**Estimating Presence Or Absence Of Smoking Through**

**Bio Signals**

**1. INTRODUCTION**

1..OVERVIEW

The main aim or goal of a machine learning project estimating the presence or absence of smoking through bio-signals is to develop an accurate and reliable system that can detect smoking behavior using physiological signals.

The system would use machine learning algorithms to analyze bio-signals, such as heart rate, skin conductance, and respiratory rate, to determine if a person is smoking or not. The system could be applied in various contexts, such as clinical settings or workplaces, to monitor smoking behavior and help individuals quit smoking.

This project focuses on developing a system that can detect the presence or absence of smoking by analyzing bio signals. Bio signals encompass physiological data such as heart rate, respiratory rate, carbon monoxide levels, skin temperature, and more. By collecting and analyzing these bio signals, the project aims to create a model that accurately identifies smoking behavior. The application of this technology may extend to health monitoring and regulatory compliance.

**2.**PURPOSE

The primary purpose of this project is to develop a system for detecting smoking behavior through bio signals. It serves the goals of health monitoring, addiction assessment, and regulatory compliance in environments where smoking is prohibited.

         Here are some of the ways it can be utilized:

* Health Monitoring
* Addiction Assessment
* Regulatory Compliance
* Public Health Research
* Privacy and Informed Consent
* Innovation and Technology Development
* Educational and Awareness Purposes

**2. LITERATURE**  **SURVEY**           2.1.EXISTING PROBLEM

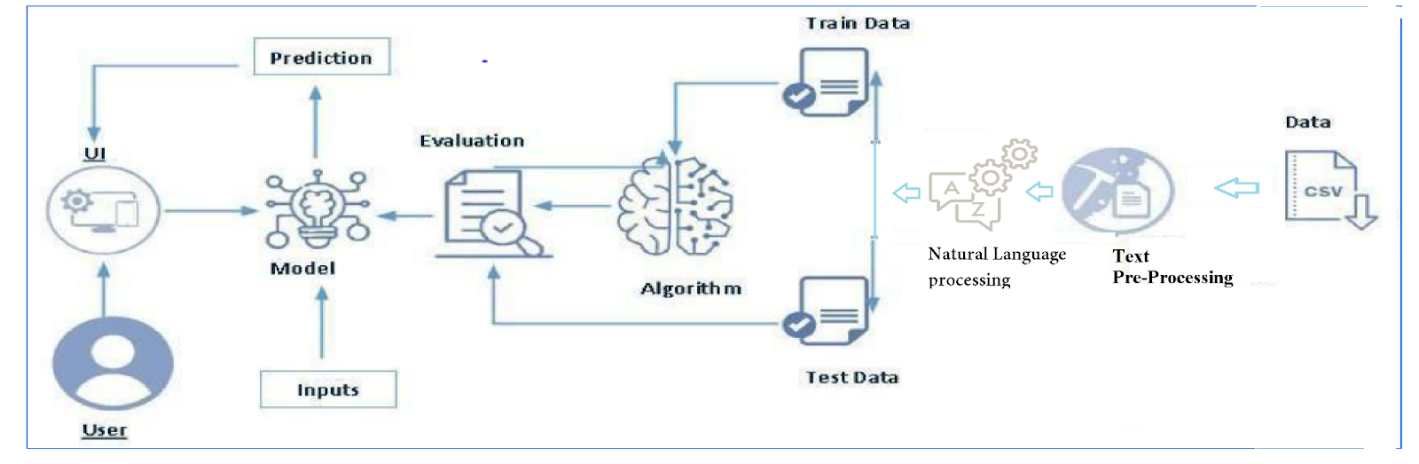
               The existing problem this project aims to address is the challenge of accurately detecting and monitoring smoking behavior in individuals. Traditional methods, such as self-reporting and visual observation, are often unreliable and subject to bias. This can hinder healthcare professionals' ability to assess and manage smoking-related health risks, impede addiction counseling efforts, and make it difficult to enforce no-smoking regulations in certain settings. The project seeks to leverage bio signals as a more objective and efficient means of detecting smoking, offering a solution to these existing challenges in a non-intrusive and data-driven manner.

2.2.PROPOSED SOLUTION

              The proposed solution is to develop a system that uses bio signals, including physiological data like heart rate, respiratory rate, skin temperature, and carbon monoxide levels in breath, to accurately estimate the presence or absence of smoking. This system will employ machine learning or deep learning models to analyze and interpret these signals, offering a reliable and objective means of detecting smoking behavior for health monitoring, addiction assessment, and regulatory compliance purposes.

**3**. **THEORITICAL ANALYSIS** 

      3.1.BLOCK DIAGRAM



3.2 HARDWARE / SOFTWARE DESIGNING

        To complete this project, you must required the following software’s, concepts and packages

1. Visual studio
2. Python packages:

* Open Jupyter in vs code then install packages
* Type “pip install numpy” and click enter
* Type “pip install pandas” and click enter.
* Type “pip install seaborn” and click enter.
* Type “pip install matplotlib” and click enter.
* Type “pip install pickle” and click enter.
* Type “pip install Flask” and click enter.

**4.EXPERIMENTAL INVESTIGATIONS**

The experimental investigations in this project will involve collecting bio signal data from a diverse group of individuals, including both smokers and non-smokers, in controlled settings. These data will be subjected to rigorous preprocessing and feature extraction techniques. Machine learning models will then be trained and evaluated on this dataset to assess their accuracy in detecting smoking behavior based on the bio signals. The experiments will aim to fine-tune the models, identify key features, and determine the system's real-world applicability, ultimately validating its effectiveness in estimating the presence or absence of smoking.

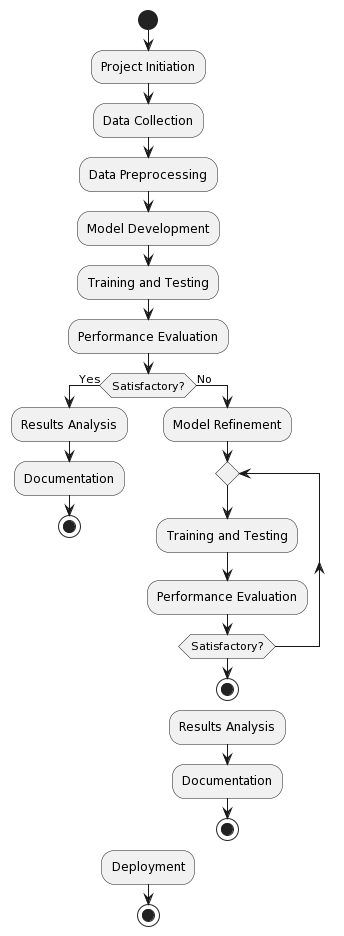
**Project Flow:**

* User interacts with the UI to enter the input.
* Entered input is analyzed by the model which is integrated.
* Once the model analyses the input the prediction is showcased on the UI

To accomplish this, we have to complete all the activities listed below,

* Define Problem / Problem Understanding
  + Specify the business problem
  + Business requirements
  + Literature Survey
  + Social or Business Impact.
* Data Collection & Preparation
  + Collect the dataset
  + Data Preparation
* Exploratory Data Analysis
  + Descriptive statistical
  + Visual Analysis
* Model Building
  + Training the model in multiple algorithms
  + Testing the model
* Performance Testing
  + Testing model with multiple evaluation metrics
* Model Deployment
  + Save the best model
  + Integrate with Web Framework

