# IPS Academy, Indore

**Institute of Engineering and Science**

**(A UGC Autonomous Institute)**

**Department of Computer Science & Engineering**

**Data Science**

**2023-24**



*A Project-I entitled*

*“DailyDharma : A Obesity Prediction Web Application ”*

*For the partial fulfillment for the award of the Bachelor of Technology (Computer Science & Engineering – Data Science) Degree by Rajiv Gandhi Proudyogiki Vishwavidhyalaya, Bhopal.*

# IPS Academy, Indore

**Institute of Engineering and Science**

**(A UGC Autonomous Institute)**

**Department of Computer Science & Engineering**

**Data Science**

**2023-24**



*A Project-I entitled*

*“DailyDharma : A Obesity Prediction Web Application”*

*For the partial fulfillment for the award of the Bachelor of Technology (Computer Science & Engineering – Data Science) Degree by Rajiv Gandhi Proudyogiki Vishwavidhyalaya, Bhopal.*

# IPS Academy, Indore

**Institute of Engineering and Science**

**(A UGC Autonomous Institute)**

**Department of Computer Science & Engineering**

**Data Science**

**2023-24**



*CERTIFICATE*

This is to certify that the Project-I entitled

*“DailyDharma : A Obesity Prediction Web Application”*

*has been successfully completed by the following students*

Dhirendra Singh , Anany Pandit & Hrishita Jain

*in partial fulfillment for the award of the Bachelor of Technology (Computer Science & Engineering – Data Science) Degree by Rajiv Gandhi Proudyogiki Vishwavidhyalaya, Bhopal during the academic year 2022-23 under our guidance.*

Dr. Prateek Nahar Dr. Neeraj Shrivastava Associate Professor HOD

Dr. Archana K. Chowdhary Principal

## TABLE OF CONTENTS

**Acknowledgement i**

[**Abstract ii**](#_30j0zll)

[**List of Tables iii**](#_1fob9te)

[**List of Figures iv**](#_3znysh7)

[**List of Abbreviations v**](#_2et92p0)

[**CHAPTER 1: INTRODUCTION 1**](#_tyjcwt)

* 1. [Objective & Scope of the Project 2](#_3dy6vkm)
  2. Theoretical Background 2
  3. Evaluation 3
  4. Data quality and features. 3
  5. Need 3
  6. [Definition of Problem](#_1t3h5sf) 3

**CHAPTER 2: SYSTEM REQUIREMENTS & ANALYSIS 5**

* 1. [User Requirements](#_4d34og8) 6
     1. [Functional Requirements](#_2s8eyo1) 6
     2. [Non-Functional Requirements](#_17dp8vu) 6
  2. [Use Case Descriptions](#_3rdcrjn) 7
  3. [Details of Hardware](#_26in1rg) 7
  4. [Software Used](#_lnxbz9) 8

**CHAPTER 3: DATA COLLECTION & PREPROCESSING 9**

* 1. [Data Source](#_35nkun2)  10
  2. [Data Quality & Quantity](#_1ksv4uv)  12
  3. [Data Preprocessing Techniques](#_44sinio) 13
  4. [Feature Selection & Engineering](#_2jxsxqh) 14

[**CHAPTER 4: EXPLORATORY DATA ANALYSIS 1**](#_z337ya)**5**

* 1. [Data Visualization](#_3j2qqm3)**16**

**CHAPTER 5: MODEL BUILDING & DEPLOYMENT 23**

* 1. [Model Selection 2](#_4i7ojhp)4
  2. [Model Training & Tuning](#_2xcytpi) 25

**CHAPTER 6: RESULTS & DISCUSSION 27**

* 1. [Model Performance](#_3whwml4) 28
  2. Result Discussion 29
  3. Deployment in Web App30

[**CHAPTER 7: CONCLUSION 3**](#_3as4poj)**3**

* 1. [Summary 3](#_1pxezwc)4
  2. [Limitations](#_49x2ik5) 35
  3. Limitations for Future Work 36

[**CHAPTER 8: REFERENCES**](#_2p2csry) **37**

[**USER MANUAL**](#_147n2zr) **40**

## ACKNOWLEDGEMENT

I would like to express my heartfelt thanks to my guide, **Dr. Prateek Nahar , CSE-DS Dept. IPS Academy, Institute of Engineering & Science** for his guidance, support, and encouragement during the course of my study for B. Tech. (CSE-DS) at IPS Academy, Institute of Engineering & Science, Indore. Without his endless effort, knowledge, patience, and answers to my numerous questions, this Project-I would have never been possible. It has been a great honor and pleasure for me to do Project-I under your supervision.

My gratitude will not be complete without mention of **Dr. Archana Keerti Chowdhary, Principal, IPS Academy Institute of Engineering & Science & Dr. Neeraj Shrivastava, HOD CSE, IPS Academy, Institute of Engineering & Science** for the encouragement and giving me the opportunity for this Project work.

I also thank my friends who have spread their valuable time for discussion/suggestion on the critical aspects of this report. I want to acknowledge the contribution of my parents and my family members, for their constant motivation and inspiration.

Finally, I thank the almighty God who has been my guardian and a source of strength and hope in this period.

**Student Name**   **Enrollment Number**

Dhirendra Singh 0808DS211039

Anany Pandit 0808DS211016

Hrishita Jain 0808DS211050

## ABSTRACT

The project "DailyDharma: An Obesity Prediction Web App" presents a unique approach to predicting obesity types by integrating ancient Vedic principles with modern medical studies. Leveraging machine learning techniques, specifically decision trees and random forest classifiers, the model analyzes general inputs and individual habits to forecast the likelihood of obesity occurrence in individuals.

Data collection involved a blend of local surveys targeting individuals aged 16-61 and curated datasets from platforms like Kaggle. By amalgamating diverse sources, the dataset was tailored to meet the project's specific requirements.

Implementation involved the development of a web application using web development languages and Streamlit, a framework for creating interactive web applications. This allowed for the seamless deployment and representation of the predictive model in a user-friendly format accessible via the web.

o

The results demonstrate the efficacy of the proposed machine learning model in accurately predicting obesity types. Notably, the model identifies key contributing factors and provides valuable insights into the determinants of obesity progression. By bridging ancient wisdom with contemporary technology, DailyDharma offers a novel approach to addressing health concerns, potentially revolutionizing how we understand and combat obesity.

## LIST OF TABLES

1. Table 1: Survey Form
2. Table 2: Cleaned Survey Excel with New Output Columns
3. Table 3 : Final Dataset
4. Table 4 : A description of numerical attributes

## LIST OF FIGURES

Fig. 1: Distribution of Obesity Classes.

Fig. 2: Pie Chart on the distribution of Adults based on health outputs.

Fig. 3: Bar Graph representing the distribution of Adults based on their overweight family history

Fig. 4: Bar Graph represents the distribution of Adults based on their Frequent consumption of high caloric food

Fig. 5: Pie Chart represents the distribution of Adults based on their Modes of Transportation Used

Fig. 6: Heatmap representing insights :

Age vs. Weight

FAF vs. Weight

FCVC vs. FAF

FCVC vs. NCP

NCP vs. CH2O

Fig. 7: A visual representation of the working of Random Forest Classifier

Fig. 8: Validation Curve for the model

Fig. 9: Screenshot of Machine Learning Model Display

Fig. 10: Screenshot of DailyDharma Website

Fig. 11 : A Screenshot of the Streamlit web app

## LIST OF ABBREVIATIONS

| **S.NO** | **Short** | **Meaning** |
| --- | --- | --- |
| 1 | pd | Panda library |
| 2 | np | Numpy library |
| 3 | plt | Matplotlib.pyplot |
| 4 | sns | Seaborn |
| 5 | cols | columns |
| 6 | EDA | Exploratory Data Analysis |
| 7 | st | Streamlit |
| 8 | csv | Comma Separated Values |
| 9 | ID | Identification Document |
| 10 | pred | Predict |
| 11 | CALC | Alcohol consumption |
| 12 | TUE | Time using technology devices |
| 13 | FAF | Physical activity frequency |
| 14 | FAVC | Frequent consumption of high caloric food |
| 15 | FCVC | Frequency of consumption of vegetables |
| 16 | NCP | Number of main meals |
| 17 | CAEC | Consumption of food between meals |
| 18 | CH2O | Consumption of water daily |
| 19 | SCC | Calories consumption monitoring |
| 20 | MTRANS | Transportation used |
| 21 | HTML | Hypertext Markup Language. |

# CHAPTER 1 INTRODUCTION

### INTRODUCTION

#### Objective & Scope of the Project

The objective of DailyDharma: Obesity Prediction ML model is to predict the likelihood of an individual being facing or can face the condition of one of the 3 types of obesity in future based on various factors such as variations in daily habitual consumptions of food and drinks, family history and other relevant data. The scope of this model is to help adults identify the potential chances of getting health issues starting from the obesity phase and to make informed decisions like changing habits and routines based on data-driven insights.

