Used libraries:

- numpy
- pandas
- matplotlib
- seaborn
- haversine
- SciPy
- scikit-learn
- Statsmodels
- prettytable
- warnings

This is a report on the ten tasks that were assigned to us in big Data Science first assignment. Its primary aim was to equip us with the expertise needed to engage in feature engineering, parametric and non-parametric modelling, and performance evaluation. The report encompasses the processes we followed, the results we obtained, and my own insights regarding each task.

TASK 1:

The first task was to download the two given data sets and to load it into my working environment. The two given files are of csv type and they provide measurements of rainfall and the enhanced vegetation index. To bring this task though completion I followed the steps outlined below:

- I loaded the two given files using a python library called pandas as the dataframes in my Jupyter notebook working environment
- I performed the data exploration by checking the length of each dataframe and the missing values

Results: This task was successfully well complete and I found that it was the monthly measured data for rainfall and vegetation index from 2000 up to 2014 for all 30 districts in Rwanda. The dataframe for the rainfall index measurements has no missing values while the one for vegetation index consisted of missing values from January up to April in each month.

Insights: Rainfall is usually quantified in terms of millimetres (mm) or inches (in) per unit area during a specific time period, which can be a day, month, or year [1]. In this scenario, I assumed the rainfall to have been measured in mm per unit area over a period of one month and the vegetation index was considered to be dimensionless.

TASK 2:

The second task was to provide two time series visualization for taken measurements. The purpose was to provide graphs that shows the behaviour of the rainfall and vegetation index from 2000 up to 2014 for each district. This section was successfully completed thanks to the following steps:

- I conducted the data pre-processing by transposing both dataframe to in order to easily manipulate it
- After performing both data exploration and data pre-processing, I cleaned the data where it was necessary by dropping the NaN values.
- I also make my two dataframe timeseries data using pandas function “to_datetime” and the interval didn’t change it was a one month interval from 2000 up to 2014

- Then I two graphs on for rainfall and the other for vegetation index. Each plot consisted of 30 subplots representing each district

Results: This task was a success as shown in figure 1 and 2 below. The results shown that the maximum rainfall occur in 2007 in almost all districts. The minimum vegetation index also was found to be in 2001 for almost all districts.

Figure 1:The rainfall index of all the districts in Rwanda

Figure 2:The vegetation index of all districts in Rwanda measured from 2000-2014

Insights: I believe that climate change may have contributed to the increase in rainfall levels in 2007. This could be due to an increase in atmospheric moisture resulting from factors such as temperature and humidity. Similarly, the decrease in the vegetation index in 2001 may have been caused by factors such as reduced precipitation or changes in irrigation practices.

TASK 3:

The third task is to provide two visualization of the statistical quantities such as mean, median, minimum and maximum against the month of the year of both the rainfall and vegetation index. This section was performed by following the steps below:

- First, I resampled the timeseries data provided into monthly average and it resulted in each month rainfall and vegetation index average
- Second, I created a python loop that calculate those statistical quantities and append their values in an empty declared variables. In fact. This I did it for both dataframes.
- Finally, I plotted the result using a visualization python library “Matplotlib”

Results: This task was successful completed as shown in figure 3 and 4. The maximum rainfall index value was found in May while the minimum was found in June.

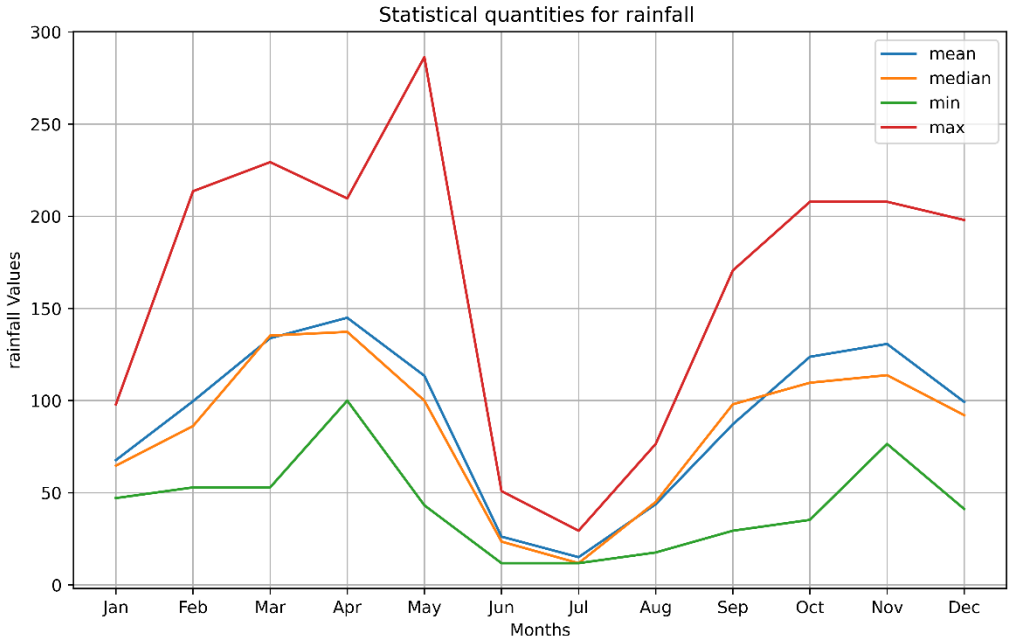


Figure 3: The Monthly statistical quantities of the rainfall index from 2000 up to 2014

