Aishwarya Kotha, Nikita Mathur, Sofia Qian, Vivian Bui, Yijie Lian

Variables Affecting Obesity in Colombian, Mexican, and Peruvian Populations

1. Introduction

Obesity is a growing problem worldwide, and it is important to examine the obesity levels of people to determine how to best remedy this global issue. The World Health Organization states that the fundamental cause of obesity is an energy imbalance between calories consumed and calories expended. In this study, we will examine how people's eating and behavioral habits will affect their obesity levels as determined by mass body index. Our study focuses on data collected via a digital survey from Colombia, Peru, and Mexico. For this study, we chose to focus on the following classification problem: determining if a person will be obese or not given a set of lifestyles and physical health conditions.

2. About The Data

Our obesity-related data was collected from populations with ages between 14 and 61 and diverse eating habits and physics conditions in Colombia, Peru, and Mexico. The research team who collected the data used an anonymous web-based survey that was available for 30 days with unbiased questions. At the end of the surveying period, 485 records were received. The survey questions used are shown in Figure 1.

Table 1
Questions of the survey used for initial recollection of information.

Questions	Possible Answers
¿What is your gender?	• Female • Male
¿what is your age?	Numeric value
¿what is your height?	Numeric value in meters
¿what is your weight?	Numeric value in kilograms
¿Has a family member suffered or suffers from overweight?	• Yes • No
¿Do you eat high caloric food frequently?	• Yes • No
¿Do you usually eat vegetables in your meals?	• Never • Sometimes • Always
¿How many main meals do you have daily?	• Between 1 y 2 • Three • More than three
¿Do you eat any food between meals?	• No • Sometimes • Frequently • Always
¿Do you smoke?	• Yes • No
¿How much water do you drink daily?	• Less than a liter • Between 1 and 2 L • More than 2 L
¿Do you monitor the calories you eat daily?	• Yes • No
¿How often do you have physical activity?	• I do not have • 1 or 2 days • 2 or 4 days • 4 or 5 days
¿How much time do you use technological devices such as cell phone, videogames, television, computer and others?	• 0–2 hours • 3–5 hours • More than 5 hours
¿how often do you drink alcohol?	• I do not drink • Sometimes • Frequently • Always
¿Which transportation do you usually use?	• Automobile • Motorbike • Bike • Public Transportation • Walking

Figure 1: Survey questions used to collect data from populations of interest.

Once data are collected and labeled into different levels of obesity, the data collectors identified that the distribution of data was imbalanced across categories, as shown in Figure 2. This imbalance would pose difficulties for our classification task, as it might result in classifiers 'with a high accuracy but very low sensitivity towards the positive class' (Elhassan et al. 2017). For this reason, synthetic data was generated using Weka and SMOTE (synthetic minority over-sampling technique). The balanced distribution of the synthetic and original data combination is shown in Figure 3.

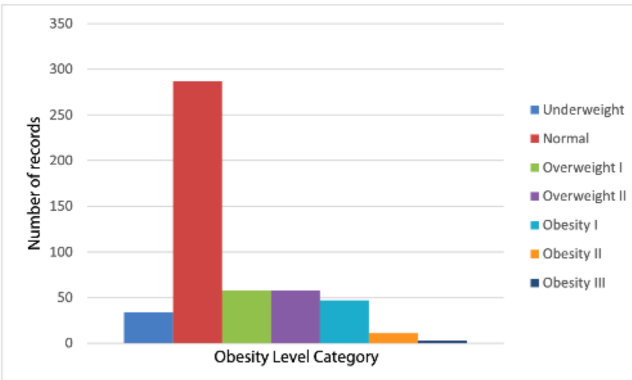


Figure 2: Imbalance in categories of data shown in a bar chart.