#### Theoretical Background

#### Foundation

The DailyDharma: Obesity Prediction ML model is built on the theoretical foundation of machine learning, which is a subfield of artificial intelligence that enables computers to learn from data and make predictions or decisions without being explicitly programmed.

#### Paradigm

The model uses supervised learning, where it is trained on labeled data, which includes information on the Personal basic data of an adult individual based on their daily habits . The model then uses this labeled data to learn patterns and relationships between a person’s habits, attributes and outcomes, mentioning which type of obesity risk he/she has high , which it can use to predict the likelihood of a new random user may have any type of obesity risk at present or future.

#### Data Quality & Features

The model's effectiveness depends on the quality and quantity of the data used for training and the features selected as input variables. The features selected for the model should be relevant to promotion decisions, such as height, age, weight , and Family history with overweight.

#### Evaluation

Additionally, the model should be evaluated for its accuracy, precision, and recall to ensure that it is making accurate predictions and not biased against certain groups of Adults. Theoretical principles of fairness, ethics, and transparency should be applied to ensure that the model does not perpetuate biases or lead to unfair promotion decisions

#### 

#### Need

Overall, the DailyDharma: Obesity Prediction ML model is a valuable tool for Health condition of people. Theoretical principles of fairness, ethics, and transparency should be applied to ensure that the model does not perpetuate biases or lead to unfair promotion decisions management and can help people make data-driven health related decisions. However, it should be used in conjunction with Doctors and health coach practices and not be the sole factor in making promotion decisions.

#### Definition of Problem

The burgeoning issue of unhealthy lifestyle habits prevalent in society today, including poor dietary choices, irregular sleep patterns, frequent alcohol consumption, and a lack of regular exercise, often leads individuals down a path towards obesity and other health complications. Despite the potential severity of these issues, existing apps predominantly focus on isolated aspects of health, failing to provide a comprehensive analysis of overall body health based on user inputs.

Furthermore, while regular checkups from healthcare professionals offer accurate health assessments, the associated costs pose a significant barrier to access for many individuals, particularly students and those facing financial constraints. This financial burden often deters individuals from seeking essential health services, leaving them vulnerable to undetected health issues and hindered preventative measures.

Furthermore, while regular checkups from healthcare professionals offer accurate health assessments, the associated costs pose a significant barrier to access for many individuals, particularly students and those facing financial constraints. This financial burden often deters individuals from seeking essential health services, leaving them vulnerable to undetected health issues and hindered preventative measures.

# CHAPTER 2

**SYSTEM REQUIREMENTS AND ANALYSIS**

### SYSTEM REQUIREMENTS AND ANALYSIS

#### User Requirements

User requirements for DailyDharma: Obesity Prediction ML model can be divided into two categories: functional and non-functional requirements.

#### Functional requirements:

1. **Data Collection and Management**: The system should be able to collect and store Adults’ data, including Body and Age specifications, Food and drinks consumptions records, past family overweight history.
2. **Prediction Mode**l: The system should be able to predict the likelihood of an adult facing obesity or can face in future.
3. **Model Training and Evaluation**: The system should be able to train the prediction model on historical data and evaluate its performance on new data.
4. **Visualization and Reporting**: The system should be able to present the predictions and insights in an intuitive and easily understandable format, such as charts, graphs, and reports.

#### Non-functional requirements:

1. **Security**: The system should ensure that employee data is stored securely and access to the system is restricted only to authorized personnel.
2. **Performance**: The system should be able to provide predictions in a timely manner, without any significant delays or errors.
3. **Scalability:** The system should be able to handle large amounts of data and support multiple users simultaneously.

#### Use Case Descriptions

1. We log into the system and upload Adults’ data, including age , body specifications and much more.
2. The system cleans and pre-processes the data to ensure its quality and consistency.
3. We select the prediction model and training algorithm to be used for the analysis.
4. The system trains the prediction model on the historical data and evaluates its accuracy and fairness.
5. The New person can then input his/her data for which an Obesity prediction is required.
6. The system uses the trained prediction model to generate predictions for each person.
7. The system presents the predictions and insights in an intuitive and easily understandable format, such as charts, graphs, and reports.
8. The system can learn from new data and feedback, and we can update the model accordingly.

#### 2.3 Details of Hardware

The hardware requirements for an employee promotion prediction ML model will depend on the size of the dataset, the complexity of the machine learning algorithms used, and the speed of predictions required. Here are some general hardware recommendations:

1. **CPU**: A multi-core processor with a clock speed of at least 2 GHz is recommended to handle the computational load of training and evaluating machine learning models.
2. **Memory:** The amount of memory required will depend on the size of the dataset and the complexity of the models. At least 8 GB of RAM is recommended for most machine learning tasks, but larger datasets may require more.
3. **Storage:** Adequate storage space is required to store the employee data, the machine learning models, and any other related files. A solid-state drive (SSD) is recommended for faster read and write speeds.

#### 2.4 Software Used

1. Programming languages: Python is a popular choice for building machine learning models due to its extensive library of machine learning frameworks and tools, such as Scikit-learn, TensorFlow, Keras, and PyTorch.
2. Integrated development environments (IDEs): IDEs such as PyCharm, Jupyter Notebook, and Spyder provide a comprehensive environment for developing and testing machine learning models.
3. Data preprocessing and visualization: Pandas, NumPy, and Matplotlib are commonly used libraries for preprocessing and visualizing data.
4. Machine learning frameworks: Scikit-learn is a popular open-source library for building machine learning models, providing a wide range of algorithms for classification, regression, and clustering. TensorFlow and Keras are deep learning frameworks that are commonly used for building neural networks.
5. Web Hosting: Using Streamlit for web hosting provides a convenient way to showcase your ML model to stakeholders, and allows non-technical users to interact with the model without needing to run code locally.

# CHAPTER 3 DATA COLLECTION &

**PREPROCESSING**

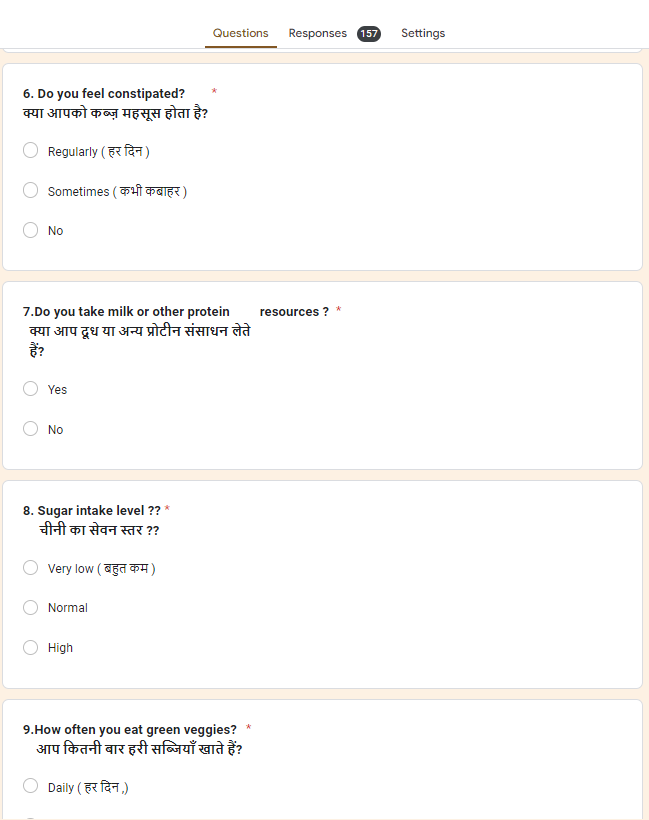
### DATA COLLECTION & PREPROCESSING

#### Data Collection Through Local Level Surveys And Internet

For this project, We first conducted a survey at our local level. The Survey comprises

of a google form with 15 major questions related to person’s habitual actions that