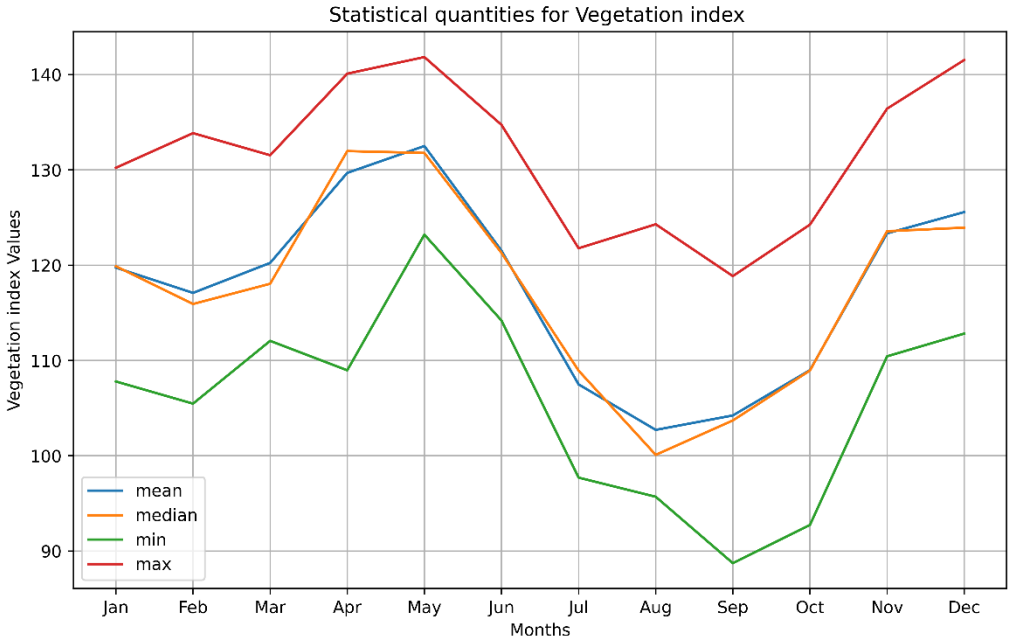


Figure 4:The Monthly statistical quantities of the vegetation index from 2000 up to 2014

Insights: According to Figure 3, the amount of rainfall during the months of June, July, and August tends to be lower, which suggests that summer is a drier season in Rwanda. The data shows that the highest amount of rainfall typically occurs in April, even though the greatest recorded amount was in May. As the level of rainfall decreases, the vegetation index also decreases, but the decline in the vegetation index is slower than that of the rainfall level because it takes time for soil moisture to evaporate.

TASK 4:

In this section the goal was to compute both the correlation and the distance between each pair of district in Rwanda. In addition we had to fit a model and plot the results to show how quickly the correlation declines or increase with distance and this was performed thanks to the following steps:

- I calculated the correlation coefficient between districts for the rainfall using a correlation matrix function
- I loaded another given file that consisted the coordinates of each district into my working environment and used the coordinates in the calculation of the distance between district with the help of SciPy function "haversine"
- Then I plotted the relationship between correlation and distance as shown in figure 5 below
- The model was fitted using a curving fitting function and the result was plotted by adding it to the figure 5 that shows the relationship between correlation and distance between districts

Results: The results showed that the longest distance is between Nyagatare and Rusizi district with a distance of 198.11 km while the shortest distance is between Kicukiro and Nyarugenge district with a distance of 12.99Km.

