**Data-Driven Prediction and Prevention of Obesity**

**by Jackson Schieber –** [**jschieber3@gatech.edu**](mailto:jschieber3@gatech.edu)

**ISYE 7406 – Data Mining and Statistical Learning**

**Group 110**

**Abstract**

Given the overwhelming empirical evidence that obesity is detrimental to public health, it is important to identify contributing factors to its existence. Insight on these factors can assist the population in avoiding the condition, and it can inform policy makers and health professionals where/how efforts should be made to combat the disease. This analysis considers obesity data from one of Kaggle’s competitions (https://www.kaggle.com/competitions/playground-series-s4e2/data) and aims to predict the obesity level (7 different classifications) based on 16 attributes generated from surveys of individuals from the countries of Mexico, Peru and Colombia with ages ranging between 14 and 61 and diverse eating habits and physical condition (Fabio and Alexis). The paper begins with exploratory data analysis (EDA) revealing that many features have a significant correlation with obesity. Next, several classification models are designed and tested on the dataset. The best performers include random forest and XGBoost which achieve accuracy scores around 90%; though, most of the models are above 70%. Then, the XGBoost hyperparameters are finetuned to enhance accuracy further. Finally, the models are inspected for possible insights regarding obesity prediction and prevention.

**Introduction**

Numerous polls will show that health is the greatest contributor to human happiness. Additionally, a healthier population is going to be more productive/capable and suffer fewer medical expenses. Therefore, society stands to benefit by maximizing this metric. While health professionals and scientists might not always agree on what is “good” health, there seems to be a strong consensus that obesity significantly and negatively impacts a person’s health. According to ourworldindata.org, 8% of global deaths are attributable to obesity, and this number is increasing over time. Thus, this analysis is intended to reveal contributing factors to obesity so that people can be aware of ways to live healthier. It aims to accomplish this through informative visualizations and examining the coefficients of the classification models. Additionally, producing the ability to predict obesity based on common and measurable demographic and behavioral features might produce an intuition in policy makers that can guide strategy and direct awareness campaigns. Prior to creating any models, this paper considers feature distribution, correlation, and descriptive statistics. After this exploration, the data is adjusted so as to be easily digestible by a machine. Finally, various models are constructed and implemented on different subsets of the data and the results are discussed. As mentioned above, this dataset is offered as part of a Kaggle competition. It is synthetically generated; though, it is designed to have the same distribution as the real/actual collected data which is found in another spot on Kaggle. Therefore, this paper experiments with using solely the provided synthetic dataset and a combination of the two when modeling; however, all EDA addresses only the synthetic dataset since the distribution and characteristics are supposed to be identical. A synthetically generated testing dataset is also provided in the competition but without the dependent feature – NObeyesdad. Only after submitting predicted values to Kaggle, can one see their overall accuracy score. Nothing else is shown.

**Exploratory Data Analysis**

This data contains 18 different attributes including an ID column and the dependent variable. The independent variables include gender, age, height, weight, family\_history\_with\_overweight, FAVC, FCVC, NCP, CAEC, SMOKE, CH2O, SCC, FAF, TUE, CALC, MTRANS. The variable of interest, NObeyesdad, represents 1 of 7 different weight classifications of each individual including Overweight\_Level\_I, Overweight\_Level\_II, Normal\_Weight, Insufficient\_Weight, Obesity\_Type\_III, Obesity\_Type\_II, and Obesity\_Type\_I." In the appendix, one can find a figure of the count of each of these classifications. Type III obesity is noticeably the most common classification, but the overall numbers are mostly even. According to Kaggle, “The attributes related with eating habits 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), and consumption of alcohol (CALC). The attributes related with the physical condition are calories consumption monitoring (SCC), physical activity frequency (FAF), time using technology devices (TUE), and transportation used (MTRANS).” The SMOKE variable indicates whether the individual is a smoker. One can find the first 3 rows of this data at the beginning of the appendix. This dataset has zero missing values and is easily downloaded as a CSV; thus, little data cleanup is needed. Also, the categorical columns are converted into additional binary columns to facilitate modeling, and the independent variable is encoded into numbers 0-6.

