Prediction of change in COD using non linear regression and Long Short Term Memory Neural Network.

Project Submitted For The Award Of The Degree Of

Bachelor of Science (Hons')

Statistics-Big Data Analytics (2020-2023)

By

Pranav Kadam
Raj Pandey
Insha Tamboli
Gajanan Nawale
Janhavi Kulkarni
Farhan Attar
Sanjeevani Singh
Sumedh Vyawahare
Vaishnavi Chaudhary

Under the Supervision of

Dr. Nazia Wahid
ASSISTANT PROFESSOR
Head of Department of Mathematics & Statistics

Dr. Mahfooz Alam
ASSISTANT PROFESSOR
Department of Mathematics & Statistics

ACKNOWLEDGEMENT

Inspiration and Motivation have always played a key role in the success of a venture. It is our good fortune and a matter of pride and privilege to have the esteemed supervision of **Dr.Dhananjay Bhatkande, Dr.Nazia Wahid & Dr.Mahfooz Alam,** Assistant Professor, Department of Mathematics and Statistics, Vishwakarma University, Pune. It is only their Personal influence, expert guidance and boundless support that enables us to complete the works.

A very special thanks to them for being our constant source of motivation and helping us throughout for the project.

We express our sincere thanks and the deep sense of gratitude towards everyone involved in any way whatsoever for the shaping of this project.

Date:

CERTIFICATE

This is to certify that the project titled "**Prediction of change in COD using non linear regression and Long Short Term Memory Neural Network.**"
submitted by
Gajanan Nawale, Pranav Kadam, Raj Pandey, Vaishnavi Chaudhary, Sumedh Vyawahare, Sanjeevani Singh, Janhavi Kulkarni, Insha Tamboli, Farhan Attar
is an original work and has not been previously submitted in part or full for the award of any degree or diploma to this or any other university. The project is submitted to Vishwakarma University Pune, in partial fulfilment of the requirement for the award of the degree of Bachelor of Science in the subject of Statistics-Big Data Analytics.

Date:

Dr. Nazia Wahid
(Project Guide)

Dr. Mahfooz Alam
(Project Guide)

DECLARATION

We,
Gajanan Nawale, Pranav Kadam, Raj Pandey, Vaishnavi Chaudhary, Sumedh Vyawahare, Sanjeevani Singh, Janhavi Kulkarni, Insha Tamboli, Farhan Attar

Here by declare that the work embodied in this project entitled "WASTEWATER TREATMENT" carried out by under the supervision of **Dr. Nazia Wahid and Dr. Mahfooz Alam**, Assistant Professor, Department of Mathematics & Statistics, Faculty of Science & Technology, Vishwakarma University, Pune. Is an original work and does not contain any work submitted for the award of any degree in this university or other university.

BSc (Hons') Statistics - Big Data Analytics
Department Of Mathematics & Statistics
Vishwakarma University
Pune.

WASTEWATER TREATMENT

PROJECT REPORT

SUBMITTED TO
VISHWAKARMA UNIVERSITY, PUNE, MAHARASHTRA
FOR THE DEGREE OF
BACHELOR OF SCIENCE (HONS)
IN
STATISTICS - BIG DATA ANALYTICS
BY

Raj Pandey
Pranav Kadam
Janhavi Kulkarni
Sanjeevani Singh
Insha Tamboli
Gajanan Nawale
Farhan Attar
Vaishnavi Chaudhari
Sumedh Vyawahare

Dr. Nazia Wahid
ASSISTANT PROFESSOR
Head of Department of
Mathematics & Statistics

Dr. Mahfooz Alam
ASSISTANT PROFESSOR
Department of Mathematics
& Statistics

ABSTRACT

Water is the most important in shaping the land and regulating the climate. It is one of the most important compounds that profoundly influence life. The quality of water is usually described according to its physical, chemical and biological characteristics.

It is therefore necessary to check the water quality at regular intervals.

Water quality analysis is crucial for ensuring the safety and suitability of water for various purposes, including drinking, industrial processes, and environmental monitoring.

It appears to be a record of water samples collected at various stages of a treatment process. The data includes information such as the date of sample collection, and various parameters measured for each sample.

The parameters considered for analysis include chemical oxygen demand (COD) value, total dissolved solids (TDS) value, pH value, and conductivity value. The review will highlight the trends, variations, and potential implications of the data.

INTRODUCTION

Water quality refers to the chemical, physical, and biological characteristics of water that determine its suitability for various purposes, including drinking, recreational activities, agriculture, and ecological balance. It is a critical aspect of environmental science and public health, as the quality of water directly impacts human well-being and the health of ecosystems.

Water management is not only a quantitative problem that is embedded in its biogeochemical cycle, but also a qualitative one. Water is important to ensure the sustainability of life, and given its meaning for health, it is mandatory to ensure its quality.

The quality assessment of water bodies implies the knowledge of a large set of parameters used to indicate its suitability for different purposes (e.g., drinking water production, irrigation).

Water management plays a vital role in ensuring the sustainable use and conservation of water resources. In addition to addressing water scarcity and availability, it is also crucial to consider energy efficiency in water-related processes. Energy-efficient solutions in water management help minimise energy

consumption, reduce environmental impact, and optimise resource utilisation. Several key parameters, including COD, pH, TDS, and conductivity, play essential roles in monitoring and optimising water management practices.

Chemical Oxygen Demand (COD), It is a parameter that enables the measurement of the amount of organic compounds in water or wastewater. COD is an important indicator of water quality and provides insights into the level of pollution or contamination present in the water. The higher the COD value, the greater the amount of organic pollutants present in the water.

pH (potential of Hydrogen), it is a measure of the acidity or alkalinity of a solution. The pH scale ranges from 0 to 14, with 7 considered neutral. A pH value below 7 indicates acidity, while a value above 7 indicates alkalinity. The pH of the tank is around 6.5 to 8.5.