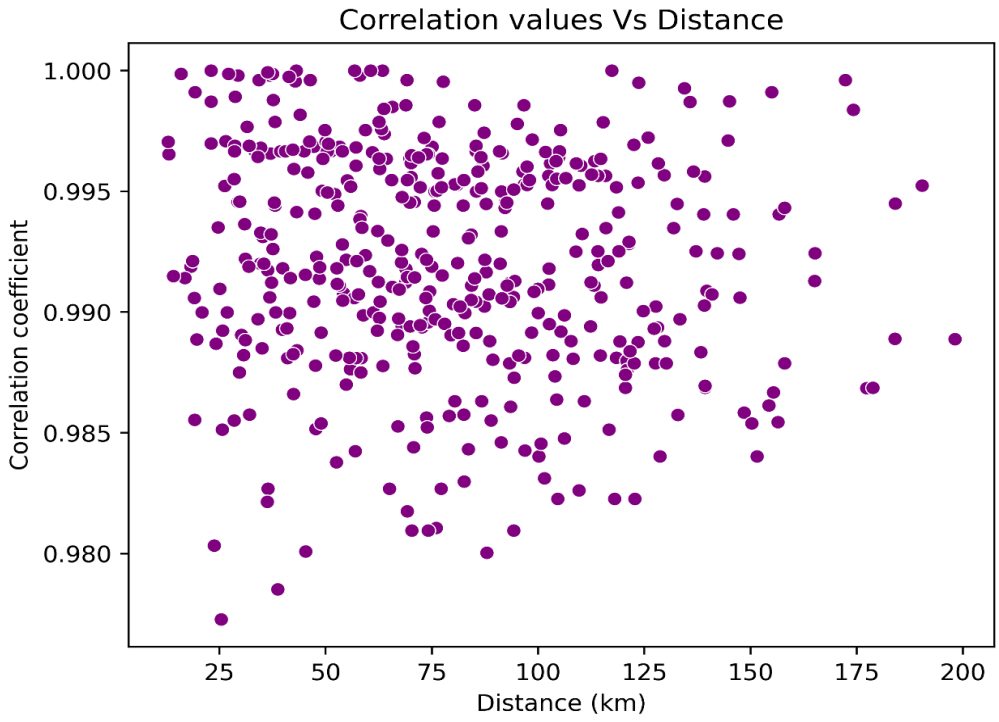


Figure 5: The relationship between correlation and distance of each pair of districts

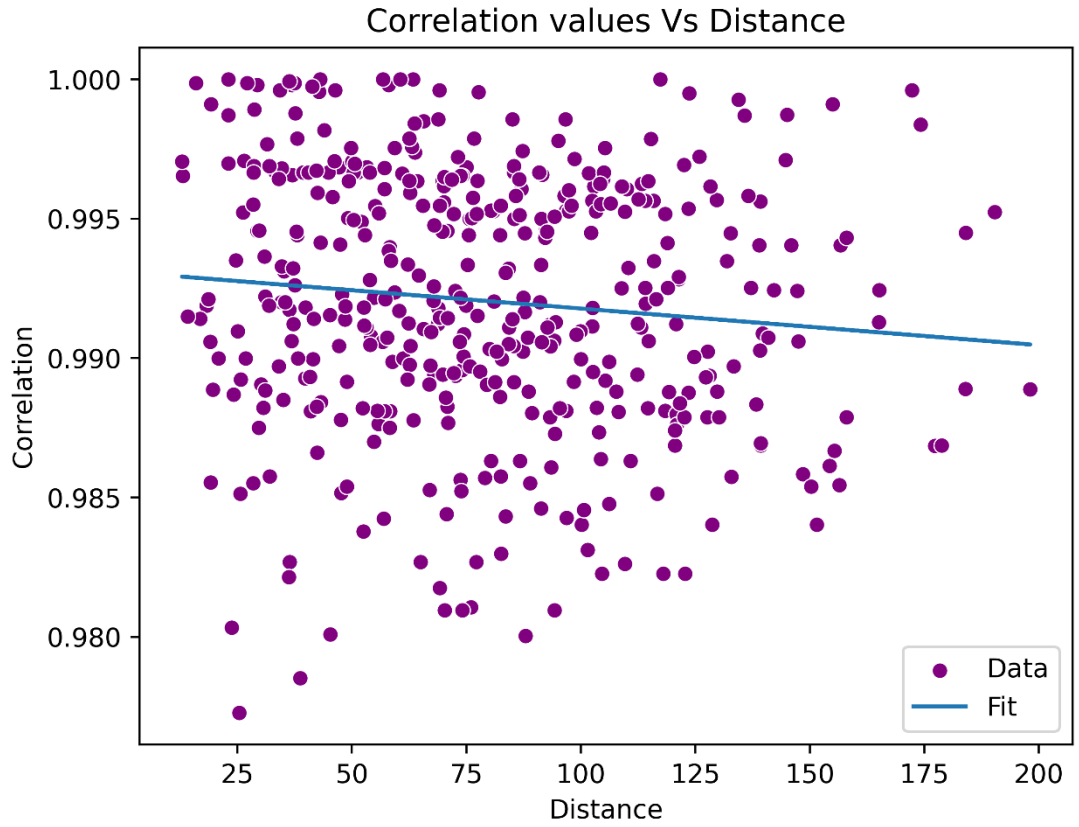


Figure 6:The behaviour of correlation versus distance based on a curve fitting model

Insights: We employed the exponential decay function, C(d) = C0exp(-ad), to model the decrease in correlation as distance increases. However, our findings indicate that this model is not the best fit, providing evidence that correlation does not decrease exponentially with distance. Therefore, we can conclude that the exponential decay function is not suitable for modelling the relationship between correlation and distance in this scenario

TASK 5:

The fifth task was to show the relationship between vegetation index and rainfall by synchronizing the dates that correspond to both timeseries and this was completed thanks to the following steps:

- First, I transformed each of the both data frames(rainfall and vegetation index dataframe) from a wide format to a long format with three columns.
- Second, I merged both dataframe in order to have four columns(Time, districts, rainfall and vegetation index)
- Finally, I created a scatter plot that shows the relationship between vegetation and rain fall index each district

Results: As shown in the figure 7 below the time by which they have had a higher rainfall was not the time they had a huge vegetation index. The maximum rainfall was measured in May 2007 in Rusizi district while the high amount of vegetation index was measured in December 2000 in Nyagatare.

