Accenture Marketing

Classifying Obesity for Lifestyle Marketing Customer Profiling

Andre Barle

Agenda

1. EDA and data preparation
2. Considerations
3. Final model
4. Conclusion and key takeaways
5. Next steps

Business Objective

How can we improve targeted marketing efforts for customers who are overweight and obese?

To improve marketing efforts for obese and overweight customers based on lifestyle metrics, age, and gender

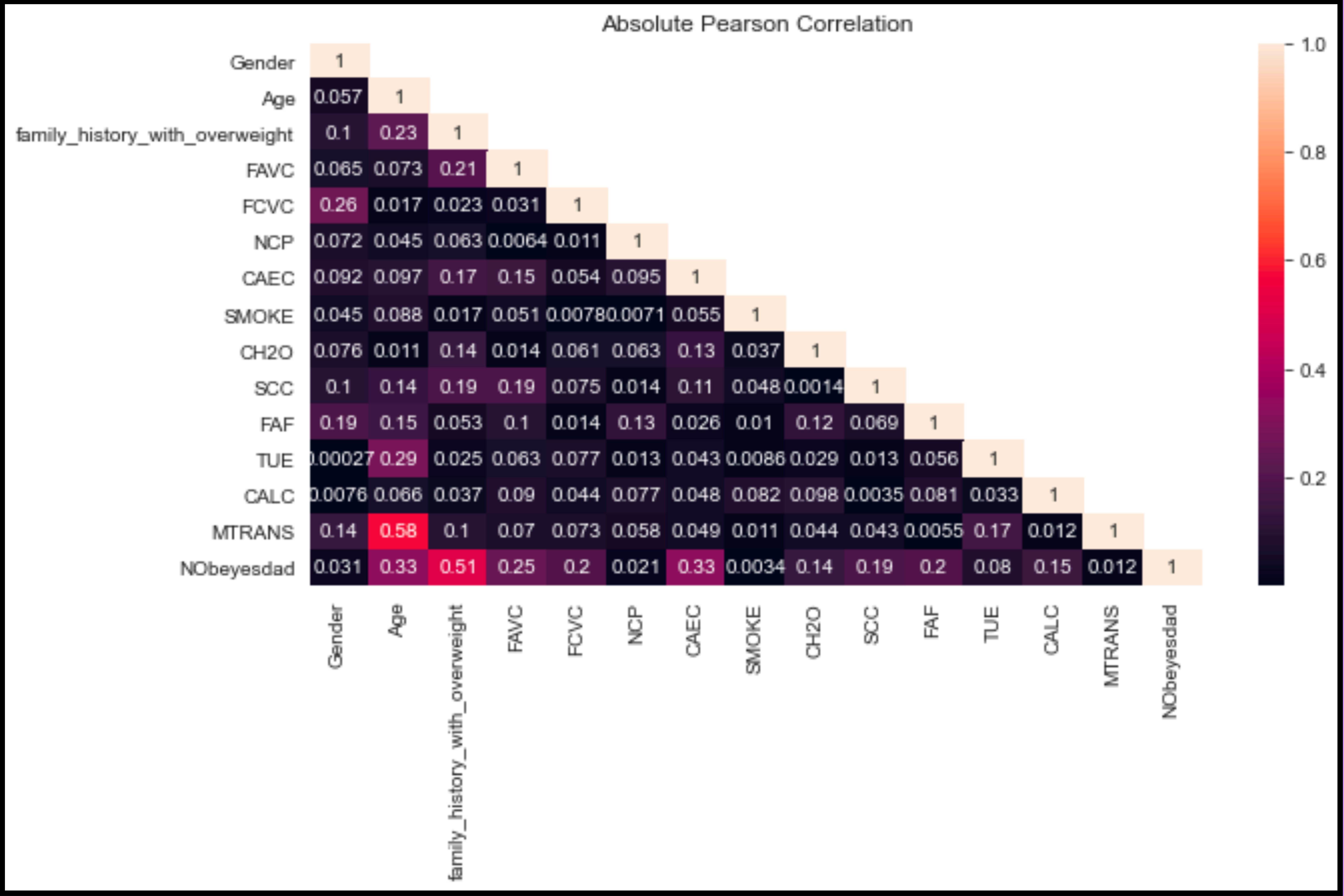
Questions we are hoping to answer:

- How can we identify if someone is Obese without asking for their height and weight?
- What lifestyle characteristics are the most important in predicting obesity?
- Can we predict if someone may be obese based on their search history profile?