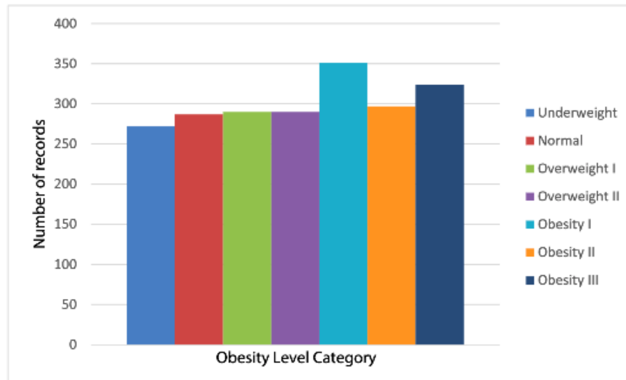


Figure 3: Balanced data categories after synthetic data generation.

The resulting dataset required cleaning, as the generated synthetic data was very noisy and did not follow the specifications as listed in the original survey. For instance, while our multiple-choice answers were recorded using whole numbers, the synthetic data contained values with decimals. As a result, we needed to round the values generated from the synthetic process. Additionally, we added a BMI column, calculated as weight/(height)^2.

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRAN
0	Female	21.0	1.62	64.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	0.0	1.0	no	Public_Transportati
1	Female	21.0	1.52	56.0	yes	no	3.0	3.0	Sometimes	yes	3.0	yes	3.0	0.0	Sometimes	Public_Transportati
2	Male	23.0	1.80	77.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	2.0	1.0	Frequently	Public_Transportati
3	Male	27.0	1.80	87.0	no	no	3.0	3.0	Sometimes	no	2.0	no	2.0	0.0	Frequently	Walki
4	Male	22.0	1.78	89.8	no	no	2.0	1.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Public_Transportati

Figure 4: Snapshot of the final dataset.

Below is the summary of the cleaned dataset. The categories of each categorical variables are listed as followed:

- Gender: Female/ Male
- Family History with Overweight: Yes/ No
- FAVC: Yes/ No
- CAEC: (1) No, (2) Sometimes, (3) Frequently, (4) Always
- SMOKE: Yes/No
- SCC: Yes/No
- CALC: (1) No, (2) Sometimes, (3) Frequently, (4) Always
- MTRANS: (1) Public_Transportation, (2) Walking, (3) Automobile, (4) Motorbike
- NObeyesdad: (1) Insufficient Weight, (2) Normal Weight, (3) Overweight Level I, (4) Overweight Level II, (5) Obesity Type I, (6) Obesity Type II, (7) Obesity Type III

```
    Gender               Age            Height          Weight       family_history_with_overweight     FAVC
 Length:2111        Min.   :14.00   Min.   :1.450   Min.   : 39.00   Length:2111                    Length:2111
 Class :character   1st Qu.:19.95   1st Qu.:1.630   1st Qu.: 65.47   Class :character               Class :character
 Mode  :character   Median :22.78   Median :1.700   Median : 83.00   Mode  :character               Mode  :character
                    Mean   :24.31   Mean   :1.702   Mean   : 86.59
                    3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.43
                    Max.   :61.00   Max.   :1.980   Max.   :173.00
      FCVC            NCP            CAEC              SMOKE                CH2O            SCC
 Min.   :1.000   Min.   :1.000   Length:2111        Length:2111        Min.   :1.000   Length:2111
 1st Qu.:2.000   1st Qu.:2.659   Class :character   Class :character   1st Qu.:1.585   Class :character
 Median :2.386   Median :3.000   Mode  :character   Mode  :character   Median :2.000   Mode  :character
 Mean   :2.419   Mean   :2.686                                         Mean   :2.008
 3rd Qu.:3.000   3rd Qu.:3.000                                         3rd Qu.:2.477
 Max.   :3.000   Max.   :4.000                                         Max.   :3.000
      FAF              TUE              CALC              MTRANS           NObeyesdad             BMI
 Min.   :0.0000   Min.   :0.0000   Length:2111        Length:2111        Length:2111        Min.   :13.00
 1st Qu.:0.1245   1st Qu.:0.0000   Class :character   Class :character   Class :character   1st Qu.:24.33
 Median :1.0000   Median :0.6253   Mode  :character   Mode  :character   Mode  :character   Median :28.72
 Mean   :1.0103   Mean   :0.6579                                                            Mean   :29.70
 3rd Qu.:1.6667   3rd Qu.:1.0000                                                            3rd Qu.:36.02
 Max.   :3.0000   Max.   :2.0000                                                            Max.   :50.81
```

