PREDICTING THE LEVELS OF OBESITY USING CLASSIFICATION, ESTIMATING WEIGHT USING REGRESSION AND HEALTH RECOMMENDER SYSTEM BASED ON DIETARY AND LIFESTYLE HABITS

Abhinav Kumar, Chaitanya Shekhar Deshpande, Nicholas Faro, Swet Shah and, Vinayaka S Gadag
Indiana University, Bloomington

ABSTRACT

Right now, obesity is abnormal or excessive fat accumulation which has several health hazards in our day-to-day life. So, to avoid this problem we can think of designing a system that would be able to identify and estimate the obesity level of a person, given certain characteristics, dietary habits, and day-to-day lifestyle. Considering this problem, we came with the machine learning approach to estimate the obesity level of a person using multiclass classification. We will also be estimating a person's weight in kgs using regression models. Along with this, we have also implemented a health recommendation system for different age groups based on the obesity level estimation dataset that we have chosen from the UCI machine learning repository. With this project, we aim to tackle this global concern.

Index Terms— Multiclass classification, Regression, probabilistic analysis, Ensemble learning.

1. INTRODUCTION

Obesity is a common problem but often ignored. A person can be classified as obese if the weight is 20% more than the ideal weight. In 2008 over 1.4 billion adults aged 20 and older were overweight. More than 65% of the population in the world live in places where obesity is more severe than being underweight. Obesity is a disease with multiple factors; its main features are unwanted fat accumulation and weight increase. This causes the person to invite the diseases like diabetes, cardiovascular problems, and even death. The research till now shows that following diet and weight loss programs have reduced obesity. But we must balance it against the future regain. Considering this, it is important to set the long-term daily habits, which we have tried to come up with the recommendation system.

In this project, we have implemented the different machine learning algorithms which address the problems and recommend the daily habits to follow.

2. PREVIOUS WORK DONE

There have been multiple studies and research done on Obesity classification as mentioned in the links to the paper in the references section [8]. These papers try to address the causes and effects of Obesity, but none of them recommended the healthy habits a person should follow. We have implemented a model in which apart from addressing the problem of Obesity classification, we are recommending the healthy habits a person must follow according to their age so that they have a healthy lifestyle and do not fall prey to Obesity. We have also implemented a regression model to estimate the Weight in kgs of a person based on the Dietary and Lifestyle habits.

3. DATASET DESCRIPTION

The dataset has been pulled from the UCI machine learning repository. It has a total of 2111 instances and 17 attributes which has been taken across 3 countries: Columbia, Peru and Mexico. There are eating habits, physical condition related, and other attributes. Namely [1], Frequent consumption of high caloric food (FAVC), Frequent consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), Consumption of alcohol (CALC), Calorie consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS), and other attributes are Gender, Age, Height, and Weight. Apart from this, there are output variables, Insufficient Weight, Normal Weight, Overweight Level 1, Overweight Level 2, Obesity Type 1, Obesity Type 2, and Obesity Type 3 [1]. The data contains numerical data and continuous data, so it can be used for analysis based on machine learning algorithms of classification, probabilistic prediction, and regression.

