COMPARATIVE STUDY OF MACHINE LEARNING AND DEEP LEARNING FOR AIR QUALITY INDEX PREDICTION

MINOR PROJECT II

Submitted by:

HARJOT SINGH PAHWA (9916103149)

PRIYAL NARANG (9916103170)

RISHIKA ARORA (9916103240)

Under the supervision of:

MR. HIMANSHU AGRAWAL

Department of CSE/IT

Jaypee Institute of Information Technology University, Noida

MAY 2019

ABSTRACT

In recent years, many countries have experienced rapid urbanization. This has led to increased air pollution level and has created a major concern among the developing countries. Therefore, demand for predicting air quality has become a necessity so that we take precautionary measures to control the air pollution levels.

This project is one specific contribution towards the issue of predicting air quality. The dataset we considered consists of major air pollutants as well as the meteorological data. We have used the standard formula followed by US EPA government for calculating the AQI using major pollutants and their concentrations. We also study the correlation and variation of one pollutant with respect to other pollutants.

We have applied machine learning algorithms like Multiple Linear Regression, Linear SVM, K-Nearest Neighbors and propose to apply Deep Learning models like CNN and LSTM to predict AQI (Air Quality Index. In LSTM, we will use past AQI and meteorological data to predict the next AQI value. The lesser the difference between the predicted and actual AQI value suggests a better model. We will further compare the above mentioned models to deduce the most efficient model. The results of the models were analyzed using evaluation metrics such as precision, recall, f1-score and accuracy. Among the models that we used, KNN outperformed the other two models by giving us a precision, recall, f1score, accuracy of 0.77, 0.82, 0.76, 81.77 % respectively.

For Deep Learning Models we used the LSTM and ANN network. Using LSTM we obtained good results where the r2 value was 0.92 but that took a long time to train. In contrast we used the ANN model which ran faster, nearly completed training in 2/3 of the time it took for LSTM to complete with a r2 score of 0.91.

ACKNOWLEDGEMENT

I would highly like to place on record my deep sense of gratitude to **Mr. Himanshu Agrawal**, Assistant Professor at, Jaypee Institute of Information Technology, Noida for his generous guidance and constant supervision as well as for providing necessary information regarding the project and also her support in completing this work.

I express my sincere gratitude to faculty, Dept. of Computer Science, and project evaluators for their stimulating guidance, continuous encouragement and supervision throughout the course of present work.

I also extend my thanks to my college for providing adequate resources for the project, my group members and seniors for their insightful comments and constructive suggestions to improve the quality of this project work.

Signature:

Harjot Singh Pahwa (9916103149)

Priyal Narang (9916103170)

Rishika Arora (9916103240)

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS AND NOMENCLATURE

AQI – Air Quality Index

SVM - Support Vector Machine

KNN – K-Nearest Neighbors

CNN – Convolution Neural Network

LSTM- Long Short Term Memory

SO_2- Sulphur Dioxide

NO_2- Nitrogen Dioxide

PM 2.5- Particulate Matter

O_3- Ozone

CO- Carbon Monoxide

Chapter 1: INTRODUCTION

Due to rapid urbanization, many environmental hazards took place in 20th century, including rise in air pollution levels. Air pollution is the introduction of chemicals, particulate matter, or biological matter that cause harm or discomfort to living organisms or damage the natural environment or atmosphere. Air pollutants are tiny and light particles and thus they stay in the atmosphere for long duration and also easily bypass the filters of human nose and throat due to their small size. According to a recent survey, the presence of particulate matter has caused 4.2 million deaths. Major air pollutants like Sulphur Dioxide (SO2), Nitrogen Dioxide (NO2), Carbon Monoxide (CO), Particulate Matter (PM2.5, PM10) and Ozone(O3) have drastic effects on human health. Thus, predicting the air quality has become a major concern.

Particle size is critical in determining the particle deposition location in the human respiratory system. PM2.5, referring to particles with a diameter less than or equal to 2.5 µm, has been an increasing concern, as these particles can be deposited into the alveoli- the lung gas exchange region. The U.S. EPA revised the annual standard of PM2.5 by lowering the concentration to 12 µg/m3 to provide improved protection against health effects associated with long- and short-term exposure. Increased mortality and morbidity rates have been found in association with increased air pollutants. Thus, we considered PM2.5 as the label for classification.

