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. Please fill the following table:

RMSE Scores (for 2014 and 2015 data):

	Linear Function	**Quadratic Function**	**Cubic Function**	**Quartic Function**
Training Data (2000 - 2013)	1.067	0.989	0.872	0.833
Testing Data (2014 and 2015)	4.269	4.151	3.822	4.193

R2 Scores (for 2014 and 2015 data):

	Linear Function	**Quadratic Function**	**Cubic Function**	**Quartic Function**
Training Data (2000 - 2013)	0.479	0.552	0.652	0.683
Testing Data (2014 and 2015)	-1.803	-1.650	-1.247	-1.704

Please submit your code (named *calculate_Afghanistan.py*). Please explain which model can be the best to predict this small dataset? why?: **The model that can be used to predict this model the best would be the cubic function (degree = 3). This is because it has the lowest RMSE and the highest R2 value which is an indicator of accuracy.**

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:

Status = Developing, Training data, degree=1, RMSE=8.528, R2=0.145
Status = Developing, Testing data, degree=1, RMSE=7.291, R2=0.125

Status = Developing, Training data, degree=2, RMSE=8.234, R2=0.203
Status = Developing, Testing data, degree=2, RMSE=6.847, R2=0.228

Status = Developing, Training data, degree=3, RMSE=8.068, R2=0.235
Status = Developing, Testing data, degree=3, RMSE=6.694, R2=0.263

Status = Developing, Training data, degree=4, RMSE=7.957, R2=0.256
Status = Developing, Testing data, degree=4, RMSE=6.675, R2=0.267

Status = Developed, Training data, degree=1, RMSE=3.637, R2=0.157
Status = Developed, Testing data, degree=1, RMSE=4.565, R2=-0.370

Status = Developed, Training data, degree=2, RMSE=3.617, R2=0.166
Status = Developed, Testing data, degree=2, RMSE=4.615, R2=-0.399

Status = Developed, Training data, degree=3, RMSE=3.614, R2=0.168
Status = Developed, Testing data, degree=3, RMSE=4.573, R2=-0.374

Status = Developed, Training data, degree=4, RMSE=3.582, R2=0.182
Status = Developed, Testing data, degree=4, RMSE=4.586, R2=-0.382

RMSE Scores:

Developing Countries	**Linear Function**	**Quadratic Function**	**Cubic Function**	**Quartic Function**
Training Data (2000 - 2013)	8.528	8.235	8.068	7.957
Testing Data (2014 and 2015)	7.291	6.847	6.694	6.675
Developed Countries	**Linear Function**	**Quadratic Function**	**Cubic Function**	**Quartic Function**
Training Data (2000 - 2013)	3.637	3.617	3.614	3.582
Testing Data (2014 and 2015)	4.565	4.615	4.573	4.586

R2 Scores:

Developing Countries	**Linear Function**	**Quadratic Function**	**Cubic Function**	**Quartic Function**
Training Data (2000 - 2013)	0.145	0.203	0.235	0.256
Testing Data (2014 and 2015)	0.125	0.228	0.263	0.267
Developed Countries	**Linear Function**	**Quadratic Function**	**Cubic Function**	**Quartic Function**
Training Data (2000 - 2013)	0.157	0.157	0.166	0.168
Testing Data (2014 and 2015)	-0.370	-0.370	-0.399	-0.374

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?: **For developing countries, the best mode would be the quartic function (degree =4). This is because it has the lowest RMSE for both training and testing data, along with the highest r2 scores. .For developed countries it**

would be the quadratic function (degree = 2), performs the best and doesn't run into over and underfitting as much as the other models

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	3.784	0.769
Developed Country	4.102	-0.036

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

	Adult Mortality	**Alcohol**	**BMI**	**GDP**	**Schooling**
Developing Country	-0.0285064	-0.1697820	0.0880479	0.0000793	1.3017623
Developed Country	-0.0257914	-0.3808841	-0.0053102	0.0000443	0.5854753

Please submit your code (named *calculate_regression.py*). Comparing developing and developed countries (two models that you build), can you find some interesting results?: **First what stuck out most to me was the fact that the r2 value for the developed countries lowered dramatically from the training to testing data. I think this may be due to overfitting in the model. The coefficients between the status of countries are relatively similar. For developing and developed countries, it looks like as GDP increases, life expectancy increases as reflected in the coefficient. Also more schooling increases life expectancy, for both developing and developed countries.**

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?

Linear regression seems to have a decent amount of limitations when it comes to prediction. One of those being that the model seems to be sensitive to noise and outliers thus causing the model to over and under fit data often, which is reflected in the r2 values obtained in some of our samples. Another limitation is linearity. The assumption that independent and dependent variables have a straight line relation is false, causing issues with the results.

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