Building a Bayesian Network

Obesity Report

Bruno Jerković, Josip Koprčina, Christos Mavrikis

Problem Domain

Obesity is defined as the accumulated fat in one's body that can cause major health risks. To classify whether or not someone is obese, a formula named BMI (Body mass index) is used. Body mass index is a person's weight in kilograms (kg) divided by the square of their height in meters (m). If one's BMI is calculated to be bigger than 30.0, they are then deemed obese [1].

According to a plethora of global estimates and reports, obesity has almost tripled since 1975 to a point where, in 2016, 13% of the world's adult population was considered obese. A review on obesity done by the World Health Organization found that an astonishing 38 million children were considered either overweight or obese as of 2019 [2]. In order to gain a better understanding of obesity, numerous research projects about it and its risks have been published. For this project we shall implement a Bayesian Network model, concerning the attributes that may cause obesity, using a publicly available dataset.

Data Explanation

For the project we have used the dataset from the Machine Learning Repository, called 'Estimation of obesity levels based on eating habits and physical condition Data set' [4]. The dataset contains records of estimated obesity levels and it is sampled from several countries of South America (Mexico, Peru and Colombia) with 2111 instances. Below (Figure 1.) is the list and explanation of features/attributes available in the dataset.

Feature Name	Feature Type
Gender	Categorical-Dichotomous: Male or Female
Age	Continuous: 14yo – 61yo
Height	Continuous: 1.45m – 1.98m
Weight	Continuous: 39kg – 173kg
Presence of a family relative with overweight history – family_history_with_overweight	Categorical-Dichotomous: Yes or No
Frequent consumption of high caloric food - FAVC	Categorical-Dichotomous: Yes or No
Frequency of consumption of vegetables - FCVC	Categorical-Ordinal: 1 or 2 or 3 (Never, Sometimes, Always)
Number of main meals - NCP	Categorical-Ordinal: 1 or 2 or 3 or 4 (One main meal, Two main meals, Three main meals, More than three main meals)

Consumption of food between meals - CAEC	Categorical-Ordinal: 1 or 2 or 3 or 4 (No, Sometimes, Frequently, Always)
SMOKE	Categorical-Dichotomous: Yes or No
Consumption of water daily - CH20	Categorical-Ordinal: 1 or 2 or 3 (Less than a Liter, Between 1L and 2L, More than 2L)
Calories consumption monitoring - SCC	Categorical-Dichotomous: Yes or No
Number of days in a week doing physical activity (Physical activity frequency) - FAF	Categorical-Ordinal: 0 or 1 or 2 or 3 (0 days, 1 or 2 days, 2 or 4 days, 4 or 5 days)
Hours a day spent on using technology devices (Time using technology devices) - TUE	Categorical-Ordinal: 0 or 1 or 2 (0-2hours, 3-5hours, more than 5hours)
Consumption of alcohol - CALC	Categorical-Ordinal: No, Sometimes, Frequently, Always
Type of transportation used - MTRANS	Categorical-Nominal: Walking, Bike, Public Transportation, Motorbike, Automobile
Obesity level - NObeyesdad	Categorical-Ordinal: Insufficient_Weight, Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Level_I, Obesity_Level_II , Obesity_Level_III

Figure 1: The table above displays the data before any possible pre-processing action took place

Data Explanation – Preprocessing

To start with, we have first observed our dataset manually, as a brief check of some possible visible inconsistencies. We found out that, after the 500th instance, some of the feature value numbers change their number type. Some variables that are integers change to decimal type (such as Age, Weight, CH20, FAF and TUE), while the Height variable gets added four more decimals. The issue can be seen in figure 2. To prevent any further problems that might occur, we decided to round the decimal variables to the closest integer number (using R's "round" function) so the data is more consistent.

Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
Female	25.90228	1.669701	104.586	yes	yes	3	3	Sometimes	no	1.570188	no	0.210351	0.881848	Sometimes	Public_Transportation	Obesity_Type_III
Female	25.54087	1.668709	104.755	yes	yes	3	3	Sometimes	no	1.412049	no	0.143955	0.753077	Sometimes	Public_Transportation	Obesity_Type_III
Female	20.52099	1.668642	124.705	yes	yes	3	3	Sometimes	no	1.15635	no	0.786828	0.366385	Sometimes	Public_Transportation	Obesity_Type_III
Male	20.87115	1.690614	129.769	yes	yes	3	3	Sometimes	no	1.154698	no	1.71836	0.959218	Sometimes	Public_Transportation	Obesity_Type_III
Female	21.76883	1.733383	135.525	yes	yes	3	3	Sometimes	no	1.485736	no	1.950374	0.869238	Sometimes	Public_Transportation	Obesity_Type_III
Female	20.89149	1.748313	133.574	yes	yes	3	3	Sometimes	no	2.874336	no	1.624981	0.825609	Sometimes	Public_Transportation	Obesity_Type_III
Male	20.94194	1.812963	138.731	yes	yes	3	3	Sometimes	no	2.641489	no	0.481555	0.735201	Sometimes	Public_Transportation	Obesity_Type_III
Female	20.98902	1.80734	155.872	yes	yes	3	3	Sometimes	no	2.417122	no	0.952725	0.573958	Sometimes	Public_Transportation	Obesity_Type_III
Female	18.36748	1.745644	133.666	yes	yes	3	3	Sometimes	no	2.900857	no	1.508897	0.625371	Sometimes	Public_Transportation	Obesity_Type_III
Female	21.05111	1.753266	133.852	yes	yes	3	3	Sometimes	no	2.95311	no	1.445148	0.67321	Sometimes	Public_Transportation	Obesity_Type_III

Figure 2: Random sample of the dataset after 500th instance – image from excel

Furthermore, since we have a dataset which consists of both continuous and categorical data, we have decided to group continuous values into categories. More precisely, we fit our data to the following groups:

- Age Grouping: “<20”, “20-34”,”35-49” and “50>=”
- Height Grouping: “<1.60”, “1.60-1.74”, “1.75-1.84” and “1.85>=”
- Weight Grouping: “<50”, “50-69”, “70-89” and “90>=”
- Weight Levels Grouping: “Normal Weight”, “Overweight” and “Obese”

Finally, the features in the dataset were assigned a numeric value. We did this as we believed it was the simplest way to handle our data. Figure 3. shows an image of what our dataset looks like in RStudio console after all of the processing has been implemented.

```
> head(train.set)
  Gender Age Height Weight family_history_with_overweight FAVC FCVC NCP CAEC SMOKE CH2O SCC FAF TUE CALC MTRANS NObeyesdad
1      1   2      2      2                                2    1    2   3    2     1    2   1   0   1    1      3          1
2      1   2      1      2                                2    1    3   3    2     2    3   2   3   0    2      3          1
3      2   2      3      3                                2    1    2   3    2     1    2   1   2   1    3      3          1
4      2   2      3      3                                1    1    3   3    2     1    2   1   2   0    3      1          3
5      2   2      3      4                                1    1    2   1    2     1    2   1   0   0    2      3          3
6      2   2      2      2                                1    2    2   3    2     1    2   1   0   0    2      5          1
```

Figure 3: A few instances after pre-processing – image from R studio

Data Explanation – About the instances & our Hypothesis

By inspecting the dataset we have calculated that the instances are in the age group of 14-61 years old (with the mean of 24.31 and the standard deviation of 6.34), in the height group of 1.45m-1.98m (with the mean of 1.7 and standard deviation of 0.09) and in the weight group of 39kg-173kg (with the mean of 86.58 and the standard deviation of 23.19).

Additionally, by inspecting the categorical data, we have seen that the dataset is well distributed between males and females, unlike for some of the other features. The categorical frequencies of the features have a higher bias towards one value, which can be seen in Figure 4.

From the mean values of the Height and Weight attributes, the BMI score is 29.8, which is the borderline score between overweightness and obesity. Therefore, we assume that most of the instances in the dataset would either be classified as overweight or obese.

