CULTURE CORRESPONDING BODY MASS INDEX PREDICTION BOLSTERED BY SUPERVISED MACHINE LEARNING MODEL USING MULTIPLE LINEAR REGRESSION ALGORITHM

Overview:

Beginning from the foetus growth to the elderliness, health analysis is always depended on two major values: The "weight" and the "height". A healthcare dataset, based on Indian region, is utilized which includes the factors based on weight and height that indicates the body mass index value (BMI). Digging the correspondence between each feature, a linear model with multiple features, trained with guidance such that it tends to form a straight line that predicts the measure of the BMI. The range of BMI is scaled between underweight, normal, overweight, obese class I, and obese class II. The foreseen outcome could provide a vision to advance in healthcare support that could help the individuals to have a balanced food control and a healthy lifestyle.

Problem Statement:

Individuals with tight schedules often neglect to track their BMI which indirectly affects their regular habits. This lowers awareness in the centre of the population, unwittingly increasing prevalence especially towards diabetes. The presence of such freely usable platforms are critical so that they does not interfere with the tasks at hand. In our everyday routine, a single minute of self-analysis would have a higher influence on increasing life expectancy and preventing diabetes or weight related diseases. This presentation would offer an approachable, data driven substitute for preventing higher and lower BMI at each person, using a machine learning model constructed with a multiple linear regression algorithm.

Main Objective(s):

- To assess each person's BMI measure based on the weight and height features mentioned.
- To train the supervised model with a multiple linear regression that sketches an appropriate straight line and predicts accurate BMI measure for the given fresh data.

DATASET DESCRIPTION:

A. Source of data: Kaggle: https://www.kaggle.com/datasets/soumyaneelsarkar/bmi-dataset

B. Structure of data:

- **a. Rows:** 10000 records.
- **b. Columns:** 3 features (Height in cm, Weight in Kg, BMI).

C. Data columns:

COLUMN NAME	DATATYPE	COLUMN DESCRIPTION
Height	Float	Height of the individual (in centimeters).
Weight	Float	Weight of the individual (in Kilograms).
BMI	Float	Body mass index of the individual (existing data)

D. Key Variables:

- **a. X – Independent variables:** Height, Weight
- **b. Y – Dependent variables:** BMI

Y DATA PROFILE:

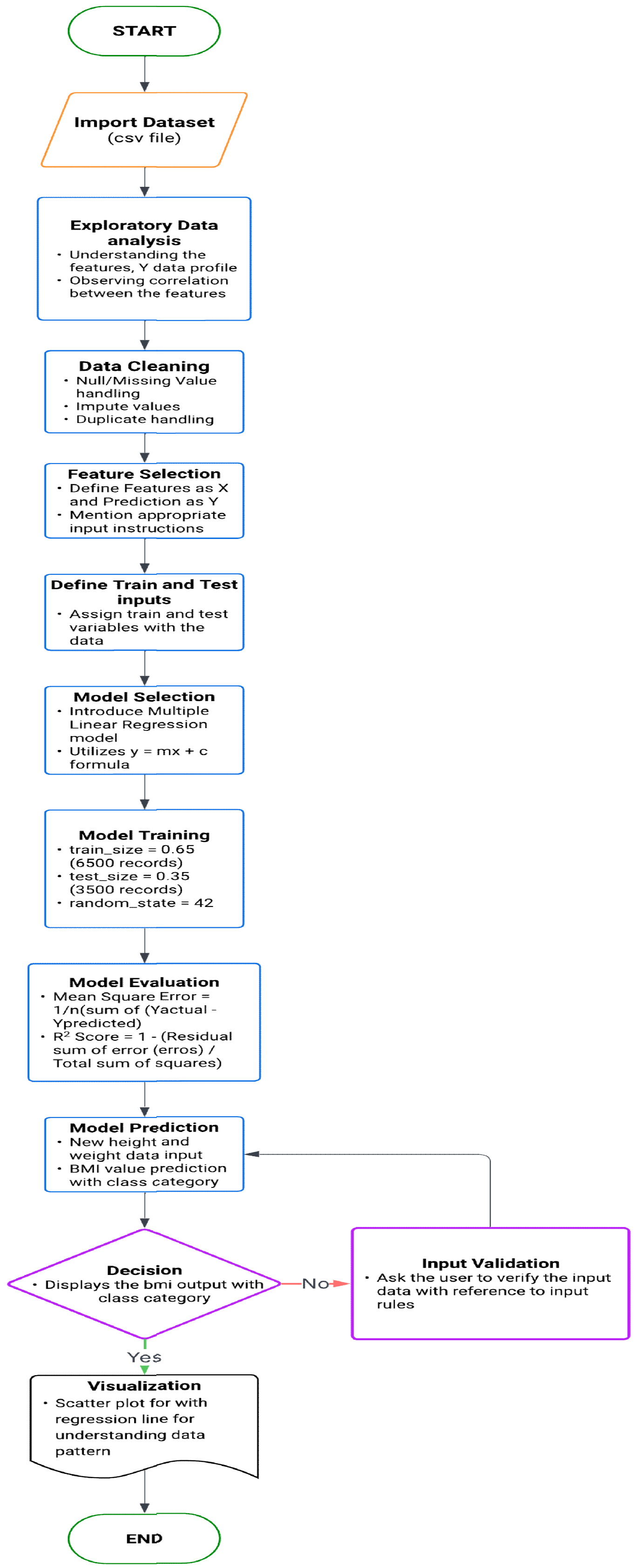
- Profile data:

bmi-edited-data-profile-report.html

- Statistical description of the data

FEATURES	**COUNT**	**MEAN**	**STD**	**MIN**	**25%**	**50%**	**75%**	**MAX**
Height	10000	160.1228	11.8196	140	150	160	170	180
Weight	10000	62.8146	21.9119	25	44	63	82	100
BMI	10000	24.9086	9.5111	7.7160	17.0392	24.5675	31.7739	51.0204

METHODOLOGY:



TRAINING – TESTING SPLIT:

Dataset training and testing size split details are as follows:

SPLIT	DESCRIPTION	SIZE (Total = 1 (10,000 records)
Training	Set of records with multiple combinations utilized by the model to understand the pattern.	0.65 (6,500 records)
Testing	Balance records are utilized to assess the model's performance following the training before a fresh input	0.35 (3,500 records)

MODEL SELECTION:

Model: Multiple Linear Regression (Supervised Model)

Description:

- Multiple linear regression is used since the BMI is a continuous variable influenced by more than one independent factor.
- With proper guidance, the model learns the patterns for each factors trained to provide accurate and expected results. Thus providing fast and reliable system.
- The output is expected to be value centric correlated with multiple factors, hence multiple linear regression model was considered.

MODEL PARAMETERS:

Parameters Reference:

PARAMETERS	Values
test_size	0.35 (Lesser the better model learns).
random_state	42

Parameters Description:

- test_size : Set of records to test the model's performance.
- random_state: Used to train the model the dataset into different combinations by shuffling the dataset. This ensures the model's reproducibility by learning data through random combinations.

MODEL TRAINING:

- Total Dataset used = 2000 records
- Test size were listed and used test_size = 0.35. Hence, 700 records were adopted as test case and 1300 records were assigned in training the model.
- To assist the model with learning multiple possibilities, random_state 42 was implemented.
- Metrics such as mean_squared_error, r2_score were imported to evaluate the performance and accuracy of the model.

MODEL EVALUATION:

Sample data values obtained:

TEST_X PARAMETERS		PRED_Y
Height (in cm)	**Weight (in Kg)**	**BMI**
168	99	36.7428
180	26	4.08839
152	87	37.0315
177	65	20.4604
172	84	29.5501
144	97	43.5051
179	64	19.4353
177	77	25.2074
172	59	19.6605
171	88	31.4472

EVALUATION METRICS:

Metrics used:

- **Mean squared error**:
    - Used to measure the average squared difference between the actual and the predicted BMI. It mitigates the large errors, optimizes the calculation, and provides a clear numerical measure of the predicted error.
    - It forms the straight line by using the below formula:

$$MSE = \frac{1}{n}\left(\sum(Yactual - Ypredicted)\right)^2$$

Where:

$$n = Number\ of\ data\ points$$
$$Yactual\ = observed\ values$$
$$Ypredicted\ = predicted\ values$$

-
    - The better value of MSE would be closer to zero (0) that marks the model's performance is at the best. Squaring ensures no cancellation of positive and negative values.
- **R^2 Score**:
    - It is helpful in finding the dependency of the variable variation that is explained by the model. It expresses how well the model fits the given data.
    - It measures the percentage of accuracy of the model using the below formula:

$$R^2 = 1 - \frac{Residual\ Sum\ of\ Squares\ (error)}{Total\ Sum\ of\ Squares}$$

