Predicting the Obesity Level

Authors: Siming Yin and Xingzhi Ma Team: N.N

- Executive Summary-

Our project aims to develop a predictive model that can identify the risk of obesity in individuals based on their eating habits and physical condition. By leveraging data-driven insights, this project seeks to contribute to early intervention and personalized health recommendations to reduce the chance of having obesity. In this project, we will mainly focus on predictive and prescriptive analysis by applying classification and visualization models.

[Decisions to be impacted]

1. **Public Investment on preventing obesity**

   Predicting the possibility of obesity and implementing preventive measures can help reduce the cost of public investments in treating obesity-related health issues. For example, the predictive model can help identify individuals at higher risk of obesity based on the physical condition, eating habits or lifestyle. With early identification, public health programs can educate people on healthy eating and exercise so that they reduce the chance of developing obesity issues and the cost of treating obesity. Also with the help of technology, we can develop a health app that combined with predictive models, can enable remote monitoring of individuals at risk of obesity. It can provide guidance, support, and feedback to individuals in order to help reduce the need for hospital visits and the related costs.

2. **Identification of health treatment**

   Predicting the possibility of obesity can be a valuable tool in identifying health treatment needs and setting achievable weight loss or weight maintenance goals. With the help of the predictive model, individuals can identify whether or not they are at the risk of obesity and provide some guidelines and recommendations on predicting obesity. The predictive model can be integrated into monitoring systems that track an individual's progress in real-time and build up personalized medication and therapies.

3. **Provide social support to maintain a healthy lifestyle**

   Since obesity is often a chronic condition that requires long-term management, predictive models can assist healthcare providers in designing treatment plans that are sustainable and adaptable as the patient's health evolves over time. It helps provide social support to people to maintain a healthy lifestyle and reduce the chance of getting obesity.

[Business value]

1. **Personal Health Improvement: Reduce the possibility of individuals suffering from obesity.**

   One of the business values of this predictive model is that it helps improving personal health involves making lifestyle changes and adopting healthy habits. For example, it helps people realize the importance of quality sleep and efficient stress management, and educate themselves with nutrition knowledge. Setting achievable and specific health goals to maintain healthy weight.
2. **Public Health Improvement: Help the government to lower the expenditure on preventing obesity.**

   The predictive model also can play an important role in public health improvement through helping the government to lower the cost of preventing obesity. By prioritizing preventive measures, such as health education, social programs, and policies targeting healthier environments, the government can effectively reduce the chance of people getting obesity in the community.
3. **Research Cost Reduction: Generate valuable research findings and insights into the relationship between eating habits, physical condition, and obesity risk.**

   Beside that, the other business value of the predictive model would be to reduce the cost of obesity research. The model employs efficient methodologies and technologies to generate valuable insights into the relationship between eating habits, physical condition, and the risk of obesity. It helps researchers uncover the patterns between these features and discover some potential research with different other fields, such as nutrition and epidemiology.

[Data assets]

Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico.

- Data Preprocessing-

[Data Description]

We are using the Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. This dataset contains the information of people from the countries of Mexico, Peru and Colombia, with ages between 14 and 61, obesity level and diverse eating habits and physical condition. The data was collected using a web platform with a survey (see Table 1) where anonymous users answered each question. It comprises 2111 rows, representing 2111 individuals, and 17 columns, representing 16 features and 1 target (obesity level) for each individual. The dataset features can be categorized into two groups based on their data type: categorical and ratio. Our ratio data are Ages, Height, and Weight. And our categorical data includes: Gender, Family History With Overweight (FHWO), Consumption of High Caloric Food (FAVC), Consumption of Vegetables(FCVC), Number of Main Meals (NCP), Consumption of Food Between Meals (CAEC), Smoke Consumption of Water Daily (CH2O), Calories Consumption Monitoring (SCC), Physical Activity Frequency (FAF), Time Using Technology Devices (TUE), Consumption of Alcohol (CALC) Transportation Used (MTRANS). Obesity level is categorized by the calculation of BMI (Body Mass Index): • Underweight Less than 18.5 •Normal 18.5 to 24.9 • Overweight 25.0 to 29.9 • Obesity I 30.0 to 34.9 • Obesity II 35.0 to 39.9 • Obesity III Higher than 40. The entire data is composed of 77% generated data by Weka tool and the SMOTE filter and 23% directly collected data from the survey. Since the obesity level is calculated by BMI, we will not use weight and height as our features to build the predictive model.

