## Placeholder image

Combinatorial Feature Selection (nCr)

Data Visualization, EDA and Plot metrics by: Mustafa Sajjad , Asif Shaikh.  
Documentation ,Plot Analysis, Plot and Model Metrics Analysis by: Tahir Mehmood, Salman Rahim, Muhammad Ghazi, Umair Tamim, Bilal Khalid  
Feature Selection Combinatorics Algorithm, Model fitting and Model Metrics by: Abdur Rehman Tariq.

# Overview

* Basic concept: use combos (nCr) and find the combo that produces the best results and use that combo in our model.
* We will use encoding here to simplify the problem since the classification target only has 3 distinct types.
* We assign an integer to each type using encoding therefore no need for logistic regression.
* If each classification or regression based feature is not linearly related or produces a subpar r^2 score, we use polynomial regression. As for continuous data, there is no need for encoding since it is already numeric.
* In addition, we will use imputing and fill any missing values with the mean although this dataset is mostly cleared and filtered. Speaking of which, when using regression, we filter out all non numeric columns for efficiency. In this way this feature selection algorithm makes it possible to fit a model for any classification or regression task.
* The code is split into 3 files each showcasing the various regression/classification types we are using to predict a target column.
* NOTE: This data set has mostly numeric data types save for Species. This is why filtering non numeric data while making combos does not affect the model score and also because we are using basic regression types only. However, the code has an imputing function since it was designed to work with most datasets.

# Goals

1. Make a feature selection algorithm that can be used for any regression task and even a bit of classification.
2. Visualize results to showcase accuracy of model fit along with providing metrics to back it up.

# Dataset Description

## iris.csv

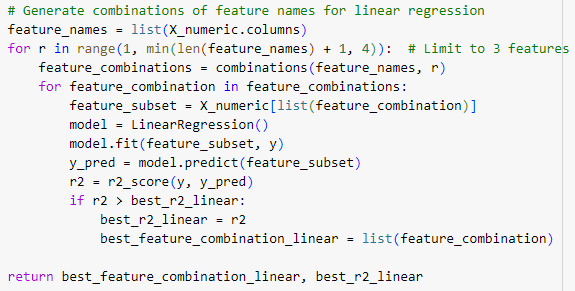
It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

The columns in this dataset are:

* **Id (int)**
* **SepalLengthCm (int)**
* **SepalWidthCm (int)**
* **PetalLengthCm (int)**
* **PetalWidthCm (int)**
* **Species (string 3 distinct types)**

# Linear Regression

Below is the code to perform linear regression. We are using numeric columns for linear regression and therefore filtered all numeric columns first. Then we are creating combinations of various columns of the data set. The Next step is to compare the results of these combinations in the loop then exit it. Finally based on the comparison results, the best r2 value is being stored in variable “”best\_r2\_linear”.

