COMPARISION BETWEEN LINEAR AND POLYNOMIAL CURVES FOR MACHINE LEARNING

An investigatory course project in the subject of Machine Learning (IC3023) undertaken by

- Q-6 Aneesh Poduval
- Q-16 Sarthak Chudgar
- Q-37 Johnathan Fernandes

Under the guidance of Prof. Archana Chaudhari

What is Machine Learning?

A subset of Artificial Intelligence, Machine Learning is the process of programming a statistical model to either predict of classify data given to it, without any explicit instructions. It is used where programming explicit instructions cannot be done, such as the weather forecast, computer vision, and even keyboard word prediction.

One such application of Machine Learning is mood prediction, which we have chosen to implement in or project.

Mood Prediction using Machine Learning:

There is a long standing belief that exercise affects the mood of a person in a positive way. Scientific research has confirmed that any amount of medium-high intensity physical exercise will heighten the amount of Serotonin in the human body. Serotonin, like Dopamine plays a large role in mood and helps humans feel pleasure, satisfaction and motivation.

Objective of this course project:

To compare the effectiveness of a linear (1-degree) and non-linear equation to plot the best fit curve for a given set of data.

The aim of the model itself is to predict mood of a person based on amount of physical exercise done in a day.

Dataset Used:

The independent feature is calories utilized, which is the end result of physical activity for the human body. This data has been collected using “Google Fit”,

which is a popular fitness tracking app that tracks it's user exercise and calculates calories used.

The feature we aim to track, "mood" is a measure of the general happiness of the test subject on a scale of 1-10. This data was manually entered by the user.

This data was collected every day over a period of 2 months from 25 August to 25 October 2019.

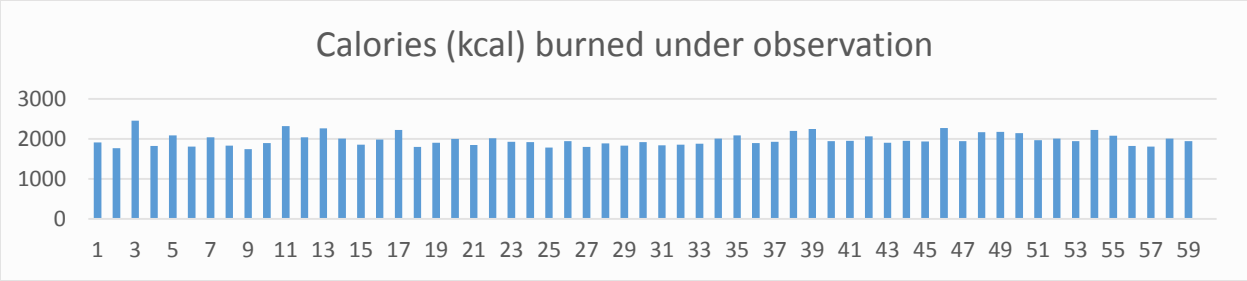

Fig. 1: Calories (kcal) burned under observation

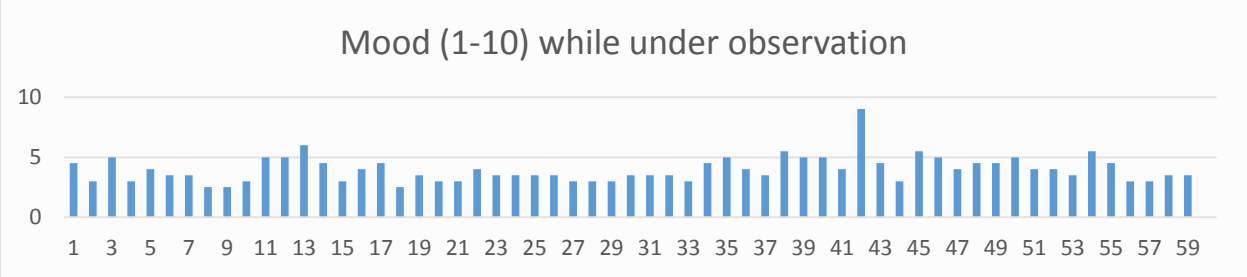

Fig. 2: Mood (1-10) while under observation

Quantifying effectiveness of curves: Goodness- of-fit & Residuals

The goodness of fit test involve multiple parameters which gives us an idea of the amount of error in our equation. These parameters are

- Sum of Squares due to Error
- R-Square & Adjusted R-Square
- Root Mean Squared Error

We can also use residuals to obtain a clear visual understanding of the deviation (error) from the curve.

Visualizing the data: MATLAB Curve Fitting Toolbox

We have utilized the built in MATLAB Curve Fitting Toolbox to plot the data along with residuals. This toolbox also calculates the goodness-of-fit variables along with the equation of the line.

The upcoming diagrams are mood plotted against calories, with the order of the equation used increasing iteratively.

