Project Scope Statement

Obesity based on eating habits & physical cond.

Yiyang Sheng, Shengdan Jin, Ying Liu
GEORGETOWN UNIVERSITY |

Table of Contents

1.1 Project #	*1.2 Project Description*	*1.3 Date Submitted*	*1.4 Project Priority*
1	Obesity has become a severe social issue. Being overweight may lead to serious health consequences such as cardiovascular disease In this case, in order to prevent bad health consequences from the beginning, it is essential to track the eating habits and physical condition of individuals and then predict their obesity levels based on those factors. We are using classification as an evaluation method since the goal is to estimate obesity levels and to study which factor/feature would contribute to the obesity level. We are commissioned to build a machine learning application that hospitals can use for predicting patients' obesity levels to maximum productivity. Data Source: https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+		Priority 01

1.5 Step 1. Project Deliverables

Please list *all project deliverables* listed in the Project Charter and, if necessary, elaborate on them. *Do not list dates*. Add more rows as necessary.

Deliverable ID#	**Description**
1	Project Charter
2	Ensemble trained model with preliminary results of the training and testing.
3	Flask and Heroku application. (customer will choose whether to deploy it locally or on the cloud)
4	Final report of the project in addition to the cloud (heroku) deployed link.

1.6 Step 2. List of Project Tasks

Please list ***all project tasks*** to be completed, based on the "Deliverables" specified in the Project Charter*. Do not list dates*. Add more rows as necessary. Optional: you may substitute a work breakdown structure (WBS) or mind-map in lieu of Step 2. Please attach WBS or mind-map to the document.

Task ID#	**Task to be completed**	**Delivery Date**	**For Deliverable #**
1	Submit Project Charter	09/16/2021	1
2	Ensemble trained model with preliminary results of the training and testing.	11/12/2021	2
3	Flask and Heroku application. (customer will choose whether to deploy it locally or on the cloud)	11/26/2021	3
4	Final report of the project in addition to the cloud (heroku) deployed link.	12/02/2021*	4

1.7 Step 3. Out of Scope

This project **will NOT accomplish or include** the following:	This project will not include any analysis of the history of obesity, further research, and other background details. 1. The features for the study might be insufficient to display or explain the problem when they contain no more than people's eating habits and physical conditions. There could be some other features making a difference on the target variable "obesity," including but not limited to people's stress level and genetic factors. 2. This project uses the dataset from 2019, which may dismiss the most updated information and may bring a deviated to some extent from the most recent ground truth. 3. This project may need help from could platform to study 2000 and more instances. 4. This project would focus on classification. Some features may need to be modified into different types for different learning. For instance, the column/feature NObesity (Obesity Level) may need to be changed from string to numeric for regression learning. 5. The ensemble methods and learning algorithms may not be sufficient to explore and learn the data. There could be more methods and algorithms be applied for comparison and checking the difference.

1.8 Step 4. Project Assumptions

Please list any project factors that will be considered to be true, real, or certain. Assumptions generally involve a certain degree of risk.

#	Assumption
1	Obesity level has a negative relationship with the time using technology devices
2	Obesity level has a positive relationship with the frequent consumption of high caloric food
3	Obesity level has a positive relationship with male

1.9 Step 5. Project Constraints

Project Start Date	08/25/2021
Launch/Go-Live Date	12/08/2021
Project End Date	12/08/2021
List any hard deadline(s)	09/16/2021 11/12/2021 11/26/2021 12/02/2021*
List other dates/descriptions of key milestones	**11/08/2021** 1. Exploratory data analysis 1.1. Include all the sub-step you will take to run the exploratory analysis (generally, it should include data frame and variable visualizations and possible suggestions for transformations). These also include cleaning and transformation steps. **11/09/2021** 2. Binary classification baseline evaluations. 2.1. List the techniques we would like to evaluate to see the performance. For the sake of this project, do a 2.1) decision tree analysis and 2.2. Logistic regression analysis. **11/10/2021** 3. Ensemble method evaluation - multiple classifiers then voting ensemble. achieve maximum accuracy **11/16/2021** 4. Exporting the model as pickel or hd5 model from TensorFlow.

	11/17/2021 5. Developing the web-based app on the flask. 6. Deploying it on AWS/Azure
Budget constraints Enter information about project budget limitations	N/A
Quality or performance constraints Enter any other requirements for the functionality, performance, or quality of the project	N/A
Equipment/personnel Constraints Enter any constraints regarding equipment or people that will impact the project	N/A
Regulatory constraints Enter any legal, policy or other regulatory constraints	N/A

1.10 Step 6. Updated Estimates

Estimate T&C hours required to complete project	50 hours	If charge-back project, list total estimated T&C cost	N/A

1.11 Step7.Approvals

Required For Project Class…	Role of Approver	Submitted for Approval on:	Approval Received on:
All classes	1. Client + Client Supervisor	Nakul R. Padalkar	09/16/2021
All classes	2. T&C Supervising Manager	Nakul R. Padalkar	11/12/2021
Class 3 + 4 only	4. VP for Technology & Communication	Nakul R. Padalkar	11/26/2021
Class 3 + 4 only	5. Project Review Board	Nakul R. Padalkar	12/02/2021

Attach any additional documentation.

