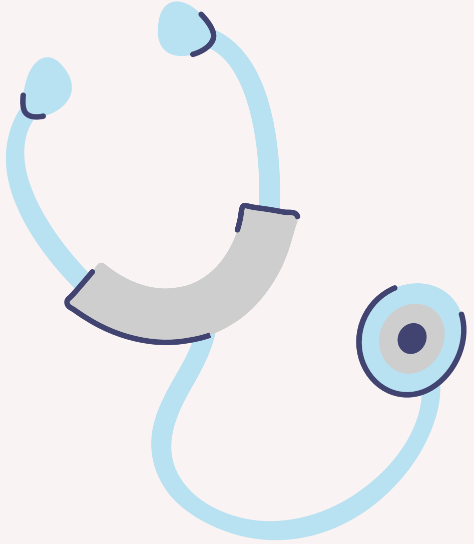
Predicting Obesity Level using Machine Learning Models: Random Forest, XGBoost and LightGBM

Nicholas Farandi Harjanto - 2602065553
Bryan Orville Audric - 2602160750
Edrico Putra Pramana - 2602078133
Darren King Wijaya - 2602145232

List of Contents



Introduction

What is obesity?

Obesity is a medical condition characterized by an excessive accumulation of body fat that presents a risk to health such as cardiovascular disease, diabetes, reducing life expectancy and causing disability.

Importance of identifying cause of obesity

Understanding obesity causes is vital for prevention and management. Tailored interventions based on accurate identification improve weight management, reduce health risks, and enhance overall health through personalized treatment plans.

Objective of this research

to create an accurate obesity level detector using machine learning models such as Random Forest, XGBoost and LightGBM

Dataset

Labels

No	Body Weight Category
1	Insufficient weight
2	Normal weight
3	Overweight I
4	Overweight II
5	Obesity type I
6	Obesity type II
7	Obesity type III

Details

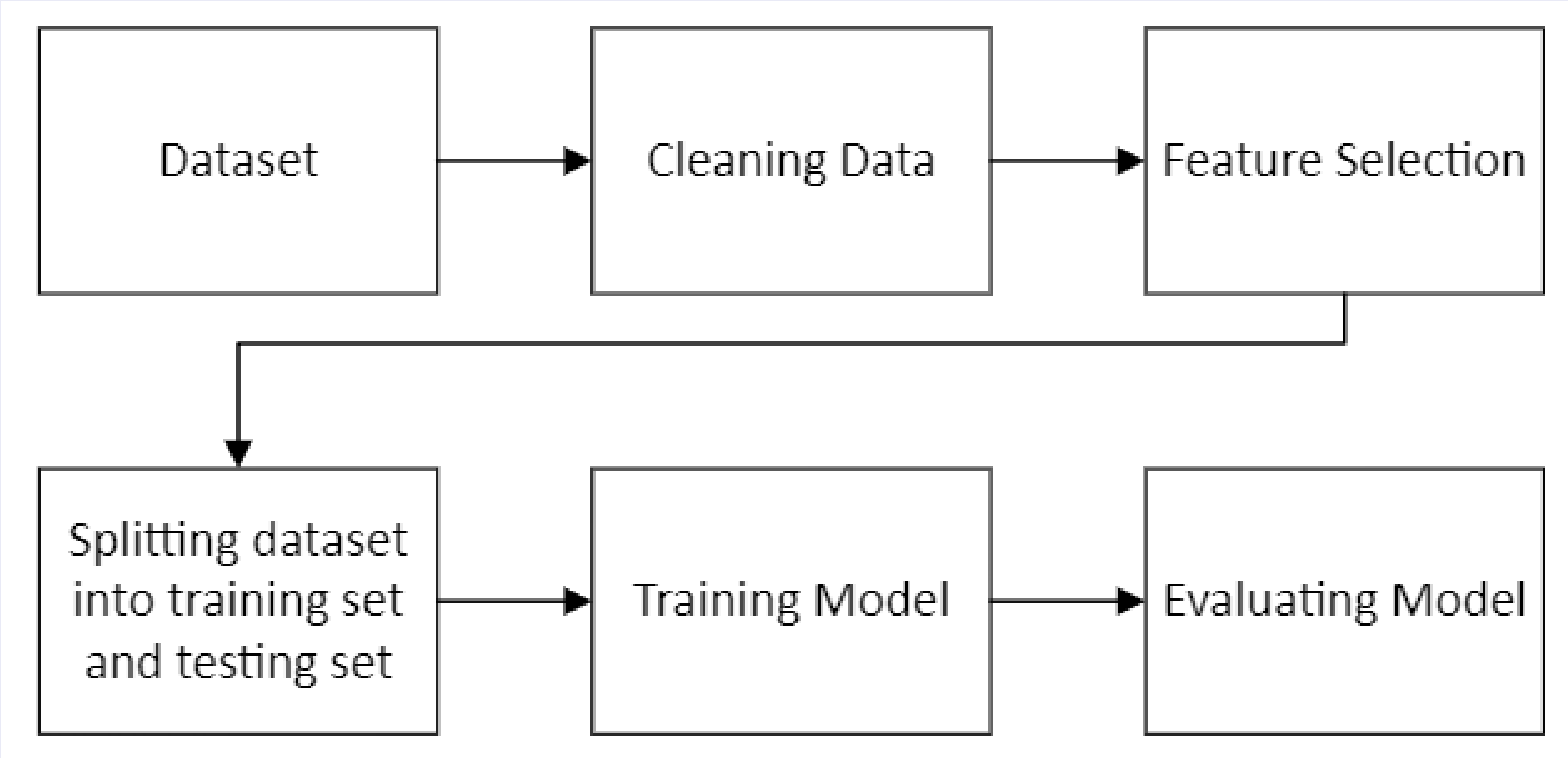
The dataset contains 20751 data which consist of 16 features and 7 labels

