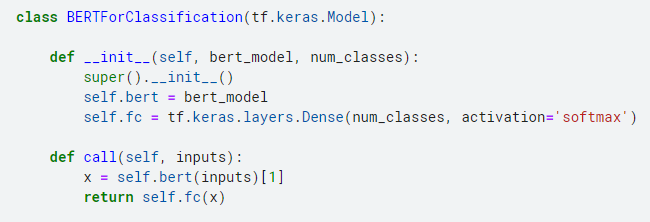
**REPORT ON TEXT CLASSIFICATION:**

**(Primary contributor: Kushaj Mallick, Swastik Mukherjee)**

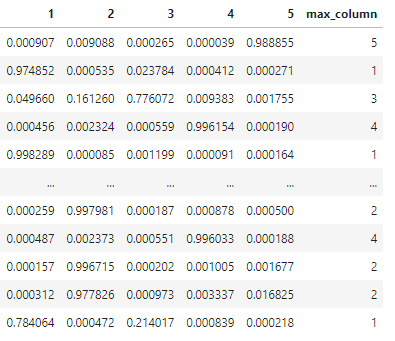
The data is preprocessed first by encoding the different mental illnesses (“Text”) as integers. It is then divided into a test train split. Both the test and train splits are tokenized.

The model used for this problem statement is a fine-tuned BERT model from the HuggingFace package.



The model is then trained with a learning rate of 10^-5 and fit with the training split of the training dataset.

Predictions are then made on the test.csv dataset which are in the form of 5 columns of numeric data. The column id of the highest value from these numeric columns is then chosen to be the final output value for the corresponding “id” and then replaced with its corresponding mental illness from the previous encoding initial encoding part.



These redundant columns other than the max\_column are removed and the max\_column is renamed as “label”. An id column is inserted into this output dataset from the test.csv dataset.

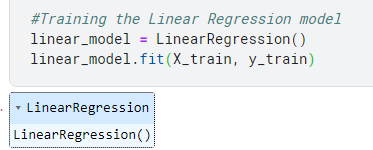
References used: Youtube channels – Pradip Nichite, codebasics, Nicholas Renotte, Pritish Mishra

**REPORT ON TRANSFER VALUE PREDICTION:**

**(Primary contributors: Nishant Rana, Swastik Mukherjee)**

The data is first preprocessed by removing the “Name” and “Country” columns and the missing values are filled with 0.

The model used is a simple Linear Regression model which trains on the entire dataset.



The data from the test.csv file is also similarly processed in order to make the number of columns similar to the train.csv dataset.

The model is used to predict the values in the test.csv dataset and this output dataset is then converted into the final dataframe and its column names are changed in accordance with the submissions.csv file.

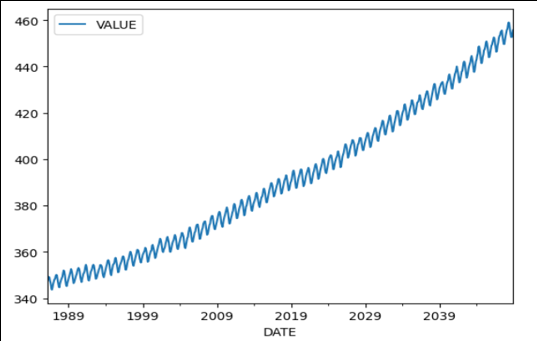
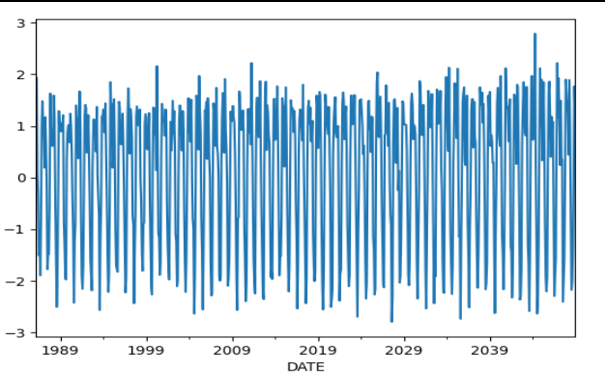
References used: Youtube channels – StudyGyaan, Analytics Summit, AK Python

**REPORT ON TIME SERIES ANALYSIS:**

**(Primary contributors: Kushaj Mallick, Nishant Rana)**

The data is preprocessed first by removing the date, month, unit, and indicator values and replacing all of them by a singular datetime column with each date set to the first of every month of the corresponding year. The “spikes” in the data are removed and all the missing data together are replaced by interpolated data with “time” method.

