Final Report

By: Baylie Schnieder, Cassady Jackson, Mady McKee, and Morgan Lind

GitHub with workflows: https://github.com/bayliemschnieder/Data-Mining-Project

As the Universal Superstar Hospital Team, we are looking to investigate factors that could lead to a patient being underweight, overweight or obese to better understand patient health on a holistic level. We are hoping that by building models that can predict what category an individual would be in based on their lifestyle choices, we can eliminate time in waiting rooms and get to the root of health issues faster. Our doctors are on busy schedules, and the faster that they are able to understand trends in a patient's life, the better they can be at understanding and diagnosing issues that may be overlooked without the predictions.

Specifically with this project we would like to understand what factors could predict if a person would be underweight, overweight or obese. These findings will be used to make better recommendations in yearly physicals for lifestyle changes and potential trends, as well as rule out certain diseases or ailments in diagnosis. As seen, these findings will not be one size fits all, but we would like to be as accurate in our modeling as possible.

The data that we are using for this exploration is a collection of data collected by data scientists at UC Irvine[1] for the estimation of obesity levels in individuals from Mexico, Peru and Colombia, based on their eating habits and physical condition. There are sixteen features in the data and 2111 instances. Twenty-three percent of the data was collected directly through users on a web platform and seventy-seven percent was generated synthetically using a Weka tool. There are no missing values.

Data Dictionary and Features:

[1] archive.ics.uci.edu/dataset

Original Name	Rename	Role	Type	Description
Gender	Gender	Feature	Categorical	
Age	Age	Feature	Continuous	
Height	Height	Feature	Continuous	
Weight	Weight	Feature	Continuous	
Family_history_with_overweight	Family_History	Feature	Binary	Has a family member suffered or suffers from overweight?
FAVC	High_Cal	Feature	Binary	Do you eat high caloric food frequently?
FCVC	Veggies	Feature	Integer	Do you usually eat vegetables in your meals?
NCP	Meals	Feature	Continuous	How many main meals do you have daily?
CAEC	Snacks	Feature	Categorical	Do you eat any food between meals?
SMOKE	SMOKE	Feature	Binary	Do you smoke?
CH20	Water_Intake	Feature	Continuous	How much water do you drink daily?
SCC	Monitor_Cals	Feature	Binary	Do you monitor the calories you eat daily?
FAF	Exercise	Feature	Continuous	How often do you have

				physical activity?
TUE	Tech_Use	Feature	Integer	How much time do you use technological devices?
CALC	Alcohol	Feature	Categorical	How often do you drink alcohol?
MTRANS	Transportation	Feature	Categorical	Which transportation do you usually use?
NObeyesdad	Obesity_Level	Target	Categorical	Obesity Level

Prepare Data

For our original data exploration, we used a separate orange file to edit domain and rename the columns to be easier to understand. We then exported this file, with the renamed columns and used it for our first data exploration. When creating our final model, we worked with the original dataset, and had to rename columns again to match this edited csv file. We did this to ensure that the data used in the final model was accurate and in the final model, split this data into training and non-training data, to have 80% training, and 10% testing and validation. For our final model we ignored height and weight as we found they were overfit.

Explore Data

To start exploring the data, we created a new orange file and imported the dataset. To clean the data, we changed the column names to be more descriptive about what the column

represents. We applied an 80 /20 split for training and non-training data. Using the training data, we examined the correlations among all variables and found a moderate correlation of 0.465 between height and weight. Given this finding, we excluded height and weight from the rest of the analysis to build a more accurate model. These two features also had potential for data leakage when trying to predict obesity levels as they are generally the main classifiers for doctors.

To gain a better understanding of the dataset, we continued with exploratory data analysis. To start this process, we created two boxplot visualizations, one showing the relationship between age and obesity level and the relationship between transportation preference by obesity level. The 'Age vs Obesity Level' plot revealed a consistent increasing trend in average age across all categories. Individuals classified as Insufficient Weight had an average age of 19.7164, while Obesity Type II category had an average age of 28.2154. This visualization demonstrates a progressive increase in age as obesity levels heightened. The transportation preference boxplot showed that across all obesity categories, most individuals preferred public transportation or automobiles. However, the preference of walking was more frequently observed among individuals in the Normal Weight category.

To further explore the class distributions, we created two distribution plots, one for obesity levels and one for transportation preferences. The obesity level distribution displayed that the most frequently occurring category was Obesity Type I with 281 individuals. The remaining categories showed a relatively even distribution, each ranging from 220 and 245 observations. In the transportation preference distribution, public transportation was the dominant choice, with 1265 individuals representing 74.90% percent of the dataset. These exploratory insights provided

a strong basis for selecting relevant features and supporting the development and evaluation for a precise model.

Additionally, we created a scatter plot displaying the relationship between ‘Exercise’ and ‘Technology Use’ with each point representing an observation from our dataset. The points on the scatter plot are color-coded based on ‘Exercise’ column values. From the distribution, we can observe a wide spread of technology use values across different exercise levels, indicating there are no obvious trends or correlations. However, further analysis would be provided to determine whether any correlations or relationships exist. We attempted to make scatterplots with many other variables in the data, but due to their proximity in distributions, we did not find anything meaningful with them.

We also explored the data by creating individual workflows for each model to test with the data. This gave us a chance to understand which would be best for our goals, without having a large, confusing workflow. Through these tests we found that linear regression and k-means testing did not work well with our data; however, logistic regression, random forest, decision trees and gradient boosting did work well with our data.

Linear Regression

When putting our variables in a linear regression the model had to be spilt into multiple numeric variables for prediction because our target variable is a categorical variable rather than numerical. We chose to create dummy variables for each obesity level and run linear regressions on each to see if it was possible to predict obesity level based on the variables given. Due to the nature of our dataset and the predictions we are hoping to make, linear regression did not work as

well as other models. It was not included in the final model, because we are not looking to predict just one of the obesity level variables, but categorize individuals based on behaviors.

Logistic Regression

In testing the data with a multiple logistic regression, we chose to continue to ignore the height and weight variables and make obesity level our target variable. We sampled data with an 80% training, 20% testing split. We also preprocessed the data, continuing discrete variables to one feature per variable. Following all this data cleaning, we were able to run a multiple logistic regression using the Ridge (L2) regularization type and a strength of C=1. For testing on training data, the logistic regression had a classification accuracy of 0.632 and an area under the curve of 0.894. These scores are available in *Figure 2* in the appendix. The classification accuracy and area under the curve mean that the model is performing somewhat well in predicting training data, but not overfit. For testing on validation data, the area under the curve was 0.891 and the classification accuracy was 0.632. These numbers are available in *Figure 4* as well. We did not test this model on testing data because it was not the best performing model of all that we tested. These scores mean that the model is performing well in predicting testing data as well, which is a good sign. However, the decision tree had better scores overall and was our final model which we tested with the additionally left out testing data.