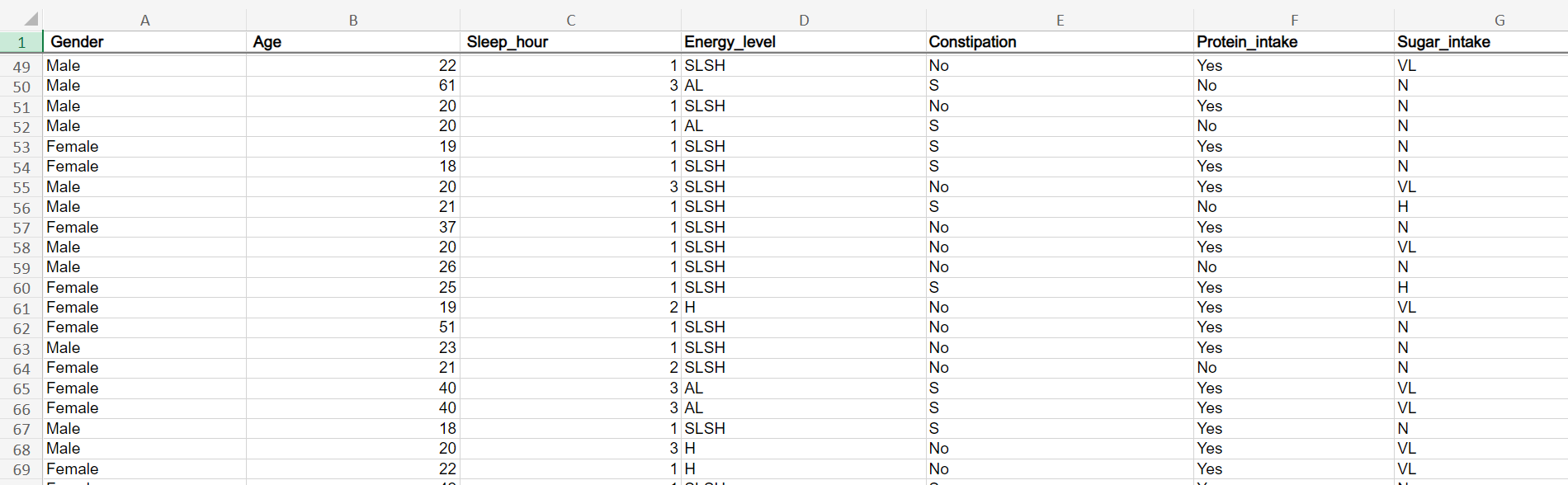
They do it on a daily basis .

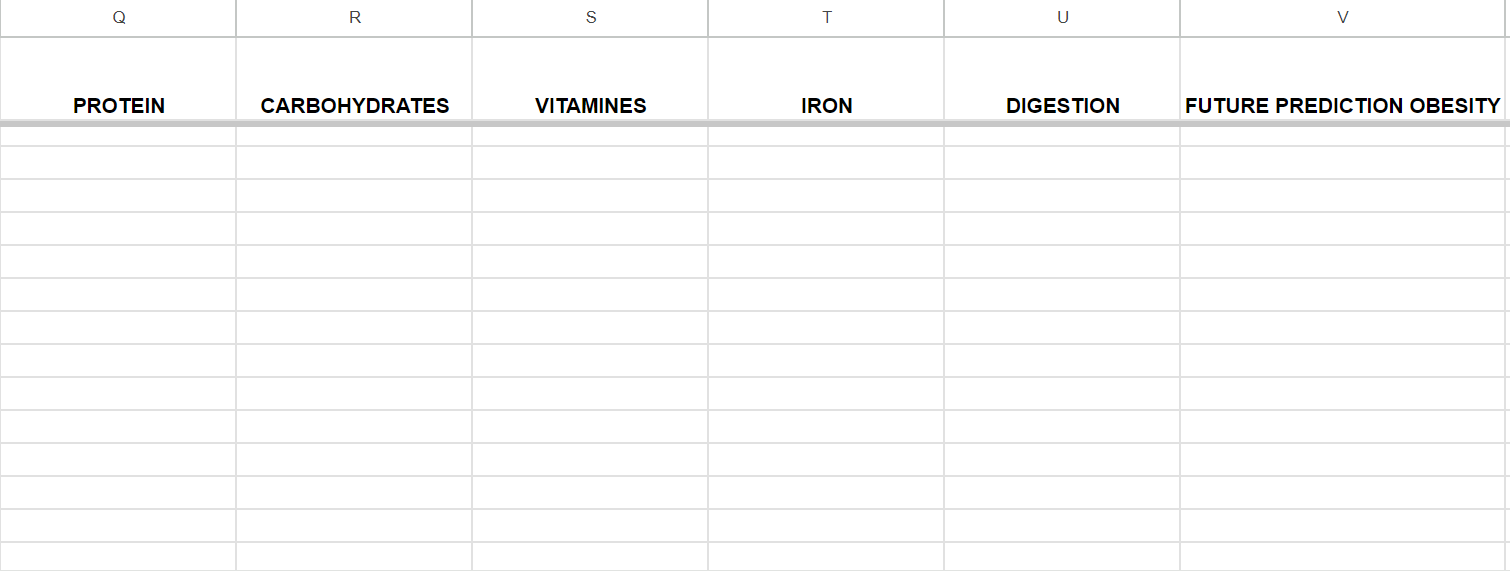


#### Table 1: Survey Form

Then managed to maintain all the responses into an excel sheet , Cleaned all the hindi

typed sentences and added 5 desired new output columns .





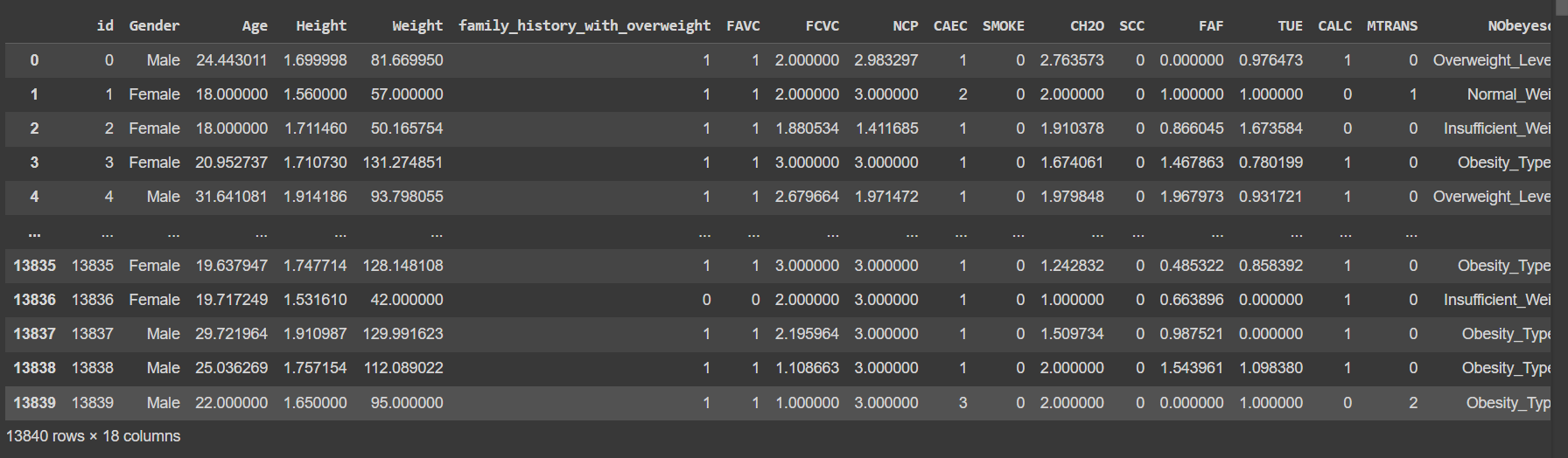
**Table 2 : Cleaned Survey Excel with New Output Columns**

At last we eliminated , some undesired data columns and added some more data

columns to our dataset by taking reference from the dataset available at kaggle .

we basically mixed both local survey and dataset from kaggle to increase row count ,

more relevant data to get our desired result.



**Table 3 : Final Dataset**

#### Data Quality & Quantity

The dataset consists of data of 13840 individuals from ages 16-61 years, with each individual being described by 16 distinct attributes regarding his/her age , body specifications, etc.

Earlier we planned to make a model that can predict 5 outputs , Carbohydrates , Iron ,Vitamines , Proteins and Future Prediction of Obesity but after researching we found there are many more imp factors that can affect the chances of obesity . And there are

3 types of obesity type1, type2, type3 , therefore regathered the data by taking reference from kaggle and decreased our desired output column to 1 .