The plot of the data proves that it is seasonal in nature and the Augmented Dickey-Fuller test helps us to understand its stationarity. Since it is non-stationary, a value first difference column is added to bring about stationarity in the data.

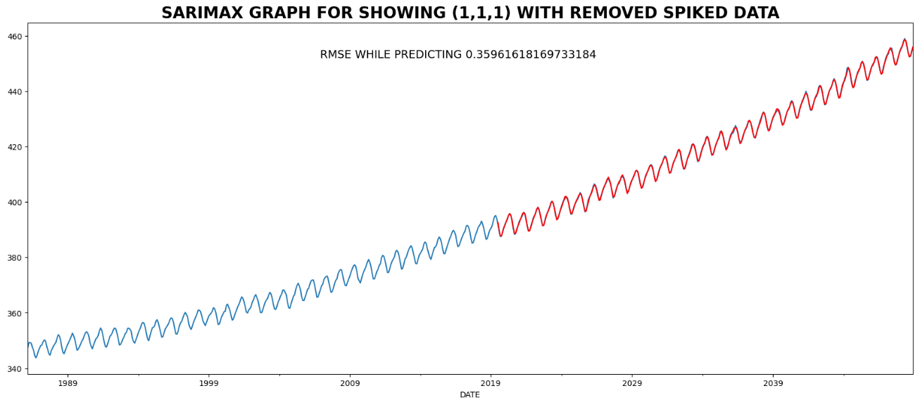
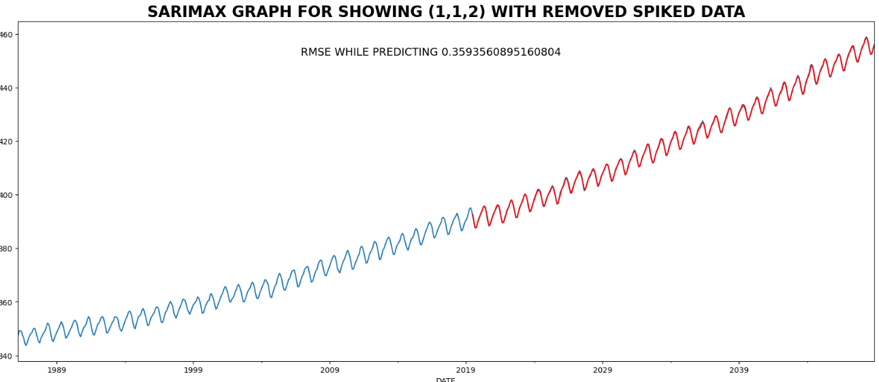
 

*(Data after interpolation) (Graph indicating stationarity)*

The model is the SARIMAX model which takes 3 input parameters, the p,d and q. Since we used value first difference, the value of d is 1. The Partial Autocorrelation graph (PACF) AND THE Autocorrelation graph (ACF) tell us more about the p and q parameters. However to be absolutely sure of the optimal parameters we also compare the AIC and BIC values of the possible cases of parameters obtained by analysing the PACF and ACF graphs.

Both AIC and BIC values for the 2 parameters are very close, we shall apply a stacking of the two models obtained by using the two sets of parameters.

We then build the 2 SARIMAX models and plot the graph of predictions it makes on the training dataset itself and also display the RMSE value accuracy with which it predicts data.

   
 *(Graph for SARIMAX model (1,1,1) (Graph for SARIMAX model (1,1,2)*

So we see that both models are pretty well trained and predict the data (in red) with an RMSE error of .359616 and .359356, respectively.

Next we predict the results for the next 24 months by using both models and finally combine the two resulting datasets by taking a weighted average of the values to get the final dataset. The graph for the predictions (in orange) is presented as follows:

