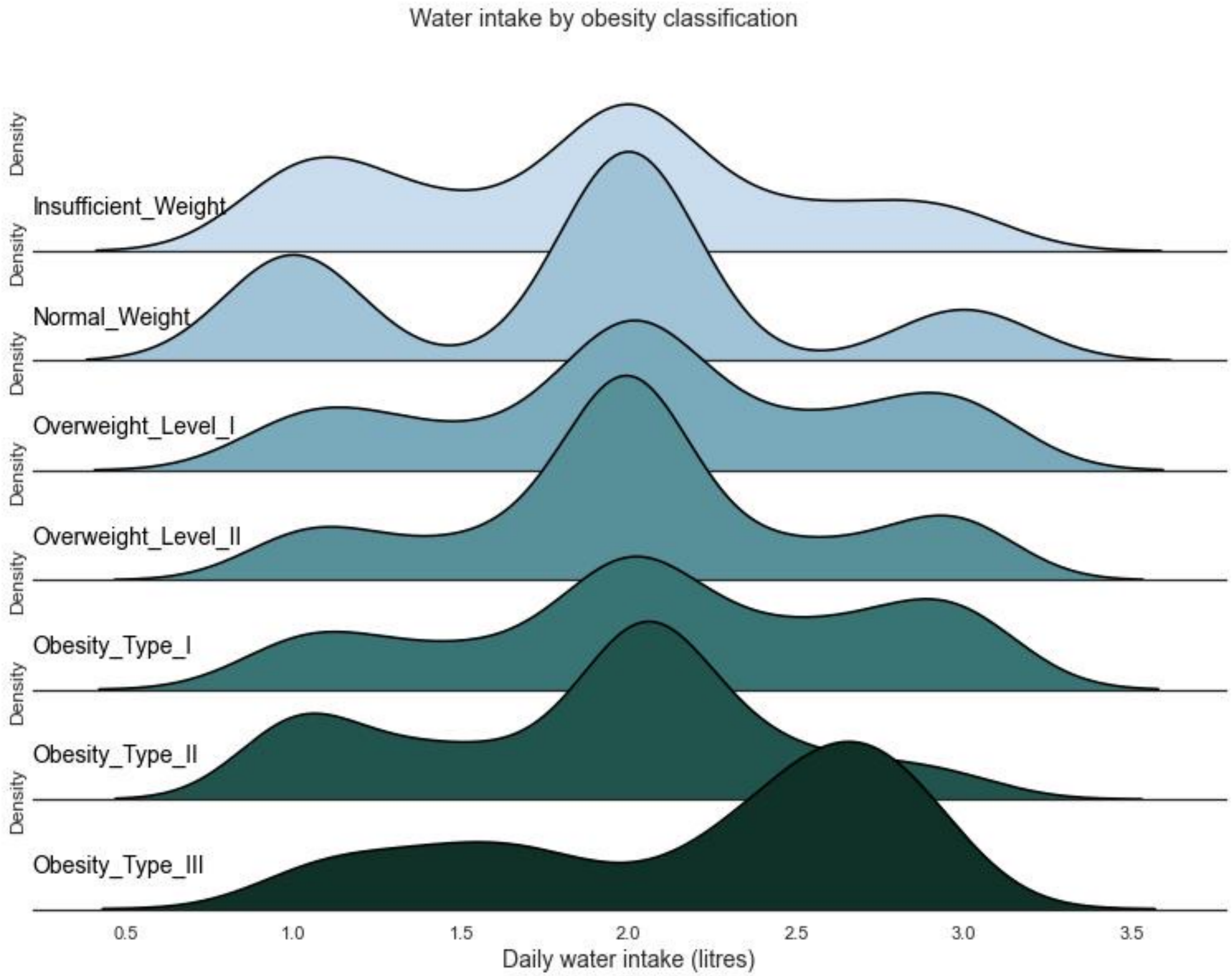
But there is no obvious association between level of obesity and time spent using 'devices'



Greater weight tends to be associated with travel by car or public transport, but this is not likely to be a predictive feature



