CURNEU- TASK-1

PROBLEM

Try to understand the dataset of Social_Network_Ads.csv and try to find the best suitable ML algorithm and write the code in python for algorithm from scratch and try to achieve the below output plot.

OBJECTIVE :

• Analyze the pattern in the data

• Build a model that gives the given graph as output

SPLITING OF DATA:

In preperation for inputting the X and y to our model, we must first split it into train and val sets.After that, the X_train and X_val are scaled with a min max scaler.

BUILDING THE MODEL:

I have fitted Random Forest, because only this model give the desired output for the given dataset. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.
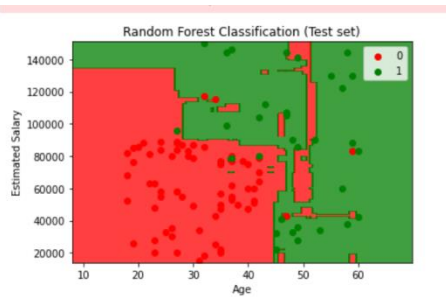
VISUALIZING THE TRAINING AND TESTING RESULTS:

Thus the required graph is being obtained by ploting randomfores for the Social networking dataset

Training Graph:

Testing Graph:

Random Forest Classification (Test set)

CURNEU- TASK-2

PROBLEM STATEMENT:

Explore whether the no. of cases, deaths, CFR of malaria increases every year

OBJECTIVE :

• Analyze the pattern in the data

• Get insights from data

• Build a model with better accuracy

• Predict the no. of cases, deaths, CFR of malaria

BACKGROUND:

Malaria is a life-threatening disease caused by parasites that are transmitted to people through the bites of infected female Anopheles mosquitoes.
• It is preventable and curable.
• In 2018, there were an estimated 228 million cases of malaria worldwide.
• The estimated number of malaria deaths stood at 405 000 in 2018.
• Children aged under 5 years are the most vulnerable group affected by malaria;
• in 2018, they accounted for 67% (272 000) of all malaria deaths worldwide.
• The WHO African Region carries a disproportionately high share of the global malaria burden.
• In 2018, the region was home to 93% of malaria cases and 94% of malaria deaths.

DATASET:

• reported_numbers.csv - Reported no. of cases across the world
• estimated_numbers.csv - Estimated no of cases across the world

• incidenceper1000popat_risk.csv - Incidence per 1000 people at risk area

FEATURE SELECTION:

- Get dummies has been applied to the 'Country' feature of the reported data, thereby one hot encoding it.

- We use a Label Encoder to convert the 'Country' and 'WHO Region' features from categorical to numerical.

- Now, an incidence feature is added to the reported data, followed by binning the cases and incidence features.

- Next, log transform and box cox features have been added to the reported dataframe.

