Assignment 2

Before working on this assignment, please read the following tutorial carefully –

1. Overfitting vs. Underfitting
2. Linear Regression in Python

In this assignment, you will use the Life Expectancy Data (from last assignment) to investigate regression problems. In case you don't have this data, you can access it via: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Task 1:

Int this task, we will use different kinds of models to explore the relationships between economic status and life expectancy. For Afghanistan for instance, as the following table shows, we can use older data (from 2000 to 2013) to train models and use the trained models to predict life expectancy of 2014 and 2015. The model input can be GDP number and the model output will be life expectancy for that year.

Country	Year	GDP	Life expectancy
Afghanistan	2015	584.2592	???
Afghanistan	2014	612.6965	???
Afghanistan	2013	631.745	59.9
Afghanistan	2012	669.959	59.5
Afghanistan	2011	63.53723	59.2
Afghanistan	2010	553.3289	58.8
Afghanistan	2009	445.8933	58.6
Afghanistan	2008	373.3611	58.1
Afghanistan	2007	369.8358	57.5
Afghanistan	2006	272.5638	57.3
Afghanistan	2005	25.29413	57.3
Afghanistan	2004	219.1414	57
Afghanistan	2003	198.7285	56.7
Afghanistan	2002	187.846	56.2
Afghanistan	2001	117.497	55.3
Afghanistan	2000	114.56	54.8

Please train 4 functions, Linear Function, Quadratic Function, Cubic Function, and Quartic Function, to fit this data (only using Afghanistan data), and then calculate RMSE and R2 scores. (You can call the functions from sklearn.metrics) Please fill the following table:

RMSE Scores (for 2014 and 2015 data):

	Linear Function	Quadratic Function	Cubic Function	Quartic Function
Training Data (2000 - 2013)				
Testing Data (2014 and 2015)				

R2 Scores (for 2014 and 2015 data):

	Linear Function	Quadratic Function	Cubic Function	Quartic Function
Training Data (2000 - 2013)				
Testing Data (2014 and 2015)				

Please submit your code (named *calculate_Afghanistan.py*). Please explain which model can be the best to predict this small dataset? why?:

_____

_____

_____

Task 2:

Please repeat this process for all the countries in this dataset. Then, you can average the RMSE and R2 scores for all the developing and developed countries. Please fill the following table:

RMSE Scores:

Developing Countries	Linear Function	Quadratic Function	Cubic Function	Quartic Function
Training Data (2000 - 2013)				
Testing Data (2014 and 2015)				
Developed Countries	**Linear Function**	**Quadratic Function**	**Cubic Function**	**Quartic Function**
Training Data (2000 - 2013)				
Testing Data (2014 and 2015)				

R2 Scores:

Developing Countries	Linear Function	Quadratic Function	Cubic Function	Quartic Function
Training Data (2000 - 2013)				
Testing Data (2014 and 2015)				

Developed Countries	Linear Function	Quadratic Function	Cubic Function	Quartic Function
Training Data (2000 - 2013)				
Testing Data (2014 and 2015)				

Please submit your code (named *calculate_all_country.py*). Please explain which model(s) can be the best to predict developing and developed countries; why?:

_____

_____

_____

Task 3:

For this task, we will use 5 variables - Adult Mortality, Alcohol, BMI, GDP, Schooling – to build regression models (Multiple Linear Regression) to predict the life expectancy of the target country for a specific year, e.g., use a model to predict "*Libya's life expectancy in year 2010*". We can train two different models (developing country model and developed country model) to predict the data. Similarly, we can use older data (from 2000 to 2013) to train models and use the trained models to predict life expectancy of 2014 and 2015.

Please fill this table (for testing with 2014 and 2015 data):

	RMSE	R2
Developing Country		
Developed Country		

Please fill the following table with the "regression coefficients" (for each variable):

	Adult Mortality	Alcohol	BMI	GDP	Schooling
Developing Country					
Developed Country					

Please submit your code (named *calculate_regression.py*). Comparing developing and developed countries (two models that you build), can you find some interesting results?:

_____

_____

_____

Task 4:

For task 3, we used the Linear Regression model to address the prediction problem. Please tell us the limitation(s) of the model, and can you improve it?

_____

_____

_____

Submission: Your code files and a PDF file (containing your solution for tasks and the results to report).