DATA SCIENCE PROJECT REPORT

On

Estimation of Obesity levels based on Eating Habits and Condition

Submitted as partial fulfillment for the Award of the

DATA SCIENCE CERTIFICATION

By

Shivam Gupta
2000320100159

Rishav Chourasia
2000320100137

Under the Supervision of

Ms. SHANU SHARMA

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

ABES ENGINEERING COLLEGE, GHAZIABAD

AFFILIATED TO

DR. A.P.J ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW

INDIA, 2023

DECLARATION

We hereby declare that the work being presented in this report entitled "**Estimation of Obesity Levels based on Eating Habits and Condition"** is an authentic record of our own work carried out under the supervision of Ms. Shanu Sharma.

The matter embodied in this report has not been submitted by **us** for the award of any other degree.

Shivam Gupta
2000320100159

Rishav Chourasia
2000320100137

Date: 19th May 2023

CERTIFICATE

This is to certify that the Project Report entitled "**Estimation of Obesity Levels based on Eating Habits and Condition"** which is submitted by Shivam Gupta and Rishav Chourasia in partial fulfillment of the requirement for the Data Science Certification in the Department of Computer Science & Engineering, ABES Engineering College (Affiliated to A.K.T.U, Lucknow) is a record of the candidate own work carried out by him under my supervision.

Shanu Sharma

Assistant Professor

Department of Computer Science & Engineering

ABES Engineering College, Ghaziabad, India

Date: 19th May 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Obesity has become a major public health issue with significant implications for individuals and society. For both people and society as a whole, obesity has emerged as a serious public health concern. The goal of this data science project is to create an estimation model for obesity levels based on indications of eating patterns and physical health. The research makes use of a dataset that contains participant data that has been anonymized and represents a wide range of people. Dietary habits, calorie intake, frequency of exercise, and body mass index are only a few of the features included in the dataset. Important patterns and relationships between these traits and obesity levels are found by exploratory data analysis and feature engineering. The development of a prediction model then uses machine learning methods like logistic regression or decision trees. By utilizing suitable methods like cross-validation or train-test splits, the model is trained and validated. Additionally, prospective difficulties are taken into account.

CHAPTER 1

INTRODUCTION

1.1. Problem Statement

Obesity is a growing concern in modern society, and it has been linked to various health issues such as heart disease, diabetes, and certain types of cancer. One way to tackle this issue is to develop effective methods for estimating obesity levels in individuals, based on their eating habits and physical condition. Eating Habits and Physical Conditions are the factors to develop a model that can accurately estimate obesity levels. The problem statement is to develop a predictive model that estimates the obesity levels of individuals based on their eating habits and physical condition. This model should be trained on a dataset that includes information on various factors such as age, gender, daily caloric intake, type of food consumed, physical activity level, and body mass index (BMI).

1.2. Motivation

The estimation of obesity levels based on eating habits and physical conditions is an important area of research as obesity is a major public health concern worldwide. Obesity is associated with a range of health problems, including diabetes, heart disease, and stroke, among others. The motivation to study the relationship between eating habits, physical condition, and obesity levels is driven by the need to address a major public health concern, promote healthy behaviors, and develop effective prevention and treatment strategies.

1.3. Objective

The objective of the estimation of obesity levels based on eating habits and physical condition is to develop a model that can accurately predict the obesity levels of individuals based on their eating habits and physical condition. Perform classification, clustering, and regression analysis on the dataset and identify the most important predictors of obesity levels. The ultimate objective of this problem statement is to develop a tool that can help individuals and healthcare professionals to better understand and address the issue of obesity.

1.4. Scope of Work The methodology involves several steps:

1. A dataset should be collected that includes information on various factors such as age, gender, daily caloric intake, type of food consumed, physical activity level, and BMI.

2. The collected data should be cleaned and preprocessed to remove any inconsistencies or missing values.
3. Feature selection involves identifying the most relevant features that contribute to the prediction of obesity levels. This can be done using techniques such as correlation analysis, and feature importance scores.

4. The selected features can then be used to train a machine learning model using regression analysis, classification analysis, decision tree, and clustering analysis techniques.

