Estimation of obesity levels based on eating habits and physical condition

IS 507 Final Presentation - Emily Chang (emilyjc4), Jung-Ching Chen (jcc11)

Table of Contents

01

Introduction

Brief introduction to the dataset for this project

Obesity dataset from UCI Machine learning Repository

This dataset includes data for the obesity levels in individuals from the countries of Mexico, Peru and Columbia, and also data for eating habits and physical condition.

About the Dataset

2111 datas 17 variables

```
$ Gender                        : chr  "Female" "Female" "Male" "Male" ...
$ Age                           : num  21 21 23 27 22 29 23 22 24 22 ...
$ Height                        : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
$ Weight                        : num  64 56 77 87 89.8 53 55 53 64 68 ...
$ family_history_with_overweight: chr  "yes" "yes" "yes" "no" ...
$ FAVC                          : chr  "no" "no" "no" "no" ...
$ FCVC                          : num  2 3 2 3 2 2 3 2 3 2 ...
$ NCP                           : num  3 3 3 3 1 3 3 3 3 3 ...
$ CAEC                          : chr  "Sometimes" "Sometimes" "Sometimes" "Sometimes" ...
$ SMOKE                         : chr  "no" "yes" "no" "no" ...
$ CH2O                          : num  2 3 2 2 2 2 2 2 2 2 ...
$ SCC                           : chr  "no" "yes" "no" "no" ...
$ FAF                           : num  0 3 2 2 0 0 1 3 1 1 ...
$ TUE                           : num  1 0 1 0 0 0 0 0 1 1 ...
$ CALC                          : chr  "no" "Sometimes" "Frequently" "Frequently" ...
$ MTRANS                        : chr  "Public_Transportation" "Public_Transportation" "Public_Transportation" "Walking" ...
$ NObeyesdad                    : chr  "Normal_Weight" "Normal_Weight" "Normal_Weight" "Overweight_Level_I" ...
```

About the Dataset

Eating habits

FAVC
Frequent consumption of high caloric food

FCVC
Frequency of consumption of vegetables

NCP
Number of main meals

CAEC
Consumption of food between meals

CH2O
Consumption of water daily

CALC
Consumption of alcohol

Physical condition

SCC
Calories consumption monitoring

FAF
Physical activity frequency

TUE
Time using technology devices

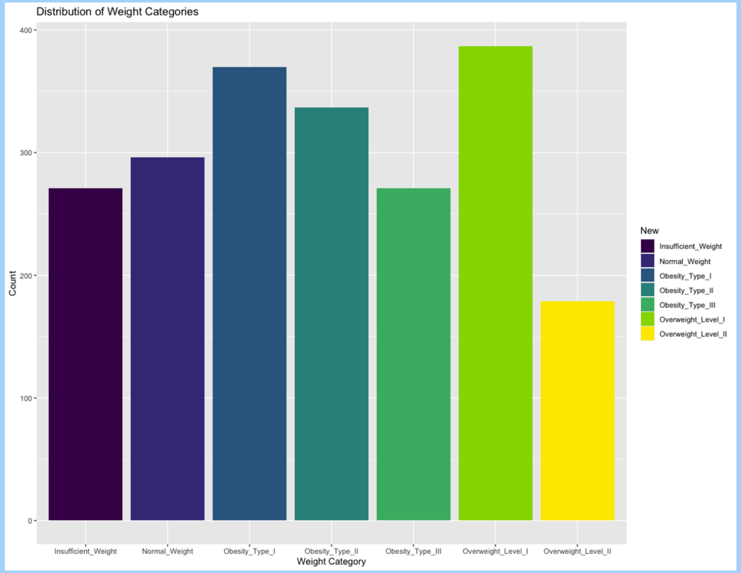
MTRANS
Transportation

Other Variables: Gender, Height, Weight, NObeyesdad

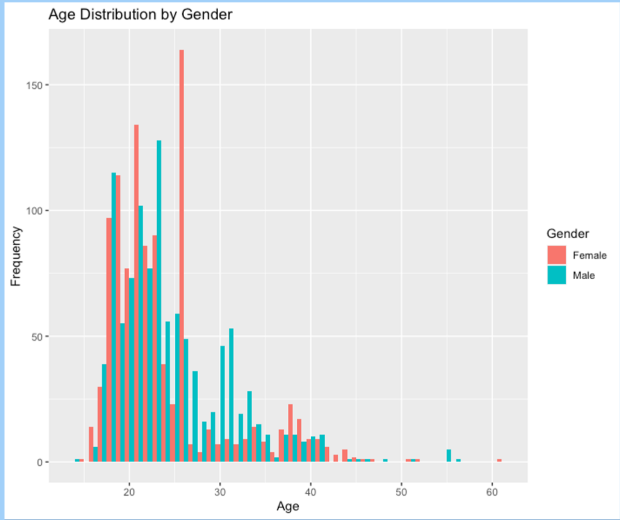
About the Dataset

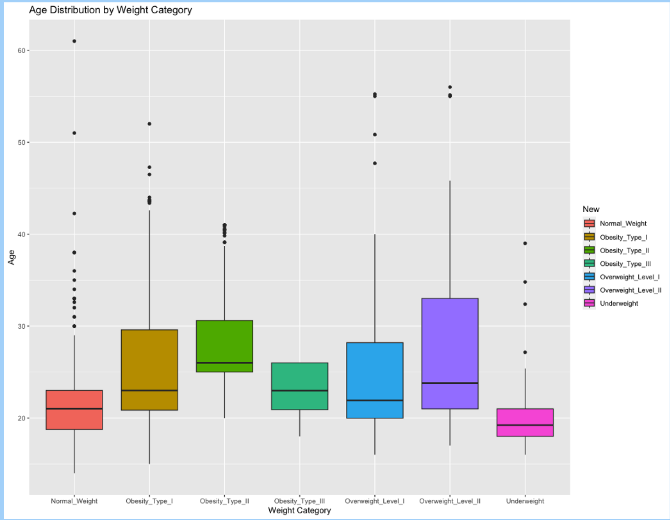
Obesity Level Category (Recategorized)

- Underweight: Less than 18.5
- Normal: 18.5 to 24.9
- Overweight I: 25.0 to 27.9
- Overweight II: 28 to 29
- Obesity I: 30.0 to 34.9
- Obesity II: 35.0 to 39.9
- Obesity III: Higher than 40



About the Dataset





Age Distribution by Gender

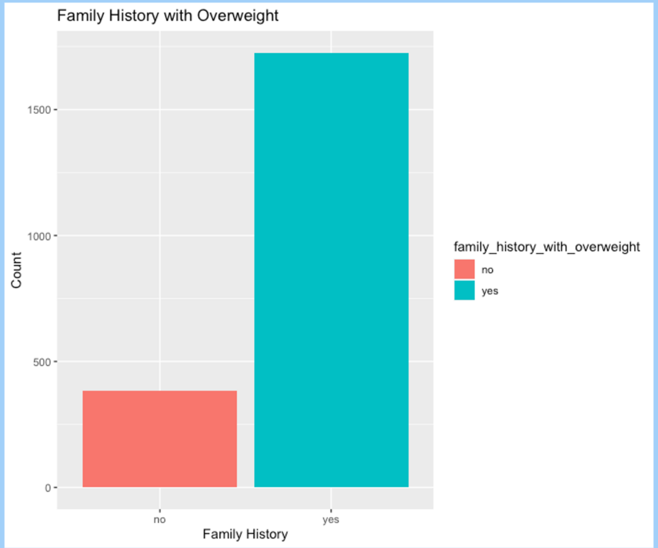
The histogram provides how ages are distributed. The height of the bars indicates the count of individuals within each age group.

Age Distribution by Weight Category

We've got seven categories on obesity which are calculated by BMI.