Dataset

Features

Features	Data type	Description
Gender	Categorical	-
Age	Numerical	-
Height	Numerical	-
Weight	Numerical	-
Family History with_Overweight	Categorical	-
FAVC	Categorical	Frequent consumption of high caloric food
FCVC	Numerical	Frequency of eating vegetables
NCP	Numerical	Number of main meals
CAEC	Categorical	Frequency of food consumption between main meals
SMOKE	Categorical	Smoking status
CH20	Numerical	Daily frequency of water consumption.
SCC	Categorical	Calories consumption monitoring
FAF	Numerical	Frequency of physical activity
TUE	Numerical	Time using technology devices
CALC	Categorical	Frequency of alcohol consumption
MTRANS	Categorical	Transportation used
NObeyesdad	Categorical	Obesity levels (Target)

Research Flow



Cleaning Data, Feature Selection

- The dataset is cleaned by removing outlier and data with null values
- Pearson Correlation is implemented to help select most relevant features, by calculating every feature’s correlation coefficient

Selected Features

The features of the dataset will be selected if it has an absolute correlation coefficient above 0.1

Features	Correlation Coefficient
Age	0.3670194416789297
Height	0.16803239219481939
Weight	0.922249548022407
Family History with Overweight	0.5086358618078486
FAVC	0.1914821991592605
FCVC	0.20358950303742987
CAEC	-0.35152261670757046
CH20	0.257157134900514 8
SCC	-0.17869528879049948
FAF	-0.211909968337177
TUE	-0.11872702963270912
CALC	0.15792181183708395

Standardizing Age And Weight

	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	CAEC	CH2O	SCC	FAF	TUE	CALC	NObeyesdad
0	-0.863345	1.673491	-1.349337	0	0	3.000000	2.0	1.000000	0	0.144950	0.000000	1.0	0.0
1	-0.181728	1.700000	-1.335948	0	1	3.000000	2.0	2.000000	0	2.000000	1.000000	1.0	0.0
2	-1.317409	1.556579	-1.564135	0	1	2.000000	1.0	1.198883	0	1.000000	0.000000	1.0	0.0
3	-1.108628	1.781543	-1.302593	0	1	1.140615	1.0	1.639524	0	0.520408	1.000000	1.0	0.0
4	-1.108628	1.691206	-1.274883	1	1	2.000000	1.0	1.000000	0	0.520407	1.560402	0.0	0.0

Splitting Dataset into testing set and training set

The dataset is divided with a ratio of 80% for training set and 20% for testing set

Machine Learning Models

Random Forest

Random Forest combines random decision trees to avoid overfitting. Each tree votes on the outcome, enhancing accuracy.