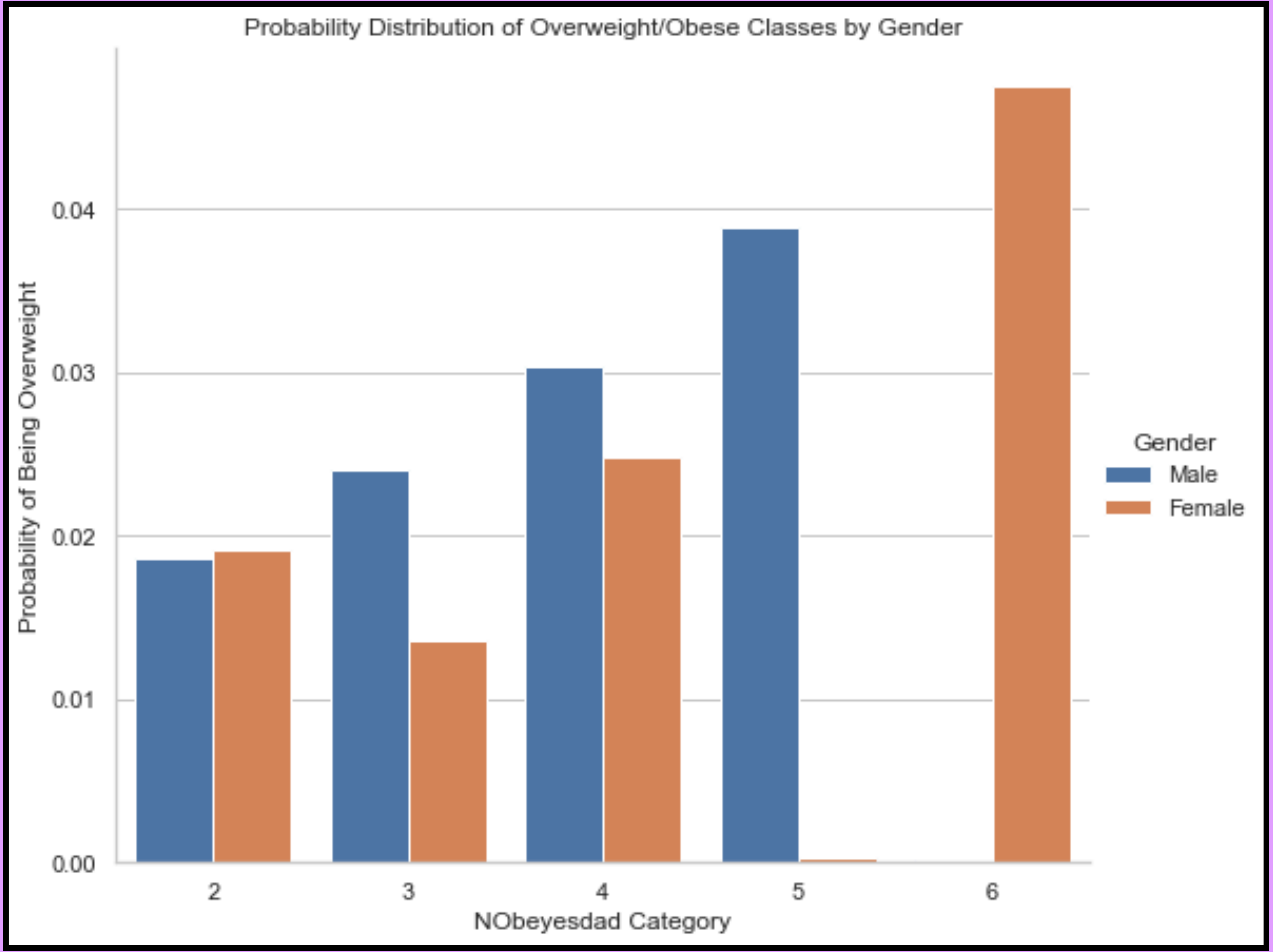
Exploring the Data

Found the strongest correlations

- Leveraged obesity data collected through a survey and synthetically expanded
- Strongest Correlations:
  - Age
  - Family History of Obesity
  - Consumption of food between meals
  - Frequent consumption of high caloric food
- Used machine learning to classify obesity - 7 classes labeled 0-6



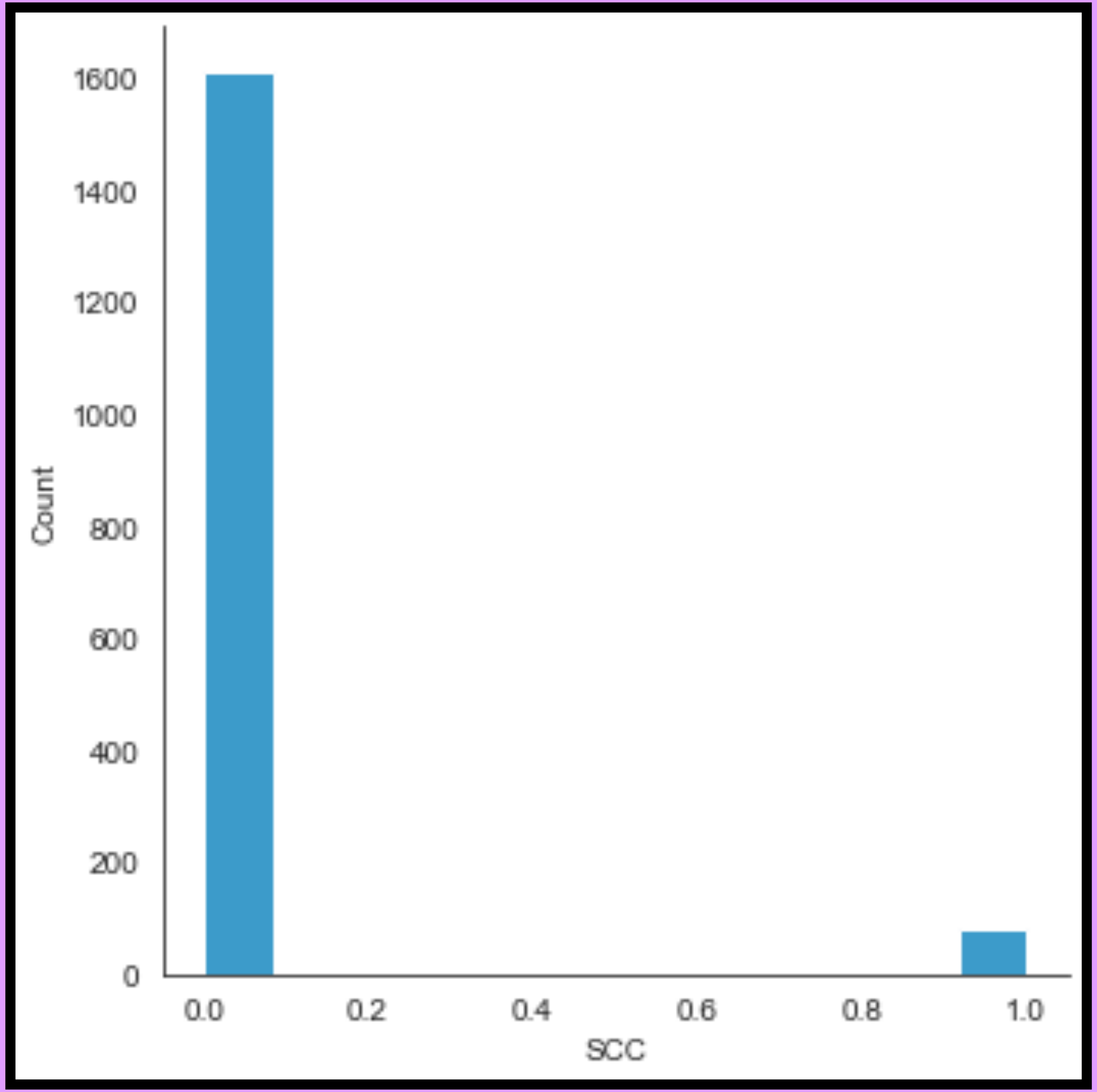
Exploring the Data cont.