Meteorological data is critical in determining the air pollutant consideration. The meteorological parameters considered by our model include: temperature, wind, relative humidity, dew point and pressure. Temperature affect air quality because of temperate inversion: the warm air above cooler air acts like a lid, suppressing vertical mixing and trapping the cooler air at the surface. As pollutants from vehicles, fireplaces, and industry are emitted into the air, the inversion traps these pollutants near the ground. Wind speed plays a big role in diluting pollutants. Generally, strong winds disperse pollutants, whereas light winds generally result in stagnant conditions allowing pollutants to build up over an area. Humidity could affect the diffusion of contaminant. Dew point indicates the amount of the moisture in the air. The higher the dew point the higher moisture content in the air at a given temperature. Dew point and the concentration of the air pollutants are inversely proportional. Pressure is also inversely proportional to the air quality.

The study of AQI started with the application of statistical models, but it was mostly been restricted to simply utilizing standard classification or regression models, which have neglected the nature of the problem itself or ignored the correlation between sub-models in different timeslots. The development of machine learning, helped to refine the model of a specific problem. This model predicts the levels of PM2.5 daily with the help of the meteorological data. For this we applied machine learning algorithms like SVM and deep learning techniques like ANN and LSTM. The model then compares and analyzes the results using the evaluation metrics and mean squared error. The model also helps us to study the correlation between various meteorological parameters and PM2.5.

LSTM is a specific recurrent neural network that doesn't only take the current input, but also what it has "perceived" previously in time, essentially using the output at time t − 1 as an input to time t, along with the new input at time t. In our case, it is understood that the air quality of the current day does depend on the air quality of the previous day. But if we use right number of hidden layers and activation layers in the ANN we obtain slightly better results in same period of time. So we can say that our ANN proved to be the best among the other models for pollution prediction.

Chapter 2: BACKGROUND STUDY

Air Quality Index prediction problem has been in limelight in recent years, due to rapid increase in air pollution. A lot of work has been done in this field to take appropriate measures to maintain the air quality.

One of the initial work by Wenjian Wang[1] presented examined the feasibility of applying SVM to predict pollutant concentrations. It presented a pioneer study of using SVM to investigate the concentration variations of six pollutants hourly measured in 1999 at the Causeway Bay Roadside Gaseous Monitory Station established by Hong Kong Environment Protection Department (HKEPD) . Six major pollutants, i.e., sulphur dioxide (SO*), nitrogen oxides (NO,), nitric oxide (NO), nitrogen dioxide (NOI), carbon monoxide (CO) and respirable suspend particles (RSP), were monitored hourly.

In feed-forward neural networks, back propagation (BP) algorithm is usually used in predicting pollutant concentration levels, some inherent drawbacks like slow convergence speed, less generalizing performance makes it difficult to put these models into practice. RBF network possesses good generalizing performance and, in the meantime, can avoid over-elaborated, lengthy computing like BP algorithm. However like other NN models, there exists issues of local minima and over-fitting problems relating to RBF . An adaptive structure of RBF was built. The number of hidden nodes were not fixed. The target error was 0.05, and the width value of the RBF network was kept 1.0.

In this paper, the Gaussian function was used as the kernel function of the SVM. Both C and **ε** were arbitrarily chosen as 100 and 10^{-3} respectively. The model was evaluated using the statistical metrics i.e., Mean Absolute Error (MAE) which is a measure of the deviation between the actual and the predicted values.

The results have shown that SVM performed better than RBF. Therefore, it can be concluded that although the meteorological variables were ignored, the SVM method possessed some advantages to RBF network a due to its features of noise-tolerance, high stability, adaptive properties and better generalization performance.

In another research by Kostandina Veljanovska[2] air quality index was predicted by comparing three supervised learning algorithms k-nearest neighbor (k-NN), Support Vector Machines (SVM) and Decision Tree (DT) and one unsupervised algorithm Neural Network (NN). Dataset that was used was based on model of official web site of Ministry of environment and physical planning of the Republic of Macedonia. Major pollutants considered were SO2 (sulphur dioxide), NO2(nitrogen dioxide), O3 (ozone), CO (carbon monoxide), suspended particulates PM2.5 (fine particles) and PM10 (large particles).

Dataset contained 365 samples (each per day of 2017), 51 of this samples were with High Air Pollution Index, other 94 were with Medium Air Pollution Index and the rest 220
samples were with Low Air Pollution Index.

