Analyzing seasonal and inter-annual turbidity of a wetland ecosystem in India using Machine Learning and Time-Series Modeling

Ashish Mishra, Santonu Goswami

Analyzing seasonal and inter-annual turbidity of a wetland ecosystem in India using Machine Learning and Time-Series Modeling

Ashish Mishra[1], Santonu Goswami[2]
[1]KIET Group of Institution, Ghaziabad, India
[2]Earth and Climate Science Area, National Remote Sensing Centre, Hyderabad, India

1. Time Series Forecasting

Time Series Forecasting is a method which uses previously stored data corresponding to a time span and predicts/estimates the futuristic values, doing so will be helpful for the business decision making in various sectors. Traditional time series models such as ARIMA and SARIMA consider the single seasonality but the modern methods such as BATS and TBATS are more effective because they consider the multiple seasonality.

Time series forecasting is an active research area which focuses on modeling future values of a parameter based on periodic observations made over a certain period of time.

Time series data set will always be a sequence of observation as follows:

Time	Feature_1	Feature_2	…….	Feature_m
Time #1	X_{11}	X_{12}	…….	X_{1m}
Time #2	X_{21}	X_{22}	…….	X_{2m}
. . .	. . .	. . .	. ……. .	. . .
Time #n	X_{n1}	X_{n2}	…….	X_{nm}

Table 1.1 General Time Series Data Set

Forecasting means to use a trained model on the historical data to predict the future values/ observation.
So the time series Forecasting means to predict the future observations using a Time series Dataset. But before making Time series prediction or forecasting it is needed to understand the nature of the dataset so to understand the nature of the data set it is needed to perform the "Time series Analysis".

2. Time series Analysis

The time series analysis is the analysis of the time series data set and it involves the understanding of the nature of the dataset and it often involves decompositions of the time series dataset into its various components.
The various 4 components in which the time series dataset can be decomposed are as follows:

1. Level :- The baseline value from the series starts when the series is straight line.
2. Trend :- It is linearly increasing or decreasing behaviour of the series over a long time.
3. Seasonality :- It is the nature of the series in which the series repeats its nature after a cyclic period of time.
4. Noise/irregularity :- It is the component which is irregular in nature and can not be defined by any model.

Here the series is time series data value and it can be calculated by using its all components i.e.

$$Series = Level + Trend + Seasonality + Noise$$

3. Why Time series forecasting is Important

Time series forecasting involves predicting the future values of a variable or feature by using a pre-trained model using historical data(of the same variable for which the model is making prediction). And in the real world we have many firms which produce a lot of Time series data which can be used for making future predictions which helps the firm economically as well as also reduce the effort of the firm.
Time series forecasting can be useful in predicting stock price values, oil price prediction, Gold price prediction, weather forecasting and predicting the covid-19 patients number etc. So time series forecasting can be useful in a wide range of areas.

4. Time series Forecasting in Climate and Environmental Science

The Application areas of time series forecasting are wider and some of them can be enlisted as follows :

4.1 Climate Science and Environmental Science:

Time series forecasting can be applied to a wide variety of areas for having almost accurate prediction of future values of some of the useful parameters. The climate science and Environmental science have a lot of parameters whose future prediction can be helpful for humankind as well as nature, some of these parameters are Global temperature, Amount of RainFall in a particular region, Rate of SnowFall, rate at which the cloud droplets convert to the

rain, Sedimentation in the wetland, pollution control, water resource management, predicting the forest fire etc.
These all parameters behave in a random way but there are some of the machine learning algorithms which made it possible that these random parameters can be predicted with a high accuracy. And if we made it possible to predict the future values of these parameters then it can make human life more ordered. If we take example of our case of predicting the Sedimentation in the chilika lake than we can made it possible to restrict the disturbance of this sedimentation on the wetland and this ultimately we can restrict the sedimentation to disturb the whole chain of the Wetland Ecosystem of the Lake which is livelihood of many peoples as well as home of various species of flora and fauna.

4.2 Other Sectors :

There are many other applications of Time series forecasting in the Sectors such as Business, Healthcare, Banking, Social media forecasting, Cyber security, Agriculture, E-Commerce, Industries and IT Sector.
Time series forecasting is useful in business in predicting the number of online users on a website, forecasting customer satisfaction as well as customer feedback, forecasting staff, salary and requirement of staff. As Global health is declining, it really becomes necessary to keep track of the diseases and here comes the role of data science and machine learning. It helps in predicting the various diseases such as cardiovascular diseases, heart disease, liver related problems and nowadays it is used in making predictions about Covis-19. Using Artificial Intelligence in healthcare doesn't mean that we are replacing the doctors with machines but it means using the technology in a proper way to help the doctors to make diagnosis decisions which ultimately reduces time and effort of doctors and also it helps doctors in having confidence in their decision. COVID-19 shows the mirror to many countries that there is a need to invest more in healthcare and education along with defence and other sectors.
Time series forecasting has applications in the banking sector such as in detecting fraudulent transactions and in management of the cash in the various branches according to their needs. It is also used in sentiment analysis using social media activities of it's users as well as it also helps in detecting the intrusion detection and preventing the cyber attack. Along with these applications there are some other applications such as Inventory management and customer recommendation system etc.

5. Introduction to Problem statement

Time series forecasting has numerous applications and its applications in climate science are very significant to conserve or protect the climate. Wetlands are the home of a wide variety of flora and fauna species as well as it's livelihood for many human beings and if the ecosystem of these wetlands gets disturbed then it leads to disturbance of a lot of humans as well as other species. Sedimentation is the main problem which causes the disturbances of the ecosystem of these wetlands so the sedimentation process need to be monitored but monitoring the sedimentation manually is a expansive process so this project is a proposal in which the data set collected by the geostationary satellites of Indian Space Research Organization(ISRO) can be used to analyse the pattern of the sedimentation in these wetlands so that they can be monitored. In this project the focus is on the chilika lake, odisha, India and the dataset is

collected for this lake which is UNESCO WORLD heritage[1]. The traditional time series models such as Autoregression (AR) and Moving Average (MA) as well as the other models such as LSTM and CNN can be applied for finding the hidden pattern inside the dataset. These models produce good results.