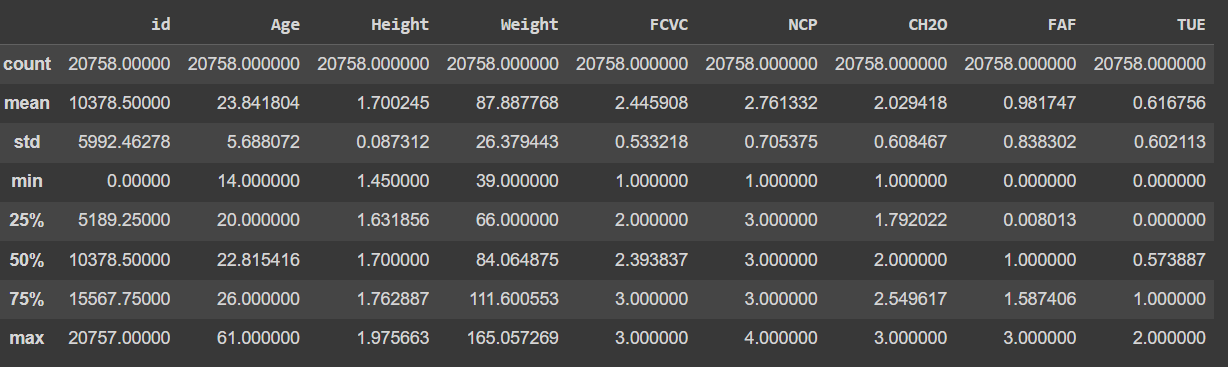
#### Data Pre-processing Techniques

Data preprocessing is an essential step in any machine learning project. It involves preparing and cleaning the data to ensure that it is in a suitable format for analysis. Here are some common data pre-processing techniques:

1. Data cleaning: This involves removing or correcting missing values, handling outliers, and correcting inconsistencies in the data.
2. Feature scaling: This involves scaling the values of the features to a standard range, such as between 0 and 1, to ensure that they contribute equally to the analysis.
3. Feature selection: This involves selecting the most relevant features that are useful for the analysis and removing any redundant or irrelevant features.
4. Encoding categorical variables: This involves converting categorical variables, such as vegetables\_consumptions or transportation\_medium, into numerical values that can be analyzed by machine learning algorithms.
5. Handling imbalanced data: This involves addressing situations where the data is imbalanced, i.e., when the number of instances in one class is much

greater than the other. Techniques such as oversampling, under sampling,

and synthetic data generation can be used to address this issue.



#### Table 4: A description of numerical attributes

#### Feature Selection & Engineering

For the prediction ML model, feature selection and engineering can involve selecting the most relevant features that are useful for predicting individuals’ Obesity Chances , such as Frequent consumption of high caloric food, Consumption of food between meals, Consumption of water daily , age, Physical activity frequency .

# CHAPTER 4 EXPLORATORY DATA ANALYSIS

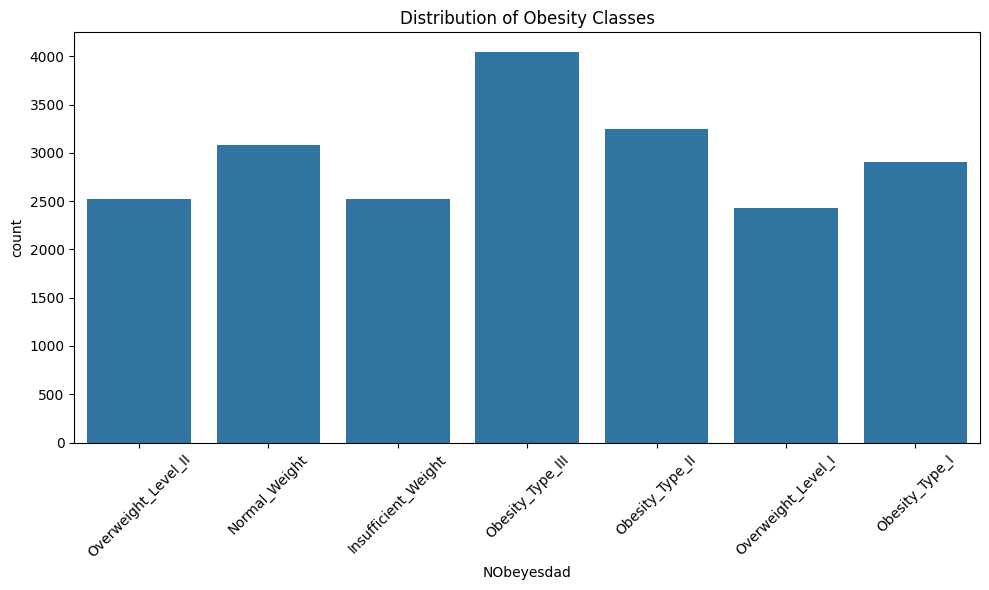
### EXPLORATORY DATA ANALYSIS

#### Data Visualization

Data Visualization is an important aspect of any data science project as it helps us to get insights from the data, and helps us to know our data better. Also, Data once visualized is easy to represent graphically and is a good way of explaining data to people from non-technical backgrounds.

In this project, extensive use of Data Visualization tools such as bar graphs and multiple bar graphs has been made with some use of Pie Charts also.

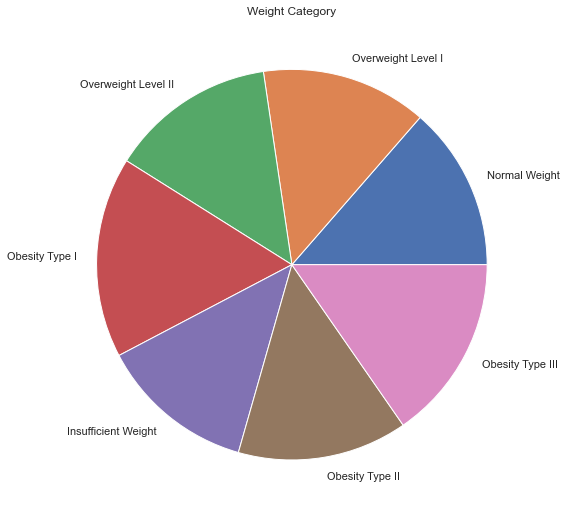
Now, let’s explore the visualizations we’ve carved from our Dataset.



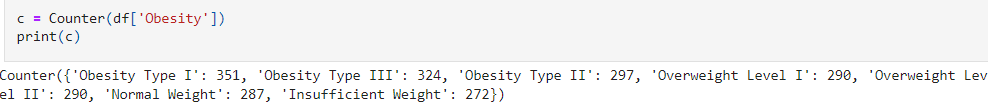
#### Figure 1

The above bar graph represents distribution of Adults according to their Obesity

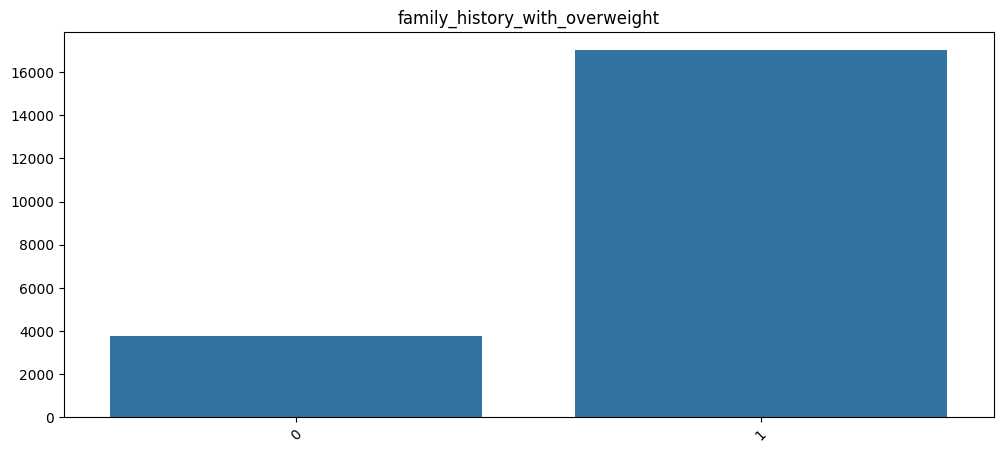
Classes or types , We see that number of People having chances of Obesity\_type3 is

highest in the dataset

#### Figure 2

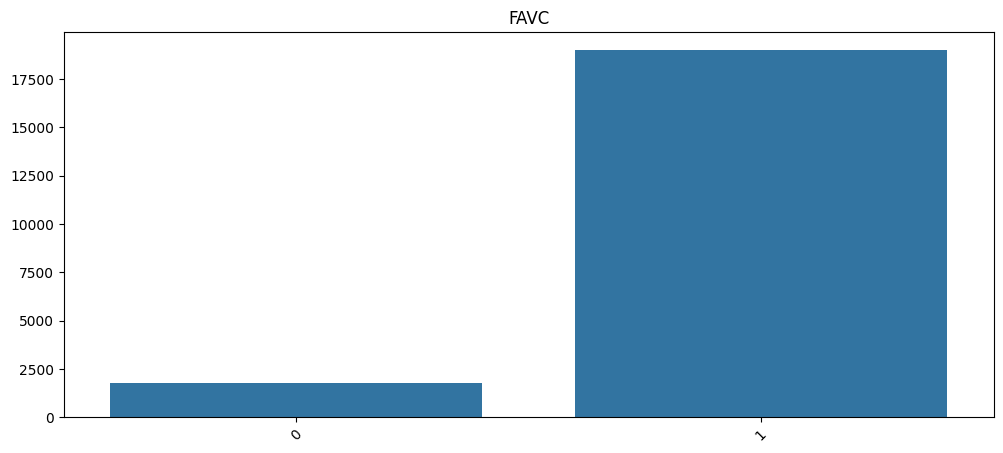
Figure 2 represents the distribution of Adults based on health outputs. 