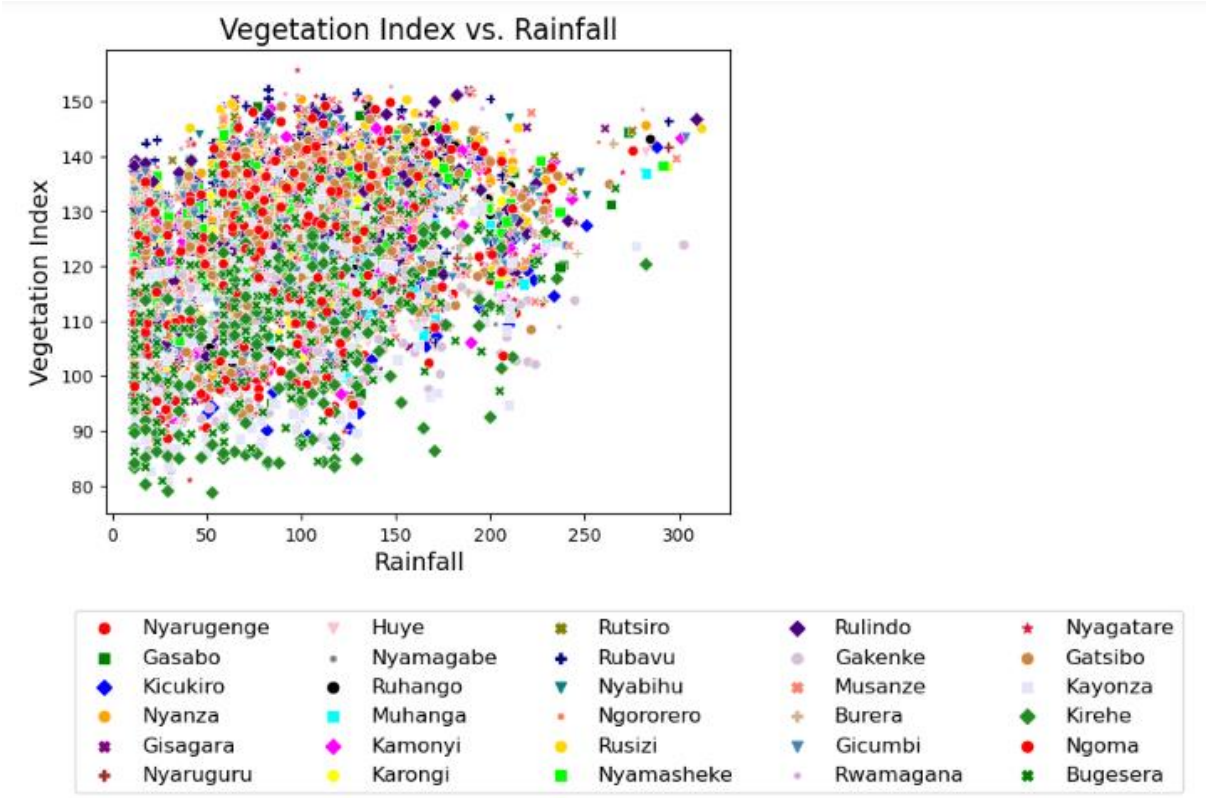


Figure 7: Relationship between vegetation index and rainfall for the same month

TASK 6:

The objective was to develop a rainfall time series feature that enhances the accuracy of predicting the vegetation index, and to identify the best month value for each district as well as noting if there is a consensus. The process involved the following steps:

- I declared the number of shift to range between zero and twelve(k[0:12]) and then I created a function that find the persistence of the rainfall values for each district by shifting it from zero to twelve
- I created another python function that calculate the correlation between the vegetation index and the predicted rainfall. The result was appended in an empty declared dictionary
- I created a loop that finds the optimal number of shift that shows us the months its takes to see the effect of rain on each district

Results: The results have showed that one month shows the effect of rain on each district except in Nyamasheke district where it takes two months. Moreover, the optimal value of K, which refers to the number of months that most strongly influence the yield, was found to be one (k1) in almost all districts, with the exception of Nyamasheke district where the optimal value of K was found to be two (k2).