6. Literature Review

Machine learning is a powerful tool which can help forecast climate and environmental changes, it has applications in the Environmental, Ecology and Climate related issues because there is a lot of data produced in this particular field. Remote sensing is one of the important applications by which data can be captured by remotely located satellites and this data can be processed using machine learning and data science fields to have good results. Okujeni et al. in their research article uses the Environment mapping and analysis program (ENMAP) for the remote sensing of the urban region, the data set used in this study is ENMAP data captured remotely over Berlin, Germany, with the resolution of 30 m. The data is generated synthetically and the model applied is Support vector Regression (SVR) [2]. Ground water is one of the important water resources mainly for the gulf countries, Naghibi et al. in their article predict the potential groundwater resource area. The models used are boosted regression tree (BRT), classification and regression tree (CART), and Random Forest (RF). They use thirteen features such as slope degree, slope aspect and plan curvature etc. [3]. Heung et al. in there research article used various algorithms (Random Forest, CART, CART with bagging, K-nearest neighbour, artificial neural network, multinomial logistic regression, Support vector machine and logistic model trees) for finding out a function which can predict the taxonomy of the soil, k-nearest neighbour and support vector machine with radial basis perform very well with accuracy of 72 %[4]. Park et al. in their research article try to predict the drought using 16 climate factors, the algorithms used are Boosted regression tree, Random Forest and cubist out of which Random forest perform best with RMSE of 0.3 and R^2 value 0.93 [5]. Appelhans et al. in a research article try to predict the monthly air temperature using 14 machine learning models but the best performance given by stochastic gradient boosting algorithm [6]. Xie et al. in their research article try to predict the Batrachochytrium dendrobatidis (BD) which is a pathogen associated with worldwide amphibian population losses, the Random forest model trained on the BD sampling record and this model easily estimate the effects of different climate factors on the BD distribution [7]. Masood et al. in their research article estimate the occupancy of Air conditioned space by using various climate factors such temperature, humidity and CO_2, these climate factors change with the change in the occupancy of the space. This research article tries to predict the occupancy level through these climate factors [8]. Hallgren et al. in their research article introduce a Biodiversity and climate change virtual laboratory (BCCVL) which can provide the open access to the biological, climate and environmental dataset, numerous species distribution modelling techniques and a variety of research types for conducting the research [9]. Lary in their research article uses the remotely sensed data as well as samples collected from 8329 geographical sites of 55 countries from 1997 to 2014, they try to predict the $PM_{2.5}$[10]. Lou et al. in there research article predict the diffuse solar irradiance of sun, for this purpose the data used is taken between 2008 to 2013 from Hong Kong, and the MAE for Hong Kong and Denver, USA comes out to be less than 21.5 w/m^2 and 30 w/m^2 [11]. O'Connor et al. in their research article try to predict the potential fire control locations for pre fire control measures which can be

decided, the boosted logistic regression model performs best with the accuracy of 69 % even though the article doesn't consider the weather conditions [12]. Soil moisture(SM) is one of the important environmental factor and it is important for the purpose of agriculture, Prasad et al. in their research article try to predict the monthly value of the SM, for this purpose the Extreme learning machine(ELM) models are applied to forecast SM of upper soil layer(0.2m depth) and lower soil layer (0.2m to 1.5m depth) [13]. Chatziantoniou et al. in their research article try to map the Land use and Land cover (LULC) using machine learning and remote sensing data and the best accuracy comes out to be 94.82 % and kappa to be 0.936 [14]. Forkour et al. in their research article use the remotely sensed data for mapping the soil properties and the models used for this purpose are multiple linear regression(MLR), Support vector Machine (SVM), Stochastic Gradient Boosting (SGB) and Random forest regression (RFR) [15] . Kleine et al. in a research article try to predict the $PM_{2.5}$ level of the Quito, Ecuador using machine learning and the regression model suggest that the prediction becomes more accurate when the weather conditions become extreme [16]. Shafizadeh-Moghadam in their research article try to predict the flood susceptibility using eighth machine learning models and the data set used consists of 201 flood events occurs in Iran and 10,000 randomly selected non-occurrence points, the highest accuracy comes with the model boosted regression model which is equal to 97.5 % [17]. Sedimentation is a big problem in the rivers and wetlands so they need to removed or predicted in advance so that they can be prevented, Choubin et al. in their research article try to predict the sedimentation in the Haraz river(Iran) using the Classification and regression tree (CART) model and the performance of CART model compared with the other models [18].

There are a lot of research application of Machine learning in other sectors such as Agriculture i.e. Bunn et al. in their research worked on to predict the suitability of the coffee crop in a particular region because coffee comes under the category of crops whose productivity depends on climate change Temperature is one of the most affecting factor from coffee, as temperature increase the production suitability of coffee decreases, and as the emission continues the production of coffee will decreases by 50%[19]. Behmann et al. in their research article work try to predict the future chances of diseases in the crops. It could detect the future chance of diseases in the crop by using spectral features and weed using shape descriptors[20]. Johnson et al. used two online machine learning models - Bayesian neural network (BNN) and model based recursive partitioning (MOB)- and Multiple linear regression model to predict the crop yields of various crops i.e. barley, canola, and spring wheat grown in canadian prairies[21]. Kuwata et al. is a research article using deep learning, machine learning and SVR for predicting corn yield in Ilinoisis. One of the algorithms used in which the network model of two InnerProductLayer outperform the other models which archive the RMSE OF 6.298 [22]. Crane-Droesch et al. in a research article used semiparametric versions of deep neural networks which outperform the fully non-parametric version of deep neural networks and other classical algorithms. The results show that the climate has a more negative impact on corn yield[23].

There are a lot of concerns in the cyber world which can be solved by the data science and machine learning applications. Buczak et al. in their research article, use the data set as well as methods which are used in the cited research article by them[24]. Beaver et al. in their research article use the data set of telemetry gas pipeline stream in mississippi state university, the data set includes the normal instance (where there is no SCADA attack occurs) and instances where SCADA attack occurs, the article mentions many algorithms which are used in the research i.e.

Naive bayes, Random forest, J48, oneR, NNge, and SVM. The results show that the algorithms are capable of classifying the type of attack[25]. Wu et al. in their research article apply machine learning models on 2 cyber Manufacturing Systems(CMS) to detect the cyber physical attack on them. The result shows that the highest accuracy of 3D printing CMS comes by applying anomaly detection algorithm which is equal to 96.1 % and highest accuracy of CNC milling machine comes by applying Random Forest and it is equal to 91.1 % [26]. Transaction frauds in E-Commerce business is one of the challenging problems and they need to be detected and tackled. Wang et al. design CLUE, which is a Novel system which can detect these fraud transactions. This system is designed and deployed on JD.com the largest E-commerce website of china with almost 220 million active users and the result shows that it helps in reducing fraud transactions[27]. Subroto et al. in a research article used the social media chat data set for the purpose of predicting the chances of the cyber attack, the highest accuracy in prediction of cyber attack given by K-Nearest Neighbor(i.e. 96.33 %)[28].

Shrivastava et al. in their research article use the Microsoft Azure cloud platform for developing and deploying the probabilistic system which categories the E-commerce products into relevant categories. The data set used is having 9 categories of products with 93 features and it uses multi decision forest for modelling and the accuracy comes out to be 68.59 % which is higher than the benchmark accuracy [29]. Dhaoui et al. in a research article use 2 machine learning packages(i.e. LIWC2015 lexicon and RTextTools) on the data set of 850 consumer comments taken from 83 facebook pages, to know the sentiments of the consumers for a particular product [30]. Isaac et al. in a research article make a web based recommendation system using python, Web2py and Model-view-controller (MVC) which recommend the products to the consumers through the website [31].

Healthcare is one of the most important fields in which data science is continuously applying to increase the level of treatment of the patients. Manogaran et al. in a research article use this DNA copy number for cancer diagnosis of patients. The model used is a hidden Markov model and Gaussian Mixture (GM) Clustering and the results are showing the improvement over the current system [32]. Poostchi et al. in their research considers the images of the blood smear of the infected cell and then tries to classify the images into the malarial cell image or not malarial cell image. The performance shows the 97 % AUC(Area under the curve) [33]. Rajaraman in their research article uses a data set which contains 27558 images of the blood smear cells out of which half are parasitized and half are non parasitized, the deep learning models used gives the highest accuracy of 95.70 % using ResNet-50 [34]. Saiprasath et al. in their research article use various models for the malaria diagnosis using the blood smear dataset, the models used are AdaBoost, KNN, Decision Tree and Linear Regression and the model which outperform the other models is AdaBoost with the accuracy of 96.20 % [35].