K-Means

When using k-means testing, we found that it was not an effective way to make predictions with our data. The scatterplots produced did not have clear groups between each obesity level, other than for height and weight, which were already correlated. When coloring by cluster there was no clear differentiation between them, and there was a lot of overlap no matter

which variable was used. Therefore, after exploring the data we chose not to continue using k-means modeling in our data predictions going forward.

Decision Tree

For the decision tree, we used a minimum number of instances in leaves of 2, not splitting subsets smaller than 5, limiting the maximal tree depth to 100, and stopping when the majority reaches 95%. The tree first splits on servings_of_vegetables, indicating it is the most important predictor. Subsequent splits involve Age and Technology_use. The decision tree with 4 levels of depth is shown in the following figure:

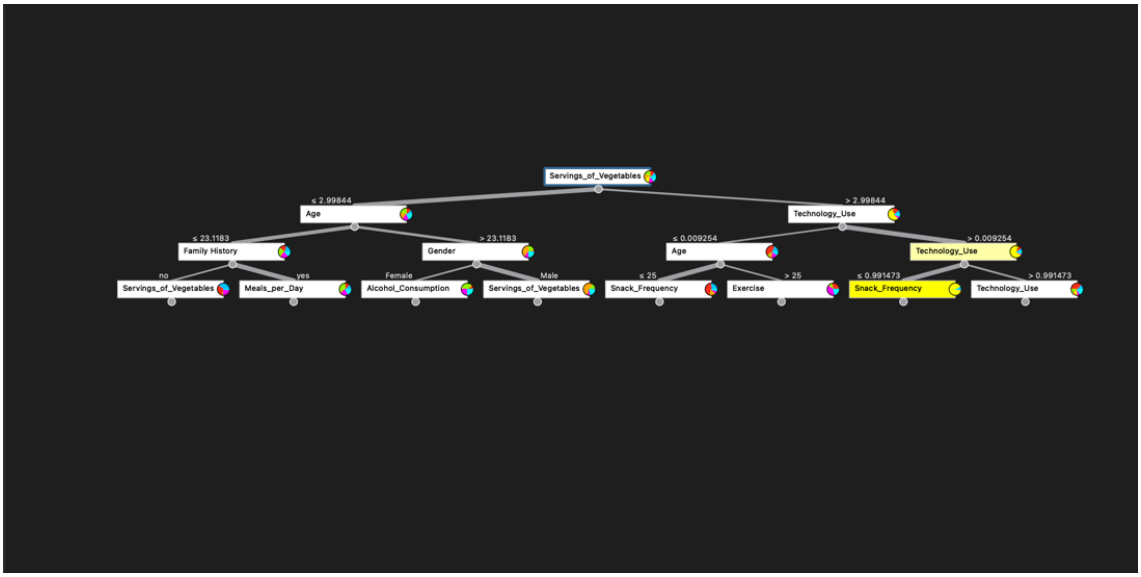


Figure 1: Final Tree Model

Random Forest

For our Random Forest model, we used a total of 200 trees, 5 attributes considered at each split, limiting depth of individual trees to 4, and not splitting subsets smaller than 5. The Test & Score results are listed in *Figure 2*, *Figure 3*, and *Figure 4*. We originally created three Random Forest models, each with different parameters, and tested their scores to see which

model performed the best. The most accurate model parameters were used for the final workflow. When comparing with the other models in the workflow, Random Forest did not perform the best for this dataset.

Gradient Boosting

For our Gradient Boosting model, we used a total of 20 trees, a learning rate of 0.100, limiting depth of individual trees to 3, and not splitting subsets smaller than 2. The Test & Score results are listed in *Figure 2*, *Figure 3*, and *Figure 4*. To decide on these numbers, we originally created three different Gradient Boosting models, with different levels of trees, and tested them for accuracy on our data. When performing these tests, the model with our current parameters was the best performing model, which is why we chose to include it in the final workflow. This gave us the best chance when comparing across models for finding a model that would work to predict the levels. When comparing across all the other techniques we used, gradient boosting was not the best for this data set.

Conclusion

Based on the models we created, the decision tree was the best fit for classifying individuals with training, testing and validation data. Therefore, when doctors are looking to better understand where a patient's level may be, they can start by following the branches of the decision tree. This model does not have perfect fit, at approximately 70-80% accuracy on all training, testing and validation, so it will not be a perfect predictor with lifestyle factors alone, but is a good place to start. Key factors to obesity level, such as height and weight were not included in building the model due to their potential data leakage and overfitting. These factors, in real prediction scenarios, would need to be included to get a full picture of the patients' health.

However, with just lifestyle choices and family history alone, the decision tree gives us a good idea of where a patient is and is heading, and doctors can use it for their predictions.