Neural Network in this project classified the samples as “low”, “medium” and “high” levels of air pollution. The six pollutants were used as inputs and the air quality index(output) was classified as low, medium and high. The hidden layer of the neural net consisted of 10 neurons as it resulted in least optimal error.

For the k-NN classifier different values of k in the range 1-21 were taken. The model was tested using several types of metrics: Euclidean, correlation, city block and cosine. The best
accuracy obtained was for k=3 with Euclidean metric. Algorithm classified samples as “high level of pollution” as follows, 17 correct and 8 in wrong (in medium class), “low level of pollution”: 32 correct and 5 wrong (in medium class) and in class of “medium level of pollution” 31 correct samples while others belonged to high and low class.

For the experiments, DT was constructed with assumption that all input functions have final discrete domains, there is one target function classification of air
pollution data (three classes). Every internal node in the DT contained input feature. Every leaf of the tree contained the class.

In SVM algorithm different kernel functions like linear, quadratic, cubic, Fine Gaussian, Medium Gaussian and Coarse Gaussian were tried to get highest accuracy result. According to the results, maximum accuracy of 80% was obtained when linear kernel function was used.
According to conclusions, the most accurate algorithm was NN with maximum accuracy of 92.3%, while KNN with a maximum accuracy of 80.0%, DT algorithm with maximum
accuracy of 78.0% and SVM algorithm with maximum accuracy of 80.0%.

Some of the recent research works consists of deep learning and spatial models. The paper by İbrahim KÖK,Mehmet[3] suggests a novel deep learning model was proposed for analyzing IoT smart city data. They proposed a novel model based on Long Short Term Memory (LSTM) networks to predict future values of air quality in a smart city. The evaluation results of the proposed model were found to be promising and they show that the model can be used in other smart city prediction problems as well.
The main contribution of this paper is two-fold: First, they presented a DL model that can be applied to SC IoT data. Second, they designed and implement a novel LSTM based prediction model that is helpful to solve future air quality problems in mart Cities.
The proposed air quality prediction model consisted of three parts. In the first part, a deep learning model consisting of LSTM neural networks was realized. In the second part, a labelling unit was created that labels data according to the daily AQI values. In the last part, a decision unit was developed which maps according to the observed and predicted alarm situations. The proposed model was trained on the Python platform by using TensorFlow and Keras DL framework. Trainings were carried out by SVR and using separate LSTM models for each type of gas. In the process of LSTM training, Adaptive Moment Estimation (Adam) optimizer which computes individual adaptive learning rate for different parameters, was used to minimize the loss function.
Consequently, the obtained results show that the employment of the LSTM based prediction model to the IoT data is effective and promising and also leaves a scope for future study and implementation of other algorithms on the data.

A semantic ETL (Extract-Transform-Load) framework on cloud platform for AQI prediction was proposed by Dan Wei[4]. The core of the ETL framework is to transform retrieved data into required instance and format. In this layer, there are three parts: Data Preprocessing, Semantic Data Model, and
Semantic Data Analytics, respectively. In the platform, they exploited ontology to concretize the relationship of PM 2.5 from various data sources and to merge those data with the same concept but different naming into the unified database. They implement the ETL framework on the cloud platform, which includes computing nodes and storage nodes.
The computing nodes were used to execute data mining algorithms for predicting, and storage modes were used to store retrieved, preprocessed, and analyzed data. They utilized restful web service as the front end API to retrieve analyzed data, and finally they exploited browser to show the visualized result to demonstrate the estimation and prediction.
The analyzing and mining algorithms were implemented using some well-known algorithms in
Apache Spark’s 1 big data framework.In this work, they exploit various statistical and machine learning methods to construct the prediction model for predicting the PM2.5 value such as Decision tree,RNN and ETL. The ETL model showed that the big data access framework on the cloud platform can work well for air quality analysis.

Chapter 3: REQUIREMENT ANALYSIS

3.1 Software Requirements

3.1.1 Libraries used:

- Numpy
- Pandas
- matplotlib
- Sklearn

3.1.2 Other Requirement:

- Anaconda Platform (spyder, Jupyter Notebook)
- Python 3.6.0

3.2 Hardware Requirements

- Microsoft Windows 10
- Processor: Intel ® Core (TM) i5 -6200U CPU @2.30GHz 2.40GHz
- Ram : 4 GB and above
- Disk Space : 1 TB

3.3 Functional Requirements

- Calculation of AQI values using pollutant concentrations.
- Analyzing the relation between the air pollutants and meteorological data.
- Predicting the AQI values using the past data.

