Project background

OBJECTIVES

Predicts calorie burn based on personal characteristics (gender, age, height, weight, exercise duration, heart rate, body temperature).

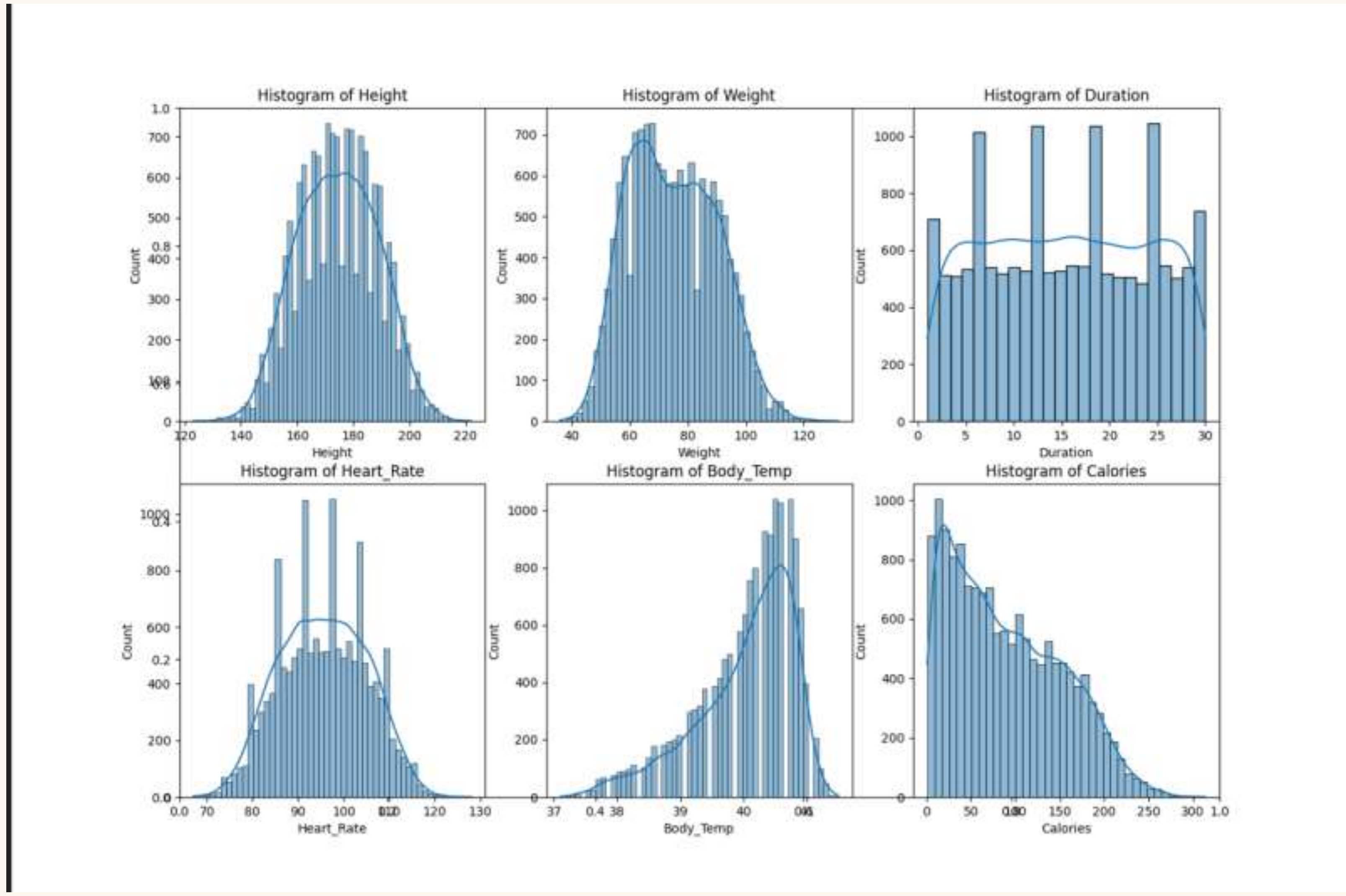
DATASET

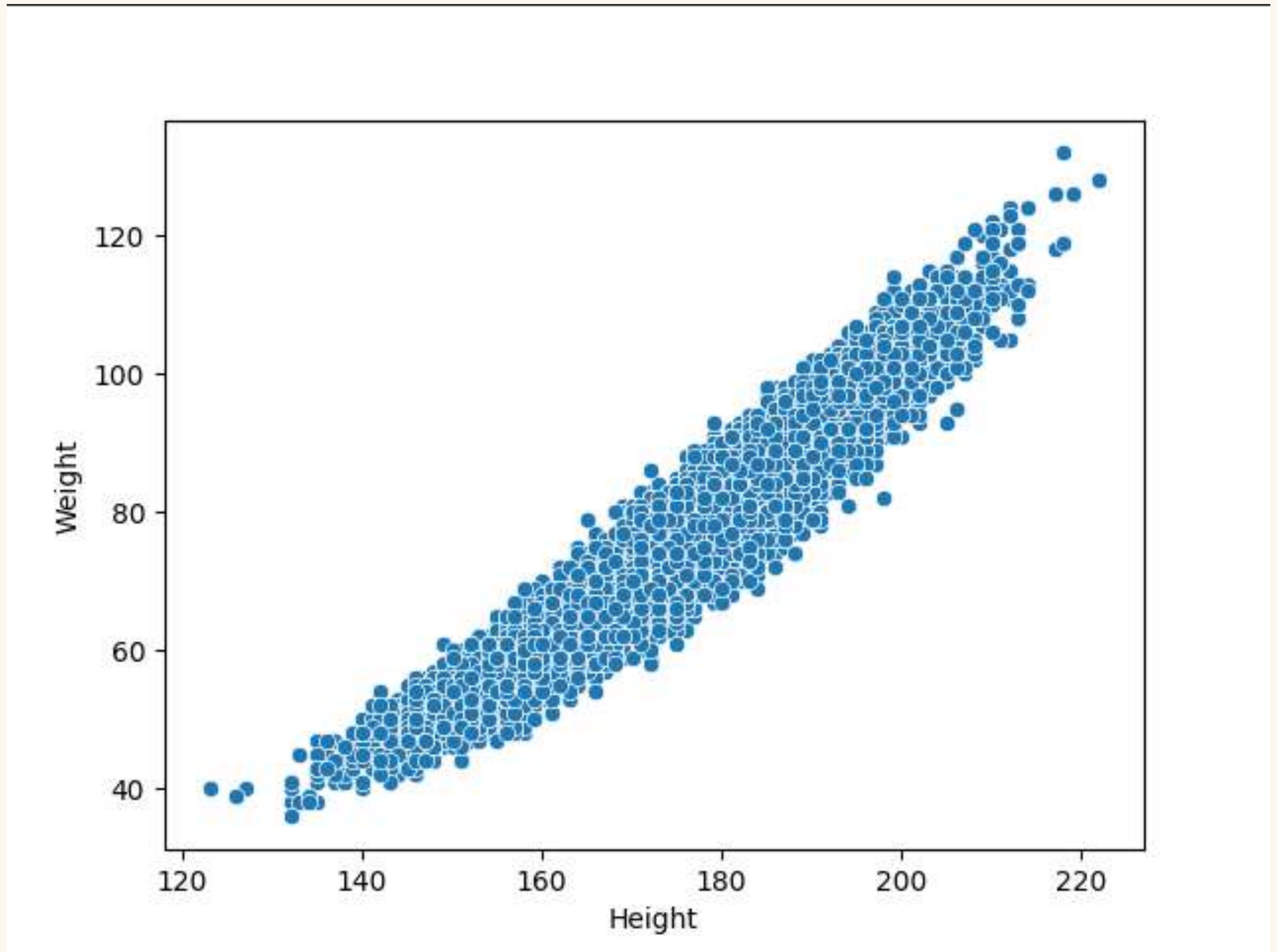
File: calories.csv
Contains 15,000 records, 7 features: Gender, Age, Height, Weight, Duration, Heart_Rate, Body_Temp
Target variable: Calories (calories burned)