Appendix

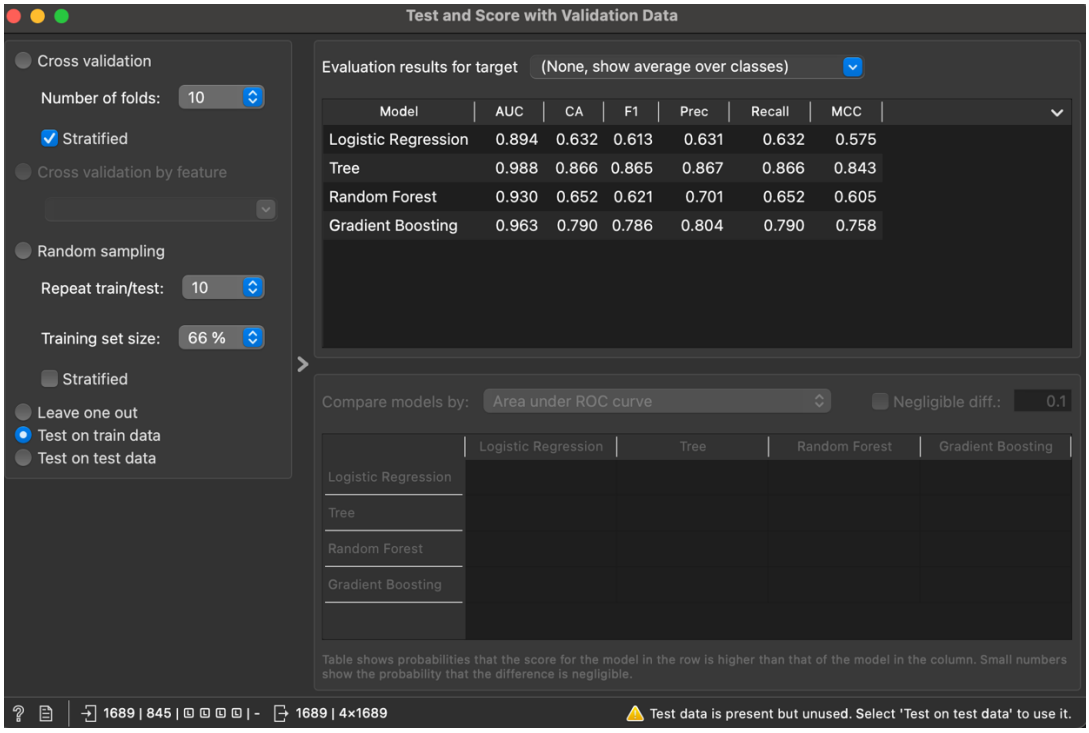


Figure 2: Test on Train Data

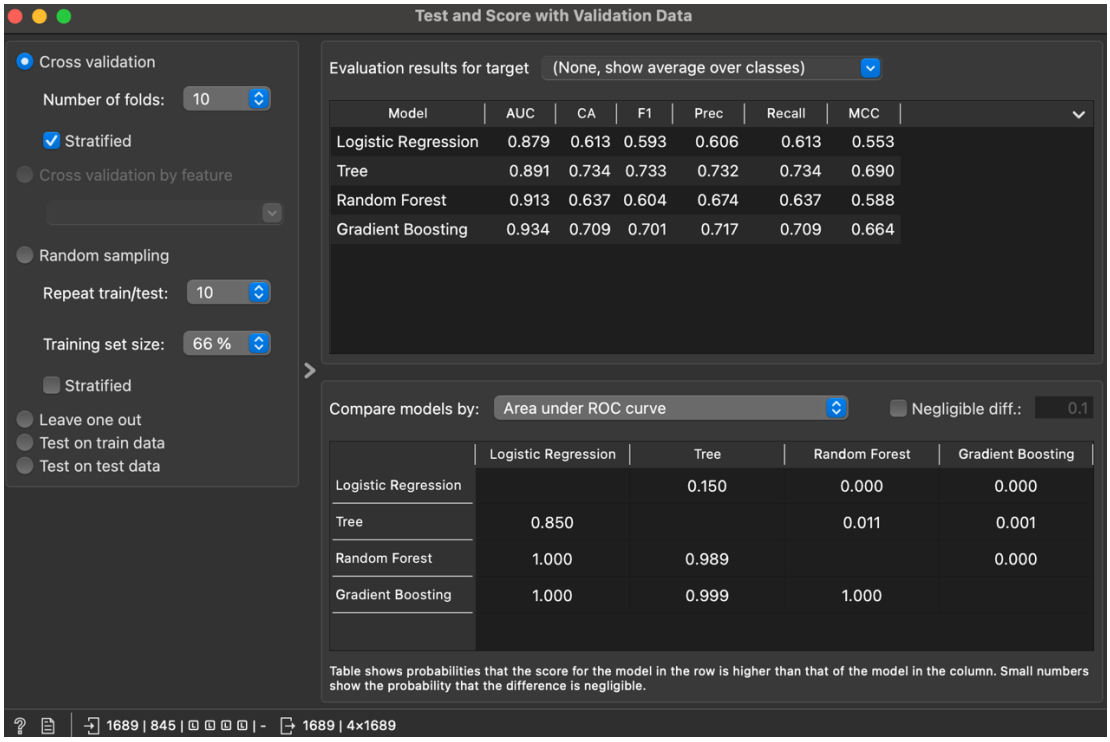


Figure 3: Cross Validation on Training and Validation Data