Feature name	Feature Value	
Family history with overweight	Yes	No
	1726	385
Smoke	Yes	No
	2067	44
Calories consumption monitoring - SCC	Yes	No
	2015	96

Consumption of food between meals - CAEC	No	Sometimes	Frequently	Always	
	51	1756	242	53	
Frequent consumption of high caloric food - FAVC	Yes		No		
	1866		245		
Type of transportation used - MTRANS	Automobile	Bike	Motorbike	Public Transportation	Walking
	457	7	11	1580	56

Figure 4: Table of categorical data with bias features

Building the Network

The network was built in the programming language R using the package *daggity* and following the book *R companion* [4]. Each node corresponds to a variable from our data and each edge represents a conditional dependency between the variables. Our dataset consists of 17 columns. From the very beginning it was evident to us that our network would start with 3 mutually independent nodes: Age, Gender and Family_history_with_overweight, because we thought that no other feature in our dataset could be connected to them. Using the prior knowledge of the BMI calculation, we decided that Height and Weight are the two only features that influence NObeyesdad (obesity) directly. Considering our prior knowledge about the features that, in our opinion, might influence height or weight, we have constructed the rest of the network. We then used d-separation to check if our assumptions about independencies were correct, and if they were not, we would remove the dependencies.

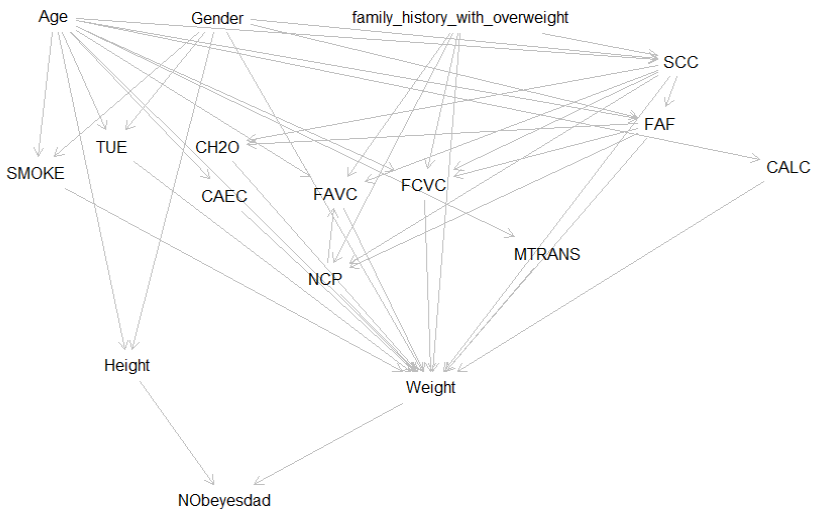

Figure 5: The Bayesian Network – plotted in R

Application – Programming Language and Libraries/Packages used

This project was implemented using R as the programming language. The imported libraries used in the project are: *lavaan*, *bnlearn*, *dagitty*, *bayesianNetworks* and *pROC*. *Dagitty* was used to create the Bayesian Network, *lavaan* to get more insight on the paths and coefficients, *bnlearn* and *bayesianNetworks* to fit the model and make predictions and *pROC* to build the ROC curve and compute the AUC. Below is a link to the script's github repository. Comments have been added to the script as well for better readability.

https://github.com/jkoprcina/BayesianNetworkFirstAssignment

Application – Phases

As we have proposed, our main objective is to predict one's obesity level, given the specific features in the dataset. In the earliest phase of the project, we divided our dataset into a training set (from instance 1:1000) and a testing set (from instance 1001:2111) and made Obesity Level predictions (based on the *R Companion* book). At that point in time, we had not yet done any form of preprocessing and we were getting some errors considering our DAG when fitting using *bnlearn*. Figure 6 shows the plot of our first result; the red line represents the train set (Area under the curve: 0.81503) and the blue line represents the test set (Area under the curve: 0.6202). It is evident that the test set predictions are worse than the train set predictions, but overall, we found the results unsatisfactory.