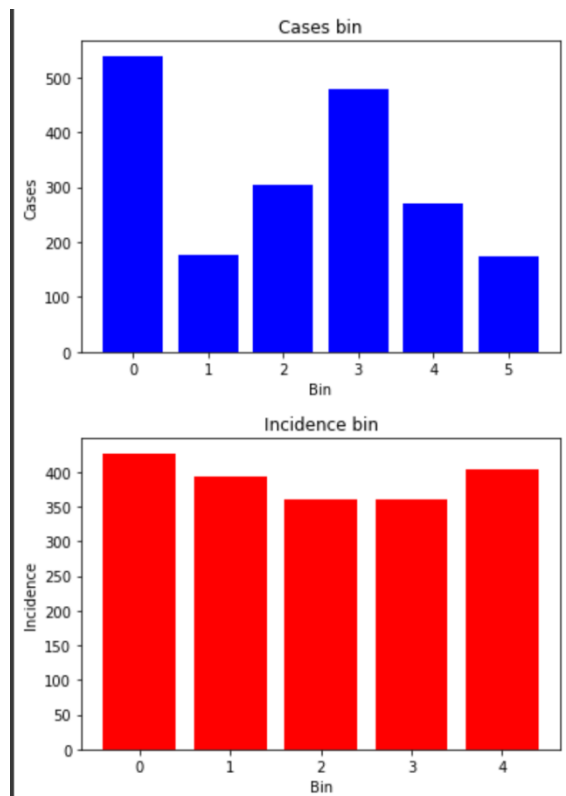- We also add a minmax, standardized and normalized set of variables
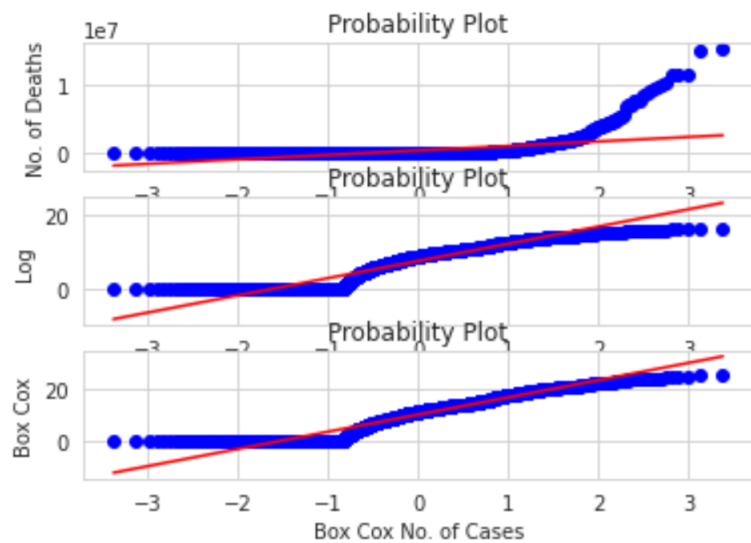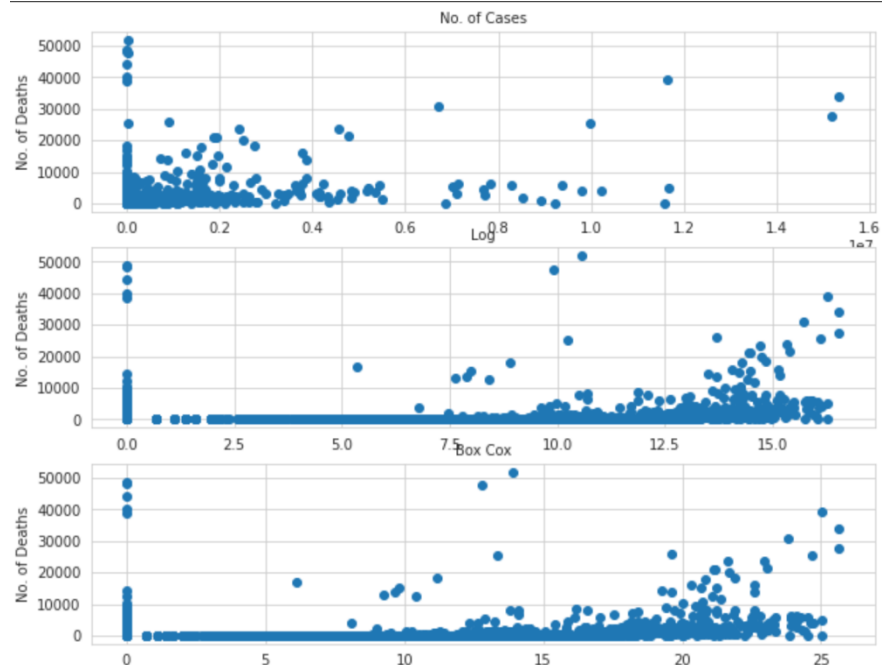
DATA VISULISATION:

Afterwards, we must now visualise our new feature engineered data with comparison to their original