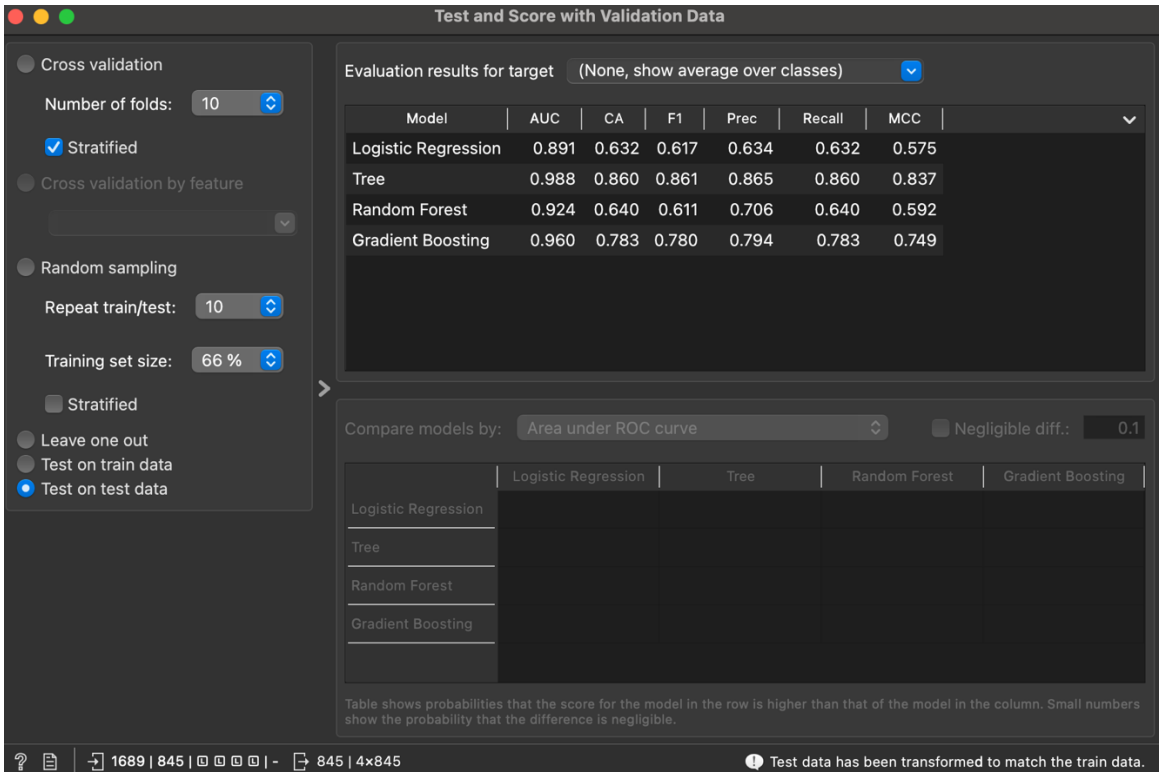


Figure 4: Test on Testing Data

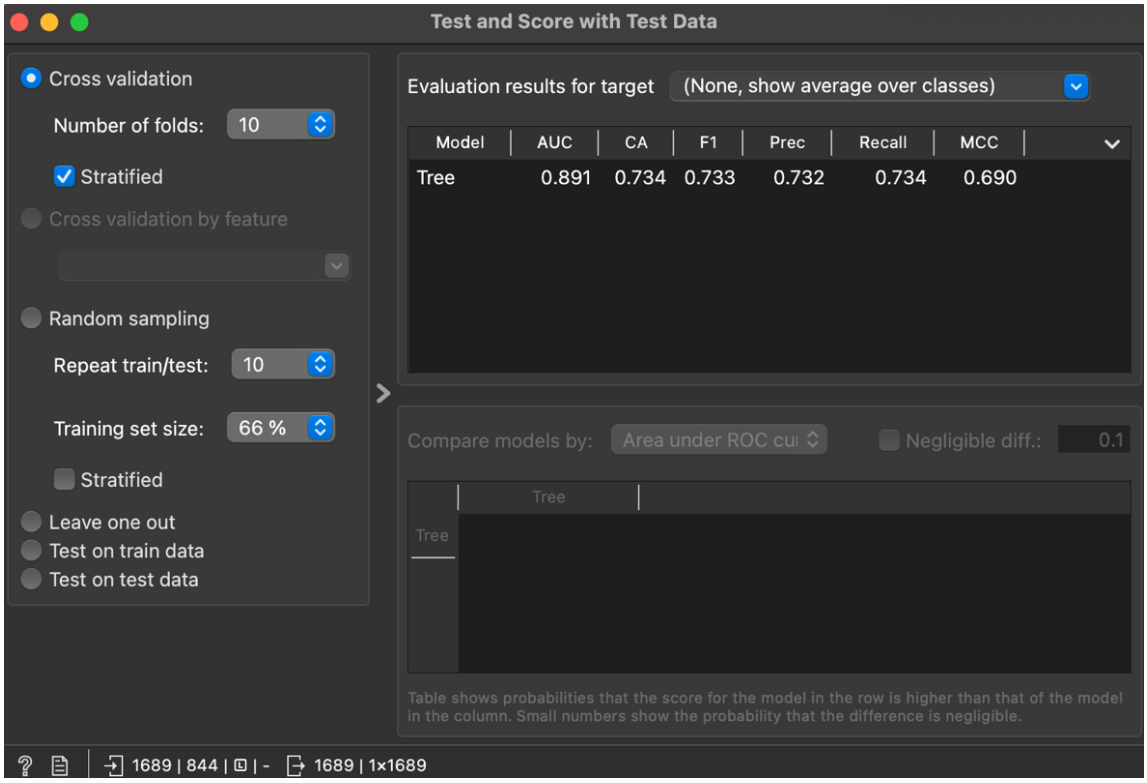


Figure 5: Cross Validation with Testing Data