3.4 Non-functional Requirements

- Accuracy: It measures the ratio of correct predictions to the total number of data points evaluated. The Machine Learning models must have a significant amount of accuracy in prediction of apps.
- Recall: It is the fraction of relevant occurrences that have been retrieved over total amount of relevant occurrences. It is also known as the sensitivity of the model.
- Precision: It is the fraction of relevant occurrences among the retrieved occurrences. It is also referred as positive predictive value.
- F1-score: It is a measure that combines precision and recall by taking its harmonic mean.

Chapter 4: DETAILED DESIGN

The dataset we used was obtained from three sources. First, meteorological and air pollution data from 2010 to 2014 from Li et. al (2014), published as a UCI dataset. Second, we took air pollution data from Dr. Xiaojing Yao for the years 2015 to 2017. We then got meteorological data from 2015 to 2017 from the US NOAA (National Oceanic and Atmospheric Administration). Finally, we built a parser in Python to extract weather data from their archived format.

The features of our dataset are cumulative rain hours, cumulative snow hours, wind speed, wind direction, dew point, air temperature, air pressure, date, year, month, day, and hour .The target variable is pollution measured in PM 2.5 .

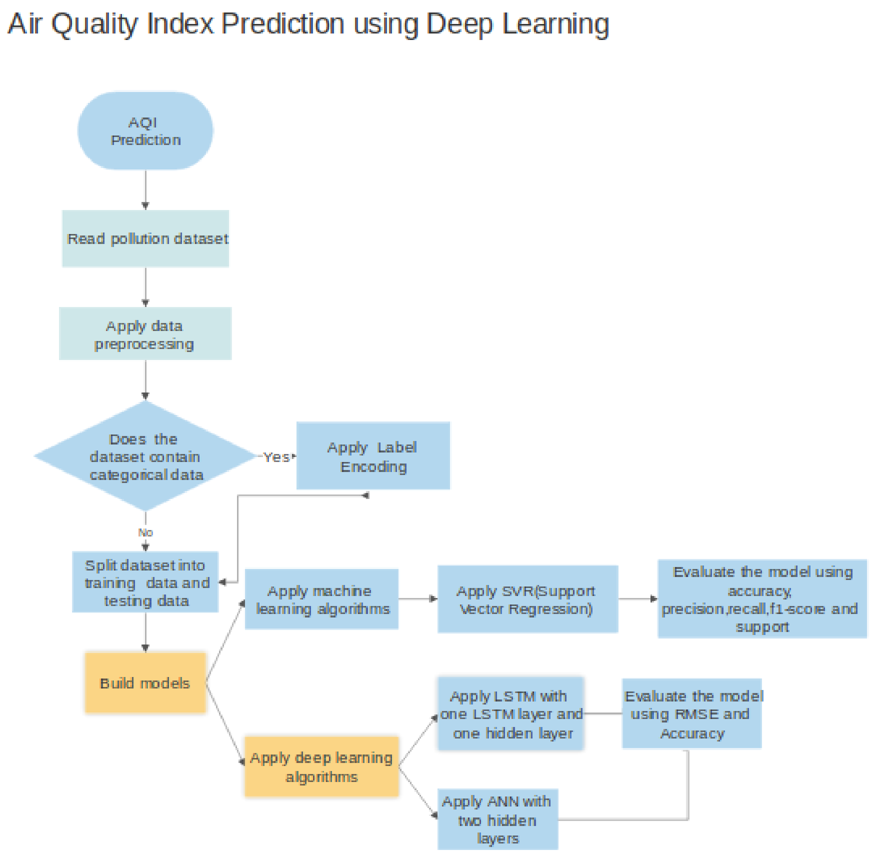

Air Quality Index Prediction using Deep Learning
AQI Prediction
Read pollution dataset
Apply data preprocessing
Does the dataset contain categorical data
Yes
Apply Label Encoding
No
Split dataset into training data and testing data
Build models
Apply machine learning algorithms
Apply SVR(Support Vector Regression)
Evaluate the model using accuracy, precision,recall,f1-score and support
Apply deep learning algorithms
Apply LSTM with one LSTM layer and one hidden layer
Apply ANN with two hidden layers
Evaluate the model using RMSE and Accuracy