APPLICATION SCENARIO

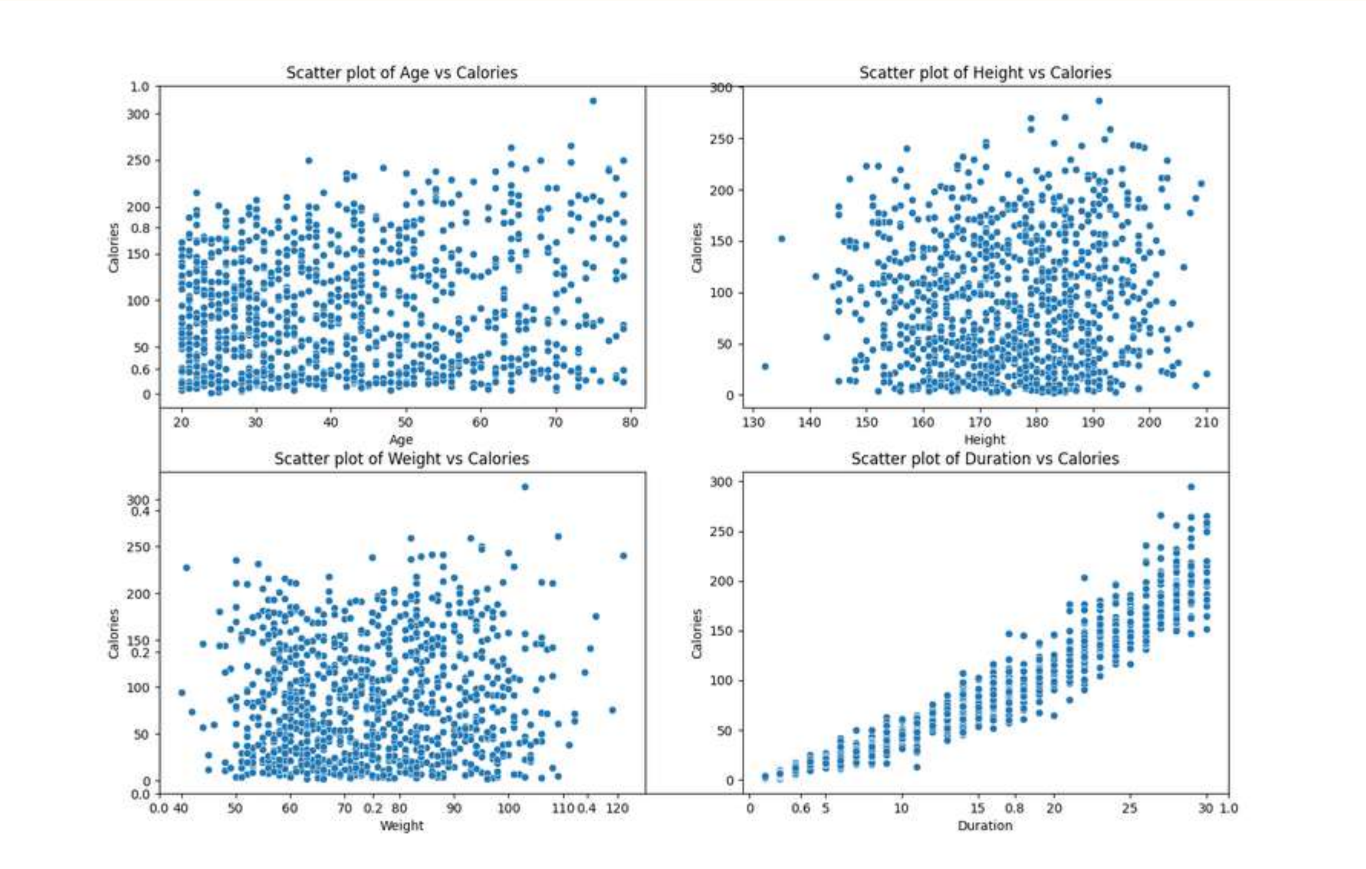
Health monitoring, fitness programme optimisation

