SemiAuto Regression Report

Dataset: GYM

Generated on: 2025-05-17 04:33:30

Project Flow

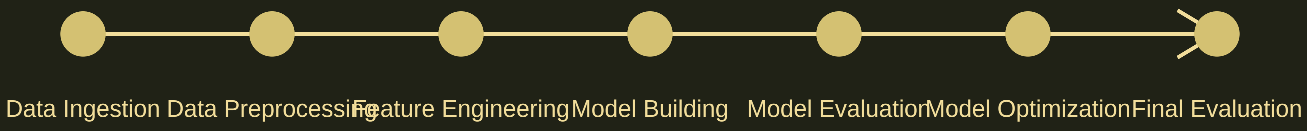


Table of Contents

1. Data Ingestion

This step involves loading and analyzing the original dataset to understand its structure and characteristics.

Dataset Overview

Dataset: gym

Train samples: 778, Test samples: 195

Target column: BMI

Column Types

Original Columns:

Age, Gender, Weight (kg), Height (m), Max_BPM, Avg_BPM, Resting_BPM, Session_Duration (hours), Calories_Burned, Workout_Type, Fat_Percentage, Water_Intake (liters), Workout_Frequency (days/week), Experience_Level, BMI

Numerical Columns:

Age, Weight (kg), Height (m), Max_BPM, Avg_BPM, Resting_BPM, Session_Duration (hours), Calories_Burned, Fat_Percentage, Water_Intake (liters), Workout_Frequency (days/week), Experience_Level, BMI

Categorical Columns:

Gender, Workout_Type

Skewed Columns:

Weight (kg), Fat_Percentage, BMI

Normal Columns:

Age, Height (m), Max_BPM, Avg_BPM, Resting_BPM, Session_Duration (hours), Calories_Burned, Water_Intake (liters), Workout_Frequency (days/week), Experience_Level

Columns with Nulls:

None

Columns with Outliers:

Weight (kg), Calories_Burned, BMI

Highly Correlated Features

Weight (kg):

- BMI: 0.8532

Session_Duration (hours):

- Calories_Burned: 0.9081
- Experience_Level: 0.7648

Calories_Burned:

- Session_Duration (hours): 0.9081
- Experience_Level: 0.6941

Fat_Percentage:

- Experience_Level: -0.6544

Workout_Frequency (days/week):

- Experience_Level: 0.8371

Experience_Level:

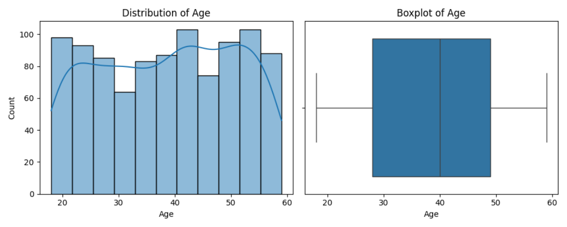
- Session_Duration (hours): 0.7648
- Calories_Burned: 0.6941
- Fat_Percentage: -0.6544
- Workout_Frequency (days/week): 0.8371

BMI:

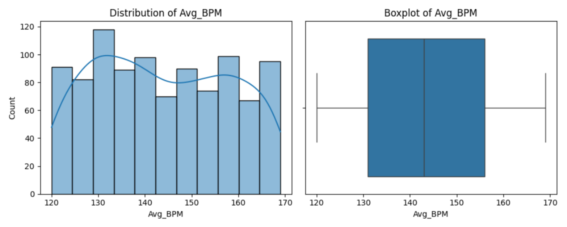
- Weight (kg): 0.8532

Feature Distributions

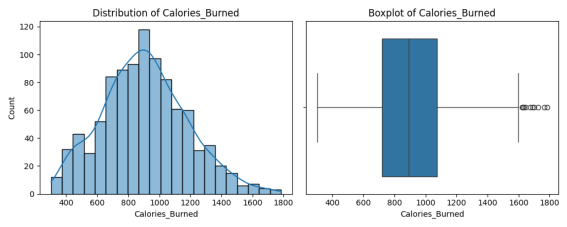
Age



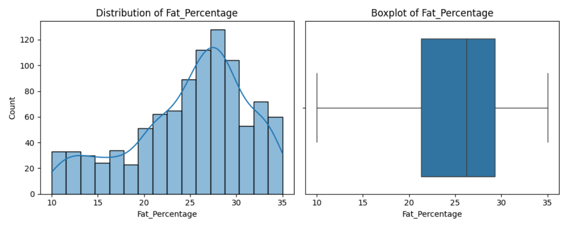
Avg_BPM



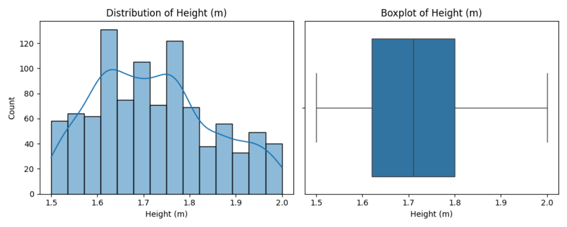
Calories_Burned



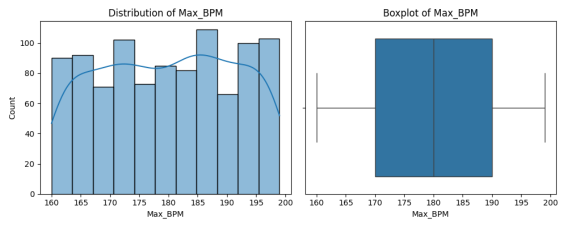
Fat_Percentage



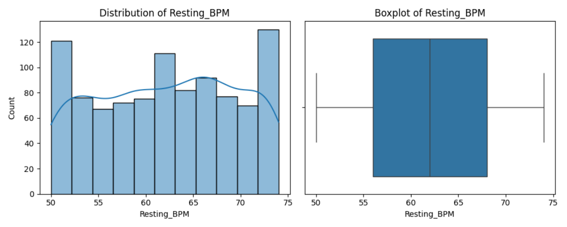
Height (m)



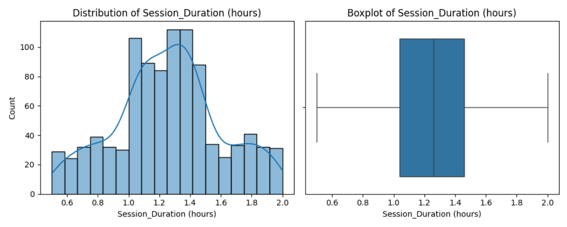
Max_BPM



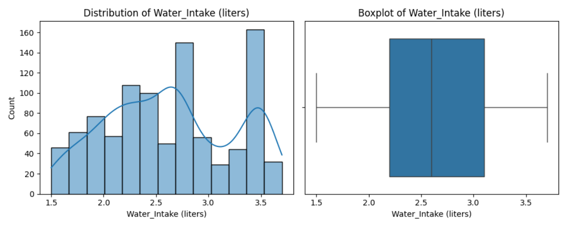
Resting_BPM



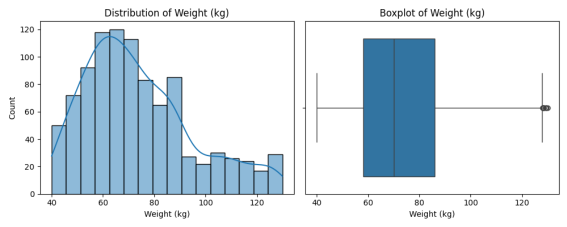
Session_Duration (hours)



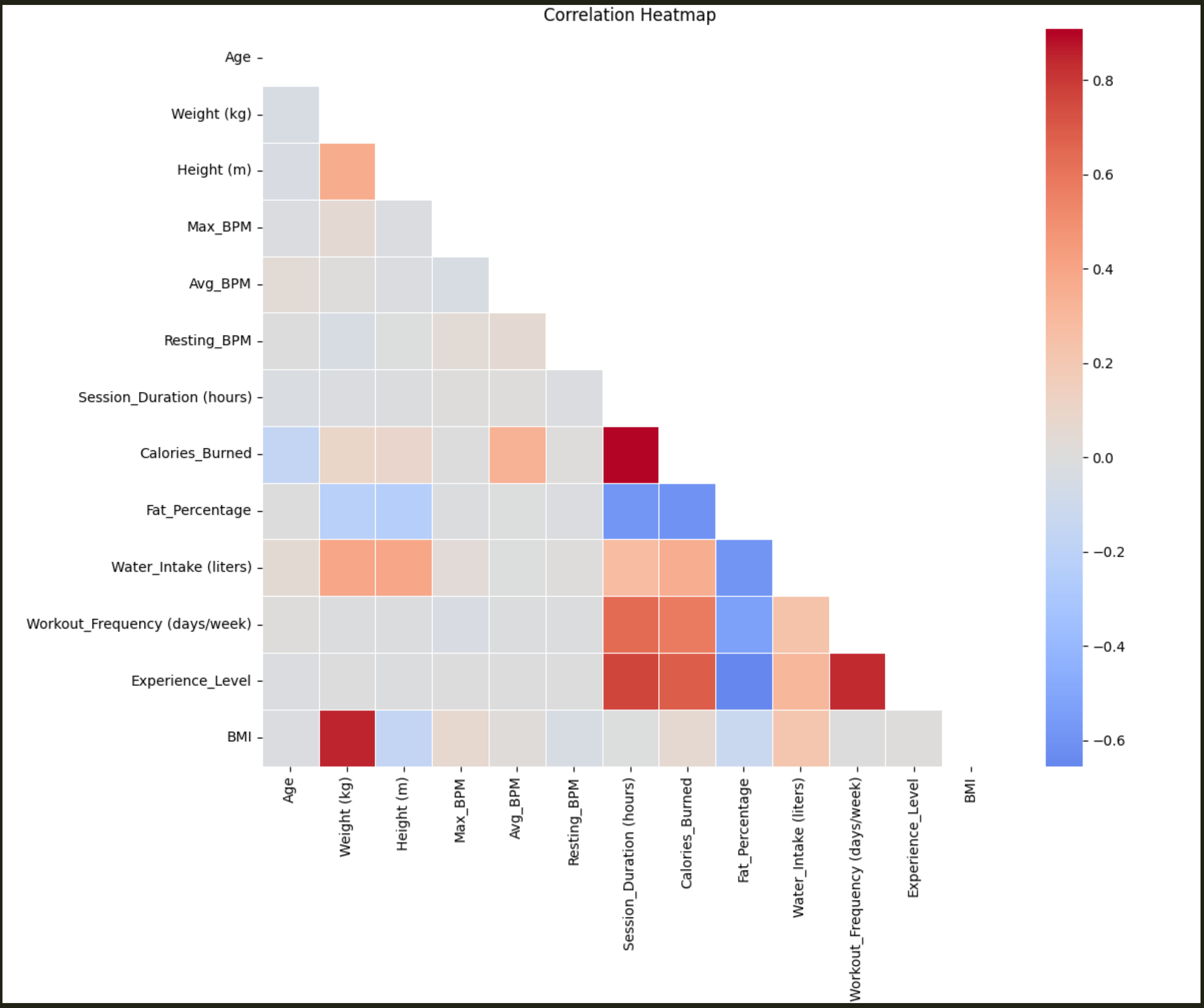
Water_Intake (liters)