CHAPTER 2

LITERATURE STUDY

1. Based on their eating habits and physical condition, people from Mexico, Peru, and Colombia were studied to estimate their rates of obesity. The data consists of 2111 records with 17 attributes, and the records are identified by the class variable NObesity (Obesity Level), which enables the classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. With the help of the Weka tool and the SMOTE filter, 77% of the data was generated artificially, while 23% was directly obtained from consumers via a web platform. Using this information, intelligent computational tools can be created to determine a person's level of obesity and to develop new treatments for obesity.

2. Physical and mental health issues are brought on by obesity, which results from the body storing too much fat. Physical impairment has been linked to numerous aspects of reduced quality of life, including social distress, sexual function, self-esteem, and work-related quality of life. Over the past few decades, there has been a steady rise in the prevalence of obesity, which is presently unparalleled. Almost all ages, genders, and races have experienced this surge. According to these findings, segments of people in the highest weight categories (BMI > 40 kg/m2) increased proportionally more than those in the lower BMI categories (BMI 35 kg/m2). There is an urgent need to address obesity given the various and significant health problems that it is associated with.

3. The WHO predicts that by 2030, more than 40% of the world's population will be overweight, and more than a fifth of them will be obese. Obesity is a disease that affects both men’s and women's health, and in recent decades it has had an increasing tendency. Because of this, scientists have worked very hard to pinpoint the causes of obesity early on. There are methods that only calculate BMI, leaving out other important criteria like whether the person has a family history of obesity, how much time they spend exercising, their gene expression profiles, and other things. Based on supervised and unsupervised data mining methods such as the Light Gradient Boosting Machine (Light GBM) classifier and random forest, a computational intelligence model is developed in this study.