4. EXPERIMENTAL SETUP

Initially, the dataset that we had, was not processed properly. Several issues were present in the dataset, for instance, there were several non-numerical values, which would make it difficult to evaluate and train our machine learning model. The attributes where we required data preprocessing were family_history_with_overweight, Frequent consumption of high caloric food, Calorie's consumption monitoring, Smoke, Consumption of food between meals, Consumption of alcohol, Gender, Transportation used, and of course Obesity Level. For features like family_history_with_overweight, Frequent consumption of high caloric food, calorie consumption monitoring, Smoke, we have only 2 possibilities: Yes and No. So, we could simply assign them with values of 1 for Yes and 0 for No. For attributes: Consumption of food between meals and Consumption of alcohol, we have attribute values as: Always, Frequently, Sometimes, and No. Now here we cannot use One Hot Encoding, because each of the above attributes has a level-based difference. So, we need to assign values like 3 for Always, 2 for Frequently, 1 for Sometimes, and 0 for No. This can be achieved through Ordinal Encoding. So, we use the OrdinalEncoder() library from sklearn.preprocessing to assign these values. By using the factorize function with sort=True, we get priority-based values. We have also used Ordinal Encoding for Obesity levels, i.e., 0.0 for Insufficient Weight, 1.0 for Normal weight, 2.0 for Overweight Level I, 3.0 for Overweight Level II, 4.0 for Obesity Type I, 5.0 for Obesity Type II, and 6.0 for Obesity Type III. For the remaining attribute- Transportation Used, we have used One Hot Encoding, as these values cannot be prioritized in any order. For all 3 Research questions, we have split the dataset into 80% for Training Dataset and 20% for Testing Dataset. After doing some exploratory data analysis, we came to know that one of the positives of our dataset is that the number of values for each class of obesity level is more or less similar in range. Such normally distributed labeled data helps us to provide more accuracy in predicting our output. We also came to know that Height and Weight attributes are well distributed in our dataset. Coming towards our final research question, to ensure a more spread-out and uniform division of age groups, we have classified them into 14 to 21 years, 21 to 30 years, and 30+ years. This kind of division will help us give more accurate outputs. After performing some analysis using box plots and lmplot, it can be seen that the average weight of the Male gender is generally higher than the Female gender, whereas the average height is more or less similar. So, we need to be careful in designing the BMI levels for these 2 genders. So generally, for 2 similar weight and height values, the BMI for the Female Gender will higher than the Male Gender. We have also implemented the Probabilistic Analysis for the Extreme Values in Attributes for Being Obese.

In this analysis, we have taken all the people for the dataset which are in Obese Categories: I, II and III, and then we have taken the extreme values for each attribute initially, like Yes for Family History with Overweight, Smoking towards the negative side. For instance, Smoking is set to “Yes” as it shouldn’t be done, whereas Calorie Consumption Monitoring has been set to “No”. However, the problem with this analysis is that it largely depends on the number of extremities in the data. For instance, if we consider Family History with Overweight, most of the samples present in the dataset have a value of "Yes", so it automatically gives an extreme probability of 99.17%. Also, the number of samples where the transportation used as Biking is very low, so we can't provide any firm conclusion toward our prediction with this, but we do get a fair idea.

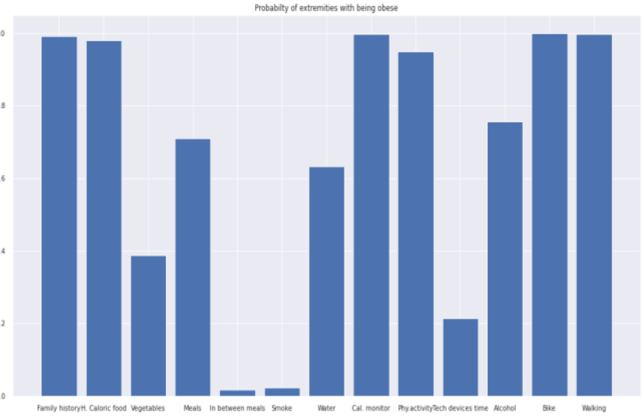

Figure 1 Probabilistic Analysis for Extreme Values in Attributes for being Obese

5. APPROACH TOWARDS SOLVING THE PROBLEM

We have used different approaches to address the Obesity classification problem and the recommendation system for healthy habits. Namely, Multiclass Classification (Decision Tree classifier, Random forest classifier, Boosting algorithms) Estimating weight in Kgs using regression (Linear, Ridge, Lasso (alpha = 0.1), Support Vector Machine, Decision Tree and Random Forest Regressors.), Health recommender system (Decision Tree, Random Forest, XGBoost, and LGBoost).

5.1. MULTICLASS CLASSIFICATION TO ESTIMATE OBESITY LEVEL

Our 1st Research Question aims towards Estimating the Obesity Level of a person given certain information of his/her dietary habits and lifestyle using Multiclass Classification. We have used 6 different algorithms to achieve this. We began with K Nearest Neighbor Classification using the value of K as 5. This gave us an accuracy of 88%, which is decent, but when we discuss about a health concern, it is essential to have more accuracy. Next, we went with Support Vector Machine Classifier with Linear Kernel. This only give us a slight increase in accuracy of 89.36%. So, we expanded further by implementing Decision Tree Classifier, which gave us 93.85% accuracy. Then, we implemented Random Forest classifier, which gave us a substantial increase in accuracy by giving 95.5% probably due to the random nature of building the Decision Trees. The XGBoost Algorithm creates a weak tree and then boosts the subsequent trees to reduce the residual errors. It also to capture and address any patterns in the errors until they start looking as if they are random. They also have great Runtime efficiency, help to prevent overfitting and also most notably help us with Feature Importance. Use of GPUs can enhance its performance. For XGBoost Algorithm, we got 95% accuracy. The Light

Gradient Boosting Machine Algorithm uses a sampling mechanism to deal with continuous values, which helps us creating trees more quickly than XGBoost algorithm and also helps us with lesser memory usage. The LGBM also grows trees depth first, i.e. leaf wise and not level wise. We use num_leaves=30 parameter to control overfitting. Use of GPUs will help in increasing the performance of our model. With Light Gradient Boosting Machine Algorithm (LGBM), we got the best accuracy with 97.16%.

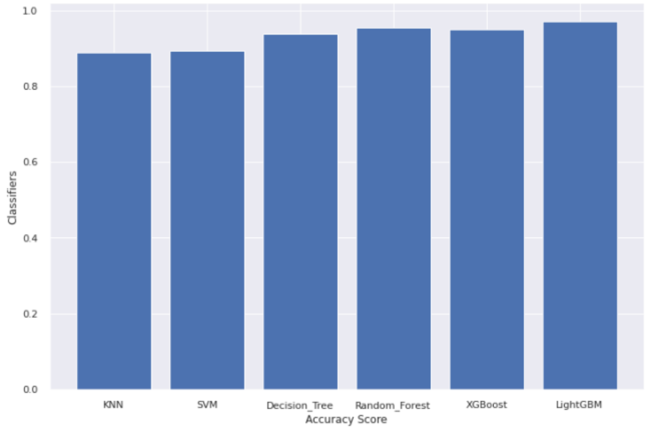

Figure 2 Accuracy Summarized for all models

Thus, we could finalize the algorithm for implementation by selecting LGBM classifier. As Decision Tree, Random Forest, XGBoost and LGBM Classifiers all gave more than 93% accuracy, we would also be evaluating the feature importance's for each of the features to identify the feature that is largely responsible for deciding the type of obesity. This would mean that the feature-importances with higher values will be suggesting that those particular attributes are the key factors towards identifying the obesity level of a person.