So in the above section we have seen the application of time series forecasting in various sectors but in this article the focus is in the Environmental section. Chilika lake is one of the UNESCO World heritage sites and it is rich by means of resources. So it needs to be preserved from unwanted changes in it. One of the main unwanted changes in the lake is the sedimentation process, it can be monitored by using remote sensing. The satellite data can be used for this purpose to find a pattern in this sedimentation process. So let’s look at the data set and try to understand it –

7. Data Set

The remote sensing datasets used in the project is described below.

7.1 Description of Data set

Remote sensing is an efficient tool for hydrological monitoring in large rivers and lakes, as it allows developing algorithms to monitor water quality parameters. The advantage of hydrological monitoring by satellite images in relation to hydrological stations is that the entire surface of the water body can be monitored with high frequency, depending on the temporal resolution of the satellite used and thus saving a lot in operational cost for a monitoring program. This project will use daily reflectance data collected from a multispectral satellite over a period of 20 years for the Chilika Lake ecosystem in India.

This raw data set is then converted in to the form as shown below -

Time	Point_1	Point_2	…….	Point_25	Average
24-02-2000	30.288413	31.881569	…….	23.949831	20.5630089
25-02-2000	30.288413	31.881569	…….	23.949831	20.58753993
. . .	. . .	. . .	. ……. .	. . .	
31-12-2019	9.509447	10.048079	…….	10.401689	8.868939873

Table 7.1 How Data set Looks Like

This data set has a total 27 columns which contains the time value (Means the dates, because the data set is recorded on a daily basis), 25 values which are the pixel values of the raw data and the final column contains the average of these 25 pixel values. Total instances used in collecting the data set i.e. the number of rows are 7247 (Almost 20 year data), Refer to table 7.1. There are some Null values inside the data set so we just have to remove them also in the data preprocessing step.

7.2 Data Preprocessing

In the data pre processing step the imported data set is stored in pandas data Frame and the steps in the data pre processing step can be seen from figure number 7.1. The following steps are as follows -

7.2.1 Null Value handling

Since a major part of the dataset contains the null value so it is needed to deal with them, so we can either remove the whole instance of the dataset which contains the null values but doing so we can also lose some of the useful information given to us. So we will try to impute these null values with some statistical value of the dataset. So we choose linear interpolation for doing so.

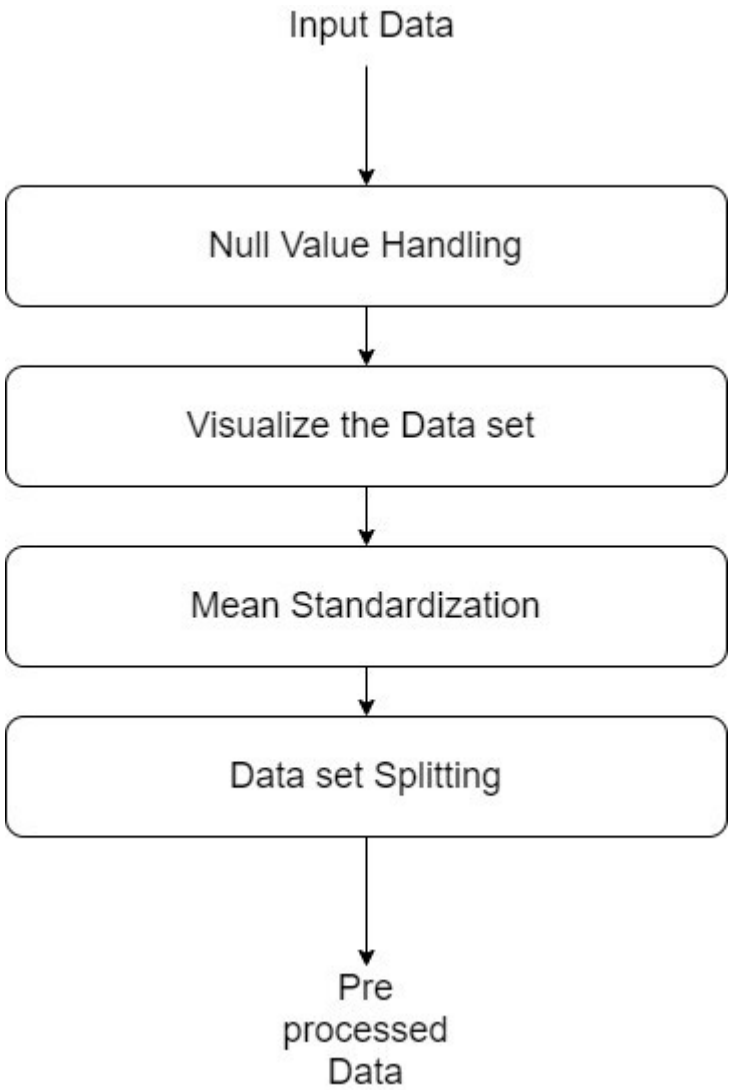

Figure 7.1 Data Pre processing

7.2.2 Visualize the Data Set

Now we have the data set without the null values but to understand the nature of the time series dataset we have to visualize the dataset. Basically we want to check the seasonality and trend inside the data so plotting the box plot monthly as well as yearly is sufficient to notice the trend and seasonality.

If we look at the Box plot of the Average column on monthly and yearly basis in figure 7.2 and figure 7.3 respectively as shown below -