CHAPTER 3

PROPOSED APPROACH

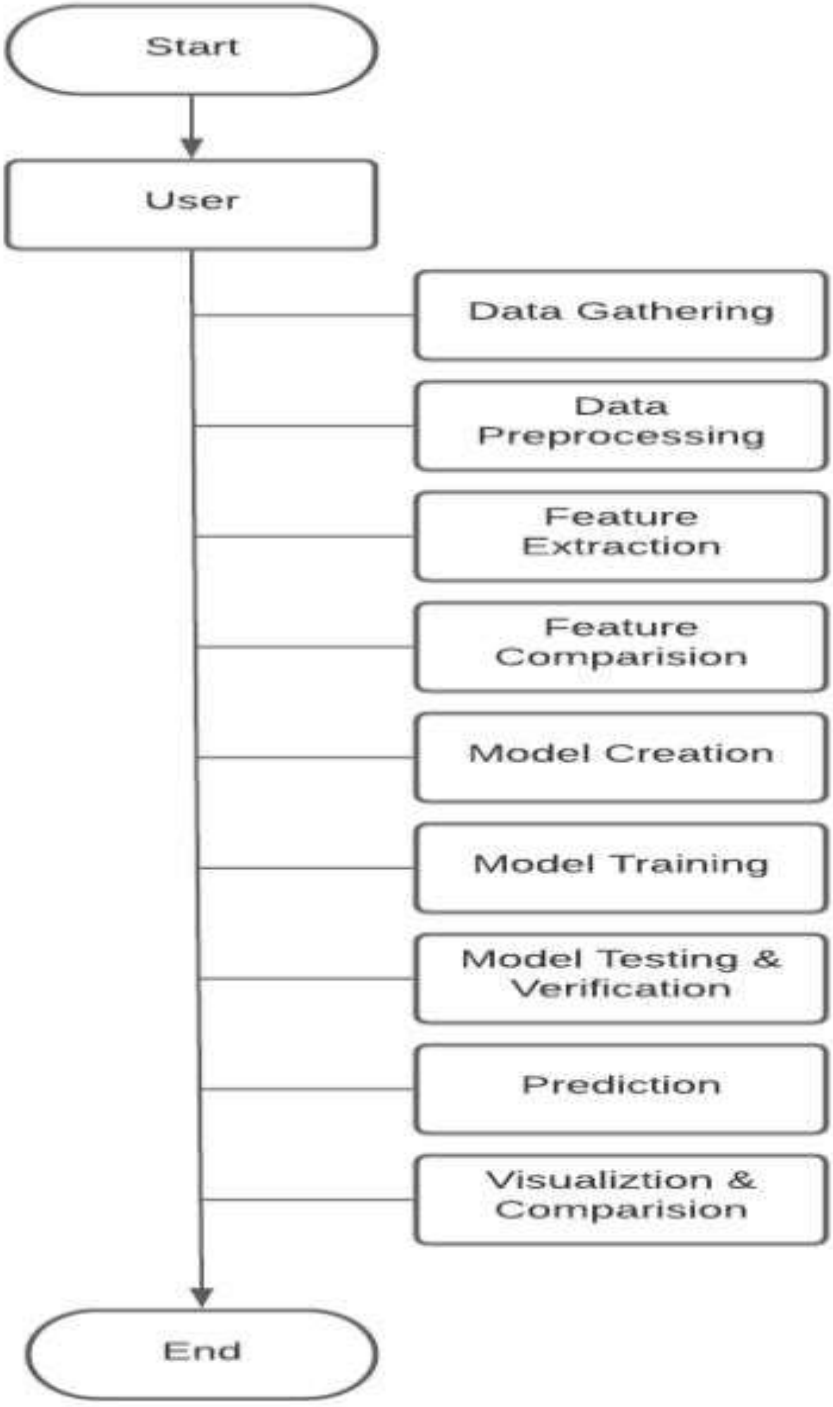

Fig 1.1. Proposed Strategy of the Model

To estimate obesity levels based on eating habits and physical condition, a data science project could follow the following proposed approach:

1. **Data Collection:** Gather a dataset that includes information about individuals' eating habits, physical activity levels, and obesity levels. This data can be collected through surveys, questionnaires, or other data sources such as health records or wearable devices. Ensure the dataset is representative and contains a sufficient number of observations.
2. **Exploratory Data Analysis (EDA):** Perform EDA to understand the characteristics of the dataset, identify missing values, and outliers, and examine the distributions and relationships between variables. This step helps in gaining insights into the data and informing subsequent data preprocessing steps.
3. **Data Preprocessing:** Prepare the dataset for modeling by addressing missing values, outliers, and data inconsistencies. Clean the data by handling missing values through imputation or deletion strategies. Additionally, normalize or standardize features as needed. Transform categorical variables into numerical representations using appropriate encoding techniques, such as one-hot encoding or label encoding.
4. **Feature Selection and Extraction**: Select relevant features that are likely to have an impact on obesity levels based on domain knowledge and statistical analysis. Consider factors such as eating habits, physical activity, demographics, and other relevant variables. Additionally, engineers new features that capture meaningful relationships or interactions between variables to improve model performance.
5. **Model Selection:** Choose an appropriate machine learning algorithm for obesity level estimation. Some possible options include logistic regression, decision trees, random forests, support vector machines (SVM), or gradient boosting algorithms like XGBoost or LightGBM. Consider the characteristics of the data, interpretability requirements, and the ability of the chosen model to handle the problem effectively.
6. **Model Testing and Verification:** Split the dataset into training and testing sets. Train the selected model on the training data and evaluate its performance using suitable evaluation metrics such as accuracy, precision,

recall, F1-score, or area under the ROC curve (AUC-ROC). Assess the model's generalization ability by evaluating it on the testing set.

7. **Model Optimization:** Fine-tune the model by tuning hyperparameters to improve its performance. Utilize techniques like cross-validation, grid search, or Bayesian optimization to find the optimal combination of hyperparameters that yield the best results.
8. **Visualization and Comparison:** Interpret the model's results to gain insights into the relationship between eating habits, physical condition, and obesity levels. Examine the coefficients or feature importance scores to identify the most influential factors contributing to obesity. Communicate and visualize the findings effectively to stakeholders.
9. **Deployment and Monitoring:** Deploy the trained model into a production environment, if applicable. Continuously monitor the model's performance and assess its accuracy over time. Update the model periodically as new data becomes available or as the problem evolves.