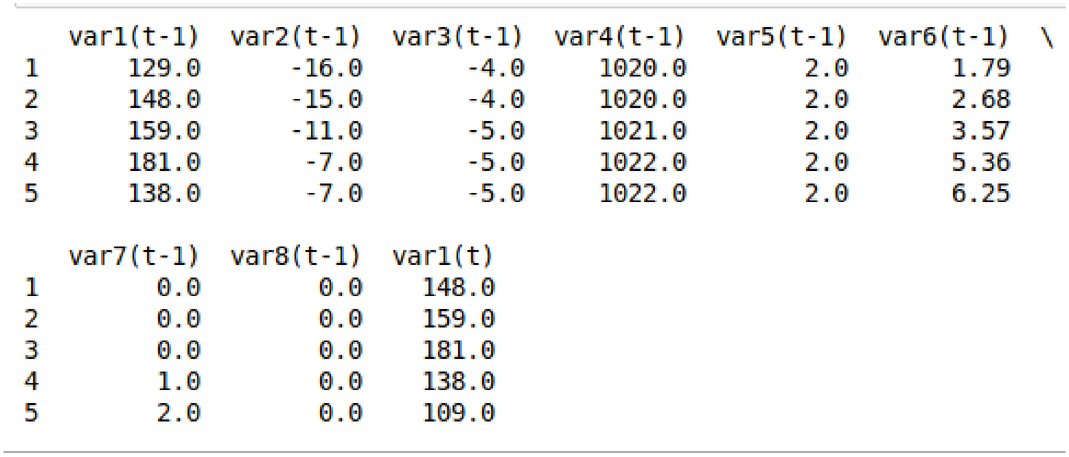
Figure 4.1: Detailed Design

Chapter 5: IMPLEMENTATION

In implementation we have performed the following tasks to get dataset on which we applied machine learning and deep learning algorithms.

5.1 Data Pre-processing

We preprocessed and converted each dataset, which had hourly information, to a time series so that it can be used towards solving a supervised learning problem. We removed columns such as date, hour, year and month. All NA values were replaced with 0. We then encoded the categorical feature direction using Label Encoder. We applied feature scaling using Min-Max Scaler.



```
   var1(t-1)  var2(t-1)  var3(t-1)  var4(t-1)  var5(t-1)  var6(t-1)  \
1      129.0      -16.0       -4.0     1020.0        2.0       1.79
2      148.0      -15.0       -4.0     1020.0        2.0       2.68
3      159.0      -11.0       -5.0     1021.0        2.0       3.57
4      181.0       -7.0       -5.0     1022.0        2.0       5.36
5      138.0       -7.0       -5.0     1022.0        2.0       6.25

   var7(t-1)  var8(t-1)  var1(t)
1        0.0        0.0    148.0
2        0.0        0.0    159.0
3        0.0        0.0    181.0
4        1.0        0.0    138.0
5        2.0        0.0    109.0
```

Figure 5.1: Pre-processed Dataset: Screenshot from our jupyter notebook (LSTM specific)

5.2. Calculation and Preparation of AQI Data Frame

We calculated the AQI values using the formula given by EPA government on the preprocessed data. These AQI values were stored in a separate data frame.

$$I = ((I_{high} – I_{low})/(C_{high} – C_{low})) *(C – C_{low}) + I_{low}$$

where:
I = the Air quality Index
C =pollutant concentration
C_{low} = the concentration breakpoint that is ≤ C
C_{high} = the concentration breakpoint that is ≥ C
I_{low} = the index breakpoint corresponding to C_{low}
I_{high} = the index breakpoint corresponding to C_{high}

Formula for AQI calculation

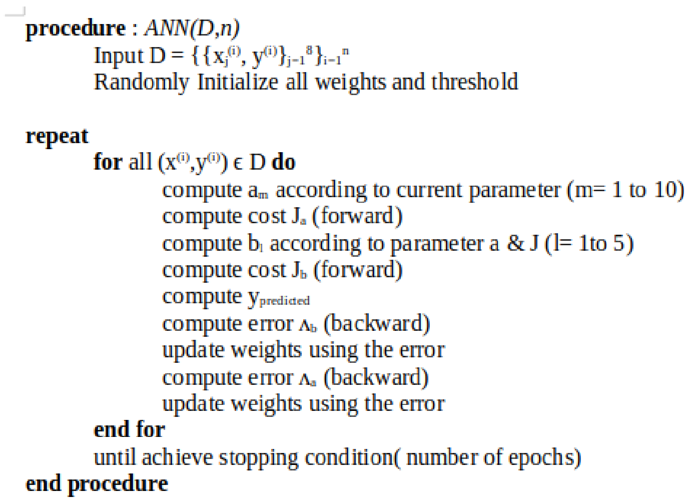
5.3 Training Models:

We first applied Machine Learning to train our dataset. We applied SVR or Support Vector Machine with radial basis kernel.