Office Use Only:

2 Introduction

Obesity has become a severe social issue. Being overweight may lead to serious health consequences such as cardiovascular disease. Therefore we plan to build a machine learning application that will predict obesity levels.

Since the outcome is to predict obesity levels, we have decided to train our machine learning model on a historical medical database. The database contains details of the eating habits and physical activity of each individual.

The dataset had 17 attributes containing information about eating habit attributes and physical attributes, and other attributes.

Eating habit attributes: 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)**Physical attributes:** Calories consumption monitoring(SCC), Physical activity frequency(FAF), Time using technology devices(TUE), Transportation used(MTRANS)
other attribute:GENDER,AGE,HEIGHT,WEIGHT

3 Analysis of the dataset and Trained Model

3.1 Exploratory Analysis and Visualization

Since we are using a secondary dataset, we have decided to perform exploratory analysis before moving forward. The following graphs show the distribution of selected variables. The dataset does not have any missing values.

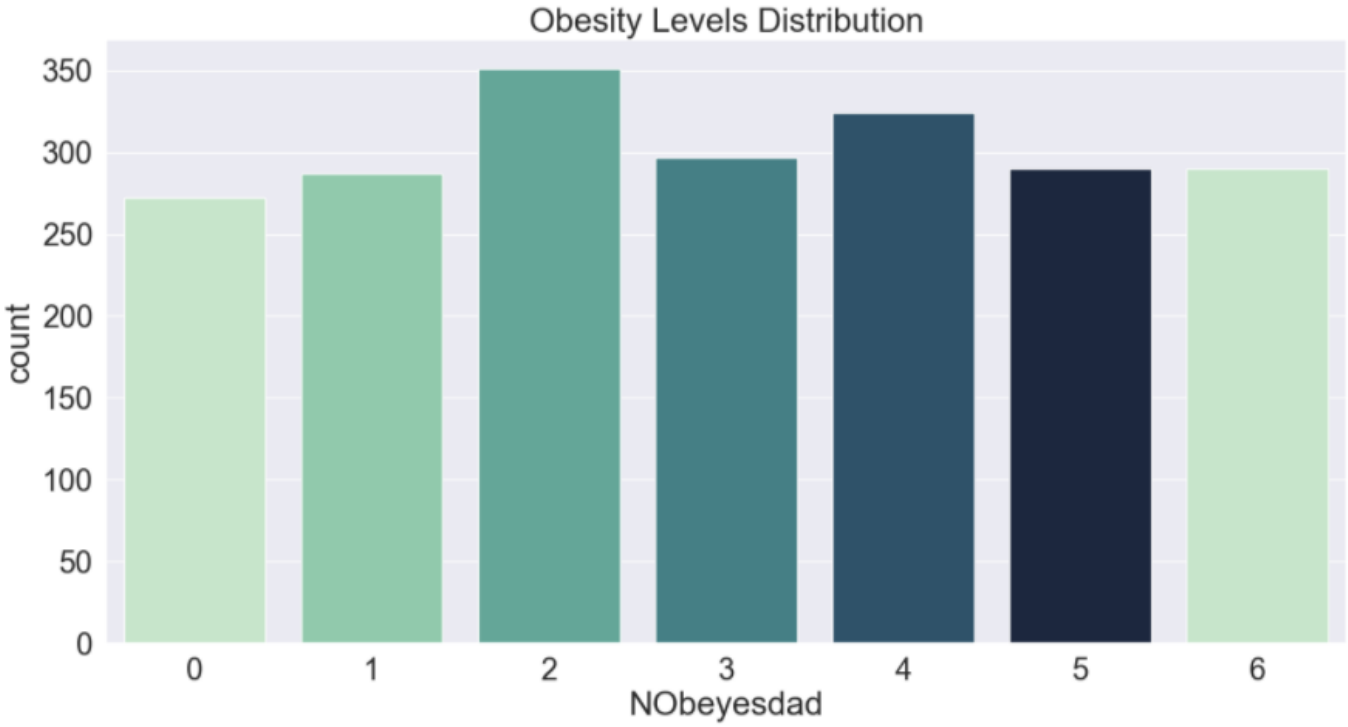


Figure 1 Obesity Levels Distribution

Figure 1 shows the distribution of the dependent variable, NObeyesdad(Obesity Level). The obesity Level is an approximately uniform distribution. Figure 2 shows pair plots, while table 2 shows the summary of all the remaining variables.