Remember that this proposed approach provides a general framework, and the specifics may vary based on the dataset, problem domain, and the chosen algorithms. It is crucial to iterate and refine each step based on the insights gained throughout the project.

CHAPTER 4

IMPLEMENTATION AND RESULTS

1. Data Exploration

```
dataset.describe()
```

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000
mean	24.312600	1.701677	86.586058	2.419043	2.685628	2.008011	1.010298	0.657866
std	6.345968	0.093305	26.191172	0.533927	0.778039	0.612953	0.850592	0.608927
min	14.000000	1.450000	39.000000	1.000000	1.000000	1.000000	0.000000	0.000000
25%	19.947192	1.630000	65.473343	2.000000	2.658738	1.584812	0.124505	0.000000
50%	22.777890	1.700499	83.000000	2.385502	3.000000	2.000000	1.000000	0.625350
75%	26.000000	1.768464	107.430682	3.000000	3.000000	2.477420	1.666678	1.000000
max	61.000000	1.980000	173.000000	3.000000	4.000000	3.000000	3.000000	2.000000

Table 1.1 Description of Data

For Numeric Data

- Made histograms to understand distributions
- Corrplot
- Pivot table comparing NObeyesdad levels across numeric variables

For Categorical Data

- Made bar charts to understand the balance of classes
- Made pivot tables to understand the relationship with NObeyesdad

Fig 1.2. Heatmap of the correlation

A heatmap of the correlation is a visual representation of the correlation matrix between Age, Height, and Weight. In data analysis and statistics, the correlation matrix measures the strength and direction of the linear relationship between pairs of variables.

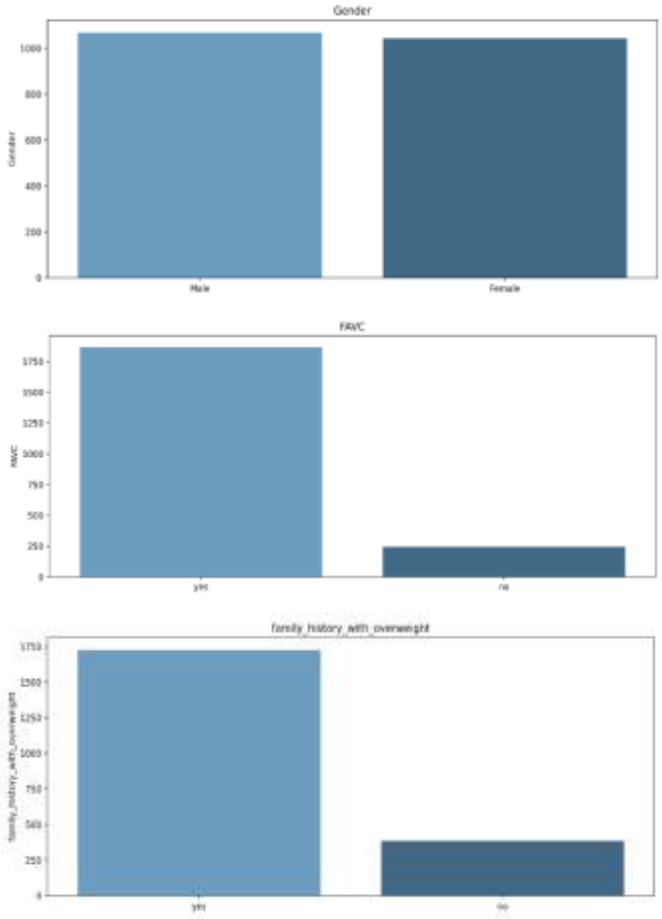

Fig 1.3. Some Categorical Graphs

Categorical graphs, also known as categorical plots or charts, are data visualizations that display the distribution or relationship between categorical variables such as Gender, FAVC, etc. Categorical variables represent qualitative or discrete data, such as categories, groups, or labels.

