D7041E _Mini Project_Group8

Anusha Talapaneni
Dipanwita Dash
Meenakshi Subhash Chippa

Project : Multivariate Time Series Forecasting with LSTMs
(Air Pollution Forecasting)

LULEÅ
TEKNISKA
UNIVERSITET

Agenda

- ➔ Project Introduction
- ➔ Data analysis & pre-processing
- ➔ Data Visualization
- ➔ Experiments
- ➔ Results

LULEÅ TEKNISKA UNIVERSITET

Project Introduction

Aim of our project is to **“forecast air pollution**” using **LSTM** which is type of RNN.

Air Quality dataset is used which contains features like pollution, dew point, temperature, pressure, wind direction, speed, date-time , snow and rain.

Our **p**roject covers various experiments which involves various steps in general:

- To transform a raw dataset into something that we can use for time series forecasting.
- To prepare data and fit an LSTM for a multivariate time series forecasting problem.
- To make a forecast on test data and rescale the result back into the original units.

LULEÅ TEKNISKA UNIVERSITET

1st Experiment

Data Preprocessing:

- Parsing Dates (datetime column)
- Dropping Unnecessary Column
- Renaming Columns: e.g., 'pollution', 'dew', 'temp', 'press', 'wnd_dir', 'wnd_spd', 'snow', 'rain
- Handling Missing Values
- Discarding Initial Rows: The first 24 hours of data are dropped, as they may not be relevant for analysis.
- Label Encoding: The 'wnd_dir' (wind direction) column, which is categorical, is encoded into numeric format using LabelEncoder.
- Data Normalization: The features are normalized using MinMaxScaler to scale them to a range between 0 and 1.

LULEÅ TEKNISKA UNIVERSITET

Transforming Data for Supervised Learning:

The supervised learning problem will be framed as predicting pollution at the current hour (t) based on previous pollution measurements and weather conditions..

Reshaping for LSTM Model:

The input data is reshaped into a 3D format required by the LSTM model, which is [samples, timesteps, features].

Model Building and Training:

LSTM model with 50 neurons & Adam SGD optimiser & Epochs = 50

Making Predictions and Inverting Transformations:

The model is used to make predictions on the test set. These predictions, along with the test set, are then transformed back to their original scale.

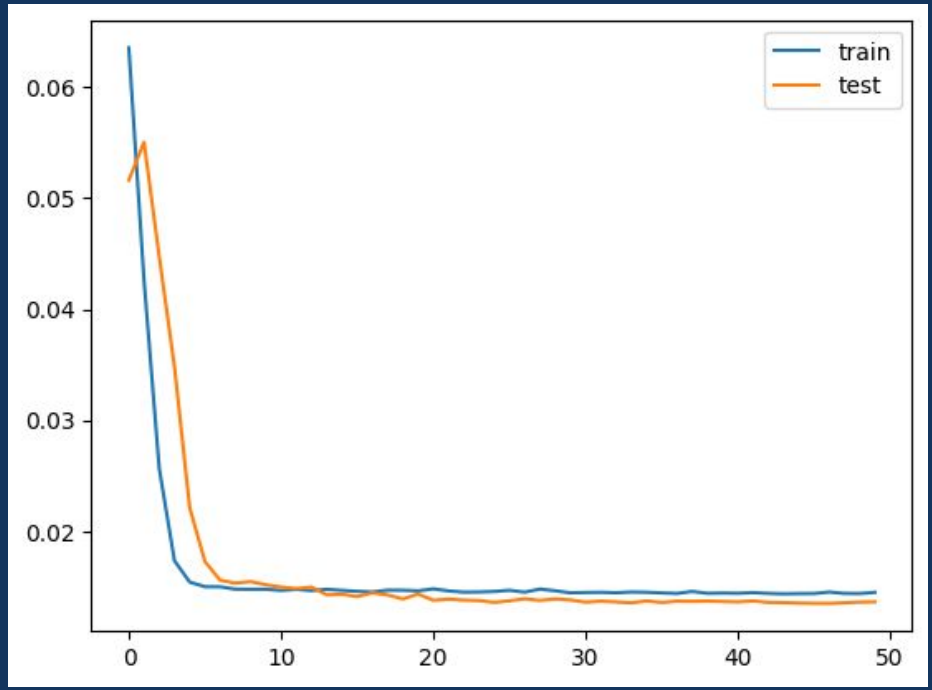
https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/