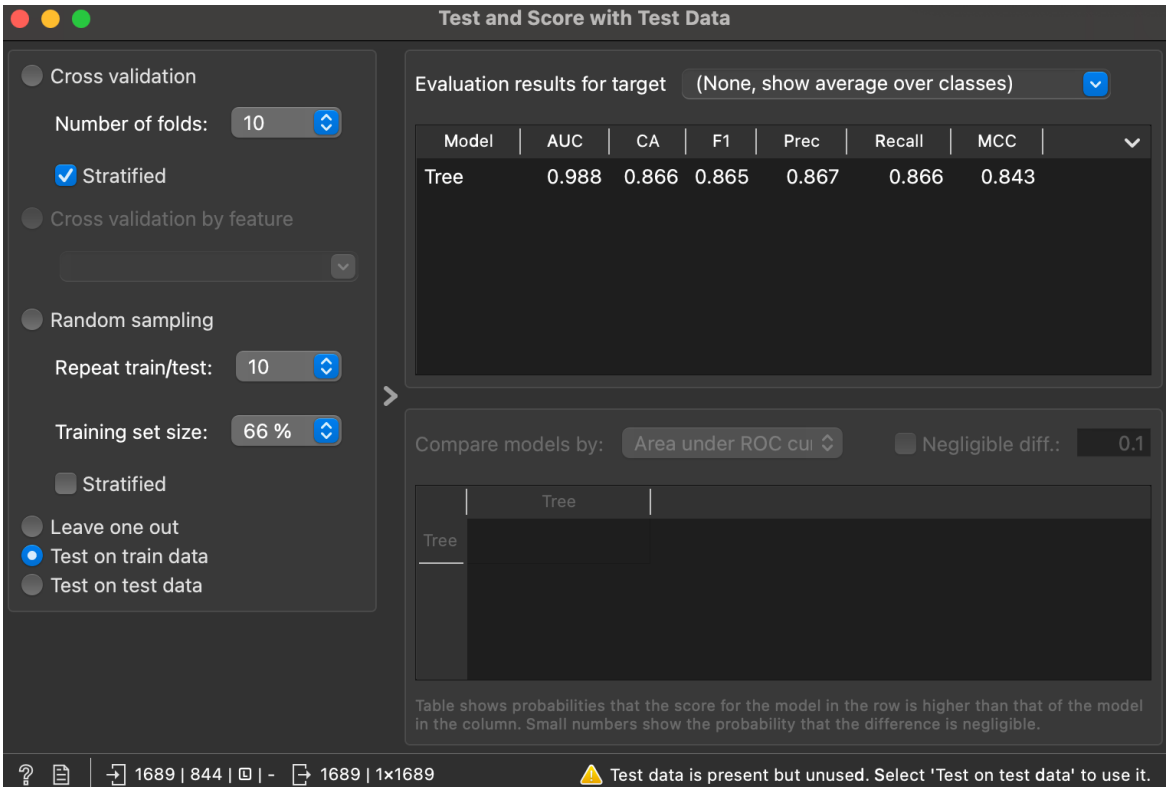


Figure 6: Test on Train Data

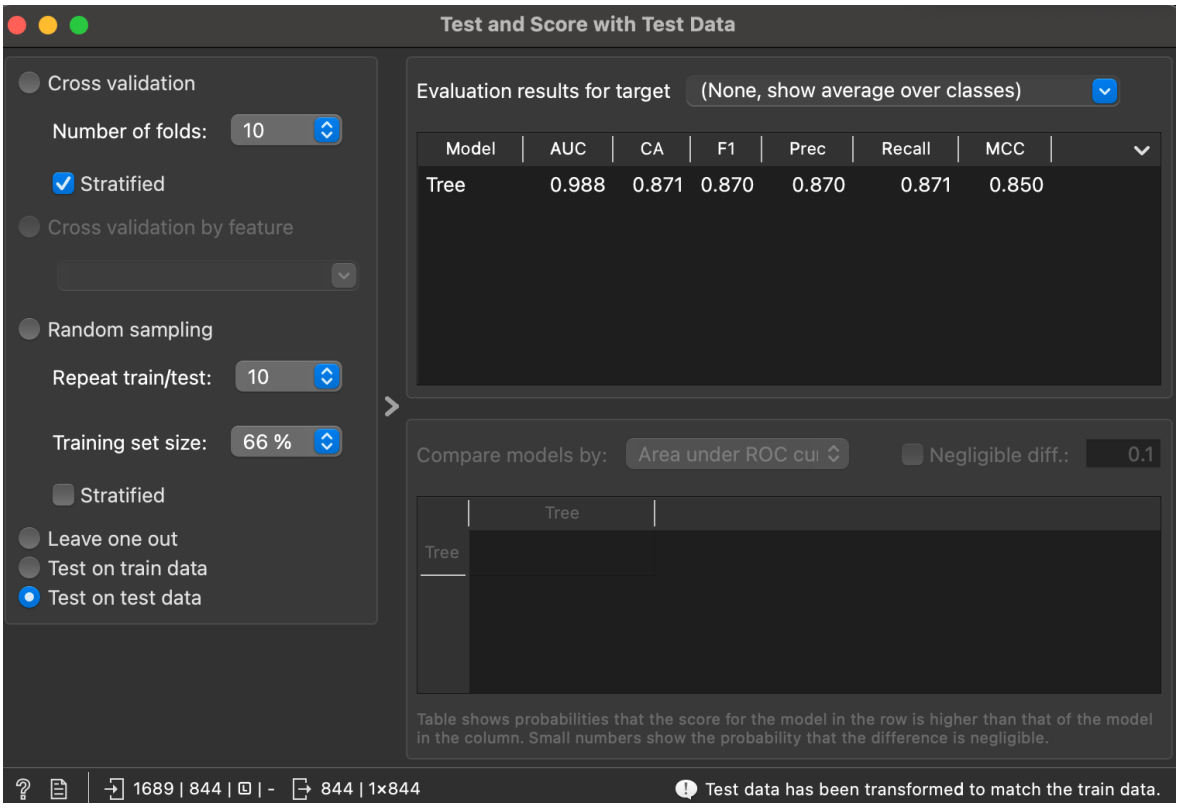


Figure 7: Test on Test Data