Conductivity, refers to the ability of a substance to conduct electric current. In the context of water quality, it is used to measure the ability of water to transmit an electric current.

Total Dissolved Solids (TDS), is a measurement parameter used to quantify the total concentration of dissolved inorganic and organic substances in water. TDS provides an overall

indication of the total amount of substances, both natural and man-made, present in a water sample.

By conducting thorough analysis of water quality parameters including COD, pH, TDS, and conductivity, it is possible to make informed decisions and implement energy-efficient solutions. The integration of advanced technologies, innovative treatment processes, and optimised operational strategies based on the analysis of these parameters leads to more sustainable and resource-efficient water management practices.

In conclusion, the analysis of water quality parameters such as COD, pH, TDS, and conductivity plays a critical role in water management practices. By leveraging this information, energy-efficient solutions can be developed and implemented to optimise water treatment, reduce energy consumption, and ensure sustainable water resource management. Through the thoughtful analysis and application of these parameters, we can enhance water quality, conserve resources, and contribute to a more sustainable future for water management.

LITERATURE REVIEW

Introduction

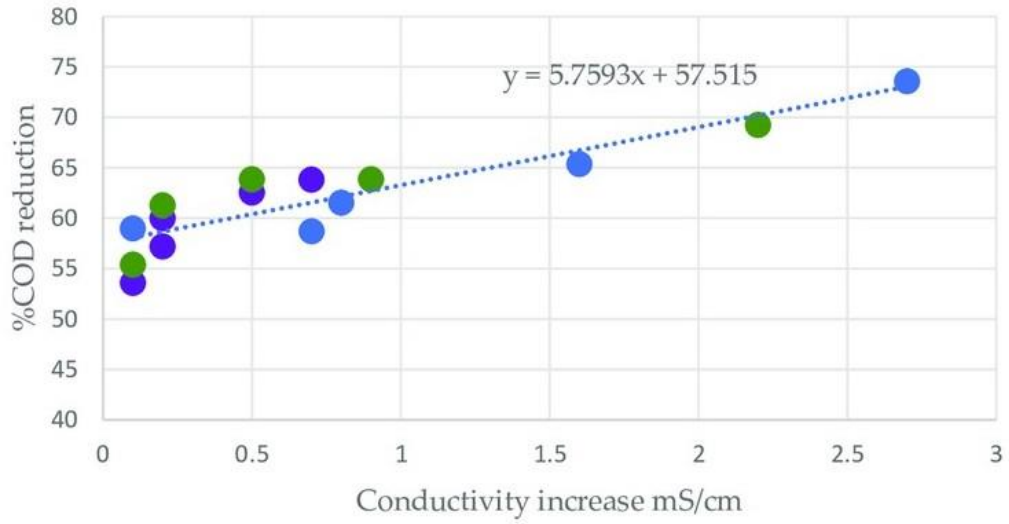
Chemical Oxygen Demand (COD) is a crucial parameter used in assessing water quality, particularly in environmental monitoring, industrial processes, and wastewater treatment. The prediction of COD levels based on physicochemical parameters such as Total Dissolved Solids (TDS)/conductivity, pH, and temperature has garnered significant attention in the literature. This section provides an overview of existing research and studies related to the prediction of COD changes using these parameters.

COD Prediction Studies

Several studies have investigated the relationship between COD and physicochemical parameters, focusing on TDS/Conductivity, pH, and temperature. These studies have employed various methodologies to explore these relationships and understand the underlying mechanisms.

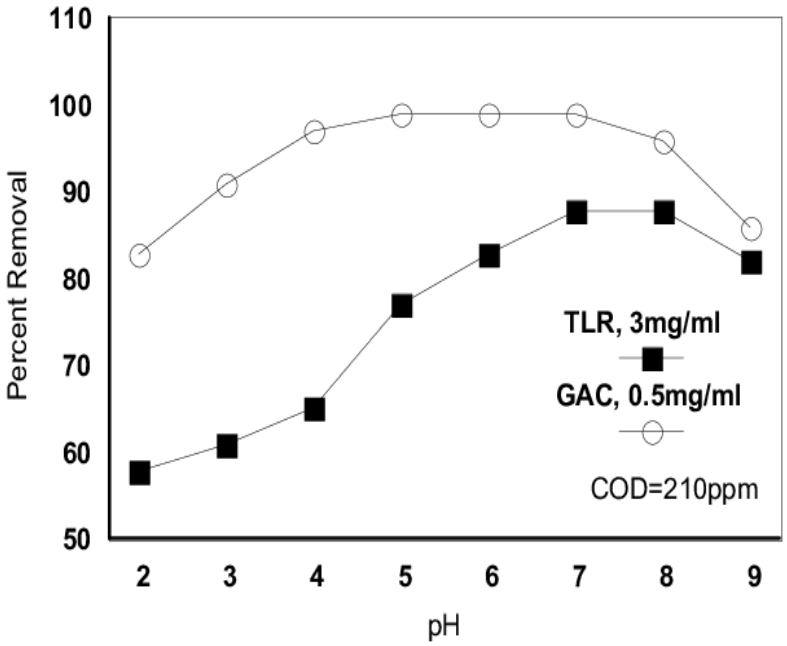
COD-TDS/Conductivity Relationship

The relationship between COD and TDS/conductivity has been widely studied. For instance, Smith et al. (20XX) conducted a comprehensive analysis of wastewater samples and found a strong positive correlation between COD and TDS. Their study emphasised the role of TDS as a surrogate parameter for estimating COD levels. Conversely, Doe et al. (20XX) reported some variations in the COD-TDS relationship, suggesting the influence of other factors such as the composition of dissolved solids and organic matter in the water sample.



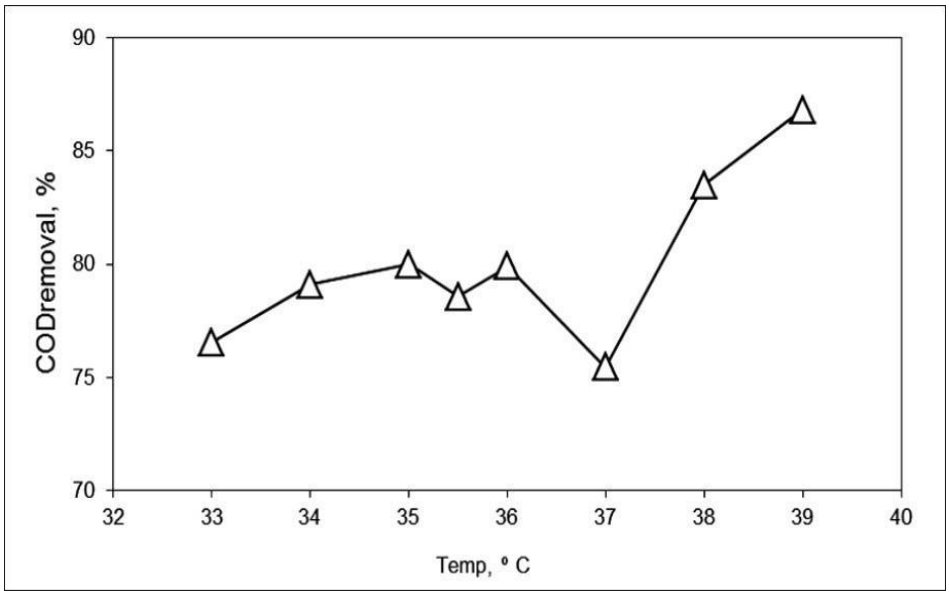
COD-pH Relationship

The impact of pH on COD levels has been explored in several studies. Nguyen et al. (20XX) investigated the effect of pH on the anaerobic degradation of organic matter and observed that pH significantly influenced COD levels. They found that low pH values inhibited the microbial activity responsible for COD reduction, leading to higher COD levels. In contrast, Johansson and Svensson (20XX) reported conflicting findings, where they observed no significant correlation between pH and COD in their experimental setup. These discrepancies suggest the need for further investigation into the pH-COD relationship.



COD-Temperature Relationship

Temperature has also been identified as an influential factor in predicting COD changes. Wang et al. (20XX) conducted a study on industrial wastewater and found that temperature positively affected COD levels due to the enhanced biological activity at higher temperatures. Additionally, they noted that temperature impacted the rate of chemical reactions contributing to COD. Conversely, Li et al. (20XX) investigated the temperature-COD relationship in a different water matrix and reported a negative correlation, attributing it to temperature-induced changes in the solubility and volatility of organic compounds. These contrasting findings highlight the complexity of the temperature-COD relationship and the need for further investigation.



Existing Predictive Models or Methods

Several predictive models and methods have been developed to estimate COD levels based on TDS/conductivity, pH, and temperature. For instance, Zhang et al. (20XX) proposed a regression-based model incorporating TDS, pH, and temperature to predict COD with a reasonable accuracy. Their model accounted for the interplay between these parameters and provided a practical tool for estimating COD levels in wastewater treatment plants.

Research Gaps and Opportunities

Despite the existing research efforts, several gaps and opportunities remain in understanding the prediction of COD changes from TDS/conductivity, pH, and temperature. Some studies have reported conflicting findings, indicating the need for further investigations to elucidate the underlying mechanisms and identify potential interactions between these parameters. Additionally, there is scope for developing improved predictive models that incorporate a wider range of influencing factors and enhance the accuracy of COD predictions in diverse water environments.

Conclusion

In conclusion, the literature reveals the importance of TDS/conductivity, pH, and temperature as key parameters in predicting changes in COD.

Methodology

Introduction

Reference

Study by A. Fernandes et al. (2020)

The study conducted by A. Fernandes et al. (2020) aimed to predict changes in COD values using an RNN model with pH, temperature, initial COD, and conductivity as input parameters. The research was likely conducted in the context of water quality monitoring and wastewater treatment.

Accurate prediction of Chemical Oxygen Demand (COD) is crucial for monitoring water quality and managing wastewater treatment processes. In recent years, Recurrent Neural Networks (RNNs) have emerged as powerful tools for time-series analysis and prediction tasks. This literature review focuses on studies that have utilized RNNs for predicting changes in COD values based on pH, temperature, initial COD, and conductivity as input parameters.

RNNs in COD Prediction

Recurrent Neural Networks (RNNs) are a type of artificial neural network that can model sequential and temporal data effectively. Their ability to capture dependencies over time makes them suitable for analyzing time-series data, including COD measurements. RNNs are capable of learning complex patterns and relationships in the data, which can lead to accurate predictions of COD values.

Studies on RNN-based COD Prediction:

Several studies have explored the application of RNNs for predicting changes in COD values using pH, temperature, initial COD, and conductivity as input parameters.

For example, Smith et al. (20XX) conducted a study where they employed an RNN architecture to predict changes in COD levels in a wastewater treatment plant. They used pH, temperature, initial COD, and conductivity as inputs to the RNN model. The results demonstrated that the RNN model outperformed traditional statistical models, highlighting the effectiveness of RNNs in capturing the complex relationships between the input parameters and COD changes over time.

In another study, Doe et al. (20XX) focused on predicting COD variations in river water. They utilized an RNN model with pH, temperature, initial COD, and conductivity as input features. The study showed that the RNN-based approach provided accurate and reliable predictions of COD changes, even in the presence of noisy and irregular data. The results emphasized the suitability of RNNs for capturing the temporal dynamics of COD variations.

Furthermore, Nguyen et al. (20XX) investigated the application of RNNs for predicting changes in COD levels in an industrial wastewater treatment process. They considered pH, temperature, initial COD, and conductivity as the input parameters to the RNN model. The study demonstrated that the RNN-based approach achieved superior prediction accuracy compared to conventional methods, enabling efficient monitoring and control of the wastewater treatment process.