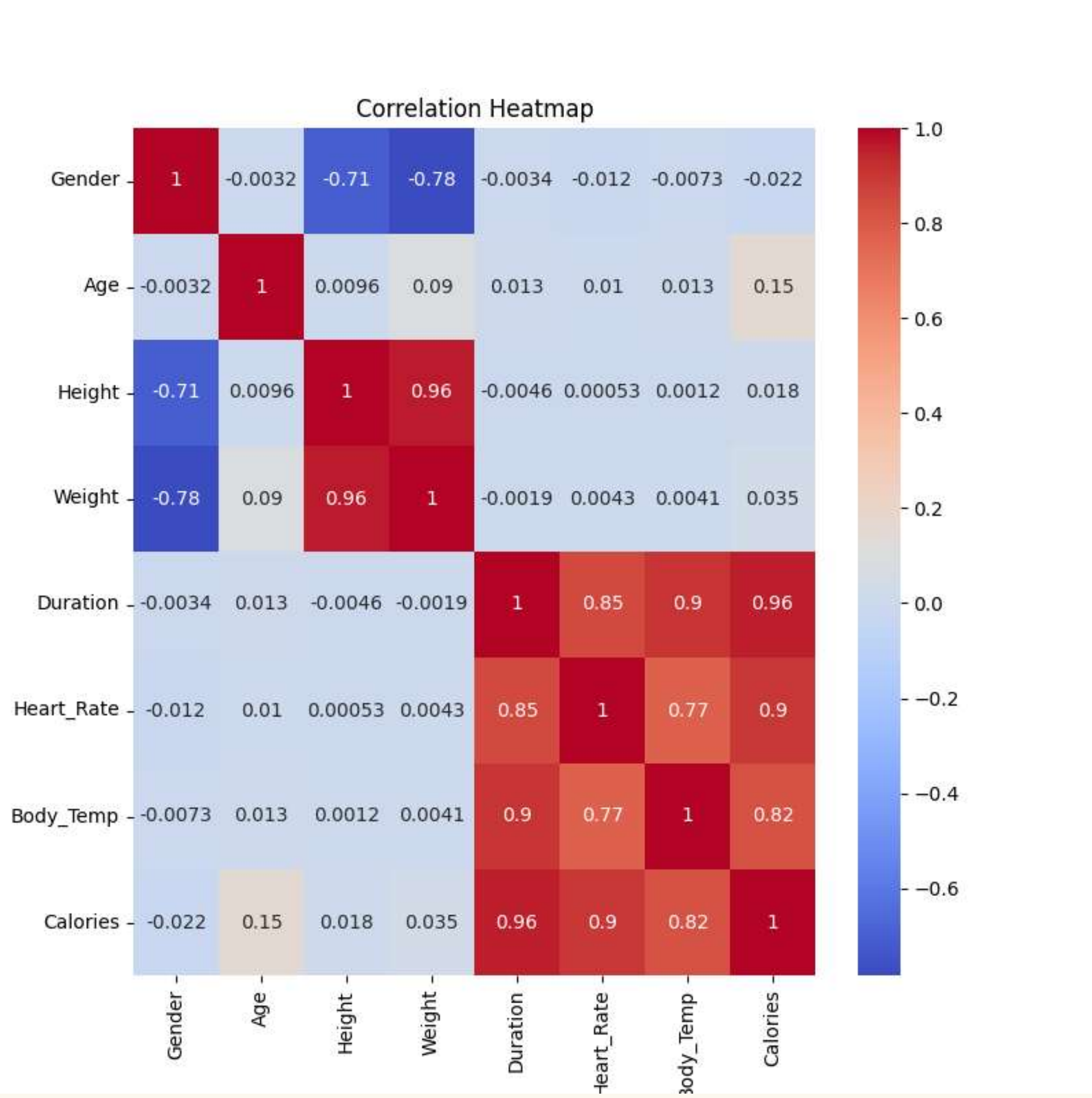
Data Exploration





Data Exploration





1 FEATURE DISTRIBUTION

Observations:

Age and Height are more evenly distributed.

Duration and Calories are somewhat skewed.

2 RELEVANCE ANALYSIS

Key Finding.

Duration and Calories were highly positively correlated (correlation coefficient of about 0.95).

Heart_Rate and Calories are also strongly correlated (~0.85).

There was some covariance between Height and Weight.

Methodologies

DATA PREPROCESSING

Feature Scaling:
Numerical features (Age, Height, Weight, Duration, Heart_Rate, Body_Temp) are standardised using StandardScaler

Dataset Segmentation:
80% training set, 20% test set (42 random seeds).

MODEL SELECTION AND HYPERPARAMETER TUNING

Models used:
Linear regression, Ridge regression, Lasso regression
Random Forest, XGBoost, Neural Networks (MLP)

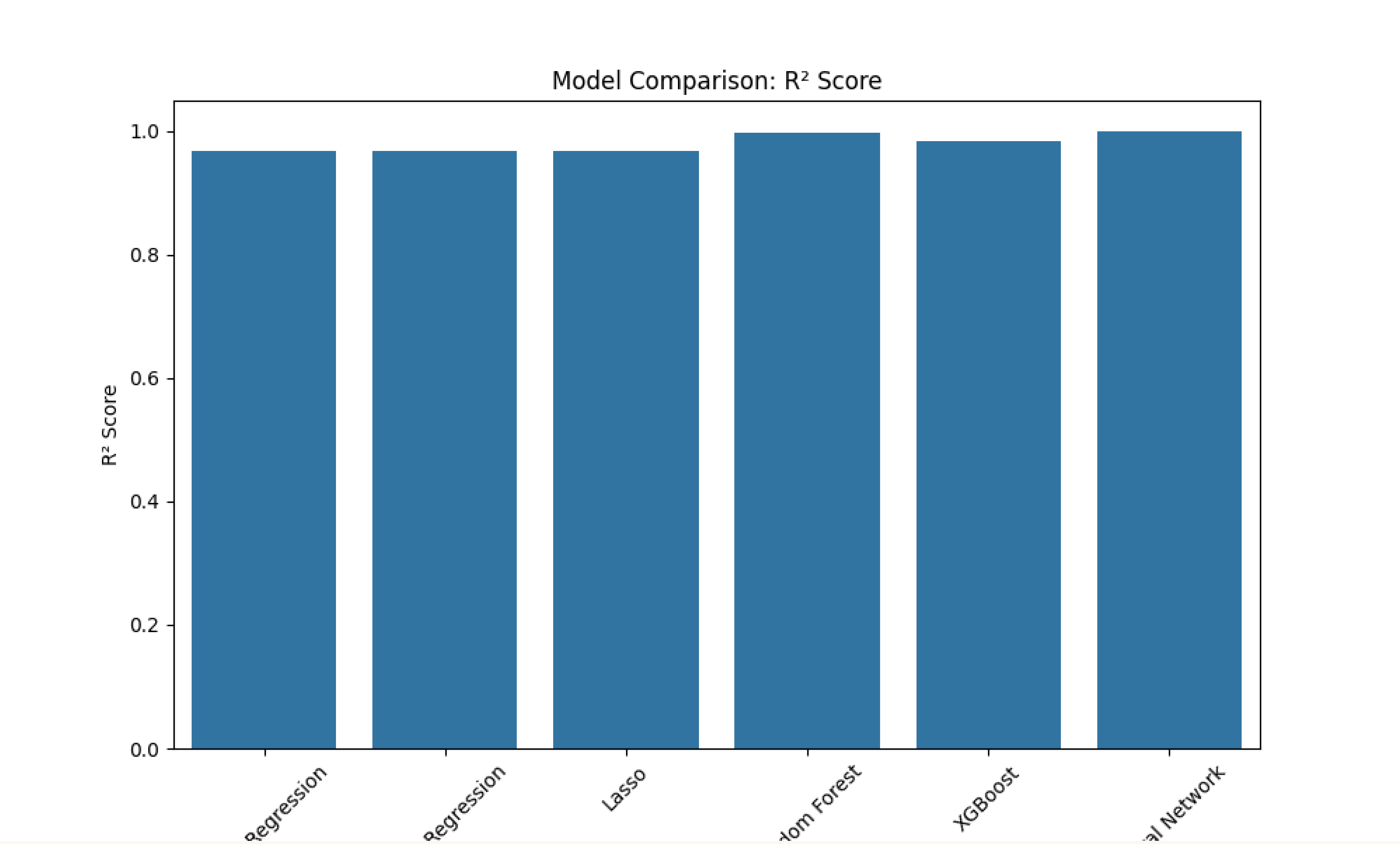
Hyperparameter Tuning:
Optimise hyperparameters using GridSearchCV (5 fold cross validation).