About the Dataset

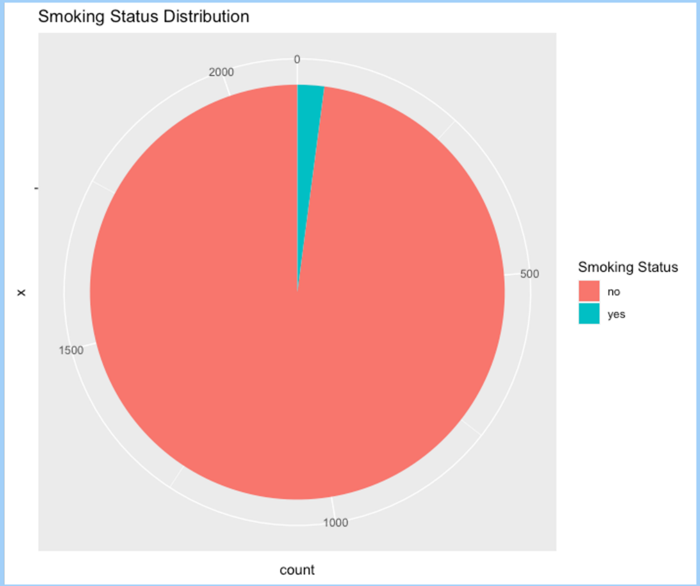
Family History with Overweight



- Yes: 1726 records
- No: 385 records

About the Dataset

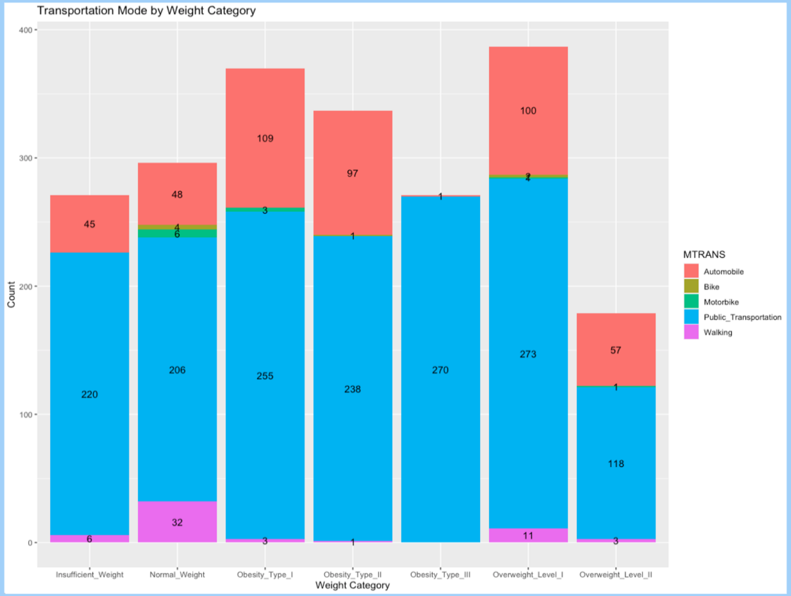
Smoking Status Distribution



- Yes: 44 records
- No: 2067 records

About the Dataset

Transportation Mode by Weight Category



02

ANOVA

Uncover significant variations in variables among diverse obesity categories

ANOVA

FCVC

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
New	2	68.2	34.10	146.2	<2e-16	***
Residuals	835	194.8	0.23			

NCP

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
New	2	10.8	5.386	10.11	4.58e-05	***
Residuals	835	444.7	0.533			

CH2O

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
New	2	35.67	17.837	49.77	<2e-16	***
Residuals	835	299.25	0.358			

FAF

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
New	2	40.5	20.267	25.71	1.46e-11	***
Residuals	835	658.2	0.788			

TUE

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
New	2	9.33	4.664	14.1	9.49e-07	***
Residuals	835	276.20	0.331			

CAEC

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
New	2	157.5	78.76	99.2	<2e-16	***
Residuals	835	663.0	0.79			

CALC

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
New	2	40.45	20.225	91.05	<2e-16	***
Residuals	835	185.48	0.222			

Post Hoc Tests

```
Fit: aov(formula = FCVC ~ New, data = data_selected)

$New
                                        diff        lwr
Normal_Weight-Insufficient_Weight    -0.1358790 -0.2312201
Obesity_Type_III-Insufficient_Weight  0.5271320  0.4297117
Obesity_Type_III-Normal_Weight        0.6630109  0.5676698
                                            upr     p adj
Normal_Weight-Insufficient_Weight    -0.04053785 0.0024636
Obesity_Type_III-Insufficient_Weight  0.62455229 0.0000000
Obesity_Type_III-Normal_Weight        0.75835205 0.0000000
```

Post Hoc Tests

FCVC

	p adj
Normal_Weight-Insufficient_Weight	0.0024636
Obesity_Type_III-Insufficient_Weight	0.0000000
Obesity_Type_III-Normal_Weight	0.0000000

FAF

	p adj
Normal_Weight-Insufficient_Weight	0.9169894
Obesity_Type_III-Insufficient_Weight	0.0000000
Obesity_Type_III-Normal_Weight	0.0000000