Weight (kg)



Correlation Heatmap



2. Data Preprocessing

This step involves cleaning the dataset and preparing it for model training.

Preprocessing Configuration

Duplicate handling:

Remove duplicates: True

Outlier Treatment:

Method: IQR

Applied to columns:

- Weight (kg)
- Calories_Burned

Skewed Data Transformation:

Method: box-cox

Applied to columns:

- Weight (kg)
- Fat_Percentage

Numerical Scaling:

Method: standard

Applied to columns:

- Age
- Weight (kg)
- Height (m)
- Max_BPM
- Avg_BPM
- Resting_BPM
- Session_Duration (hours)
- Calories_Burned
- Fat_Percentage
- Water_Intake (liters)
- Workout_Frequency (days/week)
- Experience_Level

Categorical Encoding:

Method: onehot

Drop first: True

Applied to columns:

- Gender
- Workout_Type

Preprocessed Data Preview

Training Data Sample (First 5 rows):

Age	Weight (kg)	Height (m)	Max_BPM	Avg_BPM	Resting_BPM	Session_Du..	Calories_B..	Fat_Percen..	Water_Inta..
-1.5330451..	-3.5324010..	-0.9700883..	-0.6810859..	1.70030808..	0.64940678..	-0.4037490..	0.48828893..	-4.0568311..	0.93198659..
0.03035418..	-3.5144073..	-0.5824515..	-1.5471938..	1.63084945..	0.10774130..	-0.9977108..	-0.5071114..	-3.9546773..	-1.5637186..
1.26461680..	-3.4756539..	-0.0397601..	1.22435159..	0.72788720..	-0.2985078..	-0.8492203..	-0.8650082..	-3.9902220..	-1.2309579..
-1.2039084..	-3.5196223..	-1.6678344..	0.61807601..	0.93626310..	0.78482315..	-0.6413337..	-0.3281630..	-3.8220239..	-0.2326758..
-1.5330451..	-3.5006984..	-1.5903071..	0.44485442..	-1.1474959..	-1.6526715..	-0.6710318..	-0.9321139..	-3.9424981..	-0.7318169..

Test Data Sample (First 5 rows):

Age	Weight (kg)	Height (m)	Max_BPM	Avg_BPM	Resting_BPM	Session_Du..	Calories_B..	Fat_Percen..	Water_Inta..
-0.3810666..	-3.4841766..	-0.6599789..	-0.8543075..	-0.6612854..	0.64940678..	0.01202429..	0.09311124..	-4.1245912..	-0.8981972..
0.77091175..	-3.4584482..	0.73551330..	-0.3346427..	0.38059403..	1.05565589..	1.91270222..	1.85277042..	-4.2832452..	1.43112764..
0.85319592..	-3.4678900..	-0.0397601..	1.22435159..	-1.0780372..	0.92023952..	0.60598614..	0.07819887..	-3.8823926..	1.09836694..
-0.7102033..	-3.4940894..	0.19282187..	0.01180044..	1.70030808..	0.64940678..	-0.0176738..	0.93565991..	-3.9081537..	0.76560624..
-0.5456350..	-3.4546463..	-0.9700883..	-0.3346427..	0.10275949..	-0.2985078..	2.17998505..	2.54619544..	-4.2612287..	1.43112764..

