Obesity Classification

Using Machine Learning Models

Team

- Chaitanya Shekhar Deshpande
- Swet Shah
- Vinayaka S Gadag
- Abhinav Kumar
- Nicholas Faro

CSCI-P556
Professor. Donald Williamson
Associate Instructor. Junyi Fan

Agenda

- Understanding the topic
- Dataset description
  - Data source
  - Variables or features
- Research questions
- Approach used for the analysis
- Machine Learning Models
- Future Scope and Challenges
- Conclusion
- Questions?
- References

Understanding the topic

- Obesity is abnormal or excessive fat accumulation that may impair health.
- In 2008 over 1.4 billion adults that aged 20 and older were overweight.
- 65% of the world's population live in countries where obesity kills more people than underweight.

Dataset description

- The dataset has been pulled from UCI Machine Learning repository
- We have total of 2111 instances and 17 attributes, and the eating habit related attributes 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)

Dataset description

- The attributes related with the physical condition are;
- Calorie's consumption monitoring (SCC)
- Physical activity frequency (FAF)
- Time using technology devices (TUE)
- Transportation used (MTRANS)
- And the other variables obtained were Gender, Age, Height and Weight
- Output variables: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III

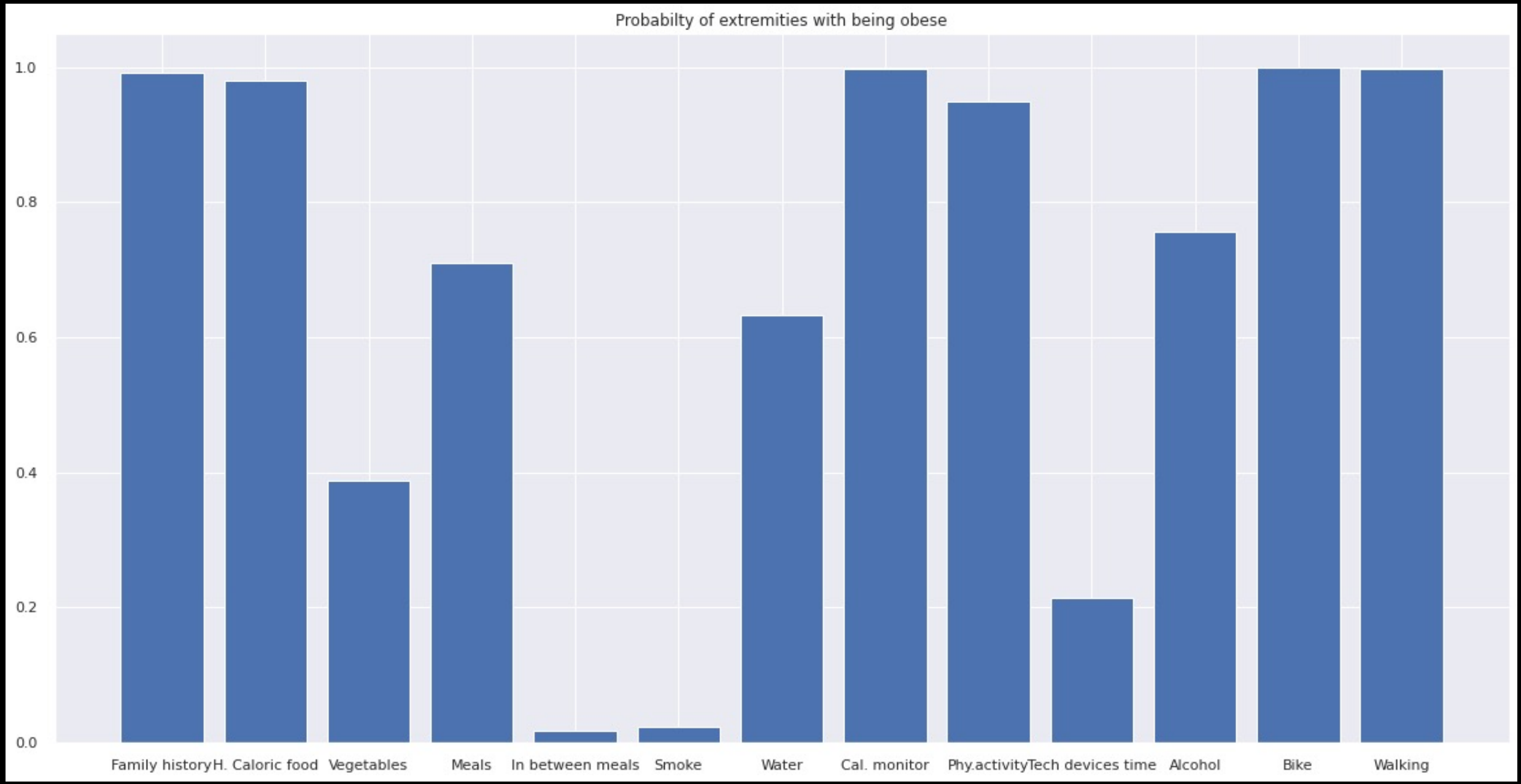
References and previous work

- Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico: https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub.
- C. Davila-Payan, M. DeGuzman, K. Johnson, N. Serban, J. Swann: Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data
- M.H.B.M. Adnan, W. Husain: A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction: Computer & Information Science (ICCIS), 2012 International Conference on, vol. 1, IEEE (2012, June), pp. 281-285
- Source Data Set: https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+