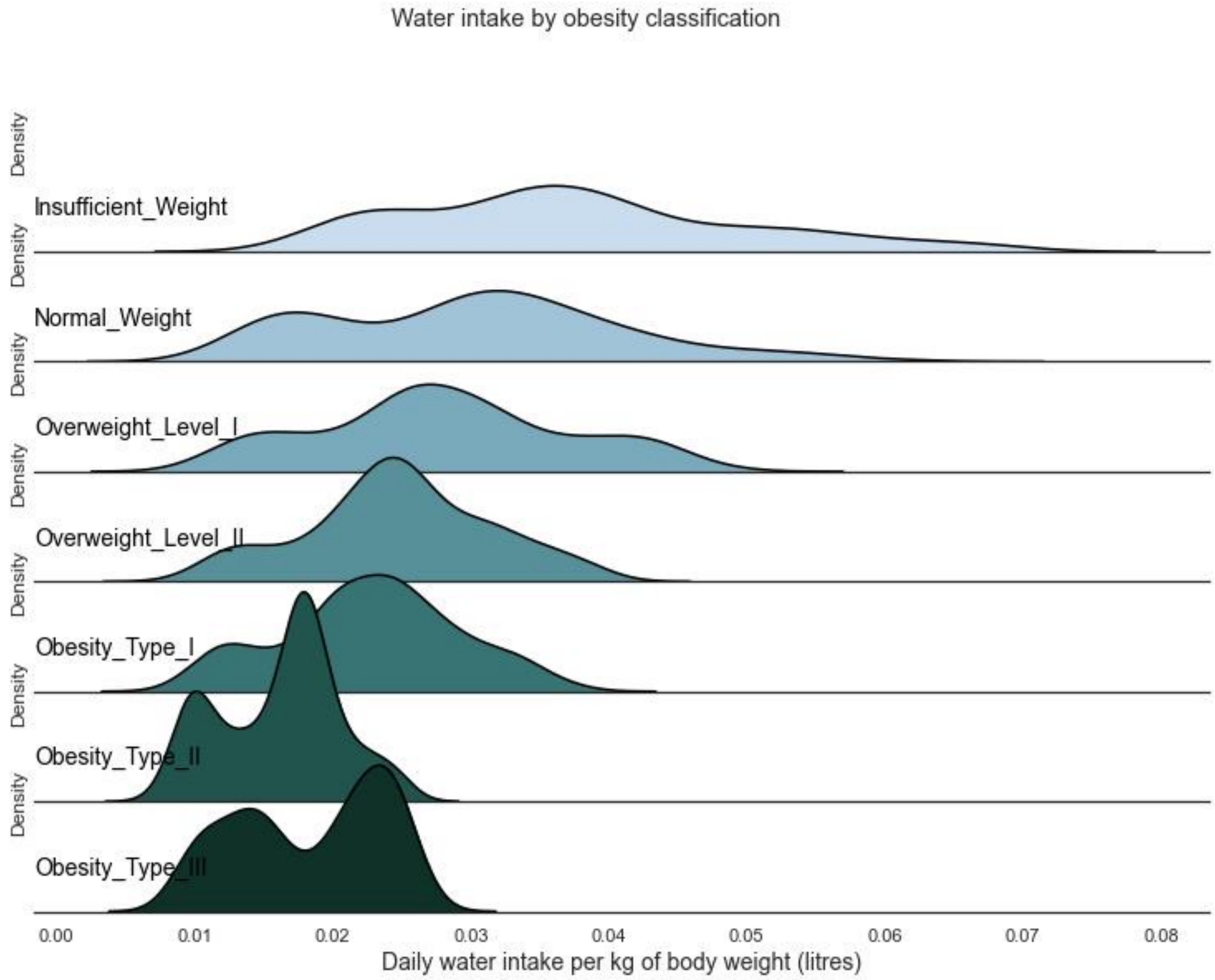
High levels of obesity appeared to be associated with higher daily water intake…



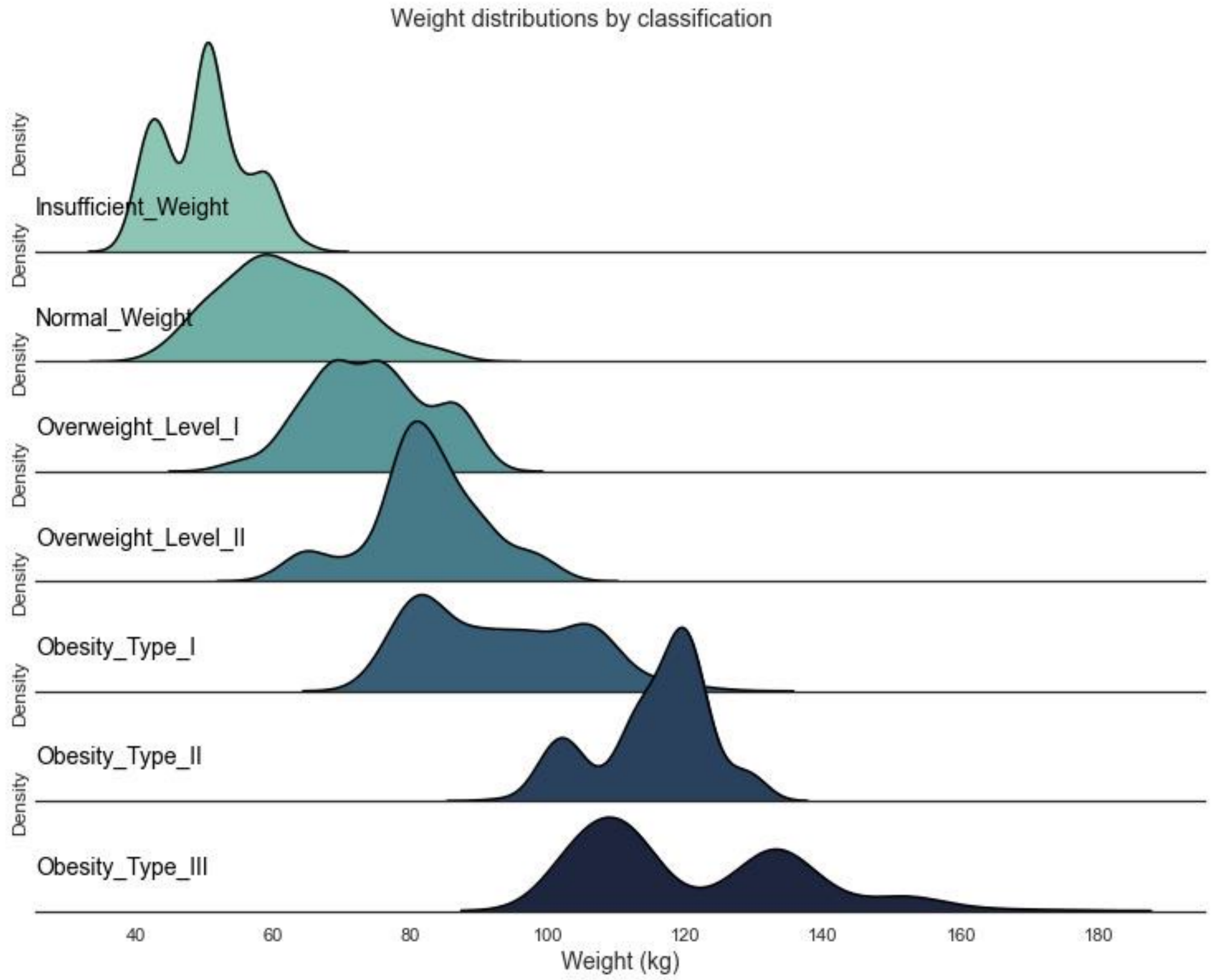
... but this was not the case when corrected for body weight