In the appendix, one can find a correlation plot between both categorical and numerical variables. However, one can see a plot of just numeric columns immediately below. It looks like height and weight are understandably the most positively correlated items. Interestingly, height is positively correlated with physical activity frequency (weight isn’t) while weight is positively correlated with consumption of water, consumption of vegetables, and age. Therefore, taller people tend to be more active, as they probably excel more at physical activity. It is interesting that weight increases with vegetable consumption. It seems that people who eat more vegetables may just eat more overall. While it is not explicitly said, it is possible that NObeyesdad measures an alternate calculation of body mass index (BMI). If this is true, these classifications may be misleading since this metric is renowned for classifying high muscle low fat individuals as obese. However, none of these correlation numbers indicate that this is happening to a serious degree, though, it is possible that the near zero correlation between weight and activity level is artificially higher because of people with higher concentrations of muscle. It appears that weight is given in kilograms and height in meters. To get a sense of how NObeyesdad is calculated, the maximum BMI value (kg/m2) is observed for the different classifications. It is standard that any person with a value over 30 is technically obese. However, there is an individual in the training data with a “BMI” of just under 33 classified as normal weight; there are many other examples of individuals not following under classic BMI classifications. There is an individual with normal weight with a BMI of 14.6 which is much below the standard underweight threshold of 18.5. Therefore, it appears that this dataset may use some form of it, but certainly not the standard form of it. See the KDE plot of BMI segmented by NObeyesdad further below for more information. This is positive, otherwise, predicting these classifications when given weight and height would be rather meaningless.

A couple observations from the full correlation plot in the appendix includes a mild positive correlation between weight and CAEC (sometimes) and CALC (sometimes). It is interesting that only the sometimes response shares the positive correlation. Anything less or more is neutrally correlated. In this instance the “moderation is key” adage does not appear to apply. Additionally, having a family history of being overweight has the most positive correlation with a person’s weight. This is likely genetic, but it could also have environmental impact as it could affect a person’s attitude toward weight/food/habits.

A screenshot of a graph

Description automatically generated

The data is comprised of 50% male and female. The first of the bar charts below displays numbers of individuals with a specific obesity type aggregated by gender. There are more females than males in both the underweight and normal weight categories; conversely, the highest obesity classification consists almost only of females. Strangely, there are just 8 females that fall in “Obesity\_Type\_2” and 5 males who fall into “Obesity\_Type\_3.” It is unknown why there is such a disparity, but this seems like an easy observation for a model to exploit. Adjacent to this figure is another bar graph depicting the enormous difference that having a family history of obesity has on one’s own weight. It is notable that the strong majority of people do qualify as having this family history. This high number makes one wonder what criteria are being used to qualify this metric or if this sample is representative of the population. Referring to the count of NObeyesdad classifications in the appendix, while the different categories are all represented well, there are only two categories that do not indicate that an individual is overweight. Thus, over two thirds of all individuals are classified as overweight to some degree. The World Health Organization estimates that about 40% of the adult population is overweight. Therefore, it appears that this sample over represents obesity. Below these two graphs are 6 more bar plots of the rest of the categorical variables stratified by obesity classification.

While an individual’s method of transportation does not appear to offer enormous predictability on weight, individuals with obesity type III exclusively use public transportation and the percent of each weight class who use automobiles increases steadily as weight class increases. Therefore, automobile usage appears to be correlated with weight class. As mentioned above, all of obesity type three is female. One might wonder if the exclusive use of public transportation in this class is related to gender; a count of transportation type segmented by gender (in the appendix) shows that many females rely on non-public methods of transportation but a significantly larger portion of them rely on public transportation than males. Thus, it appears that gender is a factor, but not the only one. The CALC plot does not provide obvious distinctions between alcohol consumption and NObeyesdad classification, but it does show that the vast majority of individuals consume alcohol periodically. The next figure shows that the percent of people who consume high caloric food increases alongside weight class, but the trend is not absolute. The graph of the consumption of food between meals among those surveyed shows that half of people underweight frequently consume food between meals. This drastically decreases as weight class increases so snacking may be conducive to a lower weight. The next plot shows that most people don’t smoke, so it is hard to glean insight from a raw count graph alone; however, the biggest proportion of smokers is in obesity class 2, so its possible this variable is a factor on weight. Finally, monitoring calorie consumption shows a large negative correlation with weight.