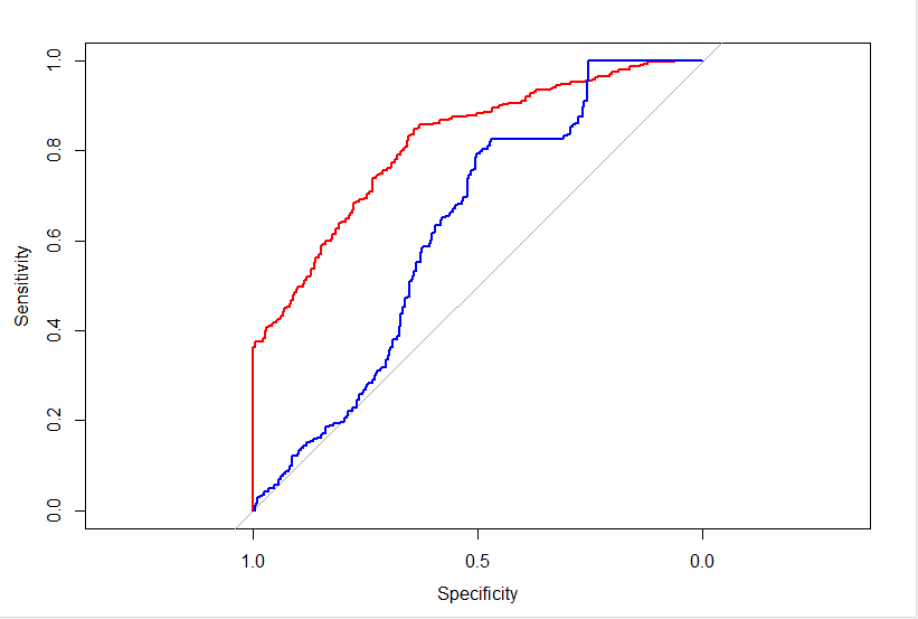

Figure 6: Our first prediction. Predicting Weight level

We realized that adaptions had to be made to improve the predictions. With time, we understood our dataset better and the types of variables, which affected our network. Many functions we used would produce errors or warnings because of different variable types (for example connecting Age

-> CALC – continuous -> ordinal). Thus, we performed the preprocessing actions described above explained in the paragraph "Data Explanation – Preprocessing ". After performing the preprocessing actions and running our script again, our results improved tremendously. Below is the plot (Figure 7.) of our improved result; the red line represents the train set (Area under the curve: 0.9962) and the blue line represents the test set (Area under the curve: 0.9009). From these results and the plot, it is evident that the predictions have improved on both sets, but especially the predictions on the testing set.