Parameters	Gender	Age	Height	Weight	family_ history	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
count	2,111	2,111	2,111	2,111	2,111	2,111	2,111	2,111	2,111	2,111	2,111	2,111	2,111	2,111	2,111	2,111	2,111

mean	0.51	24.31	1.70	86.59	0.82	0.88	2.42	2.69	1.86	0.02	2.01	0.05	1.01	0.66	2.27	2.37	3.02
std	0.50	6.35	0.09	26.19	0.39	0.32	0.53	0.78	0.47	0.14	0.61	0.21	0.85	0.61	0.52	1.26	1.95
min	0.00	14.00	1.45	39.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	19.95	1.63	65.47	1.00	1.00	2.00	2.66	2.00	0.00	1.58	0.00	0.12	0.00	2.00	3.00	1.00
50%	1.00	22.78	1.70	83.00	1.00	1.00	2.39	3.00	2.00	0.00	2.00	0.00	1.00	0.63	2.00	3.00	3.00
75%	1.00	26.00	1.77	107.43	1.00	1.00	3.00	3.00	2.00	0.00	2.48	0.00	1.67	1.00	3.00	3.00	5.00
max	1.00	61.00	1.98	173.00	1.00	1.00	3.00	4.00	3.00	1.00	3.00	1.00	3.00	2.00	3.00	4.00	6.00

Table 1 Summary of the Variables

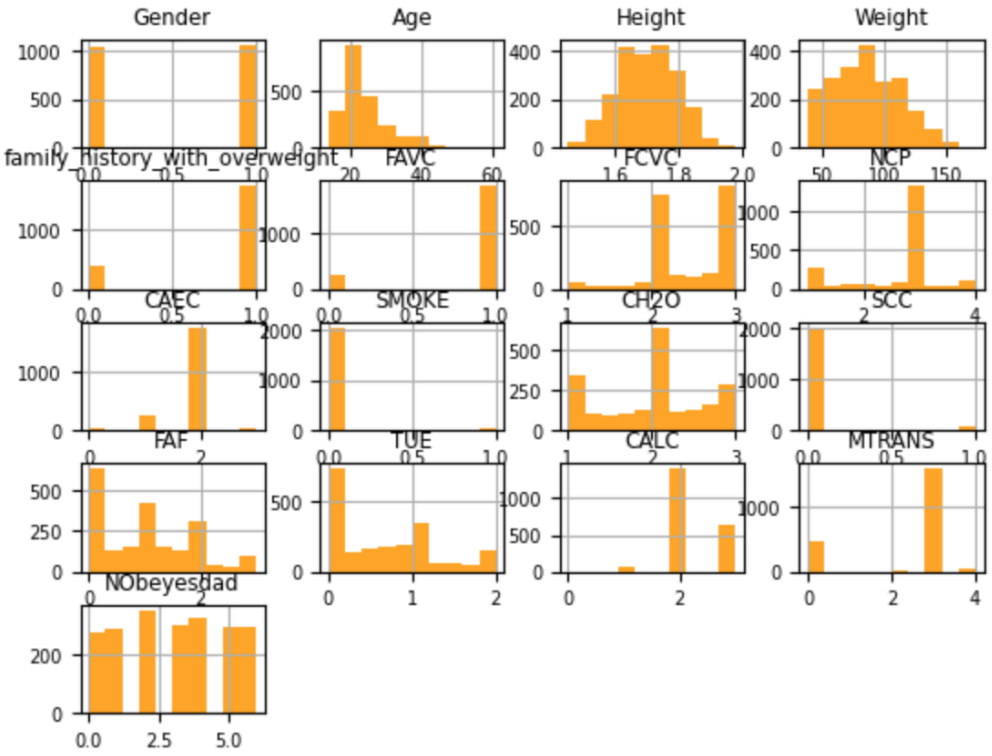


Figure 2 The distribution plots of each variable

Looking at the histograms on the diagonal, we can see that variables Height and Weight have the closest shape to normal distributions. Variables AGE, FAF, and TUE are skewed to the left. Variable Gender is uniformly distributed. For NCP, CAEC, CH2O, and SCC, we find a large distribution is around 1500, 2000, 550, 2000 correspondingly.

3.2 Baseline Model

Since this is the classification dataset, we have decided to analyze the model using logistic regression. We will also assume this to be the baseline model and improve performance using multiple ensemble techniques. For simplicity, all the model code is provided with an attached Jupyter notebook, not in the report. The R^2 for the regression is around 26.8%, suggesting that the model fits not so well to the dataset.

4 Model Selection

Obtaining a baseline provided us necessary insight for model selection. Following the preliminary classification analysis, we have evaluated Logistic Regression, Ridge Classification, Lasso Classification, Decision Tree Classification, Random Forest Classification, Bagging Classification, and Gradient Boosting Classification as candidate models. Figure 3 shows the model results, and we can see that Decision Trees, Random forests, Bagging, Gradient boosting, and light gradient boosting perform the best (have relatively smallest mean RMSE). Because of this, we will select them as the candidate models for level 0 in the stacking regression. We will use Light Gradient Boosting Regression as the level 1 combiner or metamodel to aggregate the results of the level 0 models.

