NAME: VISWANATHA REDDY KAMIREDDYGARI

ESTIMATION OF OBESITY LEVELS BASED ON EATING HABITS AND PHYSICAL CONDITION'S.

STAKEHOLDER:

The **University of California, Irvine (UCI)** is the stakeholder of this dataset, which is part of the renowned **UCI Machine Learning Repository**. The dataset is designed for the estimation of obesity levels in individuals from Mexico, Peru, and Colombia, using data about eating habits and physical conditions.

PROBLEM STATEMENT:

Obesity is a growing public health concern worldwide, significantly affecting individuals' health and well-being. This dataset aims to facilitate the estimation and classification of obesity levels in individuals from Mexico, Peru, and Colombia, based on their eating habits and physical attributes. With 17 features and 2111 records, this dataset categorizes individuals into seven obesity levels: Insufficient Weight, Normal Weight, Overweight Levels I & II, and Obesity Types I, II, & III. A substantial portion of the data (77%) was synthetically generated using the Weka tool and SMOTE filter, while the remaining 23% was collected through a web platform.

The objective is to develop and evaluate machine learning models that can accurately classify individuals into these obesity levels. Such models can be used to identify at-risk populations and guide targeted interventions to address obesity and promote healthier lifestyles.

AIM:

The aim of this project is to develop a machine learning-based classification model to accurately predict the obesity levels of individuals based on their eating habits, physical attributes, and lifestyle factors. By leveraging the dataset from the University of California, Irvine, the project seeks to identify patterns and relationships within the data to classify individuals into seven obesity levels. This classification system will enable the identification of at-risk populations, facilitate targeted health interventions, and contribute to addressing the global public health challenge posed by obesity.

DATA SET:

This dataset, provided by the University of California, Irvine (UCI), focuses on estimating obesity levels among individuals from three countries: Mexico, Peru, and Colombia

DATA SET LINK:
https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition.

The data set contains 17 features and 2111 records

KEY FEATURES OF THE DATA SET:

1. **Size and Structure:**
    - **Records:** 2111 data points.
    - **Attributes:** 17 features describing individuals' characteristics, habits, and health status.
    - **Target Variable:** NObesity (Obesity Level) with 7 categories:
        - Insufficient Weight
        - Normal Weight
        - Overweight Level I
        - Overweight Level II
        - Obesity Type I
        - Obesity Type II
        - Obesity Type III
2. **Attributes:**
    - **Categorical Features:**
        - Gender: Male or Female.
        - Family history with overweight: Indicates whether the individual has a family member who has suffered or suffers from overweight FAVC: Frequency of consuming high-calorie food.
        - CAEC: Describes the individual's habit of eating between meals (e.g., "No", "Sometimes", "Frequently", "Always").
        - SMOKE: Whether the individual smokes.
        - SCC: Indicates whether the individual monitors their daily calorie intake (Yes/No).
        - CALC: Alcohol consumption frequency.

        - MTRANS: Mode of transportation (e.g., Walking, Public Transportation, Automobile, etc.).
    - **Numerical Features:**
        - Age: Age in years.
        - Height: Height in meters.
        - Weight: Weight in kilograms.
        - FCVC: Frequency of vegetable consumption (1–3 scale).
        - NCP: Number of main meals per day.
        - CH2O: Daily water consumption in liters.
        - FAF: Physical activity frequency in hours per week.
        - TUE: Time spent using technology in hours per day.
3. **Target Variable (NObesity):**
    - This is the class label used for classification and includes the following levels:
        - Insufficient Weight
        - Normal Weight
        - Overweight Level I
        - Overweight Level II
        - Obesity Type I
        - Obesity Type II
        - Obesity Type III

MODELS WHICH I CHOOSEN AND WHY :

For this project, I explored multiple machine learning models to classify obesity levels, focusing on a diverse range of algorithms to ensure comprehensive evaluation. We started with **Logistic Regression**, a simple and interpretable model, to establish a baseline for multi-class classification. **Random Forest** was chosen for its robustness, ability to handle feature interactions, and strong performance on datasets with nonlinear relationships. Similarly, **Gradient Boosting** was selected for its iterative error-correcting nature, offering high accuracy and fine-tuning capabilities for challenging datasets. We included **Decision Trees** for their simplicity, interpretability, and ability to identify feature importance, while also serving as a

foundation for ensemble methods. Finally, we tried **Support Vector Machines (SVM)** for their effectiveness in handling complex decision boundaries and high-dimensional data. This selection of models allowed us to evaluate both linear and nonlinear approaches, ensuring a well-rounded analysis for the multi-class classification task of predicting obesity levels.