Insights: The reason why in Nyamasheke district, it takes two months for the impact of rainfall on crop yield could be due to various factors such as the different soil type, farming practices, the shape and elevation of the land, the presence of pests, or other environmental factors that differ from the other districts studied

TASK 7:

Here the task objective was somehow similar to the previous one the difference was the approach. In fact, the previous task we were delaying a rainfall timeseries while here we computed a simple moving average of the same timeseries. The process involved the following steps:

- I declared the horizon to equal to the number of months(k[1:12]) and then I created a python function that calculate a rainfall simple moving average on each district on each horizon
- I also created another python function that calculates the correlation between the result from SMA prediction and the vegetation index. The results were appended on a declared empty dictionary
- I created a loop that calculate the optimal number of horizons on that gives the highest correlation for each district and then I plotted the correlation against the horizon

Results: The results showed that k3 was the one to give the highest correlation in almost all districts except in Nyaruguru, Nyamagabe, Muhanga, Karongi, Rutsiro, Ngororero and Rusizi. As shown in the figure 8 below the variation of correlation have nearly similar shapes on all districts which show us that even though the value of correlation is not the same on each district the monthly change occurs in the same way in each month.

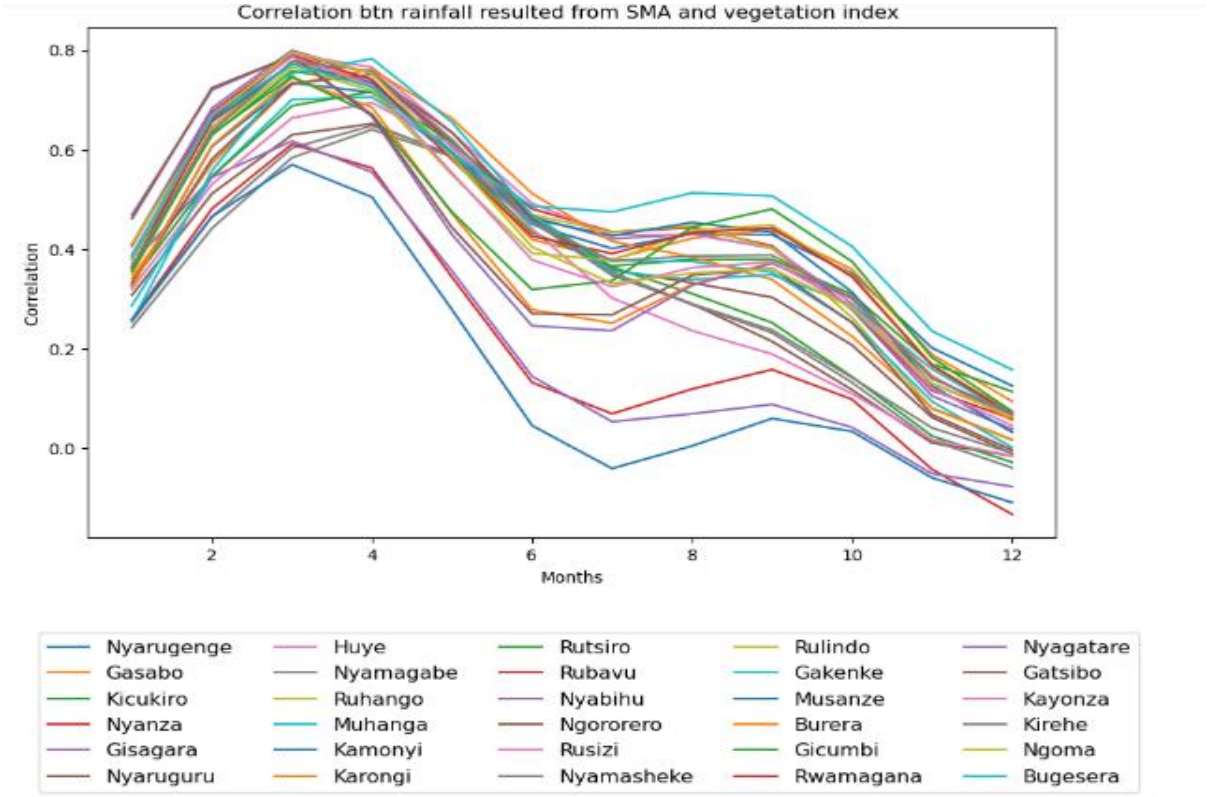


Figure 8: The monthly variation of the correlation between predicted rainfall using SMA approach and the vegetation index

Insight: According to the results there is a consensus that the monthly changes in correlation across districts have a similar pattern, but the optimal value of "k3" may vary depending on the district. In most districts, the findings indicate that a three-months lag between rainfall and its effect on yield is the most significant.

TASK 8:

The objective was to analyse the connection between vegetation index and rainfall through the use of several models, including linear, quadratic, and cubic models. The performance metrics of each model were assessed to determine their effectiveness. In fact, The steps I used on all models are similar and are shown below

- I separated the independent variables (rainfall index) from dependence variables (vegetation index). This were performed on the data given, on the optimal k resulted delayed rainfall I did in task 6, and on the resulted optimal rainfall variables resulted on SMA approach
- I created a linear regression model and then fitted both independent and dependent variables and then I predicted the vegetation index
- On quadratic and cubic regression, I firstly transformed the independent variables into my desired degree and then I fitted the transformed independent and dependent variable and after I predicted the vegetation index
- I computed the model's metrics performance evaluation (Adjusted r-squared, RMSE as well as r-squared)