Obesity Risk Factor

- **FAVC:** frequent consumption of high caloric food
- **FCVC:** frequency of consumption of vegetables (1, 2, 3)
- **NCP:** number of main meals (1, 2, 3, 4)
- **CAEC:** consumption of food between meals
- **CH2O:** consumption of water daily (1, 2, 3)
- **CALC:** consumption of alcohol

Physical Health and Condition

- **SCC:** calories consumption monitoring
- **FAF:** physical activity frequency (0, 1, 2, 3)
- **TUE:** time using technology devices (0, 1, 2)
- **MTRANS:** transportation used

Individual Variables

- Gender
- Age
- Height
- Weight
- Family History
- Smoking Habit

Obesity Variables

- Insufficient Weight
- Normal Weight
- Overweight Level I
- Overweight Level II
- Obesity Type I
- Obesity Type II
- Obesity Type III

Figure 5: Variables documented in the dataset.

3. Exploratory Analysis

3.1. Research questions

1. Which factors are most correlated to BMI?
2. Can the explanatory variables that are most correlated to BMI be used to classify a person's obesity level?
3. How well is our classifier in terms of predictive capability?

3.2. Exploratory Results

The heatmap shows that there is a correlation between BMI and FAF (Frequency of Physical Activity). For the categorical variables (Figure 7,8), it appears that Frequency of Vegetable consumption and Family History have a greater possible correlation with BMI than the other variables. A separate visualization was created for comparing Family History and BMI as well, as seen in Figure 9. Subsequently, linear regressions were conducted on each of the three variables of interest with their relation to BMI, as well as a regression with all three variables and BMI (Figure 10&11).

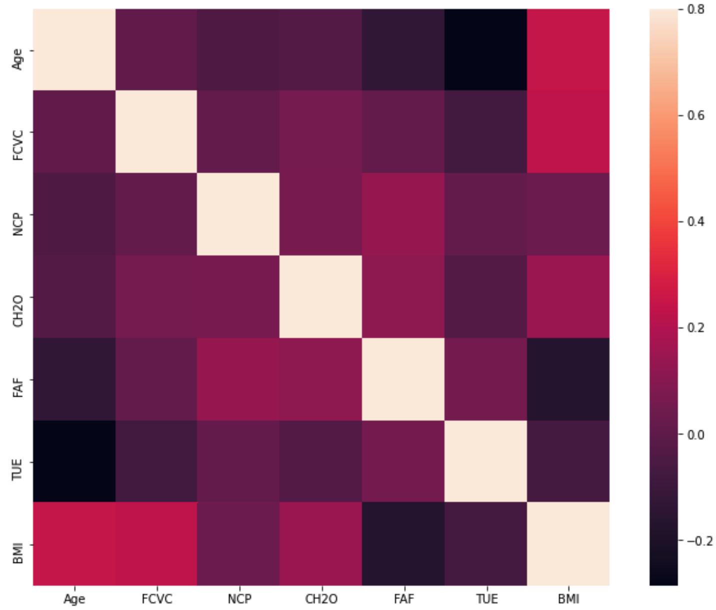
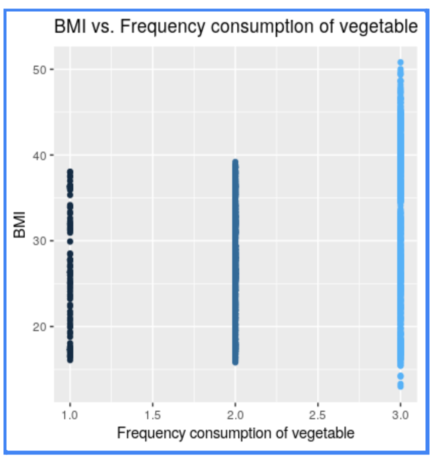
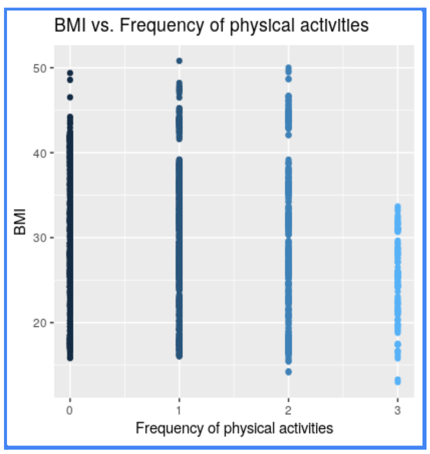
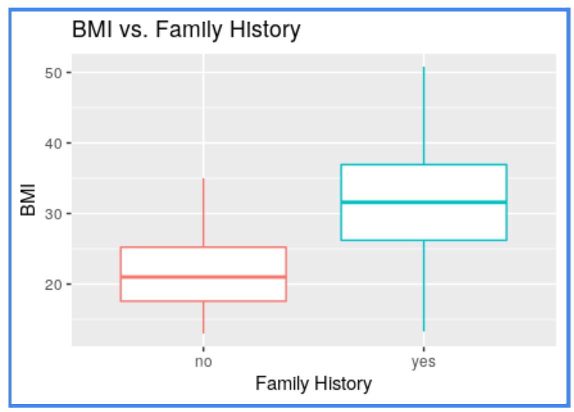