3. Feature Engineering

This step involves creating new features or selecting the most important ones.

Feature Engineering Configuration

Applied Techniques:

Automated Feature Engineering: Yes

SHAP-based Feature Selection: Yes

Transformed Data Preview

Transformed Training Data Sample (First 5 rows):

Gender_Mal..	Weight (kg)	Gender_Mal..	Gender_Mal..	Fat_Percen..	Gender_Mal..	Fat_Percen..	Height (m)..	Weight (kg..	Height (m)..
-0.2830935..	-3.5324010..	1.97008833..	-2.5324010..	-5.0269194..	4.53240101..	-3.0867427..	2.56231267..	-3.5324010..	-0.9700883..
-0.0	-3.5144073..	0.58245159..	-3.5144073..	-4.5371289..	3.51440738..	-3.3722257..	2.93195578..	0.0	-0.5824515..
-0.0	-3.4756539..	0.03976016..	-3.4756539..	-4.0299822..	3.47565397..	-3.9504618..	3.43589380..	0.0	-0.0397601..
-0.0	-3.5196223..	1.66783445..	-3.5196223..	-5.4898584..	3.51962235..	-2.1541895..	1.85178789..	0.0	-1.6678344..
-0.0	-3.5006984..	1.59030711..	-3.5006984..	-5.5328052..	3.50069842..	-2.3521910..	1.91039131..	0.0	-1.5903071..

Transformed Test Data Sample (First 5 rows):

Gender_Mal..	Weight (kg)	Gender_Mal..	Gender_Mal..	Fat_Percen..	Gender_Mal..	Fat_Percen..	Height (m)..	Weight (kg..	Height (m)..
-0.2870118..	-3.4841766..	1.65997894..	-2.4841766..	-4.7845701..	4.48417669..	-3.4646122..	2.82419774..	-3.4841766..	-0.6599789..
-0.2891470..	-3.4584482..	0.26448669..	-2.4584482..	-3.5477319..	4.45844829..	-5.0187585..	4.19396159..	-3.4584482..	0.73551330..
-0.2883597..	-3.4678900..	1.03976016..	-2.4678900..	-3.9221527..	4.46789000..	-3.8426324..	3.42812983..	-3.4678900..	-0.0397601..
-0.2861975..	-3.4940894..	0.80717812..	-2.4940894..	-3.7153318..	4.49408949..	-4.1009755..	3.68691136..	-3.4940894..	0.19282187..
-0.2894652..	-3.4546463..	1.97008833..	-2.4546463..	-5.2313171..	4.45464633..	-3.2911404..	2.48455799..	-3.4546463..	-0.9700883..

4. Model Building

This step involves training the regression model on the transformed data.

Model Selection

Selected Model:

CatBoost

Training timestamp: 2025-05-17 04:09:59

5. Model Evaluation

This step involves evaluating the performance of the trained model.

Performance Metrics

Original Model Performance:

Evaluation timestamp: 2025-05-17 04:33:27

Metric	Value
R² Score	0.99458
Explained Variance Score	0.99459
Mean Squared Error	0.26406
Root Mean Squared Error	0.51387
Mean Absolute Error	0.32438
Mean Absolute Percentage Error	0.01242
Max Error	3.50000

6. Model Optimization

This step involves tuning the hyperparameters of the model to improve performance.

Error: Could not decode hyperparameters file.

Optimization timestamp: 2025-05-17 04:33:25

7. Final Evaluation Results

This section presents the final performance of the optimized model.

Optimized Model Performance

Metric	Value
R² Score	0.99840
Explained Variance Score	0.99842
Mean Squared Error	0.07786
Root Mean Squared Error	0.27904
Mean Absolute Error	0.22171
Mean Absolute Percentage Error	0.00912
Max Error	1.03904

Evaluation timestamp: 2025-05-17 04:33:27

Performance Comparison

Metric	Original Model	Optimized Model	Improvement
R² Score	0.99458	0.99840	+0.38%
RMSE	0.51387	0.27904	+45.70%
MAE	0.32438	0.22171	+31.65%

Conclusion

Summary of the regression model development and performance.

This report summarizes the development of a regression model to predict BMI using the gym dataset. A CatBoost regression model was trained and optimized using hyperparameter tuning. The optimization process improved the model's R^2 score from 0.99458 to 0.99840, representing a 0.38% improvement.

This automatic report was generated to provide insights into the model development process and performance metrics. It includes details about data preprocessing, feature engineering, model selection, and evaluation results.