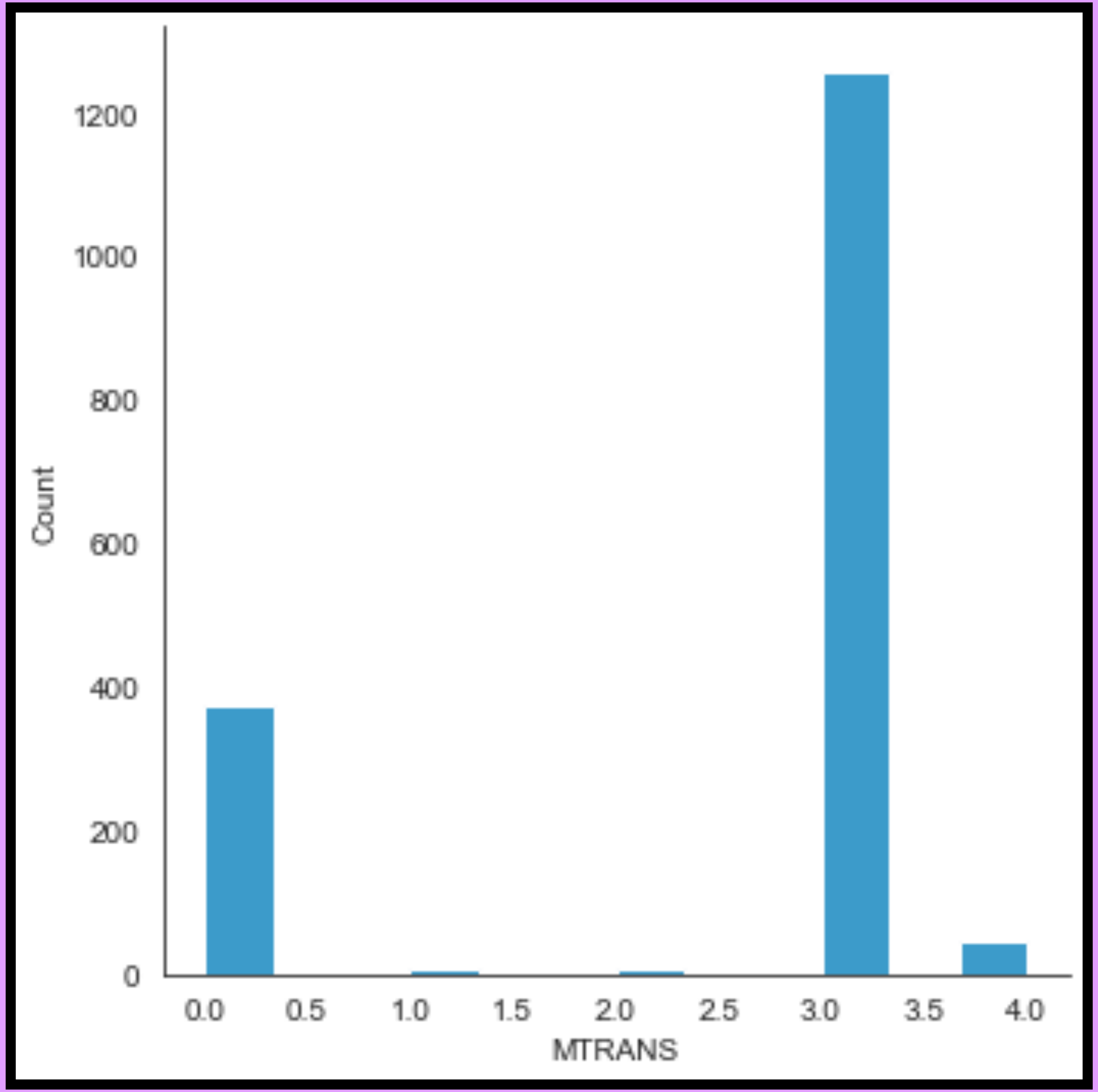
BMI = weight (kg) / [height (m)]2

Exploring the Data cont.

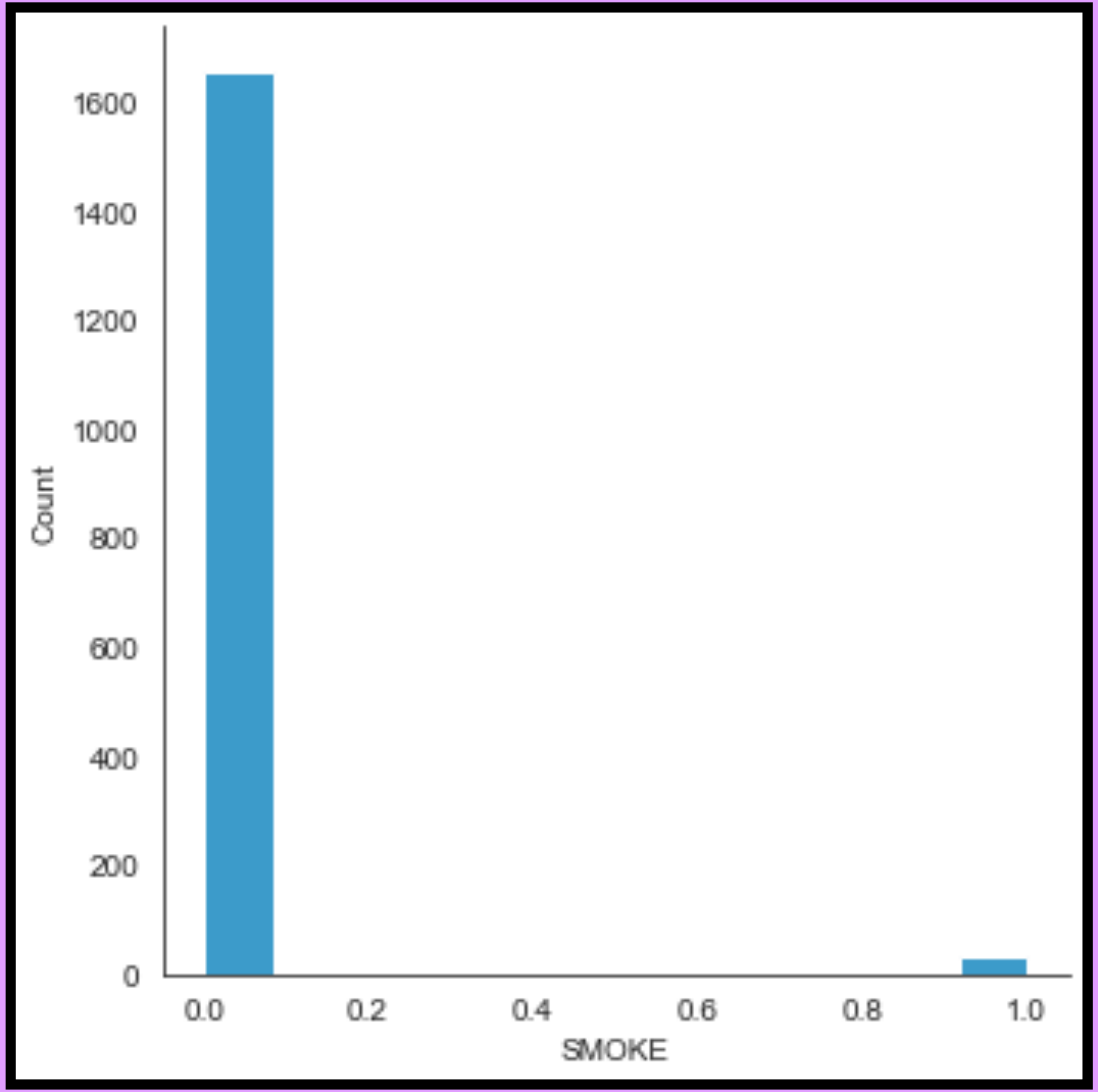
Do you monitor the calories you eat daily?