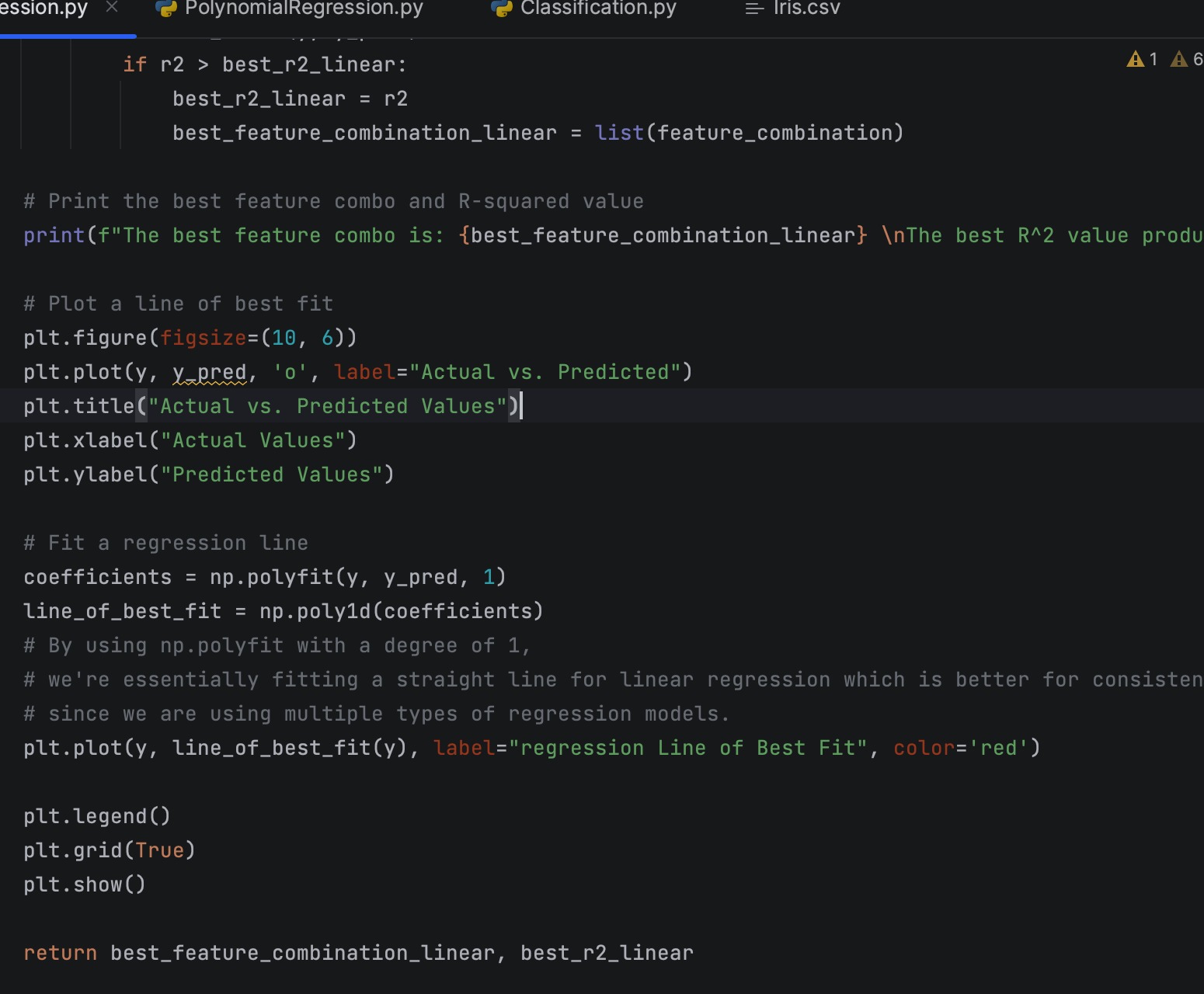


## 

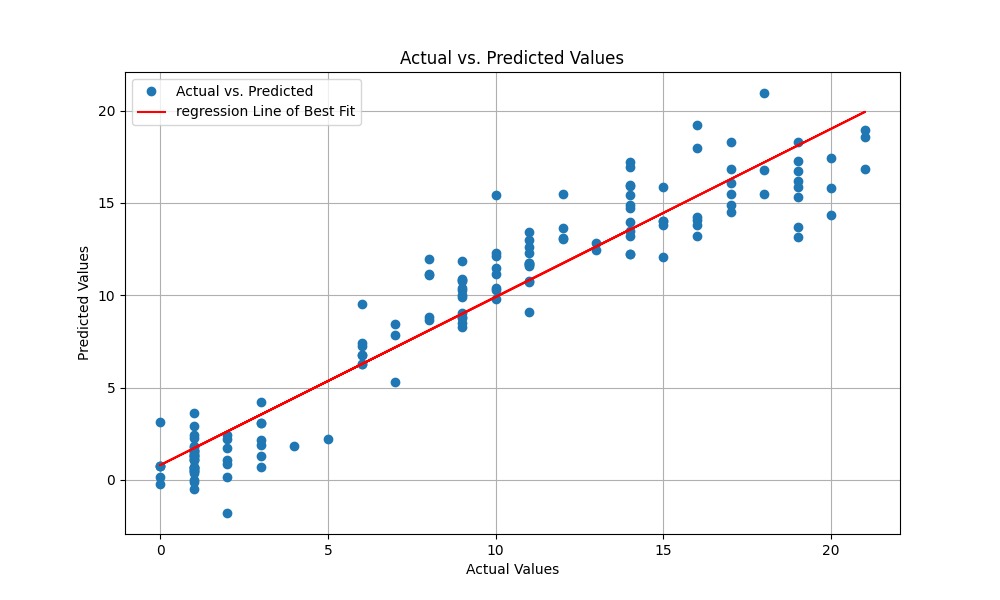
## 

## Model Fit Plot(Linear Regression)

The below image shows the code used for the plot while the image further below shows the actual plot.