Research questions

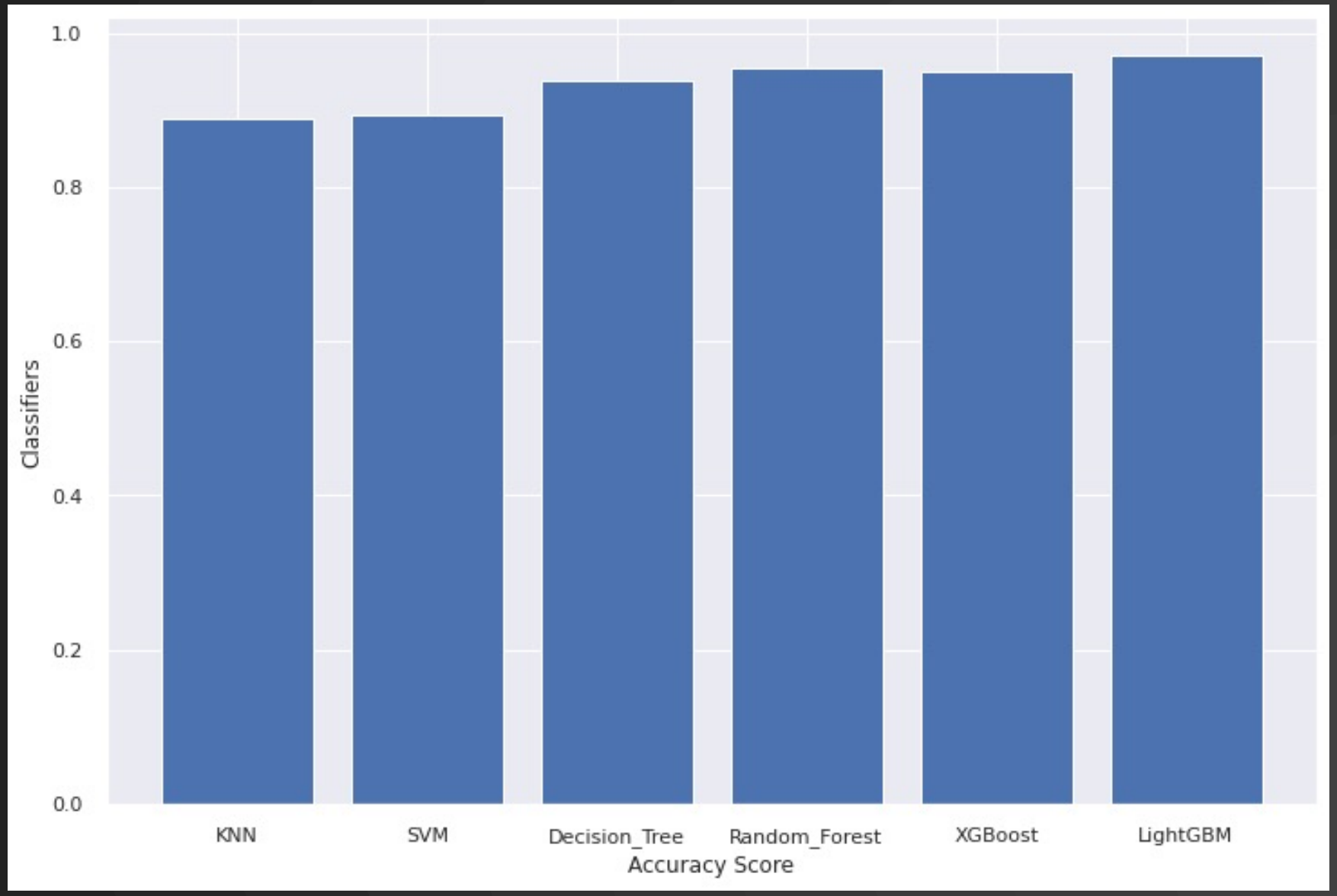
- Prediction of obesity level based on all attributes using multiclass classification
- Estimation of weight in kgs based on all eating and physical attributes using regression.
- Health recommendation system for different age groups.
  - 14-21
  - 22-30
  - 30+

Probabilistic Analysis for Extreme Values in Attributes for being Obese

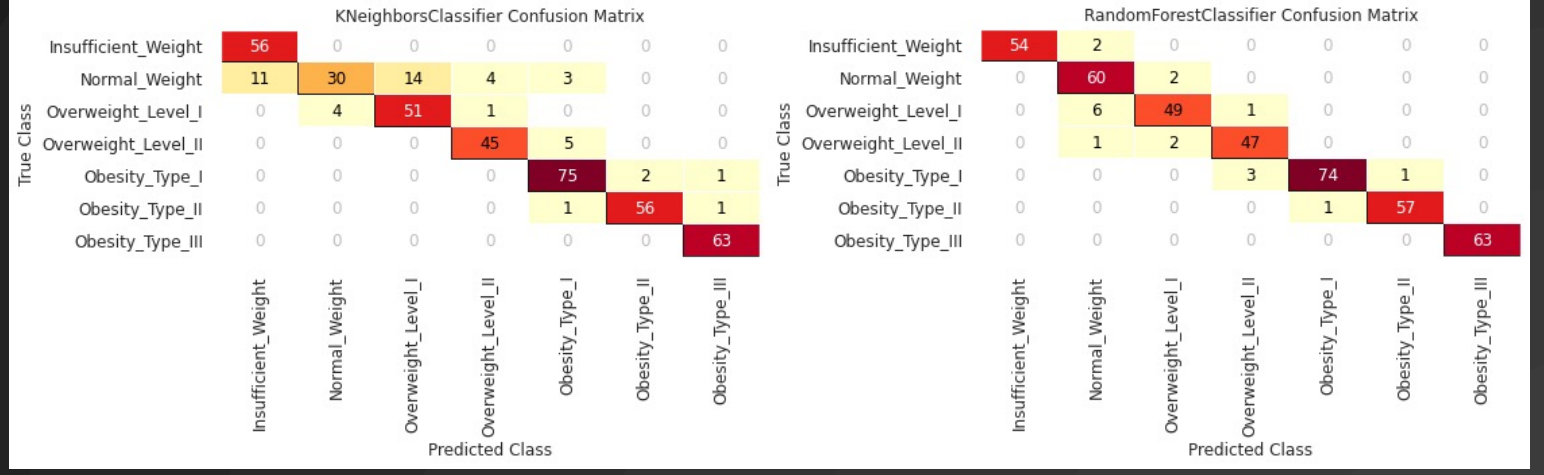


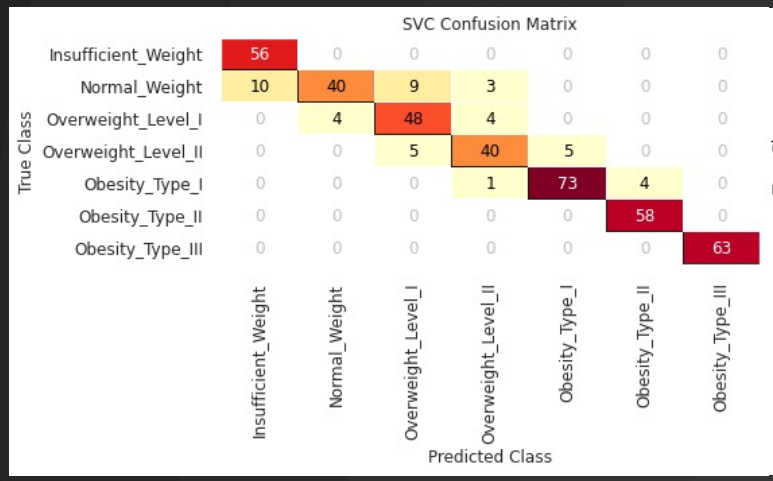
Estimation of obesity level using multiclass classification