[Data Visualization]

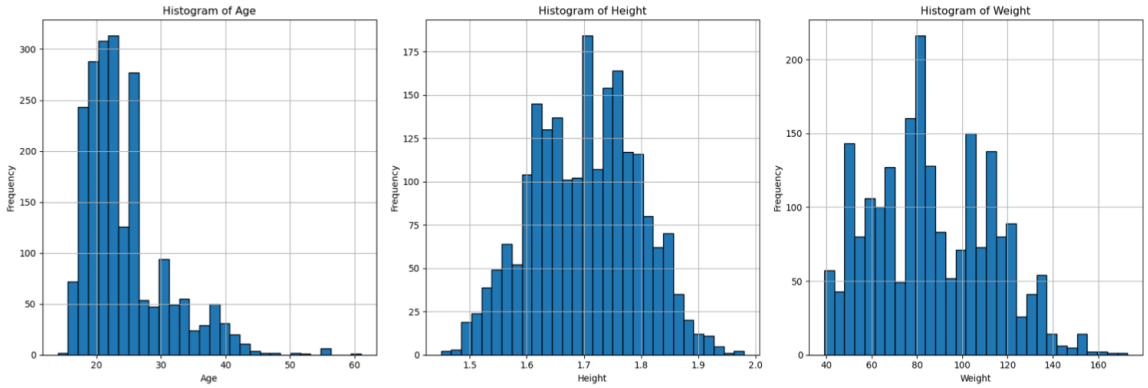

Fig: Histogram of Age (Left) Histogram of Height (Middle) Histogram of Weight (Right)

We can see that most of our samples are taken from people under 30. From the Histogram of Age on the left, we can see that most of the surveyees are aged around 15 to 28. The data of Height has a maximum of 198 and minimum of 145, and the median value is about 170. From the plot of Height, we find that most data are in the interval between 160 to 177. There is a peak around 163, a second peak at around 170, and a third peak at around 175. The data of Weight is ranged from 39 to 173. From the histogram of Weight on the right, we see that our data has a major peak at around 80, and several secondary peaks around 45, 100, and 105. From all three plots, we see that they all seem likely to form a multimodal distribution. The histogram of Age is positively skewed. The histogram of Weight is slightly right skewed. The histogram of Height has a normal-distribution-like shape compared to the other two plots because of the central tendency, but not in general. The reason lead to this could be the natural difference in Height and Weight between male and female. So we split the data by gender and generate the following plots.