A graph with blue and orange bars

Description automatically generated A graph with blue and orange bars

Description automatically generated

A graph with different colored bars

Description automatically generated A graph with different colored bars

Description automatically generated

A graph with blue and orange bars

Description automatically generated A graph with different colored bars

Description automatically generated

A graph with blue and orange bars

Description automatically generated A graph with blue and orange bars

Description automatically generated

Below, one can find kernel density estimators (KDE) of the density of the numerical variables striated by NObeyesdad. While it can be difficult to distinguish between the several colors of the various classes, focusing on green and orange (insufficient/normal weight) is helpful. The plots show that the various features are multimodal with some of them having about 4 different high concentrations areas. Surprisingly, it does not appear that the time spent using tech devices has a meaningful impact on weight. Meanwhile, physical activity frequency does appear to have a negative correlation with weight class judging from the large green peak on FAF 2 and the large red peak on FAF 0. Consumption of water has one of the more obvious distinctions of density between classes. Consumption of one unit is dominated by green and orange while the density of 3 unit consumption of water is mostly obese. It is worth noting that the middle of the range, 2, has a good mix of all weight types. The consumption of vegetables does not have an obvious relationship with weight class as one of the highest values, 3, is heavily represented by green and orange, but the minimum value among those who are obese type III is also 3. Therefore, both obese and normal/underweight eat many vegetables. It is possible that the obese eat lots of everything, including vegetables, and those with normal weight just eat lots of vegetables, but this does not seem like an easy variable to make a class prediction from. The plot of weight shows that, naturally, a higher weight is correlated with a higher NObeyesdad class. It is interesting that within each different class there is a wide range of weights. For example, type III obesity is very bimodal. The height plot is a bit messy. One takeaway is that height doesn’t appear to have a huge relationship with NObeyesdad from the KDE alone. One should remember that the towers of purple and red on opposite sides (obesity type II and 3) are a result of gender differences as men are taller. Further below, the KDE of BMI shows an incredibly strong relationship with NObeyesdad. They are closely related as suspected. However, the overlap among classes shows that these variables are not identical and that there is more to the NObeyesdad equation than just BMI. Finally, the KDE of age also shows a positive correlation with NObeyesdad. There is also a KDE on the number of main meals consumed found in the appendix; however, it looks more like a single straight line since most everyone consumes 3 meals per day.

A graph of different colored lines

Description automatically generated A graph of different colored lines

Description automatically generated

A graph of different colored lines

Description automatically generated A graph of a normal distribution

Description automatically generated with medium confidence

A graph of weight loss

Description automatically generated A graph of height and weight

Description automatically generated

A graph of age and weight

Description automatically generated with medium confidence A graph of different colored lines

Description automatically generated

To compliment the illustration of the KDE’s on the numeric variables, there are several boxplots found in the appendix which more clearly show correlation with NObeyesdad. They mostly confirm earlier conclusions. For example, height alone is not a good predictor of class, but age is. They also show that each feature has many outliers; though, it might not be necessary to scale the data since numerical spreads are small. One can more easily see the different spread of each variable by class with these boxplots by observing the difference between the 25th and 75th percentiles. Also, one should note that a variable may not be predictive by itself but can still offer valuable predictive power when in combination with other variables.

**Methodology**

Having more thoroughly understood the data, one can now consider the methods used for modeling. As mentioned, Kaggle provides clean training and testing datasets of synthetically generated data which follows the same distribution as the original copy of real data. Thus, different subsets of data are fed into the model. In the Python code, one must adjust a few lines of comments to change which data is fed into the models. One scenario includes using the full synthetic training data set to train the models and testing on the full set of original data. Cross validation is not used for this step as part of the objective is to evaluate how similar the two datasets are. Additionally, model coefficients are stored in a table found in the appendix. These coefficients indicate which features are most impactful to model predictions. This is a good estimator of which features impact obesity. However, one should interpret feature importance cautiously, as it should not be conflated with causation. It solely reflects the predictive power of features within the context of the specific model. For example, a hypothetical model could predict weight by several features including height and gender. Some versions of the model may have height set to 0 and gender set to something significant while others could have the opposite, but the predictions could be the exact same. One reason is that they are highly correlated. Only one feature may be necessary to make a prediction. Moreover, a model may not even need either of these features to make an accurate prediction, and they could both be set to 0; however, they both clearly are predictive of weight.