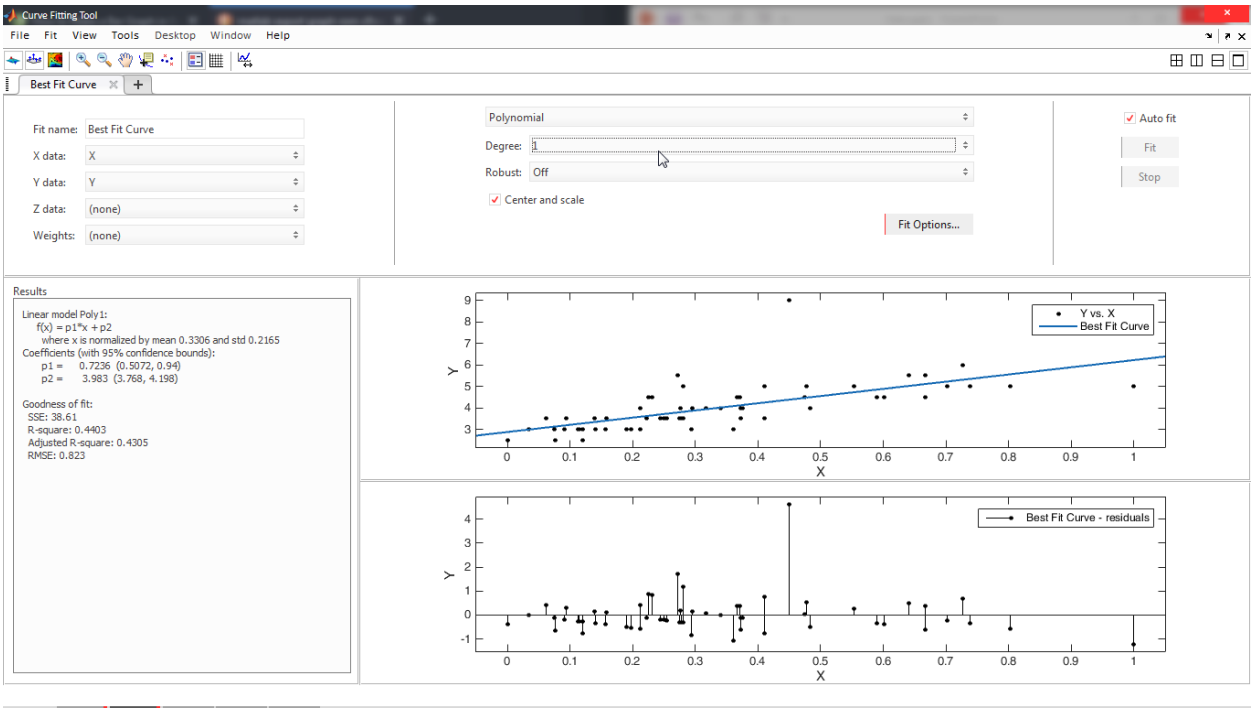

Fig. 3.1: Degree 1 (linear) curve

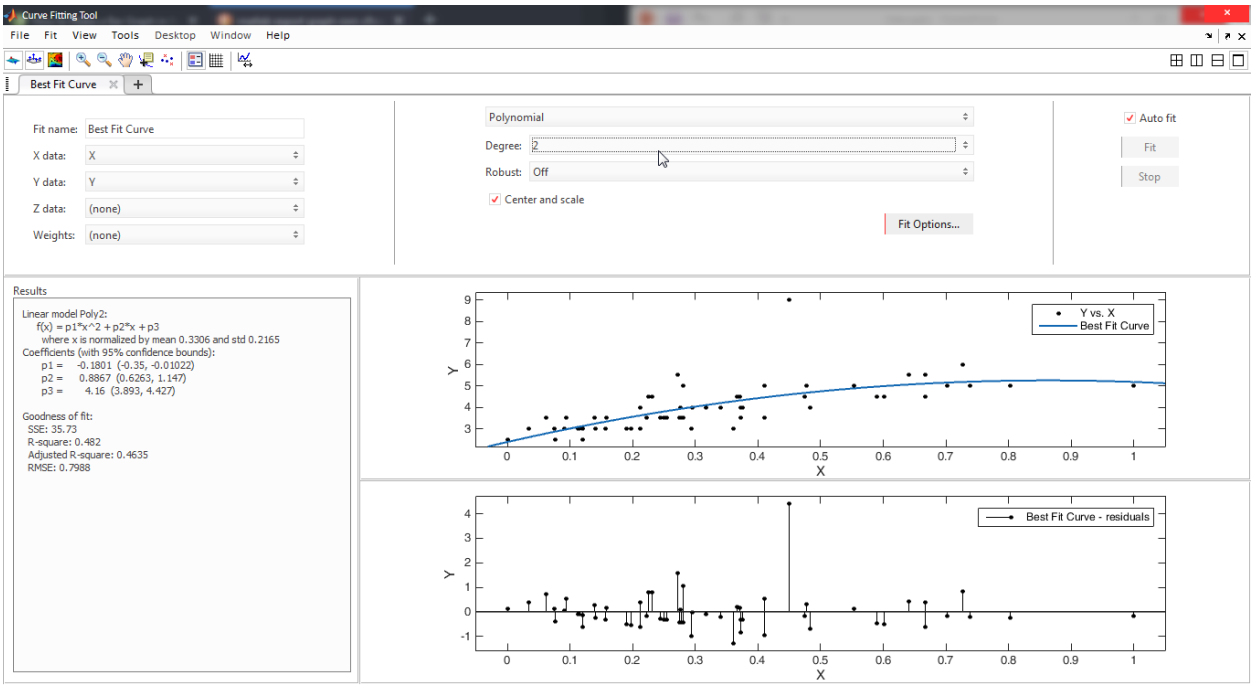

Fig. 3.2: Degree 2 curve

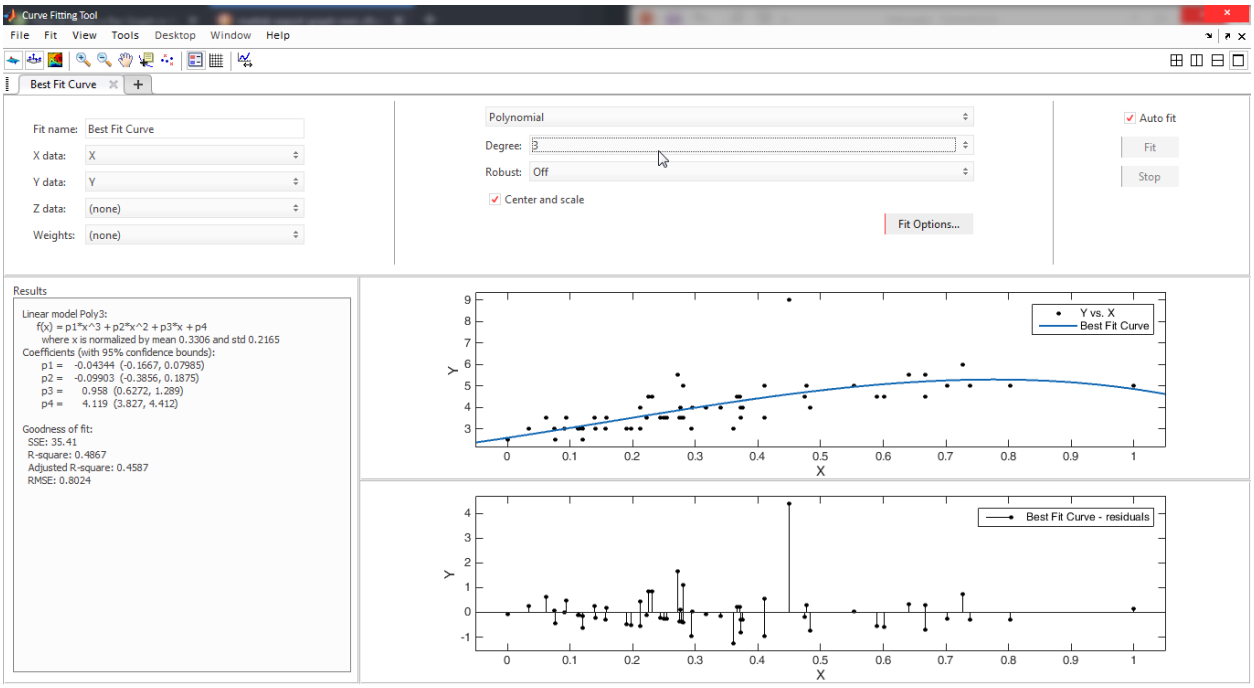

Fig. 3.3: Degree 3 curve

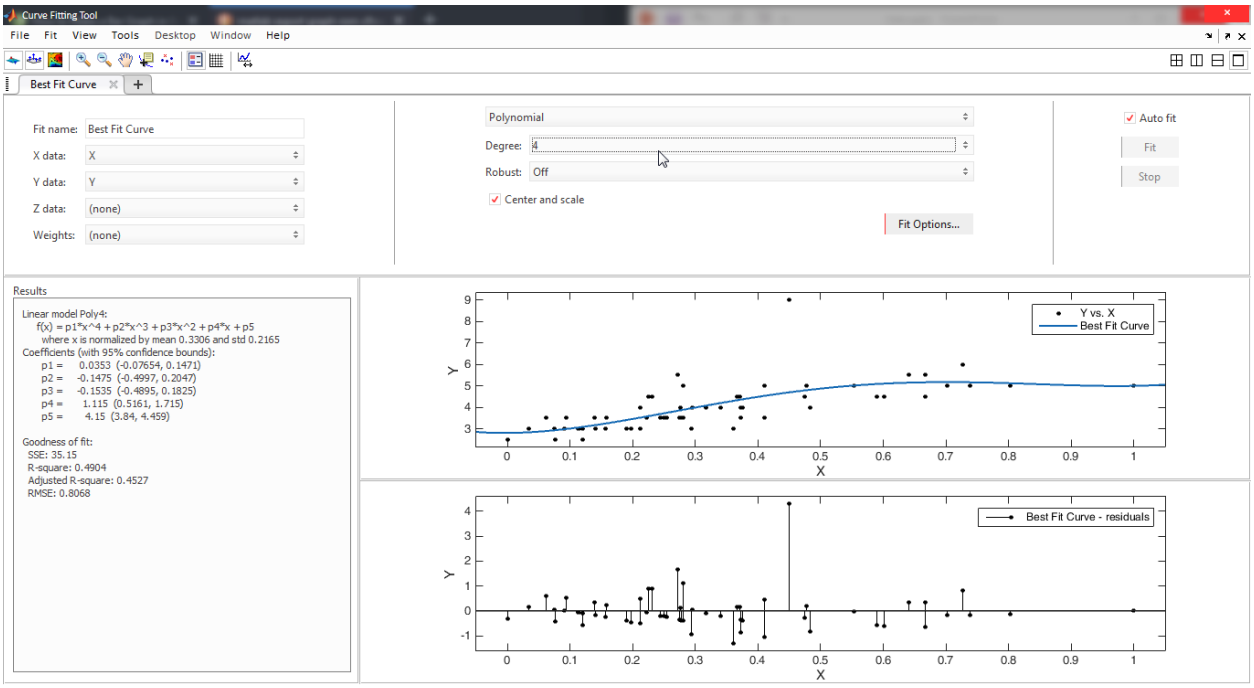

Fig. 3.4: Degree 4 curve

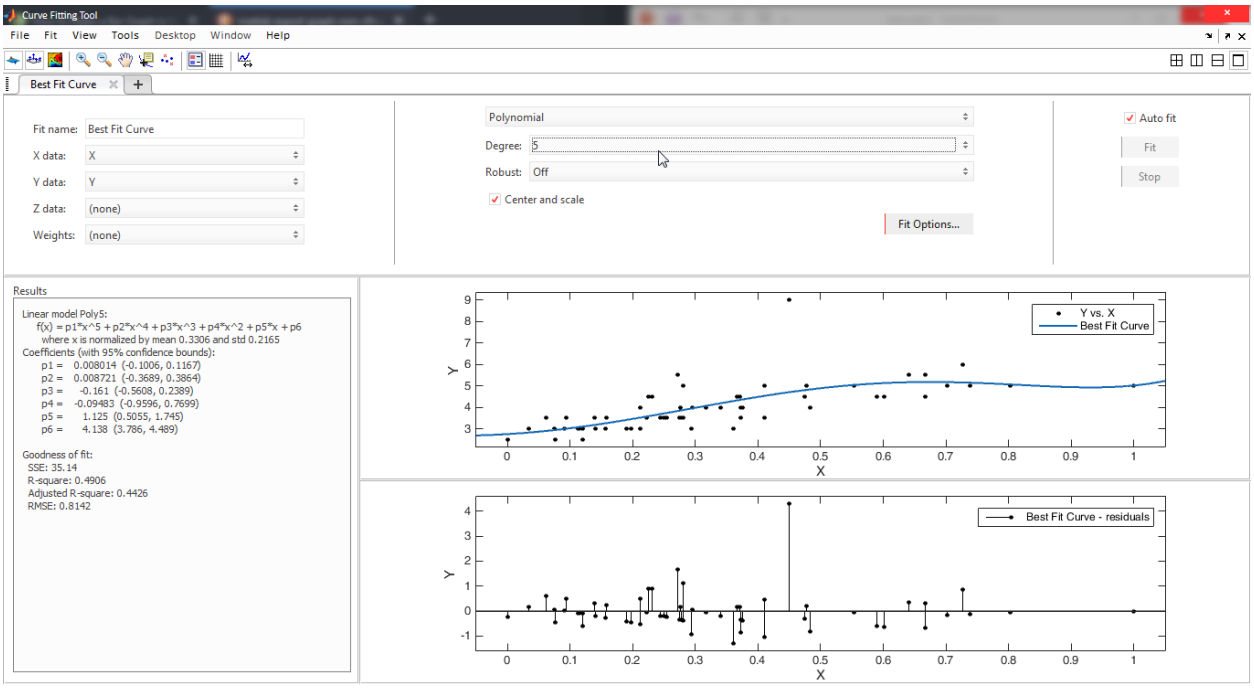

Fig. 3.5: Degree 5 curve

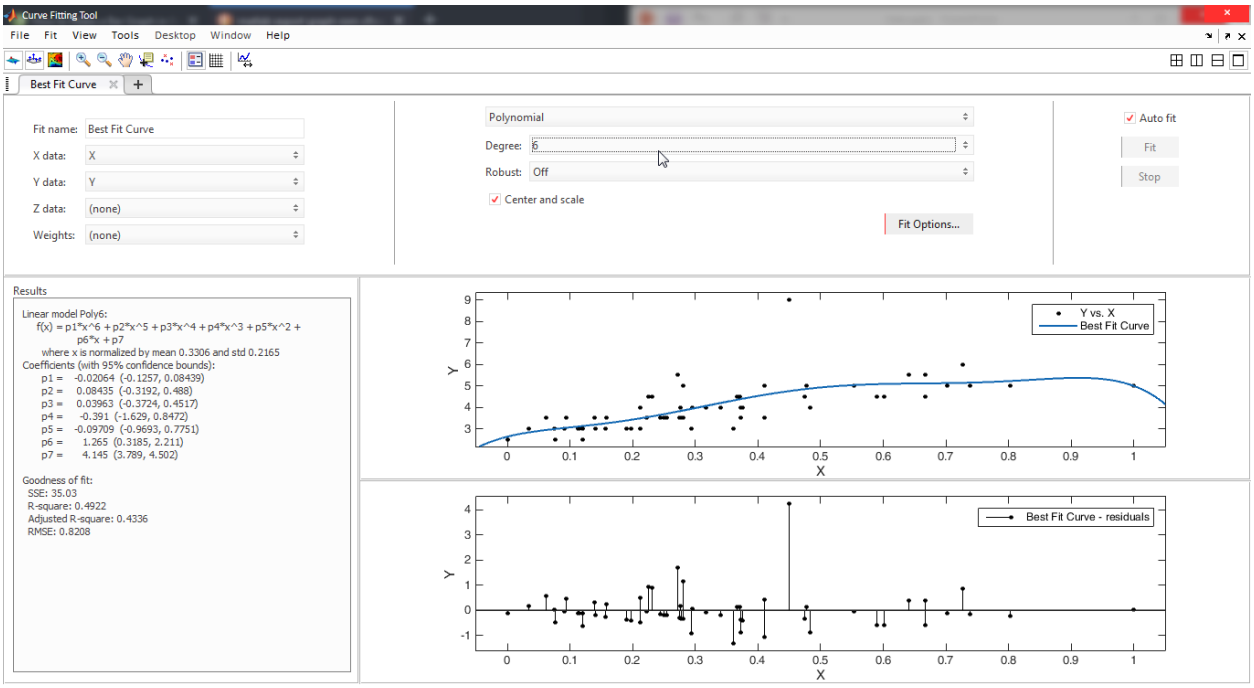

Fig. 3.6: Degree 6 curve

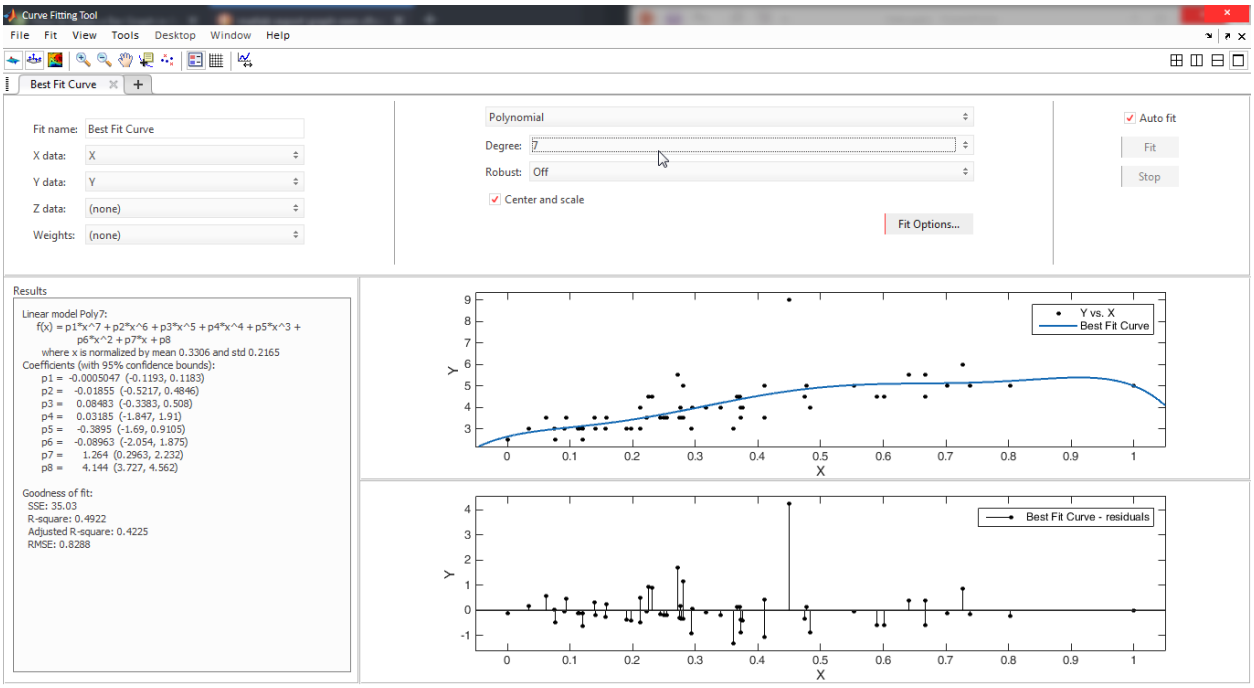

Fig. 3.7: Degree 7 curve

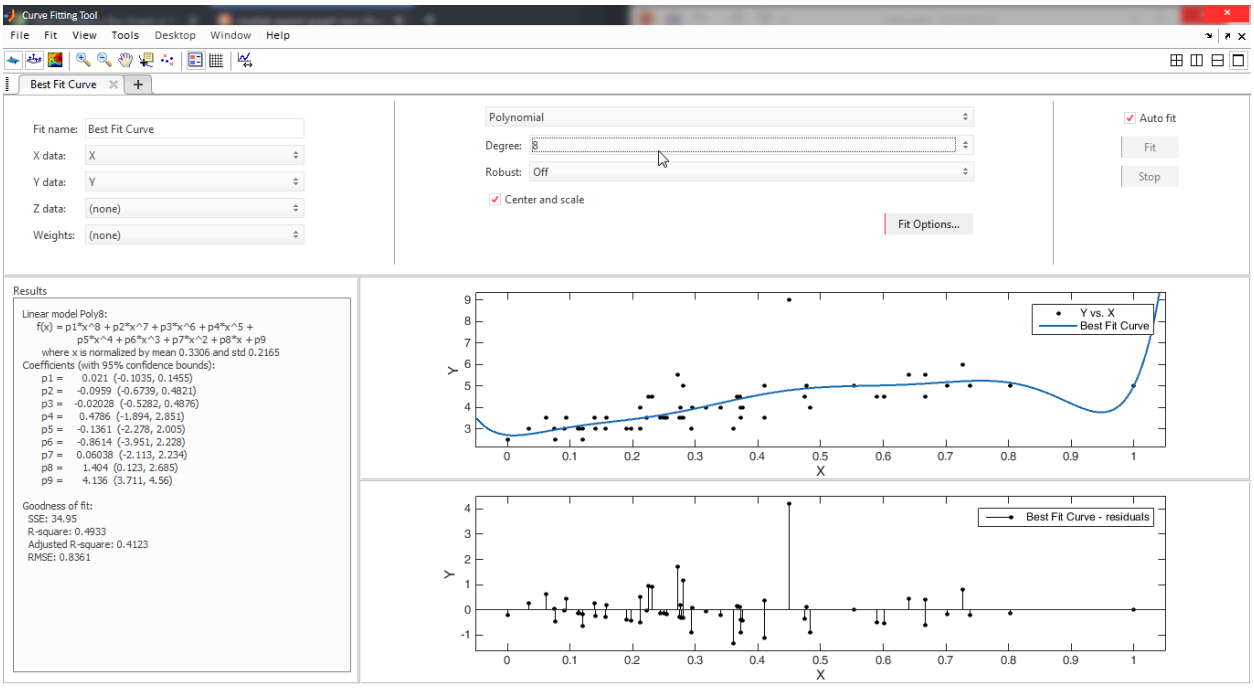

Fig. 3.8: Degree 8 curve

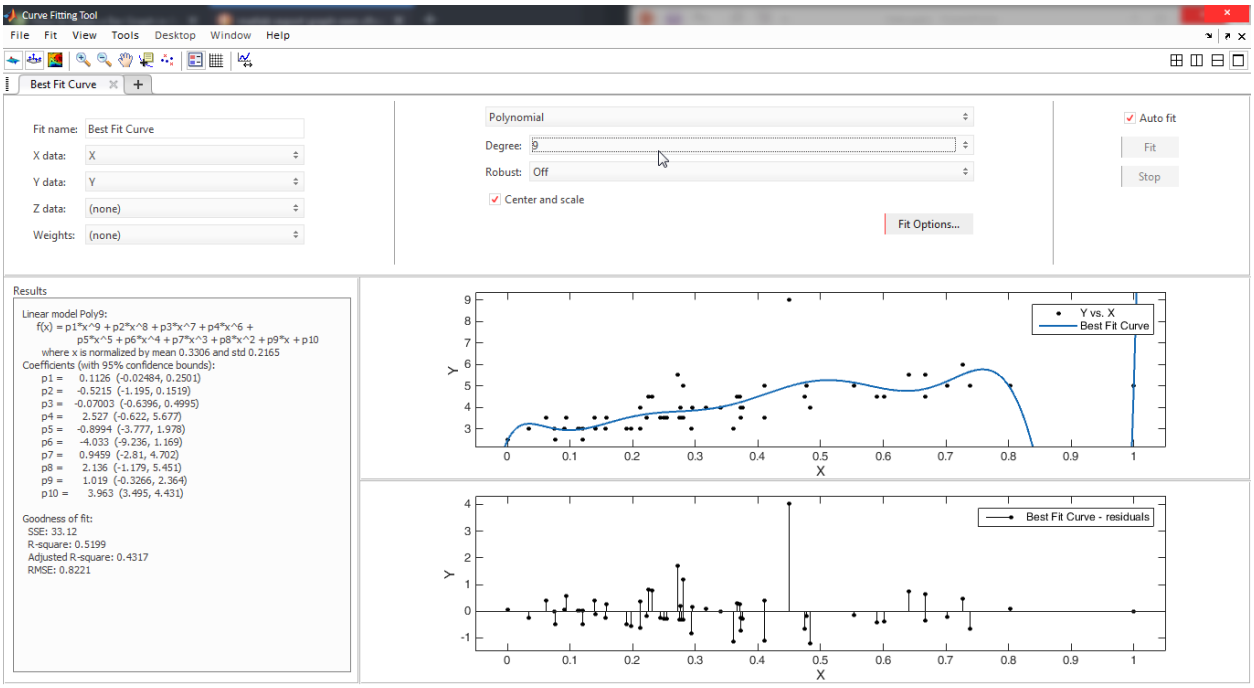

Fig. 3.9: Degree 9 curve

While there are other functions to plot curves (e.g. fitlm(), polyfit(), etc.) we use cftool() since it provides the equation, confidence bounds, residuals, and goodness of fit variables all alongside a GUI.

Observations:

Order	SSE	RMSE	R-Square	Adjusted R-Square
1	38.61	0.823	0.4403	0.4305
2	35.73	0.7988	0.482	0.4635
3	35.41	0.8024	0.4867	0.4867
4	35.15	0.8068	0.4904	0.4527
5	35.14	0.8142	0.4906	0.4426
6	35.03	0.8208	0.4922	0.4336
7	35.03	0.8288	0.4922	0.4225
8	34.95	0.8361	0.4933	0.4123
9	33.12	0.8221	0.5199	0.4317

Fig. 4: Table of Order, SSE, RMSE, R-squared and Adjusted R-Squared

Using the tool, we observe that with each iteration, the curve seems to fit better i.e. “touch” more points or get closer to them. The residuals also seem to be decreasing at various points.

We can further confirm this by plotting and comparing the values of SSE, R-Squared and RMSE graphically:

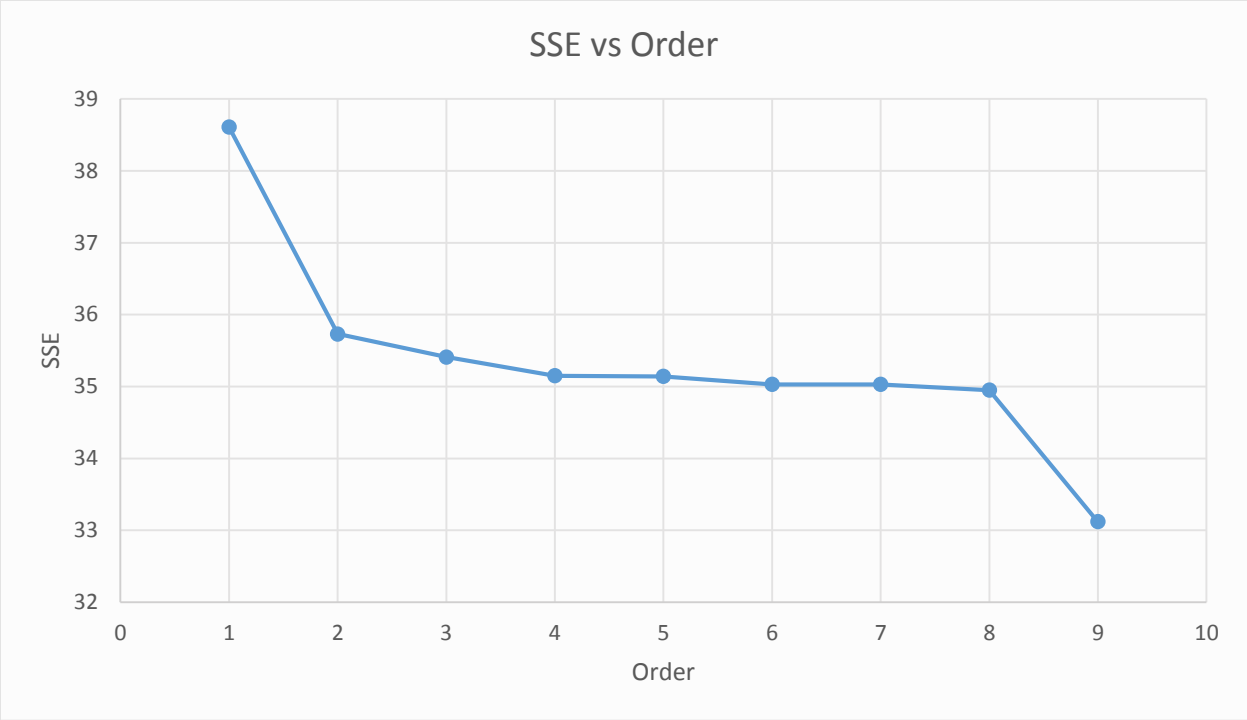

Fig. 5.1: SSE vs Order

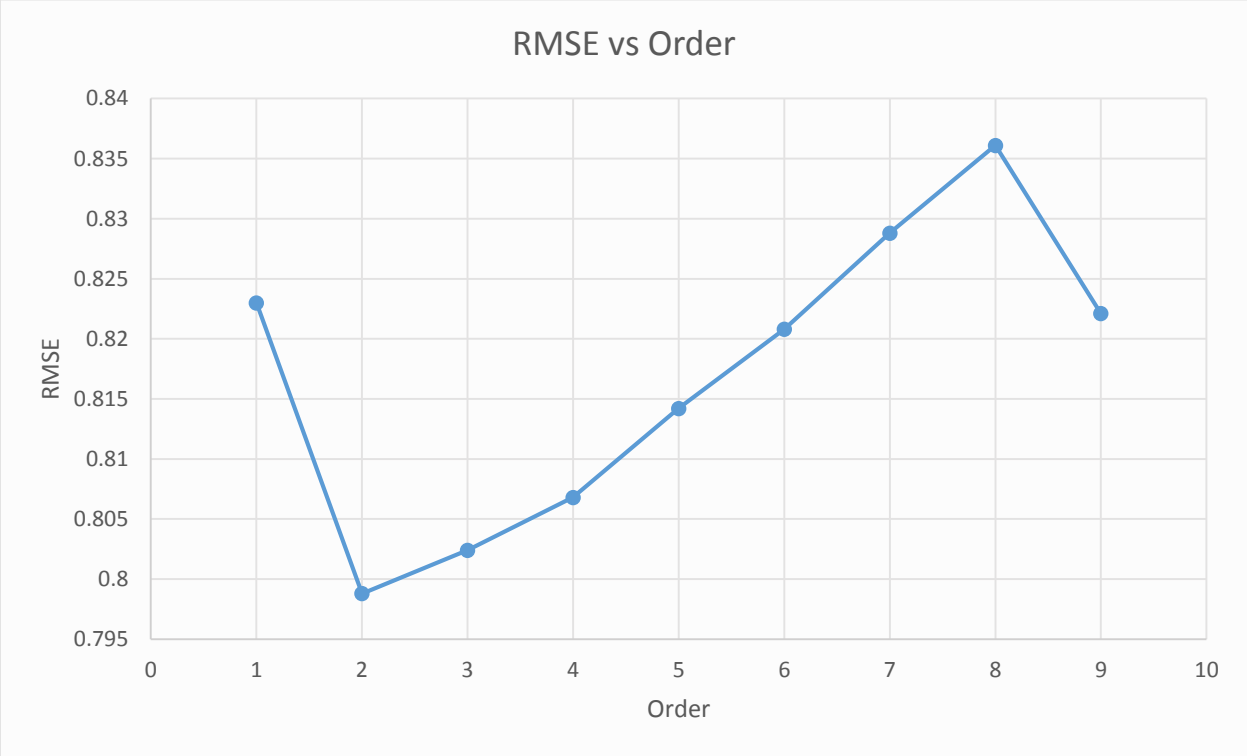

Fig. 5.2: RMSE vs Order

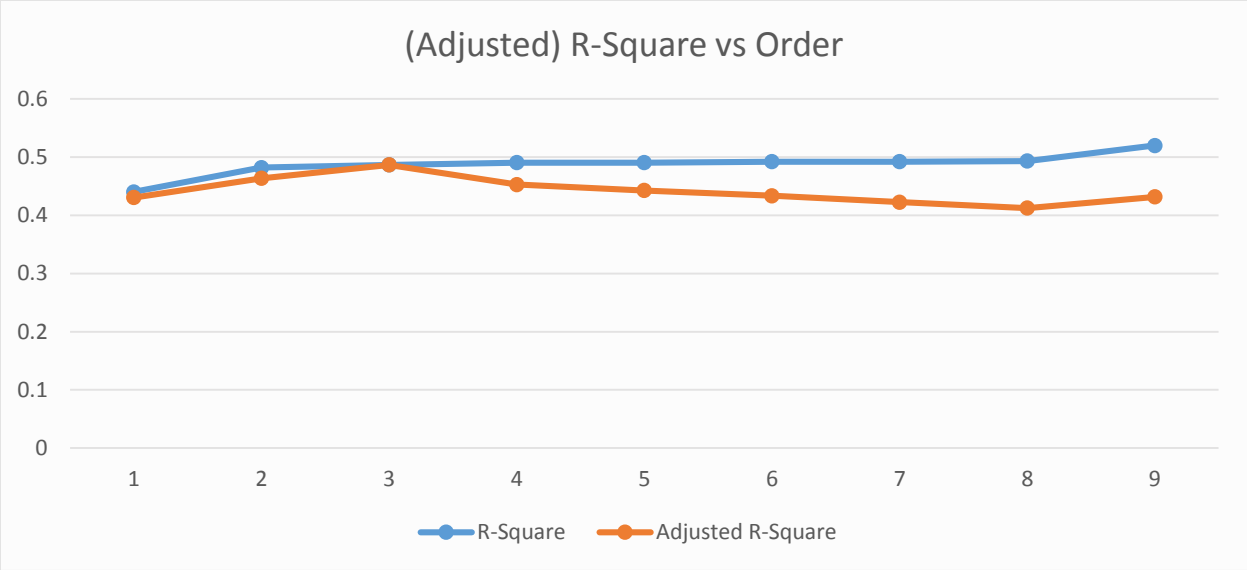

Fig. 5.3: (Adjusted) R-Squared vs Order

Observing each graph, we notice the following:

- SSE reduces drastically as order changes from 1 to 2, then reduces slowly until order 9, where is reduces drastically again.
- RMSE initially decreases when order changes from 1 to 2, then steadily increases until order 9, where it is almost identical to that of order 6.
- R-Square increases sharply at order 1 to 2 and again at 8 to 9, steadily at all other parts. Adjusted R- Square decreases as order goes from 3 to 8, then increases again.

The sharp decrease in errors when going from first to second order equations tell us that the addition of even a single degree to the equation increases the accuracy of the model by a relatively large amount. By using a higher order polynomial equation, we get marginally better performance from the model. This is attributed to the fact that a linear model is very prone to underfitting.