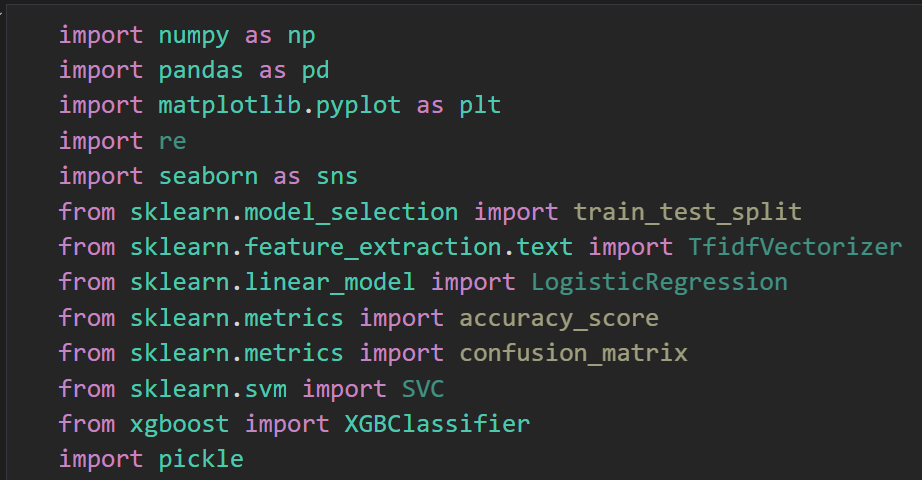
**5. FLOWCHART**



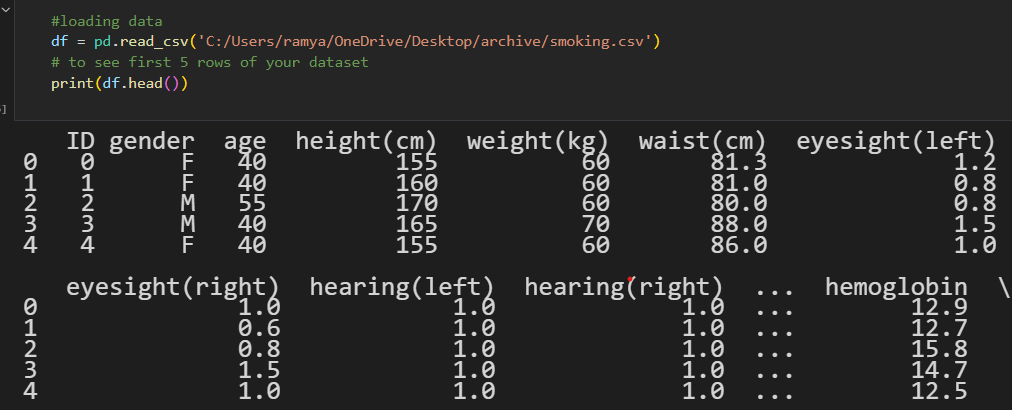
**6. RESULT**

**Activity 1: Importing the libraries** 

Import the necessary libraries as shown in the image.



**Activity 2: Read the Dataset** 

**Activity 3:Data Preparation**

Data preparation, also known as data preprocessing, is a crucial step in the data analysis and machine learning lifecycle. It involves cleaning, transforming, and organizing raw data into a format that is suitable for analysis or model training. Properly prepared data is essential for achieving accurate and meaningful insights or building robust machine learning models.As we have understood how the data is, let's pre-process the collected data.

The download data set is not suitable for training the machine learning model as it might have so much randomness so we need to clean the dataset properly in order to fetch good results.

This activity includes the following steps.

* Handle missing or null values: Impute missing values using techniques like mean, median, or interpolation.
* Remove duplicate records: Eliminate duplicate rows to avoid skewing analysis or models.
* Outlier detection and treatment: Identify and handle outliers that can distort the analysis or models.
* Handling categorical data

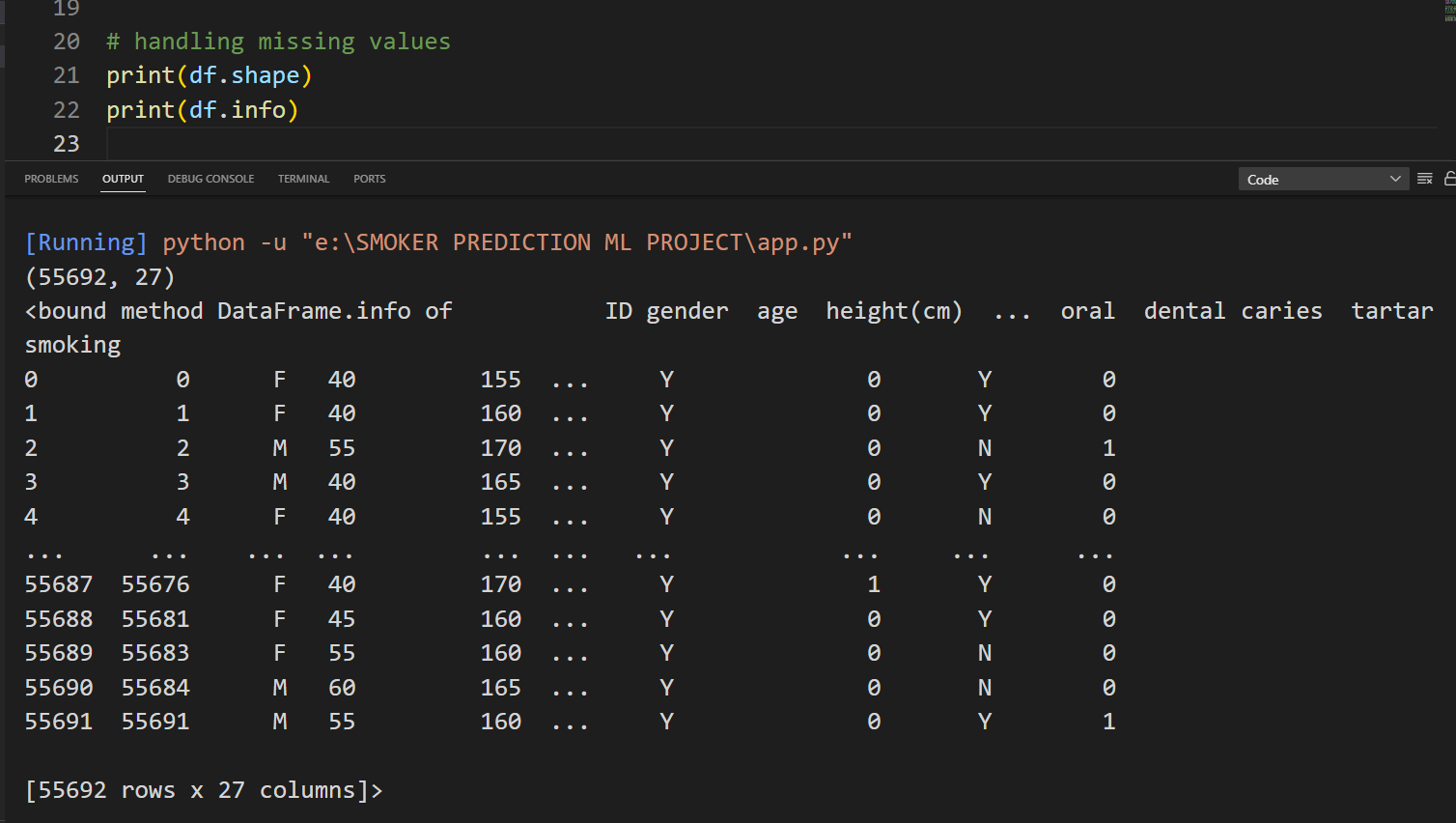
Step.1: **Handling missing values** :

The “df.shape” attribute in Pandas is used to retrieve the dimensions of a DataFrame. It returns a tuple representing the number of rows and columns in the DataFrame.

The df.info() method in Pandas provides a concise summary of a DataFrame, including essential information about the structure, data types, and presence of missing values. It's a valuable tool for quickly understanding the composition of a dataset.