Models used include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Naïve Bayes (Multinomial), Naïve Bayes (Gaussian), K-Nearest Neighbors (Optimal k=4), and XGBoost Classifier. Logistic Regression, a simple and interpretable model, is suitable for scenarios with a relatively linear relationship between features and the target. However, it may struggle to capture complexity. A decision tree is a hierarchical structure of decisions based on feature splits. For example, if predicting whether an object is an apple or an orange, an effective feature split may be color. This model is interpretable and facilitates easy visualization. Furthermore, it can effectively model for nonlinear and complex interactions. However, it is prone to overfitting. In response to the overfitting tendency of decision trees, random forest was invented. This method is an ensemble of decision trees where each tree is built on a random subset of features and data. Feature coefficients are selected from the average results of all of the trees. This method tends to be more successful, but it is computationally expensive. SVM creates a hyperplane that best separates classes in a high-dimensional space but may require careful parameter and kernel tuning. Some of these kernels can be computationally expensive. Naïve Bayes models, whether Multinomial or Gaussian, are computationally efficient but assume feature independence or a Gaussian distribution, respectively. They leverage Baye’s Theorem to make predictions. K-Nearest Neighbors is simple and non-parametric, and it is suitable for complex decision boundaries. It simply finds the total distance between data points in a multidimensional space. However, it can be computationally expensive on large datasets and can be negatively influenced by irrelevant features. Finally, XGBoost is a gradient boosting algorithm that often achieves high accuracy. Gradient boosting is similar to random forest except it is more calculated, less random, and adjusts subsequently created trees based off of the results of prior trees. Therefore, it is more prone to overfitting, and it is very sensitive to parameters. Optimizing these hyperparameters can be expensive and time consuming. XGBoost is a type of gradient boosting that includes regularization which reduces overfitting and increases generalization of the model. After creating these models and noticing that XGBoost provides the greatest performance, a couple python packages including hyperopt and optuna are used to algorithmically tune parameters for this model. As mentioned, this is computationally expensive. One of this project’s few attempts at parameter tuning for 100 iterations lasted 50 minutes.

Additionally, the numerical columns “Age” and “Weight” are scaled for some of the models since these columns have the largest spread of values. However, it has little consequence on accuracy measures, so the original data is used. Additionally, on one occasion, models are fed with Boolean values, and on another occasion these Booleans are converted to integer. Curiously, this actually effects the results a few percentage points. The reason is unclear. The Booleans produce the most accurate results, so they are used.

**Results**

The results of training on synthetic and testing on original data are immediately below. All models offer greater than 50% accuracy while the average is 77% which is quite good. Some models significantly outperformed others including XGBoost, Random Forest, Decision Tree, and KNN. Notably, XGBoost produces 95% accuracy. Interestingly, precision, recall, and accuracy are all about even. This means that the model suffers from both false positives and false negatives at an equal rate. This is partly because the dataset is well balanced with a fair proportion of each class. Models tend to struggle when one class is significantly more common than another. Note that KNN is dependent upon the number of neighbors measured for each point (K). Cross validation is used to select the optimal k which produces the highest average accuracy which is 4. The specific accuracy measures for each class for each of the better performing models (not SVM or Bayes) can be found at the bottom of the appendix. The averages for each class are immediately below. Normal\_Weight, Obesity\_Type\_I, and Overweight\_Level\_I appear to be the hardest classes to predict. Obesity type II and III are the easiest to predict which was expected given the clear gender separation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **Logistic Regression** | 0.67 | 0.66 | 0.67 | 0.66 |
| **Random Forest** | 0.94 | 0.93 | 0.93 | 0.93 |
| **SVM** | 0.67 | 0.68 | 0.68 | 0.67 |
| **Naïve Bayes (Multi)** | 0.57 | 0.57 | 0.57 | 0.56 |
| **Naïve Bayes (Gaussian)** | 0.55 | 0.57 | 0.55 | 0.52 |
| **Decision Tree** | 0.89 | 0.89 | 0.89 | 0.89 |
| **KNN (k=4)** | 0.88 | 0.88 | 0.88 | 0.87 |
| **XGBoost Classifier** | 0.95 | 0.95 | 0.95 | 0.95 |

**Results - Synthetic Training Real Testing**

|  |  |
| --- | --- |
| **Class** | **Average of Precision and Recall** |
| **Overweight\_Level\_II** | 0.924 |
| **Normal\_Weight** | 0.797 |
| **Insufficient\_Weight** | 0.846 |
| **Obesity\_Type\_III** | 0.933 |
| **Obesity\_Type\_II** | 0.986 |
| **Overweight\_Level\_I** | 0.775 |
| **Obesity\_Type\_I** | 0.774 |