LULEÅ TEKNISKA UNIVERSITET

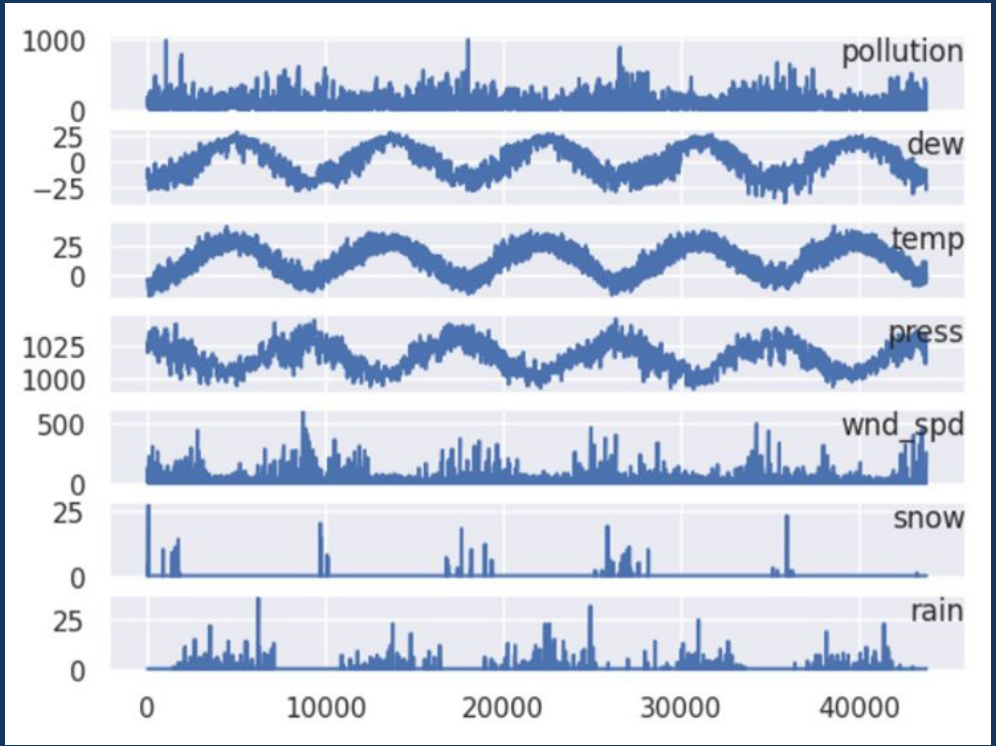
Result :

RMSE which is the difference between the predicted values and the actual values in the test dataset.