Feature Creation – 'CH2O_adj'

= CH2O (l) / weight (kg)



Mean weight increases with 'body type' classification, but there is significant overlap between the weight distributions for each class



Feature Creation - BMI

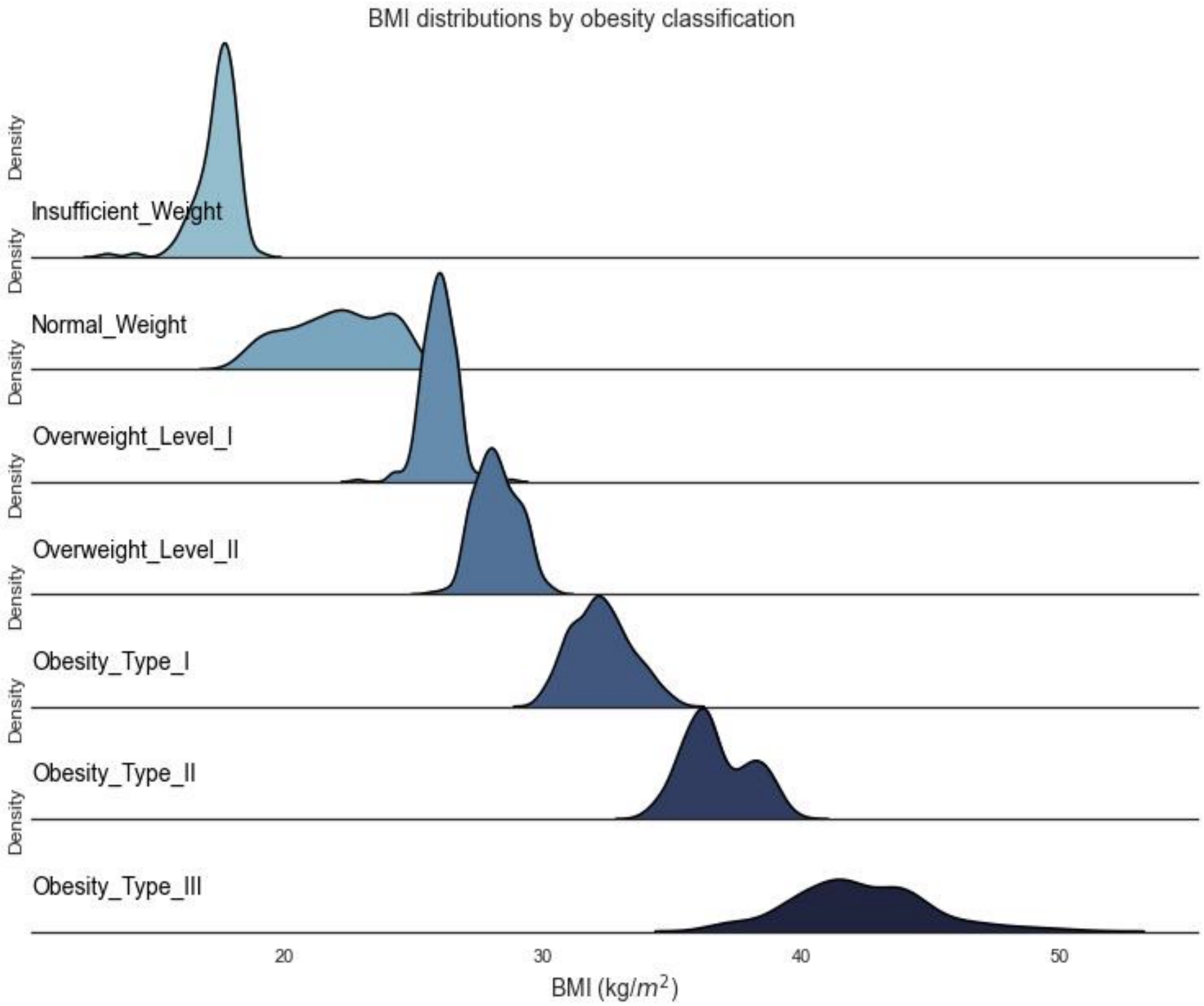
BMI = weight (kg) / (height (m))2

BMI increases with ‘body type’ classification, with much less overlap between BMI