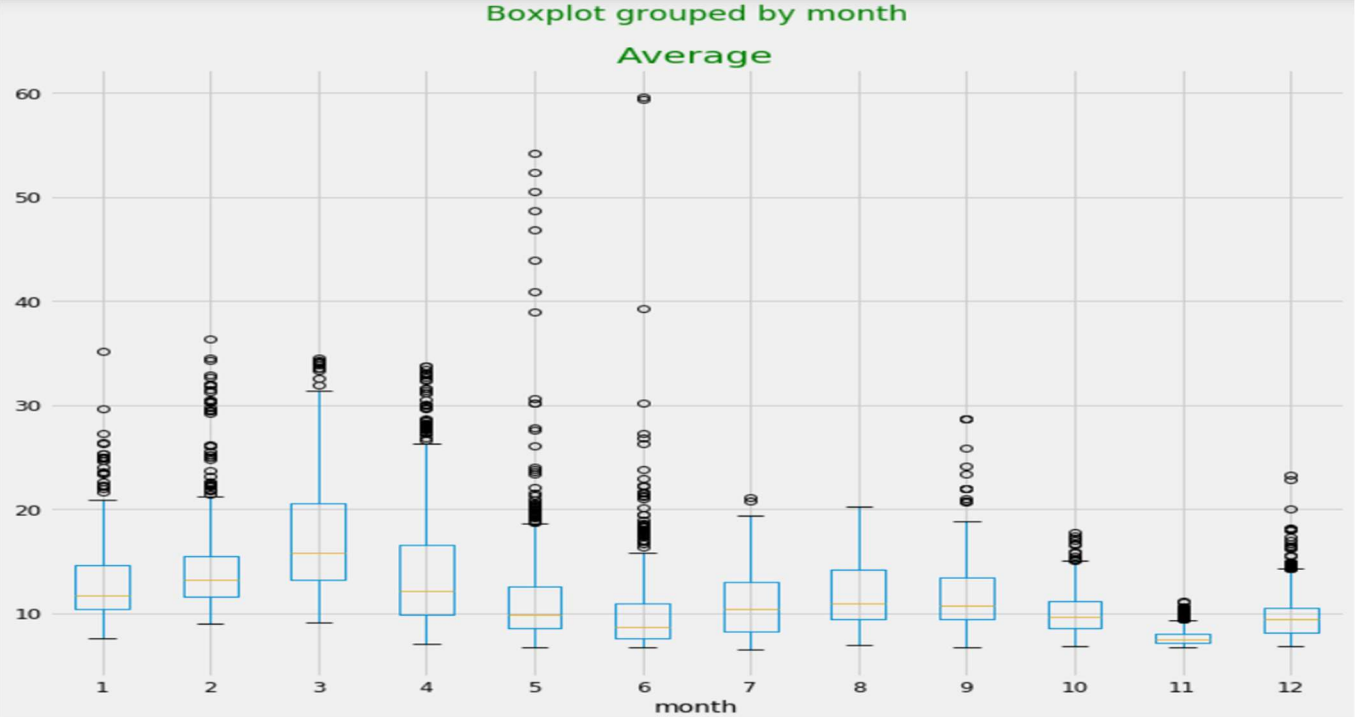

Figure 7.2

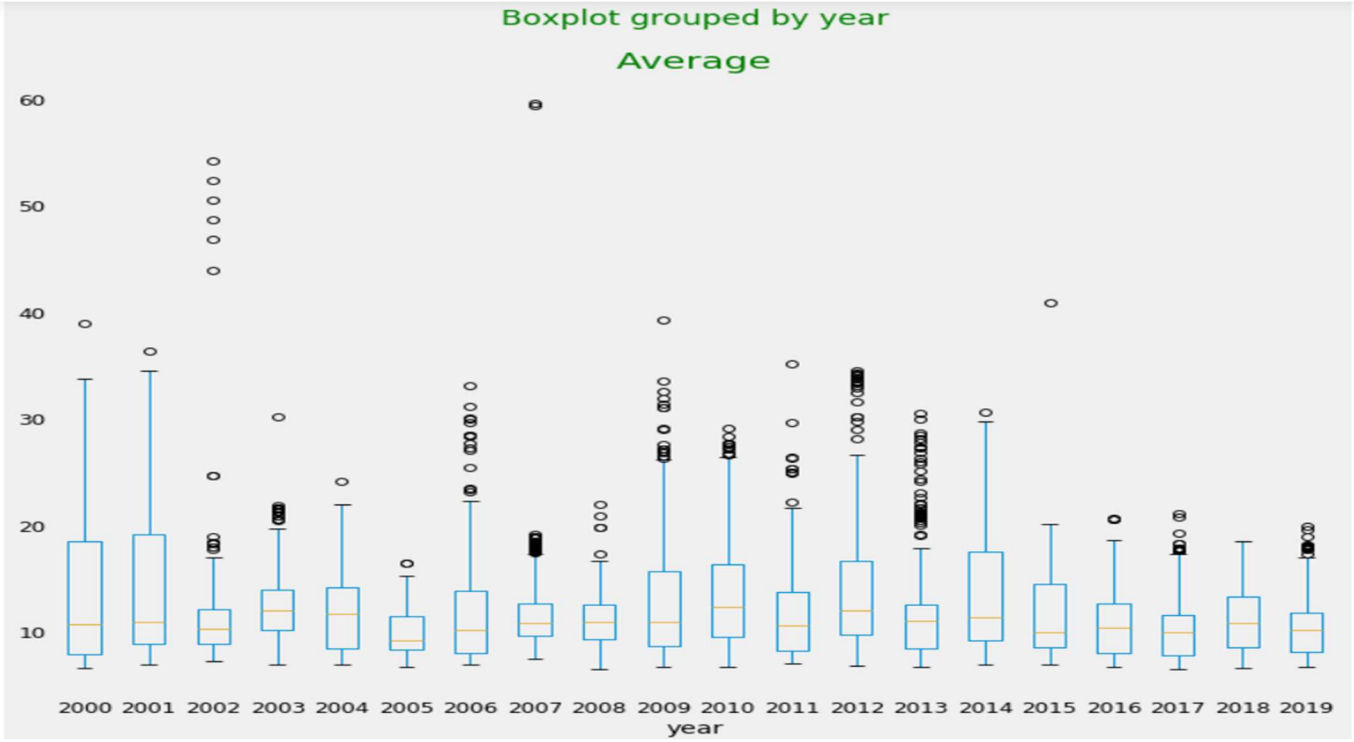

Figure 7.3

Figure 5.2 is telling that surely there is seasonality present in the data but figure 5.3 is telling that trend is absent in the data set, so overall we can conclude that the data set has seasonality but no trend. This information we also used in the imputation of the null values, we prefer the linear interpolation for imputing the null values because the data has seasonality but no trend.

7.2.3 Mean Standardization

Now it is required to normalize the data set to get more accurate results, we choose the mean standardization in which each value is subtracted by the mean of the data and divided by the standard deviation of the data, so final data will have 0 mean and 1 as its standard deviation.
The whole method can be shown as follows.
Suppose X be the given data set of single column or we can say it is a data frame with 1 column
Let is X is not in standard form so it need to be standardized, so final value of X after standardization can be find by following equation -

$$X(Standardized) = \frac{X(Non\ Standardized) - Mean(X(Non\ Standardized))}{Standard\ deviation\ (X(Non\ Standardized))}$$

So finally the mean of the data becomes 0 and the standard deviation of the data becomes 1.

7.2.4 Data splitting

Splitting of the dataset is splitting the data set into training and testing data sets so we choose 80:20 ratio for dividing the dataset which most of the research in the research as well as real time project use.

So we have 20 years(2000 - 2019) of the data we prefer to put starting 16 years(2000- 2015) data into the training set and the rest of the 4 years(2016-2019) data into the test data set.

8. Methodologies

Their are different techniques which can be used in time series forecasting i.e. Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average(ARMA), Autoregressive Integrated moving average (ARIMA), Seasonal Autoregressive Integrated moving average (SARIMA) and LSTM etc. Out of these the first five algorithms are classical/traditional algorithms. Let's look at these algorithms/techniques in detail

8.1 Autoregressive (AR) Model

Autoregression is a time series forecasting algorithm which try to forecast the next values of a variable by feeding it's previous values to the regression equation, it can be explained by the below equation –

$$X(t+1) = b0 + b1 * X(t-1) + b2 * X(t-2)$$

Here b0, b1 and b2 are parameter values and you can see that the value of X for (t+1) timestamp s predicted using the previous timestamp values (i.e. t-1, t-2) of the same variable, since it uses the previous values of the same variable that is why it is called Autoregressive Model [36].