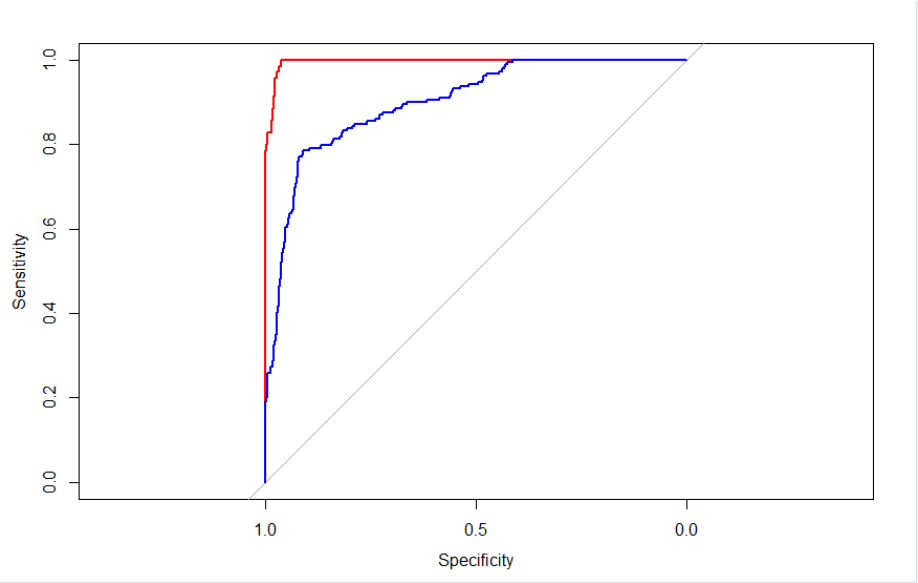

Figure 7: Our improved prediction. Predicting Weight Level

Additionally, we wanted to see if there was a better way to split the dataset into the parts used for training and testing, other than basing the decision upon the order of the instances in the dataset (as we have picked only the first 1900 instances to be in the train set). We first wanted to see how our model would work with different parts of dataset taken for testing. To that note, we performed a 10-fold cross validation while printing the ROC curve after every fold. In our opinion, the differences in the results were too small to matter, which led us to believe that the distribution of the feature values was dependent on their index in the dataset. Therefore, we have concluded the best decision should be stochastic, and we picked the instances for the train set randomly. Below is the plot (Figure 8.) of our improved result, the red line represents the train set (Area under the curve: 0.9974) and the blue line represents the test set (Area under the curve: 0.9988).

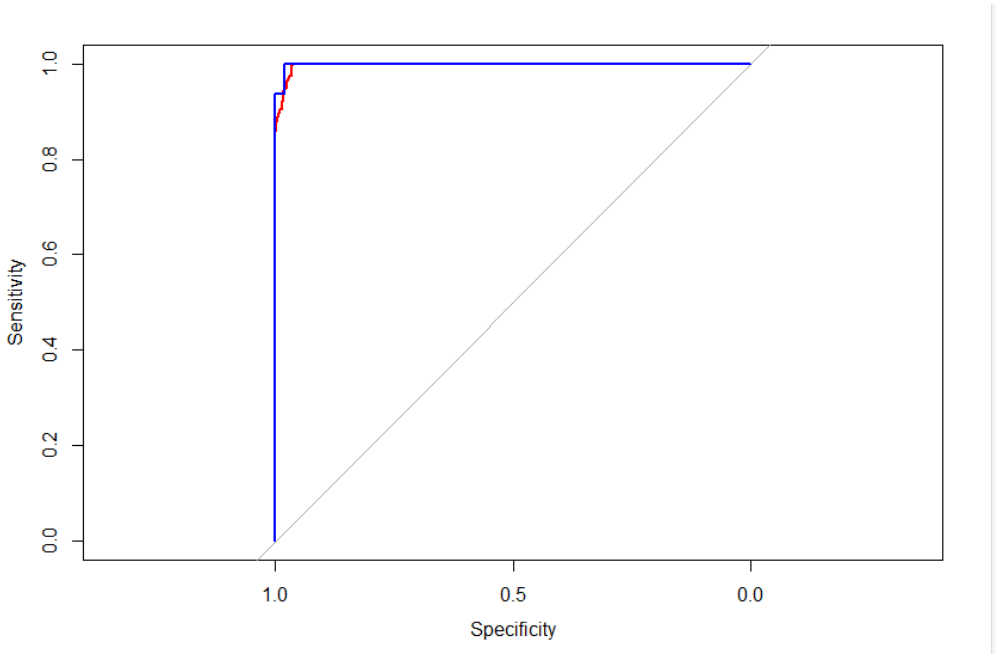

Figure 8: Prediction after random sampling. Predicting Weight Level

NOTE: because of the stochasticity in the train and test selection, the predictions may be slightly different when running

Application – Methods

The two most important decisions we had to make were choosing the methods for fitting and predicting. For fitting we used the *bn.fit* function with the method *mle* (Maximum Likelihood parameter). We chose *mle* as the other method, *bayes* (Bayesian parameter estimation), works only for discrete variables. We predicted using the predict function with the *bayes-lw* method included in the parameters. This is so we can predict any feature from any other given feature/features, even if they are not directly connected.

Analysis

We wanted to check the influence of Height and Weight on NObeyesdad as our main goal is to predict the NObeyesdad (Obesity levels) and the equation it is based on, the BMI equation, is the function of two variables Height and Weight. As we can see, weight influences the NObeyesdad feature more than Height, but as neither is close to 0, both are considered when predicting the obesity level. The figure (Figure 9) below shows the coefficients of all features.