Results: comparison of model performance

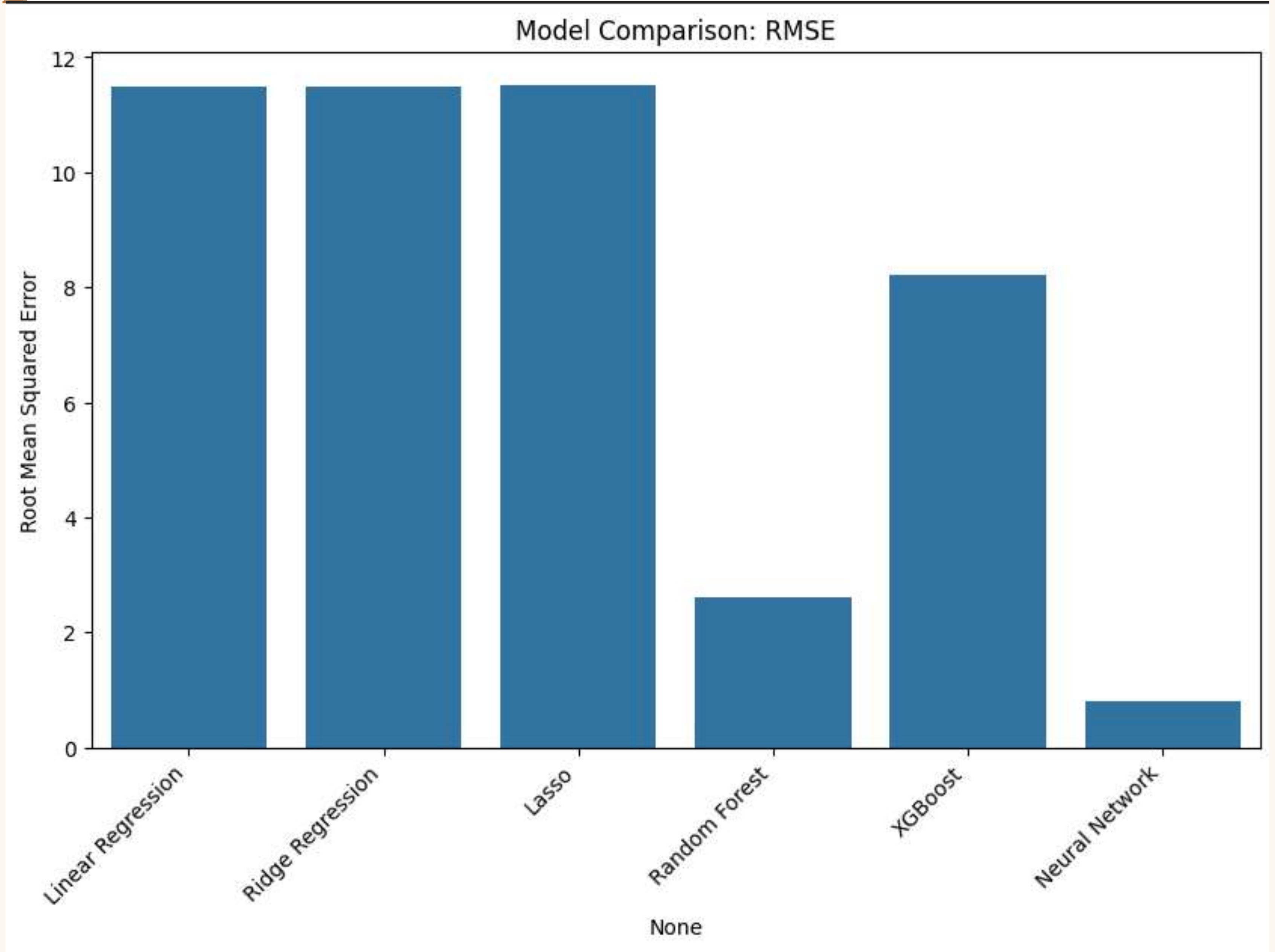
PERFORMANCE INDICATORS

Performance Metrics:

Model	**MSE**	**RMSE**	**R^2**	**CV MSE**	**CV R^2**
Linear Regression	131.80	11.48	0.9673	9.78e-29	0.9670
Ridge Regression	131.80	11.48	0.9673	8.41e-10	0.9670
Lasso	132.85	11.53	0.9671	1.00e-02	0.9669
Random Forest	6.85	2.62	0.9983	1.77e-02	0.9976
XGBoost	67.33	8.21	0.9833	1.42e-05	0.9991
Neural Network	0.63	0.79	0.9998	1.06e-03	0.9999



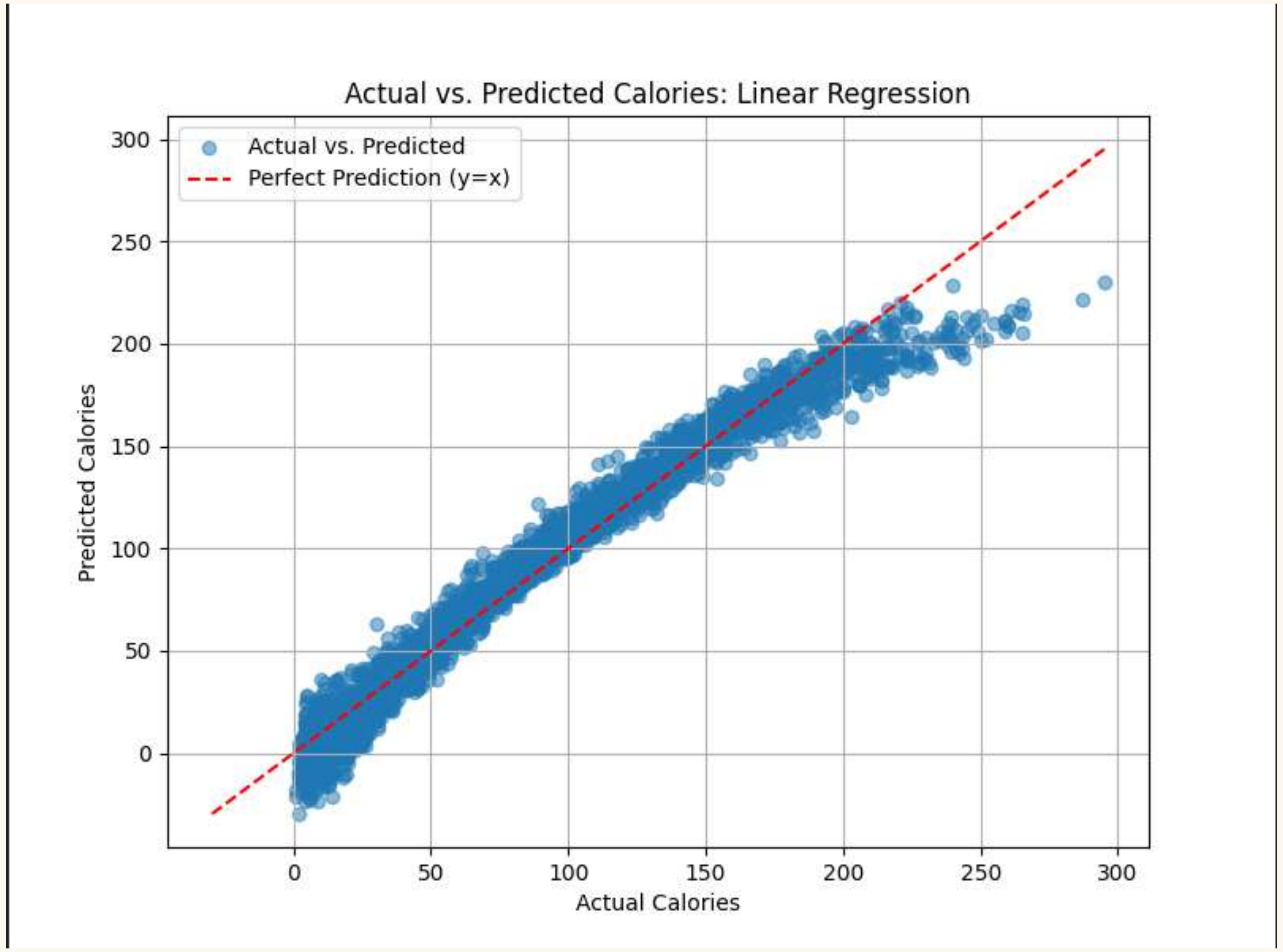
Results: comparison of model performance

