Estimation of Obesity Levels

Jasmin Karki
Masters in Cyber Physical Systems
Indian Institute of Technology Madras
Chennai, Tamil Nadu, India
ge22m019@smail.iitm.ac.in

***Abstract*—In this report, we estimate obesity levels in individuals from the countries of Mexico, Peru, and Colombia, based on their eating habits and physical condition. Descriptive analysis, correlation analysis, statistical tests, and regression analysis have been conducted using the dataset. Logistic regression has been used to perform a regression analysis to classify the obesity level of an individual. It is a multi-class classification problem where the dataset we use derives from the Obesity Level dataset. We created a logistic regression model that classifies an individual into one of seven obesity target values with a macro average f1 score of 0.71.**

***Index Terms*—descriptive statistics, correlation analysis, regression analysis**

I. Introduction

Obesity is a chronic metabolic condition influenced by social, environmental, and hereditary variables. It raises the chance of developing a number of illnesses, including hypertension, diabetes, cardiovascular disorders, and respiratory illnesses. In both industrialized and developing nations, the prevalence of obesity has risen alarmingly due to contemporary lifestyles, unhealthy diets, and declines in physical activity. This paper will focus on finding insights on which eating/physical behavior of individual is more likely to result which type of obesity and proposing a logistic regression model to perform supervised classification on it.

Our dataset is obtained from the people from Mexico, Peru and Colombia. The dataset contains individuals from age 14 to 61 with general features like age, weight, height along with attributes relating to eating habits and physical behaviour. Getting a good understanding of the Obesity dataset can help us in tackling other multi class classification problems.

We implemented descriptive analysis, correlation analysis, regression analysis and found important variables using statistical tests. Then we created a logistic regression model to capture relationships in the dataset to classify an individual into one of the seven obesity categories.

The rest of the paper is organized as follows. Section II outlines the dataset related to our problem. Section III explains the tasks performed in the project. Section IV concludes the paper with key contributions made and directions for future works.

II. Datasets

The dataset includes attributes related to eating habit, physical condition and general details of an individual. Frequent consumption of high-calorie foods (FAVC), frequent consumption of vegetables (FCVC), frequency of main meals (NCP), frequent consumption of food between meals (CAEC), daily consumption of water (CH20), and alcohol consumption are the characteristics of eating habits (CALC). Calorie consumption monitoring (SCC), physical activity frequency (FAF), time utilizing technological devices (TUE), and transportation usage (MTRANS) are the characteristics of the physical condition. Other factors retrieved were gender, age, smoking status, and a family history of obesity. While the other covariates are all categorical values, the covariate age is a continuous variable.

Our task involved performing descriptive statistics, correlation analysis, linear regression analysis and statistical tests to find important variables. A logistic regression model was created to classify an individual into one of seven target groups. The tasks have been done using data visualization libraries matplotlib and seaborn. Similarly data processing libraries scikit-learn and scientific Python was used.