Test RMSE of 26.496

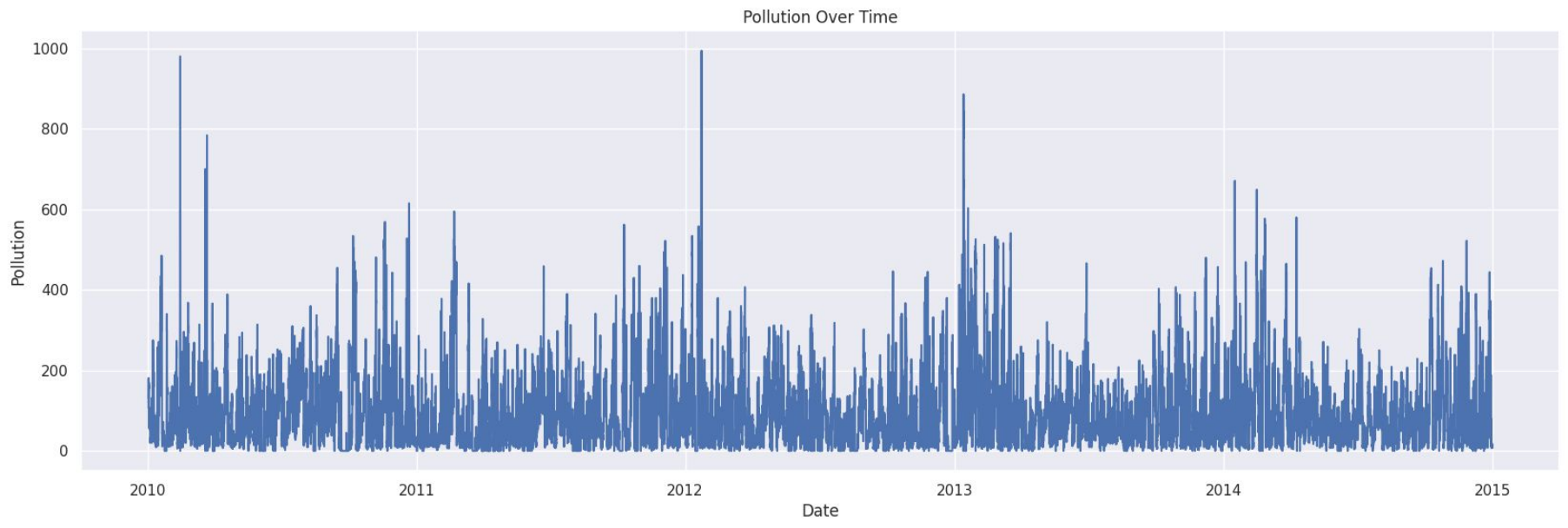


Data Visualization

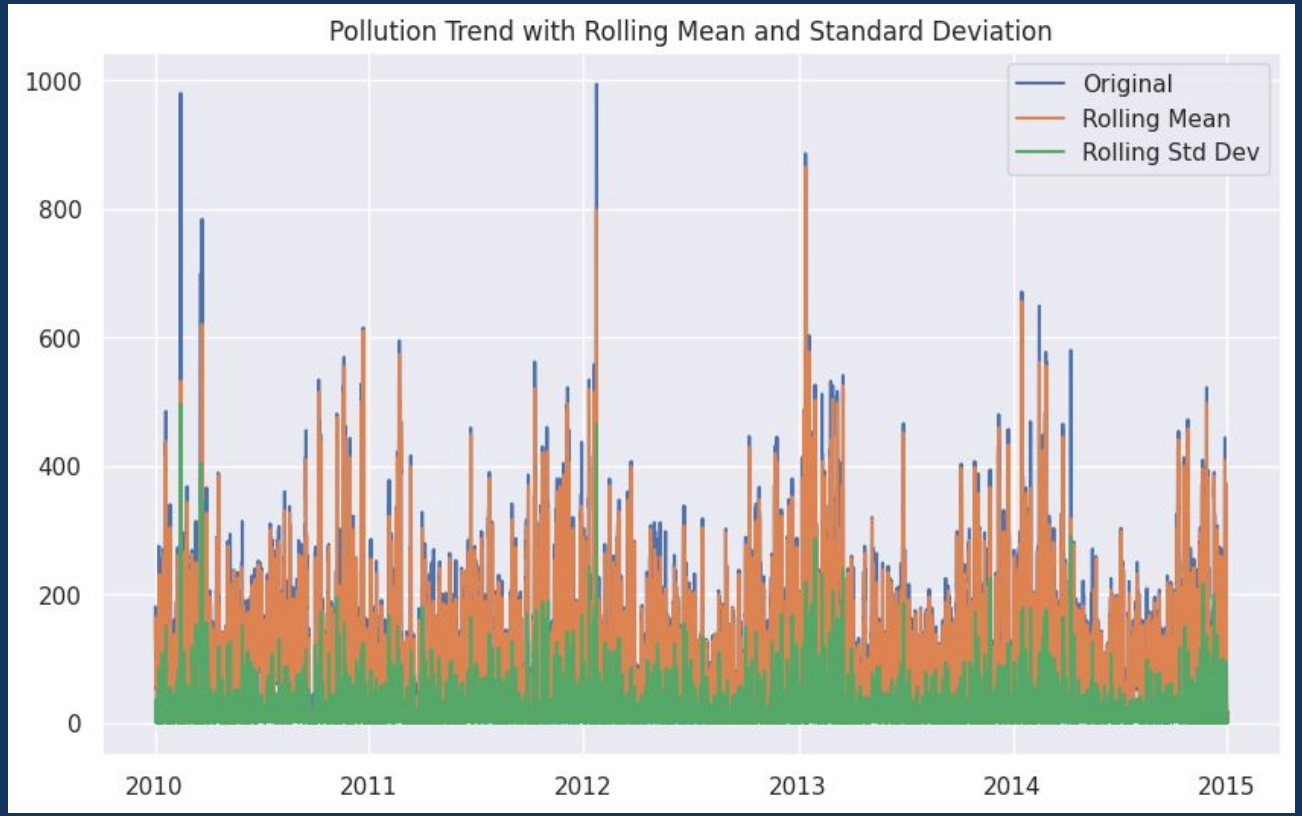


Line Plots of Air Pollution Time Series

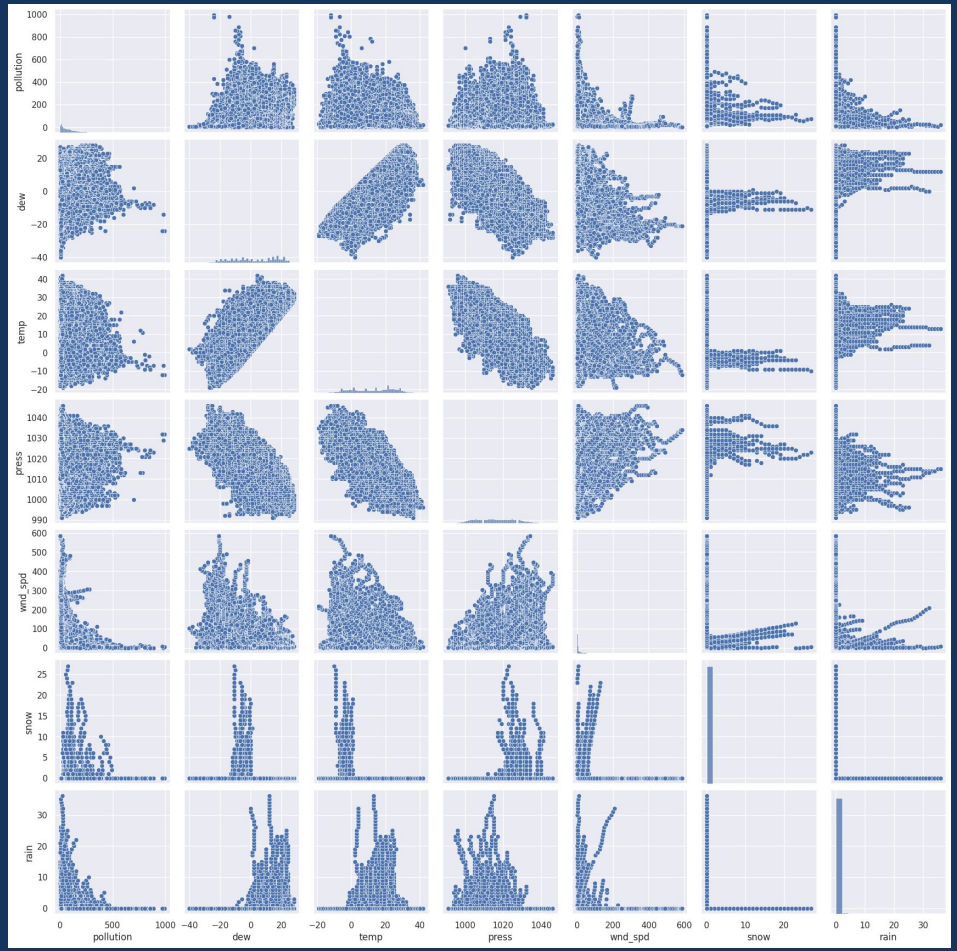
A **time series plot** for pollution over time with a title and labeled axes, providing a clear visual of the target variable trend.



A **rolling mean and standard deviation plot** is added to observe the moving average trend and volatility over time.