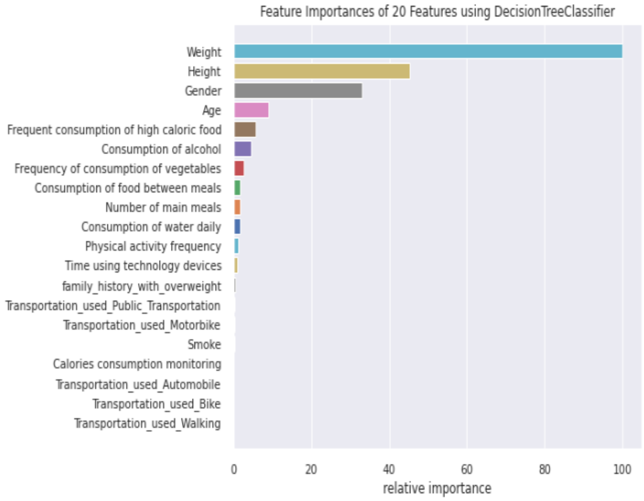

Figure 3 Decision Tree

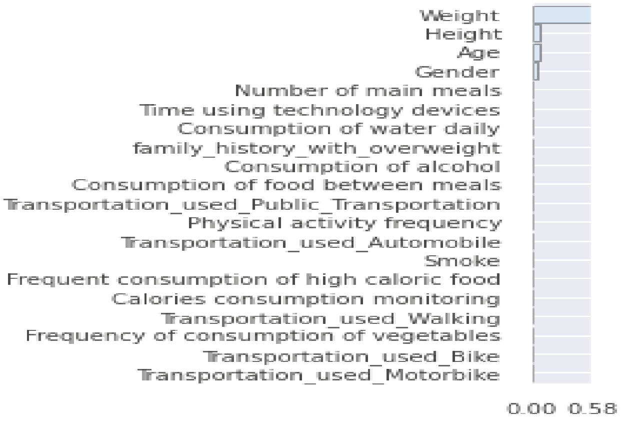

Figure 4 Random Forest

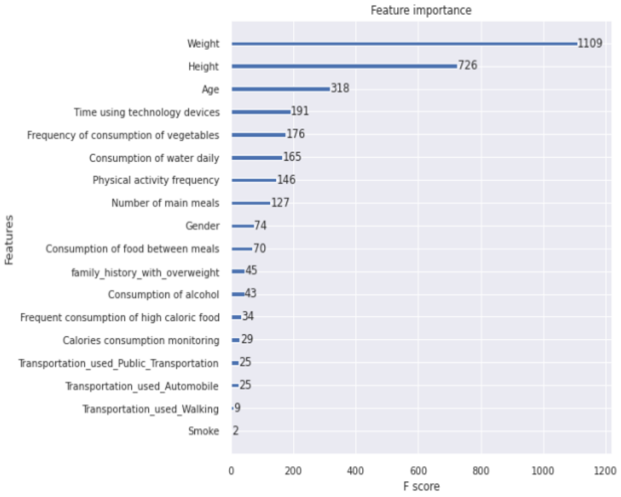

Figure 5 XG Boost

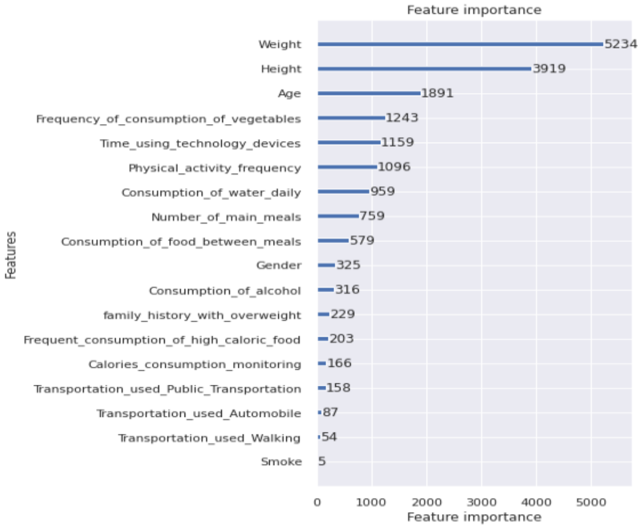

Figure 6 Light Gradient Boosting Machine

Obviously, for BMI, Height, Weight and Age will be the most important attributes, but we are more interested in answering the important attributes which are depended on our regular lifestyle. Thus, we can conclude from the 4 different graphs that the key features towards indicating obesity from our lifestyle habits are: Time using Technology Devices, Frequency of Consumption of Vegetables, Consumption of Water Daily along with Weight, Height and Age. So along with high classification accuracy and identifying the key factors towards obesity, we were able to solve or first research question.

5.2. ESTIMATING WEIGHT IN KGS USING REGRESSION

Our 2nd Research Question aims at Estimating Weight of a person based on the dietary habits, family history with overweight and lifestyle of a person using Regression Models. So, we have not included Height, Obesity Level, Age and Gender of a person for this model. We have used 6 different Regression models: Linear, Ridge, Lasso (alpha = 0.1), Support Vector Machine, Decision Tree and Random Forest Regressors.