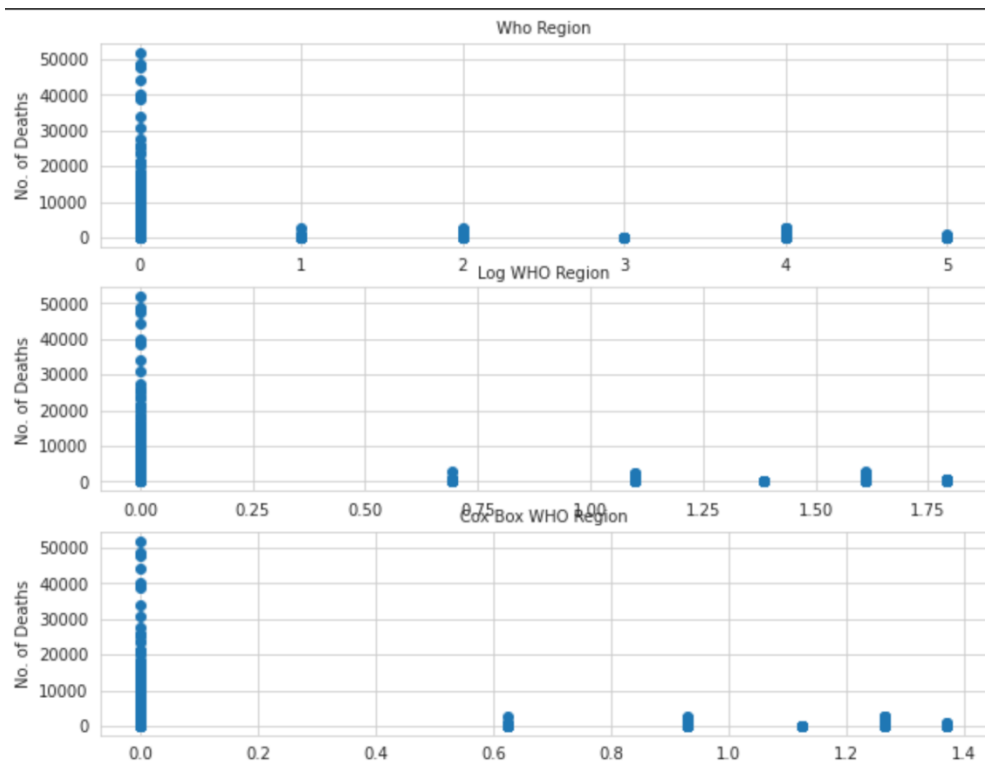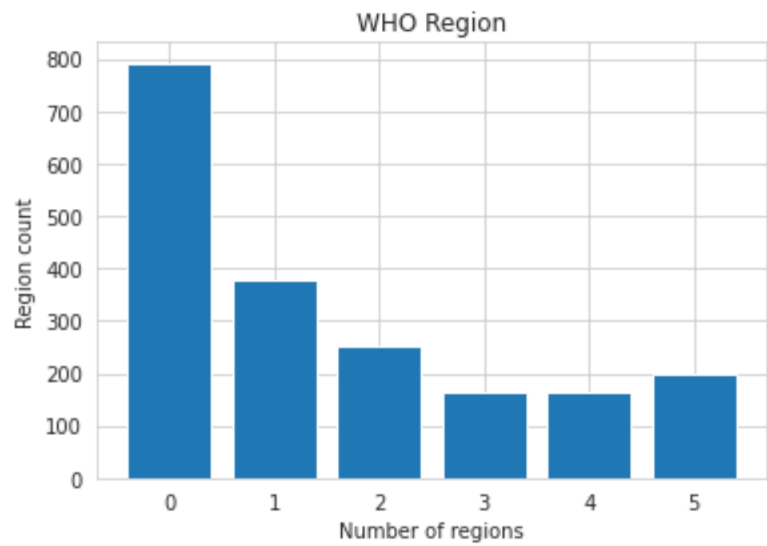Firstly, we look at the binned data which has come from the cases and incidence features
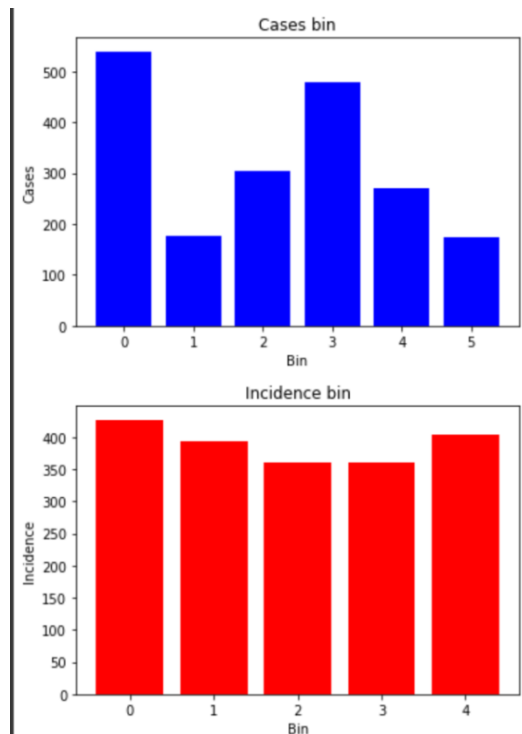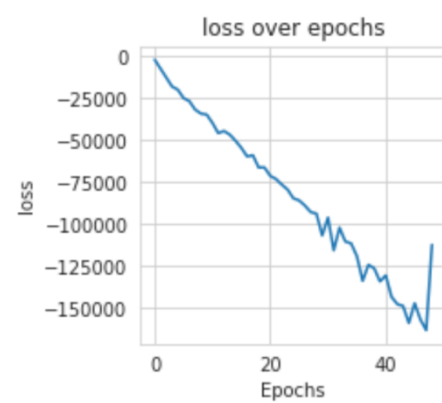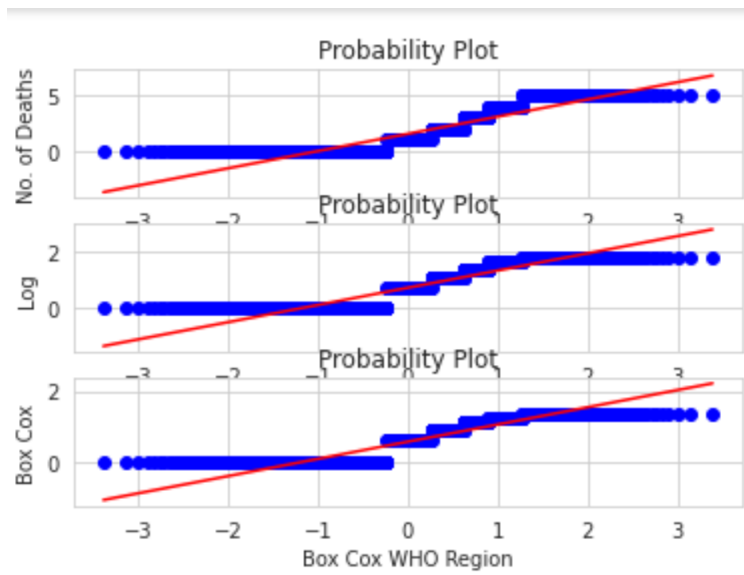
Secondly, the log transform and box cox of our X is visualised and compared to the original data.