2. Data Visualization

```
plt.figure(figsize=(13,5))
sns.countplot(x='NObeyesdad', hue='Gender', data=df, palette='mako')
```

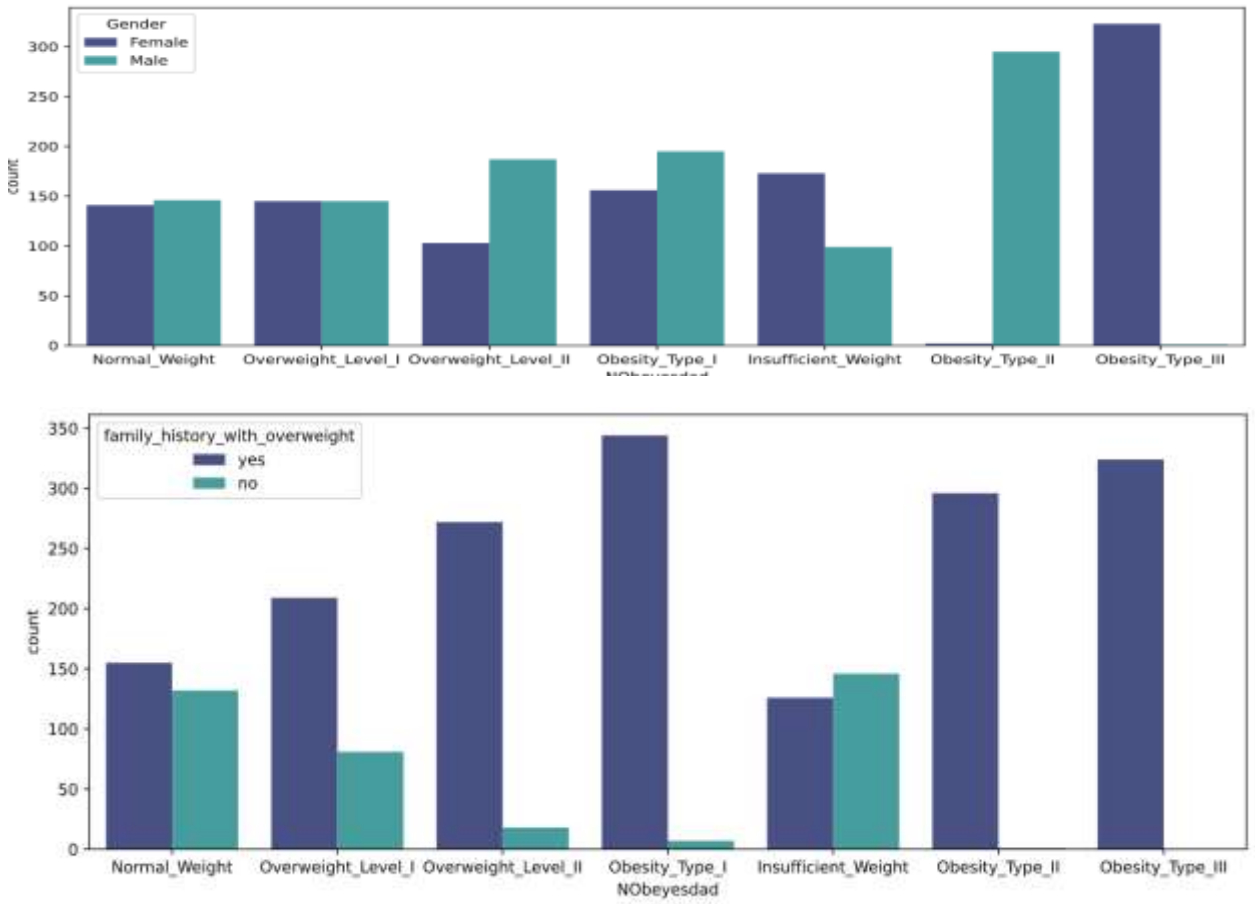

Fig 1.4. Counterplots

Counter plots are a valuable tool to gain insights into the distribution and frequency of categorical variables, allowing for easy comparison and interpretation of the data.

```
for i, col in enumerate(list(df_num.columns.values)):
    plt.subplot(1,3,i+1)
    df_num.boxplot(col)
    plt.grid()
    plt.tight_layout()
```