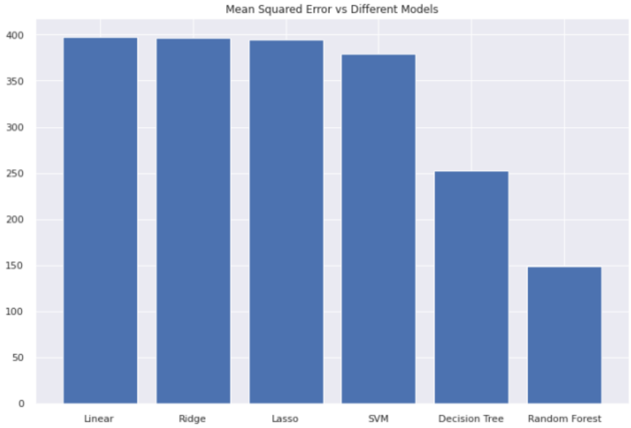

Figure 7 Mean Square Error VS Different regression models

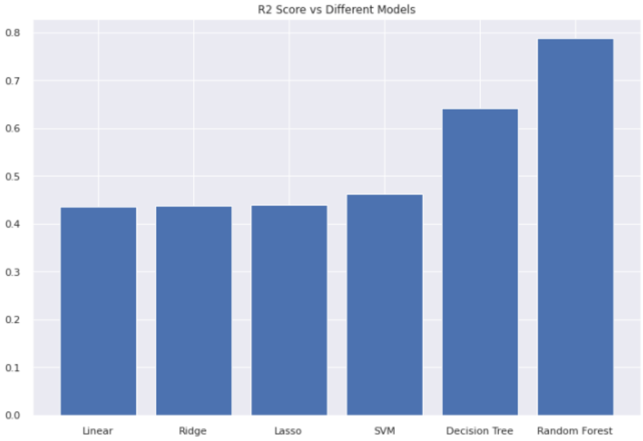

Figure 8 R2 Score VS Different Models

After Training the dataset, we then estimate the Weight value using the above models. We have also plotted the Mean Squared Error for all the models, and it turns out that the MSE values for Linear, Ridge, Lasso and SVM is high (around 400) and for Decision Tree it reduces to around 260, and for Random Forest it is the lowest, giving 150. Also, we have plotted the regression coefficients for each of these regressors, and it turns out that it is the highest for Random Forest Regressor. Thus, to estimate the Weight of a person using his/her lifestyle habits, we use the Random Forest Regressor Model for high accuracy estimation.

5.3. Health recommender system

Our 3rd Research Questions aims at recommending dietary and lifestyle habits for 3 different age groups: 14 to 21, 21 to 30 and 30+, based on Multiclass Classification and Feature-Importance's, based on only the characteristics which are in our control, i.e., the features: Weight, Height, Age, Gender and Family History with Overweight have been removed from our dataset. So here we will simply be focusing on the attributes: Frequent consumption of high caloric food, Frequency of consumption of vegetables, Number of main meals, Consumption of food between meals, Smoke, Consumption of water daily, Calorie's consumption monitoring, Physical activity frequency, Time using technology devices, Consumption of alcohol, Transportation used. Initially we run the Accuracies and Feature-Importance's for 4 classifiers: Decision Tree, Random Forest, XGBoost and LG Boost Machine learning algorithms, because these 4 classifiers gave higher accuracy in our previous models. This time we do not have the most important features: Height, Weight and Age, so our accuracy does decline a bit.

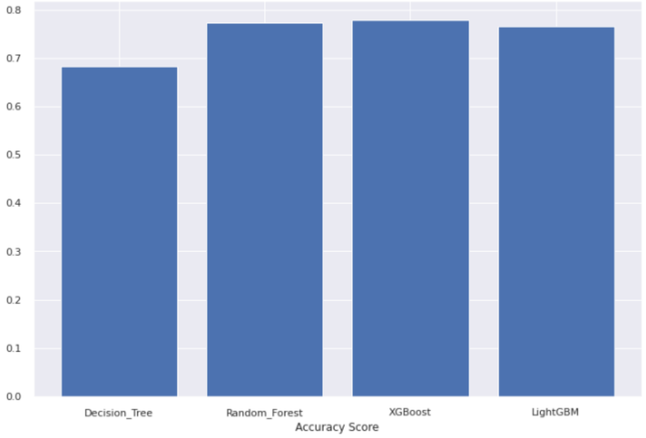

Figure 9 Accuracy score for classifiers only with dietary and lifestyle habits

We get highest accuracy for XGBoost algorithm with 77.9%, so we will be considering this classifier as the important one along with its feature-importance's. With XGBoost algorithm, we found out that Physical Activity Frequency, Time using Technology Devices, Consumption of Water Daily, Number of Main Meals, Frequency of Consumption of vegetables and Consumption of Alcohol are the most important features for this classifier. So, our task now is to identify the ideal values for the above attributes for people with Normal Weight for all 3 age groups.