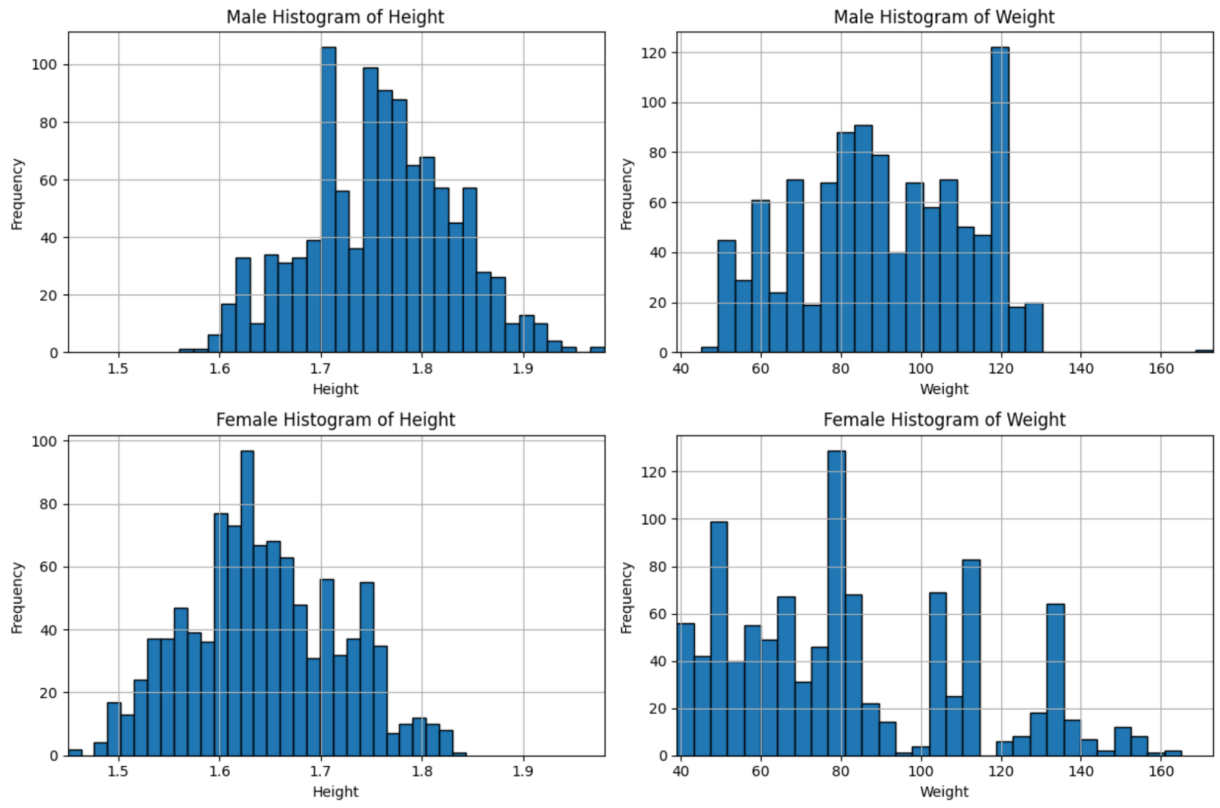

Fig: Male Histogram of Height(Up Left) Male Histogram of Weight (Up Right) Female Histogram of Height (Up Left) Female Histogram of Weight (Down Right)

From the Histograms shown above, we observe that, though we split our data by gender, the histograms of Height still display multiple peaks in both male and female. The data of Weight are still inconsistent and tend to form multimodal distribution. The potential explanation of this could be the regional difference between the sample we collected.

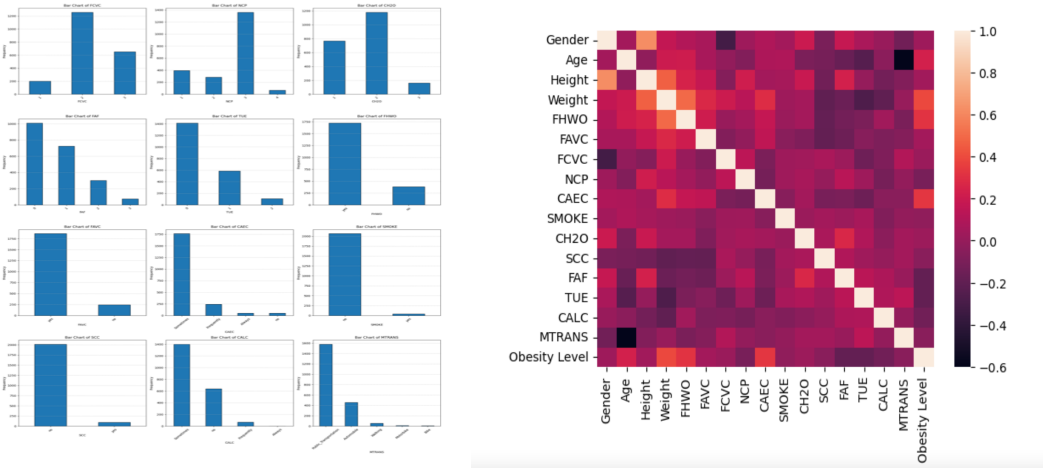

Fig: Barplot of the categorical data (Left) and correlation plot (Right)

The left bar plots show the frequency of categories of each feature. Some of the features have a dominant category that has a really high percentage. The most prominent one is the feature smoke (SMOKE). 97% of the data are in the category of *No*, and only 3% of the data are in *Yes*. The feature Calories Consumption Monitoring (SCC) has 95% of the data in the category of *Yes* and 5% of data are in *No*. The feature Consumption of High Caloric Food (FAVC) on the other hand has 88% of the data in *Yes* and 12% in *No*. So, we can say that the majority of our surveyees are non-smokers who did not monitor their calories consumption, and have consumed high caloric food. The correlation plot on the right shows the relationship between different features. We see from the plot that Weigh and Family History With Overweight (FHWO) are most correlated features with Obesity level. This is intuitive, since Obesity is very likely genetically inherited. We also see that Obesity level is correlated with Consumption of Food Between Meals (CAEC) and Age. It is unexpected that Height is not highly correlated with Obesity level, since BMI (Body Mass Index) is calculated by Height and Weight.

[Outlier Detection]

Since most of the features in our dataset are categorical data, we can use bar plot to determine the outlier. For example, the columns “Smoke” and “SCC” have a pretty high percentage of one specific value. Take “Smoke” as an example, around 97% of the population are non-smoker, which means the outlier would be those who smokes. By looking at the target value, Obesity level, we are using pie charts and bar charts to check the outliers. In the data description part, we have mentioned that the dataset has been balanced by filter SMOTE, then there are no outliers for our target value. In our dataset, there are three numerical data, we do a boxplot of these three columns.

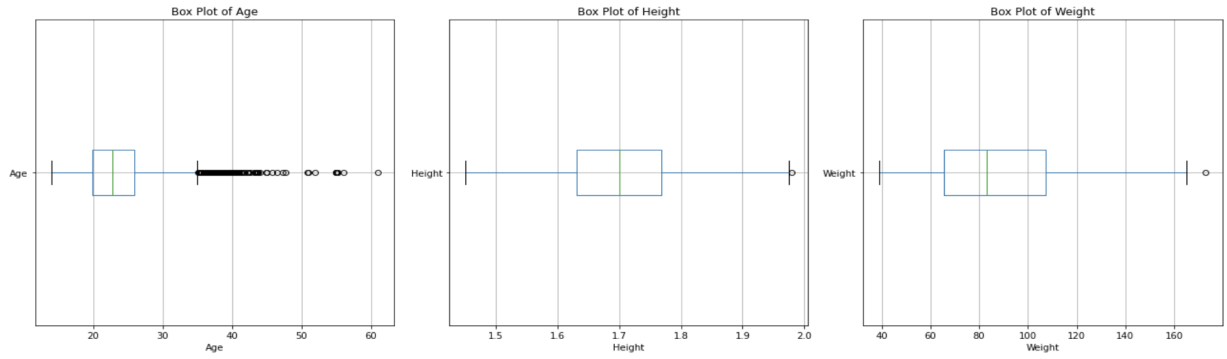

Fig: Box plot of Age (Left) Box plot of Height (Mid) Box plot of Weight (Right)

Notice that there is only one outlier for “Height” and “Weight”. In general, these two columns distribute pretty normally. However, for the “Age”, we found that with the minimum of column is 14, the 25% of Interquartile range (IQR) is 19.95, the mean is 24.31, the medium is 22.78, the 75% of IQR is 26, and the maximum is 61. We discovered that the score data is skewed to the left with plenty of older ages. Lastly, we have also performed PCA on Obesity level with the two highest correlated features (FHWO, CAEC). There's a significant overlap between several obesity levels in the central region of the plot. This indicates that, based on the first two principal components, many individuals from different obesity levels share some similar characteristics. The plot seems to show a major cluster in the center with some scatter around it. The “Insufficient_Weight" seems more distinct compared to other levels, especially in certain regions of the plot. In general, The majority of data points, particularly “Obesity_Type_I”, and “Overweight_Level_I”, are clustered around the center. The wide distribution of data points suggests there is considerable variability within the dataset.

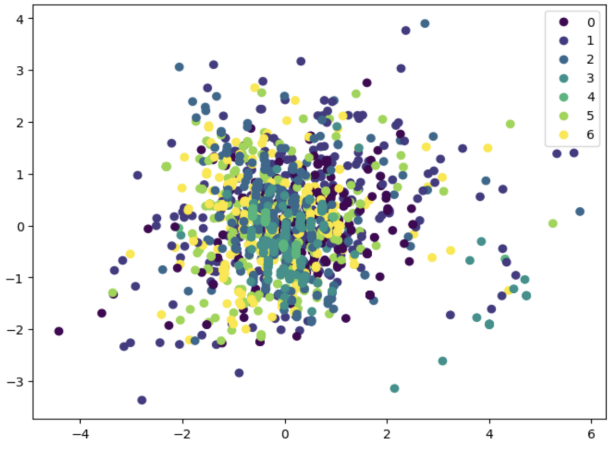

Fig: 2D PCA scatter plot for obesity level

The other thing that we do to perform the outlier detection is using the Isolation Forest to detect the anomaly data point in our data set. Isolation forest is an unsupervised learning algorithm that identifies anomalies by isolating outliers in the data. Because we have a small dataset, we have defined the expected proportion of outliers in the data set to be 1%, which is the contamination parameter in the model. There are 22 predictive outliers that have been identified. After running the model, we are using *crosstab* to create a contingency table. It is a way of summarizing and analyzing the relationship between two or more categorical variables. It computes a frequency table, showing how many times each combination of variables occurs in the dataset. We are using this table to evaluate our models. We added up