Let’s find the shape of our dataset first. To find the shape of our data, the df.shape method is used. To find the data type, df.info() function is used

In below image we can find the shape and info of our dataset



### For checking the null values, df.isnull() function is used. To sum those null values we use .sum() function.

### df.isnull(): This function returns a DataFrame of Boolean values indicating whether each element is NaN (null) or not.

### df.isnull().sum(): This calculates the sum of null values for each column. It effectively counts the number of True values for each column, as True is considered as 1 and False is considered as 0 in Python. Therefore, by summing these Boolean values, you get the count of null values for each column.

### If .sum() for a particular column is zero, it means there are no null values in that column.

### df.isnull().sum(). If the sum for a column is zero, it means there are no null values in that column. If are\_there\_null\_values is False, it indicates there are no null values in the entire DataFrame.

### From the below image we found that there are no null values present in our datasetSo we can skip handling the missing values step.

### 

### 

### Step.2: Handling Categorical Values:

### As we can see our dataset has categorical data we must convert the categorical data to integer encoding or binary encoding.

### To convert the categorical features into numerical features we use encoding techniques.

### Label Encoding: Assign a unique integer to each category. This is suitable for ordinal categorical data, where there is an inherent order.

### One-Hot Encoding: Create binary columns for each category, indicating its presence (1) or absence (0). This is suitable for nominal categorical data where there is no order.

### Ordinal Encoding: Assign numerical values according to the meaningful order of the categories. This is suitable for ordinal categorical data.

### Frequency Encoding: Replace categories with their frequency of occurrence in the dataset. This can be useful for both ordinal and nominal data.

### There are several techniques but in our project we are using manual encoding with the help of list comprehension.

### We can see that the conversion of categorical values into numerical in the below image they are successfully mapped.

### 

### Step.3: Handling Outliers:

### Identifying and handling outliers is an important step in data preprocessing to ensure that anomalies or extreme values in the data do not unduly influence analysis or models. Outliers can skew statistical measures and affect the performance of machine learning algorithms.we have different ways to handle the outliers

### 

### Here we are using Boxplot to handle the Outliers see the below image we got outliers in age column so we we need to replace them with maximum whisker point to remove outliers :

### we can see that in below image the outliers are successfully replaced with maximum whisker point like this we need to replace all the outliers present in each attribute in our dataset.

### 

### Activity 4:Descriptive Statistical:

### Descriptive analysis is to study the basic features of data with the statistical process. Here pandas has a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.

### In below image we can see that clear statistical description about each attribute in dataset like mean, median, sum etc…

### 

### Visual Analysis :

### In the context of a project or analysis, "visual analysis" refers to the process of using visualizations or graphical representations to understand, explore, and gain insights from data. It involves creating charts, graphs, plots, and other visual aids to present the data in a clear and easily interpretable format.

### Visual analysis is a crucial part of the data analysis process because it allows for a more intuitive understanding of complex data patterns, relationships, trends, and anomalies. It helps in identifying patterns that might not be immediately apparent in the raw data, aiding in better decision-making and hypothesis generation.

### Univariate Analysis:

### Definition: Univariate analysis focuses on analyzing and describing a single variable at a time.

### 

### 

### Bivariate Analysis:

### Definition: Bivariate analysis involves the simultaneous analysis of two variables to determine if there is a relationship or correlation between them.

### 

### Multivariate Analysis:

### Definition: Multivariate analysis involves the analysis of three or more variables simultaneously to understand the relationships and patterns among them.

### 

### Outliers Handling:

### At the time of data preprocessing we done outliers handling but there we only handled the one attribute of dataset so, to handle all attributes at once look at below image

### 

### 

### Here we have outliers present in different attributes so, to replace them with max and min whisker point look at below image

### 

### Separating Data For Traning And Testing:

### Separating data into training and testing sets is a fundamental step in machine learning. This division allows you to train your model on one portion of the data (the training set) and evaluate its performance on another, unseen portion (the testing set). This helps assess how well your model generalizes to new, unseen data.

### 

### Scaling the values:

### Scaling is a crucial preprocessing step in machine learning, especially for models sensitive to the scale of the input features. It ensures that all features contribute equally to the learning process, preventing features with large scales from dominating the learning algorithm.

### Common scaling techniques include Min-Max Scaling and Standardization (Z-score normalization). Let's discuss how to perform these scaling techniques using Python and scikit-learn.

### 

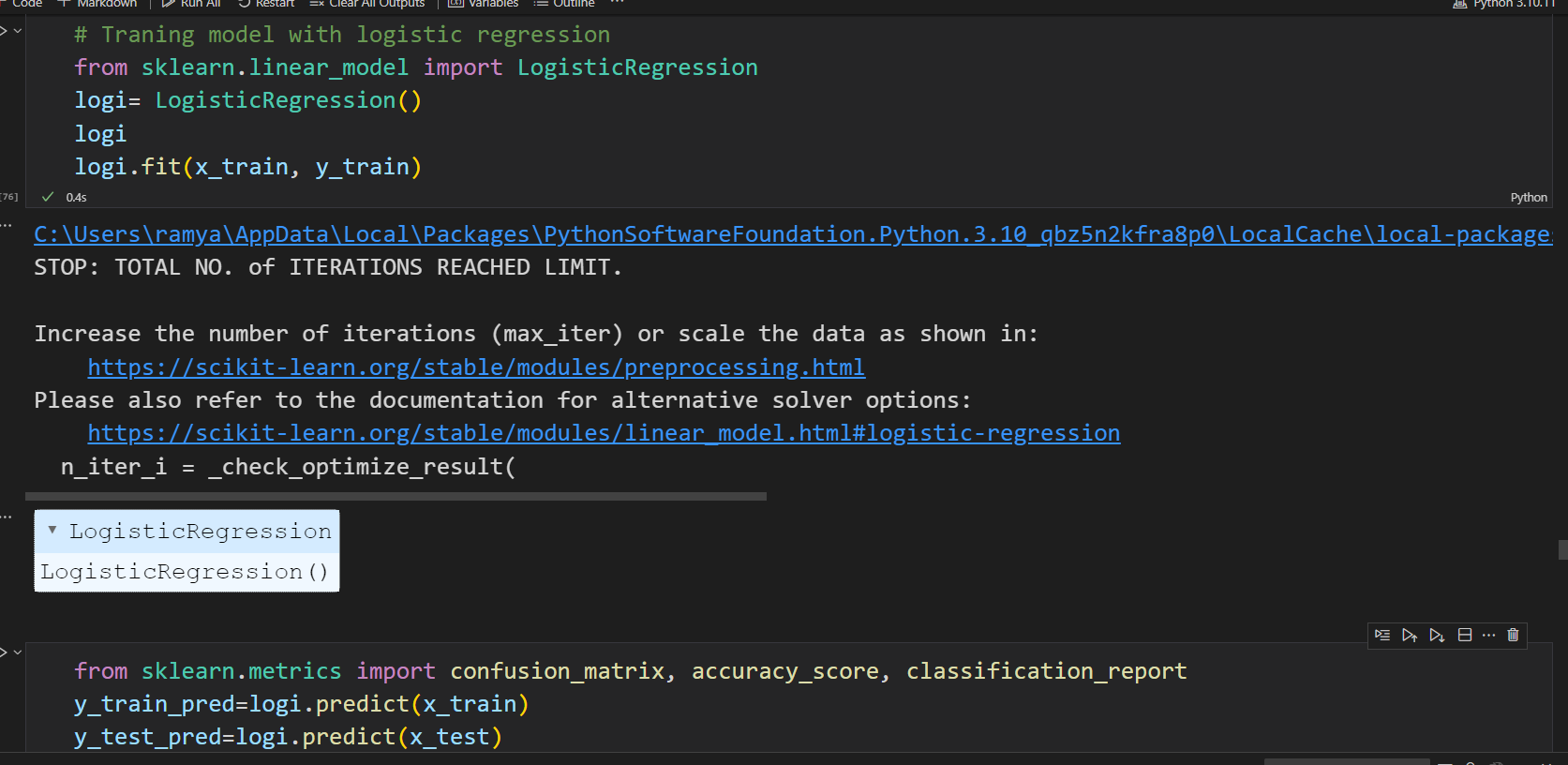
### Activity 5:Model Building:

### Training The Model In Multiple Algorithms:

### Training a model using multiple algorithms involves applying various machine learning algorithms to the training data and evaluating their performance to choose the best-performing model. This process helps you select the algorithm that best fits your specific problem and dataset.