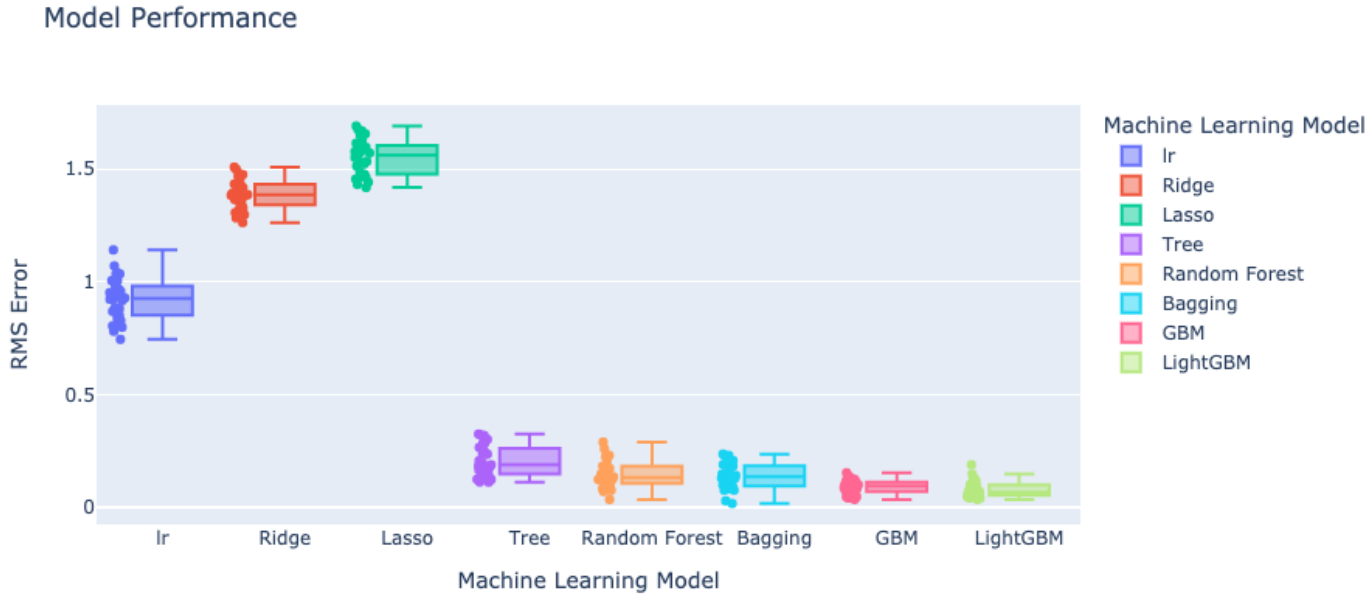


Figure 3 Machine Learning Model Result

In addition to the selected models, we have also opted for cross-validation of the data set. For the current iteration of the model, we will use five-fold validation. Figure 4 shows the performance of the standalone and stacked models. The performance of the stacked model seems to be better than the candidate model. Tune the hyperparameters can improve the stacked model's performance by tuning the hyperparameters. For now, we will focus on the obesity level classification by using a stacked model.

Figure 4 Candidate model and stacked model performance

4.1 Model Performance Evaluation

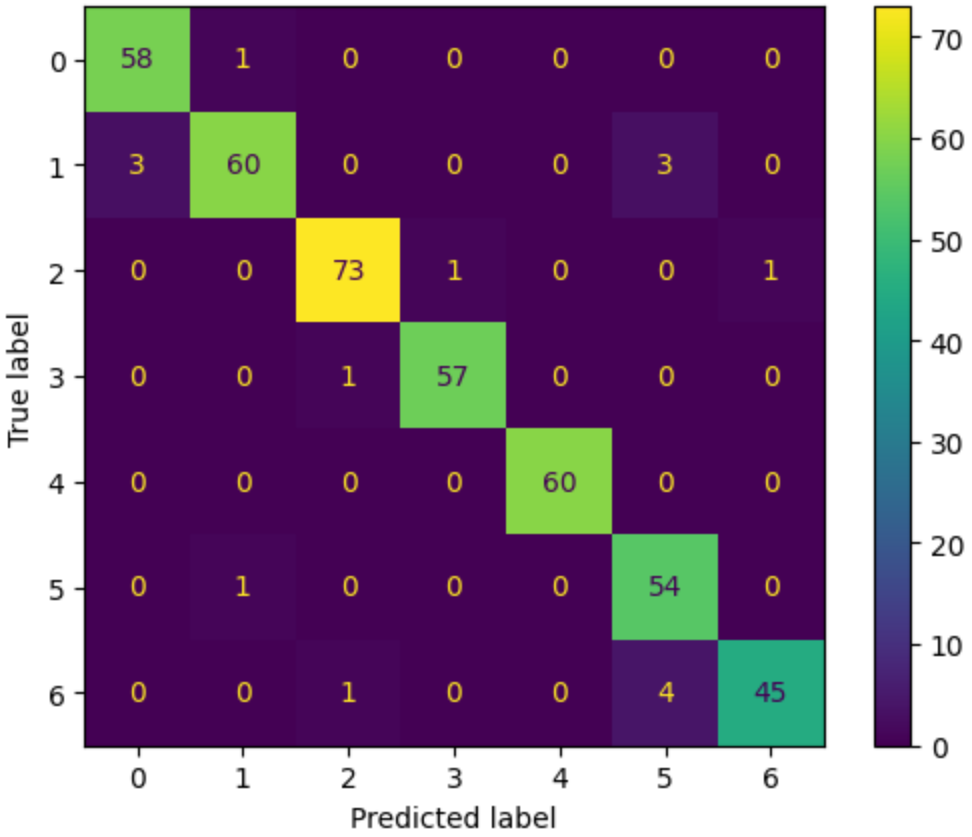


Figure 5 Confusion Matrix

The final accuracy is 96.22%. And from the confusion matrix, since most numbers are located on the diagonal line, it is indicated that most samples are predicted correctly and that the model built is reliable.