Results: The results has shown the SMA rain fall to have a better performance on its evaluation because on both Adjusted r-squared and r-squared it is the one to have a better performance compared to other and it also the one to have the low value of the root mean squared error as shown in the tables below

Adjusted R-squarred			
Variables	Linear	Quadratic	Cubic
Rainfall	0.10928384804006175	0.11603006631130253	0.11880512332205329
Delayed Rainfall	0.3886154214669141	0.4467193237878958	0.4496622786449248
SMA Rainfall	0.45353991379698033	0.47104327708627447	0.471866196731214

RMSE			
Variables	Linear	Quadratic	Cubic
Rainfall	13.197639842869853	13.147565874386844	13.126912499687728
Delayed Rainfall	10.944062896691975	10.411039435735209	10.383313841497358
SMA Rainfall	10.358588199684645	10.19134293653499	10.183412305962268

R-squarred			
Variables	Linear	Quadratic	Cubic
Rainfall	0.10945257623706128	0.11619751657341448	0.11897204790562566
Delayed Rainfall	0.3887318978583283	0.44682473066086437	0.4497671248482693
SMA Rainfall	0.45364461969585035	0.47114462920792877	0.47196739117522024

Insights: There is evidence to support the use of a quadratic model to describe the relationship between rainfall and vegetation growth, as it has demonstrated better performance compared to linear models. Additionally, utilizing higher orders of nonlinear relationships may lead to even greater performance improvements.

TASK 9:

In this section the task was somehow similar to the previous one the difference was that I had to split my variables into train and test using cross validation approach. The steps involved on each model are the same and are shown below:

- The separated variables from task 8 were splitted, 70 percent to the train set and 30 percent to the test set.
- I developed a linear regression model and utilized it to match both the independent and dependent variables of a train set. Subsequently, I made a prediction for the vegetation index.
- To perform quadratic and cubic regression, I began by converting the independent variables to the desired degree. Next, I applied the regression model to the training set, which includes both the independent and dependent variables. Finally, I used the model to predict the vegetation index using a test set of independent variables.
- I calculated the performance metrics of the model, which included the adjusted R-squared, root mean square error (RMSE), and R-squared.

Results: As shown in the tables below the SMA rainfall variable was found to be the best feature variables compared to the others for almost on all parametric models that was performed. In fact, it has a best performance on both quadratic and cubic regression models than other features.

Adjusted R-Squarred			
Variables	Linear	Quadratic	Cubic
Delayed SMA Rainfall	0.21170367653699385	0.34776212620282654	0.34909878400071404
Delayed Rainfall	0.30011286126287595	0.43588925122004374	0.4377366720657083
SMA Rainfall	0.2544379894097649	0.4417615774395002	0.4411173374173205
Rainfall	0.10397743322960407	0.11094250701105746	0.11233657523190421

RMSE			
Variables	Linear	Quadratic	Cubic
Delayed SMA Rainfall	12.245523716209167	11.138728541320834	11.127309161384868
Delayed Rainfall	11.705668824145905	10.509073124859325	10.491850795473848
SMA Rainfall	11.995499700455188	10.37973632323536	10.385724009760896
Rainfall	12.990904888844042	12.940315127589525	12.930165750599391

R-squarred			
Variables	Linear	Quadratic	Cubic
Delayed SMA Rainfall	0.21185300336470048	0.3478856795034133	0.3492220840984598
Delayed Rainfall	0.3002461984963942	0.3478856795034133	0.4378437902459207
SMA Rainfall	0.2545808447480653	0.4418685401569863	0.44122442357608316
Rainfall	0.10414716661978607	0.11111092100859288	0.11250472515135268

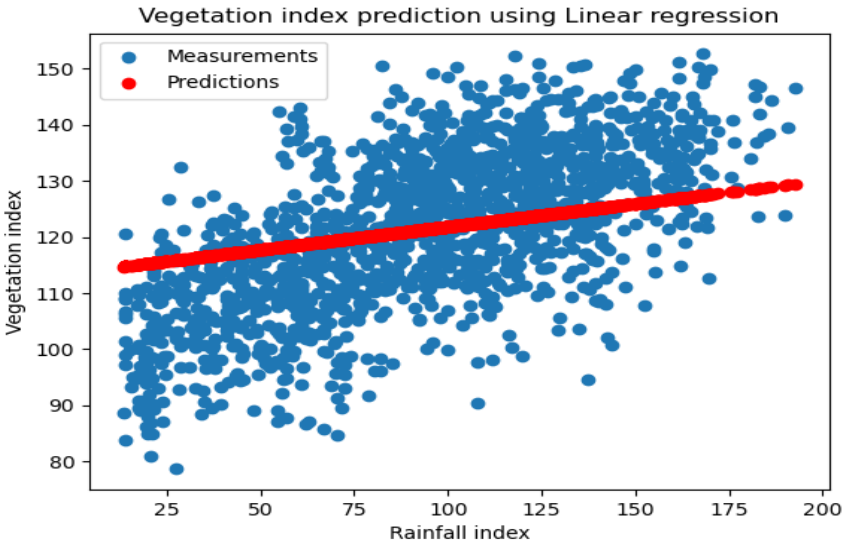
TASK 10:

In this particular section, we were required to assess both linear and non-linear models that were utilized in the previous task, along with two non-parametric models of our own choosing. The aim was to compare and contrast all the models and create visual representations of the fitted models with vegetation index and rainfall features. For my part, I selected the Support Vector Machine(SVM) and K Nearest Neighbour (SVM) non-parametric models. The following steps were taken to fulfil this task.