**Results - Synthetic Training Real Testing (by Class)**

Additionally, the coefficients of the logistic regression model and the variable importance values of Decision Tree, Random Forest, and XGBoost are at the bottom of the appendix. It was not noticed until after finishing with model creation that the class of each person was not encoded in perfect ascending order. This may throw off the calculations of the logistic regression model and make it harder to evaluate its coefficients. Doing the best that one can to find insight in this LR model, it appears that all of the variables are important to the model. The most important variables to the decision tree (in order) include Weight, Gender\_Male, Height, Age, CH2O, FAF, FCVC, NCP, TUE, and CALC\_no; however, the first four are significantly more important than the others. These variables are also in the top ten for Random Forest, though, its top four include Weight, Age, Height, and FCV. The importance weights are more evenly distributed in Random Forest than Decision Tree which seems consistent with the built in safeguards against overfitting that Random Forest has. XGBoost has similar feature importance but its top ten include Gender\_Female, CAEC\_no, FAVC\_no, and SCC\_no. Meanwhile it is without Gender\_Male, Age, FAF, and TUE. Its top four include Gender\_Female, Weight, FCVC, and CAEC\_no. Therefore, the models have distinct differences, but they do overlap.

Next, models were created after performing cross validation – a method to randomly sample training and testing data several times to avoid overfitting. One set of models is trained and tested on solely the synthetic training data set. Another is trained and tested on a combination of synthetic and real data. Providing additional data has the capability to better inform the models. The results are in the table immediately below. It seems that there is no difference between these two methods of using the data. Thus, the synthetic dataset is truly much like the original. However, one cannot conclude that the additional information that the combination of datasets provides improves model performance like hoped for. These methods also average 77% accuracy like the models trained on synthetic and tested on original which is evidence that the datasets are near identically distributed. Once again, XGBoost is the best performer, though its accuracy falls to 90/91%.

|  |  |  |
| --- | --- | --- |
|  | **Accuracy** | |
| **Model** | **Synthetic** | **Synthetic + Real** |
| **Logistic Regression** | 0.72 | 0.71 |
| **Random Forest** | 0.89 | 0.89 |
| **Support Vector Machine** | 0.74 | 0.74 |
| **Naïve Bayes (Multinomial)** | 0.64 | 0.63 |
| **Naïve Bayes (Gaussian)** | 0.59 | 0.58 |
| **Decision Tree** | 0.84 | 0.85 |
| **K-Nearest Neighbors (Optimal** | 0.84 | 0.85 |
| **XGBoost Classifier** | 0.90 | 0.91 |

**Results - Cross Validation**

Because the best performing models include XGBoost and Random Forest, they are submitted to the Kaggle competition where accuracy is measured on the synthetic testing data set. Only 20% of the dataset is initially available to test on. At the time of writing this report, the full results have not yet been released. Random forest and XGBoost received accuracy scores on this 20% of data of 88.403% and 91.329% respectively when trained on just synthetic training data. When trained on the combination of datasets, they produced accuracies of 88.692% and 91.0004%. Therefore, this extra data helped one of the models and hurt the other one slightly, but is not significantly influential. Until this point, all models (except KNN) are created with default parameters. Because the best performing model is XGBoost, additional attention is given to finetuning its parameters. The first effort uses hyperopt and attempts to tune max\_depth, gamma, reg\_alpha, reg\_lambda, colsample\_bytree, and min\_child\_weight. It achieves an awful accuracy of around 25% suggesting an error in methods. Using optuna, this paper achieves an average accuracy of about 91%. The set of parameters used include the following: {'booster': 'dart', 'learning\_rate': 0.08723360512826742, 'max\_depth': 10, 'subsample': 0.7999669669900065, 'colsample\_bytree': 0.5413275309396097, 'min\_child\_weight': 3}. When this model was tested on the competition test data, it achieved a slightly improved accuracy on all datasets. The highest achieved is 91.437% and it is just above the 75th percentile of performers (645/2754). At this stage, small improvements make a significant difference in rankings as the best performing model is 92.268%. Considering additional hyperparameters and tuning for longer periods with different random starting spots will likely yield additional improvements.