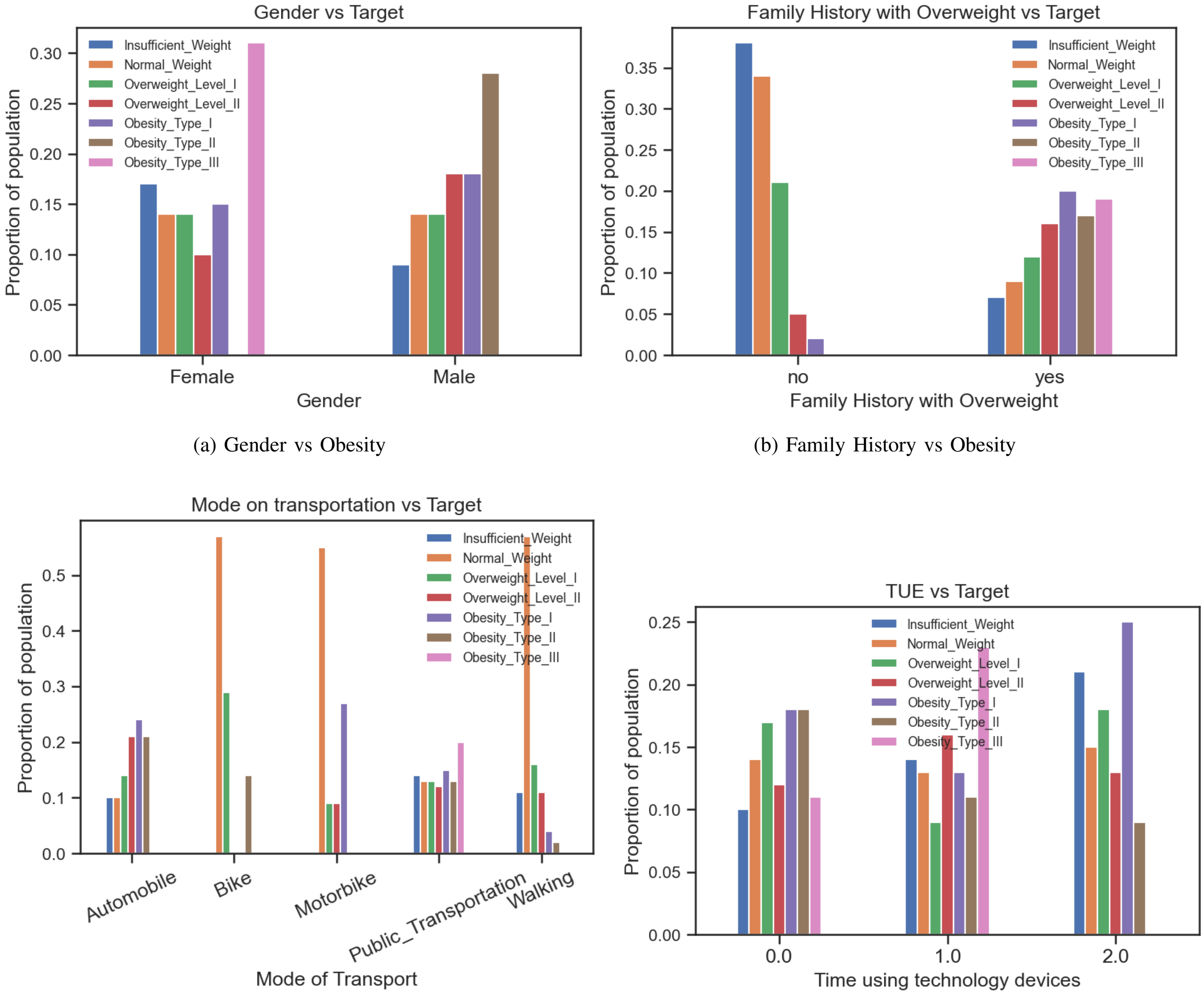
The dataset contains 24 duplicated rows which were removed. Age was grouped into teens, 20s, 30s, 40s, 50s+ to visualize the data properly.

From the bivariate analysis of the categorical features show that no women have type II obesity and no men have type III obesity. Individuals with no family history of overweight don't have much obesity while those with family history have higher chance of obesity. In case of means of transportation, those using public transport have Type III obesity. Likewise, maximum normal weight people either walk, or use bike/motorbike. Obesity is least among those who walk and bike. Time spend on technological device doesn't seem to depend upon obesity level as shown in figure 1a 1b 1c 1d.

Similarly, individuals who frequently consume high calorie food are more likely to be obese. All type III obesity individuals frequently consume vegetable. Interestingly, individuals who take four main meals per day have insufficient weight. Whereas, majority of those who always consume food between meals have normal weight and majority of those who never are level I overweight. Those who perform more physical activity are normal weight. Those who monitor their calories have least obesity. Another interesting visualization is, those who always consume alcohol have normal weight as shown in figure 2a 2b 2c2d 2e 2f.