all the counts with the condition (num_count <= 5). If the number of counts of the combination is less than five, then we will consider that as an outlier. We also get a result of 22 outliers detected.

	Gender	Age	FHWO	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	anomaly	scores
18	0	30	1	1	3	4	1	1	1	0	0	0	3	0	-1	-0.0104834
21	0	52	1	1	3	1	2	1	2	0	0	0	3	0	-1	-0.0245264
25	1	20	1	0	2	4	1	1	2	0	3	2	3	3	-1	-0.0112556
30	1	29	0	1	1	4	1	0	3	0	0	1	3	2	-1	-0.00732673
68	1	30	1	1	1	3	3	1	2	1	0	0	1	0	-1	-0.0532807
92	1	55	1	0	3	4	1	0	3	1	3	0	1	4	-1	-0.0466905
119	0	19	1	0	3	3	1	1	3	0	2	1	2	0	-1	-0.00817113
132	0	19	1	1	3	3	1	1	3	1	1	2	1	3	-1	-0.0250775
133	0	61	0	1	3	3	0	0	2	0	1	1	1	3	-1	-0.00607298
142	1	23	0	1	2	3	1	1	1	0	1	1	1	0	-1	-1.97072e-05
152	0	38	1	1	2	1	0	1	2	0	0	0	2	0	-1	-0.00580758
188	1	35	1	1	3	1	3	0	3	0	3	1	1	0	-1	-0.029988
191	1	26	1	1	3	1	1	1	2	1	2	0	2	3	-1	-0.0118799
200	0	23	1	0	3	1	2	1	3	0	1	2	2	3	-1	-0.01063
217	1	21	0	0	2	3	1	0	3	1	3	1	1	0	-1	-0.00317801
232	0	51	1	0	3	3	2	1	3	1	2	0	3	3	-1	-0.0239557
236	0	21	0	1	1	3	0	0	2	1	3	0	3	0	-1	-0.00738452
245	0	20	0	0	3	3	2	1	2	0	2	1	2	0	-1	-0.00376463
252	1	56	1	0	2	3	2	1	2	0	1	0	1	0	-1	-0.00592272
277	1	21	0	1	2	4	0	1	3	0	3	2	2	4	-1	-0.0394712
333	0	23	0	0	3	4	0	0	3	1	3	0	3	0	-1	-0.0236606
495	1	19	1	1	3	1	0	0	1	1	0	0	3	2	-1	-0.00784573

Fig: Outliers in the data set by using Isolation Forest Anomaly Detection method

[Data Cleaning]

The dataset that we are using synthetic data, most of the data preprocessing has been completed. There is no null value or missing data, atypical data has been deleted, and data normalization has been completed. The dataset has used the tool Weka and the filter SMOTE to generate the synthetic data in order to balance the categories of obesity levels. The balancing process decreases the probability of skewed learning in favor of a majority class.