NCP

	p adj
Normal_Weight-Insufficient_Weight	0.0107079
Obesity_Type_III-Insufficient_Weight	0.3039238
Obesity_Type_III-Normal_Weight	0.0000355

TUE

	p adj
Normal_Weight-Insufficient_Weight	0.0050222
Obesity_Type_III-Insufficient_Weight	0.0000005
Obesity_Type_III-Normal_Weight	0.0614452

CH2O

	p adj
Normal_Weight-Insufficient_Weight	0.8501884
Obesity_Type_III-Insufficient_Weight	0.0000000
Obesity_Type_III-Normal_Weight	0.0000000

CAEC

	p adj
Normal_Weight-Insufficient_Weight	0.9909276
Obesity_Type_III-Insufficient_Weight	0.0000000
Obesity_Type_III-Normal_Weight	0.0000000

CALC

	p adj
Normal_Weight-Insufficient_Weight	0.0003389
Obesity_Type_III-Insufficient_Weight	0.0000000
Obesity_Type_III-Normal_Weight	0.0000009

Post Hoc Tests

Frequency of eating vegetables	Obesity_Type_III > Insufficient_Weight > Normal
Number of main meals	Obesity_Type_III = Insufficient_Weight > Normal
Consumption of water daily	Obesity_Type_III > Insufficient_Weight = Normal
Physical activity frequency	Insufficient_Weight = Normal > Obesity_Type_III
Consumption of food between meals	Insufficient_Weight > Normal = Obesity_Type_III
Time using technology devices	Insufficient_Weight = Normal > Obesity_Type_III
Consumption of alcohol	Obesity_Type_III > Normal > Insufficient_Weight

03

Model

Comparing different models' predictions and accuracy

Steps

1. Data Splitting & Feature Selection

80% training 20% testing
Selected the top 10 features

2. Decision Tree Model

rpart library

3. Random Forest Model

randomForest library

4. SVM Model

e1071 library

5. Model Evaluation

Using confusion matrices and calculate accuracy

6. ROC Curves

pROC library
Visualize the performance of our models

Models

Feature Selection

```
 Variables Accuracy  Kappa AccuracySD KappaSD Selected
        10   0.9137 0.8984    0.02144 0.02522        *
        16   0.9090 0.8929    0.02489 0.02927

The top 5 variables (out of 10):
   Weight, Height, Age, FCVC, Gender
```

```
 [1] "Weight"                          "Height"                          "Age"
 [4] "FCVC"                            "Gender"                          "CH2O"
 [7] "TUE"                             "NCP"                             "FAF"
[10] "family_history_with_overweight"
```

Selected the top ten features by running RFE with Random Forest

Models

Decision Tree

```
[1] "Decision Tree Confusion Matrix:"
> print(dt_conf_matrix)
```

dt_predictions	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III
Insufficient_Weight	50	11	0	0	0
Normal_Weight	4	38	0	0	0
Obesity_Type_I	0	0	58	2	0
Obesity_Type_II	0	0	11	64	2
Obesity_Type_III	0	0	0	1	52
Overweight_Level_I	0	10	2	0	0
Overweight_Level_II	0	0	3	0	0

dt_predictions	Overweight_Level_I	Overweight_Level_II
Insufficient_Weight	0	0
Normal_Weight	5	0
Obesity_Type_I	0	2
Obesity_Type_II	0	0
Obesity_Type_III	0	0
Overweight_Level_I	64	16
Overweight_Level_II	8	17