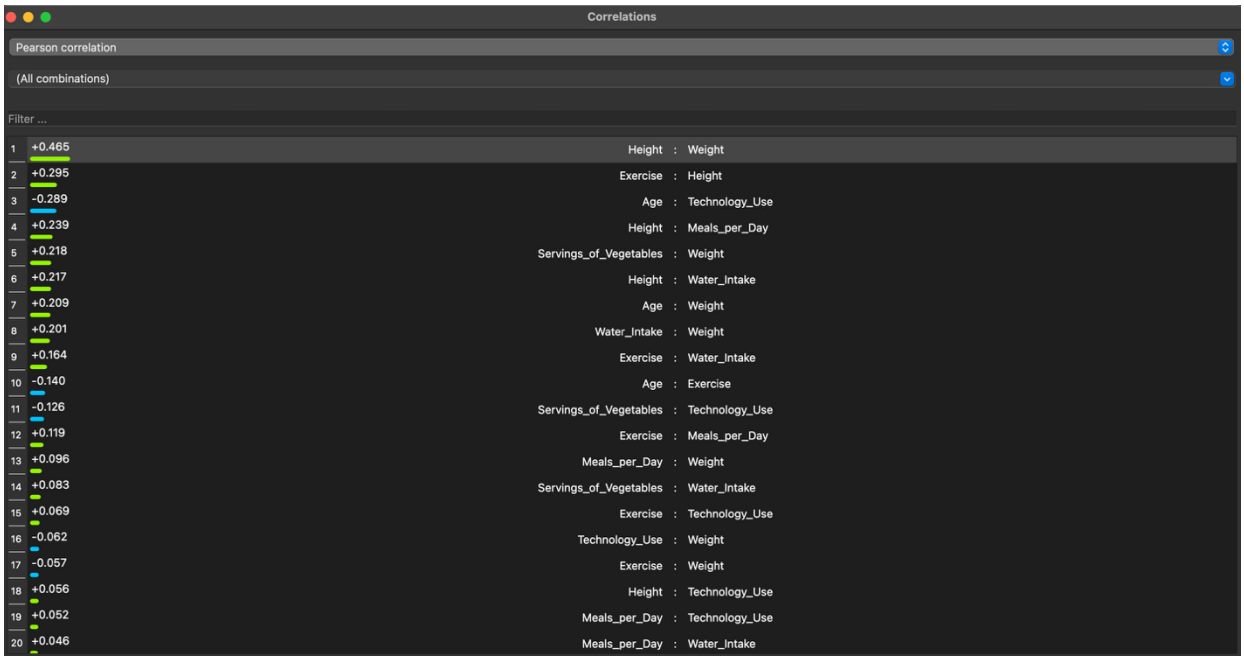


Figure 87: Correlations

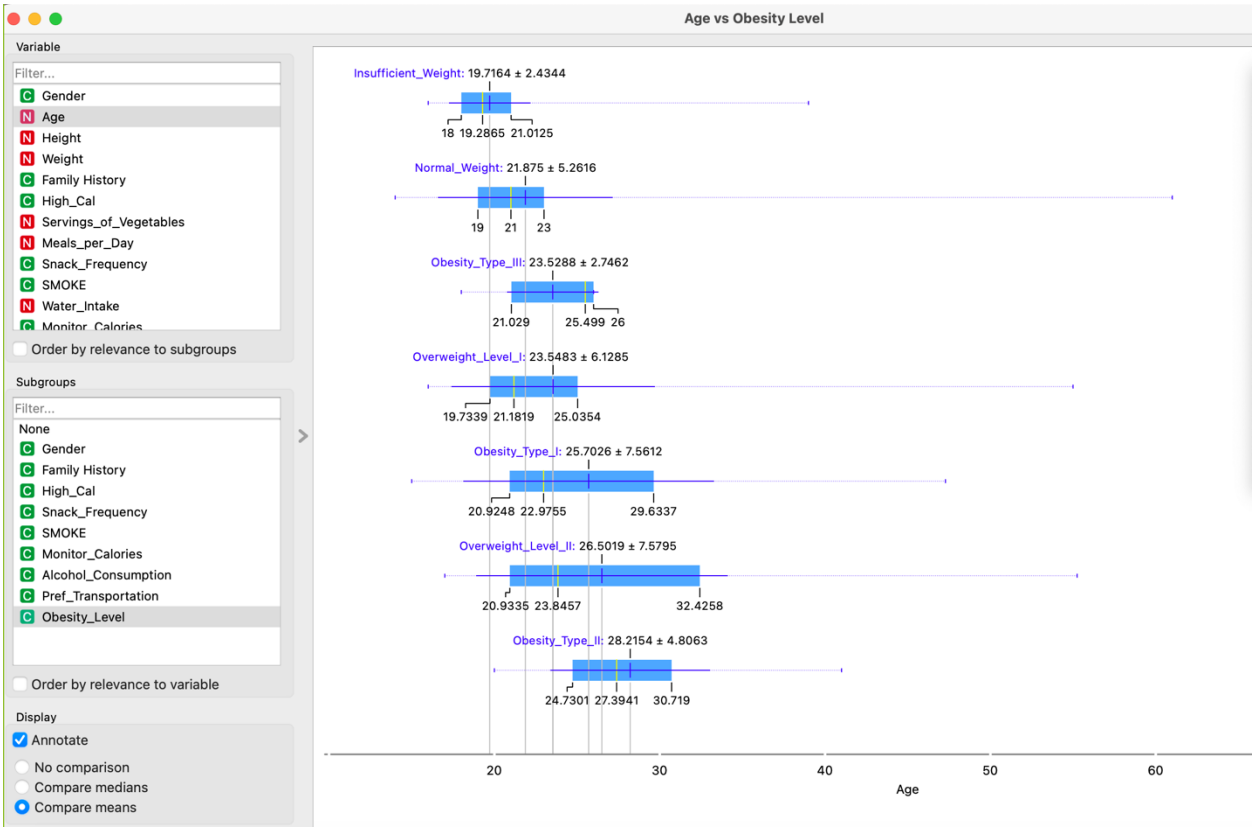


Figure 98: Box plot of Age vs Obesity Levels

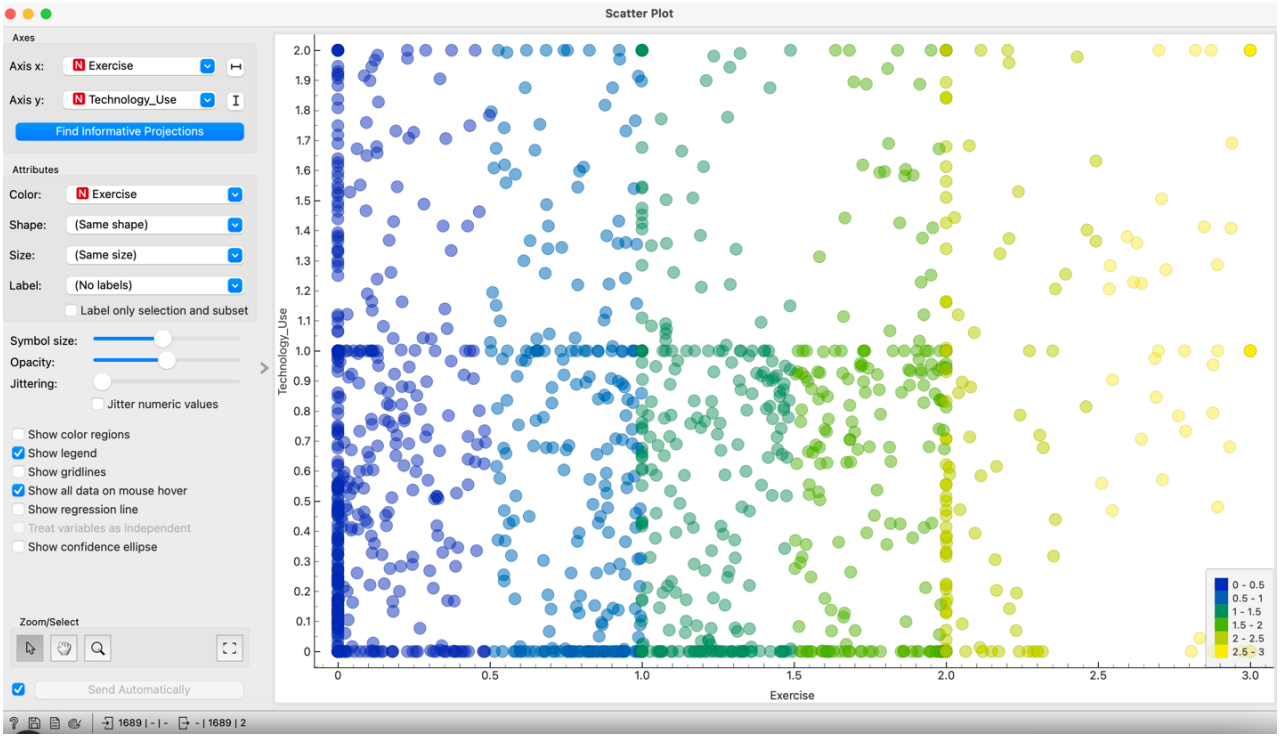