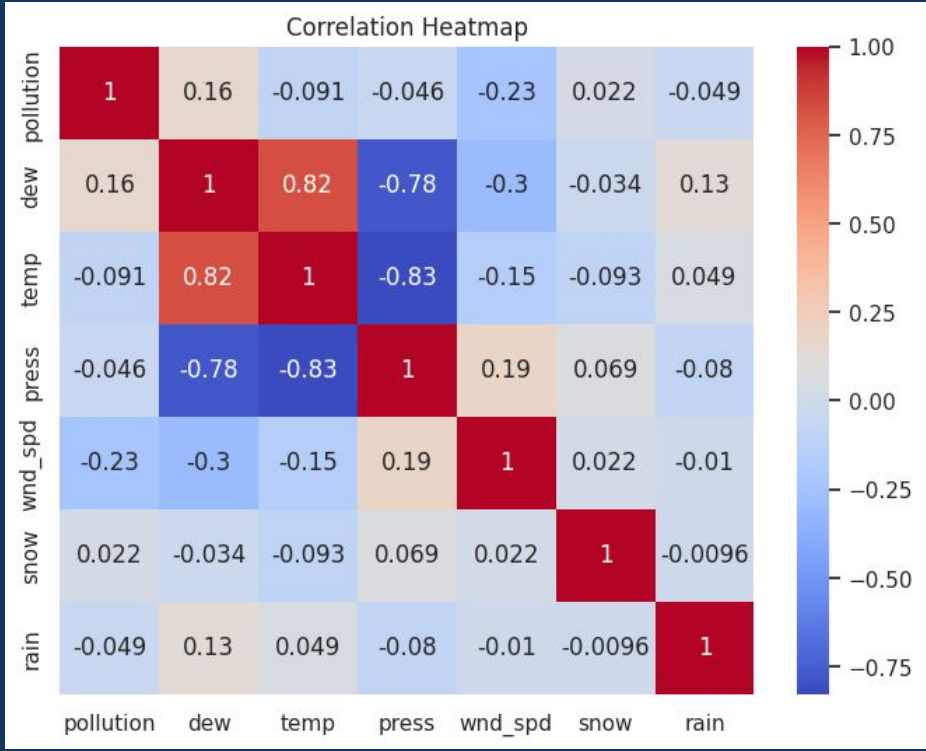


Histograms and KDE plots aid in understanding data distribution, skewedness, and outliers, while providing a smooth curve for density distribution.



A **Seaborn pairplot** is used to visualize pairwise relationships between variables, which is useful for understanding the interactions between different features.

LULEÅ TEKNISKA UNIVERSITET



A **heatmap** for correlation analysis between numerical features is added, providing a visual representation of how closely related the different variables are to each other.

2nd Experiment

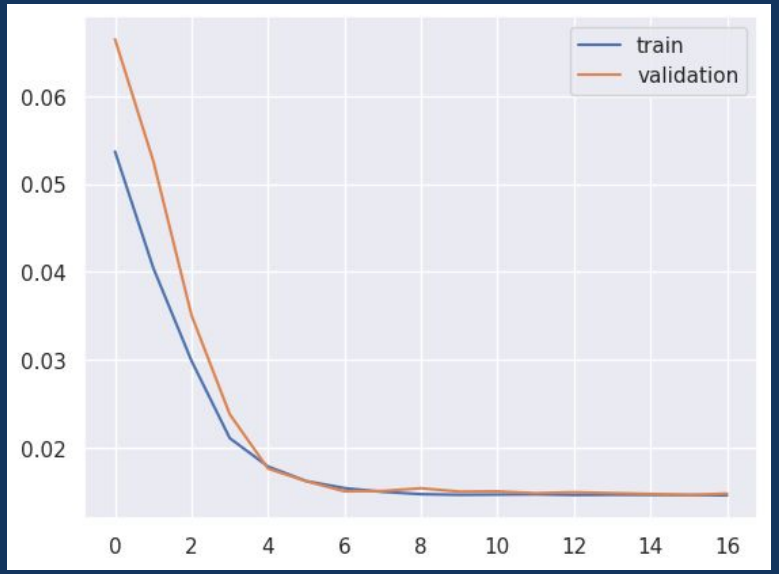
Model Building and Training:

LSTM model with 50 neurons & Adam SGD optimiser & Epochs = 50

Early stopping

Batch_Size : 72

Data splitted : Train,val,test



LULEÅ TEKNISKA UNIVERSITET

3rd Experiment

Hyperparameter Tuning:

Experimented with different numbers of LSTM units and optimizers also used different random seeds for reproducibility to ensure the robustness of the model performance

```
lstm_units = [30, 50, 100] # Neurons
optimizers = ['adam', 'rmsprop']
seeds = [42, 7, 21]
```