#### 



#### Figure 3

Figure 3 represents the distribution of Adults based on their overweight family history. Close to 16000 people said yes of having family history with overweight and close to 2000 people said no.

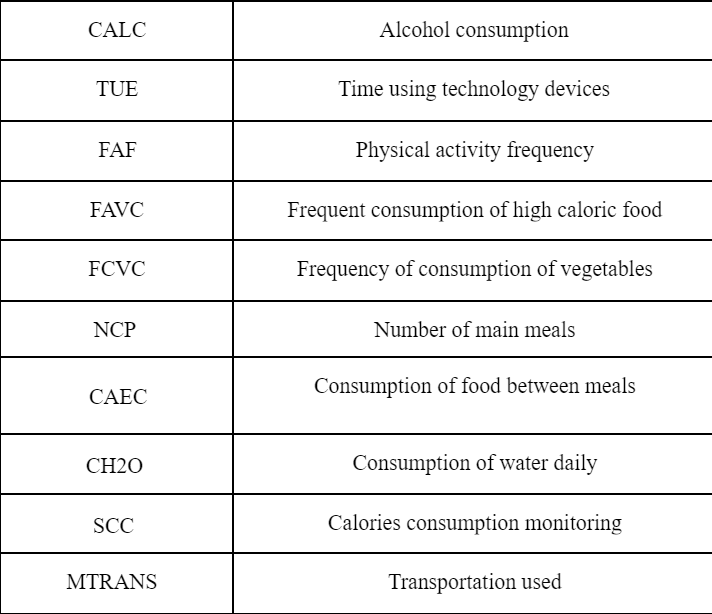


#### Figure 4

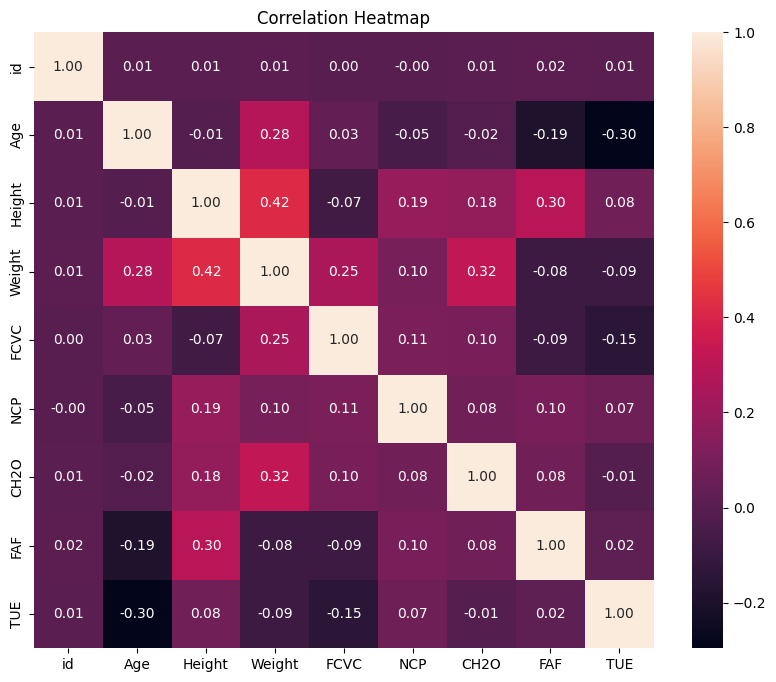
Figure 4 represents the distribution of Adults based on their Frequent consumption of high caloric food . Close to 18000 people said yes of having Frequent consumption of high caloric food and close to 2500 people said no.

#### Figure 5

Figure 5 represents the distribution of Adults based on their Modes of Transportation Used . Close to 16000 people selected 0 means “ Public Transportation “ , Close to 4000 people selected 1 means “ Automobiles “ and close to 1500 people selected 2 means “ Walking “ .



#### 



#### Figure 6

Figure 6, represents a heatmap giving following insights :-

1. **Age vs. Weight**

The correlation between age and weight is very weak and positive (0.01).

This means that there is a very slight tendency for weight to increase with

age.

1. **FAF vs. Weight**

There is a weak negative correlation (-0.19) between physical activity

frequency and weight. This means that people who exercise more tend to weigh less.

1. **FCVC vs. FAF**

There is a weak positive correlation (0.30) between frequency of vegetable

consumption and physical activity frequency. This means that people who

eat more vegetables tend to also exercise more.

1. **FCVC vs. NCP**

There is a very weak negative correlation (-0.09) between frequency of vegetable consumption and number of main meals. This means that there is a very slight tendency for people who eat more vegetables to eat fewer main meals.

1. **NCP vs. CH2O**

There is a weak negative correlation (-0.15) between the number of main meals and consumption of water daily. This means that there is a very slight tendency for people who eat more main meals to drink less water.

# CHAPTER 5

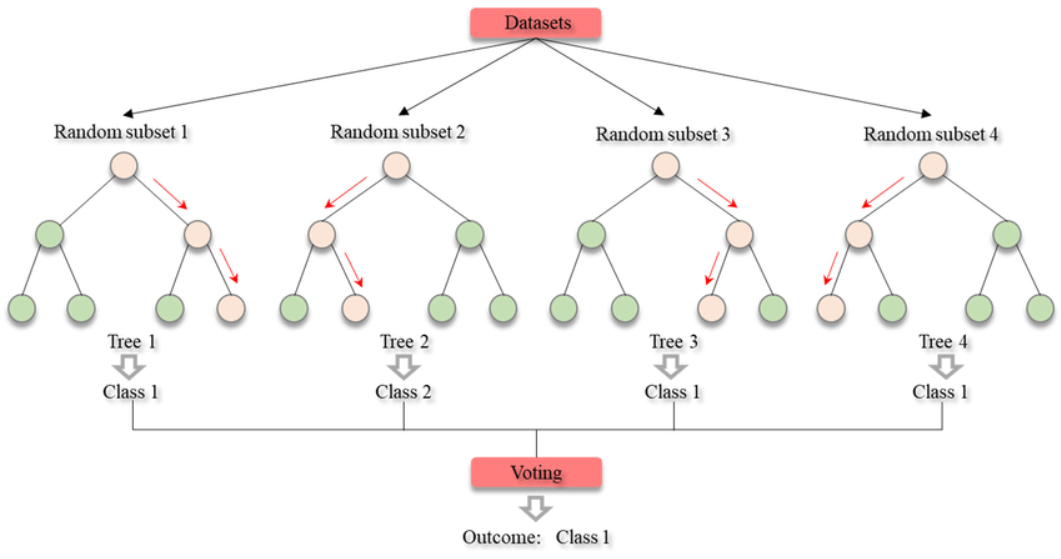
# MODEL BUILDING

### MODEL BUILDING AND EVALUATION

#### Model Selection

Random Forest Classifier was used for the purpose of this project, the reason for choosing this model is the fact that much research has been done on this problem statement concerning Obesity Prediction and random forest classifier proved to be the best Machine Learning model for predicting obesity Prediction .

Random Forest algorithm is a powerful tree learning technique in [Machine Learning](https://www.geeksforgeeks.org/ml-machine-learning/). It works by creating a number of [Decision Trees](https://www.geeksforgeeks.org/decision-tree/) during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of [overfitting](https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/) and improving overall prediction performance. In prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks) This collaborative decision-making process, supported by multiple trees with their insights, provides an example of stable and precise results. Random forests are widely used for classification and regression functions, which are known for their ability to handle complex data, reduce overfitting, and provide reliable forecasts in different environments.



#### Fig 7: A visual representation of the working of Random Forest Classifier

#### Model Training & Tuning

Model training and tuning for gradient boosting classifier involves several steps that are crucial for achieving high performance and accuracy. These steps include:

1. **Data Preprocessing:**

This step involves cleaning the data to handle missing values, normalizing or standardizing features, encoding categorical variables, and splitting the dataset into training and testing sets.

2. **Training the Random Forest Model:**

Random Forest is an ensemble learning method that constructs multiple decision

trees during training. Each decision tree is trained on a random subset of the

training data and a random subset of the features.

For each tree in the forest:

2.1 Randomly select a subset of the training data (bootstrapping).

2.2 Randomly select a subset of features (typically the square root of the

total number of features) for each split in the tree.

2.3 Grow the tree using the selected data and features until a stopping criterion is met (e.g., maximum tree depth, minimum samples per leaf).