We compared the results obtained by applying Deep Learning.We first applied LSTM having 128 nodes in the input layer and a single hidden layer with 50 nodes. We trained the model for 25 epochs and the training took nearly 48 seconds.

We then applied ANN with two hidden layers having 10 and 5 units respectively. We evaluated our results using RMSE(Root Mean Squared Error) and R2 values. We trained the model for 50 epochs and the training took nearly 32 seconds.

5.3 Pseudocode



```
procedure : ANN(D,n)
        Input D = $\{\{x_j^{(i)}, y^{(i)}\}_{j-1}^{8}\}_{i-1}^{n}$
        Randomly Initialize all weights and threshold

repeat
        for all $(x^{(i)},y^{(i)}) \epsilon D$ do
                compute $a_m$ according to current parameter (m= 1 to 10)
                compute cost $J_a$ (forward)
                compute $b_l$ according to parameter a & J (l= 1to 5)
                compute cost $J_b$ (forward)
                compute $y_{predicted}$
                compute error $\Lambda_b$ (backward)
                update weights using the error
                compute error $\Lambda_a$ (backward)
                update weights using the error
        end for
        until achieve stopping condition( number of epochs)
end procedure
```

Figure 5.2: Pseudo-Code of ANN: Screenshot from our word processor.

Chapter 6: EXPERIMENTAL RESULTS AND ANALYSIS

After training and testing models of Machine Learning and Deep Learning, we came up with a new model of ANN which performed better that LSTM RNN neural network.

	Variance (R2 value)	Root Mean Squared Error
SVR	0.79	42.14
LSTM	0.92	24.86
ANN	0.91	26.21

For the above table it's evident that ANN we made using 2 hidden layers of size 10 and 5 respectively using sigmoid activation function was able to provide better results than a more complex model LSTM. The reason we're able to obtain good results with relatively simpler model is due to direct relationship between the meteorological data provided which already depends upon the previous meteorological data so we didn't need to use a memory unit as in LSTM to predict pollution level. So just the weather data was able to give a good enough prediction of the next day pollution levels.