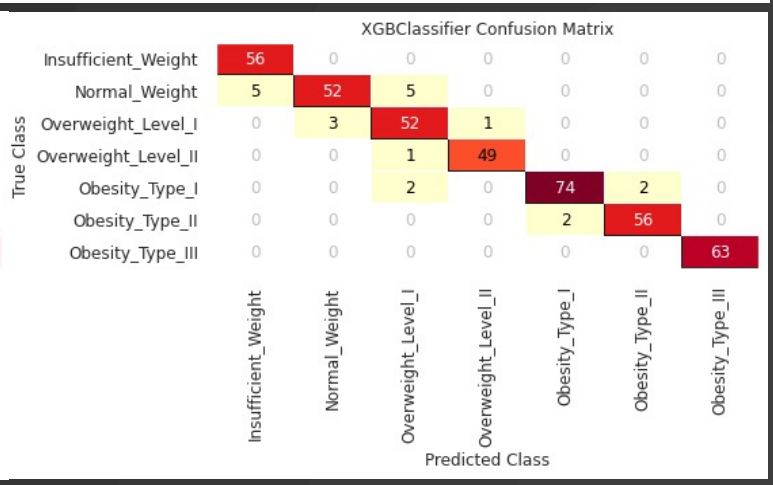
- As we have seven different obesity levels, we had to use multi class classification.
- We are using six different classification models,
  - K Nearest Neighbor
  - Support Vector Machine
  - Decision Trees
  - Random Forest
  - XG Boost
  - Light Gradient Boosting Machine