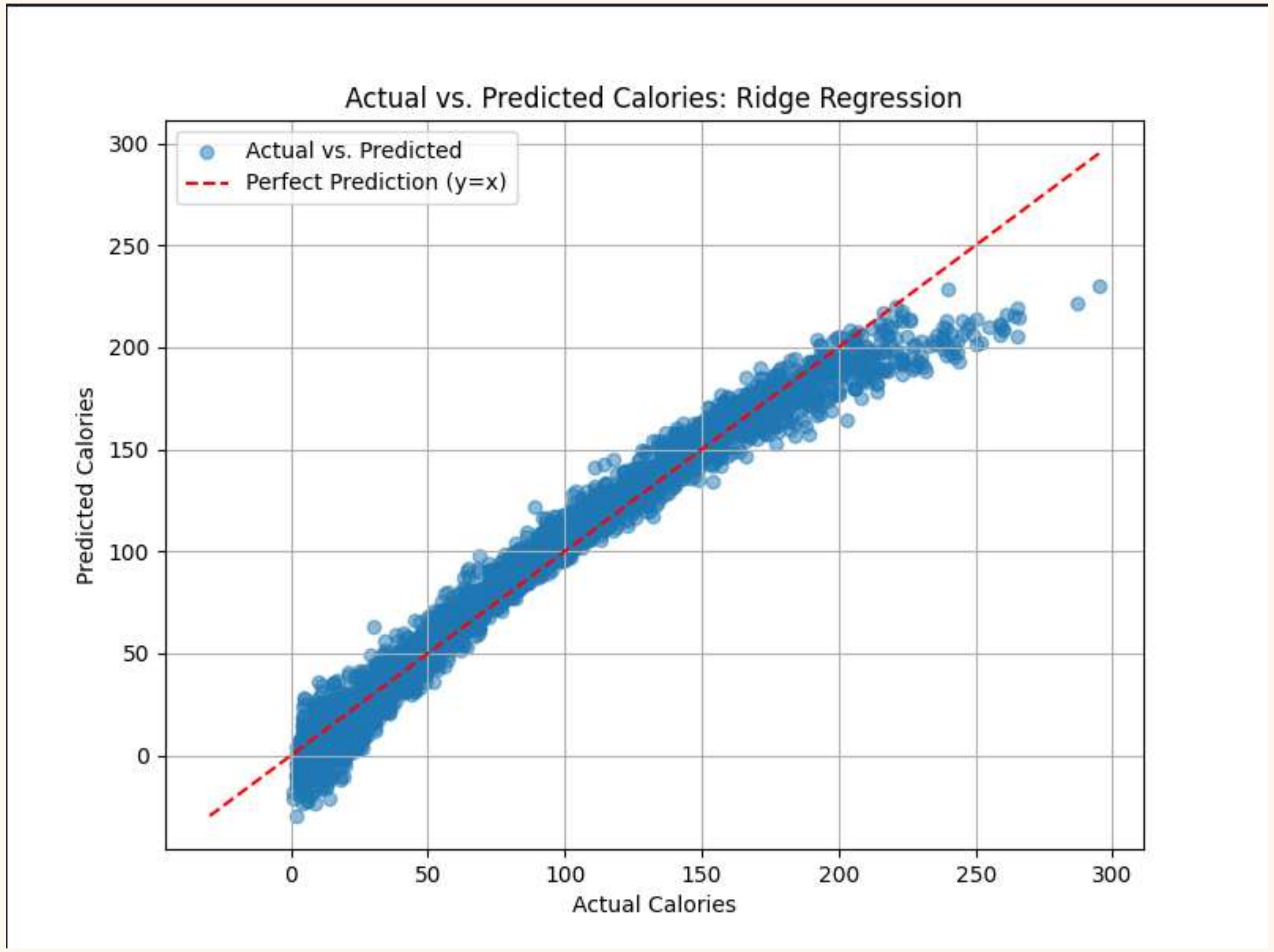


Observation:
The neural network performed best (R^2=0.9998, RMSE=0.79).
Random Forest was next best (R^2=0.9983, RMSE=2.62).

Results: actual versus projected

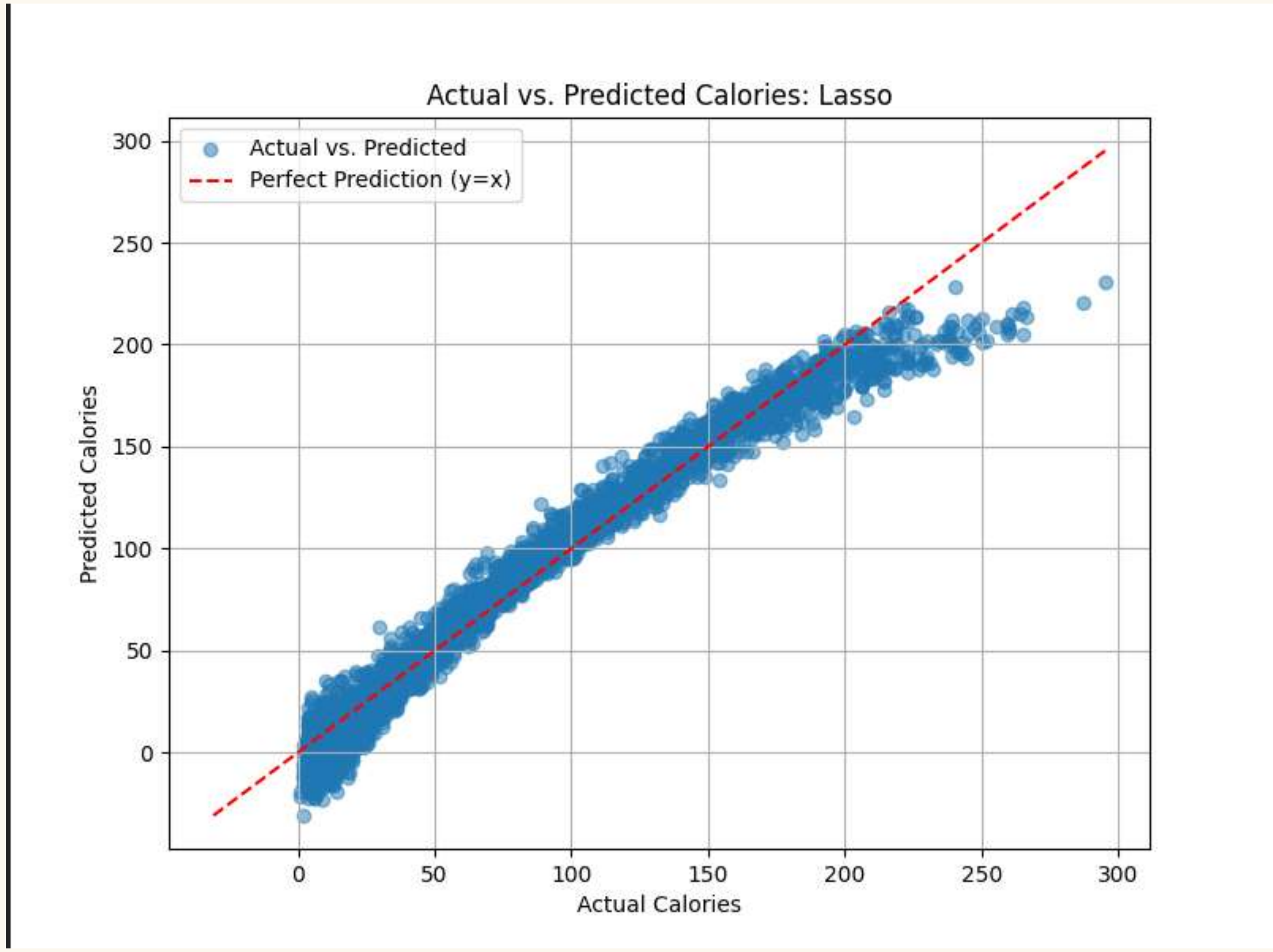
Plot actual vs. predicted scatterplot (with fitted line)