2.4 Repeat the above steps to create multiple trees in the forest.

3. **Making Predictions:**

Once the Random Forest is trained, it can be used to make predictions on new or

unseen data.

3.1 To predict the type of obesity for a new individual, the Random Forest

considers the attributes of that individual.

3.2 Each decision tree in the forest independently predicts the type of obesity

based on the individual's attributes.

3.3 Each decision tree in the forest independently predicts the type of obesity

based on the individual's attributes.

3.4 For classification, the final prediction is determined by a majority vote

among all the decision trees. The class with the most votes becomes the

predicted class for the individual.

3.5 In the case of ties, some implementations may use additional strategies to

break the tie

# CHAPTER 6

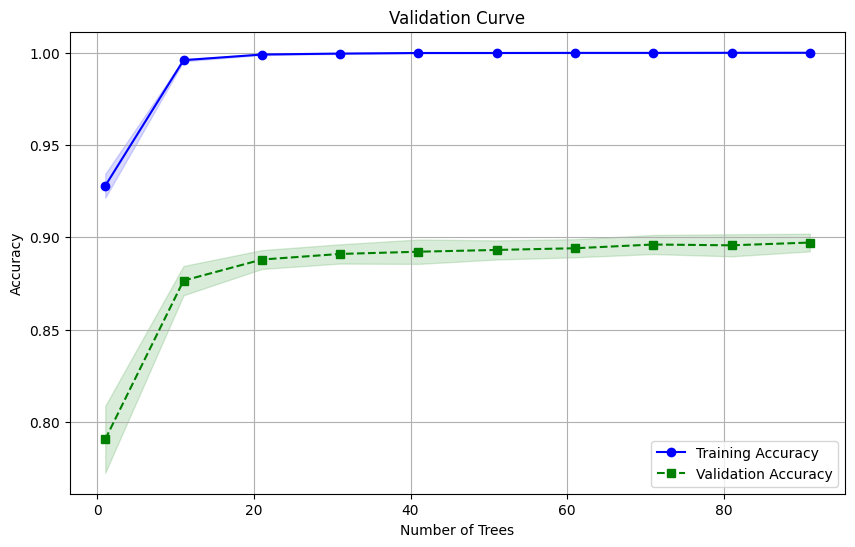
# RESULTS AND DEPLOYMENT

### RESULTS AND DEPLOYMENT

#### Model Performance

Cross-validation on the resampled dataset shows excellent results. The average Test set accuracy is about 89% with precision and recall around the same neighborhood. Thus, it can be concluded that the Random Forest Classifier is a good predictor for this dataset.

A validation curve is made. The validation curve is a plot of the train and test accuracies vs a parameter under optimization.



#### Fig. 8 : Validation Curve for the model

The validation curve for the test accuracy identifies the optimal value of max\_depth for the classifier. In this case the max\_depth value of 18 yields an average test accuracy of 89.44%. This accuracy is quite good and shows that the model has learnt the classes well and is highly trustable in its predictions.

The validation curve is intuitive when plotting for a single parameter. The GradientBoostingClassifier has many parameters such as learning\_rate, max\_depth, n\_estimators etc. When the best value for these have to be found together, GridSearchCV is used.

#### Result Discussion

Now it's time to understand the meaning and motive of the results that our machine learning model wants to convey with us.

According to our reference from the research paper

**“Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico”**

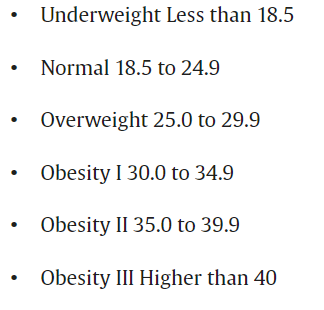
1. Our dataset “**train.csv**” , was labeled the visible outputs

according to the BMI formula .



1. So the criteria , for one of the possible type of overweight , normal

Weight & types of Obesity are based on the range of BMI scores guided by WHO.



1. So the final outputs , which is given by our DailyDharma: Obesity

Prediction Machine learning model is totally based on the combination of BMI scored labeled dataset and the experience gained

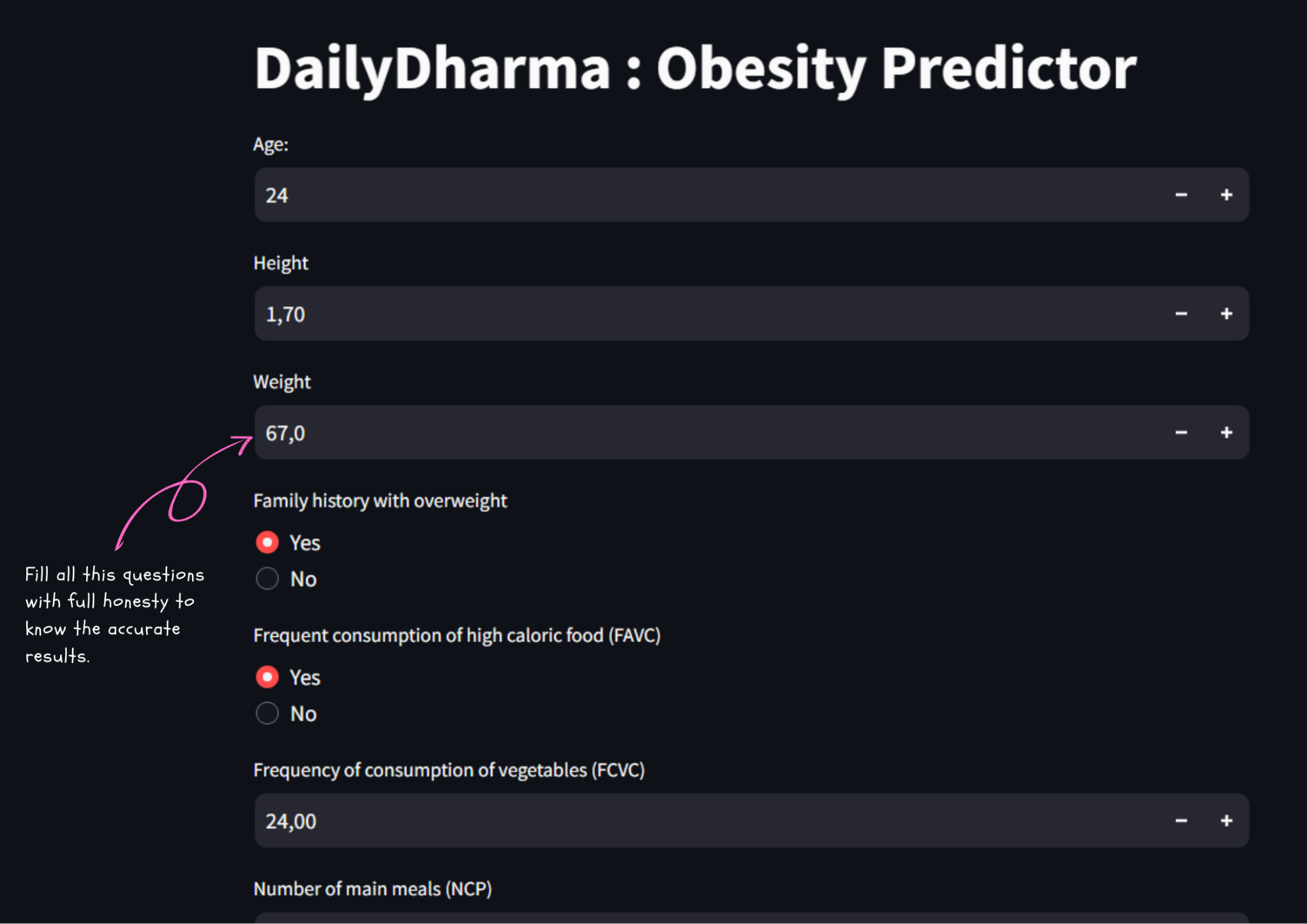
from the thousands of entries. Outputs can be in decimal and needs to convert in percentage by self.