XGBoost

XGBoost is an efficient tree boosting system that minimizes errors, prevents overfitting, and scales to large datasets effectively.

LightGBM

LightGBM is an Advanced GBDT with techniques like GOSS and EFB for optimized training and efficiency. Accelerates model training, reduces memory usage, and maintains high accuracy, ideal for large datasets in ML tasks.

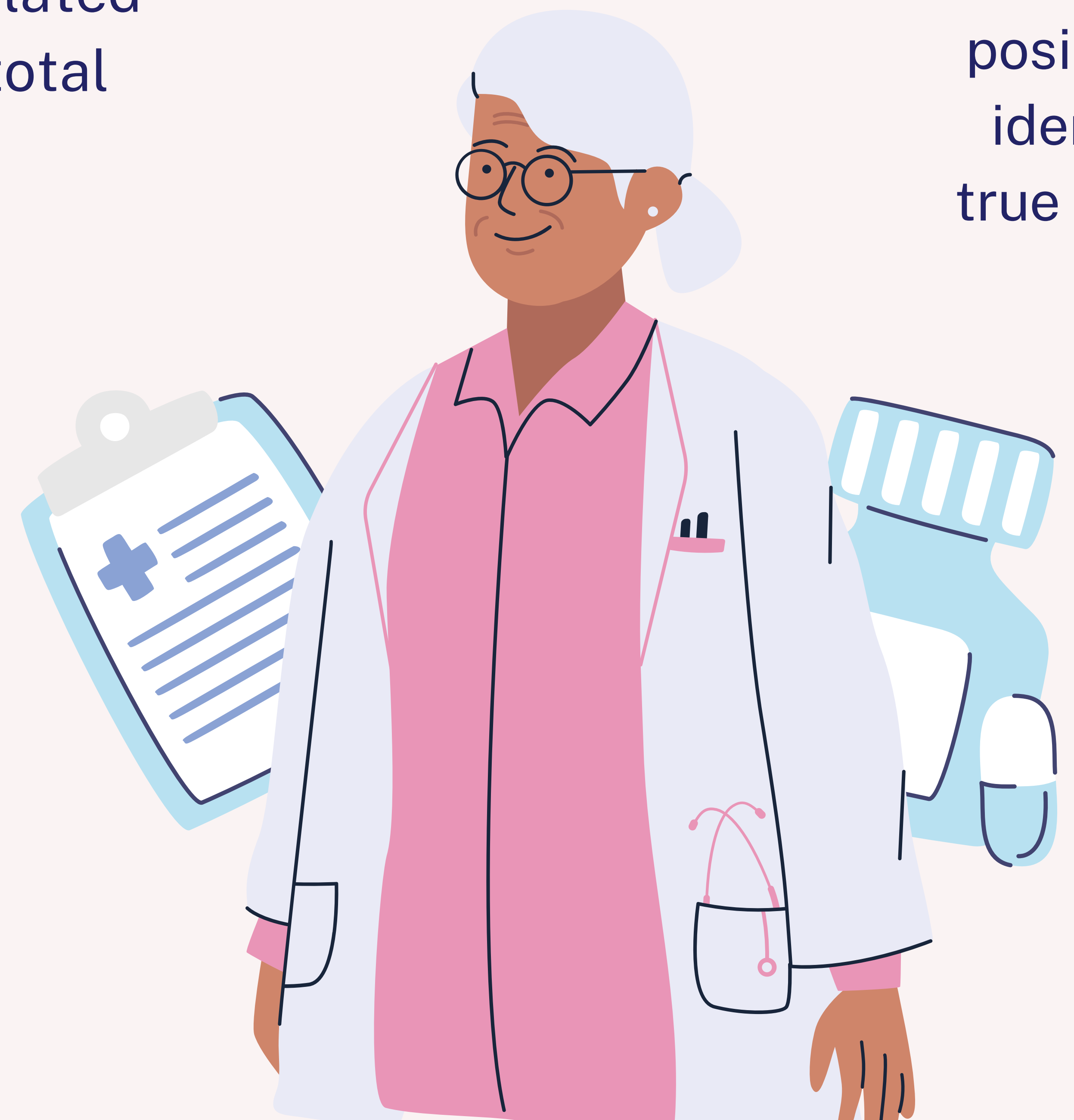
Evaluation Metrics

Accuracy

Accuracy assesses a model's overall classification correctness. It's calculated by dividing correct predictions by total predictions.

Precision

Precision is vital in classification, especially for minimizing false positives. It measures accurate positive predictions, calculated by dividing correctly predicted positives by all predicted positives.



Recall Score

Recall, or sensitivity, is crucial in classification, especially for capturing all positives. It measures the model's ability to identify positives correctly, calculated as true positives divided by all actual positives.

F1 Score

Recall, crucial in classification, captures all positives. It's true positives divided by all actual positives.

Results

Random Forest

Metric	Insufficient Weight	Normal Weight	Overweight I	Overweight II	Obesity Type I	Obesity Type II	Obesity Type III
Precision	0.94	0.83	0.79	0.78	0.88	0.96	0,99
Recall	0.94	0.86	0.72	0,83	0.86	0.97	1
F1-Score	0.94	0.85	0.76	0,80	0.87	0.97	0,99
Accuracy	0.88						

Results

XGBoost

Metric	Insufficient Weight	Normal Weight	Overweight I	Overweight II	Obesity I	Obesity II	Obesity III
Precision	0,94	0,85	0,81	0,80	0,90	0,96	0.99
Recall	0,93	0,87	0,78	0,85	0,84	0,97	1
F1-Score	0,94	0,86	0,80	0,83	0,87	0,97	0.99
Accuracy	0.89						

Results

LightGBM

Metric	Insufficient Weight	Normal Weight	Overweight I	Overweight II	Obesity I	Obesity II	Obesity III