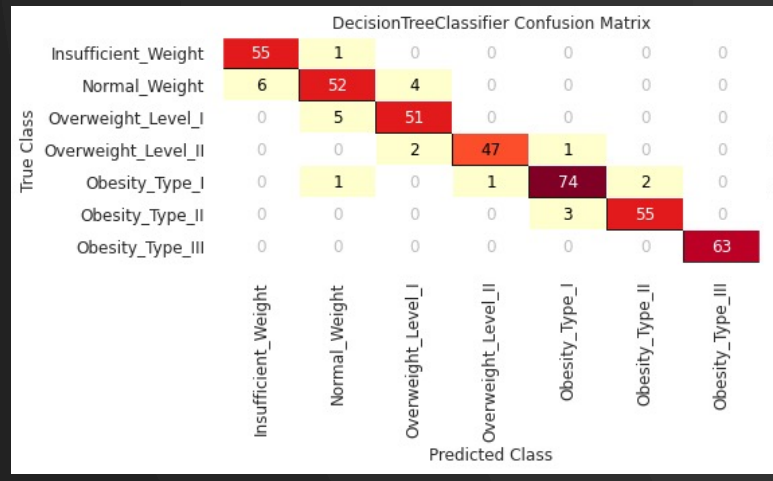


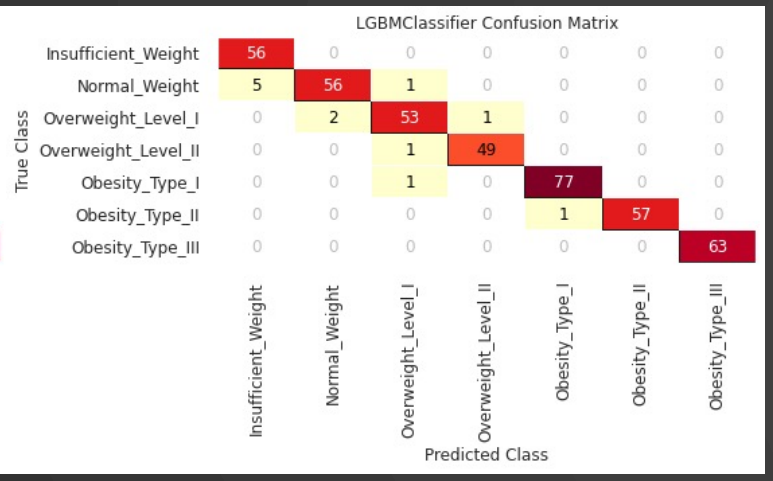
Accuracy summarized for all Models



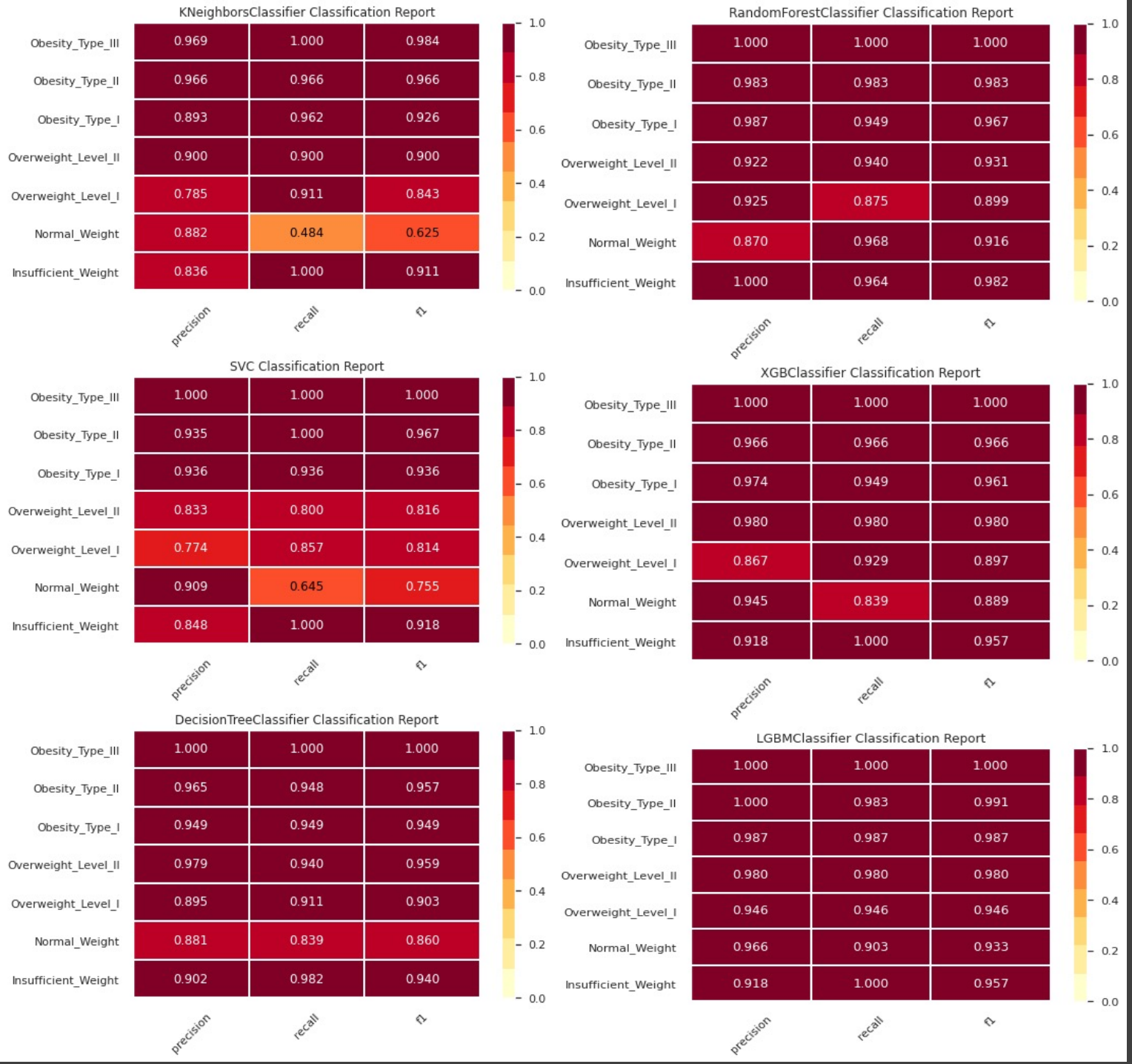






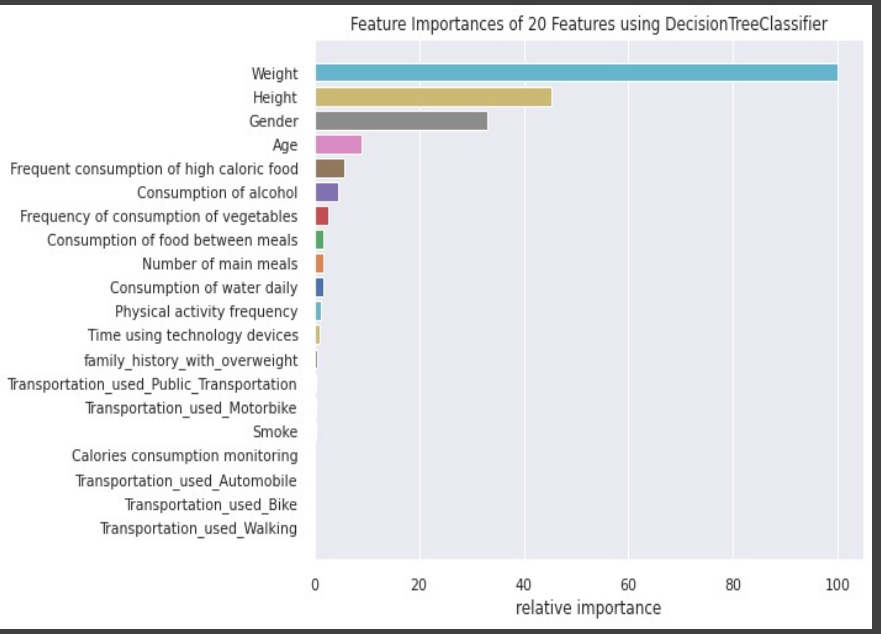


Confusion Matrix for all classifiers

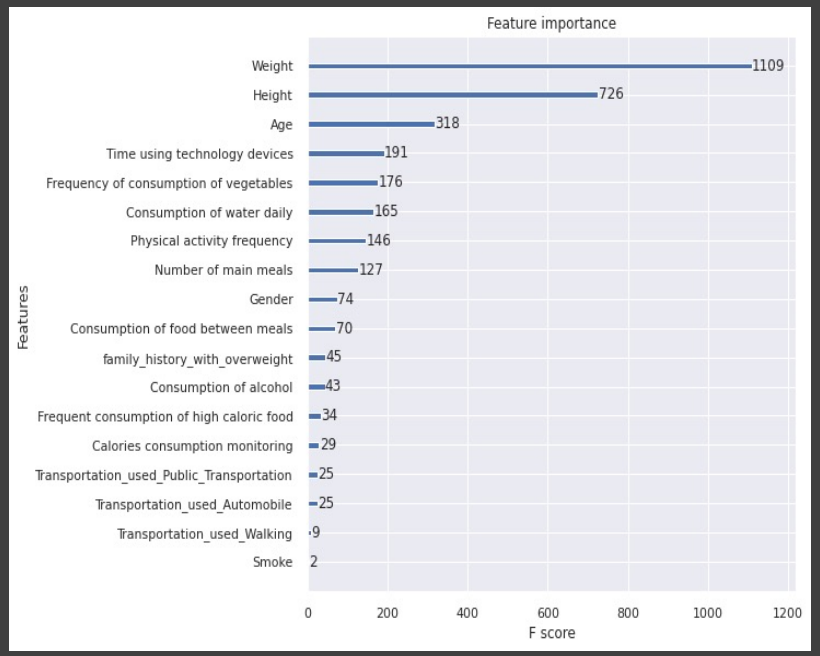


Classification report for all classifiers

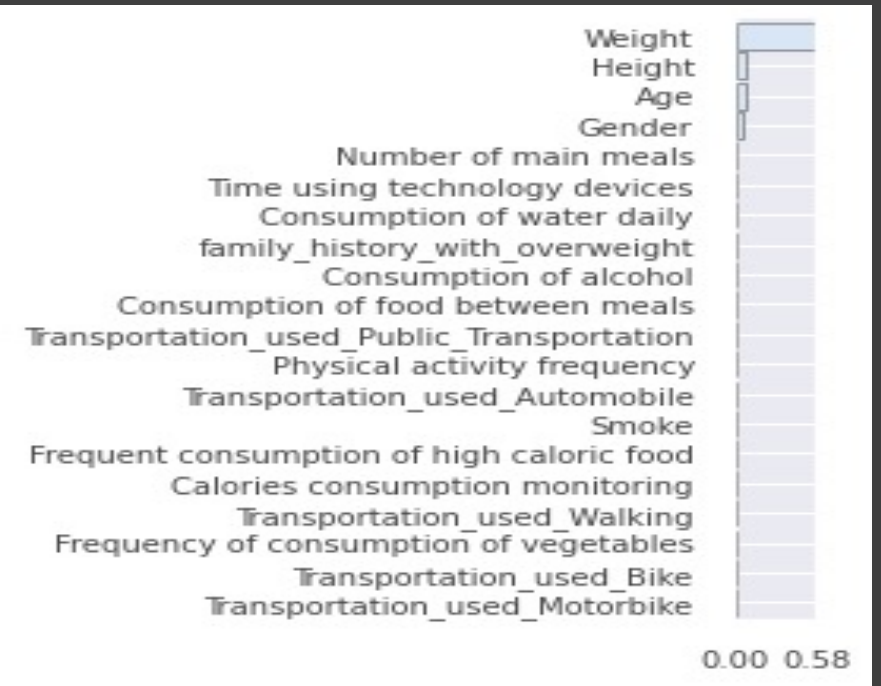
Decision Tree



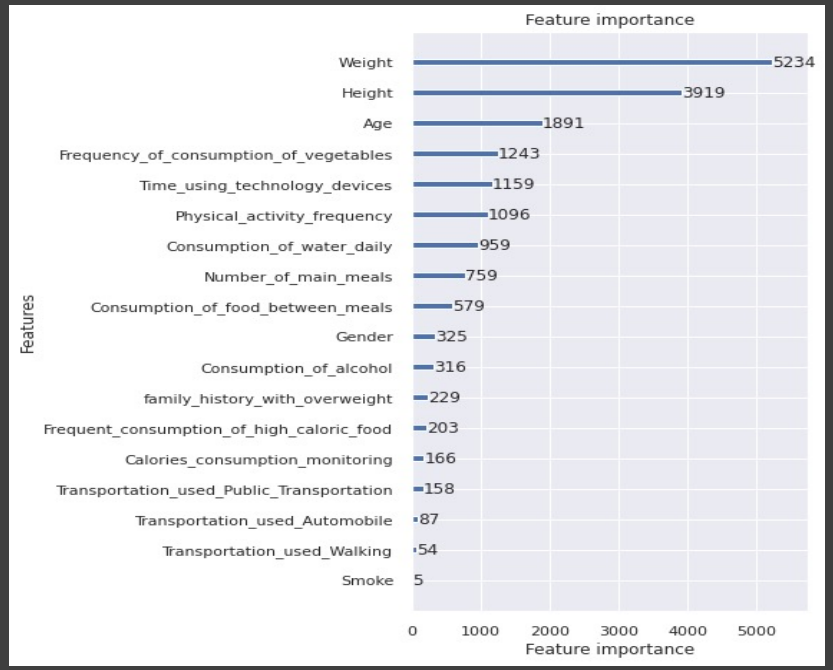
XG Boost



Random Forest



Light Gradient Boosting Machine

