Model Development and Evaluation Report

1. Data Preparation

The dataset was loaded from the *data/train.csv* file. The **id** column was removed as it served only as an identifier and carried no predictive value. No missing values were found in the dataset. Duplicates were checked after removing the **id** column and none were found.
Although outliers were present, no removal was performed as boosting algorithms (LightGBM and XGBoost) are inherently robust to outliers. Multicollinearity analysis and feature distribution checks were conducted, revealing no highly correlated features and no extreme skewness in distributions.

2. Feature Engineering

The following derived features were created to capture important interactions and ratios:

Feature	Description
Workload	Weight × Duration
Cardio_Load	Heart_Rate × Duration
BMI	Weight / (Height/100)^2
Temp_Stress	(Body_Temp – 37) × Duration
HR_Ratio	Heart_Rate / (220 – Age)
Wt_per_Height	Weight / Height
Sqrt_Body_Temp	√(Body_Temp)

Feature standardization was tested, but results indicated better generalization without standardization.

3. Models Used

Two gradient boosting models were developed and compared:

Model	Key Hyperparameters	Validation RMSE
LightGBM	n_estimators=2000 learning_rate=0.05 num_leaves=128 subsample=0.8 colsample_bytree=0.8	3.6516
XGBoost	n_estimators=2000 learning_rate=0.05 max_depth=8 subsample=0.8 colsample_bytree=0.8 tree_method="hist"	3.6518

4. Ensemble Performance

An ensemble of LightGBM and XGBoost predictions was created by averaging their outputs. This improved the validation RMSE to **3.6328**, showing that the two models captured slightly different patterns in the data and complemented each other.

5. Conclusion

The modeling pipeline successfully leveraged boosting algorithms to achieve strong predictive performance without the need for outlier removal or feature scaling. Feature engineering contributed meaningful additional predictors, and model ensembling provided a modest performance boost over the individual models.