**Conclusion**

Having thoroughly explored charts, tables, descriptive statistics, and model results from this data, one should have a more sufficient knowledge of characteristics and habits of people that impact obesity. Weight and height are blatant features that consistently make a high impact on obesity classifications and EDA. Age is another common factor. While this variable is immutable, it is helpful to know who to target for obesity intervention efforts and the young appear less susceptible. Ultimately, the results highlight the importance of both genetic and lifestyle factors in predicting obesity, as evidenced by the influence of family history and eating habits. For example, consumption of alcohol, vegetables, and water along with number of main meals all are included in the top ten most impactful features of the three best performing models – Decision Tree, Random Forest, and XGBoost. The high predictive performance of these models reveals how competent and useful machine learning can be in discovering patterns and making judgments that can far surpass a human. Additionally, the models trained on synthetic data show comparable performance to those trained on a combination of synthetic and real data, indicating the synthetic dataset's reliability. An unexpected insight might be the validity of using synthetic data to support future research efforts where data collection is costly. While there are several correlations between the various features and NObeyesdad, some of the biggest factors appear to be genetic. There does not appear to be a single lifestyle change one can make to guarantee a normal weight. Therefore, it might benefit society to focus research on genetic based prevention of obesity at a cellular and DNA level which is another domain that machine learning can potentially support. Until technology/science reaches this point, people should do what they can to achieve a healthy weight, since there do appear to be ways in which many of us can meaningfully influence it.

**Lessons Learned**

One thing this project taught and reminded me of is that it is important to be organized, neat, and write everything down. For example, I saved several pieces of code outputs, but I did not label them. Later, when transferring them to this paper, I had difficulty remembering what was what. Additionally, when programming, I tried different things (like changing the data subsets or models/parameters) by uncommenting and commenting out different lines. This is not an effective way because it also becomes too cumbersome to manage and track. I also attempted to comment and uncomment lines of code that generated plots so as not to spam my output which caused me to run into unexpected issues. Ultimately, it would be best to design everything to stand on its own. Furthermore, I should be more consistent with formatting. Another possible lesson is that the type of data fed into the model can really matter. As mentioned earlier in the report, at one point I fed the models with Boolean data, and at another point they were supported with integer datatypes. This actually changed the predictions even though True/False is supposed correspond to an integer 1/0. However, it is possible some unknown mistake was made, but this is something I will be weary of going forward. On that same note, I was reminded to take care when encoding categorical variables as the method in which they are encoded can affect the models. I also have only had one experience with finetuning a model using a public library like hyperopt, and never have I used optuna, and engaging in this process has been informative. I learned just how time-consuming tuning a model can be (50 minutes) and how important it is to be as efficient as possible.

**Appendix – Sources and EDA**

Fabio Mendoza Palechor, & Alexis de la Hoz Manotas. (2019, August 2). *Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico*. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2352340919306985

Ritchie, H., & Roser, M. (2024, January 17). *Obesity*. Our World in Data. <https://ourworldindata.org/obesity>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **id** | 0 | 1 | 2 | 3 |
| **Gender** | Male | Female | Female | Female |
| **Age** | 24.443011 | 18 | 18 | 20.952737 |
| **Height** | 1.699998 | 1.56 | 1.71146 | 1.71073 |
| **Weight** | 81.66995 | 57 | 50.165754 | 131.274851 |
| **family\_history\_with\_overweight** | yes | yes | yes | yes |
| **FAVC** | yes | yes | yes | yes |
| **FCVC** | 2 | 2 | 1.880534 | 3 |
| **NCP** | 2.983297 | 3 | 1.411685 | 3 |
| **CAEC** | Sometimes | Frequently | Sometimes | Sometimes |
| **SMOKE** | no | no | no | no |
| **CH2O** | 2.763573 | 2 | 1.910378 | 1.674061 |
| **SCC** | no | no | no | no |
| **FAF** | 0 | 1 | 0.866045 | 1.467863 |
| **TUE** | 0.976473 | 1 | 1.673584 | 0.780199 |
| **CALC** | Sometimes | no | no | Sometimes |
| **MTRANS** | Public\_Transportation | Automobile | Public\_Transportation | Public\_Transportation |
| **NObeyesdad** | Overweight\_Level\_II | Normal\_Weight | Insufficient\_Weight | Obesity\_Type\_III |

**First Three Rows of Dataset**

|  |  |
| --- | --- |
| **CAEC** | |
| **Always** | 478 |
| **Frequently** | 2472 |
| **Sometimes** | 17529 |
| **no** | 279 |

***Count***

|  |  |
| --- | --- |
| **MTRANS** | |
| **Automobile** | 3534 |
| **Bike** | 32 |
| **Motorbike** | 38 |
| **Public\_Transportation** | 16687 |
| **Walking** | 467 |

***Count***

|  |  |
| --- | --- |
| **CALC** | |
| **Frequently** | 529 |
| **Sometimes** | 15066 |
| **no** | 5163 |