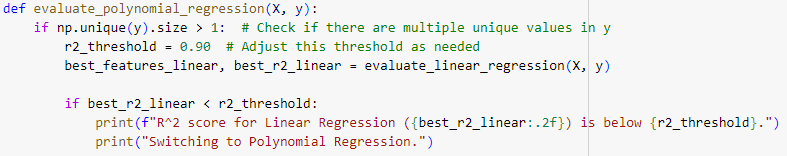


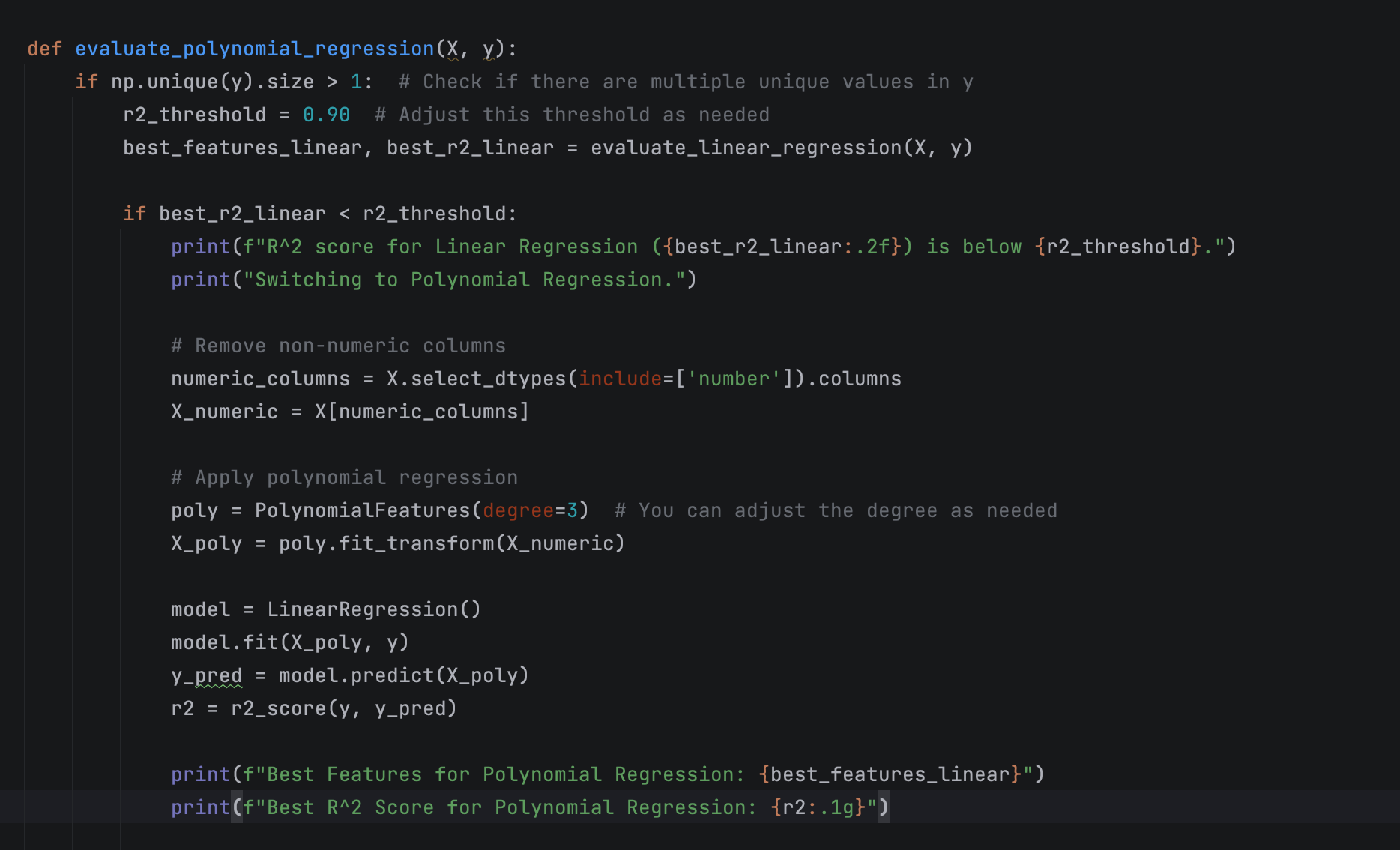
As can be seen in the scatter plot below, the actual versus predicted y axis values despite having a linear relationship have produced some outliers. The regression line of best fit indicates that the relationship is linear and hence it evidences that the ‘y’ predicted variables match up with the actual y-variable. Furthermore, this can be further backed up with the model score/accuracy (R^2 score) which is 91.1% (that is 90% when rounded to two significant figures).