Precision	0,93	0,86	0,81	0,81	0,88	0,96	0.99
Recall	0,94	0,86	0,78	0,85	0,85	0,96	0.99
F1-Score	0,94	0,86	0,79	0,83	0,86	0,96	0,99
Accuracy	0.89						

Discussion

- Both XGBoost and LightGBM achieved an accuracy of 89% higher than Random Forest.
- When predicting the categories “Insufficient Weight”, “Obesity II”, and “Obesity III” has the highest Precision, Recall, and F1-Score, the models are much more consistent at detecting the outside classes instead of the classes in between

Impact in Real Life Setting

- Some features such as frequency of vegetable consumption (FCVC) and other lifestyle-related data require detailed personal information, which may be an inconvenience for some to gather.
- However, the information needed is still feasible to be obtained easily. If the information is collected and given to the models, it can be preventive measure to detect early signs of obesity before making a further consultation with the doctor

Conclusion

- Both XGBoost and LightGBM achieved the highest accuracy of 89% amongst three machine learning models used
- By collecting several data such as frequency of vegetable consumption (FCVC) and other lifestyle-related data, the model can use the data to detect early sign of obesity

Future Direction

- Future research could go deeper on finding more accurate model as well as finetuning our method to further increase the performance in determining the classification of obesity level
- Other ideas might be to collaborate with biology or healthcare professionals to create tools that can automatically collect parameters such as frequency of vegetable consumption (FCVC) and other lifestyle-related data, which can be used for the models to detect early sign of obesity

Thank you for your attention