***FEATURES THAT I CHOOSEN**:*

For this project, I used all 17 features provided in the data set as they were highly relevant for predicting obesity levels. These included demographic attributes like Gender and Age, physical attributes like Height and Weight, and behavioral factors such as eating habits, physical activity, and smoking. Categorical features (e.g., Gender, family history with overweight) were encoded into numerical values using label encoding, and numerical features (e.g., Height, Weight, Age) were scaled using standardization to ensure compatibility with machine learning models. All features were retained as they provided valuable information, and preliminary model testing confirmed their importance in improving prediction accuracy. This careful selection and preprocessing of features helped ensure the models performed effectively.

***FEATURE ENGINEERING**:*

In this project, Feature engineering focused on encoding, scaling, and rounding the data to prepare it for machine learning models. Categorical variables such as Gender, family history with overweight, FAVC, and others were encoded using **Label Encoding** to convert them into numerical values suitable for the algorithms. Numerical features like Age, Height, Weight, FCVC, NCP, CH2O, FAF, and TUE were normalized using **Standard Scaler** to ensure they followed standard normal distribution, improving compatibility with models sensitive to feature magnitudes. Additionally, certain attributes (Age, Height, Weight, FCVC, and NCP) were rounded to their nearest integers or to a specified decimal precision to simplify their representation and better reflect their natural values. These transformations ensured the dataset was clean, standardized, and well-prepared for accurate and efficient model training.

***MODEL EVALUATION PROCESS AND METRICS USED**:*

The models were evaluated using a combination of **train and test splitting** the data to ensure robustness and generalizability.

The following evaluation metrics were used to assess the performance:

1) **Accuracy**: Accuracy measures the proportion of correctly classified instances out of the total instances. It provides a straightforward understanding of how well the model is performing overall. Accuracy was chosen because it offers a quick snapshot of the model's effectiveness, especially for balanced datasets where each class has a similar number of instances.

2) **Precision**: Precision calculates the proportion of true positive predictions out of all positive predictions made by the model. It was used to ensure that the model's positive predictions (such as specific obesity levels) are reliable. This metric is particularly important in scenarios where false positives (misclassifying a person as obese when they are not) could have significant implications.

3) **Recall (Sensitivity)**: Recall measures the proportion of true positive predictions out of all actual positive instances. It was included to evaluate how well the model captures all instances of a particular class, such as identifying individuals with a specific obesity level. This is critical in cases where missing true positives (false negatives) could lead to overlooked health risks.

4)**F1-Score**: The F1-Score is the harmonic mean of Precision and Recall, balancing the trade-off between the two. This metric was prioritized for model selection as it is particularly useful for multi-class classification and imbalanced datasets, ensuring the model performs well across all obesity levels.

Evaluation Process:

1. **Train-Test Split**:
    - The dataset was split into 80% training and 20% testing to evaluate the model on unseen data.
2. **Model Comparison**:
    - All metrics (Accuracy, Precision, Recall, F1-Score) were compared across different models, and the best model was selected based on the highest F1-Score, as it balances Precision and Recall.

Why These Metrics?

- **Multi-Class Classification**:
    - The dataset involves predicting one of seven obesity levels. Using Precision, Recall, and F1-Score for each class ensures that the model performs well across all categories.
- **Balanced Insights**:

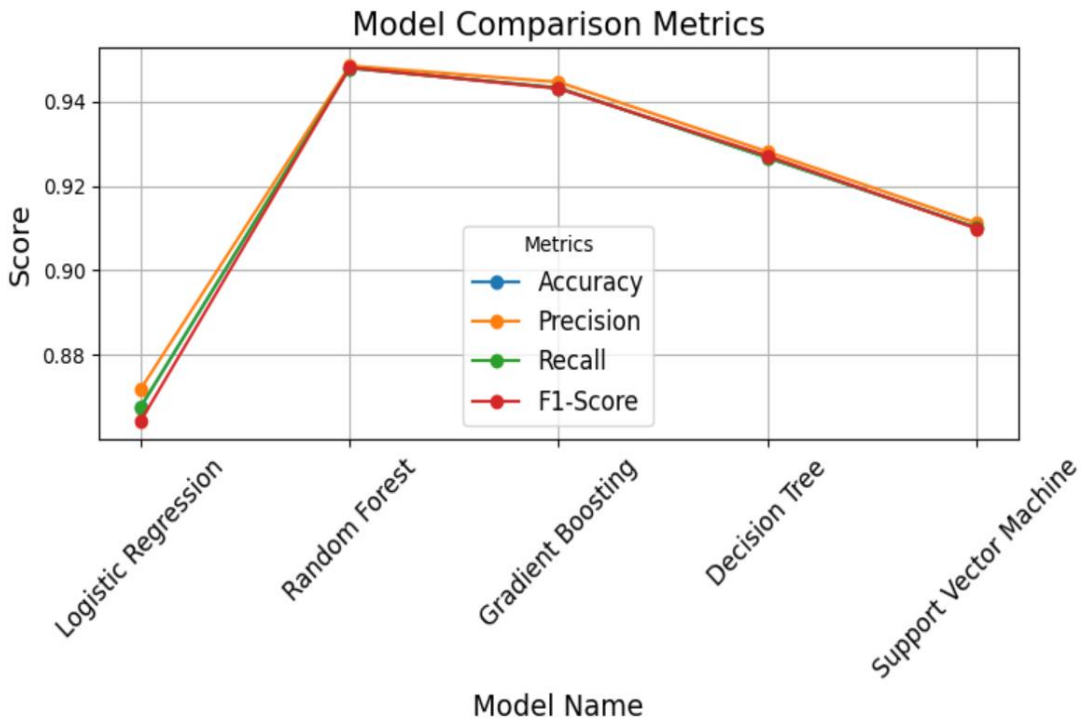
- Metrics like F1-Score handle the trade-off between Precision and Recall, providing a comprehensive measure of performance.