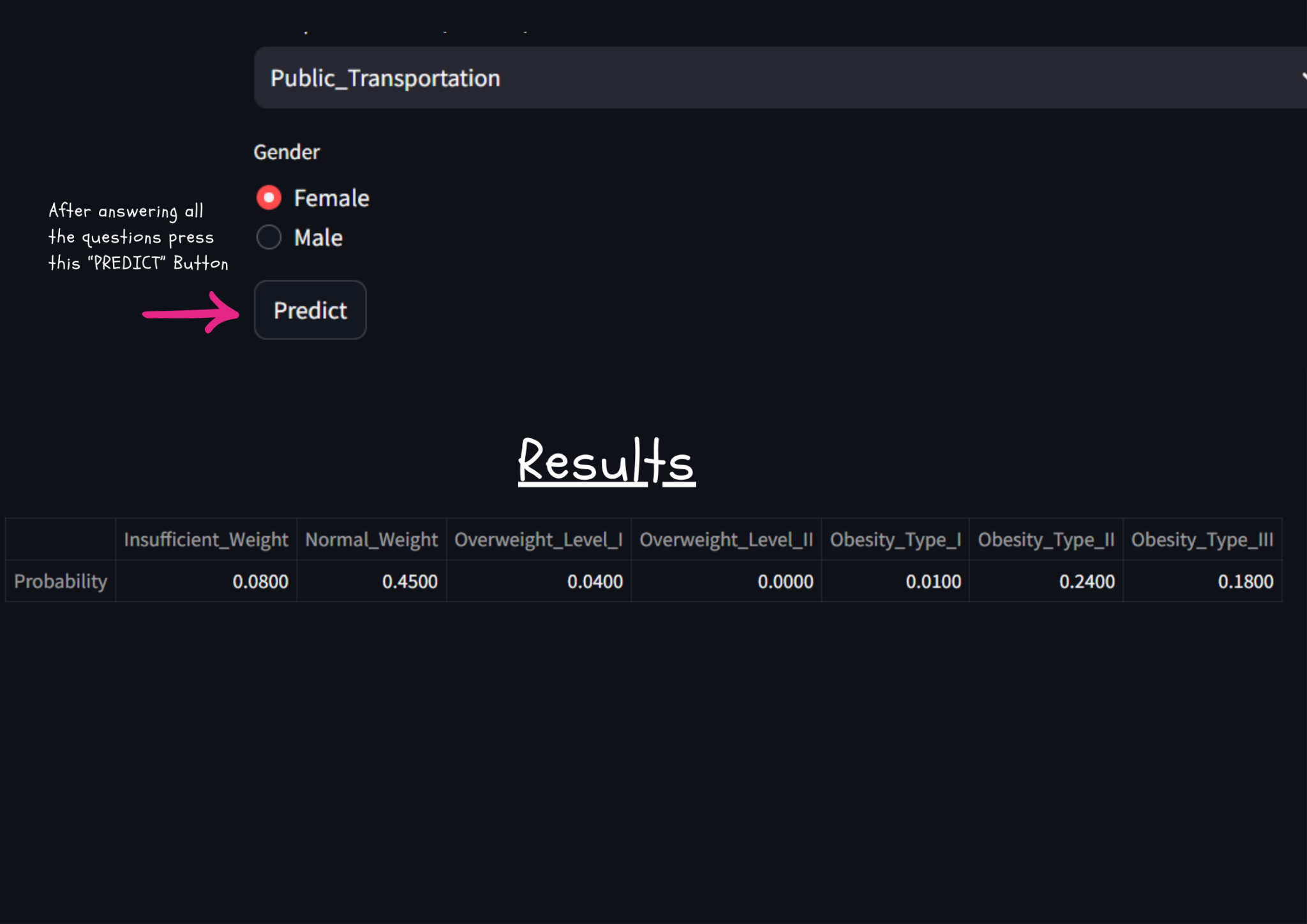
#### Deployment of Model into WebApp

After successful training and testing of the model, now here comes the time of

deploying the prediction model into a responsive representative format. We have

taken help of Streamlit. Streamlit is a free and open-source framework to rapidly build and share beautiful machine learning and data science web apps. It is a Python-based library specifically designed for machine learning engineers. The snapshots of results are here :-



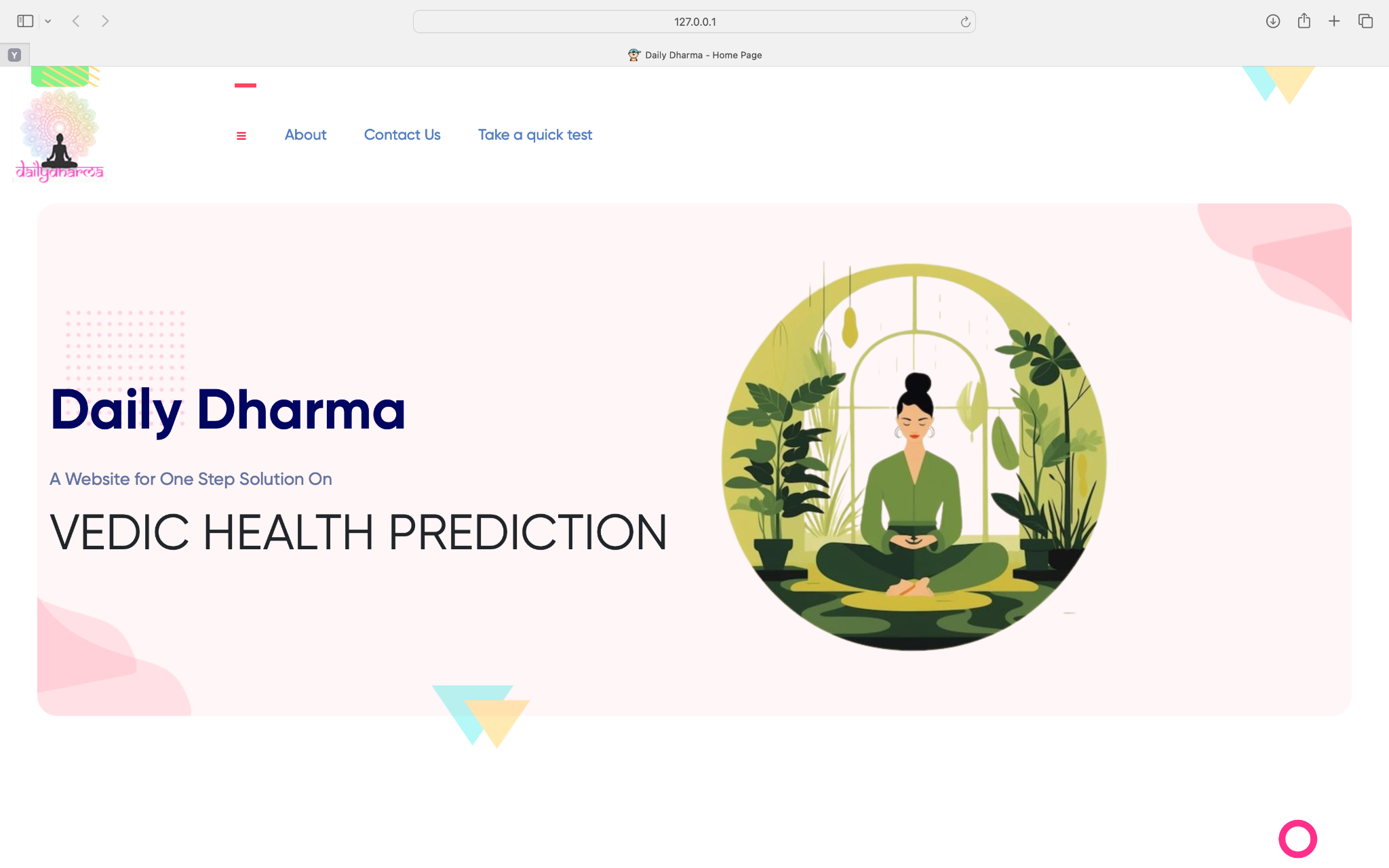
**Figure 9 : Screenshots of ML Model Display** 