8.2 Moving Average (MA)

Moving Average is a technique useful for observing the long term trend in the data, in this technique the average of subset (containing any number of observations) of observations, it can be understood by taking an example, suppose we have 200 observations and we wanted to calculate the moving average (5 years taken at a time) than we will calculate the average for every 5 values for each observation [37].

8.3 Autoregressive Moving Average(ARMA)

Autoregressive Moving Average considers both the Autoregression as well as Moving average, it has one polynomial function for Autoregression and one polynomial function for autoregression. It is used for considering weakly stochastic time series.

8.4 Autoregressive Integrated Moving Average(ARIMA)

The ARIMA Model is similar to the ARMA Model but the only difference in these two models is that the "Integrated" part which means the number of times needed to difference a series in order to achieve stationarity in it.

8.5 Seasonal Autoregressive Integrated Moving Average(SARIMA)

SARIMA Model are useful for the time series data set which is not stationary but seasonality in the data is present, the only difference between the ARIMA Model and SARIMA Model is that it can consider the seasonality factor as well.

8.6 Long Short Term Memory (LSTM)

Long Short Term Memory Network (LSTM), they are a special kind of Recurrent neural network (RNN) which are also called “LSTMs”. They are mainly used because of their capability of learning long dependencies. They Introduced in 1997 by Hochreiter & Schmidhuber.

Like RNN they have chain-like structures containing the repeated units of the neural network but the difference is that in LSTM these units are not simple but they contain 4 activation layers[38].

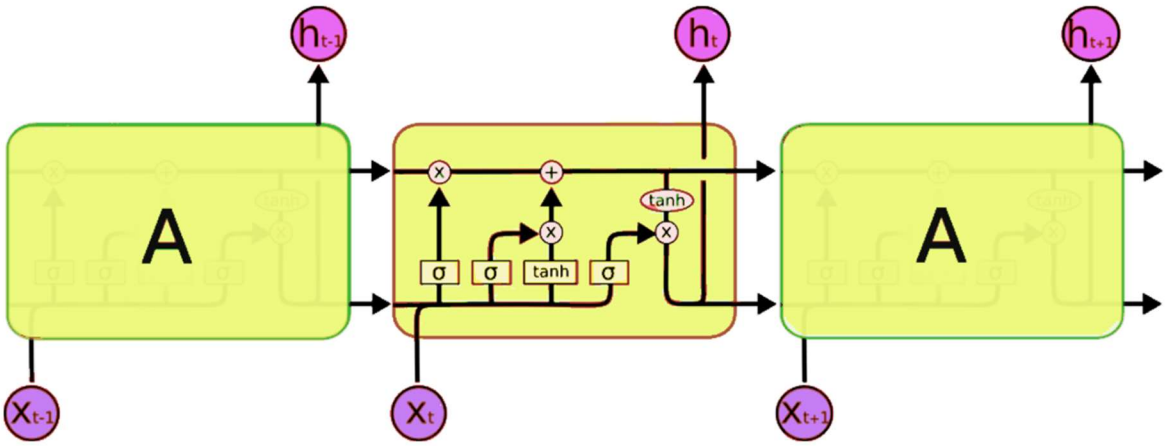

Figure 8.1 LSTM [39]

It can be seen clearly in the figure 8.1 that the LSTM are not as simple as RNN and due to these 4 activation functions they can learn the long dependencies.

9. Result and Discussion

The basic architecture of the system can be understood and visualized from Figure 9.1 as shown below. After importing the data set the data pre processing step is performed as discussed earlier in which Null value handling, data visualization and mean standardization is performed, in this step the data splitting is also performed, the data will be in processed form after these steps.

After the data is in processed form the modelling is performed on the data in which the models used are traditional/classical time series models such as Autoregression(AR), Moving Average(MA), ARMA, ARIMA and SARIMA and LSTM is also applied on the data and than results are compared on the basis of RMSE.

The machine learning models applied are Autoregression(AR), Moving Average(MA), ARMA, ARIMA and SARIMA and LSTM and the result of these models can be understand and visualized from the graphs in the Figure 9.2, 9.3, 9.5, 9.6 and 9.7. It can be visualized from these graphs that models are tested on the turbidity value of year 2016, 2017, 2018 and 2019 and prediction is made for year 2016, 2017, 2018, 2019 and 2020, so the prediction is made for one extra year.

The graphs shown in the above mentioned figure shows that models perform good as the predicted values in each case are near to the actual value in each model and the RMSE value of models is mentioned in the table 9.1, the RMSE value of model AR, ARMA and ARIMA comes out to be best i.e 0.027 and other models have RMSE value 0.155 for MA, 0.029 for SARIMA and 0.169 for LSTM.