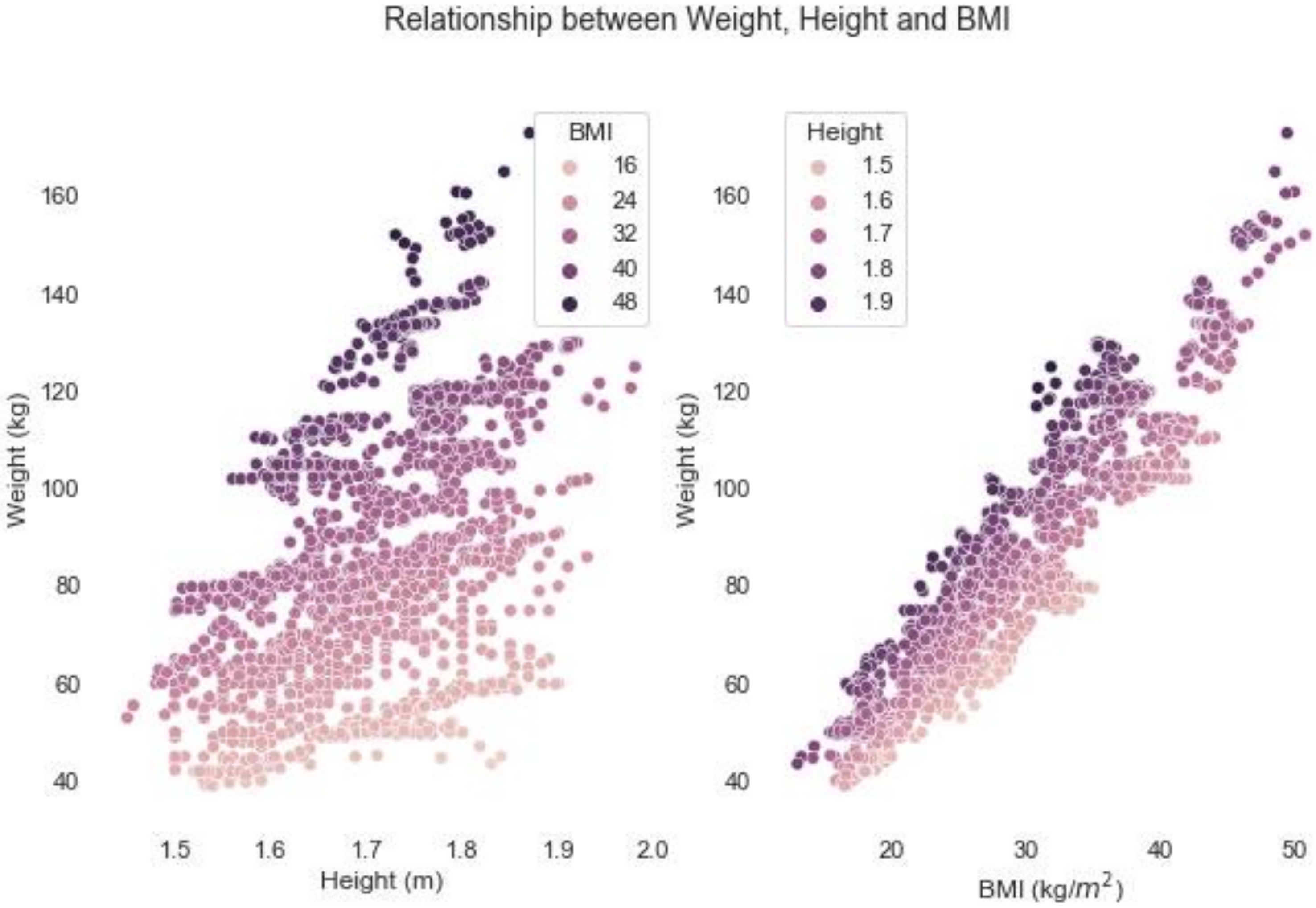
distributions for each class

Feature Creation - BMI

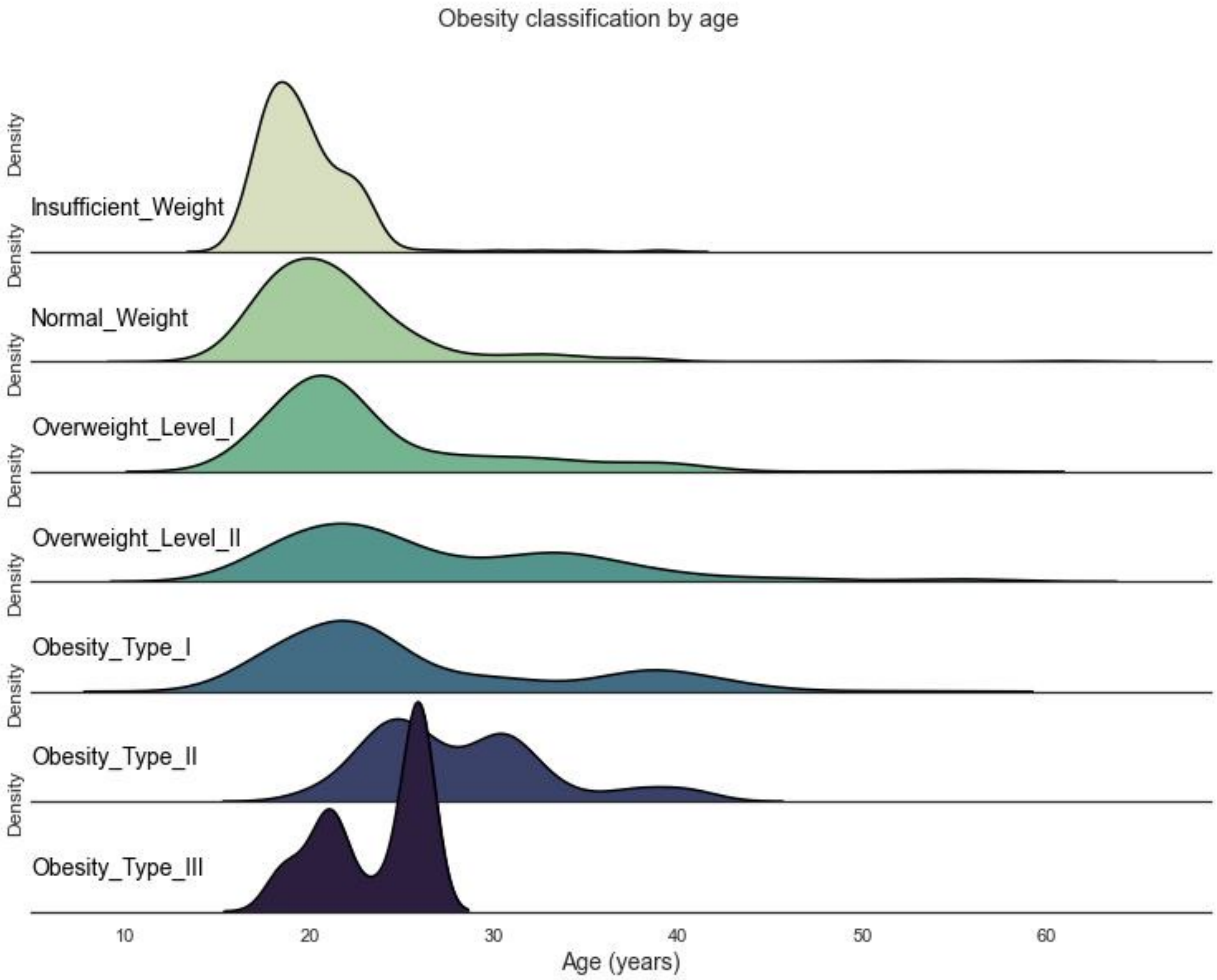
BMI = weight (kg) / (height (m))2



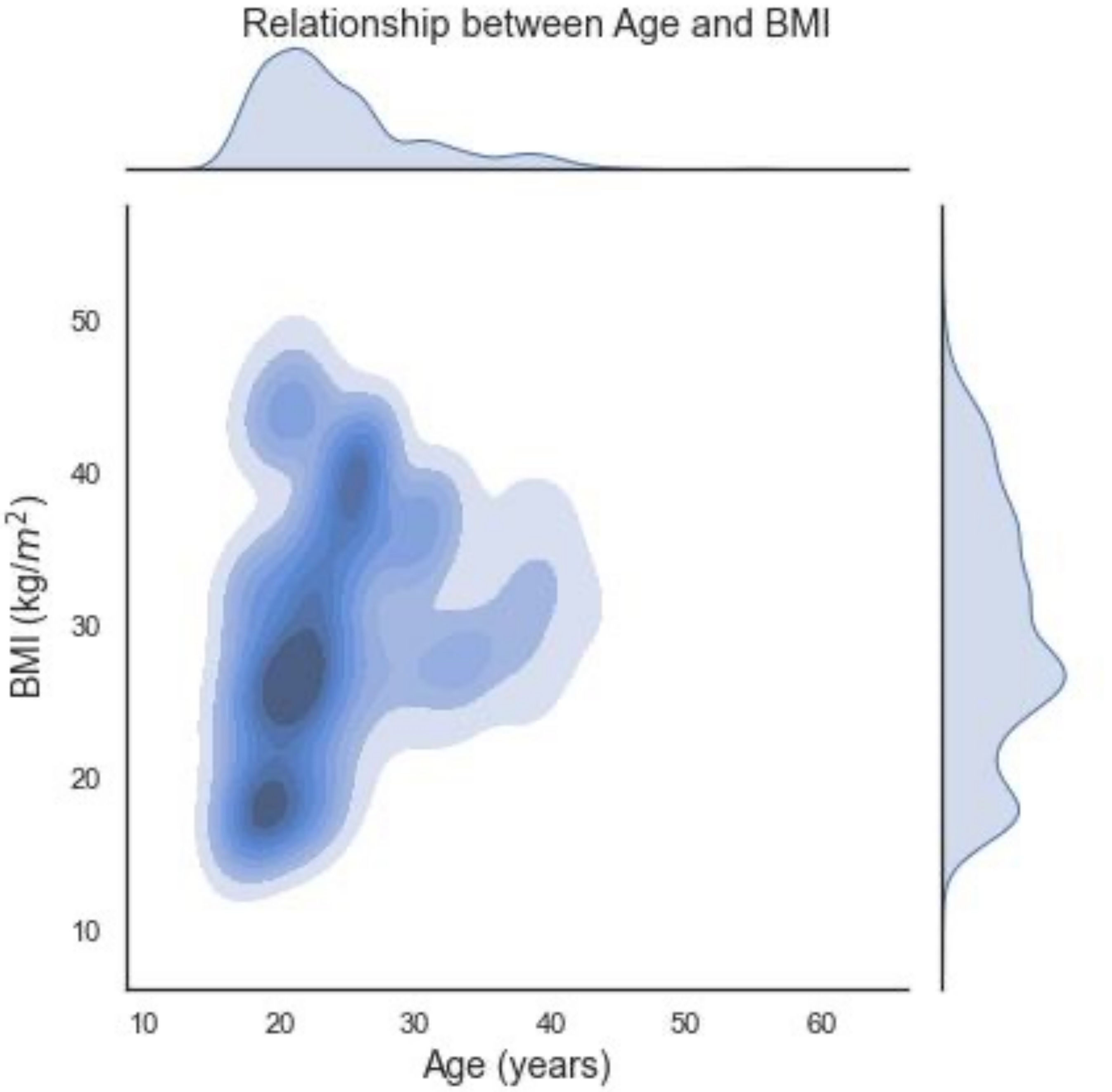
Relationship between Weight, Height and BMI



There is a slight tendency for obesity level to increase with age…



… and a slight tendency for BMI to increase with age, but with such a narrow age distribution this is not a strong correlation



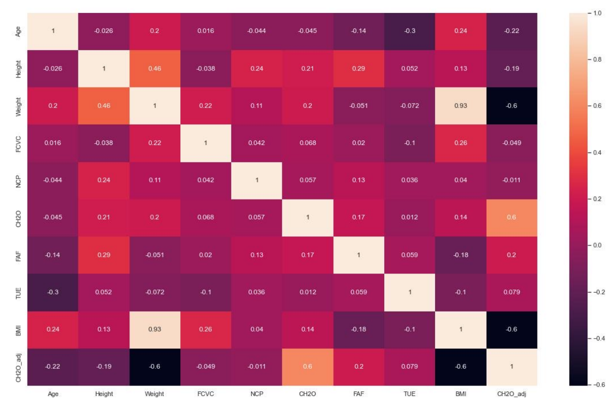
Business Question 1:	Can we predict obesity class based on lifestyle habits and/or physical condition ?
Problem type:	**Classification**