- Based on my question 9 conclusion SMA rainfall was the best feature variable
- For the linear and nonlinear models, I used the one I computed in the previous task the only thing I did with was to compute their coefficient of determination(R2) and to visualize how this models fit with vegetation index and rainfall level
- I utilized the SVM regression approach to make predictions for the vegetation index. To accomplish this, I began by splitting my feature variable into two parts, train and test. The test set was set to be 30% of the dataset. After that, I scaled the feature variables for both the train and test sets. I then fitted the train set and used it to predict the vegetation index. Finally, I calculated the coefficient of determination as well as visualised how this model fits
- I also employed the K-Nearest Neighbour (KNN) regressor for this task. To do so, I split the variables into train and test sets, and then scaled the feature variables accordingly. After fitting the train set using the KNN regressor, I predicted the vegetation index. I then calculated the coefficient of determination and created visualizations to see how well these models fit.

Results: Out of all the models, the SVM regressor model is considered the best due to its high performance as indicated by the coefficient of determination evaluation. Conversely, the KNN regressor model is the worst among all models since it showed the poorest performance based on the coefficient of determination evaluation.

R-squarred					
Variables	Linear	Quadratic	Cubic	SVM regressor	Knn regressor
SMA Rainfall	0.3002461984963942	0.4418685401569863	0.44122442357608316	0.44340516219917436	-0.11983207770429938



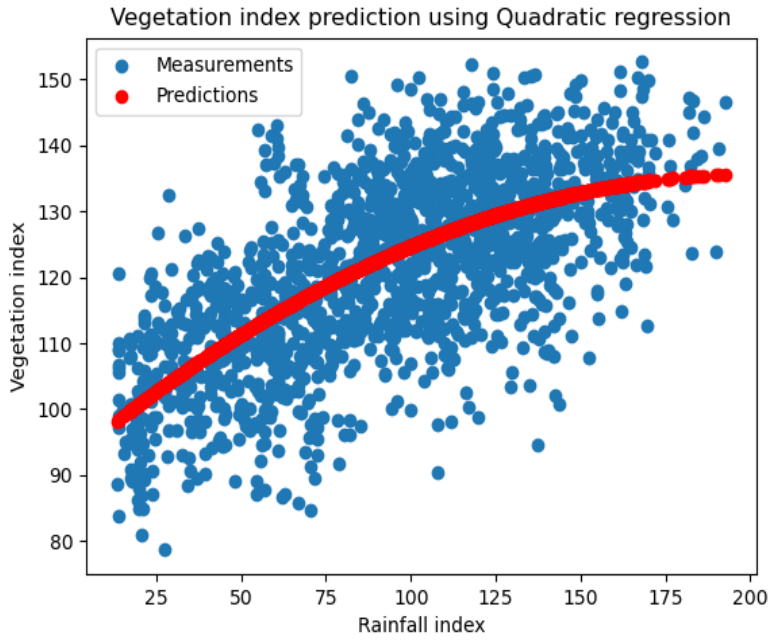


Figure 9:The fitted quadratic model with vegetation index against rainfall feature

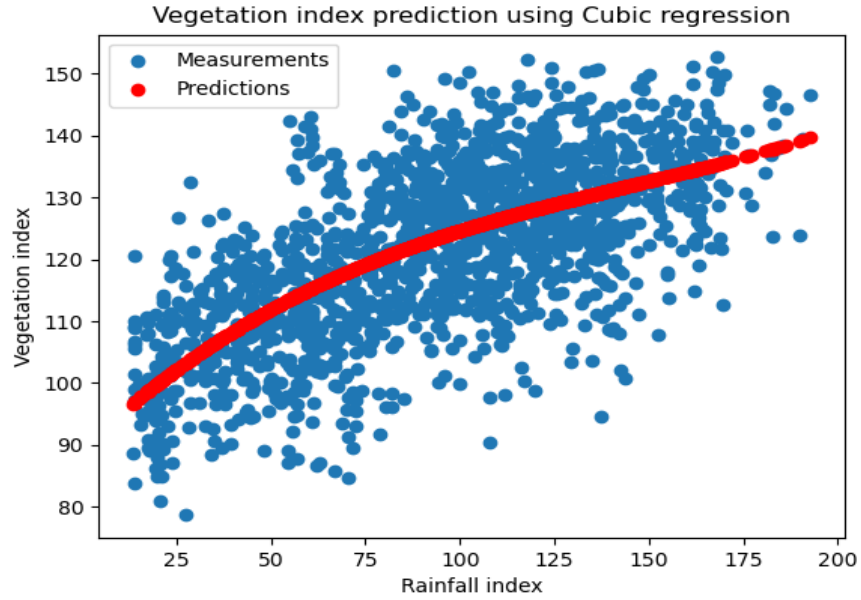


Figure 10: The fitted cubic model with vegetation index against the SMA rainfall feature