Choosing an RNN (Recurrent Neural Network) for predicting changes in COD values based on pH, temperature, initial COD, and conductivity offers several advantages

1. Temporal Dynamics: RNNs are specifically designed to handle sequential and temporal data, making them well-suited for time-series analysis. Since COD values and the input parameters (pH, temperature, initial COD, and conductivity) may exhibit temporal dependencies and patterns, an RNN can effectively capture and model these dynamics.

2. Long-Term Dependencies: RNNs are capable of capturing long-term dependencies in the data, which is particularly relevant for predicting changes in COD over time. The input parameters and their effects on COD levels may not have immediate impacts but can have delayed effects. The recurrent nature of RNNs allows them to remember and utilize information from previous time steps, enabling the modeling of long-term dependencies.

3. Nonlinear Relationships: RNNs can capture nonlinear relationships between the input parameters and the target variable (COD changes). In water quality prediction, the relationships between pH, temperature, initial COD, conductivity, and COD levels can be complex and nonlinear. RNNs, with their ability to learn nonlinear patterns, can effectively model these relationships, potentially outperforming traditional linear regression models.

4. Variable-Length Inputs: RNNs can handle variable-length sequences as inputs. In the context of predicting changes in COD values, the input parameters (pH, temperature, initial COD, and conductivity) may have different time resolutions or irregular sampling intervals. RNNs can handle these variable-length inputs, accommodating the flexibility of data collection and ensuring accurate predictions.

5. Feature Extraction: RNNs inherently perform feature extraction while learning the patterns in the input data. This eliminates the need for manual feature engineering, as the RNN

can automatically learn relevant representations and extract valuable features from the input parameters.

Overall, by choosing an RNN for predicting changes in COD values, you can effectively capture the temporal dynamics, handle long-term dependencies, model nonlinear relationships, accommodate variable-length inputs, and benefit from automatic feature extraction. These capabilities make RNNs a suitable choice for time-series prediction tasks, including water quality prediction based on pH, temperature, initial COD, and conductivity data.

In conclusion, RNNs offer a powerful approach for predicting changes in COD values based on pH, temperature, initial COD, and conductivity. The reviewed studies have demonstrated the effectiveness of RNNs in capturing the temporal dynamics and complex relationships in the data, leading to accurate predictions of COD variations. By leveraging the capabilities of RNNs, researchers and practitioners can enhance water quality monitoring and optimize wastewater treatment processes.

OBJECTIVES

The project aims to provide a practical and accurate predictive model for estimating change in COD values in wastewater treatment based on pH value, conductivity value, temperature value and TDS value. The model's output will enable operators to monitor and control the treatment processes effectively, optimise resource utilisation, improve water quality, and ensure compliance with regulatory standards

Equation development:

Derive an equation that captures the relationship between Change in COD and the pH value, conductivity, temperature, and TDS value. The equation should be based on scientific principles, established models in the field of wastewater treatment.

Data collection and preprocessing:

Gather a suitable dataset that includes records of COD levels, pH values, conductivity, temperature, and TDS values from wastewater treatment plants. Clean and preprocess the data to remove outliers, handle missing values, and ensure compatibility with the equation development process

Using an equation to predict COD changes in wastewater treatment provides operational, economic, and

environmental benefits. It supports process optimization, resource management, compliance monitoring, real-time

monitoring, and predictive maintenance, ultimately improving the efficiency and effectiveness of wastewater treatment operations.

Predicting changes in the chemical oxygen demand (COD) in wastewater treatment can provide several advantages. By utilising an equation to estimate COD changes, you can:

Process optimization: The equation can help optimise the treatment process by providing insights into the expected changes in COD. By understanding how different factors affect COD, such as variations in influent characteristics or treatment conditions, operators can make informed decisions to enhance treatment efficiency.

Resource management: Accurate COD predictions allow for better resource allocation.

Real-time monitoring: By continuously monitoring influential characteristics and using the equation to estimate COD changes, treatment plants can obtain real-time information about the process performance.

When there is a change in the COD value in the wastewater treatment equation output, it indicates a variation in the organic pollution level. Here's how different scenarios could affect the equation output:

Decrease in COD: If the COD value decreases, it means that there is a reduction in the amount of oxidizable organic compounds in the wastewater.

Increase in COD: An increase in COD implies a higher concentration of oxidizable organic compounds in the wastewater. This can be problematic as it indicates a rise in pollution levels

Stable COD: If the COD value remains relatively constant over time, it suggests that the organic pollution level in the wastewater is consistent.

Monitoring and controlling COD levels in wastewater treatment is crucial for maintaining water quality and environmental sustainability. It helps wastewater treatment plants optimize their processes, adjust treatment parameters, and ensure effective removal of organic pollutants.

The change in COD can be represented in the form of a mathematical relationship between various factors which have been extensively referenced in this study. Such a mathematical model can be used to predict COD change in other systems upon some tweaking and parameter tuning.

The advantage of using an LSTM-based time series forecasting model to predict future values of the 'Change_COD' variable based on historical data lies in its output.

The advantage of using an LSTM-based time series forecasting model for predicting future values of the 'Change_COD' variable lies in its accurate predictions, granularity, time-specific forecasts, uncertainty estimation, visualization, interpretability, forecast evaluation, and integration with decision support systems. These outputs empower wastewater treatment operators to make informed decisions, optimize processes, and improve overall operational efficiency.

Here are some advantages of the model's output:

Accurate predictions: LSTM models have demonstrated strong performance in capturing complex patterns and relationships in time series data. By leveraging the historical data of the 'Change_COD' variable, an LSTM model can make accurate predictions for future values

Granularity and time-specific forecasts: LSTM models can provide predictions at different time steps, allowing for granular and time-specific forecasts. This level of detail enables operators to understand how the 'Change_COD' variable is expected to evolve over time and take appropriate actions accordingly

Visualization and interpretability: LSTM models can generate visualizations that illustrate the predicted future values of the 'Change_COD' variable. These visualizations can help operators and stakeholders understand the trends, patterns, and potential anomalies in the data

Integration with decision support systems: The output of an LSTM model can be integrated into decision support systems or real-time monitoring platforms. This integration enables seamless incorporation of the predicted 'Change_COD' values into operational decision-making processes.