Models	Feature selection
Logistic Regression	Best 10 features (by correlation coefficient)
Support Vector Machine (SVC)	BMI only
Gaussian Naïve Bayes	All features

Workflow:

Categorical data: replaced strings with binary integers; get_dummies()	Test-Train Split Data (20:80)	Standardised continuous data (StandardScaler())	GridSearchCV() to determine optimal hyperparameters for each model

Best features



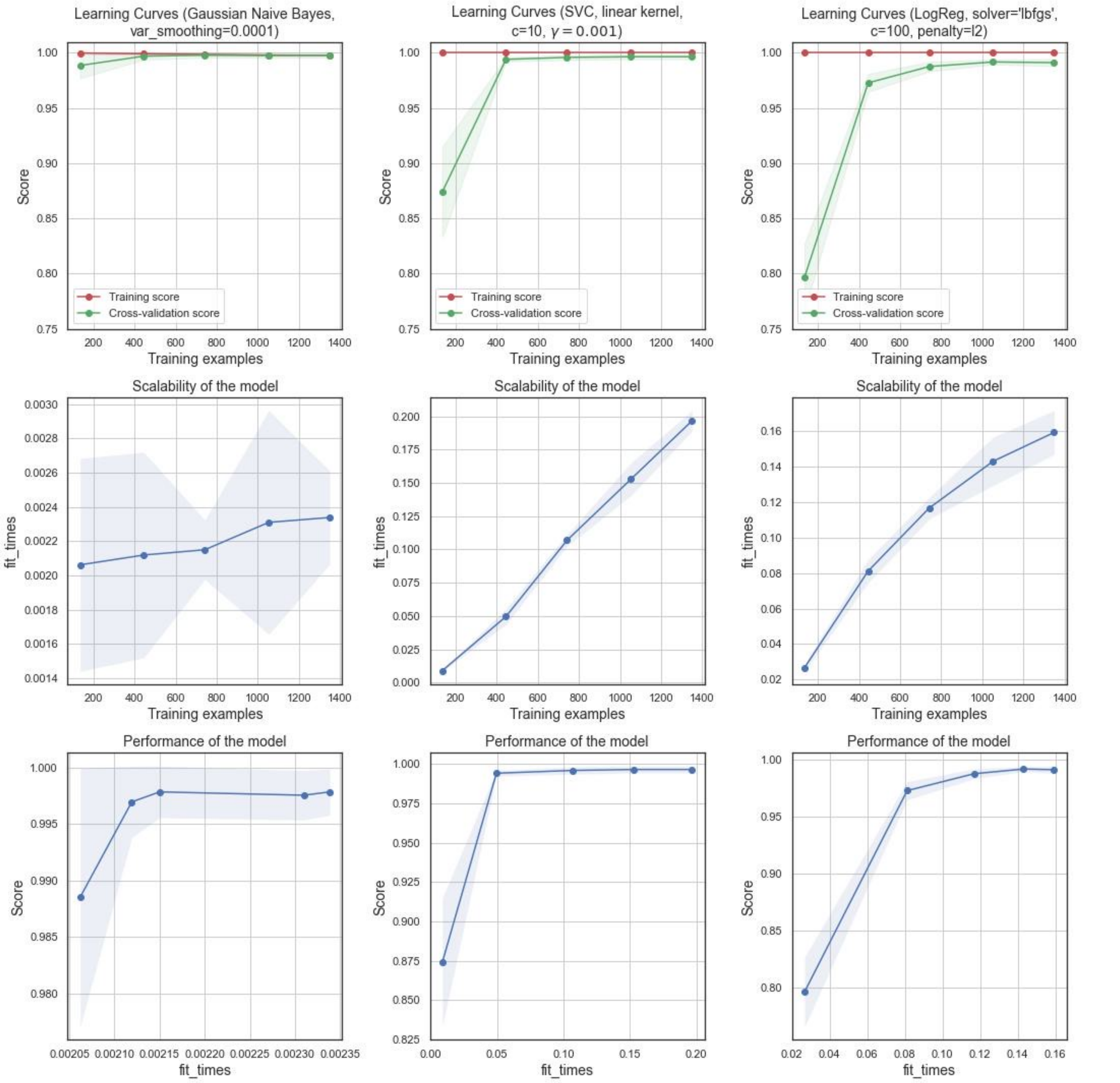
Pearson correlation coefficient

BMI	0.977826
Weight	0.913251
family_history_with_overweight	0.505148
CAEC_Sometimes	0.453188
Age	0.282913
FAVC	0.247793
FCVC	0.227759
CALC_Sometimes	0.214067
CAEC_Frequently	-0.418948
CH2O_adj	-0.609571

Learning Curves

Three models

All features



confusion matrix

	predicted_underweight	predicted_normal	predicted_overweight_I	predicted_overweight_II	predicted_obese_I	predicted_obese_II	predicted_obese_III
_underweight	61	0	0	0	0	0	0
is_normal	0	45	0	0	0	0	0
_overweight_I	0	0	61	0	0	0	0
_overweight_II	0	0	0	60	0	0	0
is_obese_I	0	0	0	0	79	0	0
is_obese_II	0	0	0	0	0	54	0
is_obese_III	0	0	0	0	0	1	62

Best Model:

SVC

classification report

	precision	recall	f1-score	support
is_underweight	1.00	1.00	1.00	61
is_normal	1.00	1.00	1.00	45
is_overweight_I	1.00	1.00	1.00	61
is_overweight_II	1.00	1.00	1.00	60
is_obese_I	1.00	1.00	1.00	79
is_obese_II	0.98	1.00	0.99	54
is_obese_III	1.00	0.98	0.99	63
accuracy			1.00	423
macro avg	1.00	1.00	1.00	423
weighted avg	1.00	1.00	1.00	423

Business Question 2:	HOW ACCURATELY CAN WE PREDICT OBESITY WITHOUT THE BMI AND WEIGHT VARIABLES?