5 Initial Deployment

We have selected Heroku as a final deployment site. The site will display 1 graph. The graph will display the target variable obesity level against the two most important features of the dataset -- height and weight.

5.1 Screens Before Prediction

Figure 6-10 shows the initial plot of the target variable against the most important features, height and weight. Users can provide input to the text box below the image. This input is processed and then used to predict the obesity level for the given feature vectors.

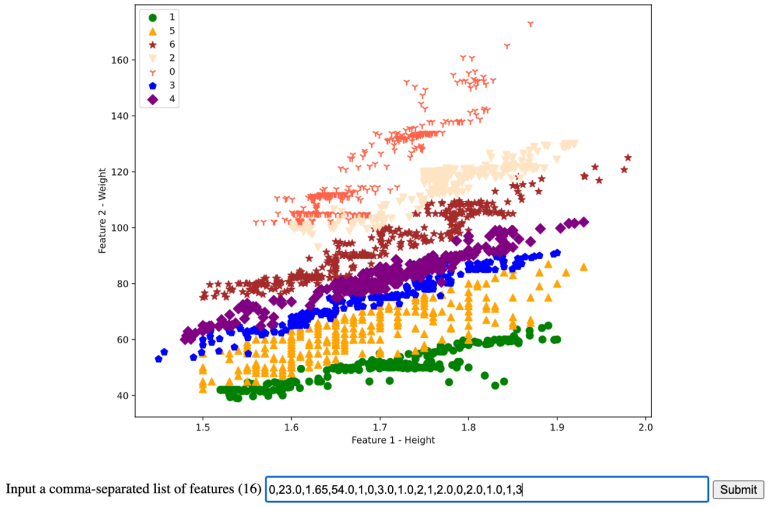


Figure 6 Screenshot of the unpredicted dataset(vector1)

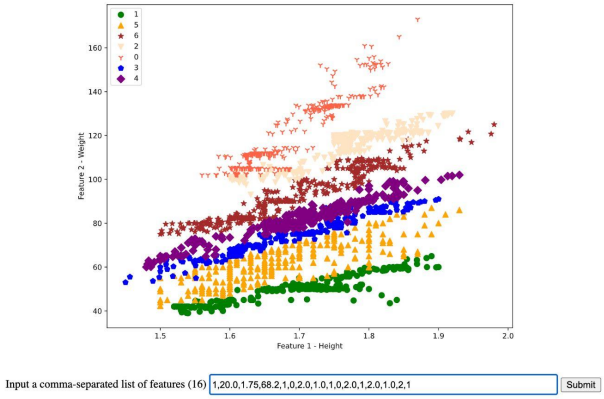


Figure 7 Screenshot of the unpredicted dataset(vector2)

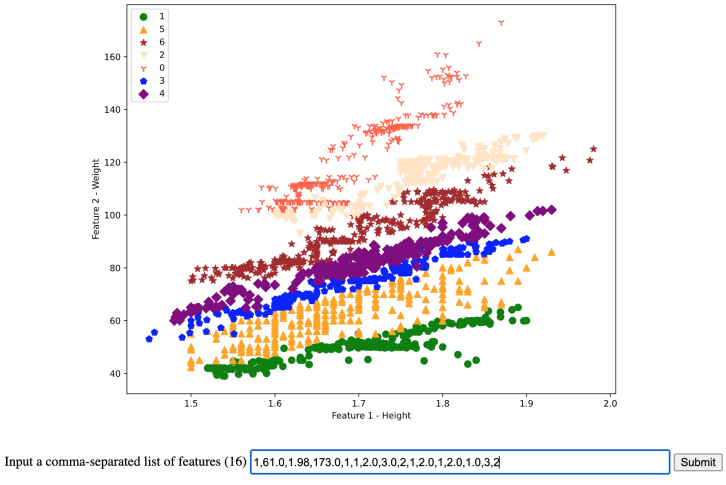
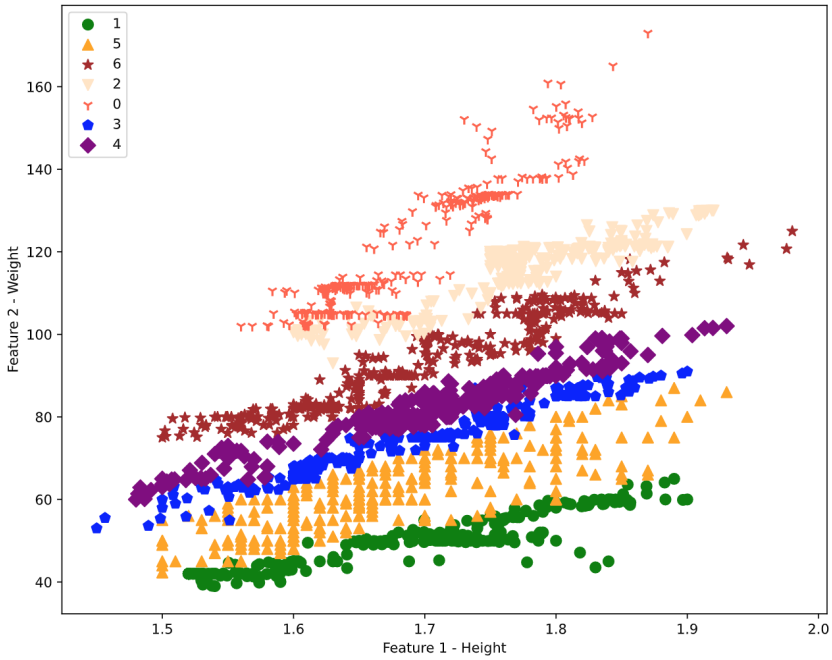
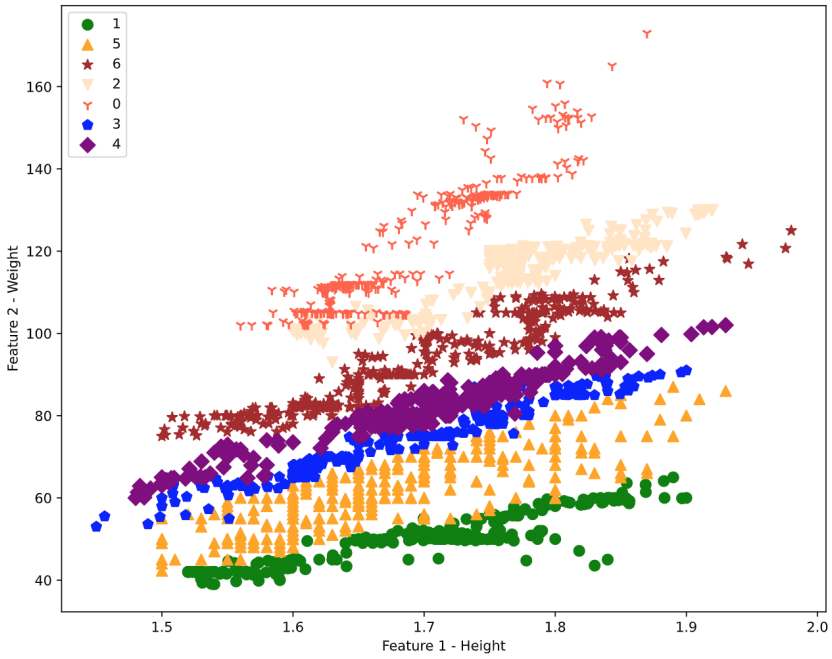