No. of Cases

Probability Plot
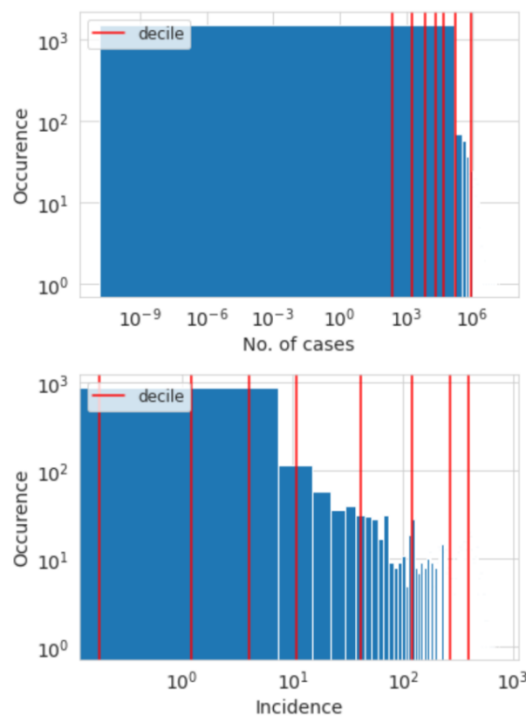
This below is done for the cases, WHO Region and incidence features
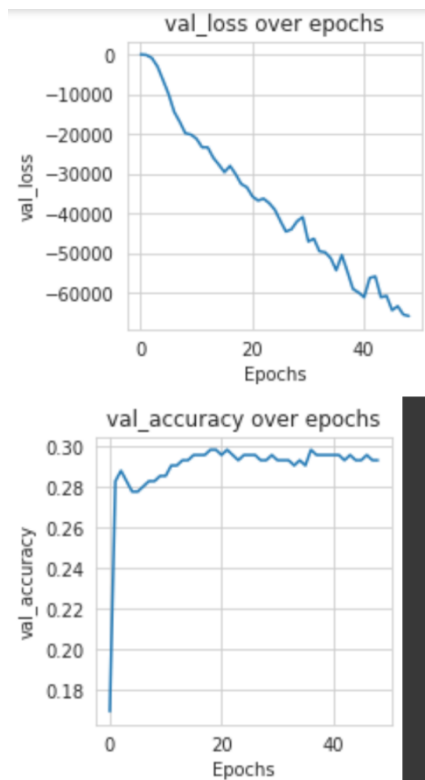
WHO Region

Who Region

Probability Plot

Probability Plot

Probability Plot

loss over epochs

accuracy over epochs

SPLITING OF DATA:

In preperation for inputting the X and y to our model, we must first split it into train and val sets.

After that, the X_train and X_val are scaled with a min max scaler.

BUILDING MODEL AND PREDICTION:

We have created a Multilayer perceptron (MLP) which is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses backpropogation for training the network.

Multilayer Perceptron with 3 hidden layers, 3 dropouts which have a dropout rate of 0.325 and an output activation of sigmoid. The loss is a binary cross entropy and the optimizer is an Adam which has a learning rate of 0.0001 and an epsilon of 2e-06. which gives moderate accuracy.