Decision Tree Accuracy: 0.8166666666666667

Models

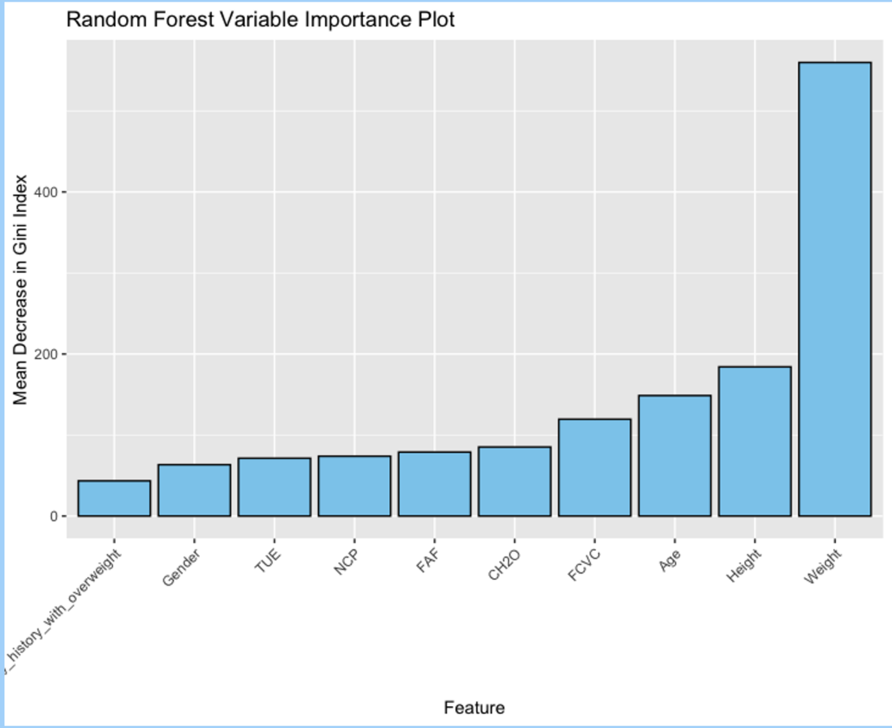
Random Forest

```
[1] "Random Forest Confusion Matrix:"
> print(rf_conf_matrix)

rf_predictions         Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III
  Insufficient_Weight                   50             2              0               0                0
  Normal_Weight                          4            51              0               0                0
  Obesity_Type_I                         0             0             69               3                0
  Obesity_Type_II                        0             0              1              64                1
  Obesity_Type_III                       0             0              0               0               53
  Overweight_Level_I                     0             6              2               0                0
  Overweight_Level_II                    0             0              2               0                0

rf_predictions         Overweight_Level_I Overweight_Level_II
  Insufficient_Weight                  0                   0
  Normal_Weight                        6                   0
  Obesity_Type_I                       0                   0
  Obesity_Type_II                      0                   0
  Obesity_Type_III                     0                   0
  Overweight_Level_I                  69                  10
  Overweight_Level_II                  2                  25
```