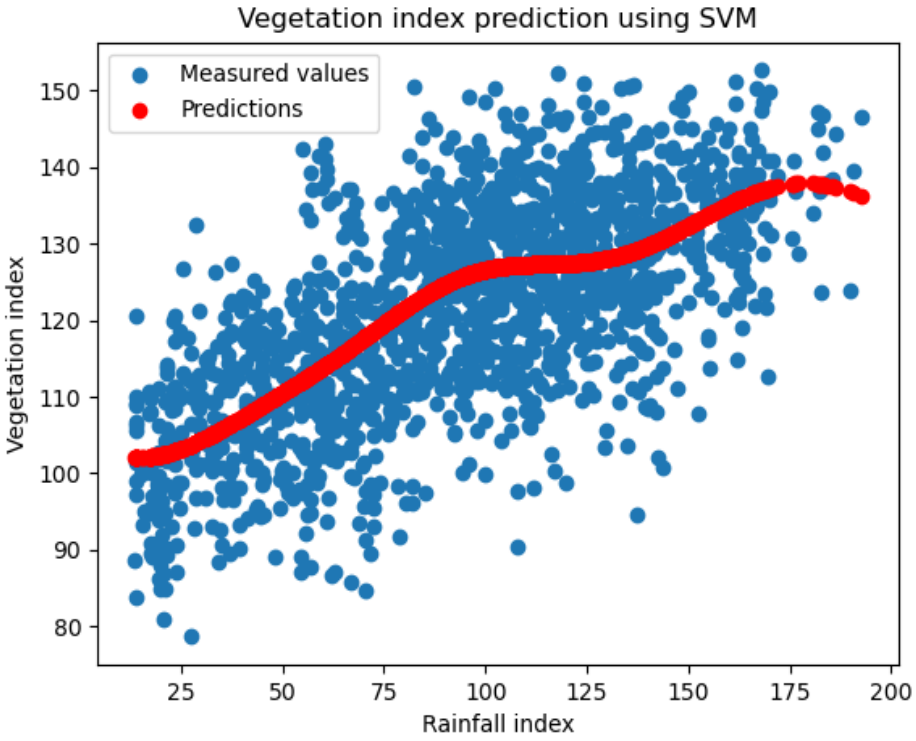


Figure 11:The fitted Support vector machine regression model with vegetation index against the rainfall feature

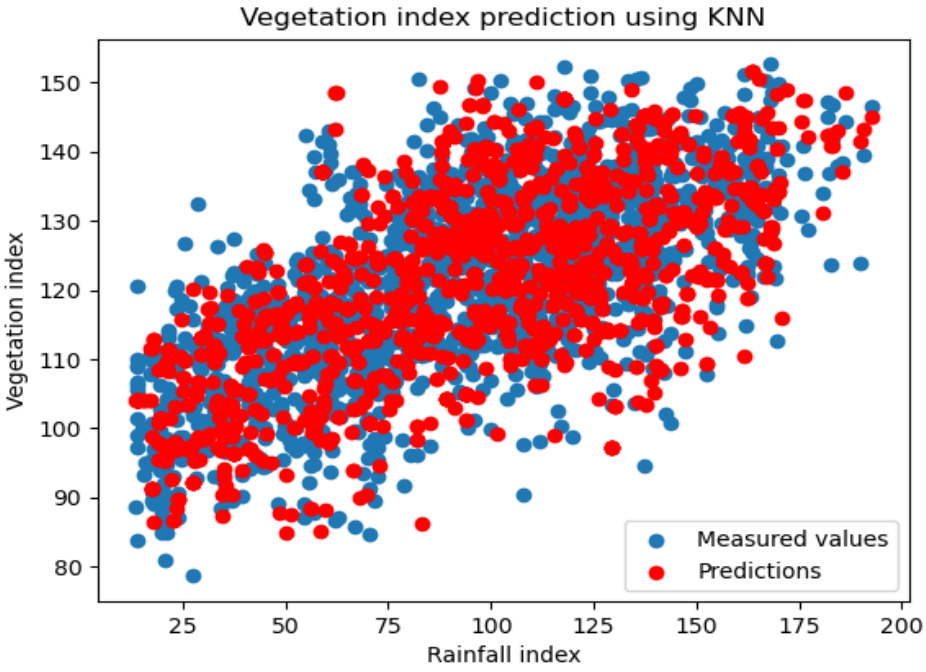


Figure 12: the fitter K nearest neighbour regression model with vegetation index against the rainfall feature

Insight: Based on the characteristics of our dataset, I suggest using SVMs for predicting the vegetation index. SVMs have demonstrated to be a robust and adaptable model that can effectively handle high-dimensional data and nonlinear relationships between the predictors and the response, making it a suitable choice for our problem.