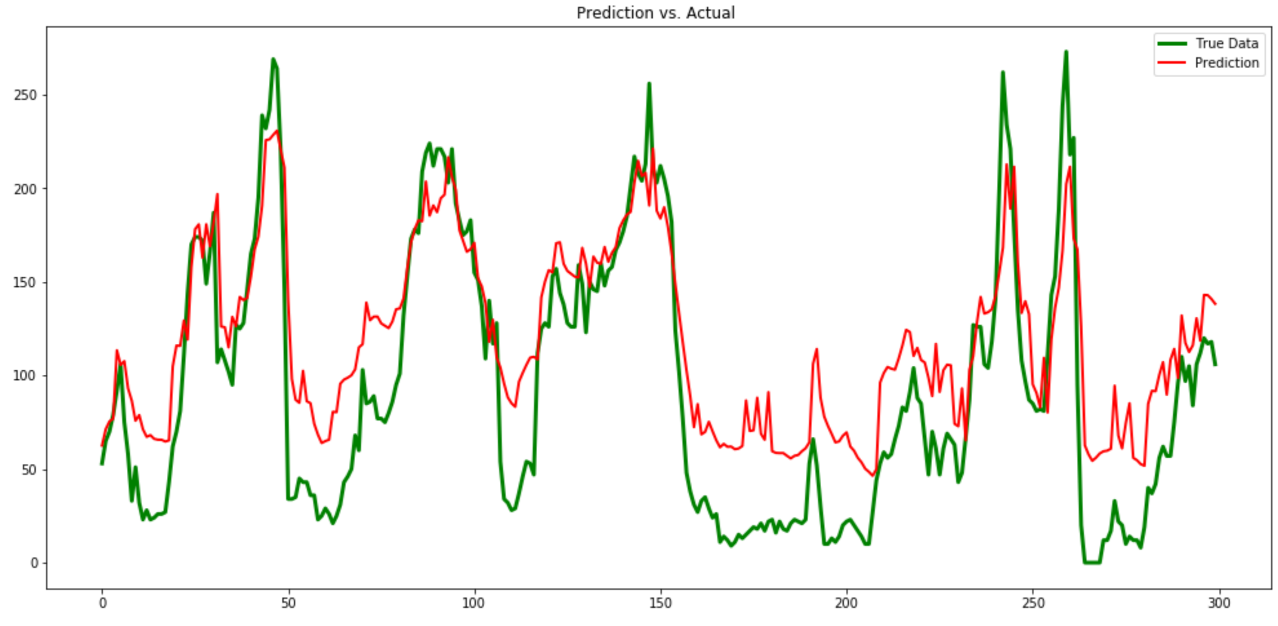


Figure 6.1: Results for SVR: Screenshot from spyder shell.

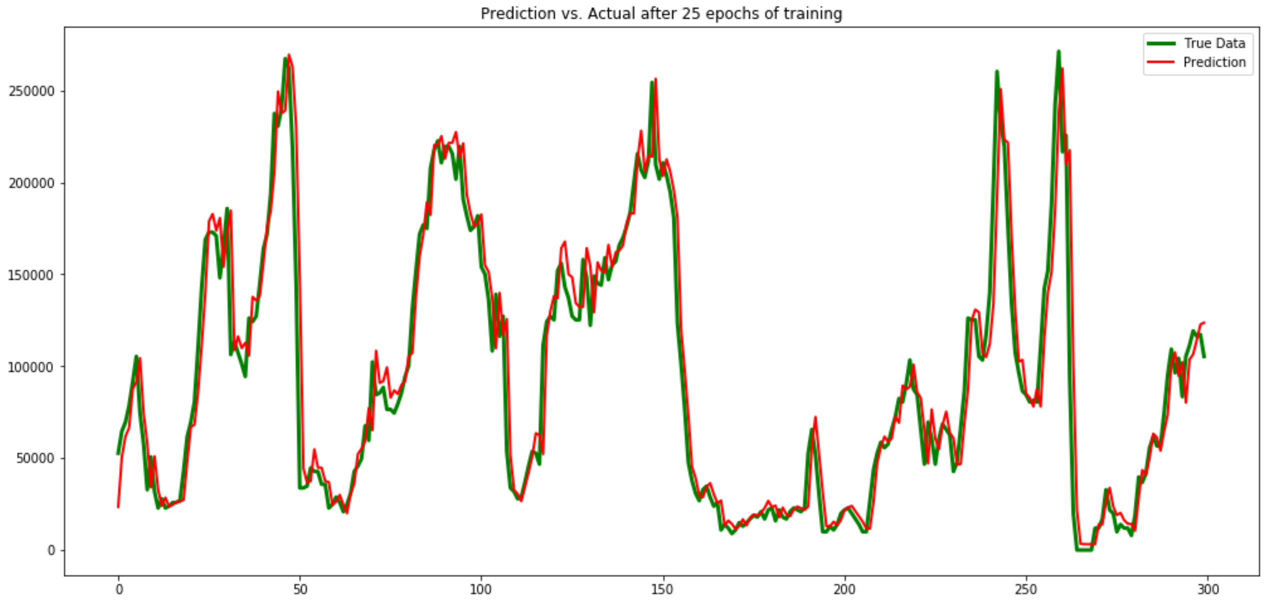


Figure 6.2: Results for LSTM: Screenshot from our spyder shell.