Another underlying fact is that, surprisingly, family_history_with_overweight and Gender influenced Weight significantly more than the other features. We can also get insight about other

dependencies. We found it surprising that the FAF feature (physical activity) and NCP feature (Number of main meals) barely influence the Weight feature at all. Additionally, even though some coefficients seem to be very low we haven't removed the edges, as the coefficients are based on our dataset, and we strongly believe that if the dataset represented the population better, the coefficients would be higher (for example the NCP influence on the Weight feature).

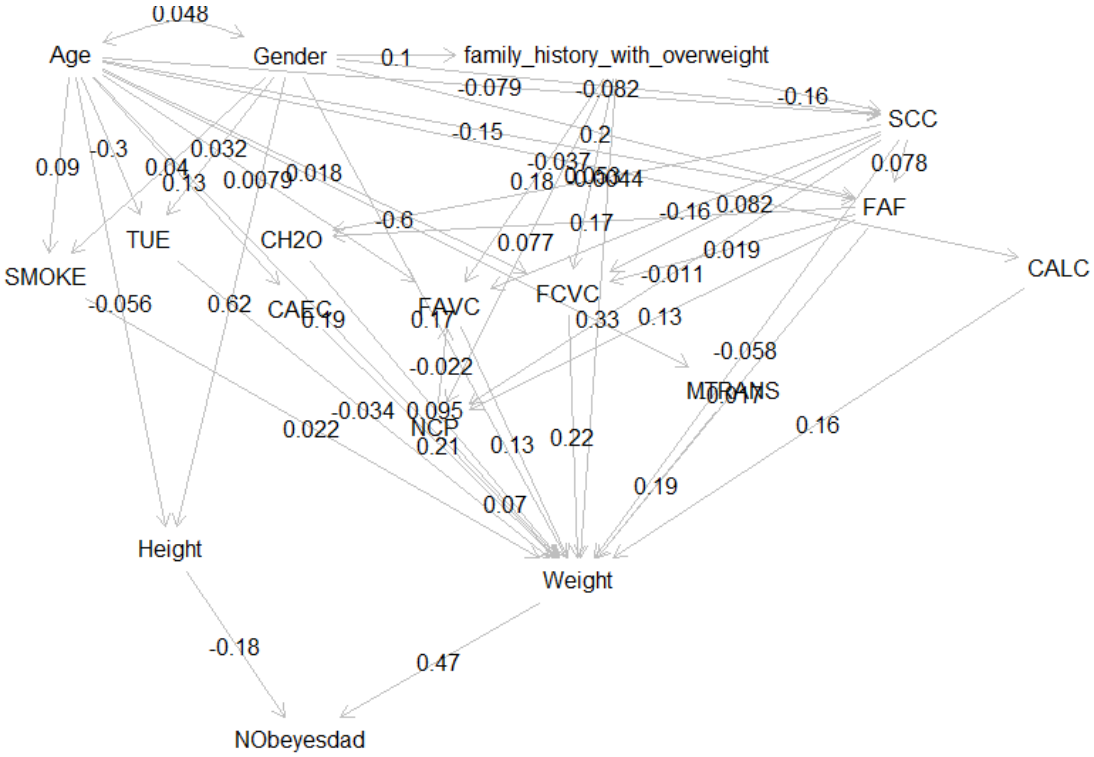

Figure 9: Network coefficients using lavaan fitting function

Discussion

Our initial goal was to be able to predict one's Obesity Level given certain features. The ROC curve that we get in our last phase shows that we have successfully completed our goal. Although, due to the feature value distribution in the dataset, and the correlation network, we believe that our dataset is not a good representation of the overall population. To achieve better, and more certain results, we would need to use a dataset that consists of a much broader population with a larger variety in features, that is not as biased as the one used here.

References

[1] C. f. D. CDS and Control, Body mass index (BMI), Sep. 2020. [Online]. Available: https://www.cdc.gov/healthyweight/assessing/bmi/index.html

[2] WHO, Obesity and overweight, Apr. 2020. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight

[3] Palechor, F. M., & de la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in Brief, 104344

[4] J. Textor, Bayesian Networks R Companion, 31.08.2020