Models	AR	MA	ARMA	ARIMA	SARIMA	LSTM
RMSE	0.027	0.155	0.027	0.027	0.029	0.169

Table 9.1

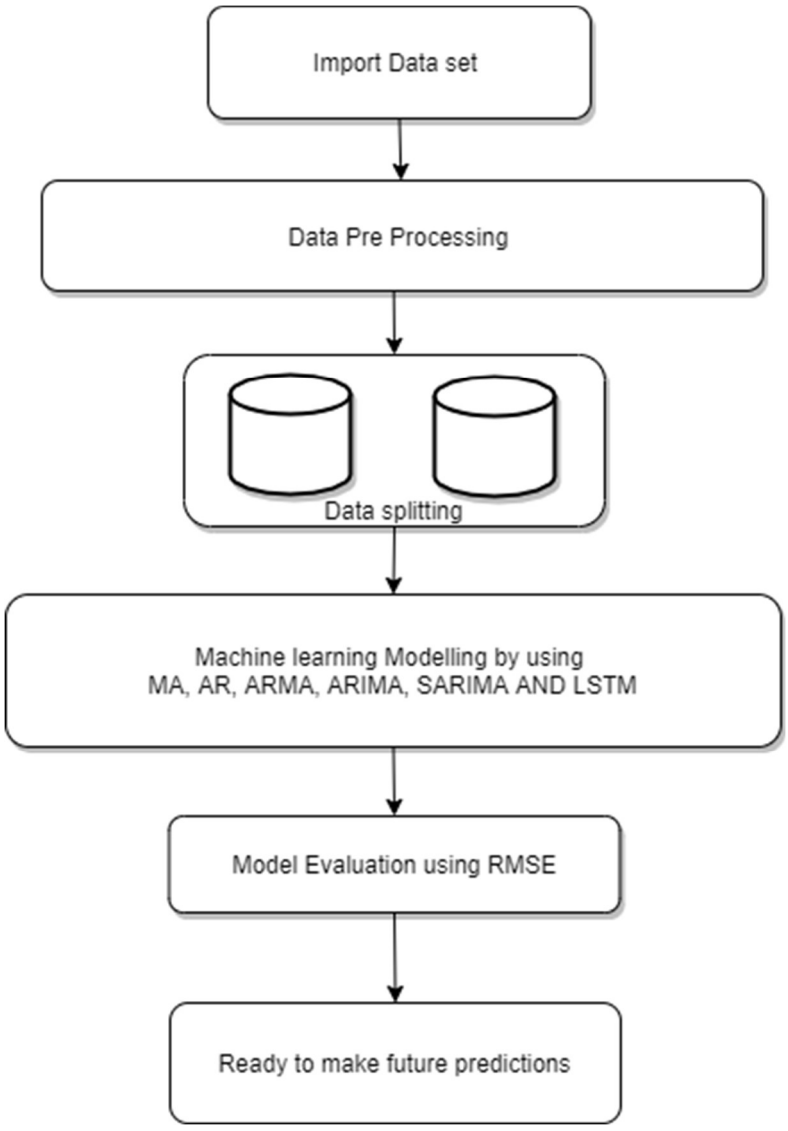

Figure 9.1 Model Architecture

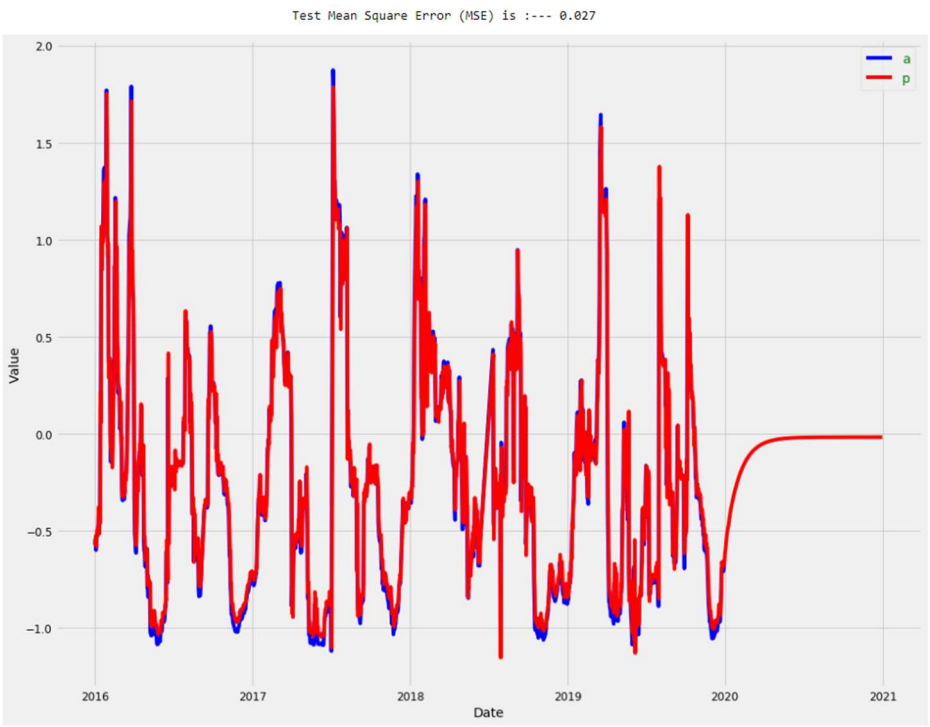

Figure 9.2 Auto Regression

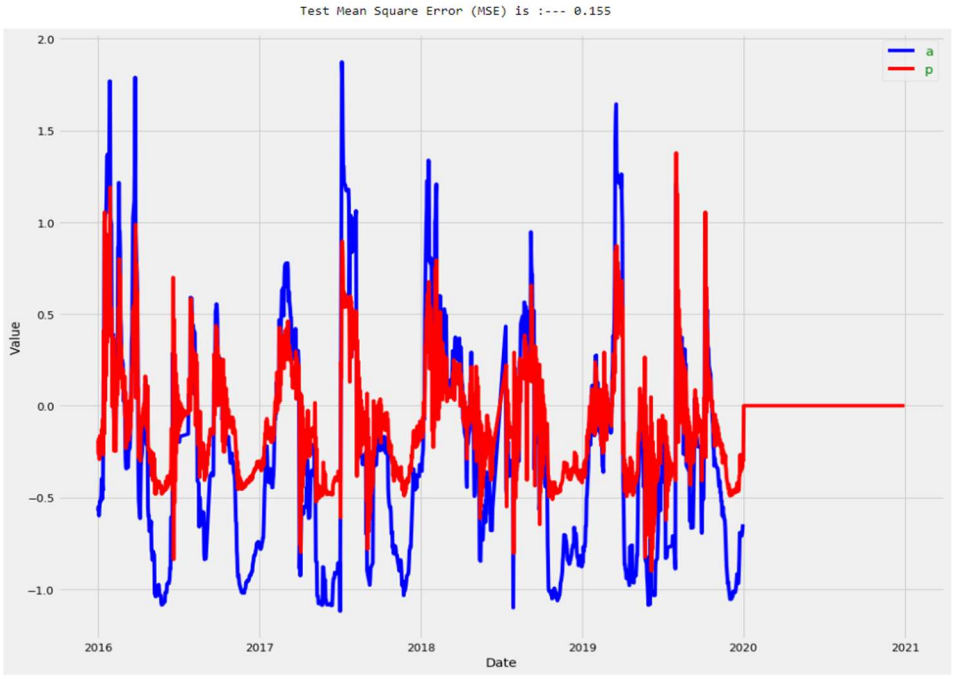

Figure 9.3 Moving Average

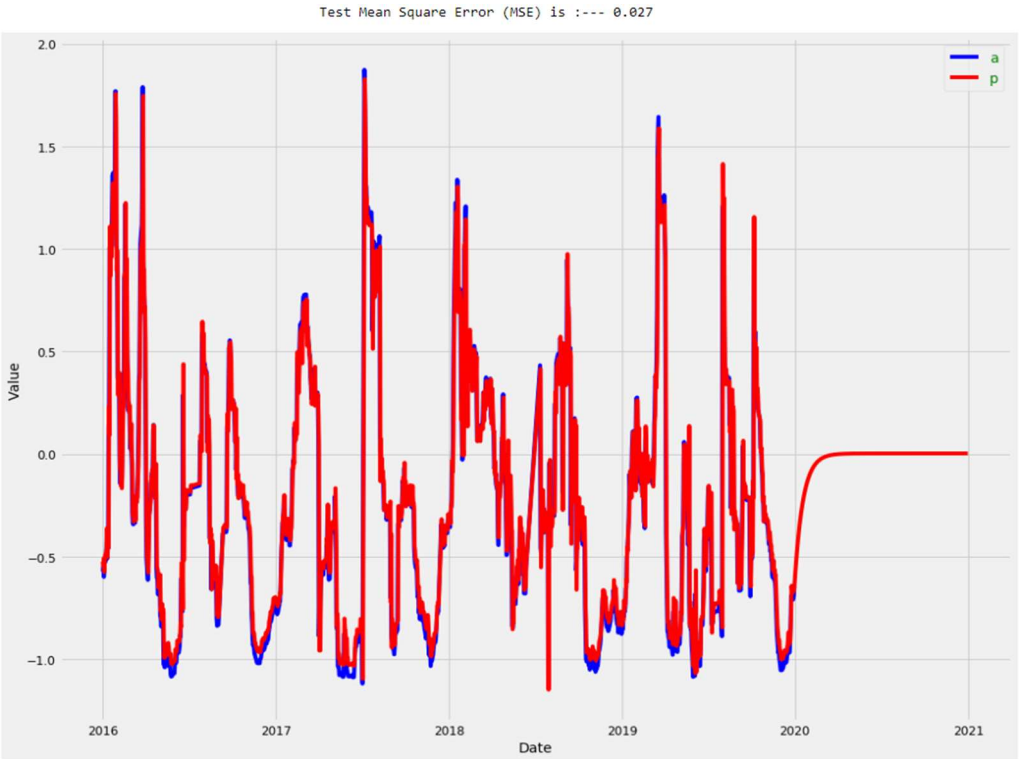

Figure 9.4 ARMA

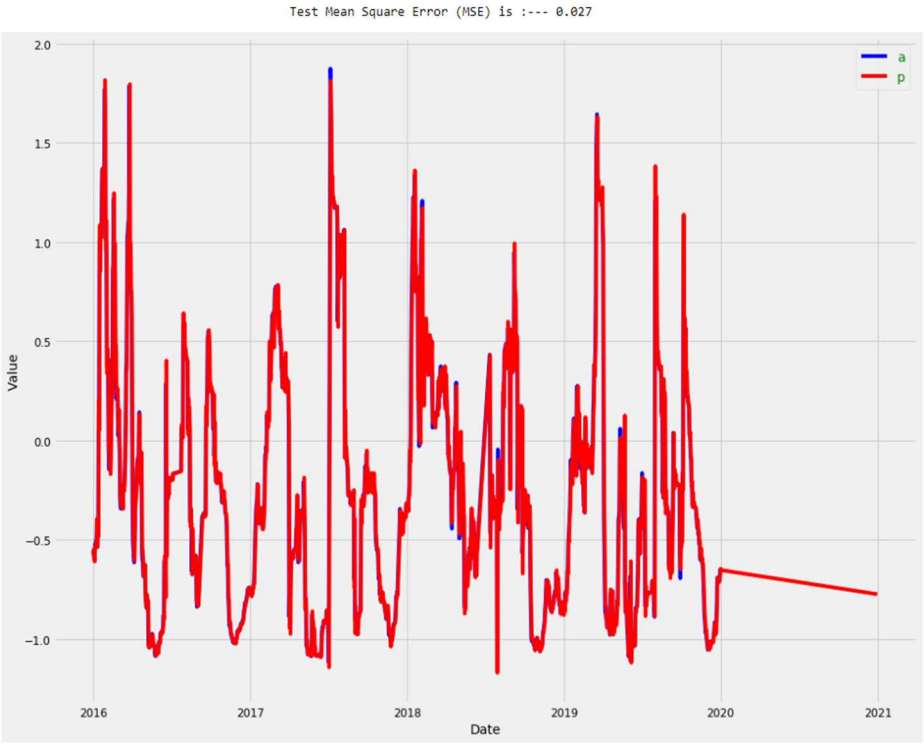

Figure 9.5 ARIMA