While the intense decrease in error as we use a 9^{th} order polynomial seems desirable, this actually indicates the issue of a special case of overfitting, where due to the high complexity of the equation, the model attempts to “touch every point”. While the calculated errors might be low, this is a problem in real world applications of this model and hence care should be taken to not use high order polynomials.

Using a 4th degree polynomial, we train the model on a separate set of test data, and obtain these results:

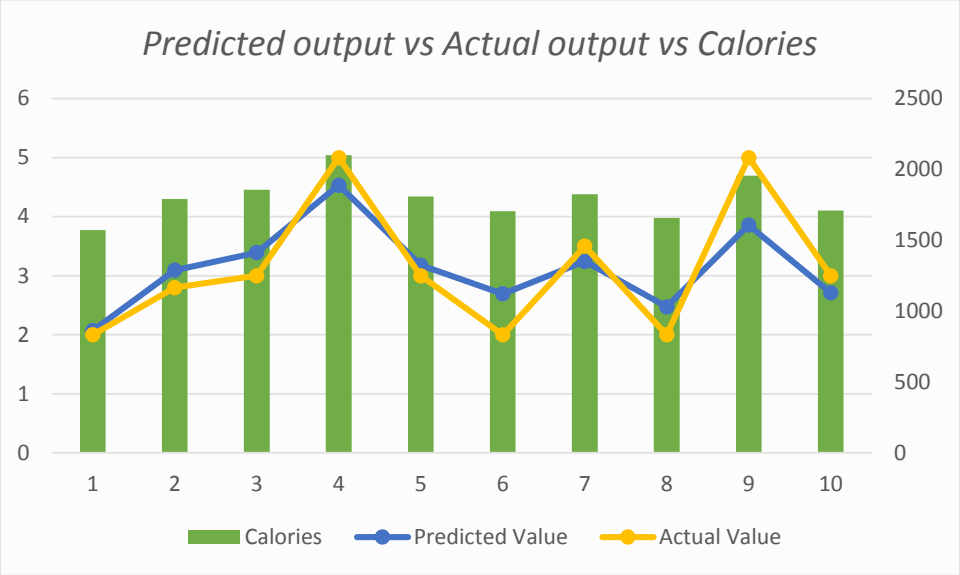

Fig. 6: Predicted output vs Actual output vs Calories

Conclusion:

Through this investigatory course project, we have determined that a linear equation is rather ineffective due to underfitting, while a high order polynomial equation can be prone to overfitting.

Hence, for optimal results, we should use an equation of 3rd to 5th order, along with suitable regularization algorithms for prediction models.

References:

https://www.apa.org/monitor/2011/12/exercise

https://www.academia.edu/10809175/Effect_of_Exercise_on_Mood_and_Self-Esteem_A_Journal_Article_Review

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1424736/

https://connect.uclahealth.org/2018/10/17/the-link-between-exercise-and-mental-health/

www.mathworks.com

https://www.livestrong.com/article/22590-effects-exercise-serotonin-levels/