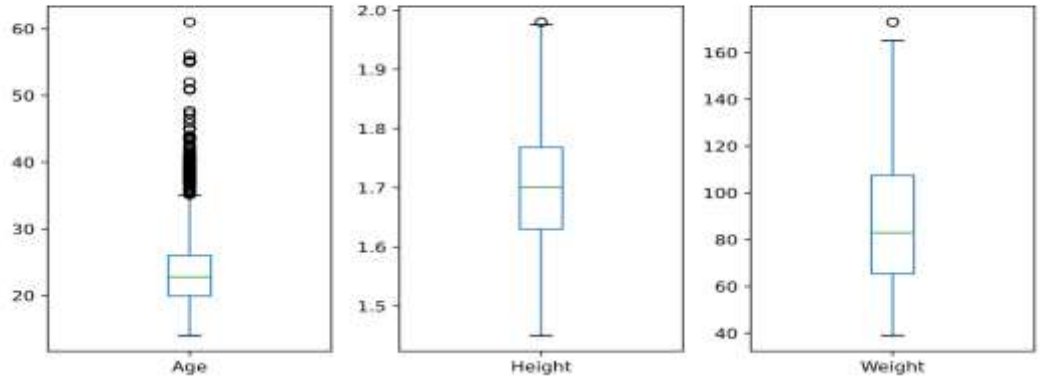

Fig 1.5. Outliers Detection

Outlier detection is a critical task in data science that involves identifying and analyzing data points that deviate significantly from the majority of the dataset. Outliers are observations that are rare, unusual, or aberrant compared to the typical patterns exhibited by the rest of the data.

	Gender_1	Gender_2	Age	Height	Weight	family_history_with_overweight_1	family_history_with_overweight_2	FAVC_1	FAVC_2
0	1	0	21	1.62	64	1	0	1	0
1	1	0	21	1.52	56	1	0	1	0
2	0	1	23	1.80	77	1	0	1	0
3	0	1	27	1.80	87	0	1	1	0
4	0	1	22	1.78	90	0	1	1	0
...	...	...	...	...	...	...	...	...	...
2106	1	0	21	1.71	131	1	0	0	1
2107	1	0	22	1.75	134	1	0	0	1
2108	1	0	23	1.75	134	1	0	0	1
2109	1	0	24	1.74	133	1	0	0	1
2110	1	0	24	1.74	133	1	0	0	1

1950 rows × 26 columns

Table 1.2 Categorical transforms on all data

Categorical transforms, also known as categorical encodings or transformations, are techniques used to convert categorical variables into numerical representations that can be utilized in data analysis and machine learning algorithms. Performing categorical transforms on all data is a common practice to prepare the categorical variables for further analysis.

3. Regression

OLS Regression Results

Dep. Variable:	NObeyesdad	R-squared:	0.955
Model:	OLS	Adj. R-squared:	0.954
Method:	Least Squares	F-statistic:	2132.
Date:	Fri, 25 Jun 2021	Prob (F-statistic):	0.00
Time:	11:38:14	Log-Likelihood:	-1139.8
No. Observations:	1950	AIC:	2320.
Df Residuals:	1930	BIC:	2431.
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.2812	0.079	28.988	0.000	2.127	2.436
Gender_1	1.1036	0.035	31.941	0.000	1.036	1.171
Gender_2	1.1776	0.048	24.711	0.000	1.084	1.271
Age	0.0319	0.003	10.872	0.000	0.026	0.038
Height	-7.4891	0.173	-43.411	0.000	-7.827	-7.151
Weight	0.0764	0.001	135.864	0.000	0.075	0.078
family_history_with_overweight_1	1.3001	0.044	29.636	0.000	1.214	1.386
family_history_with_overweight_2	0.9811	0.040	24.234	0.000	0.902	1.060
FAVC_1	1.1217	0.041	27.131	0.000	1.041	1.203
FAVC_2	1.1595	0.044	26.260	0.000	1.073	1.246
FCVC	-0.0008	0.018	-0.043	0.966	-0.037	0.035
NCP	0.0451	0.013	3.467	0.001	0.020	0.071
CAEC	-0.1432	0.023	-6.310	0.000	-0.188	-0.099
SMOKE_1	1.1856	0.048	24.596	0.000	1.091	1.280
SMOKE_2	1.0956	0.058	18.910	0.000	0.982	1.209

Fig 1.6. OLS Regression Result

OLS regression stands for Ordinary Least Squares regression, which is a widely used linear

a regression method for estimating the parameters of a linear relationship between a dependent variable and one or more independent variables.