Figure 6: Heat map shows the correlation between numeric variables and BMI

Figure 7&8: Frequency of Vegetable Consumption vs BMI and Frequency of Physical Activities vs BMI
Figure 9: Boxplot of Family History vs BMI







BMI vs Family History

Dep. Variable:	BMI	R-squared:	0.234
Model:	OLS	Adj. R-squared:	0.233
Method:	Least Squares	F-statistic:	643.5
Date:	Tue, 30 Nov 2021	Prob (F-statistic):	4.03e-124
Time:	20:47:46	Log-Likelihood:	-7106.5
No. Observations:	2111	AIC:	1.422e+04
Df Residuals:	2109	BIC:	1.423e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	21.5005	0.357	60.144	0.000	20.799	22.202
C(family_history_with_overweight)[T.yes]	10.0287	0.395	25.367	0.000	9.253	10.804

Omnibus:	27.897	Durbin-Watson:	0.477
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17.020
Skew:	0.006	Prob(JB):	0.000201
Kurtosis:	2.560	Cond. No.	4.48

BMI vs FCVC

Dep. Variable:	BMI	R-squared:	0.061
Model:	OLS	Adj. R-squared:	0.060
Method:	Least Squares	F-statistic:	136.0
Date:	Tue, 30 Nov 2021	Prob (F-statistic):	1.72e-30
Time:	20:47:46	Log-Likelihood:	-7321.6
No. Observations:	2111	AIC:	1.465e+04
Df Residuals:	2109	BIC:	1.466e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	24.2604	0.496	48.892	0.000	23.287	25.233
C(FAVC)[T.yes]	6.1540	0.528	11.660	0.000	5.119	7.189

Omnibus:	124.261	Durbin-Watson:	0.263
Prob(Omnibus):	0.000	Jarque-Bera (JB):	47.169
Skew:	0.050	Prob(JB):	5.72e-11
Kurtosis:	2.275	Cond. No.	5.71

BMI vs FAF

Dep. Variable:	BMI	R-squared:	0.032
Model:	OLS	Adj. R-squared:	0.031
Method:	Least Squares	F-statistic:	23.45
Date:	Tue, 30 Nov 2021	Prob (F-statistic):	6.32e-15
Time:	20:47:47	Log-Likelihood:	-7352.9
No. Observations:	2111	AIC:	1.471e+04
Df Residuals:	2107	BIC:	1.474e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.0088	0.294	105.504	0.000	30.432	31.585
C(FAF)[T.1.0]	-1.0723	0.408	-2.628	0.009	-1.873	-0.272
C(FAF)[T.2.0]	-2.5024	0.460	-5.438	0.000	-3.405	-1.600
C(FAF)[T.3.0]	-5.7915	0.780	-7.421	0.000	-7.322	-4.261

Omnibus:	115.984	Durbin-Watson:	0.224
Prob(Omnibus):	0.000	Jarque-Bera (JB):	51.159
Skew:	0.155	Prob(JB):	7.78e-12
Kurtosis:	2.303	Cond. No.	5.28

Figure 10: Regression results from BMI vs Family History, BMI vs FCVC, and BMI vs FAF.