**1: Logistic Regression model**

Logistic Regression is a popular statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. It's used for binary classification problems, where the goal is to predict one of two possible outcomes (e.g., yes/no, 0/1, true/false). Logistic Regression is widely used in machine learning for tasks such as spam detection, disease diagnosis, and credit risk prediction.



### 

**2: Random Forest :**

### Random Forest is an ensemble learning method that combines multiple decision trees to create a powerful and robust classification model. It's known for its ability to handle high-dimensional data and noisy or unstructured datasets.

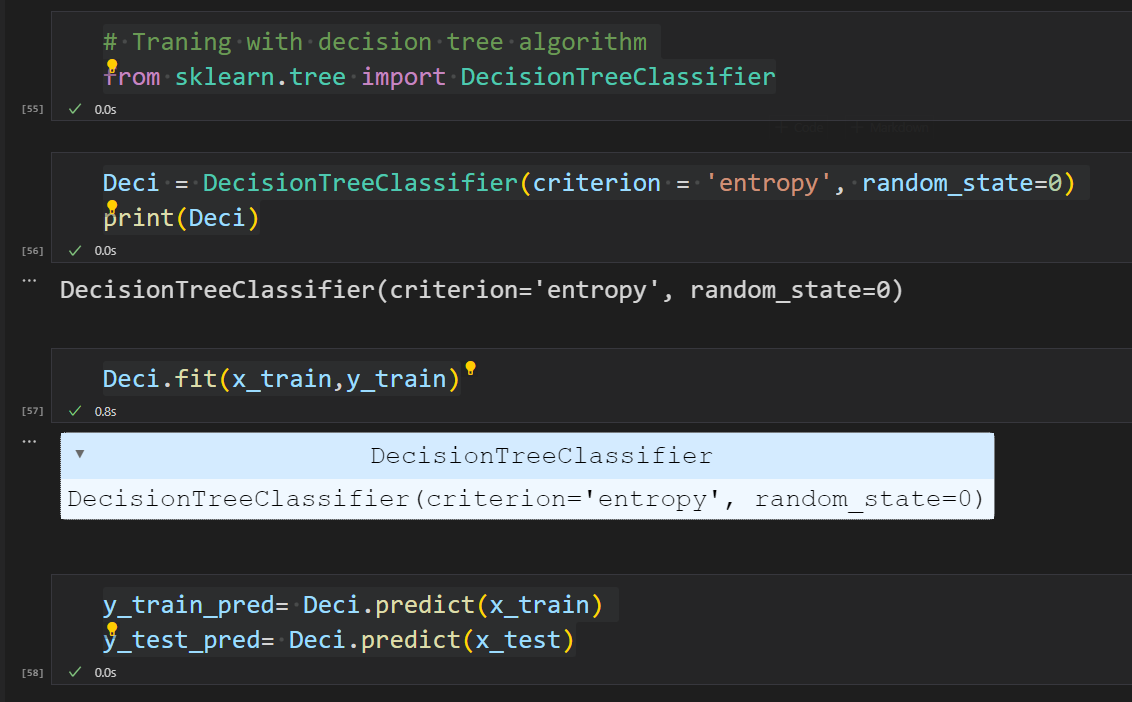
### 

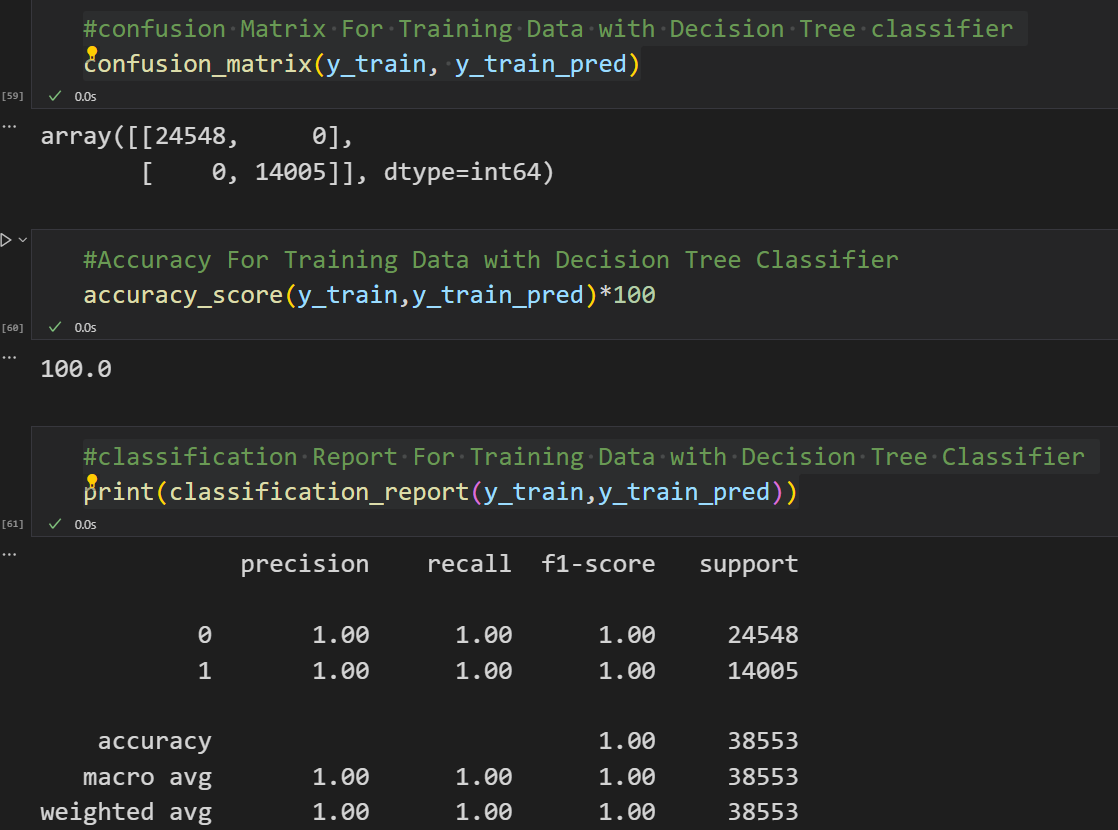
### 

### 

**3: Decision Tree :**

A Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks. It's a popular and intuitive model that mimics human decision-making. In a decision tree, the data is split into smaller subsets based on the most significant attribute at each node, leading to a tree-like structure with nodes and leaves.