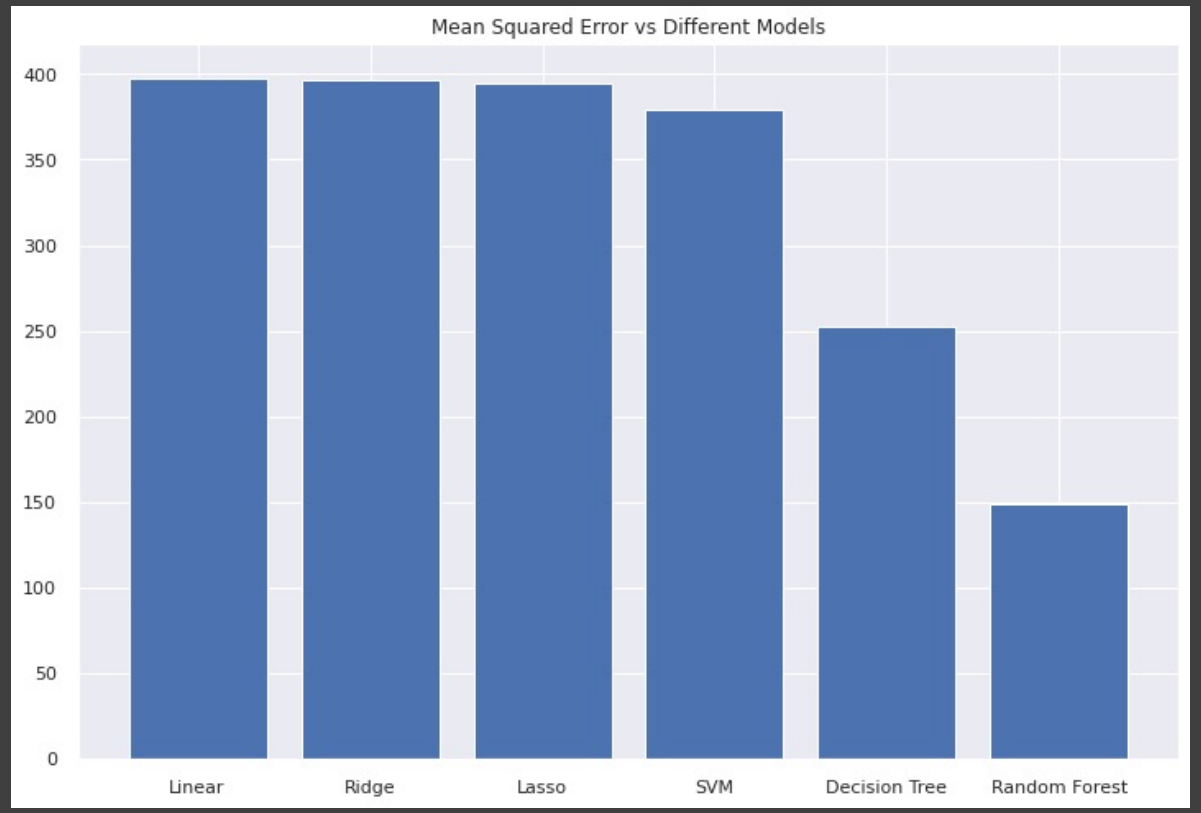


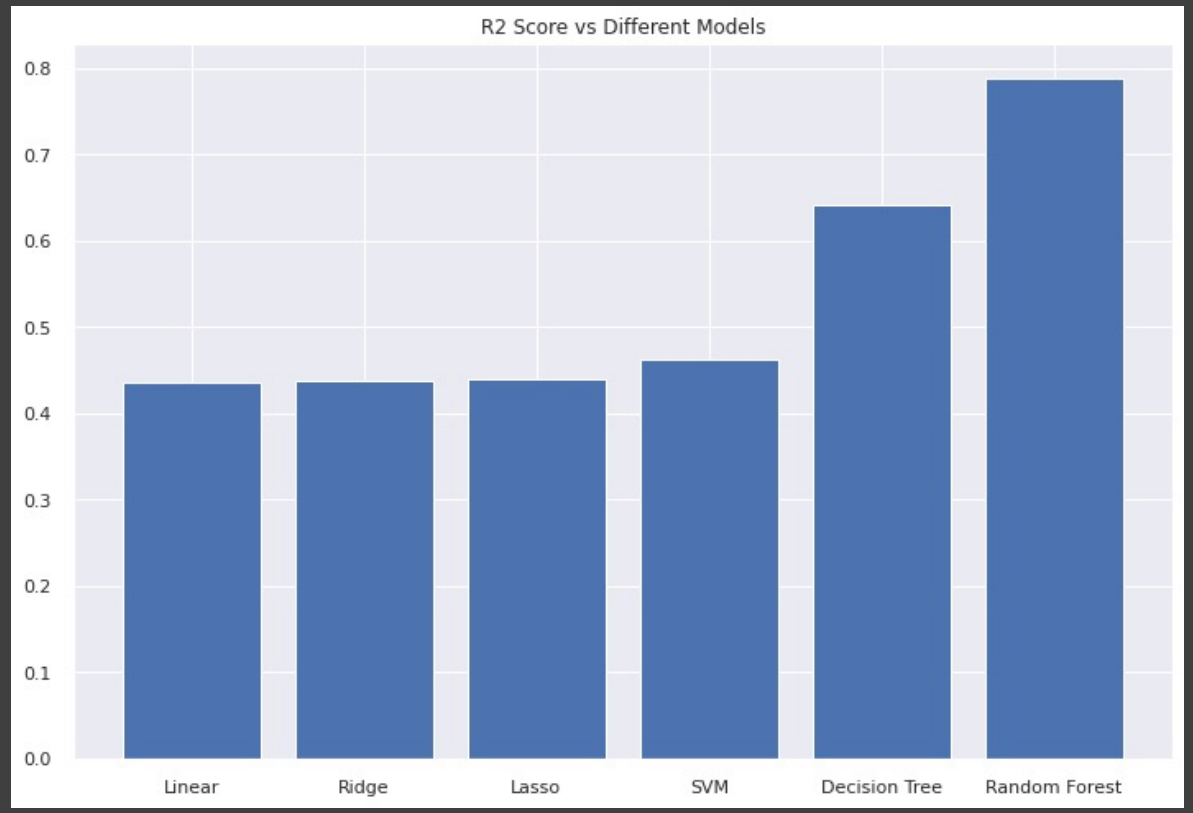
Feature Importance for all classifiers

Estimation of weight using Regression Model

- This research question focusses on estimation of weight using all the eating habits and daily lifestyle habits.
- Goal is to estimate a person’s weight in kgs given his daily habits not dependent on the gender.
- List of regression models used are,
  - Linear
  - Ridge
  - Lasso
  - SVM
  - Decision Tree
  - Random Forest