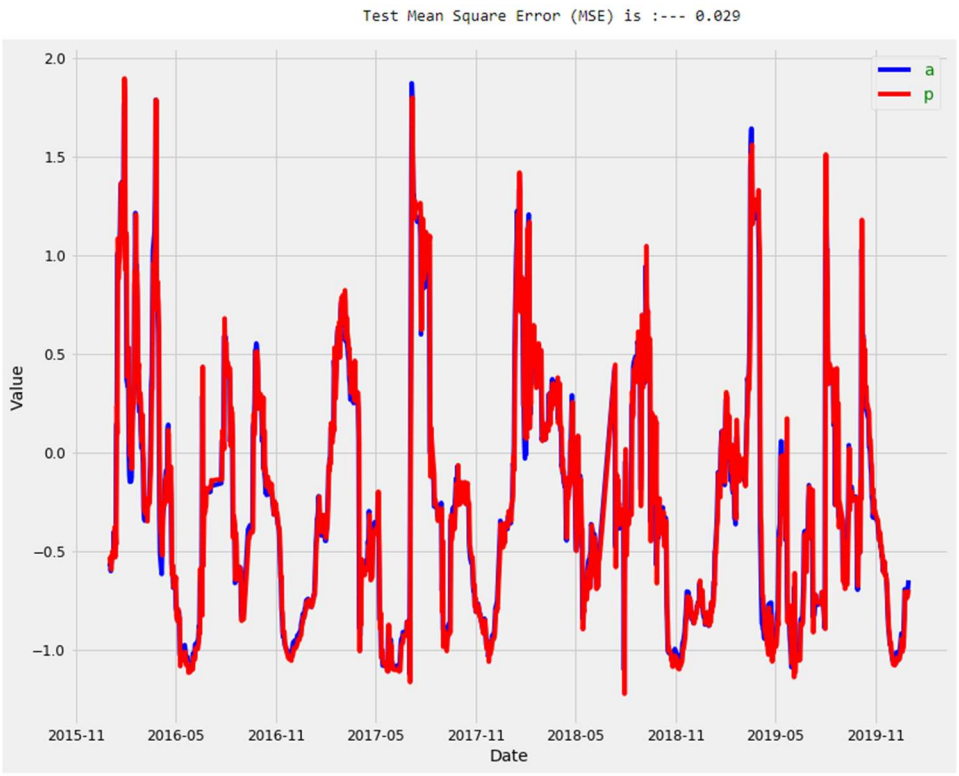

Figure 9.6 SARIMA

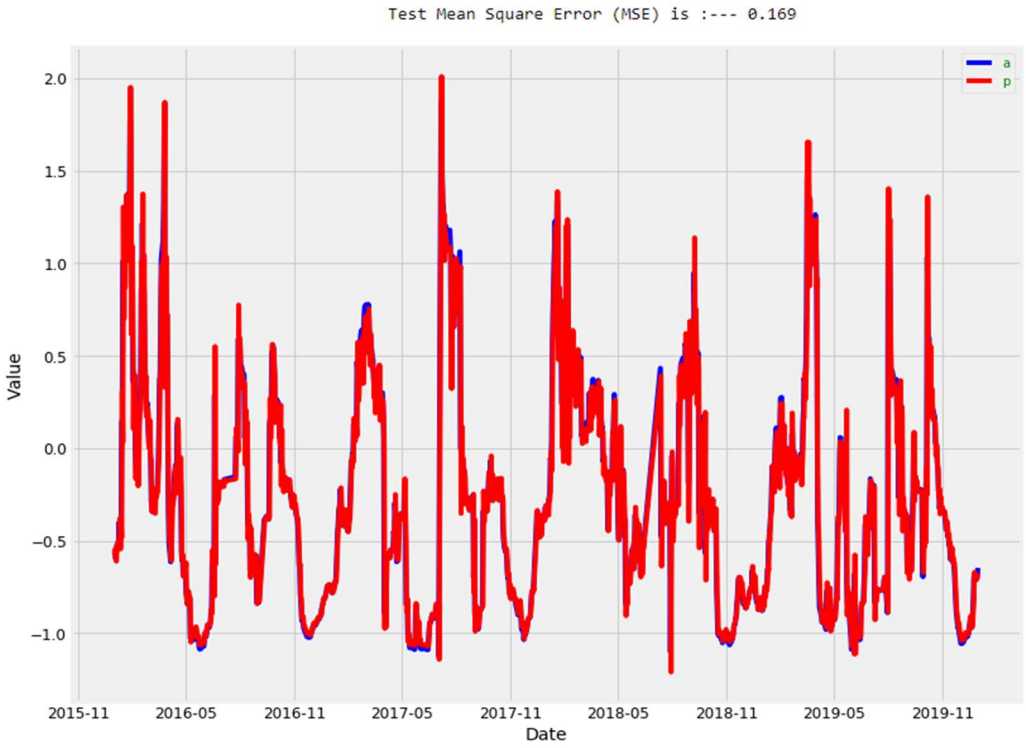

Figure 9.7 LSTM

10. Conclusion

The results above showed that machine learning models can be able to predict the turbidity in the wetlands, the most accurate models are AR, ARMA and ARIMA with RMSE value of them comes out to be 0.027. So it will be best to predict the turbidity so that it can be controlled using a manual team.