Figure 109: Scatterplot of Exercise vs Tech Use, Colored by Exercise

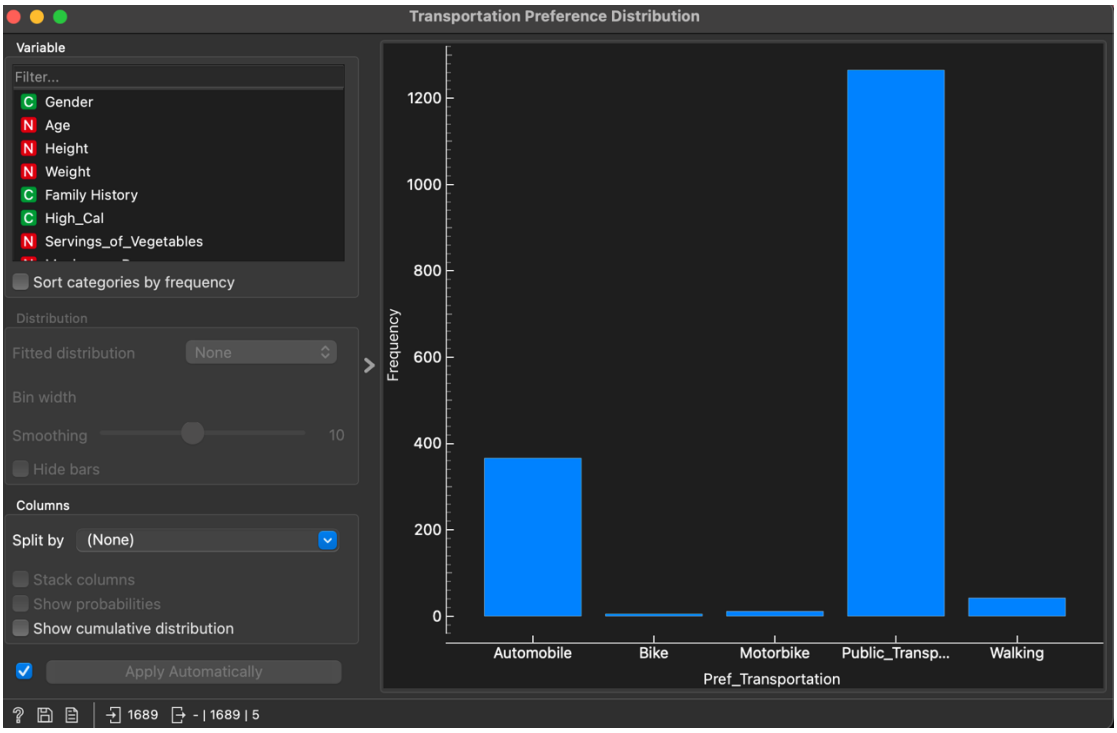
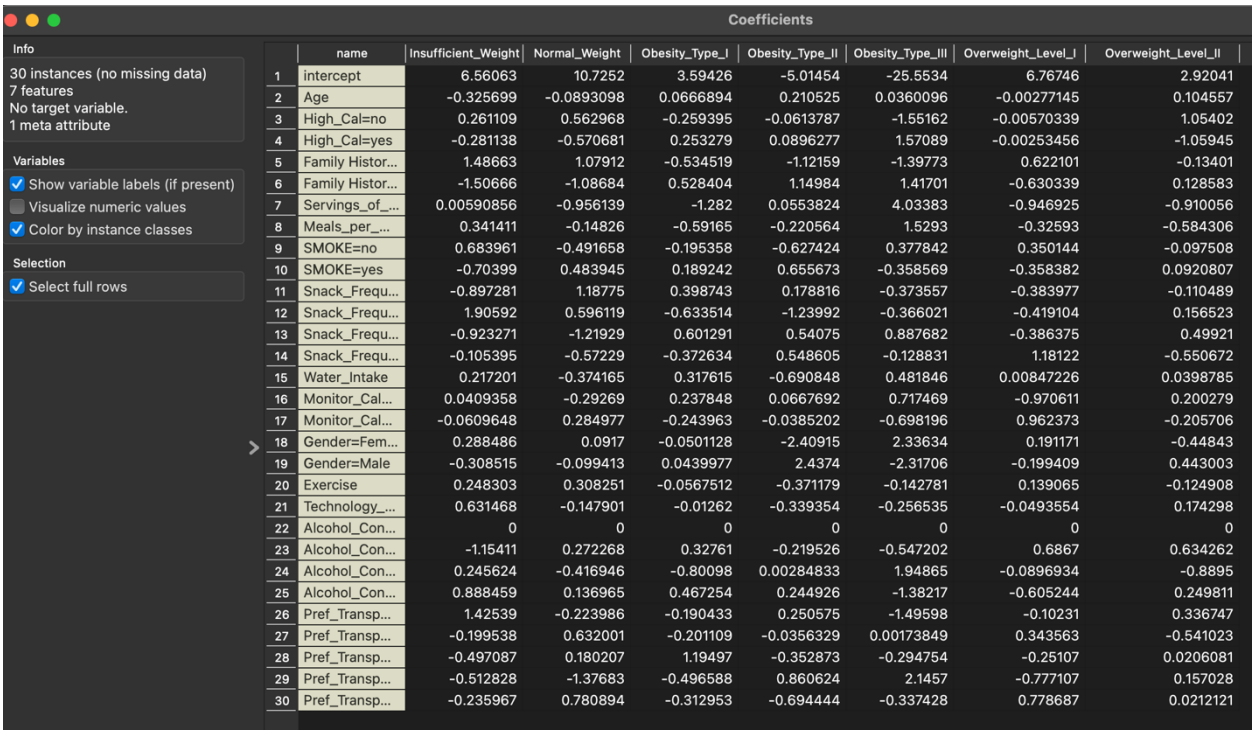