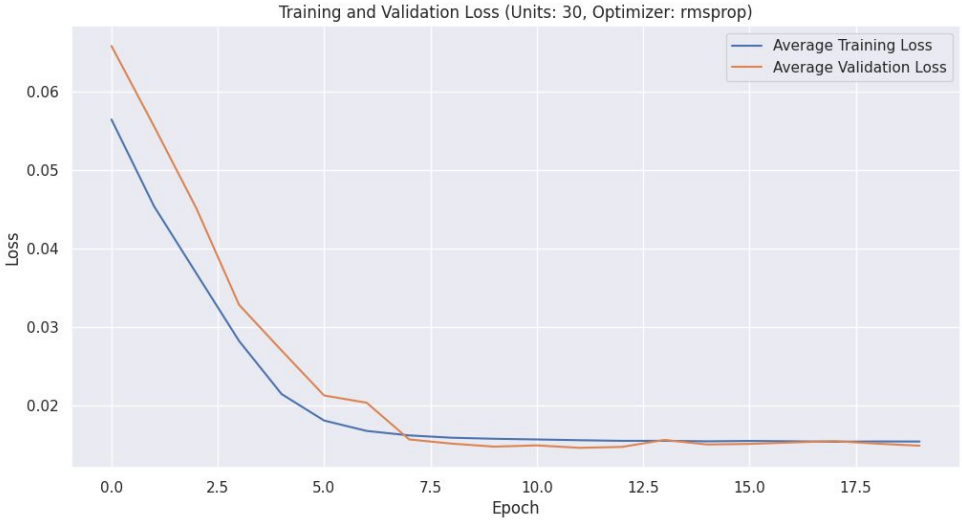
Enhanced Model Evaluation:

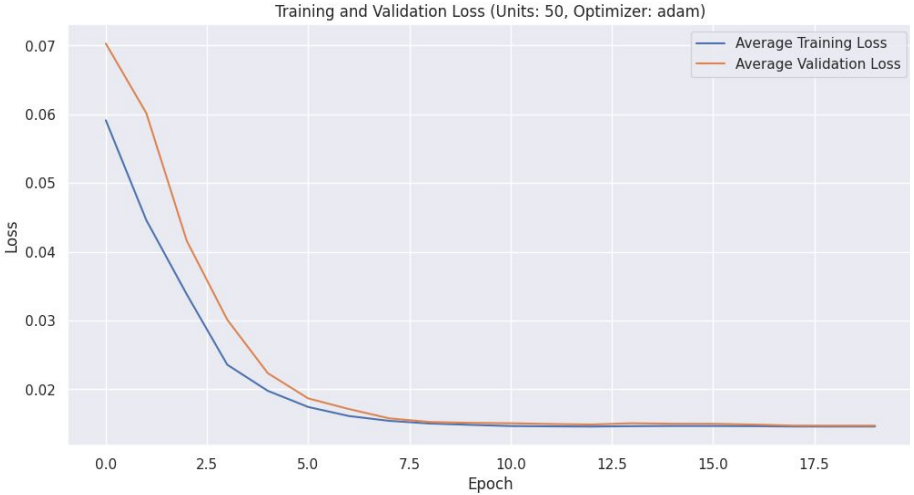
In addition to calculating the RMSE, the code also computes the Mean Absolute Error (MAE) and the R^2 score, providing a more comprehensive evaluation of model performance.

