Anum Damani
Winter 2024

Obesity Project: Supplemental Write-Up

Question 1: What dataset did you use and what are the features in the dataset?

Answer to Question 1:

For this project, I used an obesity dataset that I found on Kaggle (https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster). The original features in the dataset are 'Gender', 'Age', 'Height', 'Weight', 'family_history_with_overweight', 'FAVC', 'FCVC', 'NCP', 'CAEC', 'SMOKE', 'CH2O', 'SCC', 'FAF', 'TUE', 'CALC', 'MTRANS'. The original dataset contained 2,111 observations and 16 features.

I decided to change most of the feature names for convenience. I decided to exclude "mode of transportation" from my analysis. Since "mode of transportation" has multiple levels, it's not possible to convert it to a binary variable. Also, as part of the data cleaning/pre-processing, I converted some categorical variables to binary. Below is a table of all features that I decided to use for model training, along with variable type and description.

Name of Feature	*Variable Type*	*Description*
is_female	binary	Whether or not the person is female Yes = 1; No = 0
weight	numerical	Weight of person
age	numerical	Age of person
height	numerical	Height of person
overweight_family_binary	binary	Whether or not the person's family member is overweight Yes = 1; No = 0
binary_freq_high_calorie_food	binary	Whether or not the person frequently consumes high calorie food Yes = 1; No = 0
binary_calorie_monitoring	binary	Whether or not the person monitors their calories Yes = 1; No = 0

binary_smoke	binary	Whether or not the person is a smoker Yes = 1; No = 0
freq_vegetable	numerical	Frequency of consumption of vegetables
binary_consume_food_btwn_meals	binary	Whether or not the person consumes food between meals Yes = 1; No = 0
binary_freq_alcohol_consumption	binary	Whether or not the person consumes alcohol Yes = 1; No = 0
daily_h2o_consumption	numerical	Daily water consumption
num_main_meals	numerical	Number of main meals
freq_physical_activity	numerical	Frequency of physical activity
time_used_technology	numerical	Time spent using technology

Question 2: What method(s) did you try and why did you select your final choice of method? Please explain the underlying principles of the technique that you used.

Answer to Question 2:

My research question is: Is it possible to predict whether or not someone is obese using the following 15 predictors: is_female, weight, age, height, overweight_family_binary, binary_freq_high_calorie_food, binary_calorie_monitoring, binary_smoke, freq_vegetable, binary_consume_food_btwn_meals, binary_freq_alcohol_consumption, daily_h2o_consumption, num_main_meals, freq_physical_activity, freq_physical_activity, time_used_technology? Note that I consider the "full model" to be a model with all 15 predictors.

I would expect that the predictor "weight" would play a more significant role (in comparison to the other features) in determining whether someone is obese or not. So, in addition to my research question, I wanted to answer another question. Suppose we did not have access to "weight" as a predictor of obesity. Would we still be able to predict obesity just by looking at the other 14 features?

My target variable is "obesity_binary" which is a binary variable. I created this variable using the "obesity_level" variable that was in the original dataset. I mapped the levels "normal weight" and "insufficient weight" to 0 and all of the obesity levels to 1. Since my target variable

is a binary variable, I tried logistic regression and decision tree classifier. Specifically, I tried four models:

- Model 1: Logistic Regression (Full Model)
  - 15 predictors: is_female, weight, age, height, overweight_family_binary, binary_freq_high_calorie_food, binary_calorie_monitoring, binary_smoke, freq_vegetable, binary_consume_food_btwn_meals, binary_freq_alcohol_consumption, daily_h2o_consumption, num_main_meals, freq_physical_activity, freq_physical_activity, time_used_technology
- Model 2: Logistic Regression (Model minus Weight Variable)
  - 14 predictors: is_female, age, height, overweight_family_binary, binary_freq_high_calorie_food, binary_calorie_monitoring, binary_smoke, freq_vegetable, binary_consume_food_btwn_meals, binary_freq_alcohol_consumption, daily_h2o_consumption, num_main_meals, freq_physical_activity, freq_physical_activity, time_used_technology
- Model 3: Decision Tree Classifier (Full Model)
  - 15 predictors: is_female, weight, age, height, overweight_family_binary, binary_freq_high_calorie_food, binary_calorie_monitoring, binary_smoke, freq_vegetable, binary_consume_food_btwn_meals, binary_freq_alcohol_consumption, daily_h2o_consumption, num_main_meals, freq_physical_activity, freq_physical_activity, time_used_technology
- Model 4: Decision Tree Classifier (Model minus Weight Variable)
  - 14 predictors: is_female, age, height, overweight_family_binary, binary_freq_high_calorie_food, binary_calorie_monitoring, binary_smoke, freq_vegetable, binary_consume_food_btwn_meals, binary_freq_alcohol_consumption, daily_h2o_consumption, num_main_meals, freq_physical_activity, freq_physical_activity, time_used_technology

After training the models, testing on the testing set, and printing out accuracies, my strongest models in terms of model accuracy were the Logistic Regression (Full Model) and the Decision Tree Classifier (Full Model). Between these two, the Decision Tree Classifier (Full Model) outperformed the Logistic Regression (Full Model). I decided to focus most on model accuracy, but I still looked at precision, recall, F1 score, and Pseudo R-Squared.