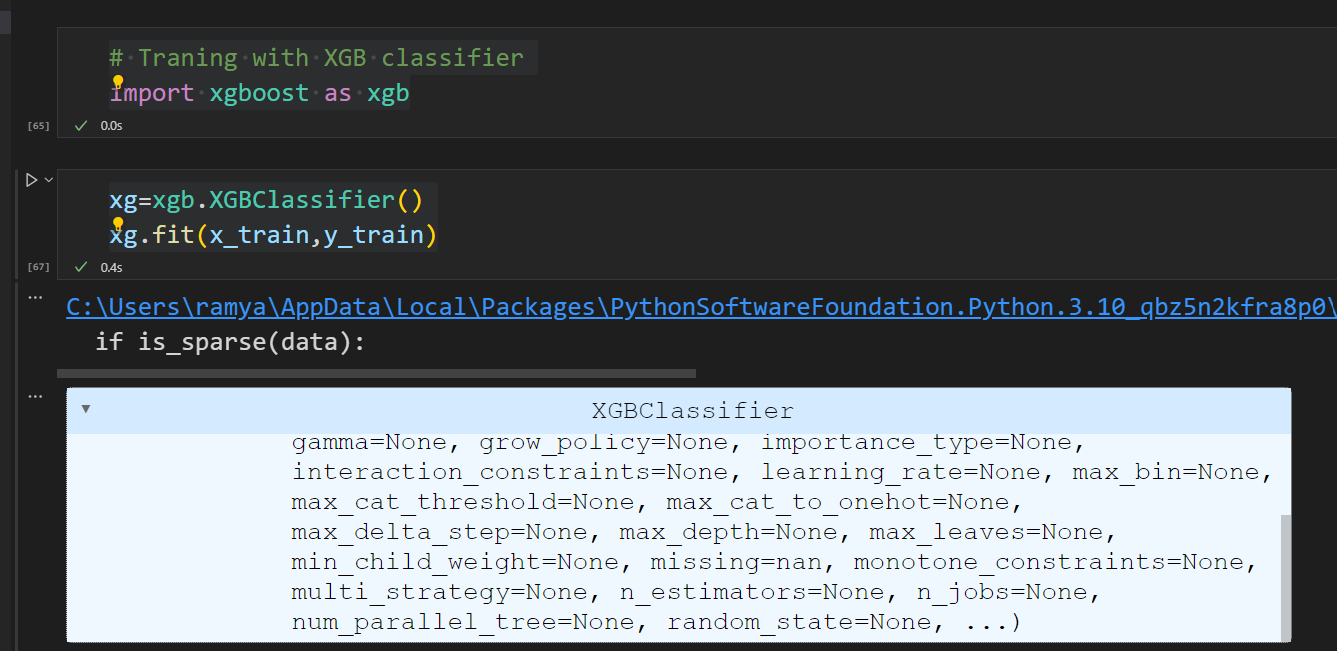




### 

**4: XG Boost Classifier :**

XGBoost (Extreme Gradient Boosting) is a popular machine learning algorithm known for its efficiency and effectiveness in a wide range of tasks. XGBoost can be used for both classification and regression problems and often performs well in data competitions and real-world applications.



### 

### 

### 

### Activity6.Performance Testing & Hyperparameter Tuning:

### Testing Model: "Testing the model" refers to the process of evaluating how well a machine learning model, which has been trained on a dataset, performs on a separate dataset that it has never seen before.

### 

### Hyperparameter Tuning for Random Forest Classifier:

### Hyperparameter tuning for a Random Forest Classifier involves the process of finding the optimal hyperparameter settings that result in the best performance of the classifier. The Random Forest algorithm has several hyperparameters that can be adjusted to fine-tune the model. The goal of hyperparameter tuning is to optimize the model's performance by systematically searching for the most suitable hyperparameter values.

### 

### 

### 

### Activity7.Model Deployment:

### Save The Best Model :

### 

### Activity8 .Integrate With Web Framework:

In this section, we will be building a web application that is integrated to the model we built. A UI is provided for the uses where he has to enter the values for predictions. The enter values are given to the saved model and prediction is showcased on the UI.

This section has the following tasks

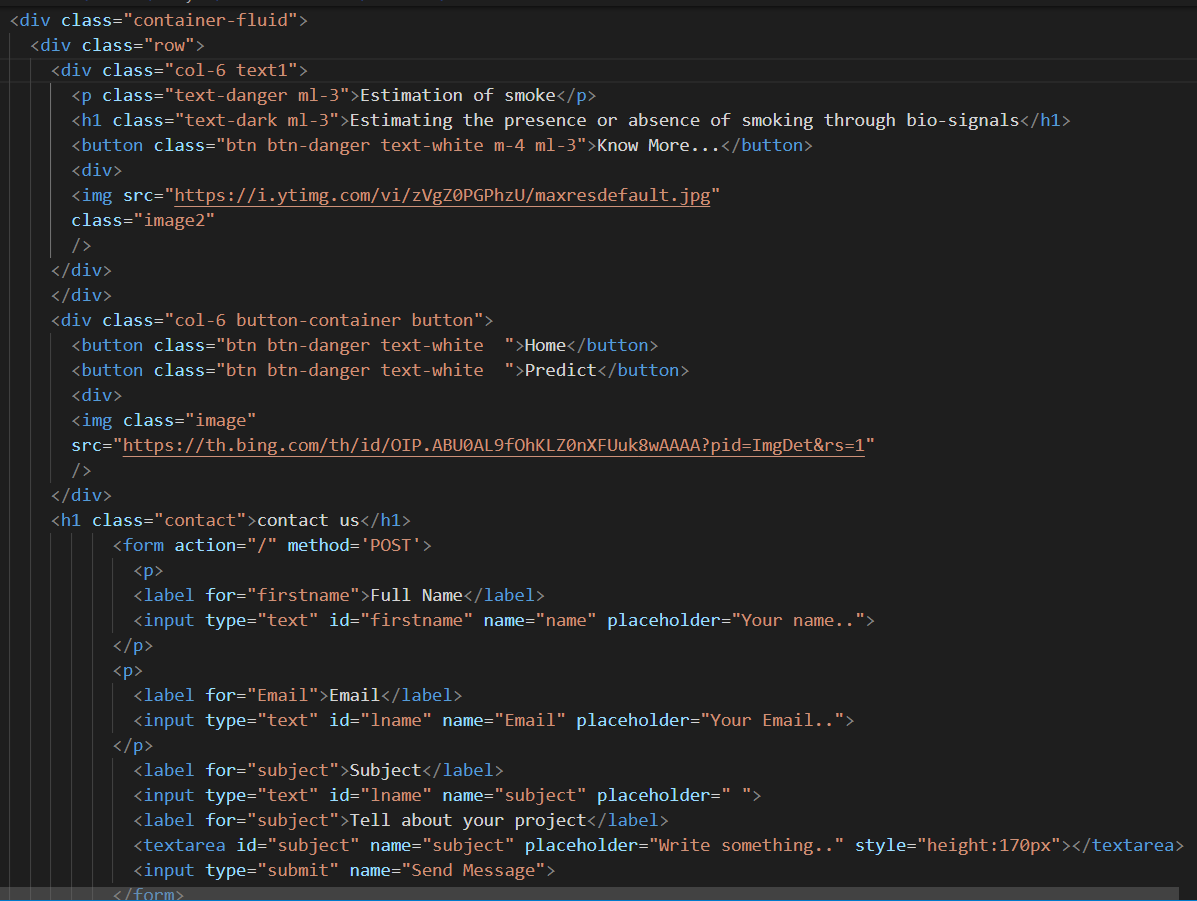
* Building HTML Pages
* Building server side script