Estimation of weight using Regression Models Summarized

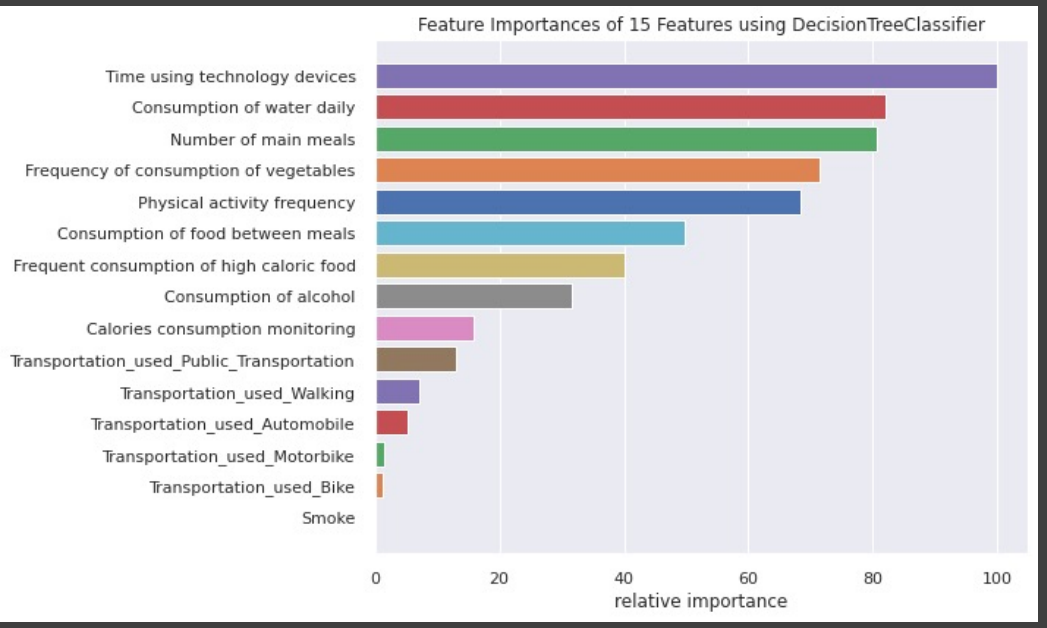




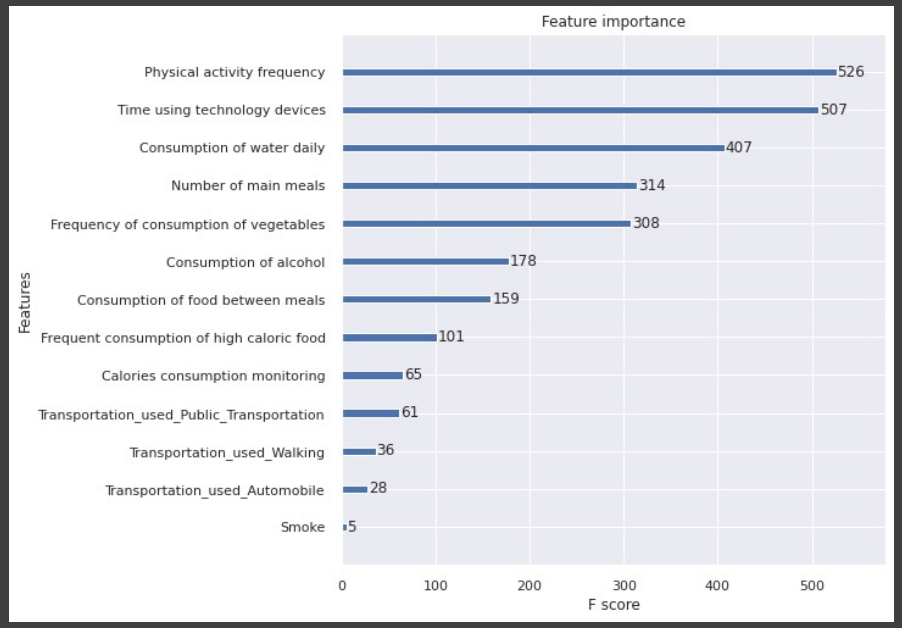
Health Recommendation System for different age groups.

- A health recommendation system basically recommends people of 3 different age groups what should be done so that they will stay in normal obesity level.
- One big challenge that we are facing for this model is that as small size of dataset. We only have around 300 people with normal weight in our dataset, but we will be evaluating the average and high frequency values of what each of these persons is doing.
- Here we will be only using the criteria which are in humans' control, like Calorie consumption, Alcohol, Smoking, Physical Activity Frequency, Number of meals, etc.
- We will be using the feature importance module to recommend people what things need to be focused with high priority as they affect obesity the most.
- Initially we will drop columns such as Age, Gender, Weight, Family History with Overweight and return the accuracy and rerun the feature importance for all 4 classifiers.: Decision Tree, Random Forest, XG Boost and Light GBM on the new dataset.