Where:

$$Residual\ Sum\ of\ Squares = (Yactual - Ypredicted)^2$$

$$Total\ Sum\ of\ Squares = (Yactual - Ymean)^2$$

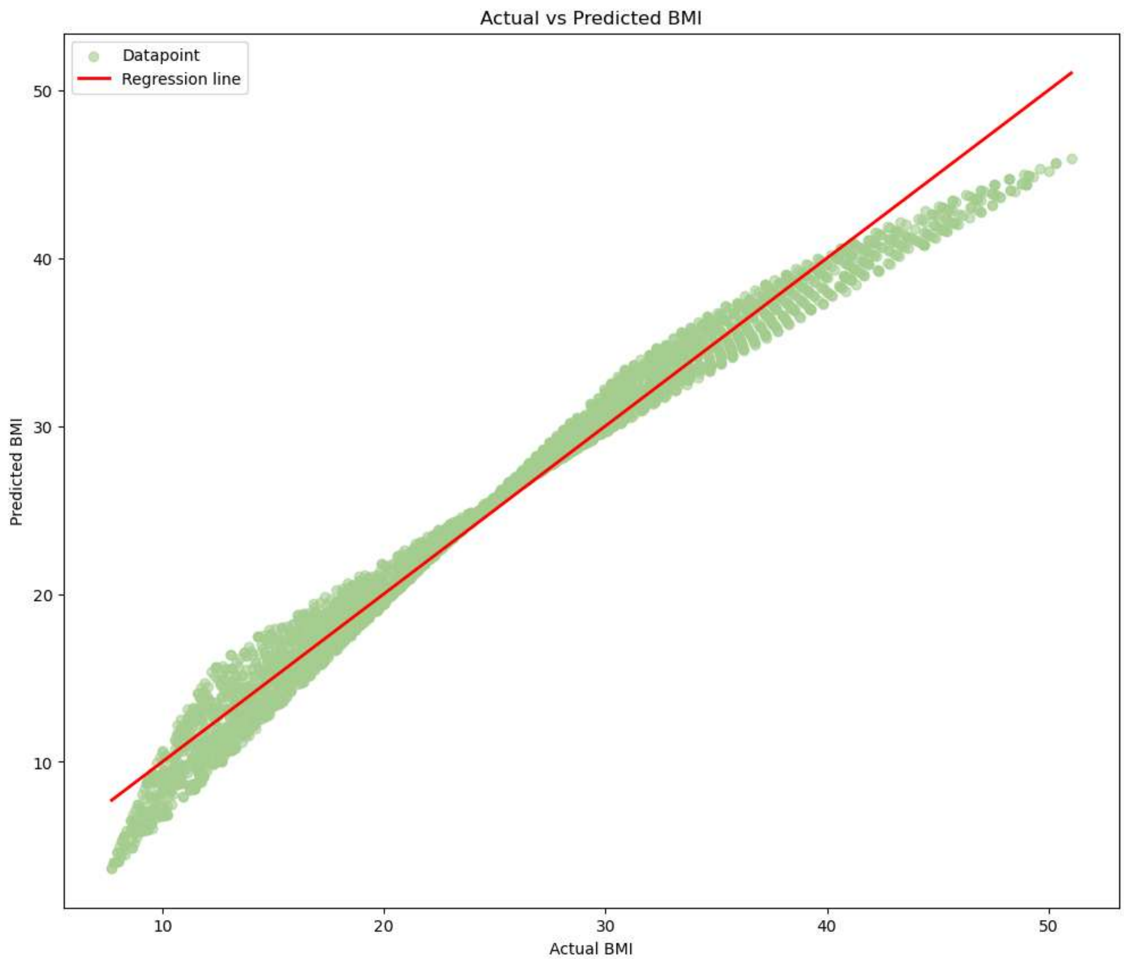
-
    - The recommended R^2 score for a best predicting model range is 0.7 to 0.9, whereas the score between 0.3 to 0.6 shows average performance and below 0.3 represents the model has to improve at a greater level.

Evaluated Metric Values from the model:

- Mean Squared Error = 1.7696
- R^2 Score = 0.98

- With the greater supporting values from the metrics evaluated, the model predicts the BMI measures with the height and weight inputs at the best.

PLOTS:



PLOT INTERPRETATION:

- From the above visible scattering plot, it is clearly visible that the data forms a linear profile and confirms the correlation between the input factors and BMI prediction.
- The red linearly increasing straight line portrays that the trained data points show less deviation since it is an actual value
- Thus the above visual stands as a supporting factor that the model predicting the BMI value might assure as the expected outcome.

NEW DATA PREDICTION:

Features	Input Values			Accuracy
	Set 1	Set 2	Set 3	
Height (in cm)	154	162	171	0.98
Weight (in Kg)	45	60	66.2	
	BMI measure			
	19.8	**23.2**	**22.8**	

CONCLUSION:

The trained multiple linear regression algorithm has successfully predicted the BMI measures from the fresh input values of height and weight of an individual. The model has showcased the best predicted accuracy of 0.98 and lowest mean squared error of 1.7696 marking the highest performance. This would also demonstrate that the model could be applied in a real time scenario for additional model analysis when considerably more data is encountered.