### Building HTML Pages:

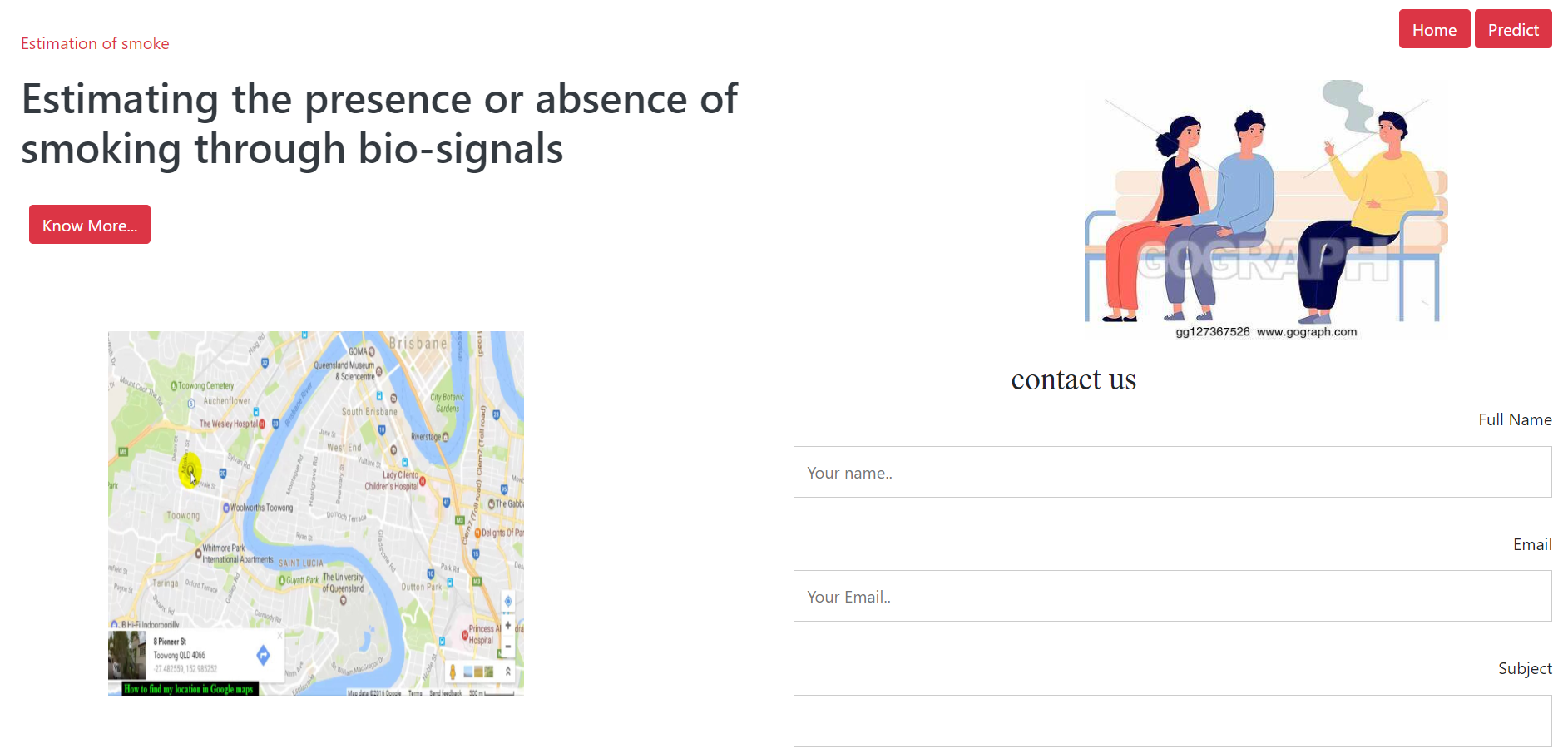
### 1.index page:

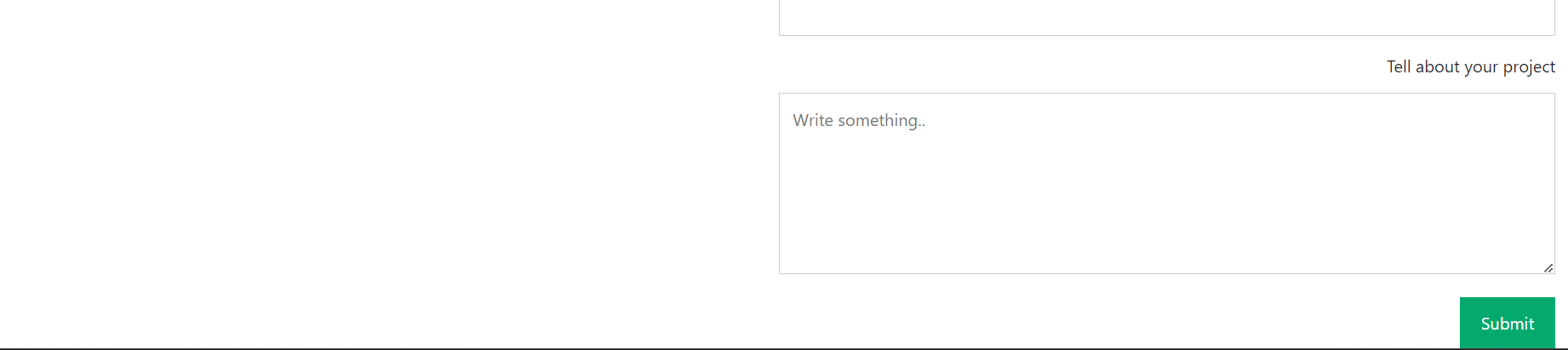
### This page consists of basic information about user and it is home page that it is going to be displayed first

### Input:



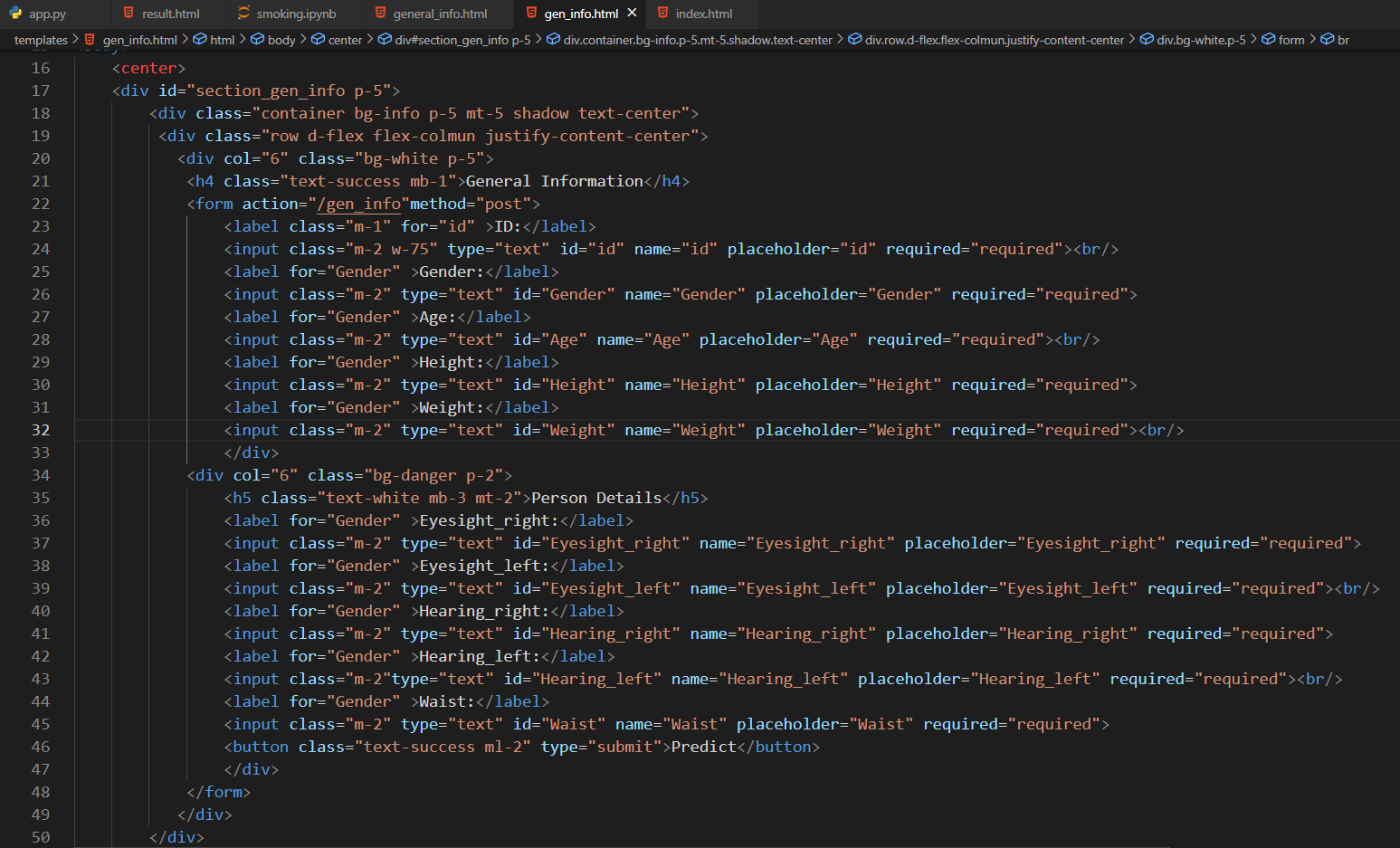
Output:



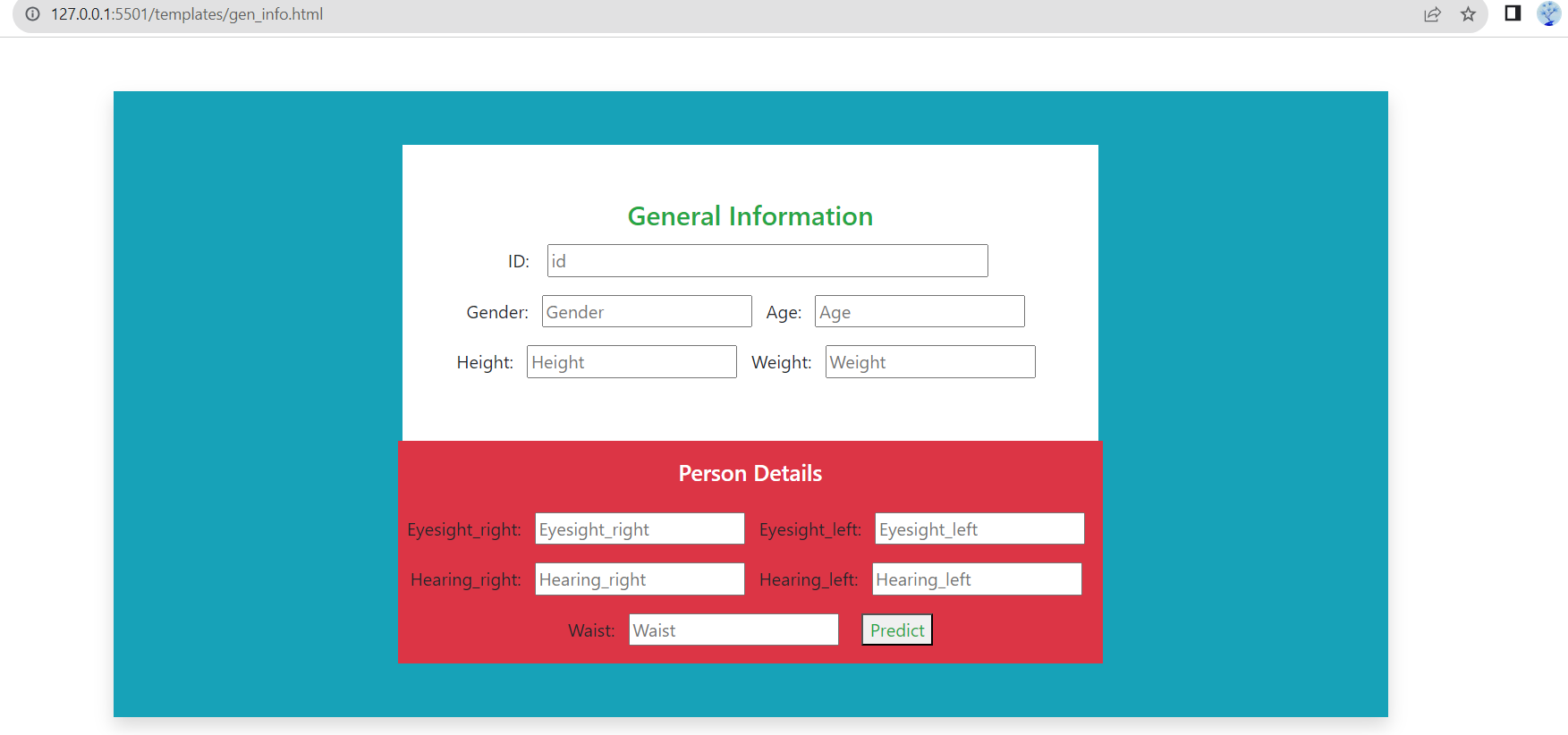


2.gen\_ingo page:

Input:

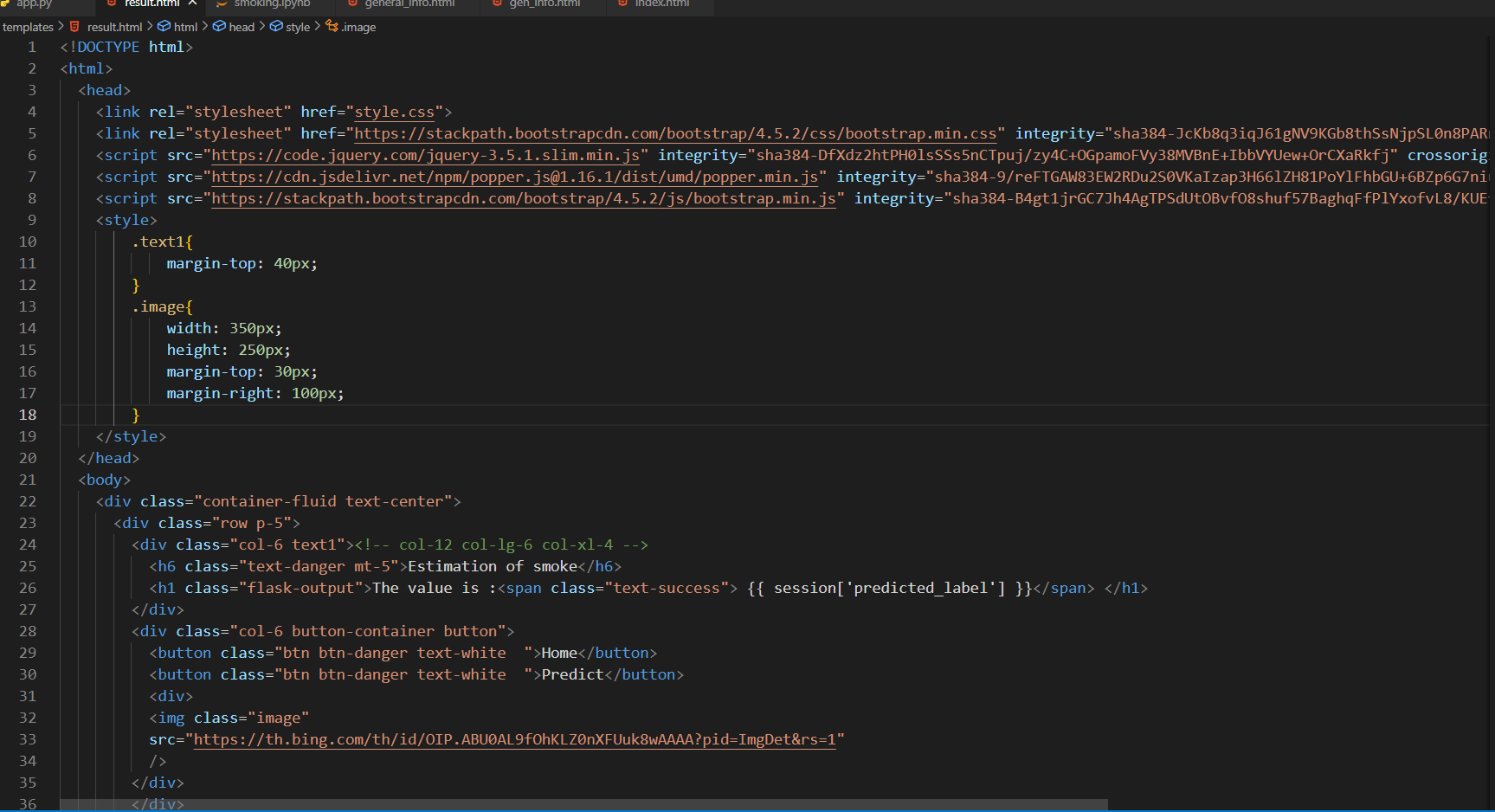


Output:

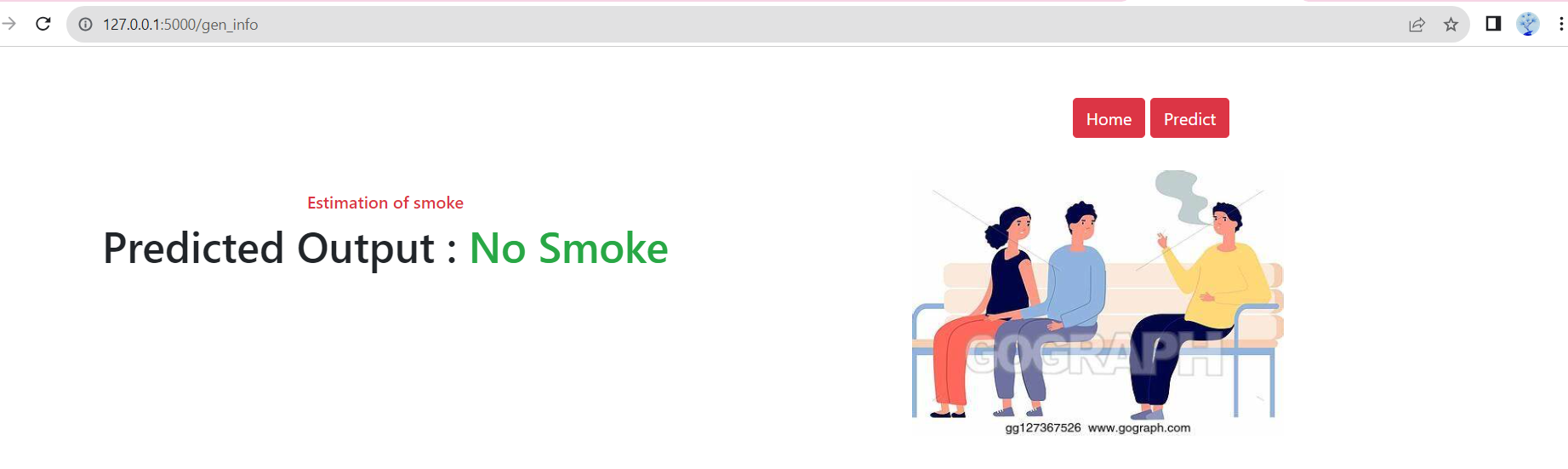


3.Result page:

Input:



Output:



**7. ADVANTAGES & DISADVANTAGE**

**Advantages**: The project to estimate smoking presence or absence through bio signals offers several advantages:

1. \*\*Objective Monitoring:\*\* It provides an objective and non-intrusive method to monitor smoking behavior, reducing the reliance on self-reporting, which can be unreliable.

2. \*\*Healthcare Benefits:\*\* The technology can assist healthcare professionals in better understanding and managing smoking-related health risks, aiding in early intervention and treatment.

3. \*\*Addiction Management:\*\* It supports addiction assessment and counseling efforts by providing real-time data on smoking habits, aiding individuals in quitting or reducing smoking.

4. \*\*Regulatory Compliance:\*\* It helps enforce no-smoking regulations in various settings, promoting a healthier and smoke-free environment.

5. \*\*Privacy and Informed Consent:\*\* It underscores the importance of privacy and informed consent when collecting and analyzing bio signal data, raising awareness about ethical considerations.

**Disadvantages**: While the project to estimate smoking presence or absence through bio signals offers several advantages, there are also potential disadvantages and challenges to consider:

1. \*\*Privacy Concerns:\*\* The collection and analysis of bio signals raise significant privacy concerns, as it involves sensitive physiological data. Proper consent and data protection measures are essential.

2. \*\*Data Accuracy:\*\* The accuracy of smoking detection may be influenced by factors like the individual's baseline physiological variations, making false positives and false negatives possible.

3. \*\*Data Volume and Storage:\*\* Managing and storing the substantial amount of data generated by bio signal monitoring can be resource-intensive.

4. \*\*Cost and Accessibility:\*\* Access to the required sensors and technology may be limited in certain regions, impacting the project's widespread implementation.

5. \*\*Real-World Variability:\*\* Real-world conditions can introduce variability in bio signal data, making it challenging to achieve consistently accurate results.

**8.APPLICATIONS**

The project to estimate smoking presence or absence through bio signals has several potential applications across various domains, including:

1. \*\*Healthcare Monitoring:\*\* It can be used in healthcare settings to monitor and assess patients' smoking behavior and its impact on their health. This data can aid healthcare professionals in providing personalized care and early intervention.

2. \*\*Addiction Management:\*\* The technology can assist addiction counselors and individuals seeking to quit smoking by providing real-time feedback on smoking habits, thereby facilitating addiction management and treatment programs.

3. \*\*Public Health Research:\*\* Bio signal data can be used for public health research to gain insights into the prevalence of smoking and its association with various health parameters. This information can inform public health policies and campaigns.

4. \*\*Workplace and Campus Compliance:\*\* It can be applied in workplaces, educational institutions, and public spaces to enforce no-smoking regulations and promote a smoke-free environment.

5. \*\*Fitness and Wellness Tracking:\*\* Individuals interested in tracking their wellness and fitness can use this technology to monitor the impact of smoking on their overall health and well-being.

6. \*\*IoT and Smart Devices:\*\* The system can be integrated into Internet of Things (IoT) and smart health devices for continuous monitoring, providing users with real-time information about their smoking habits.

**9.CONCLUSION**

In conclusion, the project focused on estimating smoking presence or absence through bio signals offers a promising solution to a range of challenges associated with smoking monitoring and addiction management. By leveraging physiological data like heart rate, respiratory rate, and carbon monoxide levels, this technology can provide objective and real-time insights into an individual's smoking behavior.

The potential applications span healthcare monitoring, addiction management, public health research, compliance enforcement, and wellness tracking, making it a versatile tool with significant societal benefits. However, it is crucial to address privacy concerns, sensor reliability, and ethical considerations when implementing this technology. As an innovative and data-driven approach, it has the potential to improve healthcare outcomes, contribute to public health research, and enhance the overall well-being of individuals by promoting healthier lifestyles and aiding in smoking cessation efforts.

**10.FUTURE SCOPE**

The future scope of the project on estimating smoking presence or absence through bio signals is promising and may involve several developments and opportunities

1. \*\*Advanced Sensors:\*\* Advancements in sensor technology may lead to more accurate and non-intrusive ways of collecting bio signals, improving the reliability of smoking detection.

2. \*\*Machine Learning Advancements:\*\* Ongoing developments in machine learning and artificial intelligence can lead to more robust models for smoking detection, reducing false positives and false negatives.

3. \*\*Personalized Health Solutions:\*\* The technology can evolve to offer personalized feedback and recommendations for individuals attempting to quit smoking or improve their health, tailoring interventions to individual needs.

4. \*\*Wearable Health Devices:\*\* Integration into wearable devices and smartphones can provide continuous health monitoring, making it easier for users to track their smoking habits.

**11.BIBILOGRAPHY**

* www.google.com
* www.wikipedia.org
* https://ieeexplore.ieee.org

**APPENDIX**

SOURCE CODE OF FLASK :

https://github.com/Ramyaerumalla/smoking