Logistic Regression is a supervised classification method that can help us predict whether an observation belongs to a certain class or not. Logistic regression estimates the probability that an event will occur, based on independent features. The actual value of the observation y will either be 0 or 1. This method is appropriate for my dataset because I want to predict whether or

not a person is obese. My features are a mix of numerical and categorical, and logistic regression can handle both types of features.

Decision Tree Classifier is a supervised classification method that can be used to predict or model categorical data. The tree considers all possible partitions of the data and chooses the optimal split that minimizes the Gini index or entropy. In other words, each node in the decision tree tests an attribute. Similar to logistic regression, this method is appropriate for my dataset because I want to predict whether or not a person is obese. My features are a mix of numerical and categorical, and decision tree classifiers can handle both types of features.

To evaluate model performance, I looked at accuracy, precision, recall, and F1 Score. I also looked at Pseudo R-Squared for Logistic Regression models, and MSE for Decision Tree Classifiers. After analyzing these metrics, I determined that Model 3 Decision Tree Classifier (Full Model) is the best model. The accuracy score is measurement for how well the model generalizes to new, unseen data (the test set); so, accuracy score is the (number of accurate predictions)/(total number of predictions). Precision is a measurement for how well the model avoids false positives; so the formula for precision is (true positives)/(true positives + false positives). Recall is another metric for the performance of the model, and can be calculated this way: (true positives)/(true positives + false negatives). F1 Score is calculated using this formula: 2(Precision X Recall)/(Precision + Recall). F1 Score is a metric that is the average of precision and recall. Higher values of model accuracy, precision, recall, and F1 score indicate better model performance. In logistic regression, Pseudo R-Squared is a measure of the goodness of fit of the model, and a higher Pseudo R-squared means a better fit. Smaller MSE is better for models since MSE measures the average squared difference between predictions and actual values. (Please see my answer for Question 5 for more details).

Question 3: Were there any considerations or choices you made with regards to missing data and/or other data processing/cleaning?

Answer to Question 3:

The original obesity dataset did not have any missing values. In terms of data pre-processing/cleaning, I changed multiple categorical variables to binary:

- I created “obesity_binary” using the “obesity_level” variable
- I created “overweight_family_binary” using the “overweight_family” variable
- I created “binary_freq_high_calorie_food” using the “freq_high_calorie_food” variable
- I created “binary_calorie_monitoring” using the “calorie_monitoring” variable
- I created “binary_smoke” using the “smoke” variable
- I created “binary_consume_food_btwn_meals” using the “consume_food_btwn_meals” variable

- I created “binary_freq_alcohol_consumption” using the “freq_alcohol_consumption” variable

I had to convert these categorical variables to binary so that I could run logistic regression.
I also dropped the “mode_transportation” variable because it had multiple levels and it did not make sense to convert it to a binary variable.

Question 4: Were there any challenges you faced in your implementation?

Answer to Question 4:
I did not face many challenges in the implementation.

Question 5: What were the outcomes of training your model? Was there a good fit? Did it work well with the training data? What was your accuracy score? Does it give good predictions?

Answer to Question 5:
Below, I have tables of performance metrics for the four models that I tried.

Performance Metrics for Logistic Regression Models

	Model 1: Logistic Regression (Full Model)	Model 2: Logistic Regression (minus Weight variable)
Accuracy	0.9551	0.8534
Precision	0.9604	0.8711
Recall	0.9813	0.9470
F1 Score	0.9707	0.9075
Pseudo R-Squared	0.9813	0.3391

Model 1 (full model) had an accuracy of 95.51%. Model 2 (without weight variable) had an accuracy of 85.34%.

Comparing Model 1 and Model 2, Model 1 outperformed Model 2. Model 1 has a higher accuracy, precision, recall, F1 Score, and Pseudo R-Squared than Model 2.

Performance Metrics for Decision Tree Classifier Models

	Model 3: Decision Tree Classifier (Full Model)	Model 4: Decision Tree Classifier (minus Weight variable)
Accuracy	0.9716	0.8510
MSE	0.0284	0.1489
Precision (Weighted Average)	0.97	0.85
Recall (Weighted Average)	0.97	0.85
F1 Score (Weighted Average)	0.97	0.85

Model 3 has an accuracy of 97.16%. Model 4 (without weight variable) has an accuracy of 85.10%.

Comparing Model 3 and Model 4, Model 3 outperformed Model 4. Model 3 has a higher accuracy, precision, recall, and F1 Score than Model 4. Also, Model 3 has much lower MSE than Model 4.

Comparing Model 1 and Model 3, Model 3 outperformed Model 1 in terms of model accuracy. Model 1 has an accuracy of 95.51% and Model 3 has an accuracy of 97.16%.

Therefore, the best model is Model 3 Decision Tree Classifier (Full Model) because it has the highest model accuracy. It was a good fit and worked well with the training data. The accuracy score was 97.16%, which is very strong. In other words, this model can predict whether or not someone is obese with 97.16% accuracy. So, it gives good predictions. Also, the MSE is 0.0284 which is quite low and good. The precision, recall, and F1 Score are 0.97, which is also quite strong. Therefore, Model 3 is the best model to use to predict obesity.