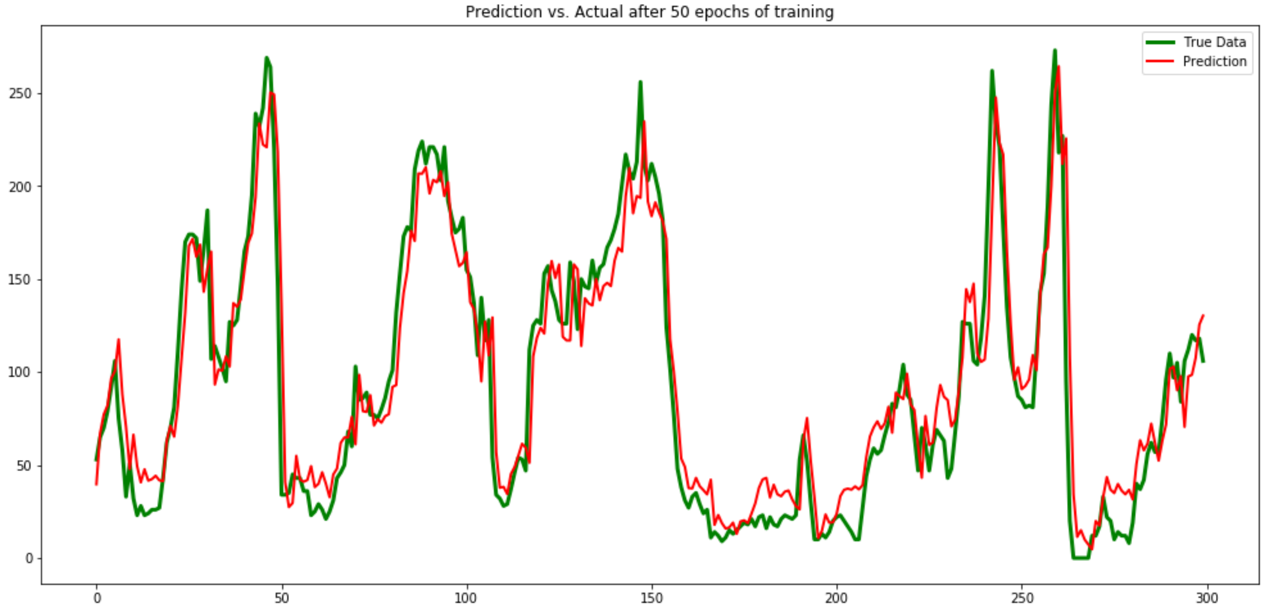


Figure 6.3: Results for ANN: Screenshot from our spyder shell.

Chapter 7: CONCLUSION AND FUTURE SCOPE

This project performs a comparative study of machine learning and deep learning models for the prediction of AQI values. We used the Air Quality data of China collected by the Air Pollution Prediction System. Three machine learning algorithms: SVM, KNN and Multiple linear regression were used. The models were evaluated using metrics like precision, recall, support, f1-score and accuracy. The best results were obtained by using KNN classifier which provided us with an accuracy of 82%.

For Deep Learning Models we used the LSTM and ANN network. Using LSTM we obtained good results where the r2 value was 0.92 but that took a long time to train. We then applied ANN with two hidden layers having 10 and 5 units respectively. We evaluated our results using RMSE(Root Mean Squared Error) and R2 values. We trained the model for 50 epochs and the training took nearly 32 seconds. The ANN model which ran faster, nearly completed training in 2/3 of the time it took for LSTM to complete with a r2 score of 0.91.

AQI can be used as a precautionary measure to control the pollution levels in a city. Thus, predicting it beforehand is beneficial. This would help in reducing the environmental hazards like global warming and many health issues to which people are prone due to bad air quality.

With increased factors of pollutants and multiple unpredictable sources of pollution, we plan to use multiple models to predict pollution levels of multiple pollutants. We also plan to find spatial relationship between pollutants and hence give better predictions.

REFERENCES

Research Paper

[1] Wenjian Wang ; A.Y.T. Leung ; Siu-Ming Lo ; R.K.K. Yuen ; Zongben Xu ;Huiyuan Fan, *Air pollutant parameter forecasting using support vector machines*,2002 International Joint Conference On Neural Networks (IJCNN'02), Volume 1, Year 2002

[2] Kostandina Veljanovska, Angel Dimoski , *Air Quality Index Prediction using Simple Machine Learning Algorithms*, Internation Journal Of Emerging Trends & Technology in Computer Science(IJETTCS), Volume 7 ,Issue 1,Jan-Feb 2018.

[3] İbrahim KÖK,Mehmet Ulvi ŞİMŞEK, Suat ÖZDEMİR, *A deep learning model for air quality prediction in smart cities*, IEEE International Conference on Big Data,2017.

[4] Dan Wei ,Predicting air pollution level in a specific city.

[5] Yi, X., Zhang, J., Wang, Z., Li, T., & Zheng, Y. . Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 965-973). ACM, July 2018