Which transportation do you usually use?

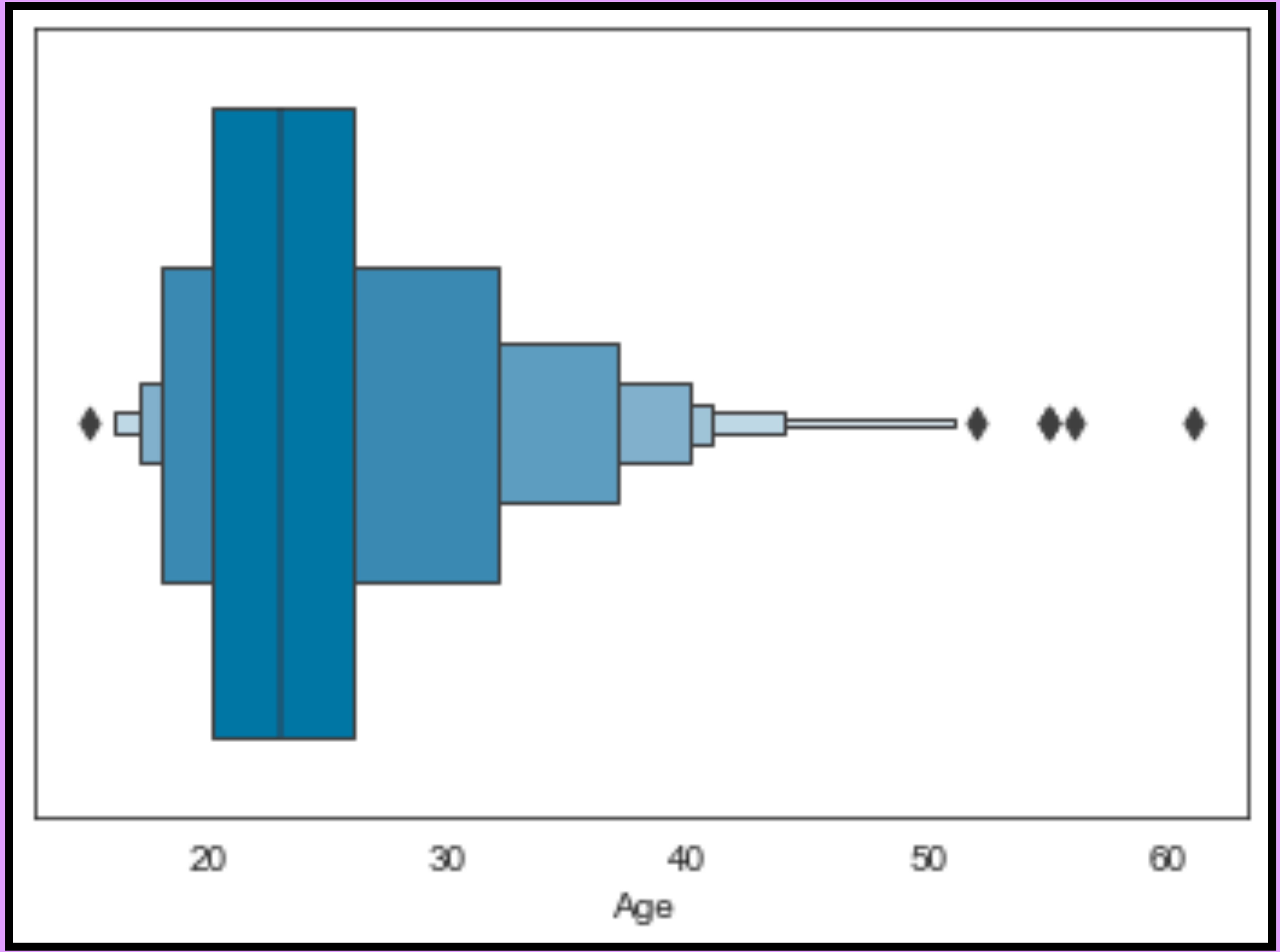


Do you smoke?