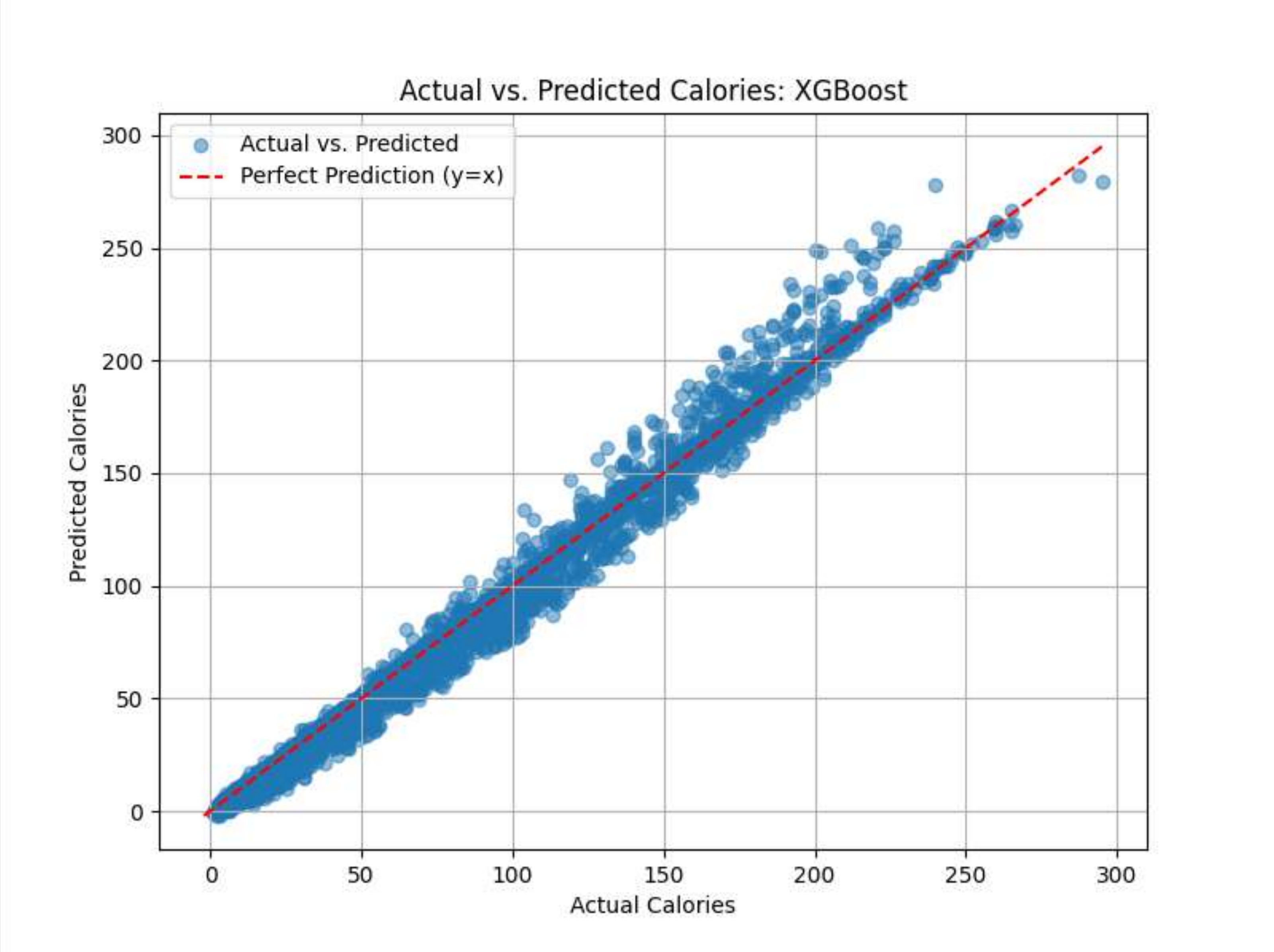




Results: actual versus projected

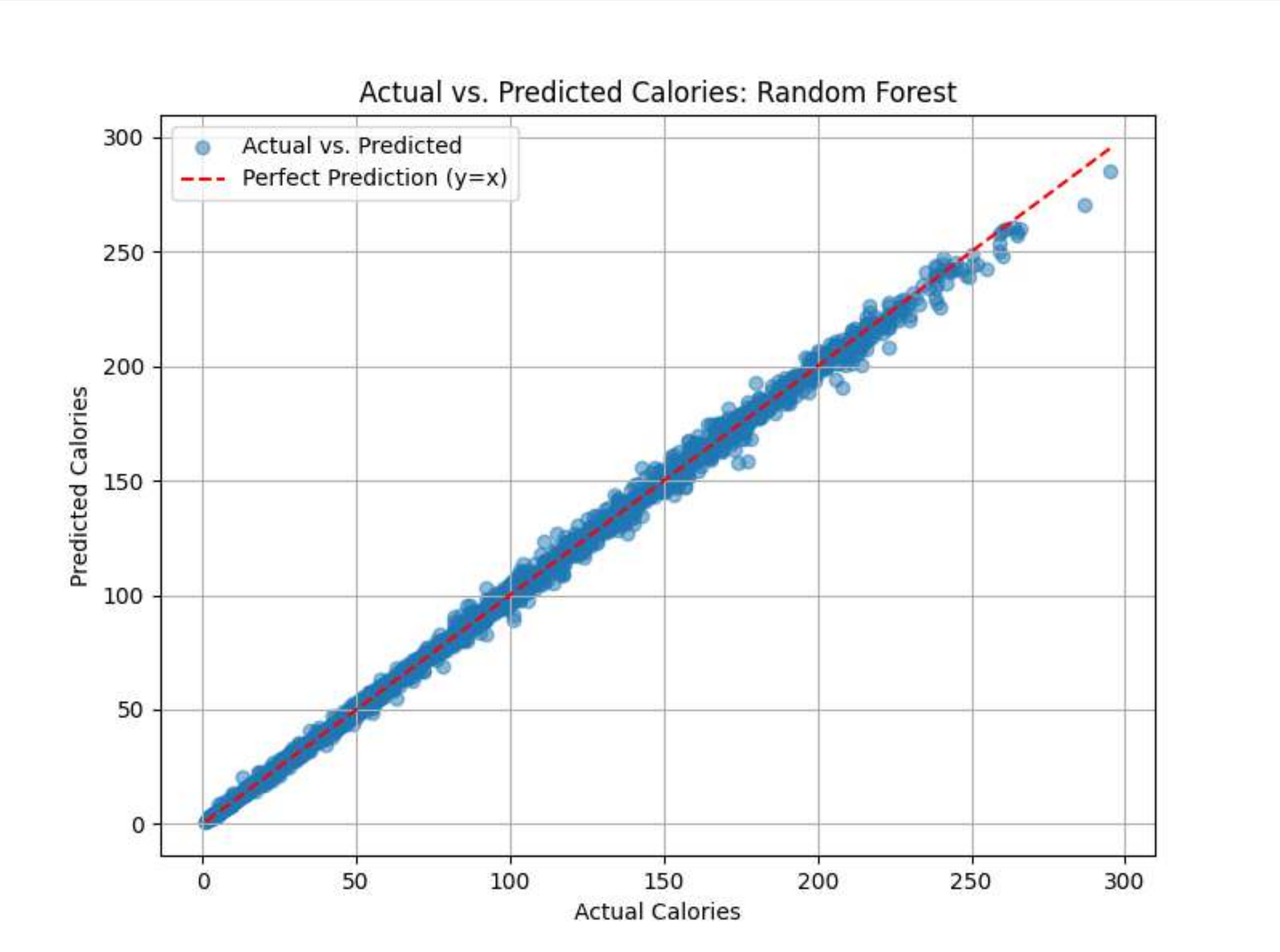
Plot actual vs. predicted scatterplot (with fitted line)