***Count***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **family\_history\_with\_overweight** | **FAVC** | **SMOKE** | **SCC** |
| **no** | 3744 | 1776 | 20513 | 20071 |
| **yes** | 17014 | 18982 | 245 | 687 |

***Count***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***id*** | ***Age*** | ***Height*** | ***Weight*** | ***FCVC*** | ***NCP*** | ***CH2O*** | ***FAF*** | ***TUE*** |
| ***count*** | 20758 | 20758 | 20758 | 20758 | 20758 | 20758 | 20758 | 20758 | 20758 |
| ***mean*** | 10378.5 | 23.8418 | 1.700245 | 87.88777 | 2.445908 | 2.761332 | 2.029418 | 0.981747 | 0.616756 |
| ***std*** | 5992.463 | 5.688072 | 0.087312 | 26.37944 | 0.533218 | 0.705375 | 0.608467 | 0.838302 | 0.602113 |
| ***min*** | 0 | 14 | 1.45 | 39 | 1 | 1 | 1 | 0 | 0 |
| ***25%*** | 5189.25 | 20 | 1.631856 | 66 | 2 | 3 | 1.792022 | 0.008013 | 0 |
| ***50%*** | 10378.5 | 22.81542 | 1.7 | 84.06488 | 2.393837 | 3 | 2 | 1 | 0.573887 |
| ***75%*** | 15567.75 | 26 | 1.762887 | 111.6006 | 3 | 3 | 2.549617 | 1.587406 | 1 |
| ***max*** | 20757 | 61 | 1.975663 | 165.0573 | 3 | 4 | 3 | 3 | 2 |

**Summary Statistics of Numerical Columns**

**A graph of blue bars

Description automatically generated A graph of a number of people

Description automatically generated**

**A graph with red and blue squares

Description automatically generated**

A graph with text on it

Description automatically generated

**A graph of a graph with blue squares and black text

Description automatically generated with medium confidence**

**A diagram of a graph

Description automatically generated**

**A diagram of a graph

Description automatically generated with medium confidence**

**A graph of blue rectangular objects

Description automatically generated**

**A graph of blue rectangular objects

Description automatically generated with medium confidence**

**A diagram of a graph

Description automatically generated**

**A graph of blue rectangular objects

Description automatically generated**

**A graph of blue rectangular objects

Description automatically generated**

**Appendix-Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **LR Coefficient** | **Decision Tree Importance** | **Random Forest Importance** | **XGBoost Importance** |
| **Age** | 0.170 | 0.068 | 0.097 | 0.018 |
| **Height** | 1.456 | 0.142 | 0.094 | 0.035 |
| **Weight** | -0.411 | 0.448 | 0.328 | 0.220 |
| **FCVC** | 2.147 | 0.017 | 0.073 | 0.103 |
| **NCP** | 2.789 | 0.017 | 0.032 | 0.020 |
| **CH2O** | 1.194 | 0.025 | 0.047 | 0.022 |
| **FAF** | 1.416 | 0.019 | 0.045 | 0.011 |
| **TUE** | 1.248 | 0.014 | 0.046 | 0.012 |
| **Gender\_Female** | 0.445 | 0.005 | 0.039 | 0.300 |
| **Gender\_Male** | 0.370 | 0.190 | 0.048 | 0.000 |
| **family\_history\_with\_overweight\_no** | 0.976 | 0.002 | 0.015 | 0.016 |
| **family\_history\_with\_overweight\_yes** | -0.161 | 0.005 | 0.021 | 0.000 |
| **FAVC\_no** | 0.118 | 0.007 | 0.007 | 0.032 |
| **FAVC\_yes** | 0.697 | 0.004 | 0.008 | 0.000 |
| **CAEC\_Always** | -0.051 | 0.002 | 0.003 | 0.012 |
| **CAEC\_Frequently** | 0.946 | 0.005 | 0.013 | 0.013 |
| **CAEC\_Sometimes** | -0.051 | 0.003 | 0.015 | 0.011 |
| **CAEC\_no** | -0.029 | 0.002 | 0.004 | 0.051 |
| **SMOKE\_no** | 0.831 | 0.000 | 0.001 | 0.006 |
| **SMOKE\_yes** | -0.016 | 0.001 | 0.001 | 0.000 |
| **SCC\_no** | 0.666 | 0.001 | 0.004 | 0.021 |
| **SCC\_yes** | 0.149 | 0.001 | 0.003 | 0.000 |
| **MTRANS\_Automobile** | -0.318 | 0.004 | 0.009 | 0.015 |
| **MTRANS\_Bike** | -0.003 | 0.000 | 0.000 | 0.004 |
| **MTRANS\_Motorbike** | -0.004 | 0.000 | 0.000 | 0.002 |
| **MTRANS\_Public\_Transportation** | 1.113 | 0.003 | 0.010 | 0.010 |
| **MTRANS\_Walking** | 0.027 | 0.001 | 0.003 | 0.009 |
| **CALC\_Always** | 0.000 | 0.000 | 0.000 | 0.000 |
| **CALC\_Frequently** | -0.059 | 0.001 | 0.004 | 0.008 |
| **CALC\_Sometimes** | 0.493 | 0.004 | 0.015 | 0.010 |
| **CALC\_no** | 0.381 | 0.008 | 0.014 | 0.040 |