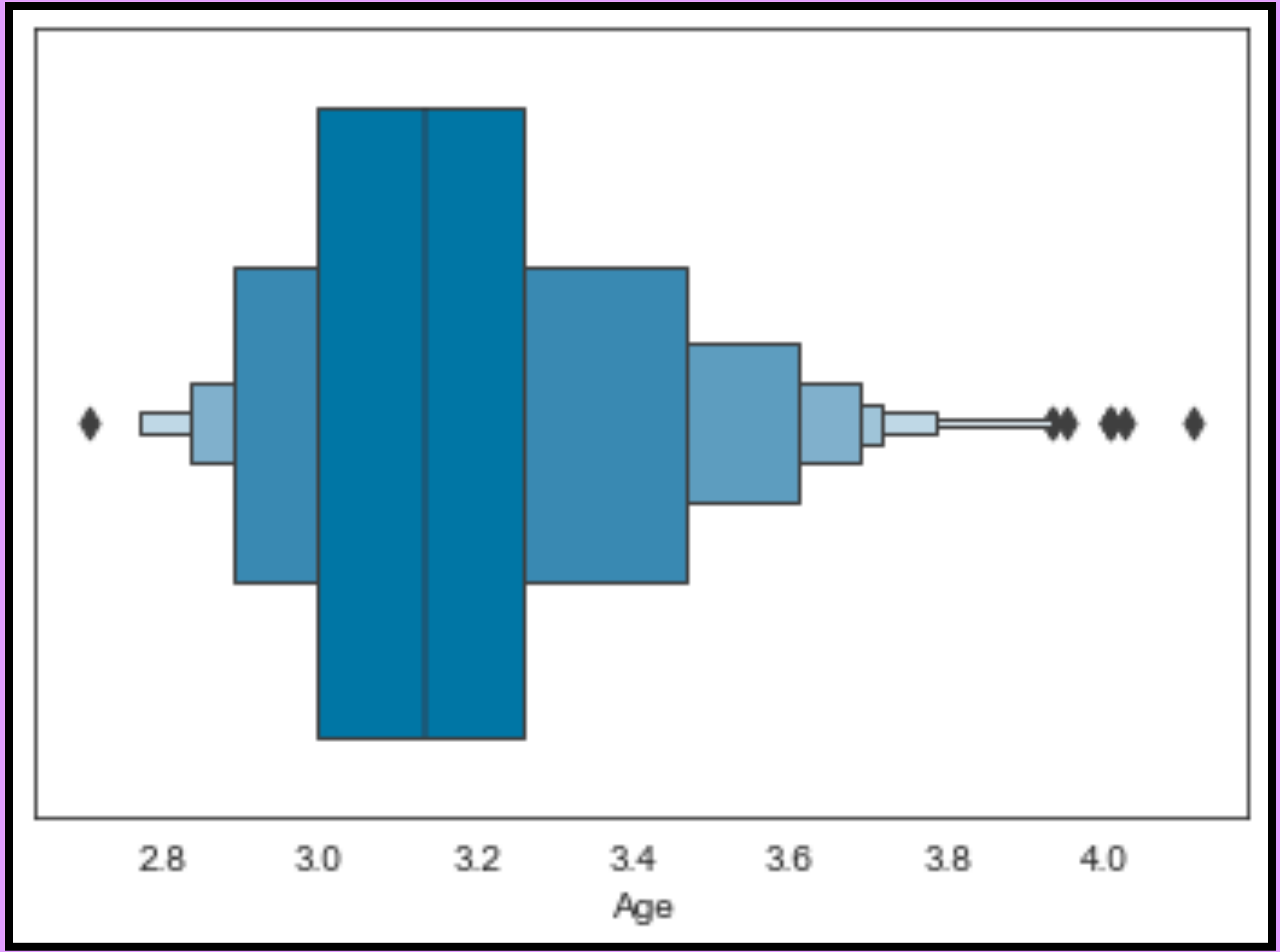


Feature Normalization

Age



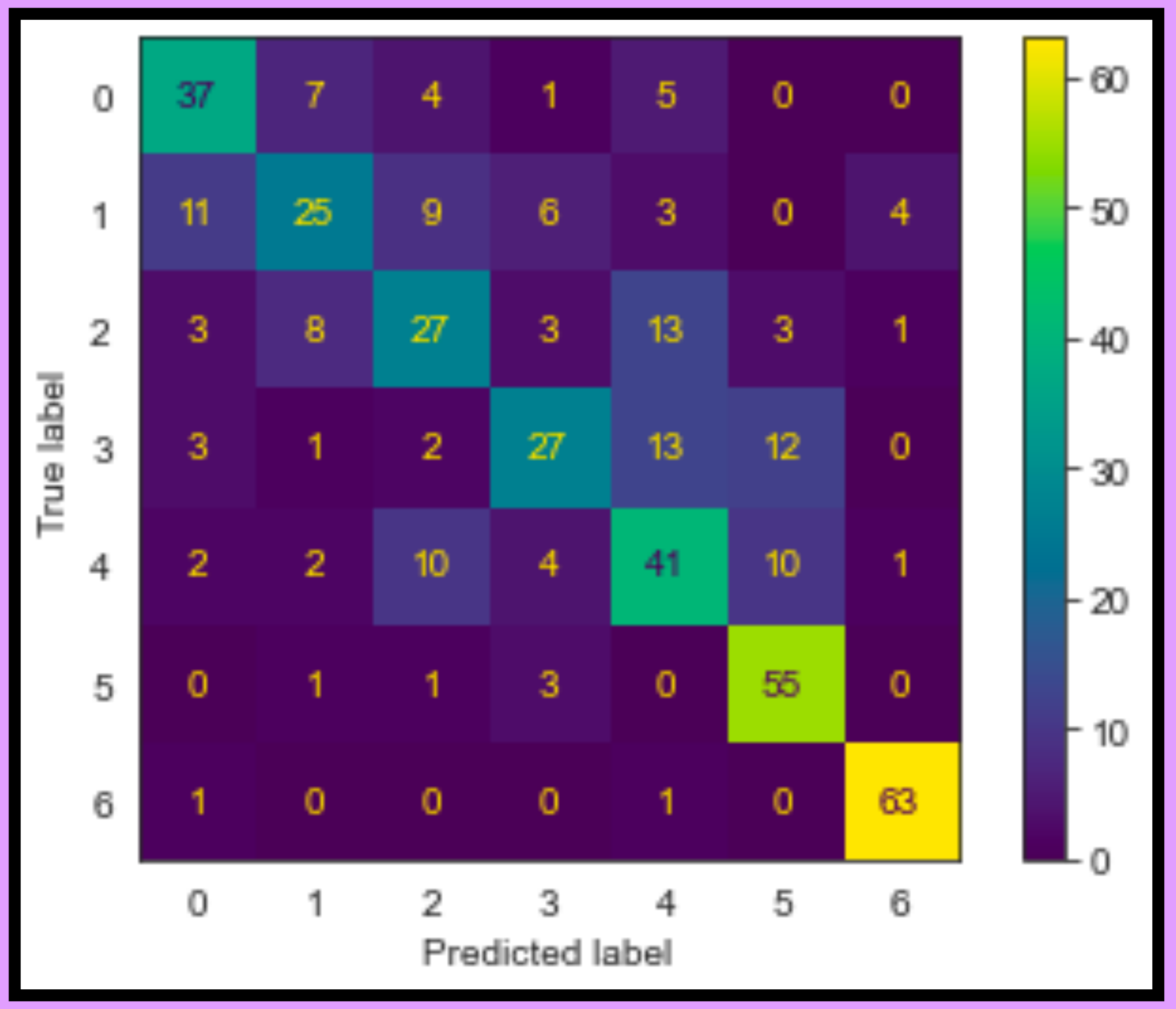
Log transformed Age



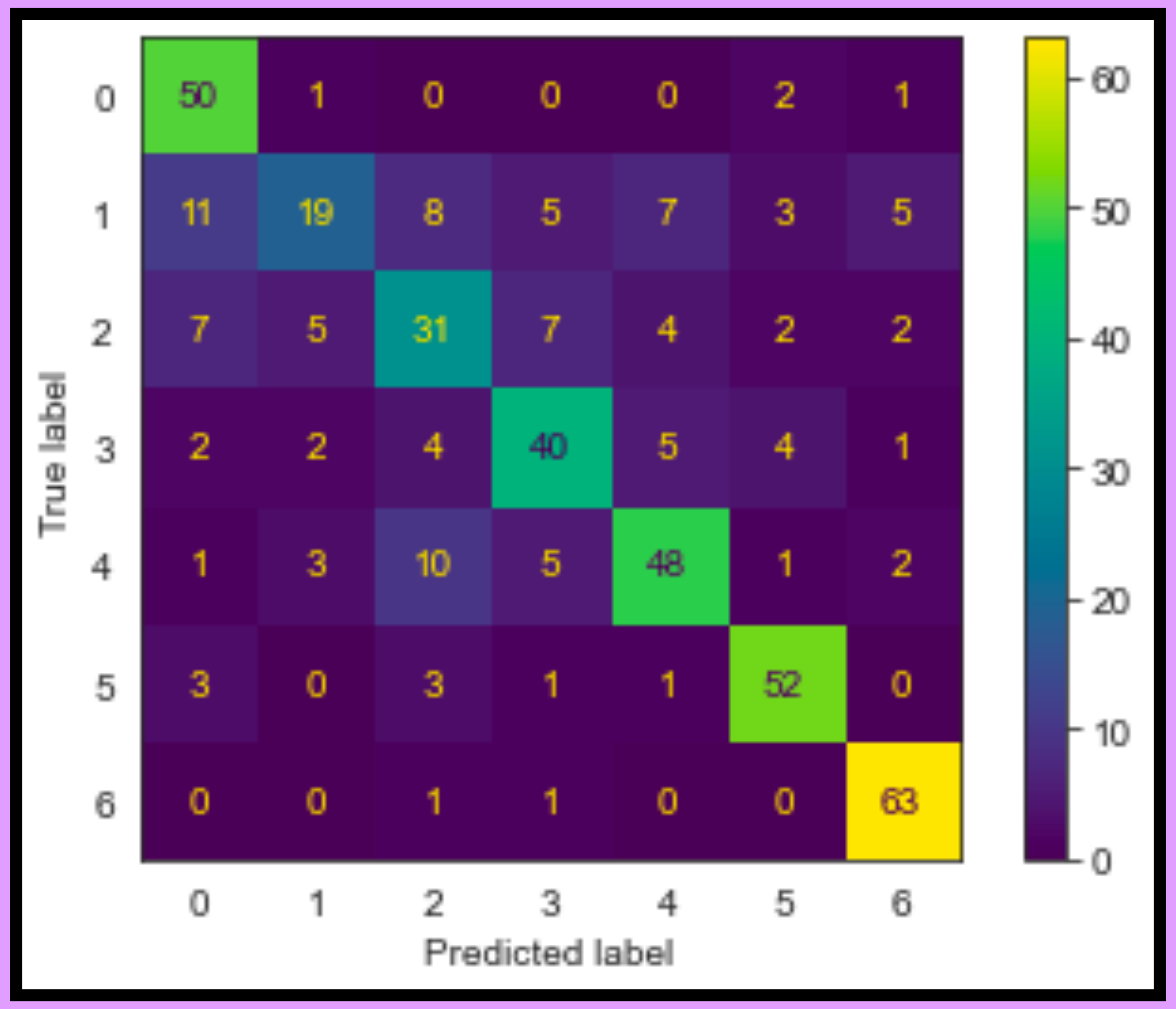
How we arrived at our model