Figure 8 Screenshot of the unpredicted dataset(vector3)



Input a comma-separated list of features (16) 1,55.0,1.75,160.0,0,0,2.0,4.0,0,0,2.0,2,2.0,0.0,0,4 Submit

Figure 9 Screenshot of the unpredicted dataset(vector4)



Input a comma-separated list of features (16) 1,50.0,1.88,100.0,1,0,2.0,3.0,1,0,2.0,1,2.0,0.0,1,3 Submit

Figure 10 Screenshot of the unpredicted dataset(vector5)

5.2 Screens After Prediction

Figure 11-15 shows the updated image with the predicted obesity level in black on the plots. Here are five feature vectors that the user can use for testing.

1. 0.23,0,1.65,54.0,1,0,3.0,1.0,2,1,2.0,0,2.0,1.0,1,3

2. 1,20.0,1.75,68.2,1,0,2.0,1.0,1,0,2.0,1,2.0,1.0,2,1

3.1,61.0,1.98,173.0,1,1,2.0,3.0,2,1,2.0,1,2.0,1.0,3,2

4.1,55.0,1.75,160.0,0,0,2.0,4.0,0,0,2.0,2,2.0,0.0,0,4

5.1,50.0,1.88,100.0,1,0,2.0,3.0,1,0,2.0,1,2.0,0.0,1,3

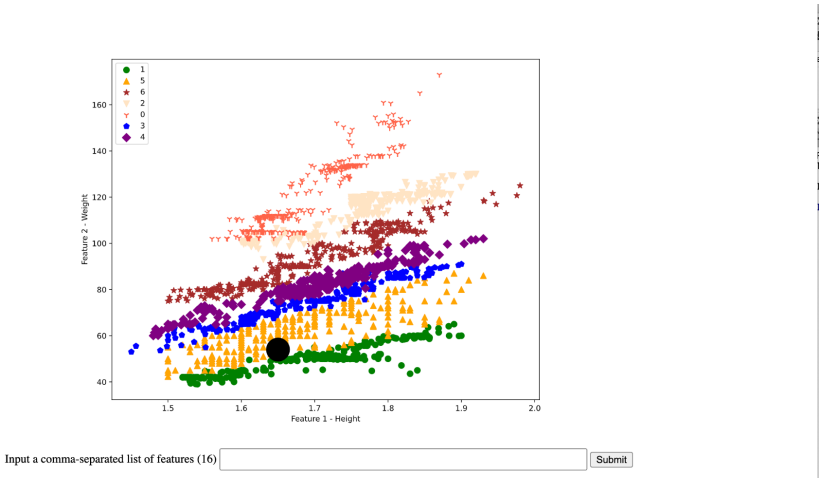


Figure 11 Screenshot of the predicted dataset(vector 1)