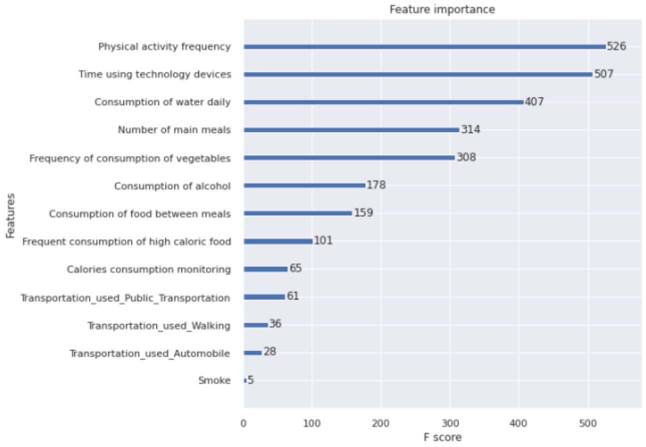

Figure 10 Feature Importances for XGBoost only with Dietary and Lifestyle habits as features

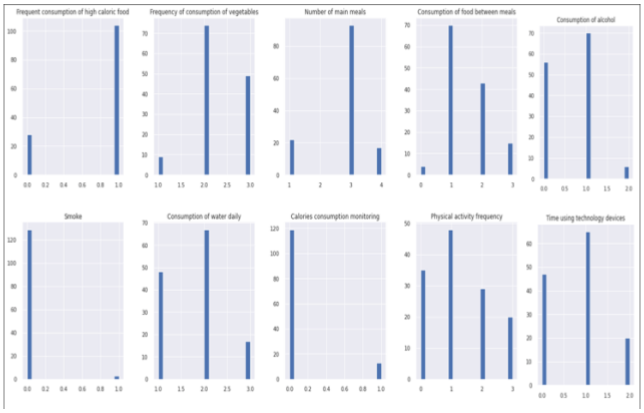

Figure 11 Plots for the important features Age Group-1

We have filtered out the 3 age groups into age1, age2 and age3 Data-Frames. After plotting the bar plots for each of the attributes, we are able to make certain conclusions on our research. For age group 14-21, we can observe that for Physical Activity Frequency, we have a maximum frequency of value as 1, i.e., a person of that group must often have 1 to 2 days of Physical activity [1]. The time using technology devices has maximum frequency of value 1, i.e., a person of that age group can have an average of 3 to 5 hours [1] of time on technology devices, but not more than that. Similarly, after analyzing all the plots, we can conclude that the Amount of Daily Water consumption must be at least 1 to 2 liters [1], Number of main meals must be 3 [1], as lesser meals can lead to Insufficient Weight, Frequency of Consumption of vegetables must be in category of "Sometimes" [1] and Consumption of Alcohol also comes under "Sometimes" [1] category to maintain Normal Weight for age group 14 to 21.

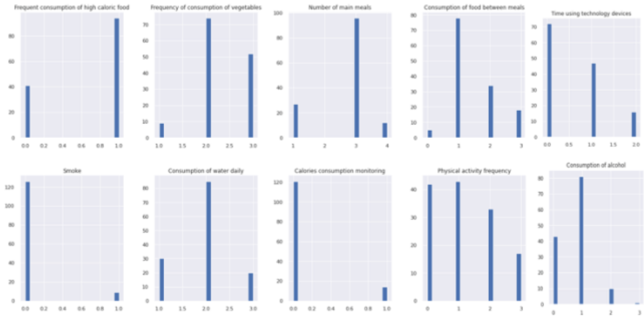

Figure 12 Plots for Important features Age Group-2

For age group 21-30 group, they can have Physical Activity Frequency of None or 1 to 2 days [1], time using technology devices must be lesser than above age group, i.e., 0 to 2 hours [1], Amount of Daily Water consumption must be at least 1 to 2 liters [1], number of main meals must be 3[1], Frequency of Consumption of vegetables and Frequency of Consumption of Alcohol must be in category of "Sometimes" [1]. It can be noticed that all the values for the above 2 age groups are quite similar, because in general, this age group belongs to Teenagers and Young Adults.