Tested Logistic Regression, KNN, Decision Tree, Gaussian Bayes

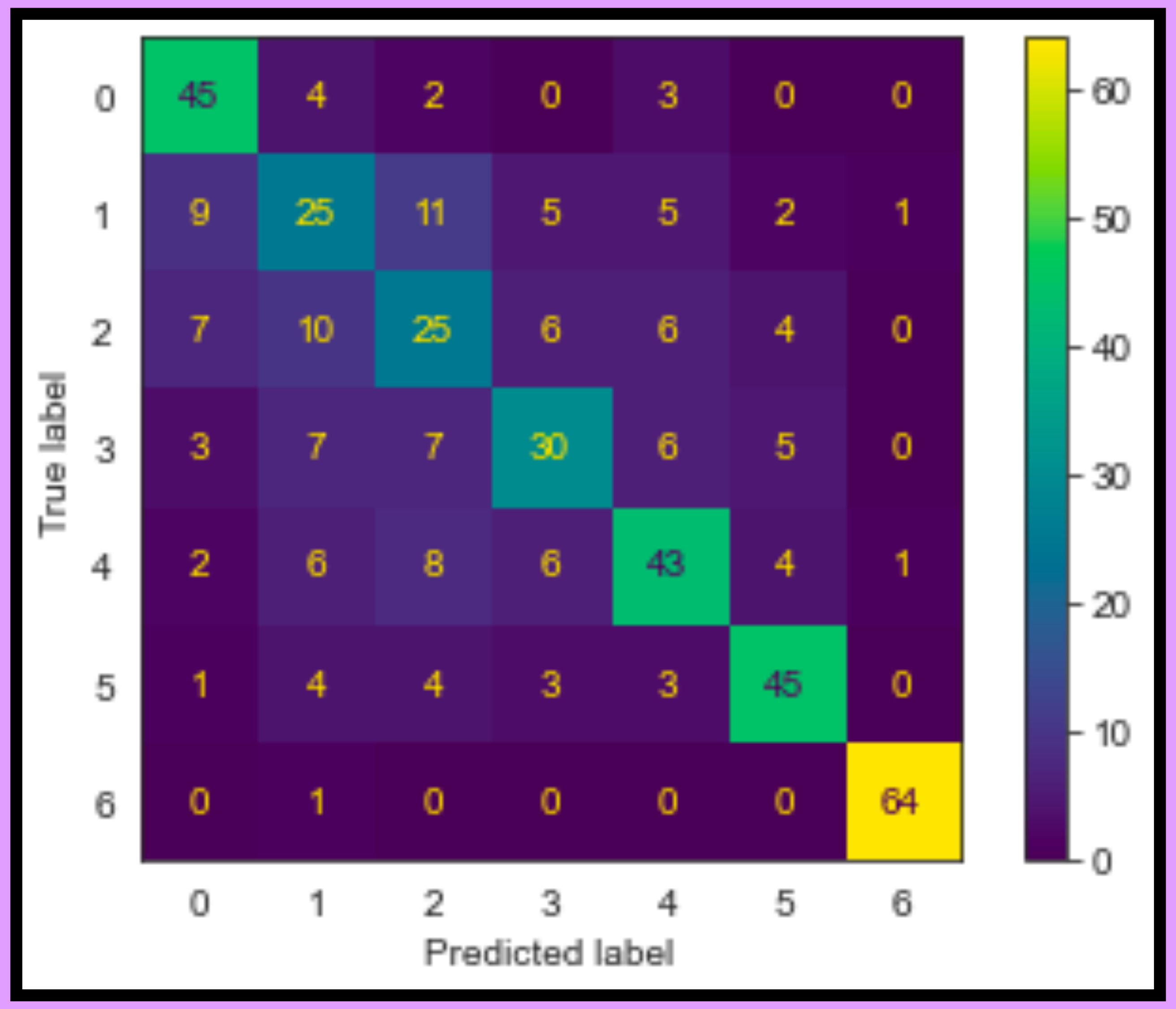
Logistic Regression



K Nearest Neighbor



Decision Tree

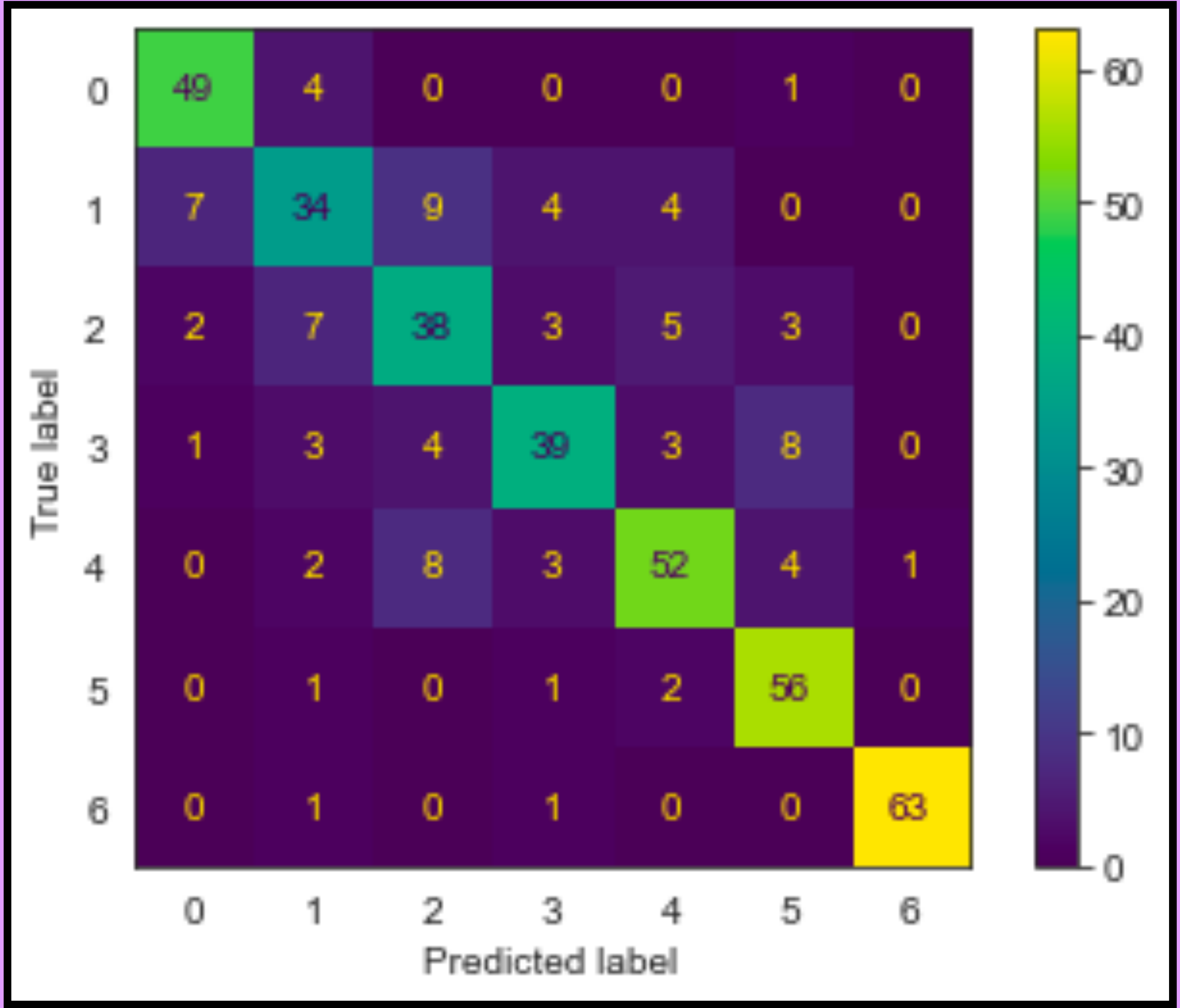


Obesity Classes

Base model

- 0: Underweight less than 18.5
- 1: Normal 18.5 to 24.9
- 2: Overweight I: 25.0 to ~27.5
- 3: Overweight II: ~27.5 to 29.9
- 4: Obesity I 30.0 to 34.9
- 5: Obesity II 35.0 to 39.9
- 6: Obesity III Higher than 40