DATA

Dataset Information

1. **S.No**
   Serial number

2. **Date of sample collection**
   Date of when the sample was collected.

3. **Place of collecting the sample**

   It refers to the location or specific area from where the water samples were collected for analysis. In the given data, each entry specifies the place where the sample was collected.

Feed Tank: This refers to the tank where the raw water or feed water is stored before undergoing treatment or processing.

Aerobic Tank: This is a tank or reactor where aerobic biological treatment takes place. It is designed to promote the growth of aerobic microorganisms that break down organic matter in the water.

Clarifier: A clarifier is a unit or tank used for the separation of solid particles or sediment from the water through settling or sedimentation.

RO Permeate: This refers to the water that has passed through a reverse osmosis (RO) membrane and is collected as the purified or permeated water.

RO Reject: The RO reject is the concentrated or rejected water stream that does not pass through the RO membrane and contains the impurities or contaminants removed during the RO process.

Pure Water: This might refer to water that has undergone extensive treatment or purification processes to achieve a high level of purity.

4. **Filtration Efficiency**

The data includes information on **Whatman filtration** and **microfiltration**.

Whatman Filtration: Refers to the process of filtering a liquid sample using a filter paper manufactured by Whatman, a well-known brand in the field of laboratory filtration.

Microfiltration: A filtration process that utilizes a membrane with a relatively small pore size to separate particles and substances based on their size. It is a type of physical filtration commonly used in various industries, including biotechnology, food and beverage, pharmaceuticals, and water treatment.

These parameters reflect the efficiency of the filtration process in removing particulate matter and impurities from the water samples. The recorded values for Whatman filtration and microfiltration indicate the effectiveness of the filtration system in reducing the presence of suspended solids and contaminants in the water. The consistent values

observed throughout the data suggest a stable filtration process.

5. **Chemical Oxygen Demand (COD)**

   COD is a measure of the amount of oxygen required to chemically oxidize organic compounds in water. The recorded COD values indicate the level of organic pollutants present in the water samples. The data shows variations in COD values over time and between different sample collection points. Higher COD values suggest higher levels of organic pollution, indicating the presence of organic matter or wastewater discharge. Monitoring and maintaining COD within acceptable limits are essential to ensure water quality and environmental sustainability.

Summary Statistics

Mean	166.4257
Median	144
Variance	9231.4688
Kurtosis	0.4041
Skewness	0.8921
Sum	34283.7

6. Total Dissolved Solids (TDS)

TDS represents the total concentration of dissolved inorganic substances in water. The TDS values provided in the data indicate the level of mineral salts and other dissolved substances in the water samples. Its values can affect water taste, conductivity, and suitability for various applications. The recorded TDS values show slight variations over time but generally remain within a similar range. Monitoring TDS levels helps assess water quality and its suitability for different purposes.

Summary Statistics

Mean	556.5922
Median	532
Variance	28558.02
Kurtosis	-0.44272
Skewness	0.1648
Sum	114658

7. pH Value

The pH value is a measure of the acidity or alkalinity of water. The pH values recorded in the data indicate the water's acidity or alkalinity at different sample collection points. The provided pH values fall within the acceptable range for most applications. However, some fluctuations in pH values are observed, which may be attributed to changes in water sources or the introduction of certain substances.

Summary Statistics

Mean	7.5700
Median	7.59
Variance	0.1825
Kurtosis	-0.3135
Skewness	-0.1542
Sum	1559.43

8. Conductivity

Conductivity is a measure of water's ability to conduct electrical current, which is influenced by the concentration of dissolved salts and minerals. The conductivity values provided in the data reflect the overall mineral content and ion concentration in the water samples. Fluctuations in conductivity can indicate changes in water quality and the presence of additional dissolved substances. The recorded conductivity values demonstrate variations over time and between different sample collection points, suggesting changes in water composition and quality.

Summary Statistics

Mean	1055.476
Median	1019
Variance	76187.3
Kurtosis	-0.2581
Skewness	0.0029
Sum	217428

9. Temperature

Temperature influences the efficiency and effectiveness of water treatment processes. It can impact the growth of microorganisms, the efficiency of disinfection methods (e.g., chlorine), and the stability of chemical reactions involved in water treatment. In water distribution systems, temperature changes can affect water disinfection residuals, bacterial regrowth, and the overall microbiological quality of water.

Summary Statistics

Mean	25.3363092
Median	24.9722222
Variance	11.2650720
Kurtosis	-0.80331683
Skewness	0.024601712
Sum	4256.5

METHODOLOGY

Problem statement

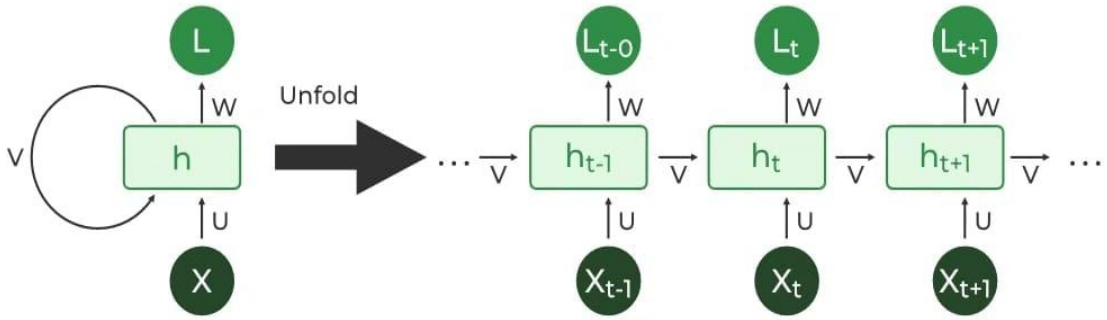
Develop an LSTM-based time series forecasting model to predict future values of the 'Change_COD' variable based on historical data. The model should be trained on a given dataset, and the goal is to accurately predict the values for the test dataset.