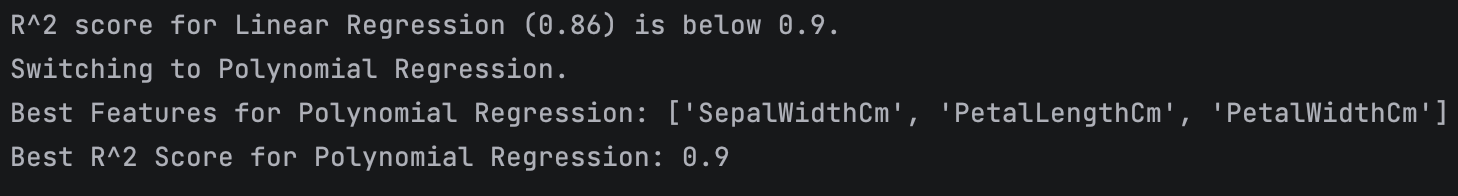


# Polynomial Regression

A function to calculate polynomial regression is defined in code as shown in below code. We have defined a threshold value of r2. Next we compared the value of r2 we calculated in linear regression with the threshold and if that value is less than threshold, we switch to Polynomial regression to see the results.



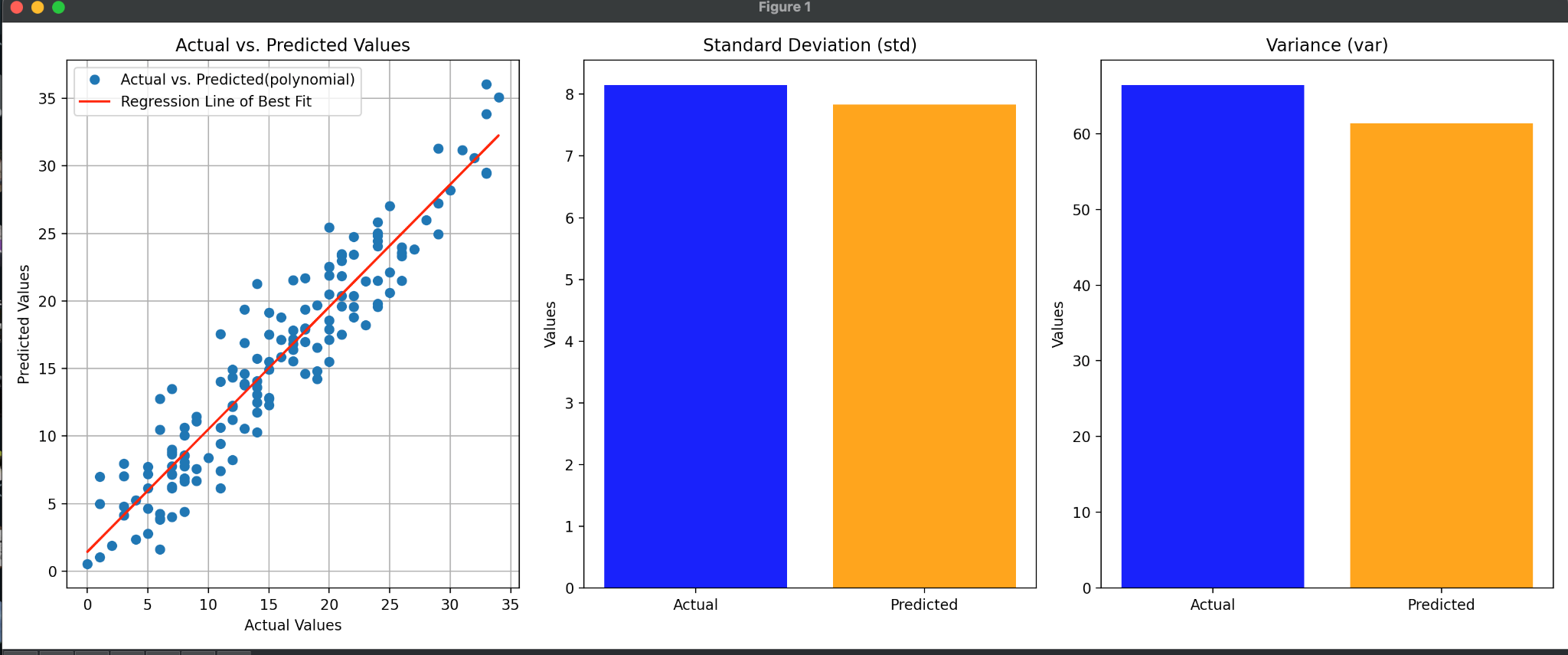
Below section of code reflects application of polynomial regression with a given degree value equal to 3. Best features of the polynomial regression and the r2 score is being calculated through this code and compared with the linear regression best features of linear regression and r2 score.

We can further back up these claims with metrics for the model fit / accuracy given in the image below:  


As you can see, the initial output does not meet the threshold for model accuracy so we try using polynomial regression for a small increase in model accuracy.

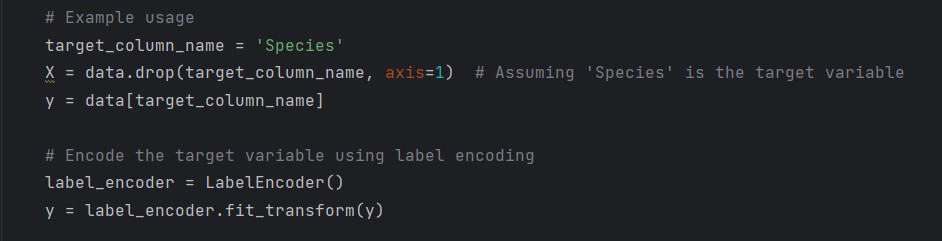
## Model Fit Plot(Polynomial Regression)

As you can see in the scatter plot below the, actual vs predicted y axis values despite having a linear relationship have produced some outliers. The regression line of best fit indicates that the relationship could be improved ever so slightly if it was non linear (again due to outliers) this is why using polynomial regression gives us a small but welcome boost in model accuracy (from 86% to 90%) to meet our threshold. In addition, comparing the standard deviation and variance of the actual and predicted values (then visualizing them) shows that our model is working to produce good results with minimal difference..



# Classification Through Linear/Polynomial Regression via Encoding

The dataset used contains a column ‘Species’. For the classification approach, we are encoding the ‘Species’ (i.e. converting it into a numeric value) to enable it to be used for linear regression. This is to maintain consistency.



## Model Fit Plot(Classification through Logistic Regression)

The regression line of best fit indicates that the relationship is linear and hence it evidences that the predicted ‘y’ variables are directly proportional to the actual y-variables. Furthermore, this can be further backed up with the classification score/accuracy (R^2 score) which is 95.6% (that is 96% when rounded to two significant figures). Since the feature to predict was categorical, we will use a categorical plot (or just cat plot). The below plot shows a linear relationship with concentrated categorical values for each type of Species (there are 3 distinct types)