Dep. Variable:	BMI	R-squared:	0.275
Model:	OLS	Adj. R-squared:	0.273
Method:	Least Squares	F-statistic:	159.4
Date:	Wed, 01 Dec 2021	Prob (F-statistic):	5.85e-144
Time:	11:17:12	Log-Likelihood:	-7048.6
No. Observations:	2111	AIC:	1.411e+04
Df Residuals:	2105	BIC:	1.414e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	20.1732	0.551	36.628	0.000	19.093	21.253
C(family_history_with_overweight)[T.1]	9.2656	0.394	23.515	0.000	8.493	10.038
C(FAVC)[T.1]	3.4552	0.477	7.243	0.000	2.520	4.391
C(FAF)[T.1.0]	-1.0223	0.354	-2.891	0.004	-1.716	-0.329
C(FAF)[T.2.0]	-2.0528	0.399	-5.141	0.000	-2.836	-1.270
C(FAF)[T.3.0]	-4.3431	0.680	-6.384	0.000	-5.677	-3.009

Figure 11: Regression results from BMI vs Family History, FCVC, and FAF

4. Modeling
4.1. Modeling

We chose to perform classification tasks to build the model as all of our attributes are categorical and the labels are provided in the data. Overall, we tested five different classifiers: KNN, SVM, Random Forest, Decision Tree, and Naive Bayes. For kNN classification, we need to specify the number of the nearest neighbors k. To find the best-fitted k, we calculate the error rate at each different k given the trained datasets. The results showed that k = 3 has the lowest error rate. For each classification, we split the data into two parts: training data and test data. Training data were used to build the model, and test data were used to evaluate the model.