Base Model - Random Forest

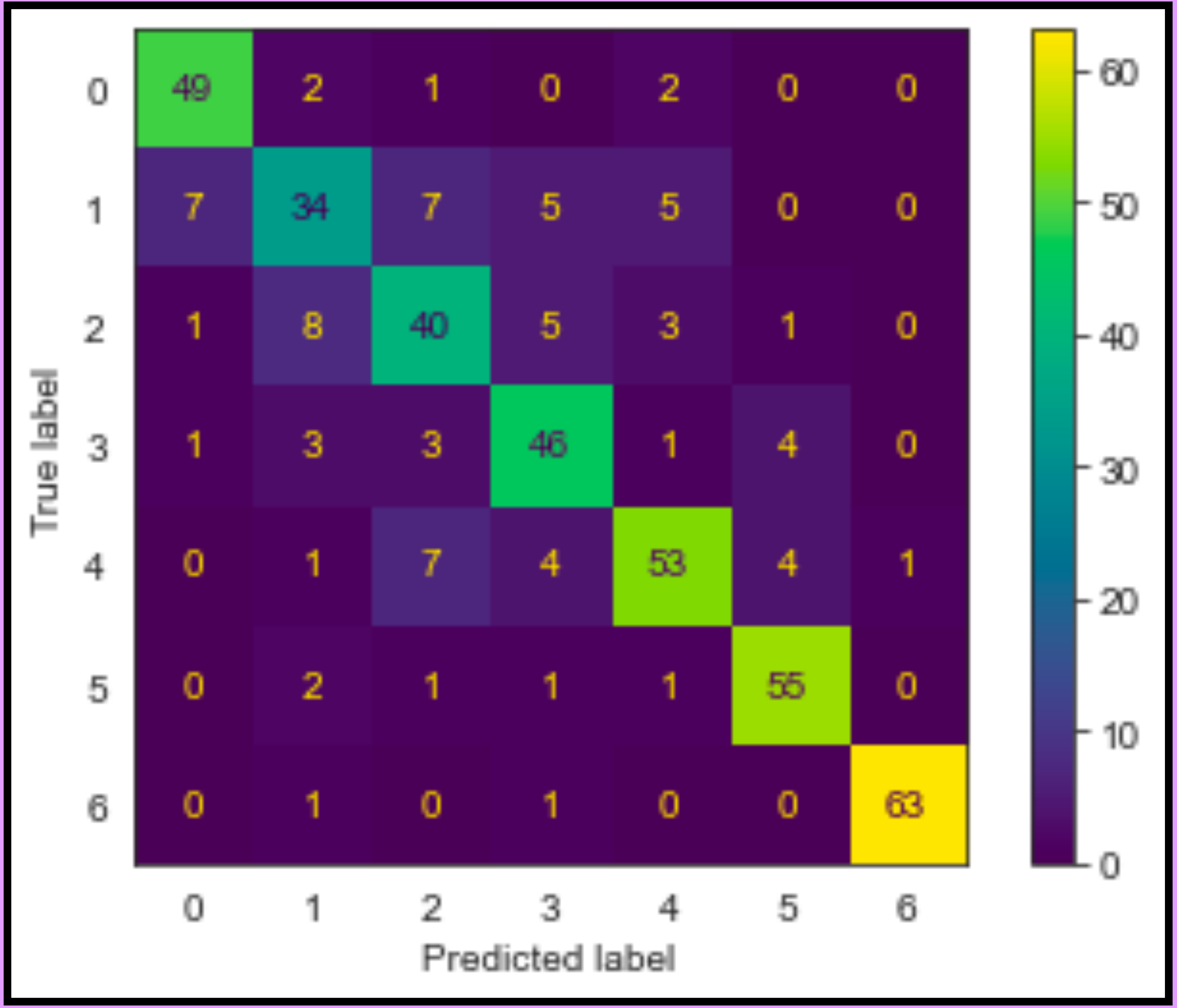


Final Predictive Model

- This model is 80% accurate in classifying the specific class of obesity the customer falls in

- **Cross Validation score** = 79.50%
- **Test Accuracy** = 80.38%

Final Model - XGBoost



Key takeaways

Our predictions

- The most important predictors are:
  - Gender
  - Family History with obesity
  - Consumption of food between meals
- All other categories are somewhat evenly important
- **F-1 score** = 79.98%
- **Recall score** = 80.28%

Final Model - XGBoost

```
Pipeline

StandardScaler
StandardScaler()

XGBClassifier
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=0.6, gamma=0, gpu_id=-1,
              importance_type='gain', interaction_constraints='',
              learning_rate=0.1, max_delta_step=0, max_depth=7,
              min_child_weight=1, missing=nan, monotone_constraints='()',
              n_estimators=200, n_jobs=0, num_parallel_tree=1,
              objective='multi:softprob', random_state=0, reg_alpha=0,
              reg_lambda=1, scale_pos_weight=None, subsample=0.8,
              tree_method='exact', validate_parameters=1, verbosity=None)
```

Conclusion

Assign importance to family history, gender, consumption of food between meals

- Obesity seems to be inherited
- Males are predicted to be more obese than females
- Eating food between meals is a predictor of obesity

- Gender: 0.1143202
- Age: 0.0627237
- family_history_with_overweight: 0.12640324
- FAVC: 0.07093425
- FCVC: 0.07636473
- NCP: 0.06553428
- CAEC: 0.11182177
- SMOKE: 0.04112323
- CH2O: 0.03709772
- SCC: 0.06637838
- FAF: 0.04186976
- TUE: 0.04007904
- CALC: 0.07396457
- MTRANS: 0.07138525

Next Steps

How we could improve this model

- We could feed more data into this model, such more survey results, more questions, add height or weight
- This model can be used to predict obesity based on a set of questions, not including height or weight metrics
- Useful for marketing segmentation based on a customer profile from search history
- Can use the weights assigned to the signals in the data to better classify if someone may be obese

Thank you

https://github.com/Cer001/Obesity_Classification

Andre Barle

Appendix

- If you are **male** and in Overweight I the probability is: 0.01865118861962575
- If you are **female** and in Overweight I the probability is: 0.019098244914439406
- If you are **male** and in Overweight II the probability is: 0.024053601874965618
- If you are **female** and in Overweight II the probability is: 0.013566339490946612
- If you are **male** and in Obese I the probability is: 0.03035863507896039
- If you are **female** and in Obese I the probability is: 0.024869048716647902
- If you are **male** and in Obese II the probability is: 0.03886144805965541
- If you are **female** and in Obese II the probability is: 0.00026978257976639737
- If you are **male** and in Obese III the probability is: 0.0001437095151666764
- If you are **female** and in Obese III the probability is: 0.04753078541702529

Appendix II

- Legend:
  - 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)
  - Calories consumption monitoring (SCC)
  - Physical activity frequency (FAF)
  - Time using technology devices (TUE)
  - Transportation used (MTRANS)
  - other variables obtained were: Gender, Age, Height and Weight.

Data Sources:

Downloaded from here

Click here for survey questions and how they were asked

Appendix III

How the survey questions were asked for this dataset:

Source: Here

Table 2: Dataset description

Attributes	Values
Sex	H: Male M: Female
Age	Integer Numeric Values
Height	Integer Numeric Values (Mt)
Weight	Integer Numeric Values (Kg)
Family with overweight / Obesity	Yes No
Fast Food Intake	Yes No
Vegetables Consumption Frequency	S: Always A: Sometimes CN: Rarely
Number of main meals daily	1 to 2: UD 3: TR More than 3: MT
Food intake between meals	S: Always CS: Usually A: Sometimes CN: Rarely
Smoking	Yes No
Liquid intake daily	MU: Less than one liter UAD: Between 1 and 2 liters MD: More than 2 liters
Calories Consumption Calculation	Yes No
Physical Activity	UOD: 1 to 2 days TAC: 3 to 4 days COS: 5 to 6 days NO: No physical activity
Schedule dedicated to technology	CAD: 0 to 2 hours TAC: 3 to 5 hours MC: More than 5 hours
Alcohol consumption	NO: No consumo de alcohol CF: Rarely S: Weekly D: Daily
Type of Transportation used	TP: Public transportation MTA: Motorbike BTA: Bike CA: Walking AU: Automobile
IMC	WHO Classification
Vulnerable	Based on the WHO Classification