**Comparison of Model Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Class** | **precision** | **recall** | **f1-score** | **support** |
| **Logistic Regression** | **Insufficient\_Weight** | 0.83 | 0.8 | 0.82 | 272 |
| **Normal\_Weight** | 0.52 | 0.55 | 0.54 | 287 |
| **Obesity\_Type\_I** | 0.57 | 0.52 | 0.54 | 351 |
| **Obesity\_Type\_II** | 0.76 | 0.91 | 0.83 | 297 |
| **Obesity\_Type\_III** | 0.9 | 0.99 | 0.95 | 324 |
| **Overweight\_Level\_I** | 0.55 | 0.47 | 0.5 | 290 |
| **Overweight\_Level\_II** | 0.48 | 0.44 | 0.46 | 290 |
| **Random Forest** | **Insufficient\_Weight** | 0.97 | 0.96 | 0.97 | 272 |
| **Normal\_Weight** | 0.91 | 0.9 | 0.9 | 287 |
| **Obesity\_Type\_I** | 0.96 | 0.93 | 0.95 | 351 |
| **Obesity\_Type\_II** | 0.94 | 0.99 | 0.96 | 297 |
| **Obesity\_Type\_III** | 1 | 0.99 | 1 | 324 |
| **Overweight\_Level\_I** | 0.88 | 0.87 | 0.87 | 290 |
| **Overweight\_Level\_II** | 0.87 | 0.9 | 0.89 | 290 |
| **Decision Tree** | **Insufficient\_Weight** | 0.93 | 0.95 | 0.94 | 272 |
| **Normal\_Weight** | 0.89 | 0.79 | 0.84 | 287 |
| **Obesity\_Type\_I** | 0.9 | 0.9 | 0.9 | 351 |
| **Obesity\_Type\_II** | 0.9 | 0.97 | 0.94 | 297 |
| **Obesity\_Type\_III** | 1 | 0.99 | 1 | 324 |
| **Overweight\_Level\_I** | 0.79 | 0.82 | 0.8 | 290 |
| **Overweight\_Level\_II** | 0.82 | 0.8 | 0.81 | 290 |
| **KNN** | **Insufficient\_Weight** | 0.87 | 0.97 | 0.92 | 272 |
| **Normal\_Weight** | 0.86 | 0.7 | 0.77 | 287 |
| **Obesity\_Type\_I** | 0.85 | 0.92 | 0.89 | 351 |
| **Obesity\_Type\_II** | 0.95 | 0.97 | 0.96 | 297 |
| **Obesity\_Type\_III** | 1 | 1 | 1 | 324 |
| **Overweight\_Level\_I** | 0.78 | 0.79 | 0.78 | 290 |
| **Overweight\_Level\_II** | 0.82 | 0.77 | 0.79 | 290 |
| **XGBoost** | **Insufficient\_Weight** | 0.98 | 0.98 | 0.98 | 272 |
| **Normal\_Weight** | 0.95 | 0.9 | 0.92 | 287 |
| **Obesity\_Type\_I** | 0.96 | 0.95 | 0.96 | 351 |
| **Obesity\_Type\_II** | 0.95 | 0.99 | 0.97 | 297 |
| **Obesity\_Type\_III** | 1 | 0.99 | 1 | 324 |
| **Overweight\_Level\_I** | 0.88 | 0.92 | 0.9 | 290 |
| **Overweight\_Level\_II** | 0.92 | 0.92 | 0.92 | 290 |

**Evaluation: Train = Synthetic | Test = Original**