

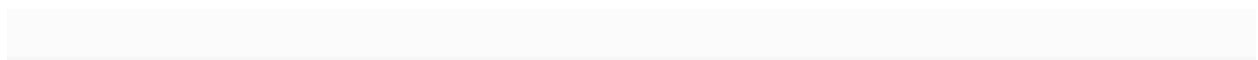
Final Project Report
Introduction to Data Analytics

Project Title:
Prediction/Analysis of
Daily Calories Burned using FitBit Device

Prepared by:



ITE 5201 – Winter 2022
Humber College



1. Problem Statement

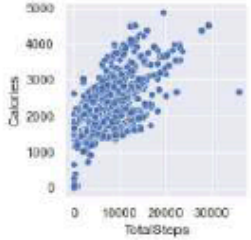
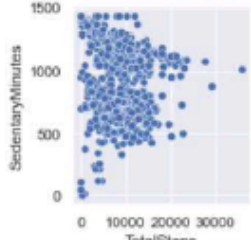
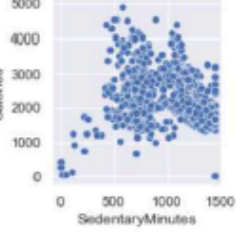
⇒ Prediction/Analysis of Daily Calories Burned using FitBit Device

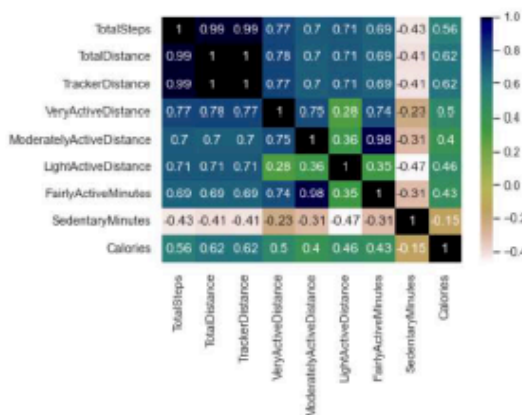
2. Dataset Description

- ⇒ The Daily Activity using FitBit device Dataset contains several dependent information of user's activity which is tracked by the FitBit device.
- ⇒ Dependent variables such as TotalSteps, TotalDistance, TrackerDistance, VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance, FairlyActiveMinutes, SedentaryMinutes which I used to train the model for predicting burned calories.

3. Dataset Analysis and Observations

- ⇒ For dataset analysis, I used pairplot for univariate and bivariate analysis and Heatmap for finding Correlation Coefficient Rank using the above listed columns (few column's visualize below).
- ⇒ Observation:

	TotalSteps Vs Calories	TotalSteps Vs SedentaryMinutes	SedentaryMinutes Vs Calories
⇒ Pairplot			
⇒ Direction:	Positive	Negative	Nor Positive neither Negative
⇒ Form:	Linear	Linear	Linear
⇒ Strength:	Strong	Strong	Strong
⇒ Outlier:	Yes	Yes	Yes



This heatmap plot using spearman method which denotes relationship between two data variables and returns its Correlation Coefficient rank (R)

- ⇒ R = 1 Strong Positive relationship
- ⇒ R = 0 Not linearly correlated
- ⇒ R = -1 Strong negative relationship

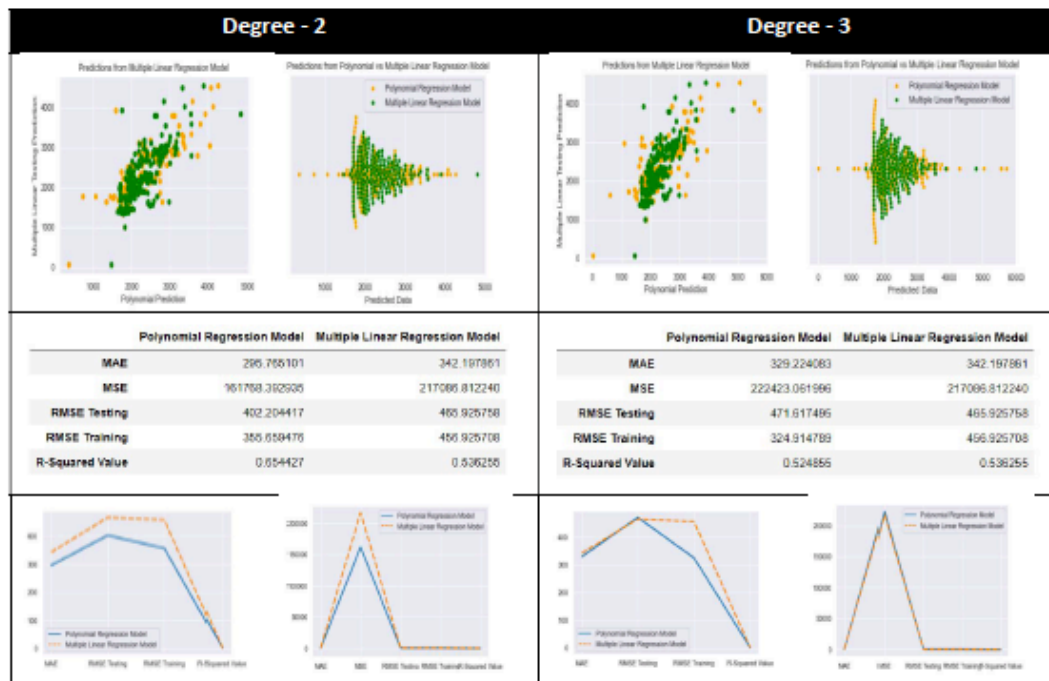
- ⇒ From this heatmap, I conclude that Calories Strongly Positively correlated with TotalSteps, TotalDistance, TrackerDistance, On the contrary, Calories Strongly Negative correlated with SedentaryMinutes.
- ⇒ Factors like TotalSteps, TotalDistance, VeryActiveDistance contribute a major role in burning daily calories.

4. Proposed Analytical/Prediction Model

- ⇒ As above analysis, the dataset has dependent variables for calculating burned calories.
- ⇒ That's why, I trained Polynomial and Multiple Linear Regression Model for this dataset.
- ⇒ For both models, I trained dataset in degree 2 and 3, then compare both model and its retrieve error index as well as R-Squared Value for better prediction.

5. Results and Discussions

- ⇒ In the 1st row, the first plot is scatter plot for Polynomial Vs Multiple Linear Regression Model, the exact next is swarmplot for the same for better comparison.
 - From the plot, I conclude that Polynomial model gives better predicted data and less outliers as compared to Multiple Linear Regression Model for both degrees.
- ⇒ Then in the 2nd row, you can find the MAE, MSE, RMSE for testing and training, R-Squared value for both models.
 - In both degrees result, I conclude that Polynomial Regression model has less error numbers in (MAE, MSE, RMSE) and high R-squared value as compared to Multiple Linear Regression Model.
- ⇒ In the last row, there is visualization of 2nd row's data using lineplot.



Conclusion:

- ⇒ The lowest MAE, MSE, RMSE and the highest R-Squared Value ⇒ Minimal error or difference found between original and predicted data.
- ⇒ In addition, there is MORE DATA ACCURACY found in predicted data from the Polynomial Regression Model
- ⇒ So, I can say that "Polynomial Regression Model" is better for this dataset for evaluation