Random Forest Accuracy: 0.907142857142857

Models

Random Forest



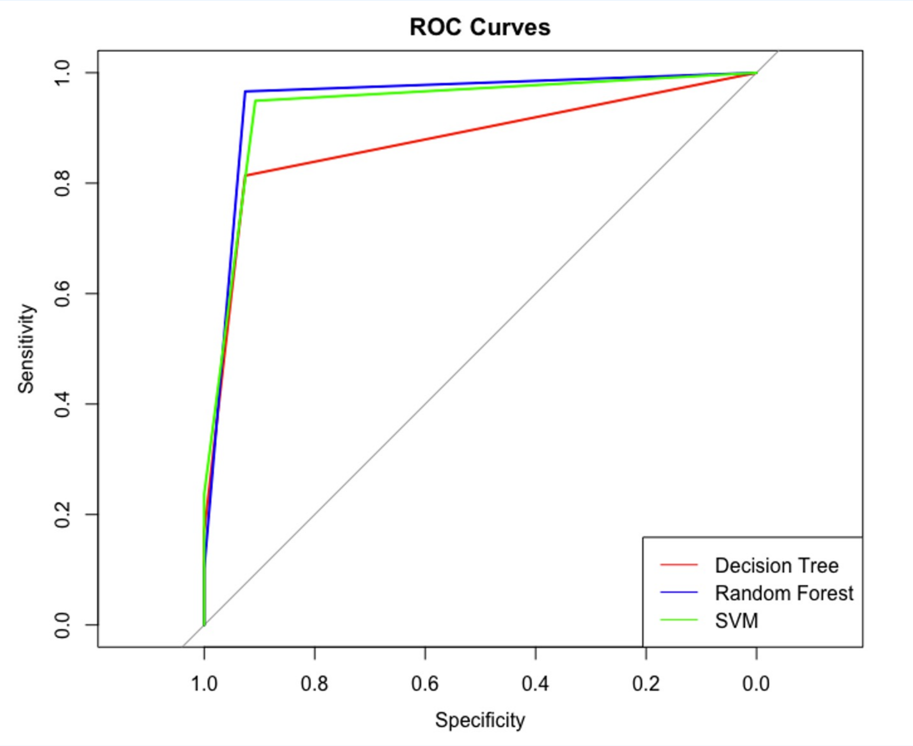
Models

Support Vector Machine

```
[1] "SVM Confusion Matrix:"
> print(svm_conf_matrix)

svm_predictions        Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III
  Insufficient_Weight                   49             3              0               0                0
  Normal_Weight                          5            42              0               0                0
  Obesity_Type_I                         0             0             66               0                0
  Obesity_Type_II                        0             0              4              66                2
  Obesity_Type_III                       0             0              0               1               52
  Overweight_Level_I                     0            14              1               0                0
  Overweight_Level_II                    0             0              3               0                0

svm_predictions        Overweight_Level_I Overweight_Level_II
  Insufficient_Weight                   0                   0
  Normal_Weight                         4                   0
  Obesity_Type_I                        0                   4
  Obesity_Type_II                       0                   0
  Obesity_Type_III                      0                   0
  Overweight_Level_I                   70                  17
  Overweight_Level_II                   3                  14
```

SVM Accuracy: 0.854761904761905

ROC Curves



Random Forest Model has the highest accuracy.

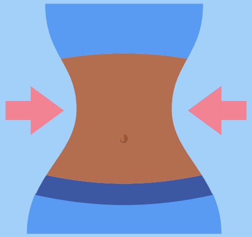
	Decision Tree	Random Forest	SVM
Accuracy	0.8166667	0.9071429	0.8547619
AUC	0.8760201	0.9497803	0.9392655

04

Conclusion

Conclusions for the project

Conclusions



Conclusion A

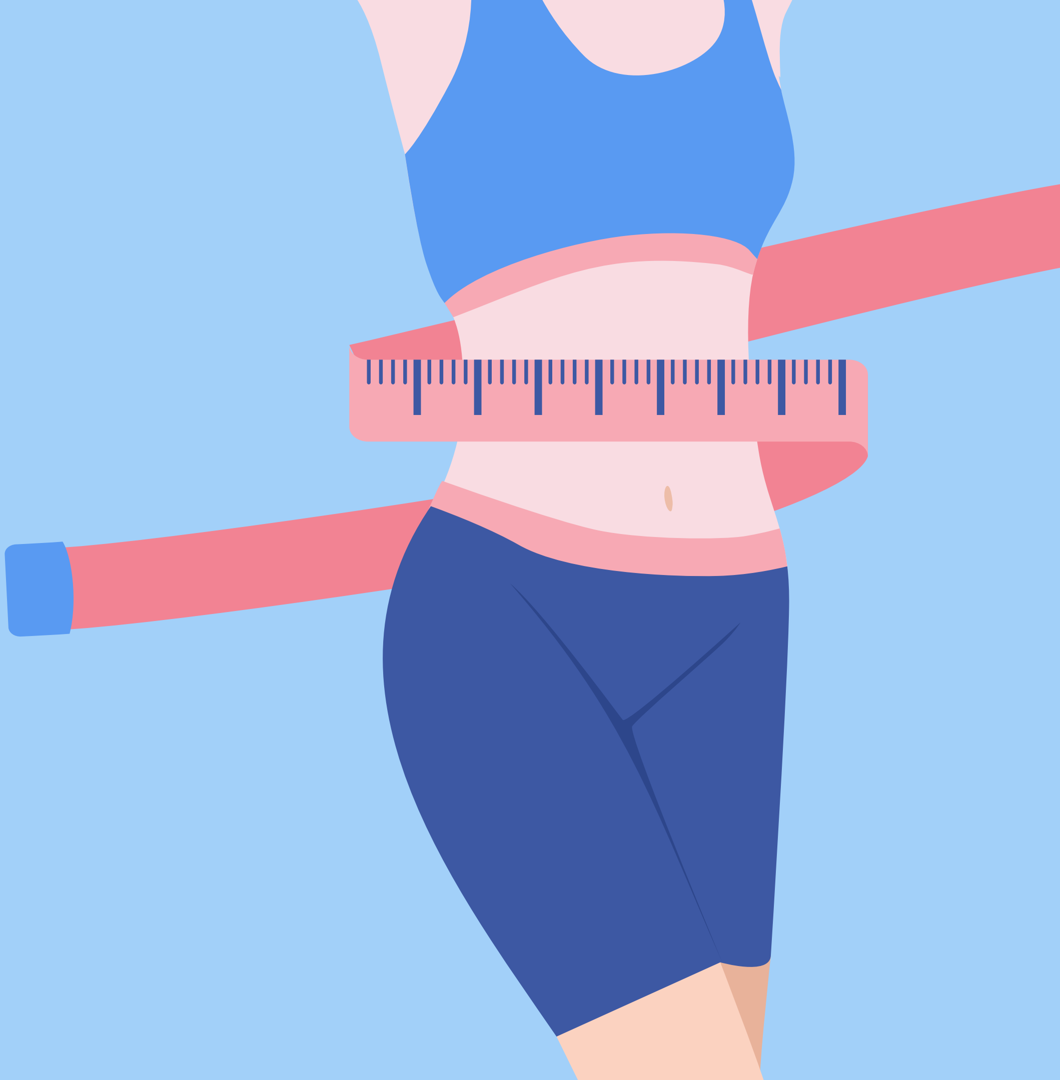
For Physical Activity Frequency and Number of Main Meals, no significant difference is observed between Obesity and Insufficient Weight. However, a significant difference exists between Obesity and Normal Weight.