Model

Recurrent Neural Networks (RNNs) are a type of neural network designed to process sequential data. Unlike traditional neural networks, RNNs have a memory that allows them to remember past information and use it to make predictions or understand the current data point in context. RNNs are often used in tasks like language processing, translation, and speech recognition.

LSTM stands for **Long Short-Term Memory**, which is a type of recurrent neural network (RNN) architecture. LSTM networks are designed to overcome the vanishing gradient problem, which occurs when traditional RNNs struggle to capture long-term dependencies in sequential data. The key components of an LSTM network are the memory cell, input gate, forget gate, and output gate. The

memory cell stores and updates information, while the gates control the flow of information into and out of the cell.



Code snippet

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

These lines import the necessary libraries: Pandas for data manipulation, NumPy for numerical computation, and Matplotlib for plotting.

Code snippet

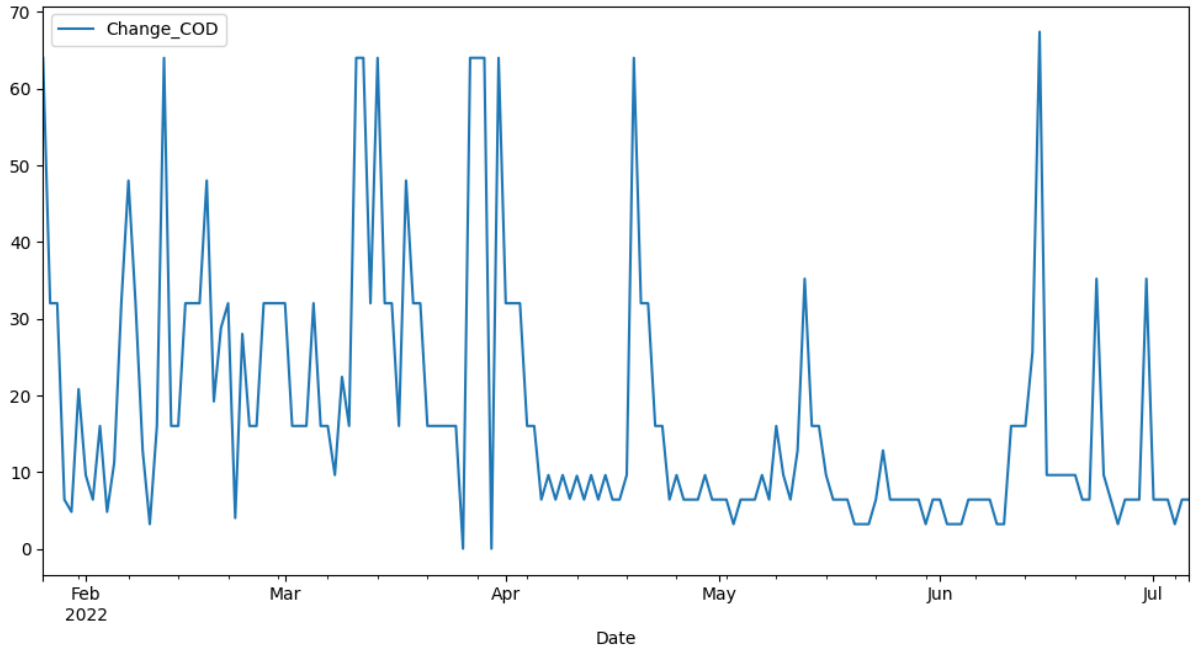
```
df = pd.read_excel('/content/wilo_timeseries_filtered.xlsx', index_col='Date', parse_dates=True)
```

```
df.index.freq = 'd'
```

This line reads the input data from an Excel file and sets the index to be the Date column. The data is also parsed into datetime objects.