TABLE I: Descriptive Statistics

index	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	2111.0	2111.0	2111.0	2111.0	2111.0	2111.0	2111.0	2111.0
mean	24.312	1.701	86.586	2.419	2.685	2.00	1.010	0.657
std	6.345	0.093	26.191	0.533	0.778	0.612	0.850	0.608
min	14.0	1.45	39.0	1.0	1.0	1.0	0.0	0.0
25%	19.947	1.63	65.473	2.0	2.658	1.584	0.124	0.0
50%	22.777	1.700	83.0	2.385	3.0	2.0	1.0	0.625
75%	26.0	1.768	107.430	3.0	3.0	2.477	1.666	1.0
max	61.0	1.98	173.0	3.0	4.0	3.0	3.0	2.0



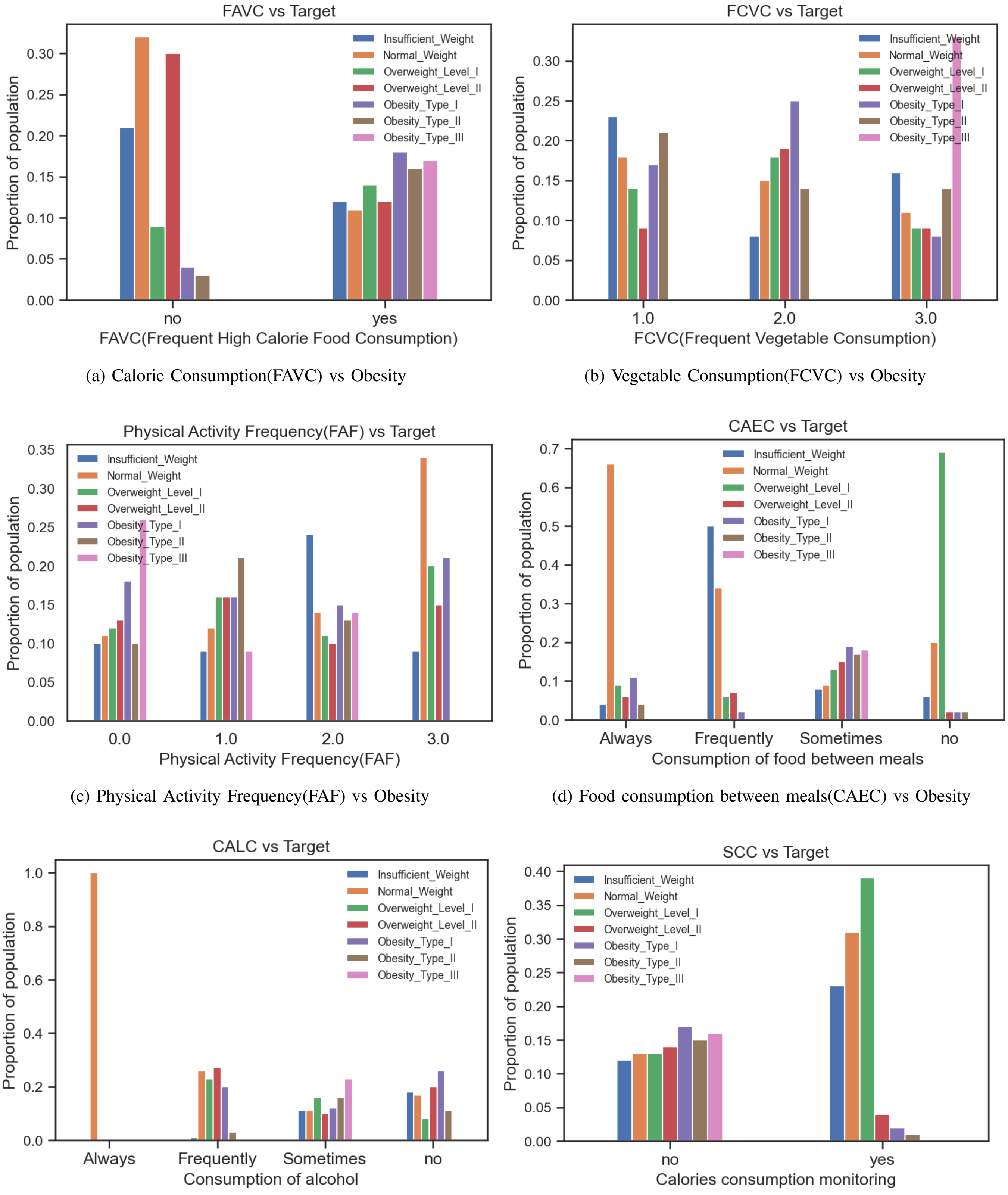
(a) Gender vs Obesity

(b) Family History vs Obesity

(c) Mode of Transport vs Obesity

(d) Tech Devices Usage vs Obesity

Fig. 1: Bivariate Analysis Plot 1



(a) Calorie Consumption(FAVC) vs Obesity

(b) Vegetable Consumption(FCVC) vs Obesity

(c) Physical Activity Frequency(FAF) vs Obesity

(d) Food consumption between meals(CAEC) vs Obesity

(e) Alcohol Consumption(CALC) vs Obesity

(f) Calories monitored vs Obesity

Fig. 2: Bivariate Analysis Plot 2

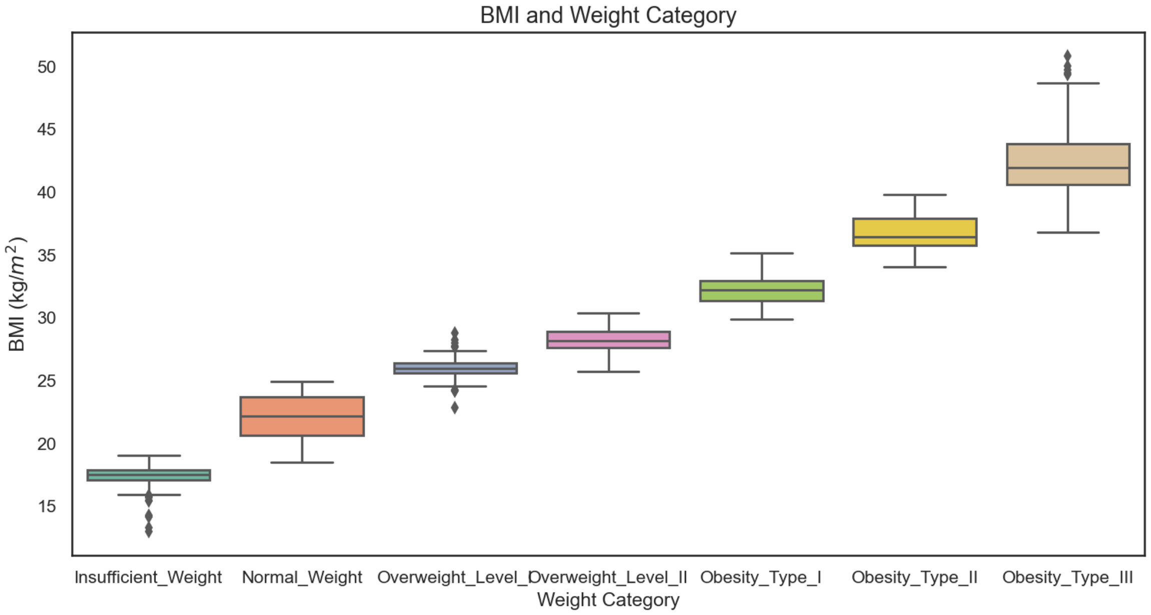