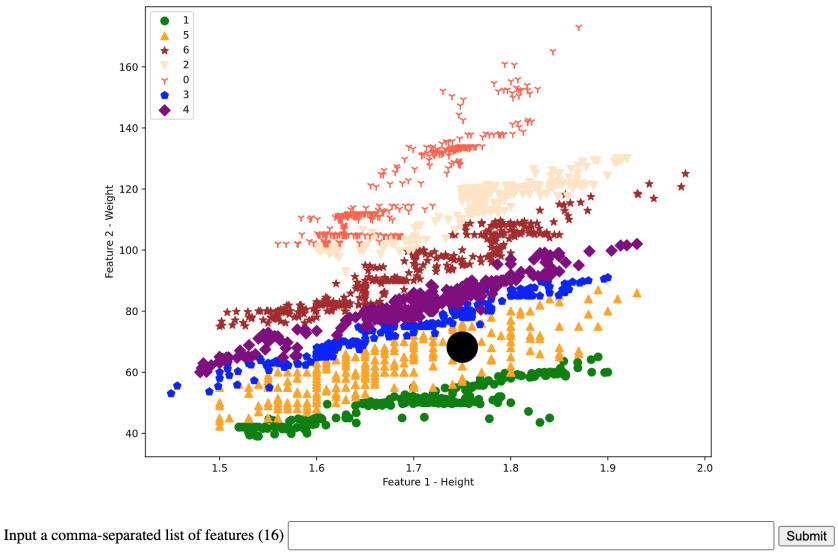
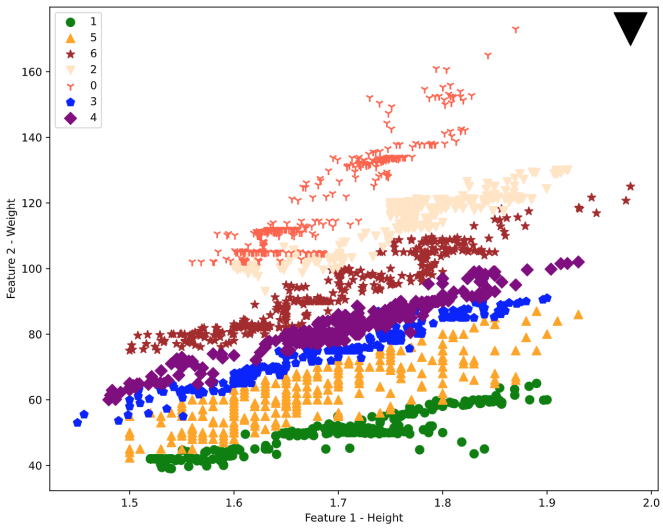
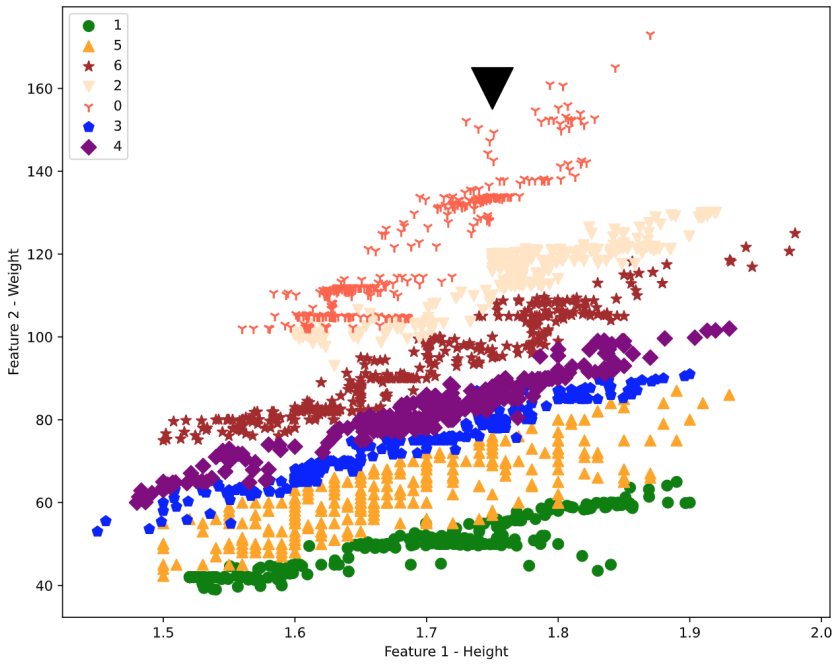


Figure 12 Screenshot of the predicted dataset(vector 2)



Input a comma-separated list of features (16) Submit

Figure 13 Screenshot of the predicted dataset(vector 3)



Input a comma-separated list of features (16) Submit

Figure 14 Screenshot of the predicted dataset(vector 4)