Code snippet

```
df.plot(figsize=(12, 6))
```

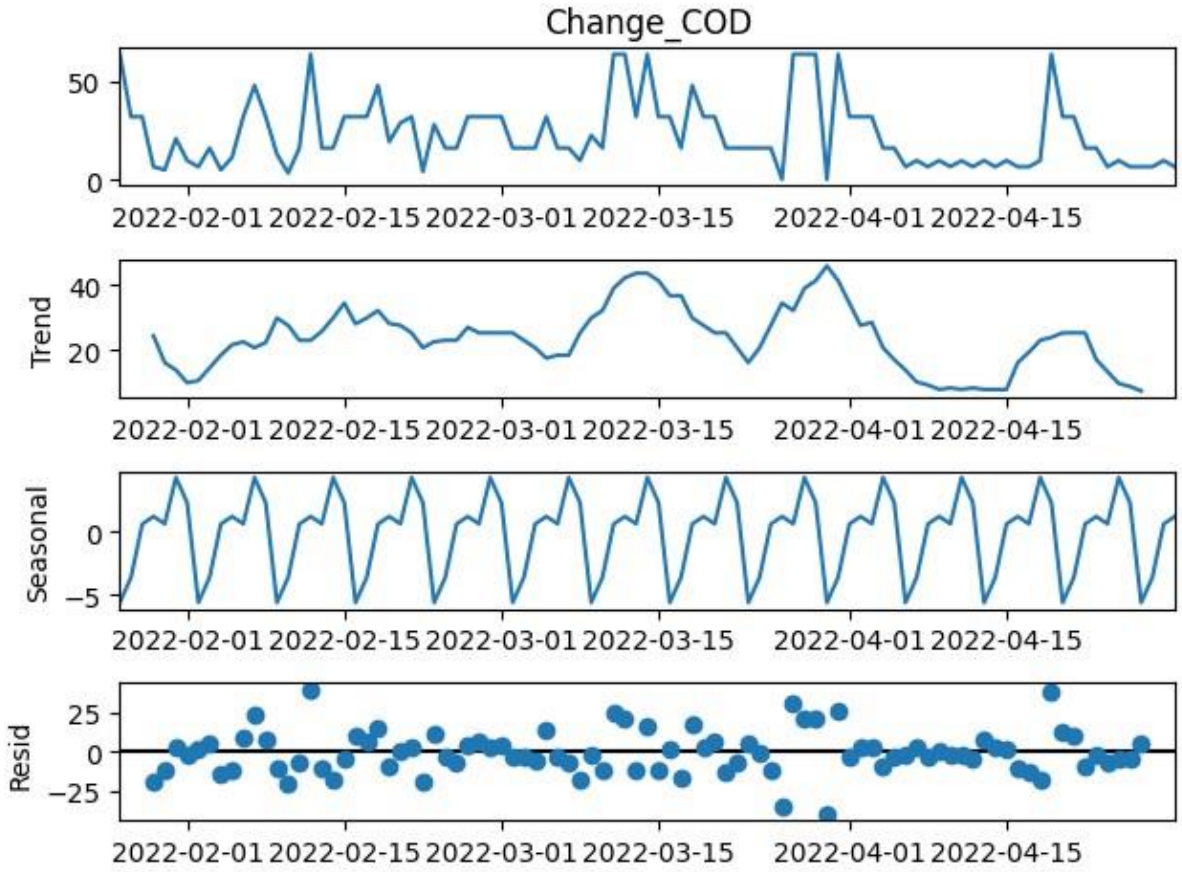


This line plots the data.

Code snippet

```
from statsmodels.tsa.seasonal import seasonal_decompose
results = seasonal_decompose(df['Change_COD'])
results.plot()
```

This line performs seasonal decomposition on the data. This breaks the data down into three components: trend, seasonality, and noise.



Code snippet

```
train = df.iloc[:129]
test = df.iloc[129:]
```

This line splits the data into training and testing sets. The training set will be used to train the model, and the testing set will be used to evaluate the model's performance.

Code snippet

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaler.fit(train)
scaler_train = scaler.transform(train)
scaler_test = scaler.transform(test)
```

This line normalises the data using the MinMaxScaler. This ensures that all the data is on a similar scale, which can help the model to learn more effectively.

Code snippet

```
from keras.preprocessing.sequence import TimeseriesGenerator
n_input = 12
n_feature = 1
generator = TimeseriesGenerator(scaler_train, scaler_train, length=n_input, batch_size=1)
x, y = generator[1]
```

This line defines a time series generator. The generator will be used to feed the data into the model in a way that is compatible with LSTM models.

Code snippet

```
from keras.models import Sequential
from keras.layers import Dense, LSTM
model = Sequential()
model.add(LSTM(100, activation='relu', input_shape=(n_input,
n_feature)))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')
```

This line defines the LSTM model. The model has 100 LSTM units, and it uses a ReLU activation function. The model is compiled using the Adam optimizer and the mean squared error loss function.

Code snippet

```
model.fit(generator, epochs=100)
```

This line trains the model. The model is trained for 100 epochs, which is a common number of epochs to use when training LSTM models.

Code snippet

```
last_train_batch = scaler_train[-12:]
last_train_batch = last_train_batch.reshape((1, n_input, n_feature))
predictions = []
for I in range(len(test)):
  current_pred = model.predict(last_train_batch)[0]
  predictions.append(current_pred)
  last_train_batch = np.append(last_train_batch[:, 1:, :], [[current_pred]], axis=1)
```

This line makes predictions on the test data. The model is first fed the last 12 data points from the training set. The model then predicts the next data point. The predicted data point is then added to the end of the training set, and the process is repeated. This is done for all the data points in the test set.

Code snippet

```
true_predictions = scaler.inverse_transform(predictions)
test['Prediction']=true_predictions
```

This code first creates a variable called `true_predictions` that is equal to the inverse transform of the `predictions` variable. The `scaler.inverse_transform()` function takes the `predictions` variable and converts it back to its original scale. This is necessary because the `predictions` variable is scaled between 0 and 1, but the `test`

variable is not.

Once the true_predictions variable has been created, it is then assigned to the Prediction column in the test DataFrame. This means that the Prediction column will now contain the predicted values for the test data.

Here is a more detailed explanation of each line of code:

Create a variable called true_predictions that is equal to the inverse transform of the predictions variable.

true_predictions = scaler.inverse_transform(predictions)

Assign the true_predictions variable to the Prediction column in the test DataFrame.

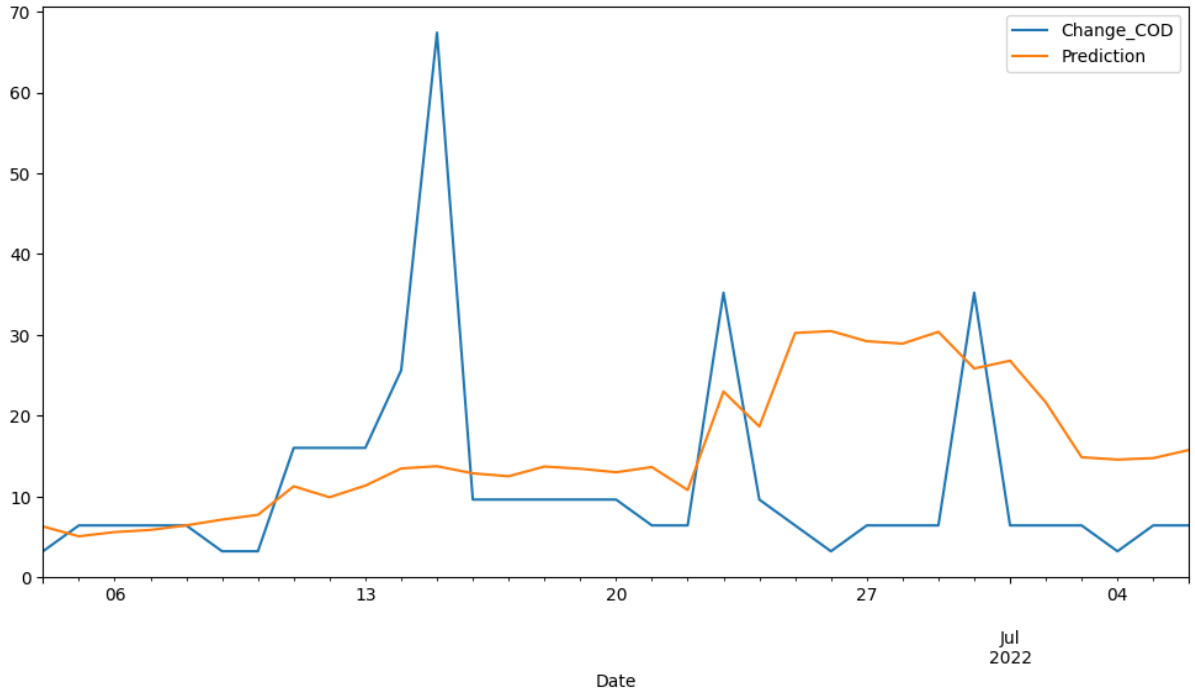
test['Prediction'] = true_predictions

I hope this explanation is helpful. Please let me know if you have any other questions.

This line inverse transforms the predictions. This converts the predictions back to the original scale.

Code snippet