Fig. 3: BMI vs Obesity

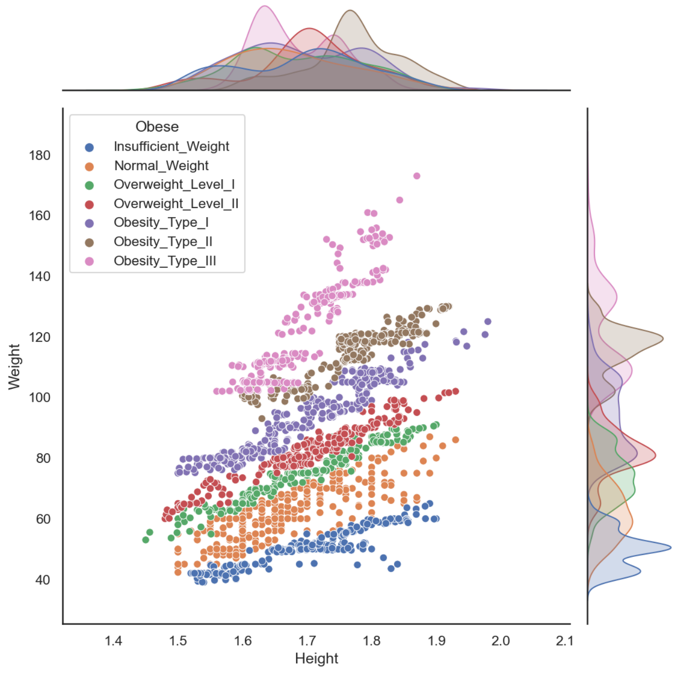


Fig. 4: Height and Weight

III. Tasks

A. Descriptive Analysis

The basic features of the dataset was summarized using descriptive statistics. Variability in the dataset was measured using means, mode, median, variance and standard deviation. The symmetricity and skewness of the dataset was also identified.

The average age of the individuals in the dataset is 24.31 years with height of 1.7 metres and weight of 86.58 kgs. The maximum number of participants are of age 18. The standard deviation of the age is 6.34 years as shown in table I. The interquartile range of age is 6.05, height is 0.138, weight is 41.957. Age and weight is positively skewed while height is negatively skewed. Skewness for categorical dataset have not been considered. Age is leptokurtic i.e heavy tailed with higher kurtosis than normal distribution, has longer tails and contains outliers.

B. Correlation Studies

Figure 4 shows the scatterplot and densityplot of height and weight. We can see linear pattern in the figure. Using this data, BMI was calculated and its boxplot was created as shown in figure 3. We can see that as BMI increases, the level of obesity also increases which is obvious. All the categorical variables were converted to ordinal variables and correlation analysis was done. The heatmap shown in figure 5 shows that BMI weight and height are correlated. CAEC(Food consumption between meals) is negatively correlated with weight. CAEC, SCC, FAF, TUE are negatively correlated with obese. Family history, FAVC, CH2O, CALC are positively correlated with obese. SMOKE seem to have no correlation with the target variable. Whereas, using Spearman's rank correlation coefficient, FCVC is not correlated to the target variable.

C. Statistical Tests

The number of male and female individuals is 1068 and 1043 which is comparable. The hypotheses to check if the means of the male and female population are equal are:

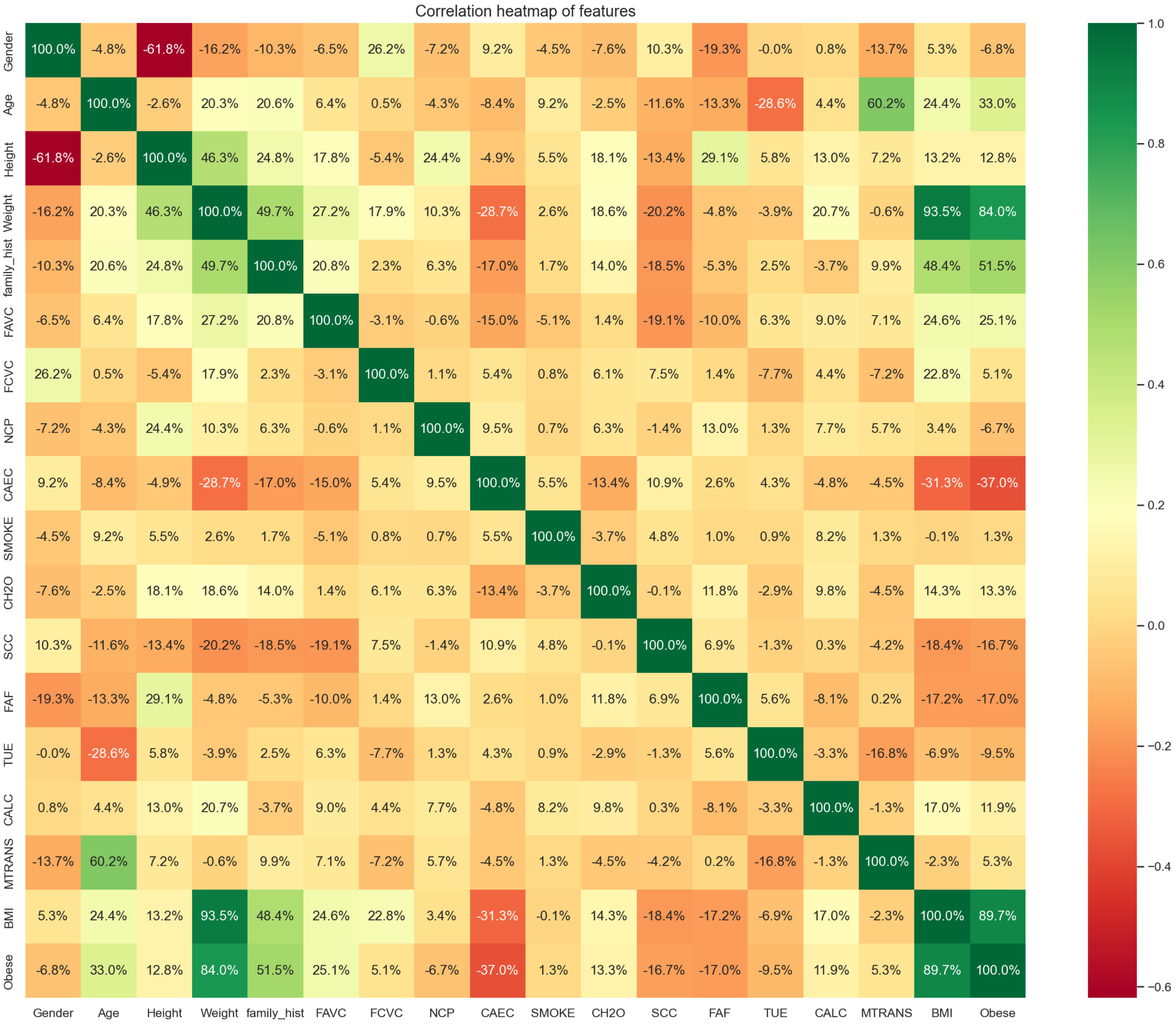


Fig. 5: Correlation on features (categorical features and target converted to ordinal)

H0: There is no difference between the means of age of male and female.

H1: There is a difference in the means of age of male and female.

Using two sample Z test, we rejected the null hypothesis which signifies that there is difference in means of the age of male and female.

Chi-squared test is the test when we have categorical variables from a population. It determines association or relationship between the two variables.

The relevant null hypothesis and alternative hypothesis are:

H0: There is no significant association between two features. i.e. independent

H1: There is significant association between two features. i.e. dependent

The chi-square test is used to check for independence. It is also used to find out the correlation between categorical variables in the data. The result obtained from chi2_contingency function of sciPy library is shown in the table below.

The null hypothesis stating that there is no relation between two variables is rejected if p-value $\leq$0.05. So from the table II we can say that there is association of target variable with all categorical variables.