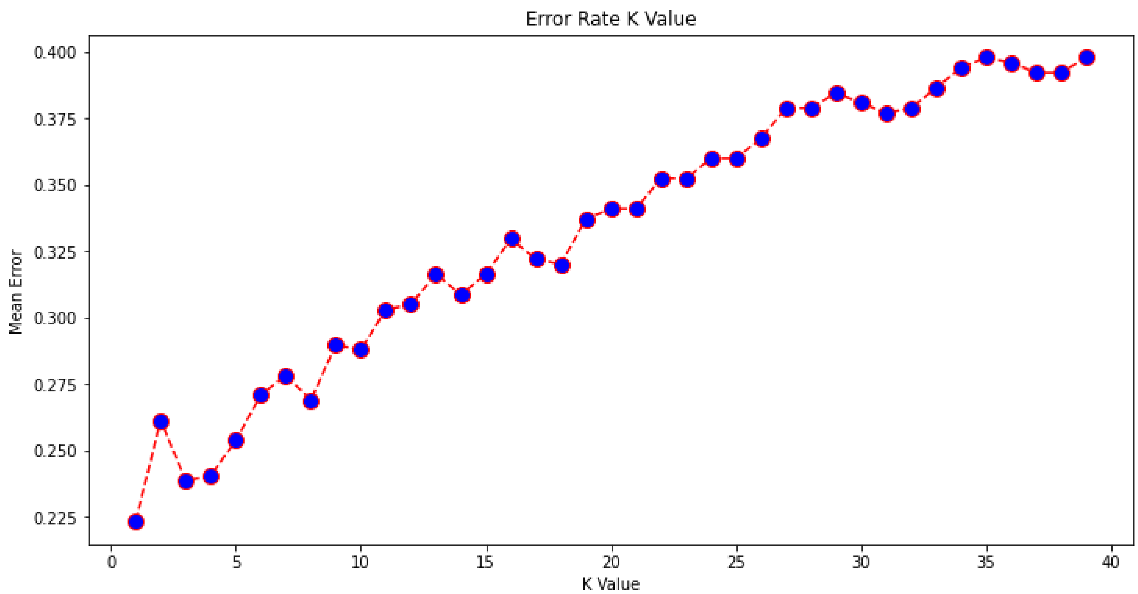


Figure 12: Error rates of various K values in KNN model

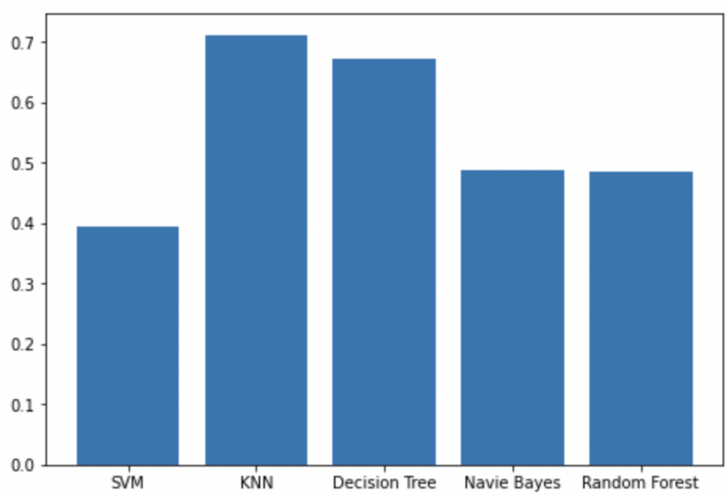
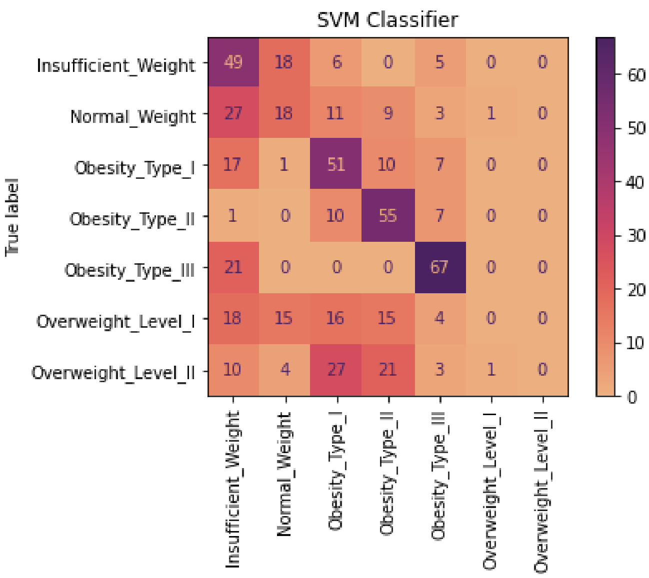


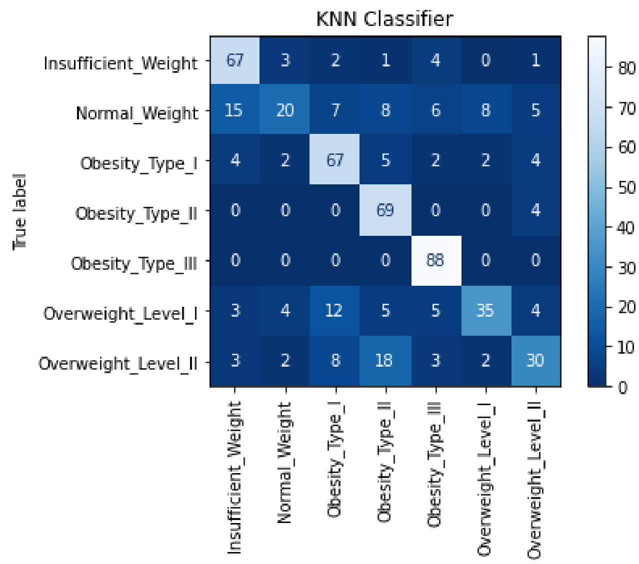
Figure 13: Accuracy Score of Classifiers