Decision Tree

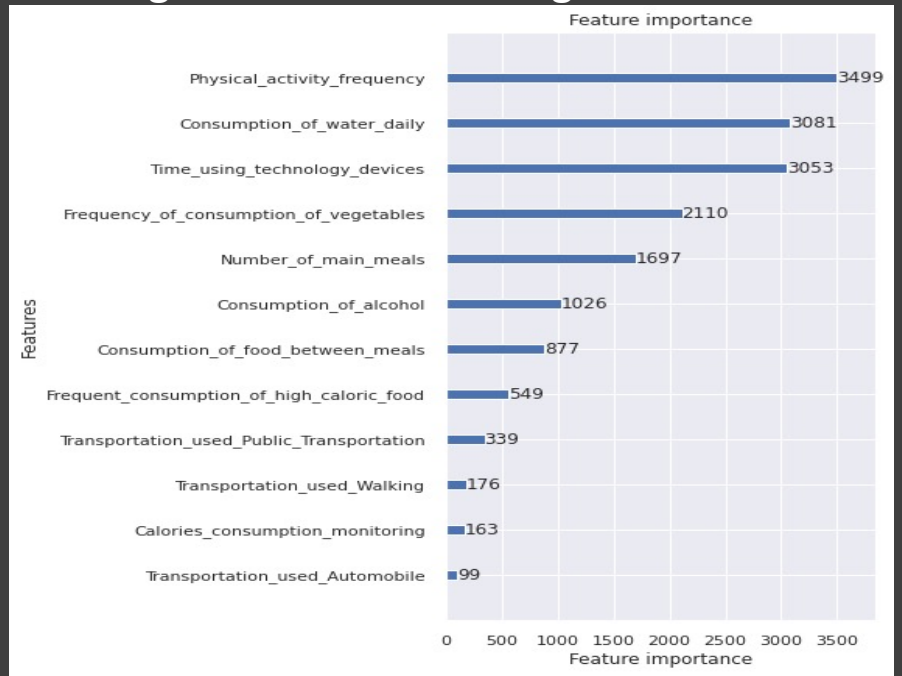


XG Boost



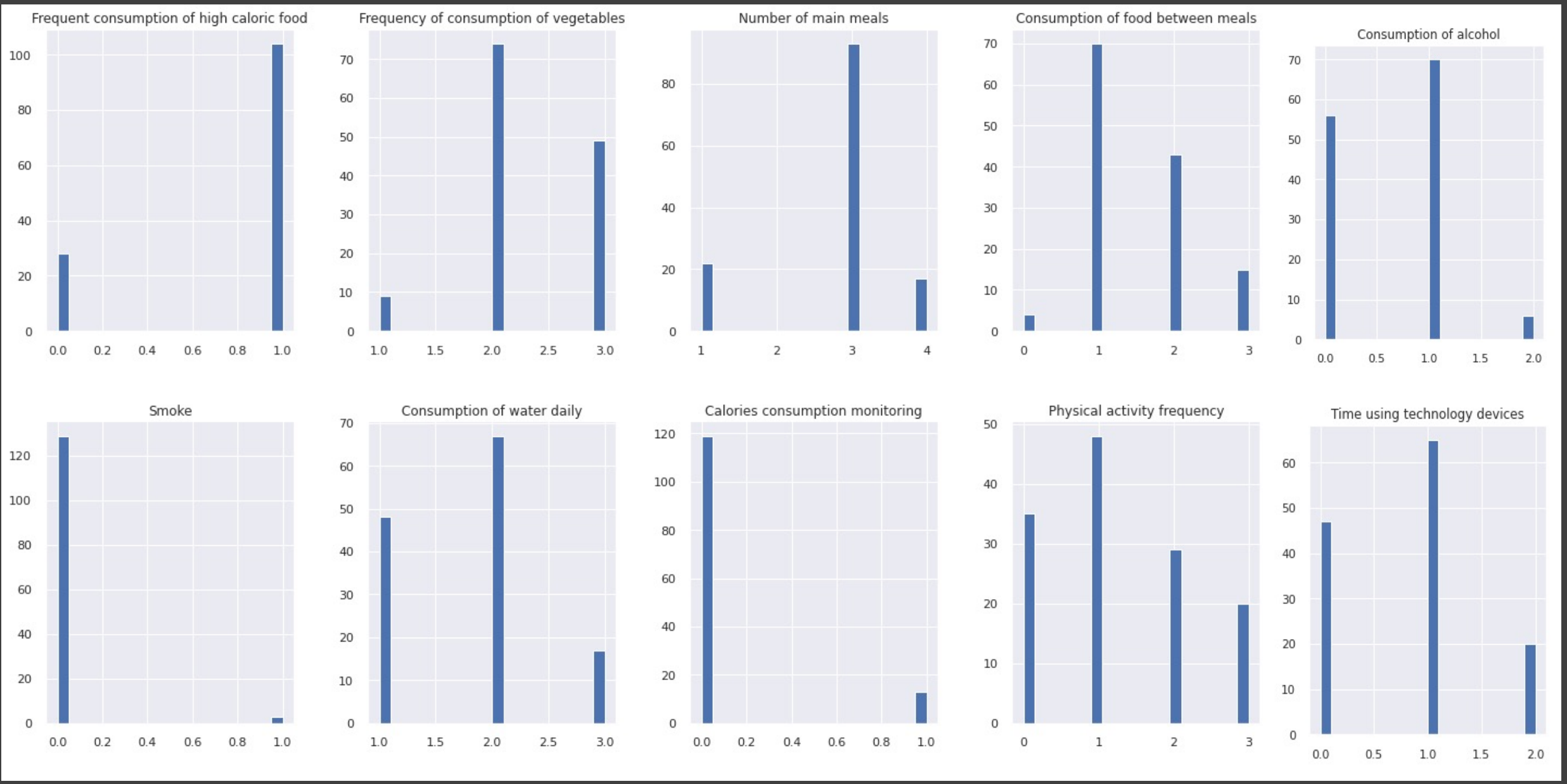
Random Forest

Light Gradient Boosting Machine

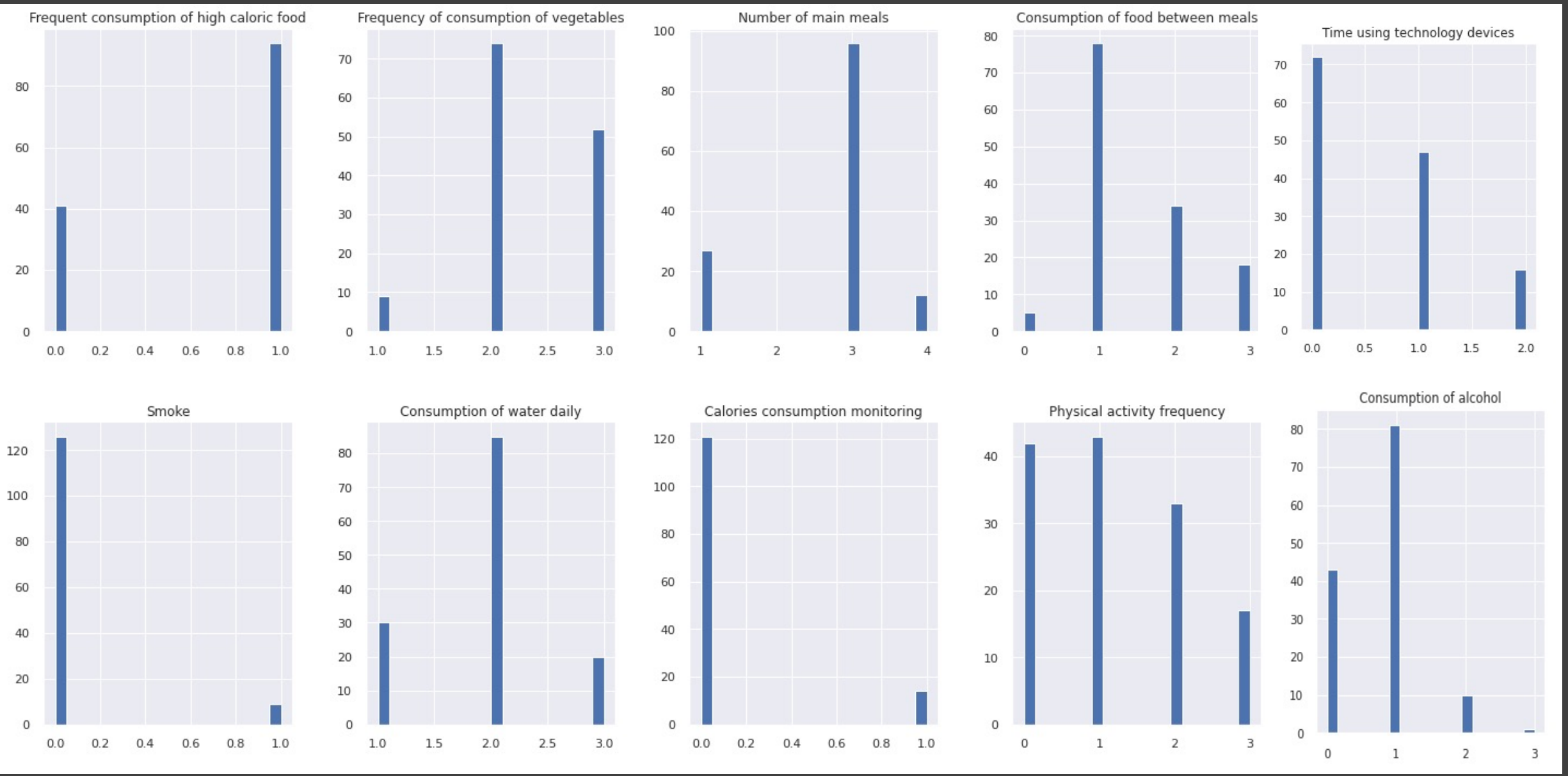


Feature Importance with only lifestyle attributes

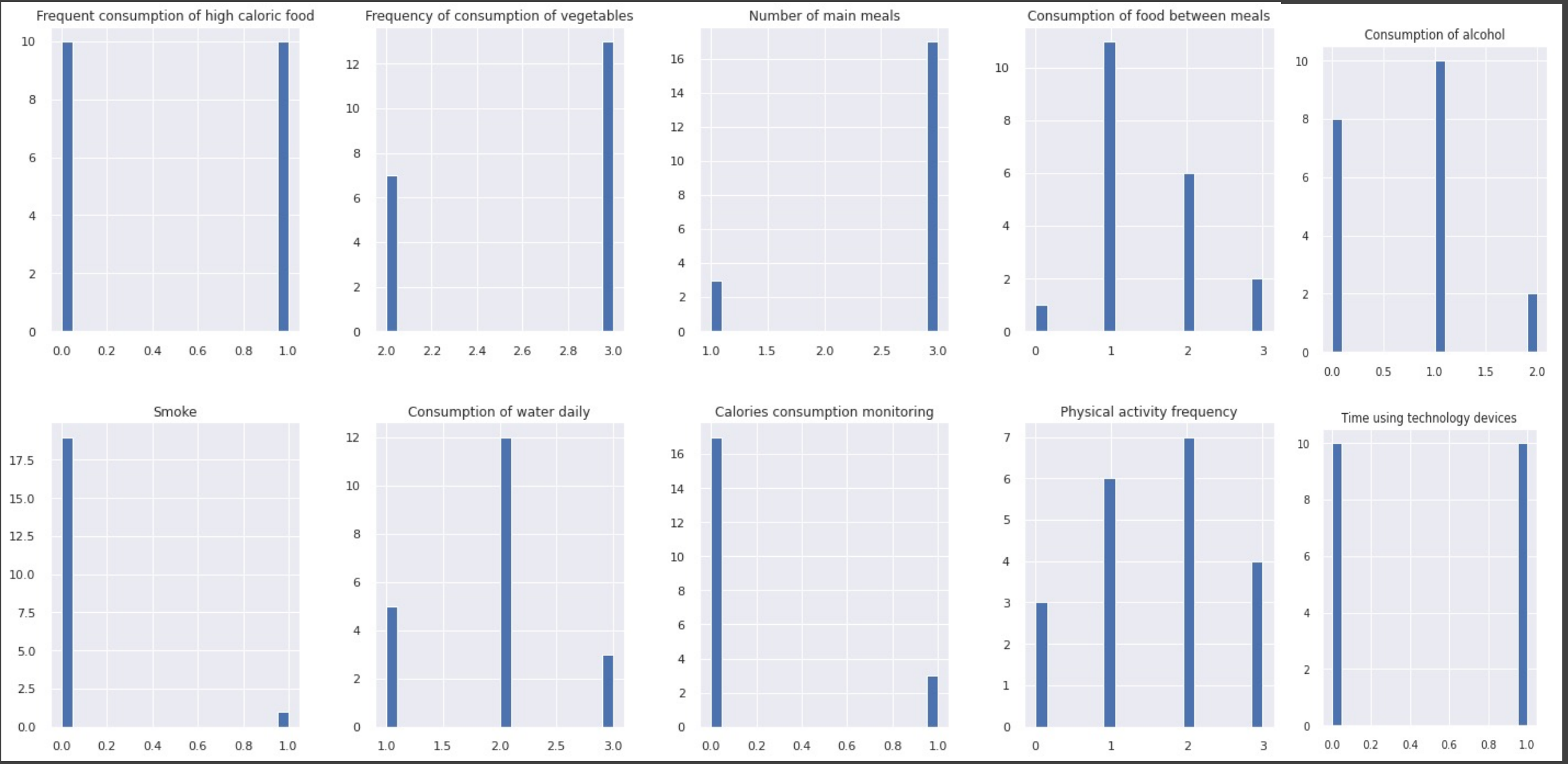
Plots for the important features Age Group 1



Plots for the important features Age Group 2



Plots for the important features Age Group 3



Conclusion

- Thus, we were able to successfully answer the 3 Research Questions that we had placed.
- The best Machine Learning algorithms we used were:
- Light Gradient Boosting Machine for Obesity Level Classifier with accuracy about 97%
- Random Forest Regressor with MSE value of about 150.
- We were also able to identify the features that were largely responsible for high obesity and were also able to suggest/recommend what different age group people should do such that they stay in the Normal Weight Category.
- The best thing we loved about this project was that it tries to solve a big problem that we face in our day-to-day life. With more data, we could try and improve our Health Recommendation system to provide us with more accurate outputs.

Challenges Faced and Future Scope

- The size of our dataset(2111 people), might have made it a bit difficult and slightly inaccurate to give perfect health recommendation.
- With small dataset, implementation of Neural Networks was not possible as it is likely to give very less accuracy.
- If the survey of taking health details increases its size, it will help us a lot to implement a more robust recommendation system.

THANK YOU