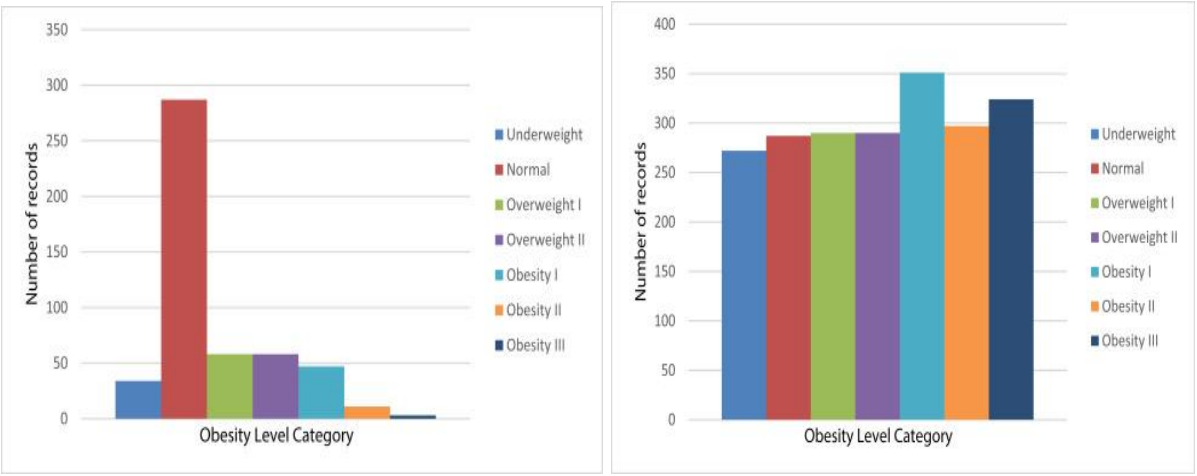

Fig: Unbalanced distribution of data regarding the obesity levels category (Left) Balanced distribution of data regarding the obesity levels category (Right)
(*https://doi.org/10.1016/j.dib.2019.104344*)

To make the column name consistent, we renamed some of the column names to make it easy to understand and check out the number of the features and decide the target column, which is the obesity level. Since the data are collected by the survey, most of the columns are categorical data. Some of them have been encoded into numerical data and some are not. We have transformed those numerical data into integers and the rest. For example, one column is called “Frequency of consumption of vegetables” which has the values 'Always', 'Frequently', 'Sometimes', 'no'. Then we labeled them into number 0, 1, 2, 3.

- Model Updates-

Because the target value obesity level is calculated by BMI formulate which is using the height and weight. The correlation between height and weight should be highly correlated to Obesity level, which can cause an issue in our predictive model. Thus we decided to use all the features except weight and height in our model.

[Predictive Model]

Random Forest

In this project, we are going to first implement a Random Forest algorithm. Among all other supervised classification models, Random Forest has several outstanding properties that give us more accurate and explicit outputs, particularly in the topic of predicting obesity and classifying obesity level. Random Forest is capable of dealing with multifactorial conditions,

overfitting, and able to process categorical data and aggregate feature importance. At the data preparation and preprocessing stage, we have set the feature Obesity Level as the target value, and converted all the categorical data into numerical format. When we do train test split, we use bootstrap to randomly select training dataset and testing dataset. We are going to use the grid search, the random search and the bayesian optimization to find the optimized value of hyperparameters. Our hyperparameter includes the number of decision trees, maximum depth of each tree, and proportion of train test split. By using grid search, we test all possible combinations of hyperparameters from predefined ranges, training and evaluating the model using each combination, and selecting the one that yields the best performance. For the predefined ranges, we can set a list of values for each hyperparameter, and we apply GridSearchCV from sk-learn package, and then we fit GridSearchCV on our training data. We will calculate an accuracy score to evaluate the performance. Instead of testing all possible values of hyperparameters, we will also apply random search to explore the best hyperparameters combination in a more efficient way. Bayesian Optimization, on the other hand, also tests different combinations of hyperparameters, but will also estimate the past performance of your past hyperparameter that affects the future decision.

KNN (K-nearest Neighbors)

The other model that we plan to use is KNN. KNN is used to classify different obesity levels, it counts the number of data points in each obesity level among the K nearest neighbors and assigns the obesity level label that is most common among them to the new data points. First we decided to use Minkowski distance to calculate the distance in order to find the nearest neighbors of given data points. This distance measure is the general form of Euclidean and Manhattan distance metrics. Since we have 16 features in our dataset, Minkowski distance can measure the distance between two data points in N-dimensions.

$$Minkowski\ Distance\ = \ (\sum_{i=1}^{n} |X_i - Y_i|^{1/p})$$

The other thing that we have decided is to find the number of K. The k value in the k-NN algorithm defines how many neighbors will be checked to determine the classification of the obesity level. Deciding the K-value is important because different k-values can lead to overfitting or underfitting. If the k-value that we choose is too small then it will lead to low bias but high variance; and if the k-value is too large then it will lead to high bias and lower

variance. From the previous data preprocessing stage, we noticed that our dataset contains more outliers and noise, so it will likely perform better with higher values of k. Also choosing an odd value for k helps avoid ties when determining the class with a majority vote. In order to make sure the k-value works the best for our model, we will experiment with different k values and observe how they affect the model's performance. We may plan to apply Leave-One-Out Cross-Validation (LOOCV) to find the proper k-values. In LOOCV, we will train the model with k set to the number of data points minus one and validate it on the remaining point. Repeat this process for all data points and compute the average error. We can also apply cross validation to find k-values. We will estimate the performance of the KNN model for different values of k. Then select the k that results in the best average performance across the folds.