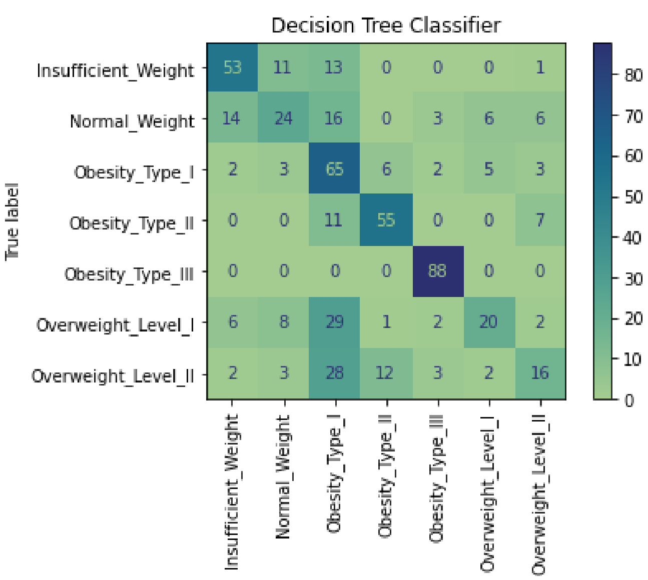
Classifier	Accuracy Score
SVM	0.454
KNN	0.759
Decision Tree	0.752
Naive Bayes	0.504
Random Forest	0.564

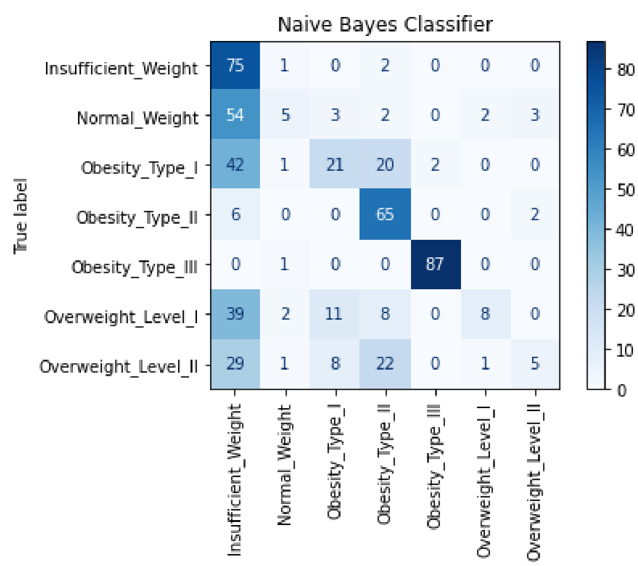
4.2. Model Evaluation

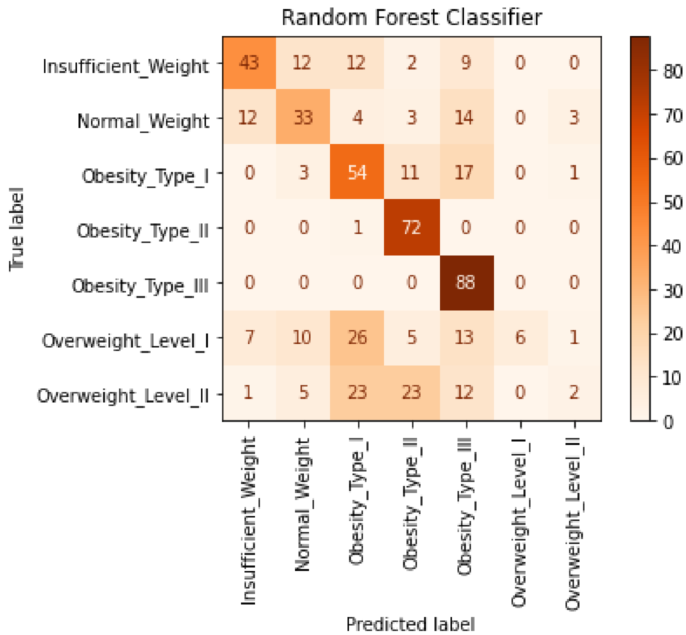
Evaluating models with confusion matrices indicates that KNN classifier has an overall better predictive performance compared to other classification models. The total number of true positive and true negative correctly identified by KNN overall is higher than other classifiers, while its number of falsely predicted data points is also less than that resulted from other models.