When performing OLS regression, the result typically includes several key components that provide insights into the model and its performance.

4. Classification

```
In [413…
# Logistic Regression
lor = LogisticRegression(max_iter = 2000)
cv_lor = cross_val_score(lor,scaled_X_train,y_train,cv=10)
print(cv_lor)
print(cv_lor.mean())
```

```
[0.81751825 0.84671533 0.89051095 0.77372263 0.86861314 0.84558824
 0.90441176 0.84558824 0.83088235 0.86029412]
0.8483844997853156
```

```
In [415…
# Decision Tree
dt = tree.DecisionTreeClassifier(random_state = 1)
cv_dt = cross_val_score(dt,scaled_X_train,y_train,cv=10)
print(cv_dt)
print(cv_dt.mean())
```

```
[0.94890511 0.95620438 0.89781022 0.9270073  0.91240876 0.91176471
 0.94117647 0.94117647 0.94117647 0.94852941]
0.9326159295835122
```

Fig 1.7. Logistic Regression and Decision Tree

Logistic Regression is a statistical model used for binary classification problems, where the goal is to predict the probability of an instance belonging to a particular class. Decision Trees are non-parametric models that make predictions by recursively partitioning the input feature space based on a series of if-else decision rules.

```
# Decision Tree
dt = tree.DecisionTreeClassifier(random_state = 1)
cv_dt = cross_val_score(dt,scaled_X_train,y_train,cv=10)
print(cv_dt)
print(cv_dt.mean())
```

```
# Random Forest
rf = RandomForestClassifier(random_state = 1)
cv_rf = cross_val_score(rf,scaled_X_train,y_train,cv=10)
print(cv_rf)
print(cv_rf.mean())
```

