

Comparative Analysis of Linear and Polynomial Regression for Accurate Prediction of Calories Burned During Workouts

A.K.S.Sandaruwan wimalasena

January 22, 2025

Abstract

This study analyzed the predictive performance of multiple linear regression and second-degree polynomial regression models. The evaluation on test data showed that polynomial regression achieved a higher R-squared value (0.997) compared to linear regression (0.975), indicating a better fit for capturing complex patterns. To facilitate practical use, the final model was deployed through a graphical user interface (GUI), allowing users to input data and receive predictions easily. Findings show that polynomial regression is one of the effective forecasters for correctness while it is crucial for access to actual real-world implementation.

Keywords

Polynomial Regression, Multiple linear regression, Regression

1 Introduction

Understanding the relation between physical activities and the amount of calories burnt will help manage health and fitness in today's world. Precise prediction of the calories burnt during different activities helps individuals monitor their fitness, make informed choices, and improve their exercise regimen. This project will analyze the factors that influence calorie expenditure and develop a predictive model that will estimate the number of calories burned based on age, gender, weight, height, duration, heart rate, and body temperature. The dataset used is from a fitness center that collected data from their customers.

1.1 Objective

Develop and train a predictive model to estimate the number of calories burned based on the iden-

tified factors of a new member.

2 Materials & Methods

2.1 Dataset

The dataset used in this project is collected from a fitness center, which gathered data from its customers during their workout sessions. The dataset includes the following variables for each recorded instance:

- Age: The age of the individual in years.
- Gender: The gender of the individual (male/female).
- Weight: The weight of the individual in kilograms.
- Height: The height of the individual in centimeters.
- Duration: The duration of the physical activity in minutes.
- Heart Rate: The heart rate of the individual during the activity in beats per minute.
- Body Temperature: The body temperature of the individual during the activity in degrees Celsius.
- Calories Burned: The dependent variable, representing the number of calories burned during the activity.

2.2 Exploratory Data Analysis (EDA)

Most of the people in this analysis were between 30 and 50 years old, so the Age data was right-skewed, and no outliers were found in the Age data. The majority of people had a height of 174

cm, and the distribution of height was symmetric and slightly followed a normal distribution, although not statistically confirmed. No outliers were found in the Height data. However, outliers were identified in the Weight data, as some individuals were heavier, requiring the outliers to be addressed. Additionally, outliers were found in the Heart Rate, Body, and Calories data.

2.3 Data Preprocessing

The gender data was encoded for analysis, and during the EDA stage, the Duration data was identified as categorical, so the Duration column was one-hot encoded. Then, outliers were considered. Since weight and height were found to be highly positively correlated during the EDA process, the outlier entries were forecasted using linear regression, with height as the independent variable and weight as the dependent variable. The outliers in heart rate were replaced by fitting a line with calories, as both were highly correlated, ensuring that the outliers in calories burned were not used. Similarly, the outliers in calories were replaced by fitting a line with heart rate and calories burned, ensuring that only non-outlier heart rate data was used. The outliers in body temperature were replaced by fitting a line with calories burned, as both variables were found to be highly correlated.

2.4 Modeling

Data modeling was done using multiple linear regression and polynomial regression. The R-squared value was used to assess the accuracy and goodness of fit of the models. Multiple linear regression was used in modeling linear relationships between dependent and independent variables, while polynomial regression was applied in modeling any nonlinear trend that may appear in the data. The R-squared value showed the proportion of dependent variable variation explained by the different models, which would be helpful in choosing the best approach to make good predictions.

3 Results

It was found that the second-degree polynomial regression achieved an R-squared value of 0.997, while the multiple linear regression model attained an R-squared value of 0.975 when evaluated on the test data. These values indicate how well each model was able to explain the variability of the dependent variable in unseen data. Its R-squared for polynomial regression

stands higher than in the multiple linear regressions. It follows that this fit was superior and caught underpinning patterns much more appropriately compared to the multiple linear regression model. Therefore, polynomial regression will provide a better approach towards establishing confident predictions relating to the test data.

3.1 Project files

<https://github.com/SadaruwanSRI/Analyzing-and-Predicting-Calories-Burned.git>.

4 Discussion

Evaluation of the models on test data shows that second-degree polynomial regression achieved an R-squared value of 0.997 while the multiple linear regression model obtained an R-squared value of 0.975. It shows that both models did an excellent job of capturing the relationship between dependent and independent variables, but polynomial regression had a better fit than the former as its R-squared value is more significant.

This thus indicates that the superior performance of the polynomial regression model may suggest a non-purely linear relationship in some variables, and the addition of non-linear terms enabled it to grasp more complicated patterns within the data. Even though the multiple linear regression model was somewhat less accurate, it still yielded strong predictive performance and might be preferred when model parsimony and interpretability are a priority.

While the polynomial regression model showed a higher R-squared value, it is important to consider potential drawbacks such as overfitting, where the model captures noise rather than the true underlying trend. Therefore, additional validation techniques, such as cross-validation and residual analysis, should be conducted to ensure the model's generalizability.

Overall, the analysis demonstrates that polynomial regression is a more suitable choice for the given dataset when focusing on maximizing accuracy, but practical considerations such as computational complexity and interpretability should also be taken into account when selecting the final model.

Conclusions

The analysis showed that the second-degree polynomial regression model performed bet-

ter than the multiple linear regression model, achieving an R-squared value of 0.997 compared to 0.975 on the test data. This indicates that the polynomial regression model captured non-linear patterns more effectively, making it the better choice for accurate predictions.

The model was then deployed on a GUI to make it accessible. The GUI will allow the user to input values and get instant predictions, hence making it very user-friendly for practical applications of the model.

Further improvements, such as additional validation and optimization, can be done to enhance the performance and reliability of the model.

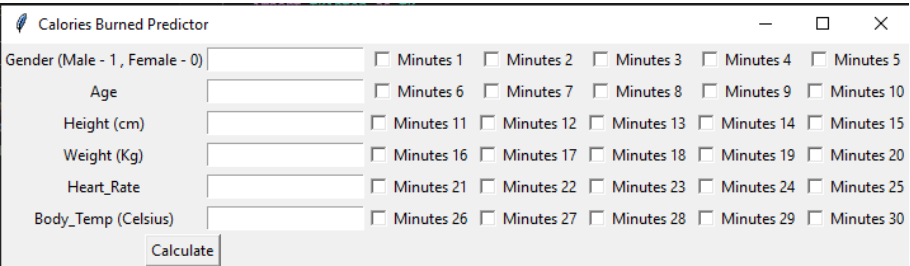


Figure 1: GUI of the program

References

[1] <https://www.geeksforgeeks.org/machine-learning-projects-using-regression/#3-calories-burnt-prediction>