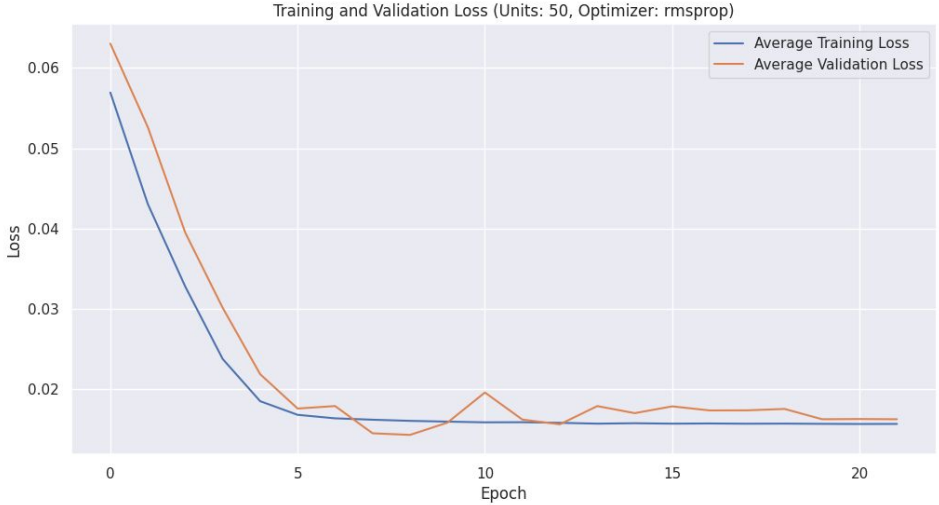
LULEÅ TEKNISKA UNIVERSITET


Training and Validation Loss (Units: 30, Optimizer: adam)
Average Training Loss
Average Validation Loss
Loss
0.07
0.06
0.05
0.04
0.03
0.02
0
5
10
15
20
25
30
Epoch

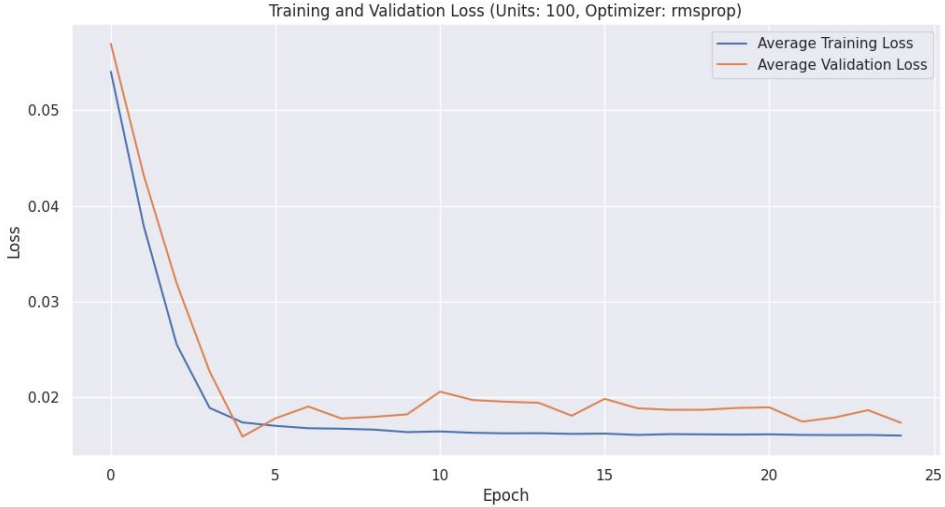
Results:


Training and Validation Loss (Units: 30, Optimizer: rmsprop)
Average Training Loss
Average Validation Loss
Loss
0.06
0.05
0.04
0.03
0.02
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epoch


Training and Validation Loss (Units: 50, Optimizer: adam)
Average Training Loss
Average Validation Loss
Loss
Epoch


Training and Validation Loss (Units: 50, Optimizer: rmsprop)
Average Training Loss
Average Validation Loss
Loss
Epoch