By combining these metrics and evaluation techniques, we ensured a thorough assessment of the model's strengths and weaknesses. This process guided us in selecting the best-performing model for the obesity prediction task.

The picture below gives information about the results which I got after evaluating the model using different machine learning algorithms.

```
Model Comparison:
```

	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.867612	0.871849	0.867612	0.864269
1	Random Forest	0.947991	0.948515	0.947991	0.948070
2	Gradient Boosting	0.943262	0.944635	0.943262	0.943096
3	Decision Tree	0.926714	0.928105	0.926714	0.927143
4	Support Vector Machine	0.910165	0.911221	0.910165	0.909929



Why Random Forest as the Final Model?

- **Reason for Selection:** Random Forest was chosen because it outperformed other models (Logistic Regression, Gradient Boosting, Decision Tree, and SVM) across all evaluation metrics, especially the F1-Score (0.948). The F1-Score is crucial for multi-class classification since it balances Precision and Recall, ensuring robust predictions across all obesity levels.
- **Pros:**
    - Handles both categorical and numerical features effectively.
    - Resistant to overfitting by averaging multiple decision trees.
    - Provides feature importance, helping to identify key predictors.
- **Cons:**
    - Computationally intensive compared to simpler models like Logistic Regression.
    - Less interpretable than single decision trees or linear models**.**

FUTURE WORK*:*

Given more time, I would focus on advanced hyperparameter tuning, exploring additional ensemble methods like XGBoost or LightGBM, and conducting deeper feature engineering to

uncover complex interactions. Expanding the dataset to include more diverse populations and improving class balance would enhance model generalizability.

RECOMMENDATION TO CLIENT:

Yes, I recommend the client use this model, particularly the Random Forest model, which demonstrated excellent performance across all metrics, including high Precision, Recall, and F1-Score. The F1-Score of 0.948 indicates a strong balance between Precision and Recall, ensuring the model is both reliable and accurate for predicting obesity levels. This makes it suitable for the intended use case of identifying at-risk individuals and supporting targeted health interventions. However, the client should be informed that the model's performance may vary slightly with different populations, and periodic retraining with updated data is advisable for sustained accuracy.

MODEL DEPLOYMENT USING STREAMLIT AND JOBLIB LIBRARAY:

The deployment process involved creating a web-based application using Streamlit to interact with the trained Random Forest model for obesity prediction. Here's a step-by-step breakdown:

1. **Model and Preprocessing Objects Loading**:
    - The trained Random Forest model, along with the Scaler (for numerical features) and Label Encoders (for categorical features), was loaded using the joblib library.
    - These objects were saved during the model training process to ensure consistent preprocessing and predictions.
2. **Interactive User Interface:**
    - A Streamlit web application (opa.py) was developed to allow users to input their data interactively.
    - The interface includes input widgets such as select box for categorical features (e.g., Gender, FAVC) and number input for numerical attributes (e.g., Age, Height, Weight).
3. **Preprocessing User Input:**
    - User-provided inputs are preprocessed to match the format of the training data:
        - Label Encoding: Converts categorical inputs (e.g., Gender, CAEC) into numerical representations using the same encoders used during training.
        - Scaling: Normalizes numerical inputs (e.g., Height, Weight) using the Standard Scaler to ensure consistency with the model's expectations.

4. **Prediction Pipeline:**
    - The preprocessed user data is passed to the loaded Random Forest model for prediction.
    - The predicted class (obesity level) is decoded back to its original label using the corresponding label encoder for the target variable (NObesity).
5. **Real-Time Results:**
    - The predicted obesity level is displayed on the web interface as a success message, providing immediate feedback to the user.
6. **Deployment Process:**
    - The app is run locally using the “streamlit run opa.py” command.