References

[1] https://whc.unesco.org/en/tentativelists/5896/, Accessed on Sept 1, 2020.

[2] Okujeni, A., van der Linden, S., & Hostert, P. (2015). Extending the vegetation–impervious–soil model using simulated EnMAP data and machine learning. Remote Sensing of Environment, 158, 69-80.

[3] Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environmental monitoring and assessment, 188(1), 44.

[4] Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, 265, 62-77.

[5] Park, S., Im, J., Jang, E., & Rhee, J. (2016). Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions. Agricultural and forest meteorology, 216, 157-169.

[6] Appelhans, T., Mwangomo, E., Hardy, D. R., Hemp, A., & Nauss, T. (2015). Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spatial Statistics, 14, 91-113.

[7] Xie, G. Y., Olson, D. H., & Blaustein, A. R. (2016). Projecting the global distribution of the emerging amphibian fungal pathogen, Batrachochytrium dendrobatidis, based on IPCC climate futures. PLOS one, 11(8), e0160746.

[8] Masood, M. K., Soh, Y. C., & Chang, V. W. C. (2015, July). Real-time occupancy estimation using environmental parameters. In 2015 international joint conference on neural networks (IJCNN) (pp. 1-8). IEEE.

[9] Hallgren, W., Beaumont, L., Bowness, A., Chambers, L., Graham, E., Holewa, H., ... & Vanderwal, J. (2016). The biodiversity and climate change virtual laboratory: where ecology meets big data. Environmental Modelling & Software, 76, 182-186.

[10] Lary, D. J., Lary, T., & Sattler, B. (2015). Using machine learning to estimate global PM2. 5 for environmental health studies. Environmental health insights, 9, EHI-S15664.

[11] Lou, S., Li, D. H., Lam, J. C., & Chan, W. W. (2016). Prediction of diffuse solar irradiance using machine learning and multivariable regression. Applied energy, 181, 367-374.

[12] O'Connor, C. D., Calkin, D. E., & Thompson, M. P. (2017). An empirical machine learning method for predicting potential fire control locations for pre-fire planning and operational fire management. International journal of wildland fire, 26(7), 587-597.

[13] Prasad, R., Deo, R. C., Li, Y., & Maraseni, T. (2018). Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. Geoderma, 330, 136-161.

[14] Chatziantoniou, A., Psomiadis, E., & Petropoulos, G. P. (2017). Co-Orbital Sentinel 1 and 2 for LULC mapping with emphasis on wetlands in a mediterranean setting based on machine learning. Remote Sensing, 9(12), 1259.

[15] Forkuor, G., Hounkpatin, O. K., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. PloS one, 12(1), e0170478.

[16] Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters. Journal of Electrical and Computer Engineering, 2017.

[17] Shafizadeh-Moghadam, H., Valavi, R., Shahabi, H., Chapi, K., & Shirzadi, A. (2018). Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. Journal of environmental management, 217, 1-11.

[18] Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F., & Kløve, B. (2018). River suspended sediment modelling using the CART model: a comparative study of machine learning techniques. Science of the Total Environment, 615, 272-281.

[19] Bunn, C., Läderach, P., Rivera, O. O., & Kirschke, D. (2015). A bitter cup: climate change profile of global production of Arabica and Robusta coffee. Climatic Change, 129(1-2), 89-101.

[20] Behmann, J., Mahlein, A. K., Rumpf, T., Römer, C., & Plümer, L. (2015). A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. Precision Agriculture, 16(3), 239-260.

[21] Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., & Bédard, F. (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agricultural and forest meteorology, 218, 74-84.

[22] Kuwata, K., & Shibasaki, R. (2015, July). Estimating crop yields with deep learning and remotely sensed data. In 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 858-861). IEEE.

[23] Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003.

[24] Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications surveys & tutorials, 18(2), 1153-1176.

[25] Beaver, J. M., Borges-Hink, R. C., & Buckner, M. A. (2013, December). An evaluation of machine learning methods to detect malicious SCADA communications. In 2013 12th International Conference on Machine Learning and Applications (Vol. 2, pp. 54-59). IEEE.

[26] Wu, M., Song, Z., & Moon, Y. B. (2019). Detecting cyber-physical attacks in CyberManufacturing systems with machine learning methods. Journal of intelligent manufacturing, 30(3), 1111-1123.

[27] Wang, S., Liu, C., Gao, X., Qu, H., & Xu, W. (2017, September). Session-based fraud detection in online e-commerce transactions using recurrent neural networks. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 241-252). Springer, Cham.

[28] Subroto, A., & Apriyana, A. (2019). Cyber risk prediction through social media big data analytics and statistical machine learning. Journal of Big Data, 6(1), 50.

[29] Shrivastava, A., Sondhi, J., & Kumar, B. (2017). MACHINE LEARNING TECHNIQUE FOR PRODUCT CLASSIFICATION IN E-COMMERCE DATA USING MICROSOFT AZURE CLOUD.

[30] Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: lexicon versus machine learning. Journal of Consumer Marketing.

[31] Isaac, G., Meacham, S., Hamzeh, H., Stefanidis, A., & Phalp, K. T. (2018, June). An adaptive E-commerce application using web framework technology and machine learning. BCS.

[32] Manogaran, G., Vijayakumar, V., Varatharajan, R., Kumar, P. M., Sundarasekar, R., & Hsu, C. H. (2018). Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. Wireless personal communications, 102(3), 2099-2116.

[33] Poostchi, M., Silamut, K., Maude, R. J., Jaeger, S., & Thoma, G. (2018). Image analysis and machine learning for detecting malaria. Translational Research, 194, 36-55.

[34] Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., ... & Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ, 6, e4568.

[35] Saiprasath, G. B., Babu, N., ArunPriyan, J., Vinayakumar, R., Sowmya, V., & Soman, K. P. (2019). Performance comparison of machine learning algorithms for malaria detection using microscopic images.

[36]https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/, Accessed on Sept 1, 2020.

[37]https://www.statisticshowto.com/moving-average/#:~:text=A%20moving%20average%20is%20a,for%20any%20period%20of%20time., Accessed on Sept 1, 2020.

[38] https://colah.github.io/posts/2015-08-Understanding-LSTMs/, Accessed on Sept 1, 2020.

[39] https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/, Accessed on Sept 1, 2020.