Models
Logistic Regression
Support Vector Machine (SVC)
Gaussian Naïve Bayes

Feature selection
All features except 'Weight' and 'BMI'

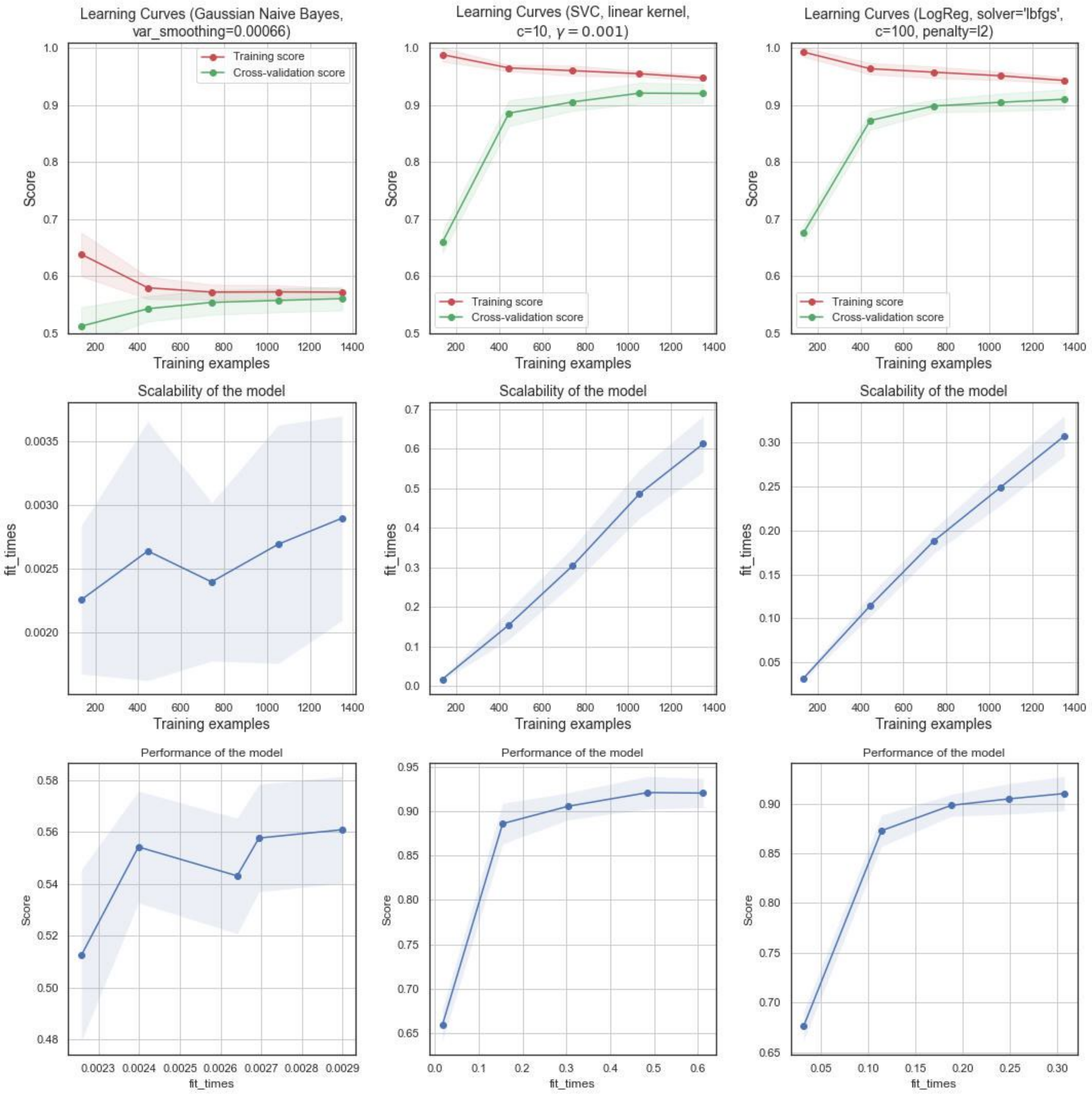
Workflow:

Categorical data: replaced strings with binary integers; get_dummies()	Test-Train Split Data (20:80)	Standardised continuous data (StandardScaler())	GridSearchCV() to determine optimal hyperparameters for each model

Learning Curves

Three models

All features *except* ‘Weight’ and ‘BMI’



confusion matrix

	predicted_underweight	predicted_normal	predicted_overweight_I	predicted_overweight_II	predicted_obese_I	predicted_obese_II	predicted_obese_III
_underweight	57	4	0	0	0	0	0
is_normal	5	32	6	2	0	0	0
_overweight_I	0	6	50	5	0	0	0
_overweight_II	0	0	4	55	1	0	0
is_obese_I	0	0	0	10	66	3	0
is_obese_II	0	0	0	0	0	54	0
is_obese_III	0	0	0	0	1	1	61