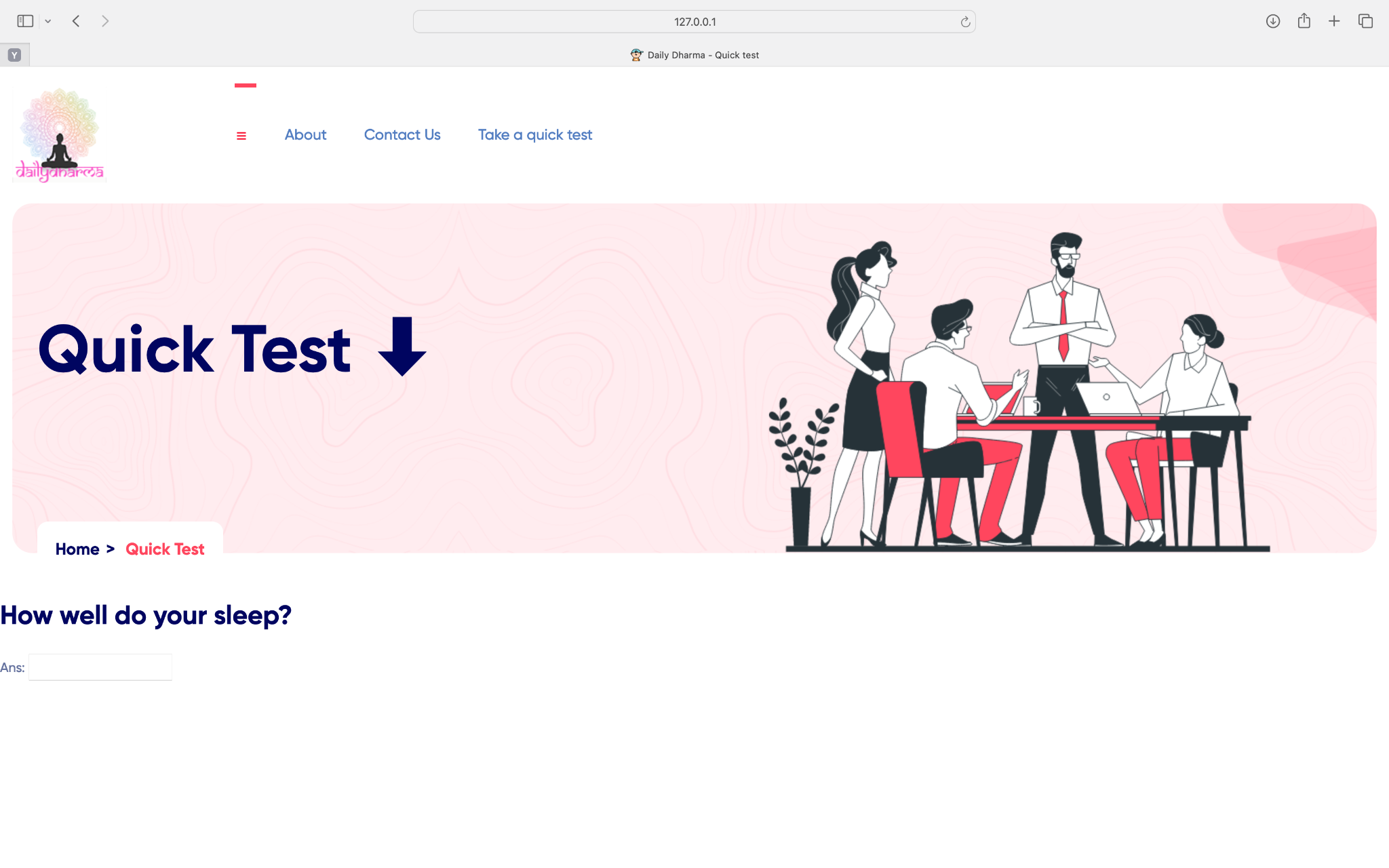
It's just a Streamlit Display or representation of the ML model which is made by us , but to launch it in future or make a hub for blogs , test and other features combined ,

We needed a beautiful dynamic website.

Therefore, We made a beautiful Vedic Themed Website from scratch using HTML,

CSS & JS.





**Figure 10 : Screenshot of DailyDharma Website**

# 

# CHAPTER 7 CONCLUSION

### CONCLUSION

#### Summary

The goal of the obesity health prediction project was to develop a predictive model using a random forest classifier that could accurately predict whether an Individual is at his/her health phase or obesity phase.. The project used data from a kaggle dataset and local survey that included information about Adults between 16-61 years old such as their age, gender, eating habits , etc ..

Data cleaning was the first step in the project. The raw data contained missing values. These issues were addressed by imputing missing values, and standardizing the formatting of the data.

Exploratory data analysis was then performed to gain insights into the data and identify patterns and relationships between the variables. This involved visualizing the data using plots and charts, and conducting statistical tests to determine correlations and significance levels between variables. The analysis revealed that variables such as eating and drinking related questions were strongly correlated with the outcome variable of Obesity, while variables such as age and gender had a weaker correlation. A random forest classifier model was then built using the cleaned and transformed data. The model was trained using a training set and evaluated using a validation set.. The final model had an accuracy of more than 89%, indicating that it was able to predict obesity level with a high degree of accuracy.

The model was then tested on a separate test set to assess its performance on unseen data. The test results showed that the model had an accuracy of 89%, indicating that it was able to generalize well to new data. A classification report was also generated to provide more detailed information on the model's performance, including precision, recall, F1 score, and support. The report showed that the model had high precision and recall for both the promoted and unpromoted classes.

In conclusion, the DailyDharma Obesity prediction project demonstrated the effectiveness of using a random forest classifier model to predict obesity level base

transformation, exploratory data analysis, model building using hyperparameter tuning, and test results analysis. The resulting model had a high accuracy and was able to generalize well to new data, providing valuable insights into the factors that contribute to employee promotions.

#### Limitations

Although the Obesity prediction project was successful in developing a predictive model for prediction of obesity in adults , there are several limitations to the project that should be considered:

**Limited Data**: The dataset used in the project may be limited in size and scope, which could impact the generalizability of the results. A larger and more diverse dataset could provide a more accurate representation of the project in different contexts. we could only gather 130 responses from our local peoples in limited then , therefore we were compelled to find a similar dataset from the internet and got the same from kaggle which is gathered from columbian people.

**Limited Model Performance**: While the gradient boosting classifier model used in the project achieved high accuracy, it is possible that other machine learning models could perform better on this task. Using a wider variety of models could improve the accuracy and robustness of the predictions.

**Limited Insights**: The project only provides insights into the relationships between the available input features and employee promotions. However, it may not provide insights into why these relationships exist or how they can be influenced. Further investigation and analysis may be necessary to understand the underlying causes of these relationships.

**Limited Applicability**: The project was focused on predicting the obesity type and rate in an adult , but mentioned parameters may be not enough for some cases. In other words it may or may not be applicable in some people or can be proven wrong in some cases.

#### Limitations for Future Work

**Limited Parameters**: Our obesity prediction model relies on a set of parameters that

are currently known to influence obesity in adults. However, it's crucial to recognize

Our understanding of obesity is continually evolving. New research and discoveries

about the multifaceted nature of obesity may reveal additional factors that our current

model does not account for. Therefore, while our model may be accurate based on

existing knowledge, it could be prone to inaccuracies as our understanding of obesity

expands.

**Disease Interactions:** Obesity often interacts with various diseases and medical

conditions in complex ways. While our model considers some of these interactions

based on available data, there may be other disease-obesity relationships that are not

fully captured by our current knowledge. As a result, the model's predictions may not

accurately reflect cases where obesity is influenced by coexisting medical conditions

that are not accounted for in our dataset.

**Population Specificity:** Our model predominantly relies on data collected from a

large population of Colombian citizens to make predictions about obesity. However,

the factors contributing to obesity can vary significantly across different populations

and ethnic groups.Therefore, the applicability of our model to other populations, such

as Indian citizens, may be limited. Cultural differences, dietary habits, genetic

predispositions, and environmental factors unique to specific populations may not be

adequately represented in our training data, leading to potential inaccuracies in

predictions for populations outside the dataset's scope.

# CHAPTER 8 REFERENCES

### REFERENCES

#### Research Papers

# [1] “Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico” by Fabio Mendoza Palechor and Alexis de la Hoz Manotas , University of Columbia.

# [2] Charaka Samhita Vedic Book , the core text of ancient Indian medical system - [Ayurveda](https://www.carakasamhitaonline.com/index.php?title=Ayurveda).

#### Websites

[1] <https://www.kaggle.com/>

Kaggle is a platform for data science competitions and projects.

[2] <https://github.com/>

GitHub is a platform for hosting and sharing code.

[3] <https://www.analyticsvidhya.com/>

Analytics Vidhya is a community of data science enthusiasts and professionals.

[4] <https://machinelearningmastery.com/>

Machine Learning Mastery is a website that offers tutorials and resources on machine learning algorithms.

<https://pandas.pydata.org/>

<https://numpy.org/>

<https://matplotlib.org/>

<https://seaborn.pydata.org/>

<https://scikit-learn.org/>

<https://imbalanced-learn.org/>

# USER MANUAL

## USER MANUAL

Here is a user manual for running theDaily Dharma prediction project on your local machine using Streamlit.

#### Prerequisites:

Python 3.7 or higher Pip package installer Git (optional)

#### Steps:

**Step 1:** Clone the project repository from GitHub to your local machine using git clone or download the repository as a zip file and extract it to a folder.

**Step 2:** Open a command prompt or terminal in the project folder.

**Step 3**: Install the required packages by running the command: pip install -r requirements.txt

**Step 4:** Start the Streamlit app by running the command: streamlit run app.py

**Step 5:** Wait for the Streamlit app to start up, and you should see a web page with the DailyDharma: Obesity prediction app.

**Step 6:** To use the app, simply fill the questions asked with full honesty and

Once you have selected the input values, click the "Predict" button to see the predicted obesity type outcome.

The app will display the predicted obesity level outcome.

To exit the app, simply close the web page or press "Ctrl + C" in the command prompt or terminal.

Note: The project is designed to run locally on your machine, so you can modify the code and input data as needed. If you encounter any issues while running the app, please refer to the error messages in the command prompt or terminal .

**Fig. 17: A Screenshot of the Streamlit web app**