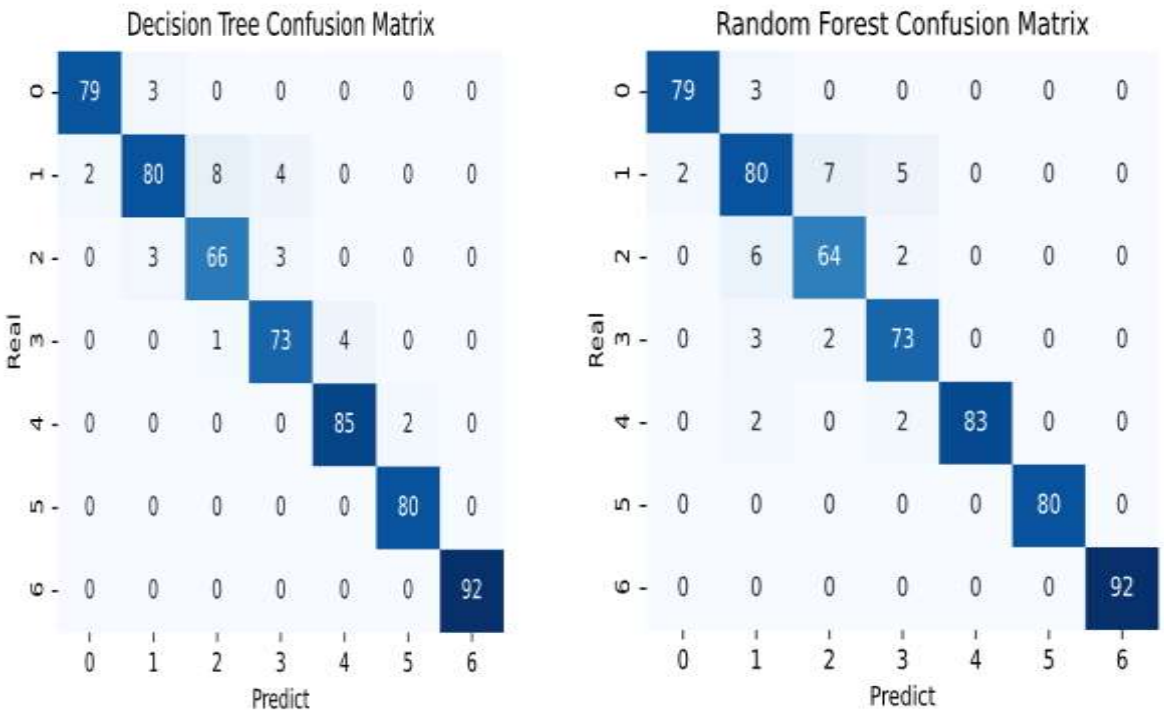

Fig 1.8. Decision Tree and Random Forest Confusion Matrix

For a Decision Tree classifier, the confusion matrix provides insights into the model's performance in terms of classifying instances into different categories. A Random Forest is an ensemble method that combines multiple Decision Trees to make predictions. The confusion matrix for a Random Forest aggregates the individual predictions from each Decision Tree within the ensemble.

```
# K-means: centroids-based clustering
X_clust = data_ord_enc.iloc[:, [17,25]].values
print(X_clust)
```

```
[[0 1]
 [3 1]
 [2 1]
 ...
 [1 6]
 [1 6]
 [1 6]]
```

Fig 1.9. K-Means

K-means is a popular unsupervised machine learning algorithm used for clustering, which aims to group similar data points together based on their feature similarity. It is a simple and efficient algorithm that partitions a dataset into K clusters, where K is a pre-defined number chosen by the user.

CHAPTER 5

CONCLUSION

In conclusion, the estimation of obesity levels based on eating habits and physical condition is a significant topic in data science projects aimed at understanding and addressing the obesity epidemic. Through the analysis of various data points related to individuals' eating habits and physical characteristics, it is possible to develop models that can predict and estimate obesity levels.

Some Conclusions based on analysis:

1. FCVC, NCP, CH2O, FAF, and TUE charts are very messy on barplot.
2. Looking at the mean and standard deviation, it is noted that in the Age and Weight columns, there are very different values and different scales.
3. Standard Errors assume that the covariance matrix of the errors is correctly specified.
4. The smallest eigenvalue is 8.83e-26. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.
5. 0.95 is a good result for the r2_score (coefficient of determination) regression score function, and 0.21 MAE does not affect the forecast as much.
6. Accuracy of Decision Tree: 0.95
7. Accuracy of Random Forest: 0.94

Classifier performance:

- Decision Tree (93.2%)
- Random Forest (93.1%)
- Logistic Regression (84.8%)
- Support Vector Classifier (83.1%)
- K Nearest Neighbor (77.0%)
- Naive Bayes (49.1%)

In summary, the estimation of obesity levels based on eating habits and physical condition holds promise for understanding and combating the obesity epidemic. By leveraging data science techniques, researchers and practitioners can gain insights that contribute to the development of personalized interventions and public health strategies to address this critical public health concern.

REFERENCES

1. 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*, *25*, 104344.
2. Kivrak, M. (2021). Deep learning-based prediction of obesity levels according to eating habits and physical condition. *The Journal of Cognitive Systems*, *6*(1), 24-27.
3. Quiroz, J. P. S. (2022). Estimation of obesity levels based on dietary habits and condition physical using computational intelligence. *Informatics in Medicine Unlocked*, *29*, 100901.
4. Cui, T., Chen, Y., Wang, J., Deng, H., & Huang, Y. (2021, May). Estimation of Obesity levels based on Decision trees. In *2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM)* (pp. 160-165). IEEE.
5. Yagin, F. H., Gülü, M., Gormez, Y., Castañeda-Babarro, A., Colak, C., Greco, G., ... & Cataldi, S. (2023). Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique. *Applied Sciences*, *13*(6), 3875.
6. Celik, Y., Guney, S., & Dengiz, B. (2021, July). Obesity Level Estimation based on Machine Learning Methods and Artificial Neural Networks. In *2021 44th International Conference on Telecommunications and Signal Processing (TSP)* (pp. 329-332). IEEE.