Conclusion B

Random Forest Model appears to be the most accurate among the three, suggesting that its ensemble nature helps improve predictive performance.

Thanks!

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.



05

Response

Response to questions received

Response to Questions

Figure 1

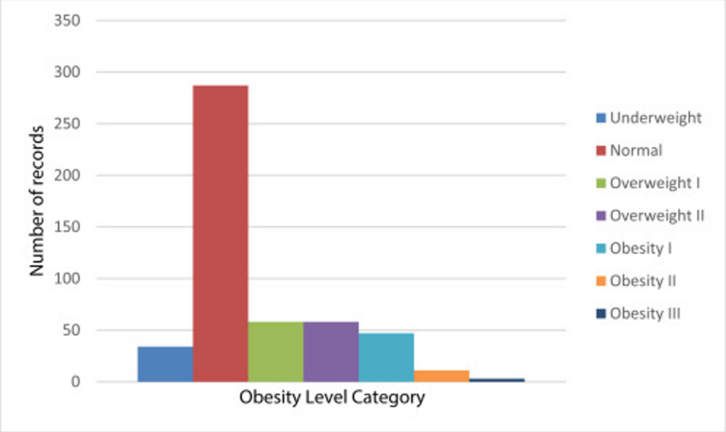
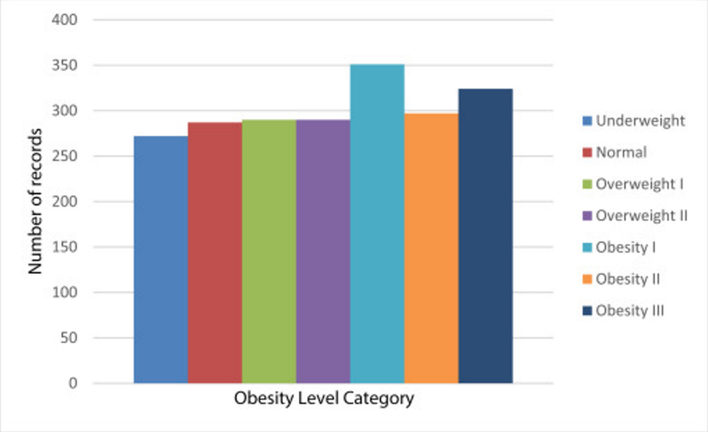


Figure 2



According to the paper in the data set, the original data was also unbalanced. Since we found that the data classification was incorrect when pre-processing the data, we reclassified the data but the data was slightly unbalanced. According to the original literature, after the labeling process was finished, the categories of obesity levels were unbalanced (Fig. 1), and this presented a learning problem for the data mining methods since it would learn to identify correctly the category with most records compared with the categories with less data. After the balancing class problem was identified, synthetic data was generated, up to 77% of the data, using the tool Weka and the filter SMOTE. Although the data we reclassified did not have serious data imbalance problems , but according to the literature, if the above method is used, there is a chance that the model accuracy can be improved.