[Machine Learning Morphism]

Since our simple model is only taking use of 14 features, including , we decided to drop columns that are not being used (weight and height).

$MLM1: ML_{clean1} = (R^{2111\times14}, R^{2111\times14}, F_1(x;\theta) = \theta x, P_\theta(\theta) = [110011111111111], L_1: trivial)$

We perform data cleaning and data preprocessing to transfer categorical data into numerical data.

$MLM2: ML_{clean2} = (R^{2111\times14}, R^{2111\times14}, F_2: embedding\ parameter, L_2: trivial)$

K-nearest neighbors (KNN):

Input Space: $X = R^{2111\times n}$

Output Space: $Y = R$

Parameter Pior: K (number of neighbors), Distance Metric

Learning Morphism: To get the highest kth terms using their obesity level data point

Loss Function: MAE = $\frac{\sum_{i=1}^{n}|s_i|}{n}$

$MLM3: ML_{model1} = (R^{2111\times14}, R^{2111\times14}, F_3(F_1(x:\theta), O_x) = x, P_\theta(O_x) = S_x, L_3 = (-O_x))$

where O_x = *the similarity* obesity level of data point and input

MLM4:

$ML_{result1} = ML_{model1} \circ_o ML_{clean1} \circ_o = (R^{2111\times14}, R^{2111\times14}, F_3(F_1(x:\theta), O_x) = x, P_\theta(O_x) = O_x, L_3 = (-O_x))$

Random Forest:

Input Space: $X = R^{2111\times n}$

Output Space: $Y = R$

Parameter Pior: Number of decision trees, Maximum depth of each tree
Proportion of train test split.

Learning Morphism: Multiple Trees

Loss Function: Gini Impurity $= 1 - \sum_{i=1}^{k} p_i^2$

$MLM5: ML_{model2} = (R^{2111\times n}, R^{m\times n}, F_5(F_1(x:\theta), O_x) = x, P_\Theta(O_x) = O_x, L_5 = (-O_x))$

$MLM6: ML_{result2} = ML_{model2} \circ_o ML_{clean1} \circ_o = (R^{2111\times n} R^{m\times n}, F_5(F_1(x:\theta), O_x) = x, P_\Theta(O_x) = O_x, L_5 = (-O_x))$

- Source Code -

https://github.com/simiy1/predicting_obesity

- Next Steps-

For the next step, we are going to perform feature selection to decide what features we should apply in our model. After all the data preprocessing and feature selection finished, we are going to work on our model by using Random Forest algorithm and KNN algorithm.

10/23 - 10/27 Midterm Presentation

10/30 - 11/3 Perform KNN and Random Forest algorithm

11/6 - 11/10 Fix model

11/13 -11/17 Try to build another model for weight control recommendation system

11/20 - 12/1 Review the project

Reference:

Kho, J. (2019, March 12). Why random forest is My Favorite Machine Learning Model. Medium. https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706

Palechor, F. M., & Manotas, A. de. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in Brief, 25, 104344. https://doi.org/10.1016/j.dib.2019.104344

Centers for Disease Control and Prevention. (2022, September 24). Health effects of overweight and obesity. Centers for Disease Control and Prevention. https://www.cdc.gov/healthyweight/effects/index.html

What is the K-nearest neighbors algorithm?. IBM. (n.d.). https://www.ibm.com/topics/knn#:~:text=Next%20steps-,K%2DNearest%20Neighbors%20Algorithm,of%20an%20individual%20data%20point

K, D. (2021, April 9). Anomaly detection using isolation forest in python. Paperspace Blog. https://blog.paperspace.com/anomaly-detection-isolation-forest

Gupta, R. (2023, May 6). The power of Crosstab function in pandas for data analysis and visualization. Medium. https://medium.com/geekculture/the-power-of-crosstab-function-in-pandas-for-data-analysis-and-visualization-6c085c269fcd#:~:text=One%20of%20the%20most%20useful,the%20relationships%20between%20the%20variables.