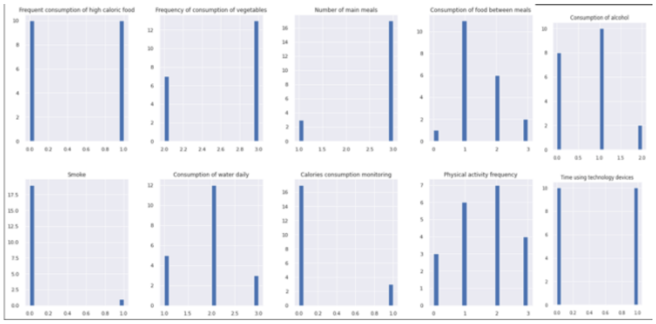

Figure 13 Plots for Important features Age Group-3

For age Group above 30 years, the Physical Activity Frequency is higher, with value of 2 to 4 days [1], Time Using Technology Devices also reduces to 0 to 2 hours [1], Amount of Daily Water consumption must be at least 1 to 2 liters [1], number of main meals must be 3 [1], Frequency of Consumption of vegetables increases to "Always" [1], Frequency of Alcohol is "Sometimes" [1]. We can conclude that the Vegetable consumption increases along with the Physical Activity Frequency to maintain Normal Weight, whereas the Time using Technology Devices reduces to 0 to 2 hours [1].

We can also draw certain conclusions on other attributes, even though they don't show with high importance value on the Feature-importance graph. For smoking, in almost all cases we have value of “No” [1] for Normal Weight, Consumption of Food between meals must be "Sometimes" [1], Calorie Consumption Monitoring doesn't matter much, Smoking has been set for 0 [1] for Normal Weight and Frequent Consumption of High Caloric Food can be allowed for age groups.

6. APPROACH SUMMARY AND FUTURE DIRECTIONS

Thus, we have successfully answered our 3 research questions. For Obesity Level Classification, we have used Light Gradient Boosting Machine Algorithm to produce 97.16% accuracy to estimate the Obesity Level of a person. Then we have also estimated the weight of a person in kgs using his/her dietary and lifestyle habits using Random Forest Regressor with Mean Squared Error value of around 150. And then finally, we were also able to implement a Health Recommender System to ensure that a person maintains himself/herself in the Normal Weight category for 3 different age groups.

For future directions, we can have more things which can be implemented. We could implement a Full-Stack project, where user can provide more details about different attributes. We could also differentiate our dataset based on each country, add new features like Stress levels for providing more details. With more entries in Dataset, we could also implement a more accurate health recommendation system, and could also go ahead with more advanced classifying and regression algorithms like Neural Networks, etc. With more uniform distribution of data, we can also categorize the age groups in our health recommender system from Kids to Old people in a more detailed manner, thereby giving more accuracy in our model.

7. CONCLUSION

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. Also, the future scope of this project mentioned before also makes it scalable and can further solve the recurring problem of obesity and help us to implement a more robust health recommendation system for different countries.

The initial approach only had estimation of Obesity level using classification, but we have implemented Gradient Boosting Algorithms to extend our accuracy to 97.16% which is really good. We were also able to estimate the weight of a person based on the lifestyle habits of a person, which wasn't done in the previous method.

One of the biggest challenges we faced was the size of the dataset, which was with 2111 values, which isn't small, but could have been bigger for our health recommender system. To implement a more robust recommendation system, we need more inputs, helping to provide us with more accuracy in our system.

We also learned a lot from this project, including Data Analysis, implementation Boosting Algorithms and proper utilization of FeatureImportances module as well.

8. REFERENCES

[1] 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=ihub

[2] 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.

[3] 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.

[4] Source Dataset:
https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+