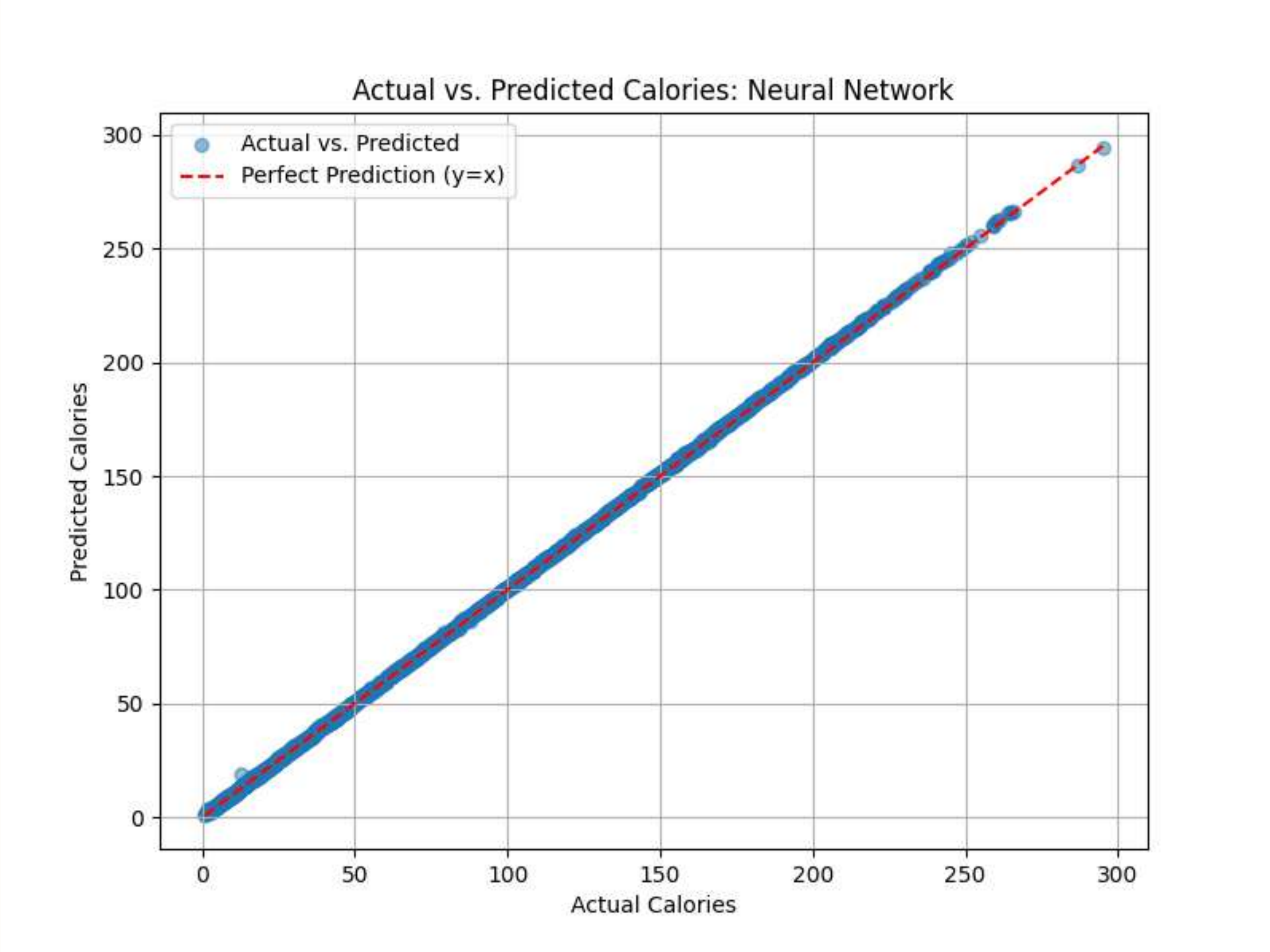




Results: actual versus projected

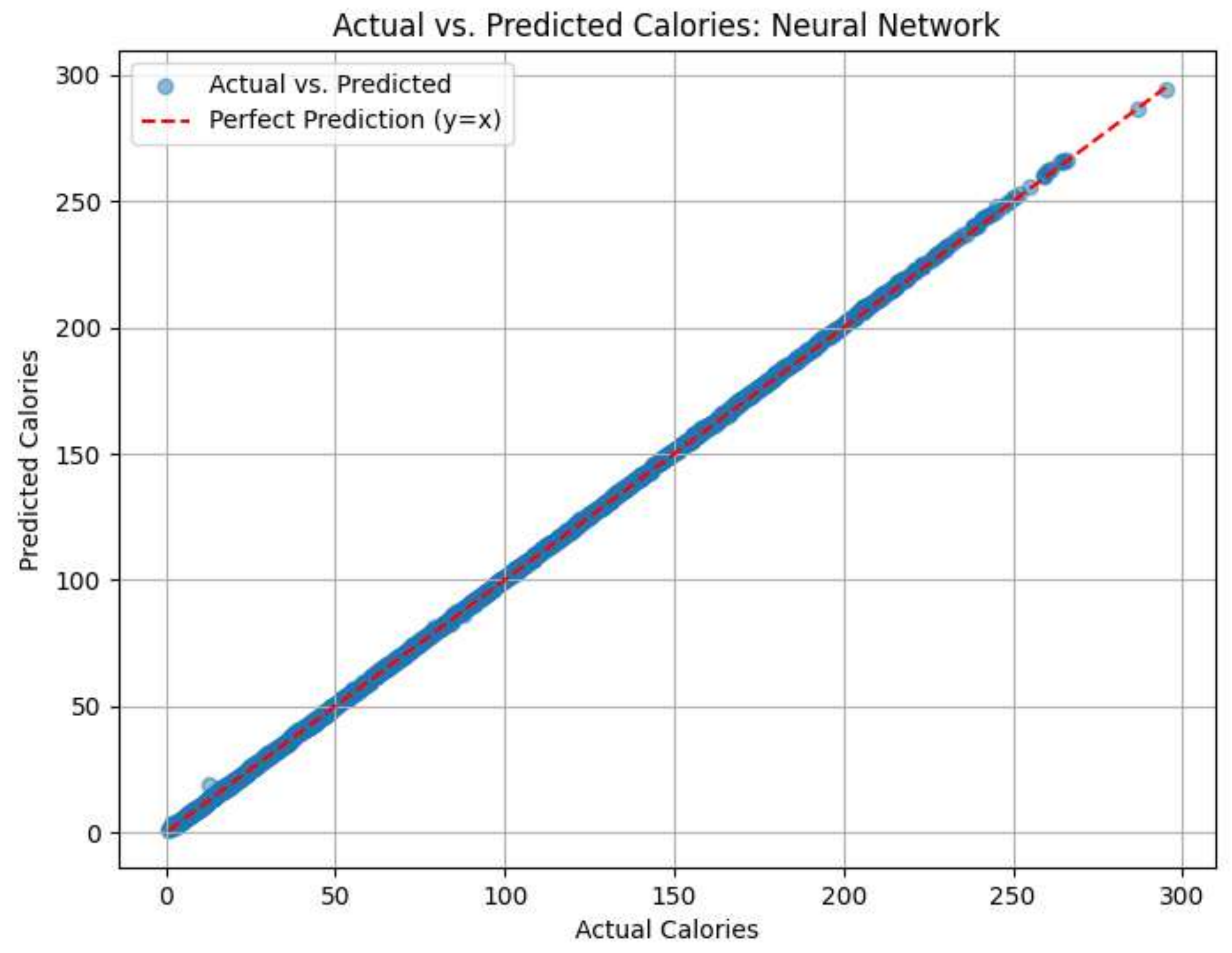
Plot actual vs. predicted scatterplot (with fitted line)

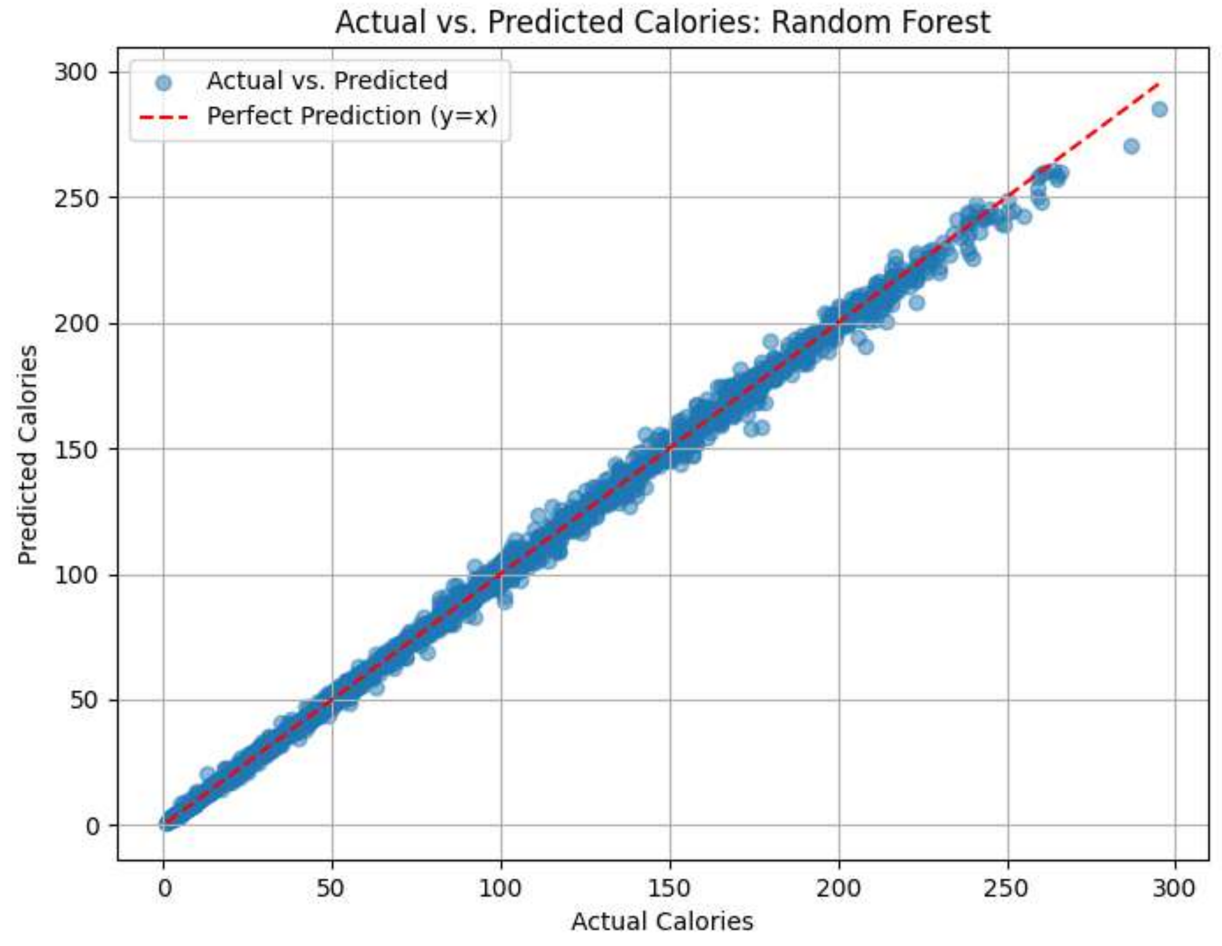




Results: actual versus projected

The neural network and random forest predictions are close to the actual values (slope of the fitted line ≈ 1).
The linear regression class of models is more biased ($R^2 \approx 0.967$).





DISCUSSION: MODELLING PERFORMANCE ANALYSIS

Neural Networks and Random Forests:

Capturing non-linear relationships with excellent performance (R^2>0.998).

Higher computational cost and long training time.

Linear regression type models :

Simple and explanatory, but assumes a linear relationship and limited performance ($R^2 \approx 0.967$).

Performs poorly on non-linear patterns (e.g. Age and Calories).

CV MSE Exception.

Cross-validation MSE values are abnormally small (e.g., 9.78e-29), there may be a data preprocessing problem that requires further examination.

SUMMARY AND FUTURE WORK

Summary:

A calorie consumption prediction model was successfully constructed with the best performance of the neural network (R^2=0.9998).

Features such as Duration and Heart_Rate contributed the most to the prediction.

Future work:

Introducing polynomial features to improve linear regression-like models.

Optimise neural network structure to reduce computational cost.

Solve potential problems in cross-validation to ensure the reliability of results.