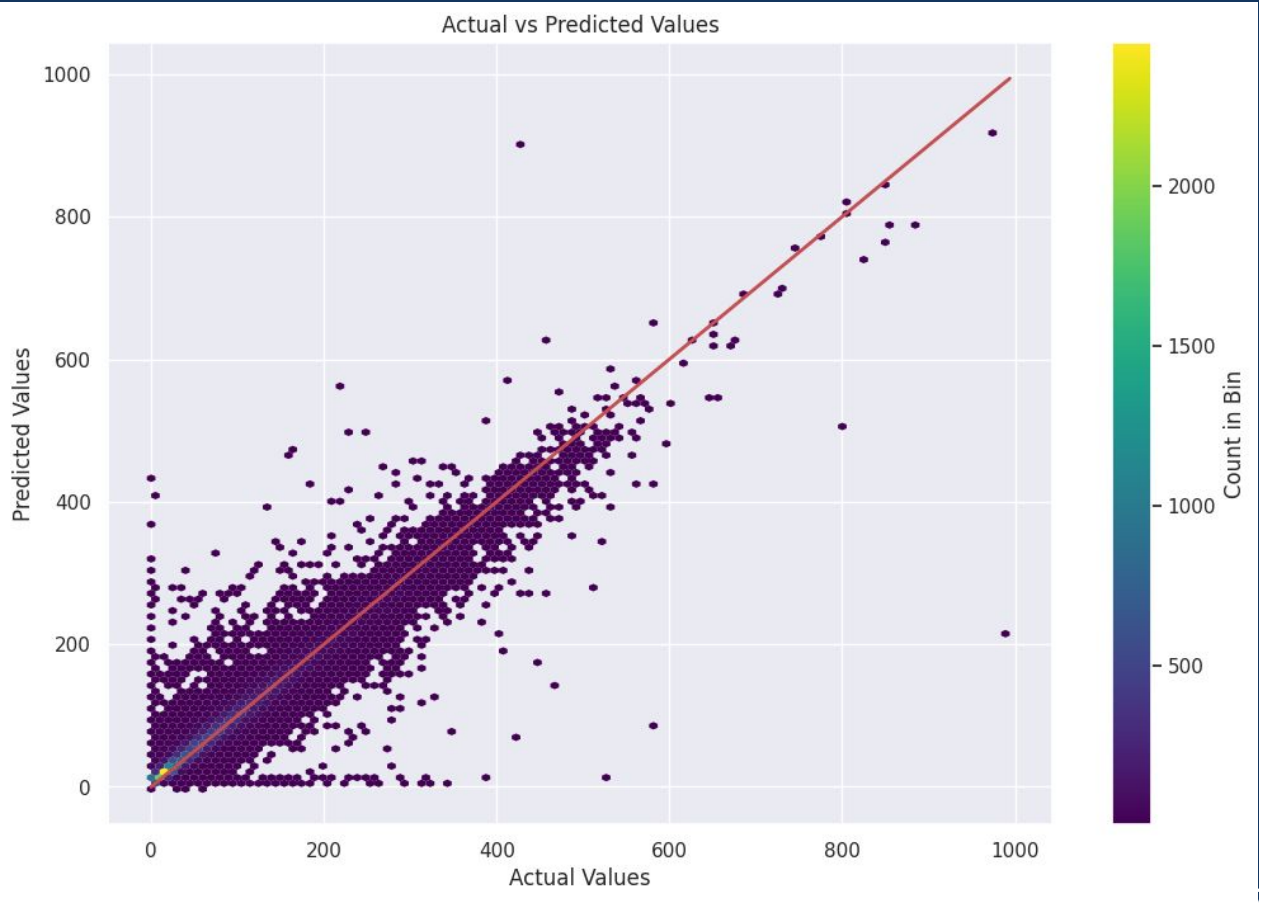

Training and Validation Loss (Units: 100, Optimizer: adam)
Average Training Loss
Average Validation Loss
Loss
Epoch


Training and Validation Loss (Units: 100, Optimizer: rmsprop)
Average Training Loss
Average Validation Loss
Loss
Epoch

Test RMSE: 26.371

Mean Absolute Error: 14.037

R^2 Score: 0.918



4th Experiment

Model Parameters:

Data splitted : Train,Validation,Test

Random Forest:

n_estimators=100

loss function: Mean Squared Error (MSE)

Validation RMSE for Random Forest: 73.14355054783394

Support Vector Regressor:

kernel = rbf

degree=3 (polynomial kernel function)

C=1.0

epsilon=0.1

Validation RMSE for SVM: 82.43127051948618

LULEÅ TEKNISKA UNIVERSITET

Results

Experiment 1: Test RMSE of 26.496

Experiment 2: Test RMSE: 26.185

Experiment 3: Test RMSE: 26.371
Mean Absolute Error: 14.037
R^2 Score: 0.918

Experiment 4: SVM - RMSE: 83.06
Random Forest - RMSE : 74.71

LULEÅ TEKNISKA UNIVERSITET

Conclusion

- Thorough exploration of multivariate time series forecasting
- Detailed analysis of LSTM networks and comparison with traditional models
- Results highlighting LSTM's capability with a strong R^2 score of 0.918
- RMSE values indicating the challenging nature of environmental forecasting
- Insightful comparative analysis against Random Forest and Support Vector Regressor models
- Potential practical applications in environmental monitoring and policy making.

Github link : https://github.com/dipanwitadash/Applied-AI-miniproject-group8

LULEÅ TEKNISKA UNIVERSITET


LULEÅ
TEKNISKA
UNIVERSITET