D. Regression Analysis

After the dataset was cleaned, processed, analyzed, encoded and visualized; the data samples were shuffled randomly and split in standard 80:20 ratio i.e. 80% data for training and

TABLE II: Chi-square score between categorical variables

Attribute 2	**CAEC**	**CALC**	**FAVC**	**Gender**	**MTRANS**	**SCC**	**SMOKE**	**family_hist**	**Obese**
Attribute 1									
CAEC	NaN	0.0	0.0	0.0	0.0	0.0	0.059556	0.0	0.0
CALC	0.0	NaN	0.0	0.149593	0.0	0.023429	0.000011	0.346822	0.0
FAVC	0.0	0.0	NaN	0.003552	0.0	0.0	0.036638	0.0	0.0
Gender	0.0	0.149593	0.003552	NaN	0.0	0.000004	0.05726	0.000003	0.0
MTRANS	0.0	0.0	0.0	0.0	NaN	0.006492	0.419484	0.000001	0.0
SCC	0.0	0.023429	0.0	0.000004	0.006492	NaN	0.067637	0.0	0.000015
SMOKE	0.059556	0.000011	0.036638	0.05726	0.419484	0.067637	NaN	0.547493	0.0
family_hist	0.0	0.346822	0.0	0.000003	0.000001	0.0	0.547493	NaN	0.0
Obese	0.0	0.0	0.0	0.0	0.0	0.000015	0.0	0.0	NaN

remaining 20% for testing. A logistic regression model was fit on the training dataset and predictions were made on the test dataset. Python's scikit-learn library was used to create, predict and evaluate the logistic regression model. The performance of the model was defined and summarized in the confusion matrix shown in table III.

TABLE III: Confusion Matrix

Target Labels	Confusion Matrix							
Insufficient	46	5	2	1	0	0	0	54
Normal	11	28	11	4	4	0	0	58
Overweight I	0	6	37	10	5	0	0	58
Overweight II	0	2	11	25	17	3	0	58
Obesity I	0	2	4	6	43	12	3	70
Obesity II	0	0	0	0	0	60	0	60
Obesity III	0	0	0	0	0	0	65	65
Total	57	43	65	46	69	75	68	423

Model prediction and evaluation was done on test data i.e. 423 samples split from the dataset, and tabulated in the table III. Based on these predictions, classification matrix was created with accuracy, precision, recall and f1 score measure of the model. The macro f1 score was considered as the evaluation metric for the model as it calculates the average f1 score of all target variables treating them equally regardless of their counts or support values [1]. The logistic regression model generated is able to correctly classify an individual into one of seven level of obesity with macro f1 score of 0.71 and accuracy of 0.72.

IV. Conclusion

A. Contributions

In this paper, the obesity level dataset has been used to perform descriptive statistics, correlation analysis, statistical test and regression analysis. A logistic model as a regression model was proposed which classifies an individual into one of seven obesity level categories. Univariate and bivariate analysis of the predictors with target variable was performed to visualize the dependence of the target variable on predictors. Likewise, features that are correlated to target variables were found and verified with statistical tests. The logistic regression model built predicted the obesity leve with an average f1 score of 0.71.

TABLE IV: Classification Report

	precision	recall	f1-score	support
0	0.81	0.85	0.83	54
1	0.65	0.48	0.55	58
2	0.57	0.64	0.60	58
3	0.54	0.43	0.48	58
4	0.62	0.61	0.62	70
5	0.80	1.00	0.89	60
6	0.96	1.00	0.98	65
accuracy			0.72	423
macro avg	0.71	0.72	**0.71**	423
weighted avg	0.71	0.72	0.71	423

B. Avenues for further research

The categorical features available in the dataset can be explored to find further insights. Furthermore, more detailed statistical tests can be done to find and verify more associations from the dataset. Likewise, multivariate analysis can be conducted to discover the relationship between the predictors and target with greater confidence.

References

[1] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.