5. Discussion

For the obesity level, we started our exploratory analysis by investigating the distribution of obesity level by gender: maximum female respondents fell under Obesity Type III and maximum male respondents fell under Obesity Type II. Next, we created a heatmap, in which we identified that there are strong correlations between BMI and Frequency of Physical Activity. We identified relatively strong relations between (1) BMI and Family History, (2) BMI and Frequency of Consumption of Vegetables, (3) BMI and Frequency of Physical Activities, (4) BMI and Alcohol Consumption, (5) BMI and Frequency of Consumption of high-calorie food. In the regression results from BMI vs Family History, BMI vs FCVC, and BMI vs FAF, Family History has the highest coefficient (9.266) compared to other variables, meaning that the change in one unit of Family History will have greater influence on BMI compared to the other variables. At the same time, the adjusted R-squared of the final regression model also points out that only 27.5% of the variation in BMI can be explained by our explanatory variables. We do not, however, have enough evidence to state that any of these factors directly affect the BMI.

The results of our regression further indicate that Family History, FCVC, and FAF alone are not sufficient for the predictive task. Therefore, in choosing the explanatory variables to train our data for classification, we took into account all the possible factors. We calculated the accuracy scores for each classification model and found the KNN model has the highest accuracy score compared to the other 4 models. As a result, we choose KNN as the most fitted classifier for our study. The accuracy score of 76% implies that given a person's set of lifestyle habits (ones that accounted as our explanatory variables), the KNN model will be 76% correct in predicting the person's level of obesity. At the same time, the results of our model evaluation using confusion matrices implies that KNN model has an overall better predictive performance compared to other classifiers. Compared to other classifiers, KNN yields a higher general number of correctly classifying data points, while also holding the least number of falsely predicted testing data.

The first limitation of our study lies within our selection of the KNN classifier. To be more specific, the KNN algorithm is computationally expensive, as we need to calculate the proximity measures during the process. The KNN model is also not good at identifying edge cases, which can adversely be affected by outliers. Additionally, feature selection is critical in KNN, as irrelevant features can dominate a decision, so the model will not perform as well if there is a class imbalance. The second limitation is with our process of data collection. To begin with, 77% of the data is synthetic and consequently might not be representative of the original populations of interest. At the same time, the original survey was only available online and for 30 days, which could cause response and volunteer bias in the data. As a consequence, we lack randomization in the process of collecting data and our findings also lack the external validity as it might not be representative for the population.

6. Conclusion

Given the results from our models, KNN appears to be the best fit for our data since it has the highest accuracy. Using our KNN model, we can predict a person's obesity level based on their lifestyle attributes, age, gender, and family history with obesity. To further advance this study, we can try to improve the model's accuracy and sensitivity by using the attributes that are most relevant to the obesity levels instead of using all attributes in its development. Furthermore, to increase the external validity of our study, we can collect more human data and expand our sample size instead of using generated synthetic data.

7. Acknowledgment

- Vivian Bui: Actively attended meetings, provided useful insights for the project, implemented R and Python for data visualization and classification modeling, presented project to the class, contributed to the report
- Xinying Qian: Actively attended meetings, provided useful insights for the project, implemented R and Python for data visualization and classification modeling, contributed to the report
- Aishwarya Kotha: Actively attended meetings, provided useful insights for the project, worked hard with the presentation slides, presented project to the class, contributed to the report
- Nikita Mathur: Actively attended meetings, provided useful insights for the project, worked hard with the presentation slides, contributed to the report
- Yijie Lian: Contributed to the report

8. Reference

Elhassan et al. (2017), Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. Global Journal of G Technology & Optimization.

Fabio Mendoza Palechor, Alexis de la Hoz Manotas, Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico, Data in Brief, Volume 25, 2019, 104344, ISSN 2352-3409, https://doi.org/10.1016/j.dib.2019.104344.