Figure 1110: Transportation Preference Distribution

Figure 1211: Box Plot of Transportation Preference by Obesity Level



	name	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
1	intercept	6.56063	10.7252	3.59426	-5.01454	-25.5534	6.76746	2.92041
2	Age	-0.325699	-0.0893098	0.0666894	0.210525	0.0360096	-0.00277145	0.104557
3	High_Cal=no	0.261109	0.562968	-0.259395	-0.0613787	-1.55162	-0.00570339	1.05402
4	High_Cal=yes	-0.281138	-0.570681	0.253279	0.0896277	1.57089	-0.00253456	-1.05945
5	Family Histor...	1.48663	1.07912	-0.534519	-1.12159	-1.39773	0.622101	-0.13401
6	Family Histor...	-1.50666	-1.08684	0.528404	1.14984	1.41701	-0.630339	0.128583
7	Servings_of_...	0.00590856	-0.956139	-1.282	0.0553824	4.03383	-0.946925	-0.910056
8	Meals_per_...	0.341411	-0.14826	-0.59165	-0.220564	1.5293	-0.32593	-0.584306
9	SMOKE=no	0.683961	-0.491658	-0.195358	-0.627424	0.377842	0.350144	-0.097508
10	SMOKE=yes	-0.70399	0.483945	0.189242	0.655673	-0.358569	-0.358382	0.0920807
11	Snack_Frequ...	-0.897281	1.18775	0.398743	0.178816	-0.373557	-0.383977	-0.110489
12	Snack_Frequ...	1.90592	0.596119	-0.633514	-1.23992	-0.366021	-0.419104	0.156523
13	Snack_Frequ...	-0.923271	-1.21929	0.601291	0.54075	0.887682	-0.386375	0.49921
14	Snack_Frequ...	-0.105395	-0.57229	-0.372634	0.548605	-0.128831	1.18122	-0.550672
15	Water_Intake	0.217201	-0.374165	0.317615	-0.690848	0.481846	0.00847226	0.0398785
16	Monitor_Cal...	0.0409358	-0.29269	0.237848	0.0667692	0.717469	-0.970611	0.200279
17	Monitor_Cal...	-0.0609648	0.284977	-0.243963	-0.0385202	-0.698196	0.962373	-0.205706
18	Gender=Fem...	0.288486	0.0917	-0.0501128	-2.40915	2.33634	0.191171	-0.44843
19	Gender=Male	-0.308515	-0.099413	0.0439977	2.4374	-2.31706	-0.199409	0.443003
20	Exercise	0.248303	0.308251	-0.0567512	-0.371179	-0.142781	0.139065	-0.124908
21	Technology_...	0.631468	-0.147901	-0.01262	-0.339354	-0.256535	-0.0493554	0.174298
22	Alcohol_Con...	0	0	0	0	0	0	0
23	Alcohol_Con...	-1.15411	0.272268	0.32761	-0.219526	-0.547202	0.6867	0.634262
24	Alcohol_Con...	0.245624	-0.416946	-0.80098	0.00284833	1.94865	-0.0896934	-0.8895
25	Alcohol_Con...	0.888459	0.136965	0.467254	0.244926	-1.38217	-0.605244	0.249811
26	Pref_Transp...	1.42539	-0.223986	-0.190433	0.250575	-1.49598	-0.10231	0.336747
27	Pref_Transp...	-0.199538	0.632001	-0.201109	-0.0356329	0.00173849	0.343563	-0.541023
28	Pref_Transp...	-0.497087	0.180207	1.19497	-0.352873	-0.294754	-0.25107	0.0206081
29	Pref_Transp...	-0.512828	-1.37683	-0.496588	0.860624	2.1457	-0.777107	0.157028
30	Pref_Transp...	-0.235967	0.780894	-0.312953	-0.694444	-0.337428	0.778687	0.0212121

Figure 1312: Multiple Logistic Regression Coefficients

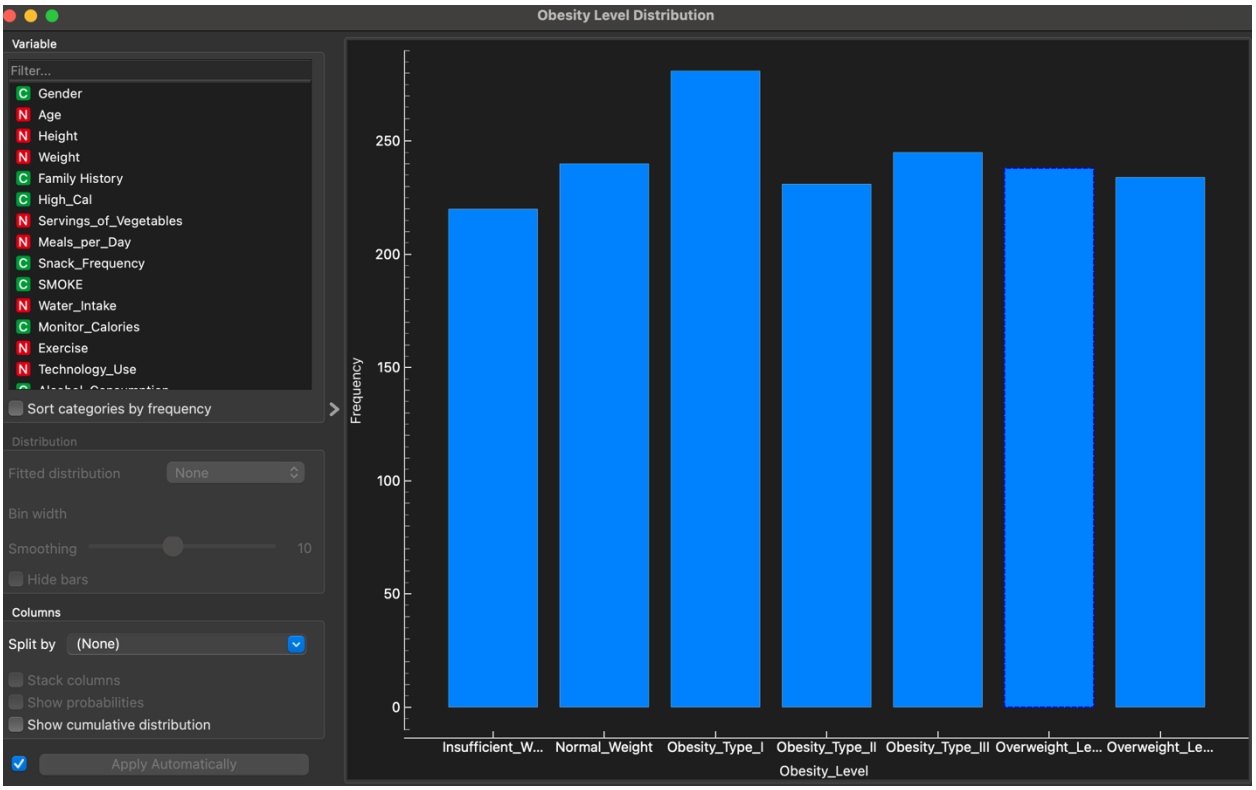


Figure 14: Obesity Level Distribution