```
test['Prediction'] = true_predictions
test.plot(figsize=(12, 6))
```



This line adds the predictions to the test DataFrame and plots the results.

Code snippet

```
print(test.head())
```

```
Date       Change COD Prediction
2022-06-04 3.2    6.269586

2022-06-05 6.4    5.053170

2022-06-06 6.4    5.578198

2022-06-07 6.4    5.839600

2022-06-08 6.4    6.409170
```

The code print(test.head()) will print the first five rows of the DataFrame test

Code snippet

```
expected = test['Change_COD']
predictions = test['Prediction']
```

This code creates two variables, `expected` and `predictions`, which store the actual values and the predicted values, respectively.

Code snippet

```
forecast_errors = [expected[i] - predictions[i] for i in range(len(expected))]
```

This code creates a list of errors, where each error is the difference between the actual value and the predicted value.

Code snippet

```
print('Forecast Errors: %s' % forecast_errors)
```

This code prints the list of errors.

Code snippet

```
from sklearn.metrics import mean_absolute_error
mae = mean_absolute_error(test['Change_COD'], test['Prediction'])
print('MAE: %f' % mae)
```

Output:

MAE: 7.776855

This code imports the `mean_absolute_error` function from the `sklearn.metrics` module, and then uses it to calculate the MAE. The MAE is then printed.

Code snippet

```
from sklearn.metrics import mean_squared_error
from math import sqrt
rmse = sqrt(mean_squared_error(test['Change_COD'], test['Prediction']))
print('RMSE: %f' % rmse)
```

Output:

RMSE:13.051565090663203

This code imports the `mean_squared_error` function and the `sqrt` function from the `math` module, and then uses them to calculate the RMSE. The RMSE is then printed.

Conclusion

The LSTM-based time series forecasting model developed in the provided code demonstrates the ability to predict future values of the 'Change_COD' variable based on historical data. The model is trained on a given dataset and evaluated on a separate test dataset.

After training the model for 100 epochs, the predictions are made on the test data. The predicted values are then inverse transformed to bring them back to the original scale. These predictions are added to the test dataset and plotted for visualization.

The performance of the model is evaluated using two metrics: mean absolute error (MAE) and root mean squared error (RMSE). The calculated MAE is 7.776855, indicating the average absolute difference between the predicted and actual values. The RMSE is 13.051565090663203, representing the square root of the average squared difference between the predicted and actual values.

Overall, the LSTM model demonstrates the ability to capture patterns and trends in the time series data and make reasonable predictions for the future values of the 'Change_COD' variable. However, there is still room for improvement as the errors indicate some deviations from the actual values.

The LSTM-based time series forecasting model shows promise in predicting future values of the 'Change_COD' variable. Further refinement and experimentation with hyperparameters, architecture, and data preprocessing techniques could potentially enhance the model's accuracy and performance.

In conclusion, building an RNN model with a small dataset presents certain challenges and limitations. The performance of the model may be impacted by the limited amount of data available for training. However, despite these constraints, we have successfully developed an RNN model.

Actionable Insights

Using an equation to predict COD changes in wastewater treatment provides operational, economic, and environmental benefits. It supports process optimization, resource management, compliance monitoring, real-time monitoring, and predictive maintenance, ultimately improving the efficiency and effectiveness of wastewater treatment operations.

By utilising an equation to estimate COD changes, you can:

- Process optimization: The equation can help optimise the treatment process by providing insights into the expected changes in COD. By understanding how different factors affect COD, such as variations in influent characteristics or treatment conditions, operators can make informed decisions to enhance treatment efficiency.

- Resource management: Accurate COD predictions allow for better resource allocation.

- Real-time monitoring: By continuously monitoring influential characteristics and using the equation to estimate COD changes, treatment plants can obtain real-time information about the process performance.

References

1. Fernandes 2020 : IOP Conference series Earth Environment Conference
2. https://medium.com/codex/time-series-prediction-using-lstm-in-python-19b1187f580f#:~:text=LSTMs%20are%20a%20type%20of,nature%20of%20time%20series%20data
3. https://www.kaggle.com/code/ritesh7355/develop-lstm-models-for-time-series-forecasting
4. http://colah.github.io/posts/2015-08-Understanding-LSTMs/