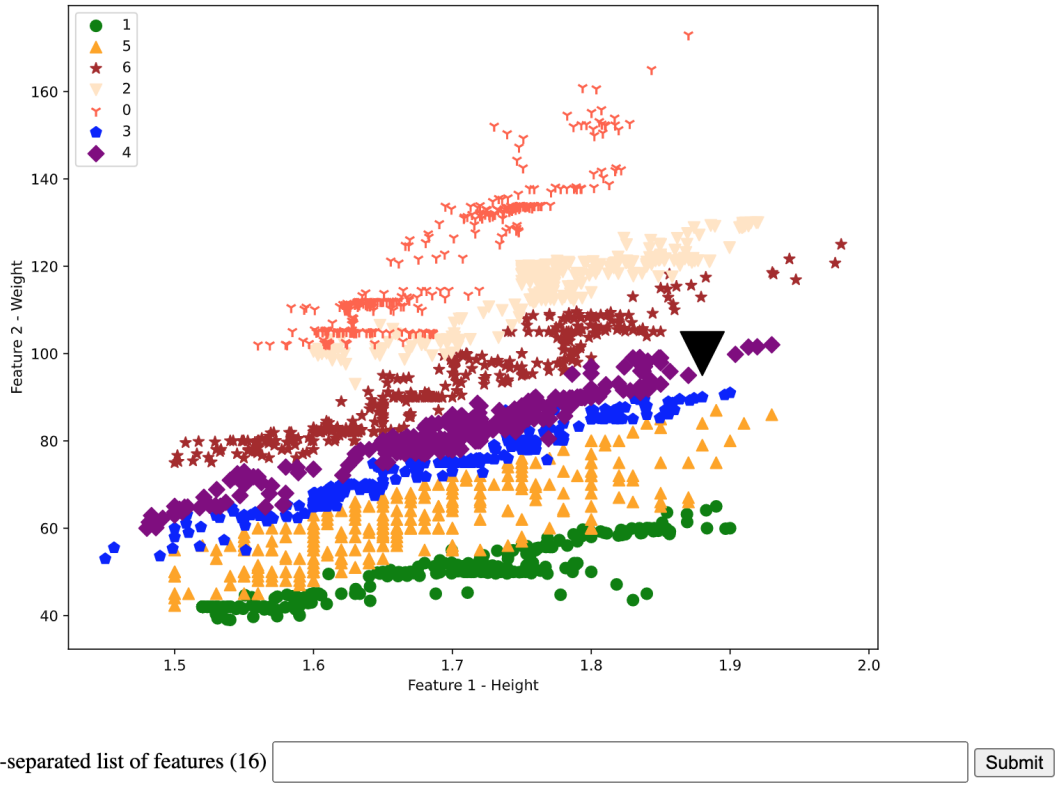


Figure 15 Screenshot of the predicted dataset(vector 5)

5.3 Heroku Application

5.3.1 The Heroku Platform

The Heroku network runs the customer's apps in virtual containers, which execute on a reliable runtime environment. Heroku calls these containers "Dynos." These Dynos can run code written in Node, Ruby, PHP, Go, Scala, Python, Java, or Clojure. Heroku also provides custom buildpacks with which the developer can deploy apps in any other language. Heroku lets the developer scale the app instantly by either increasing the number of dynos or changing the type of dyno the app runs in.

5.3.2 Application path

Users can go to https://anly605app-g04.herokuapp.com/ to test the model and the build. Please utilize the following five feature vectors to test the model. A table showing the variable and example value for each variable is also included below the input field. Users may choose to create any feature vectors to feed to the model.

5.3.3 Deployment, debugging, and updates

We used Herkou's deployment feature to connect a Github repository and used it to deploy the model. Heroku provides a command-line tool to debug the error and the code. We utilized it to modify the model to make it usable on the Heroku platform.

6 Conclusion

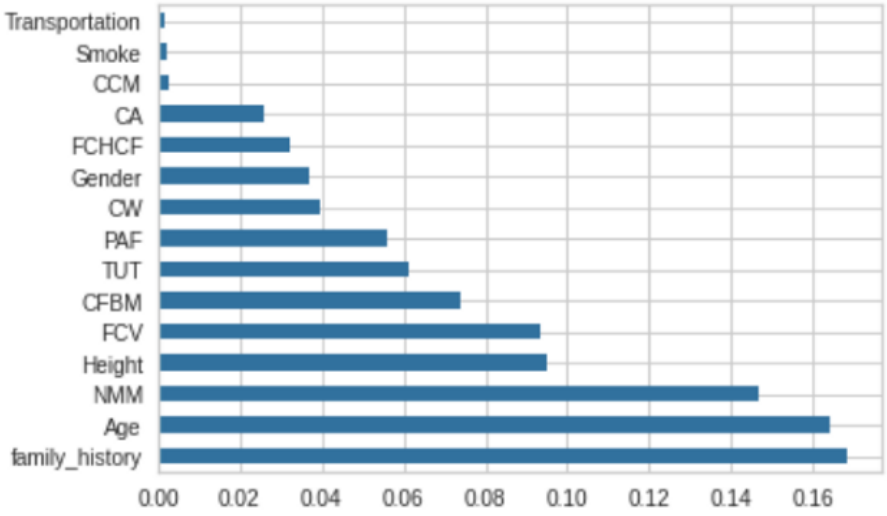


Figure 16 Feature Importance Plot

The feature importance calculations show that age, family history, and height had the largest influence on the obesity level. Other factors that had a large influence on obesity level were the consumption of food before meals, physical activity frequency, and the number of main meals. Factors that were not impactful across the board were smoking, calorie consumption monitoring, gender, and transportation.