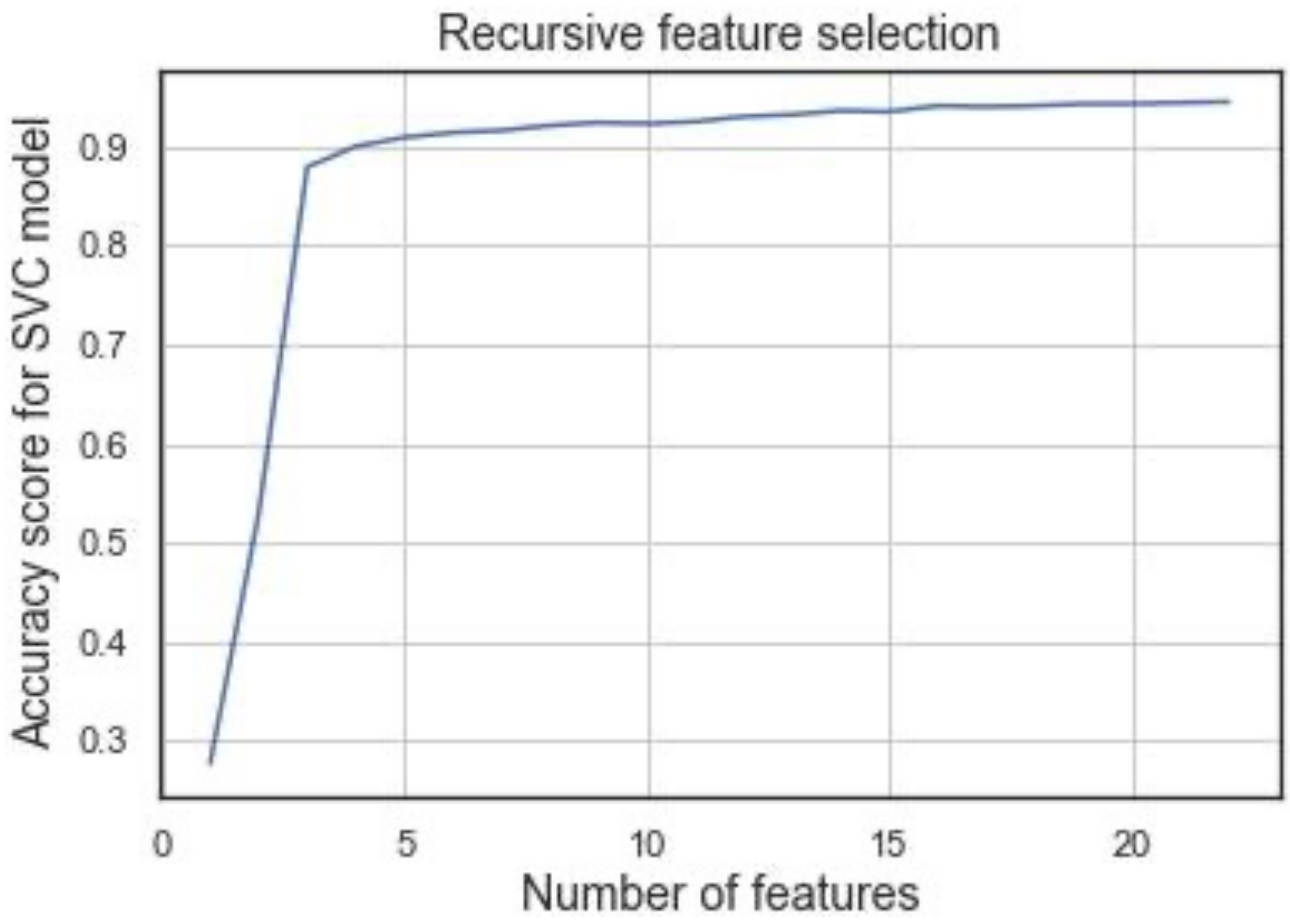
Best Model:
SVC

classification report

	precision	recall	f1-score	support
is_underweight	0.92	0.93	0.93	61
is_normal	0.76	0.71	0.74	45
is_overweight_I	0.83	0.82	0.83	61
is_overweight_II	0.76	0.92	0.83	60
is_obese_I	0.97	0.84	0.90	79
is_obese_II	0.93	1.00	0.96	54
is_obese_III	1.00	0.97	0.98	63
accuracy			0.89	423
macro avg	0.88	0.88	0.88	423
weighted avg	0.89	0.89	0.89	423

Can the model be good enough with fewer features?

Top 5 Features (score = 0.909)
CH2O_adj
CH2O
Height
Gender
Age



Business Question 3:	HOW ACCURATELY CAN WE PREDICT OBESITY WITHOUT THE BMI, WEIGHT & CH2O_adj VARIABLES?

Models
Logistic Regression
Support Vector Machine (SVC)
Gaussian Naïve Bayes

Feature selection
All features except ‘Weight’ , ‘BMI’ & ‘CH2O_adj’

Workflow:

Categorical data: replaced strings with binary integers; get_dummies() →	Test-Train Split Data (20:80) →	Standardised continuous data (StandardScaler()) →	GridSearchCV() to determine optimal hyperparameters for each model

```
family_history_with_overweight      0.505148
CAEC_Sometimes                      0.453188
Age                                 0.282913
FAVC                                0.247793
FCVC                                0.227759
CALC_Sometimes                      0.214067
CAEC_Frequently                    -0.418948
```

Jupyter Notebook problems have prevented me from completing this!

Top 5 Features (score (SVC) = 0.466)
Freq of vegetable consumption (FCVC)
Age
Family history
Freq of alcohol consumption (sometimes)
Eating between meals (sometimes)

Score (SVC) – All features except ‘BMI’, ‘Weight’, ‘Height’ and ‘CH2O_adj’
0.5195