Assignment in

Machine Learning

Machine learning Model for AI-Driven Bioremediation: Optimizing Microbial Degradation of Organic Pollutants in Soil and Water

Submitted By

Nirai Anbu 21BCE0965

K. Tarun 21BCE3082

N.Anish Balaji 21BCE3885

1. Introduction

Bioremediation is a natural process that uses microbial activity to break down organic pollutants in the environment, transforming them into less harmful substances. This project leverages artificial intelligence (AI) to optimize the bioremediation process, ensuring effective and efficient degradation of organic pollutants in soil and water. The focus is on using AI to monitor conditions, predict pollutant levels, and provide insights that enhance microbial degradation.

2.Project Scope and Functionalities

The project integrates microbiology, environmental engineering, data science, and synthetic biology to create a comprehensive system that can effectively and efficiently degrade organic pollutants, making it a promising approach to address environmental contamination challenges.

1. **Microbial Identification and Selection**:

The system identifies and selects optimal microbial strains based on their capacity to degrade specific organic pollutants. These microbes are chosen from a diverse set of bacteria, fungi, and other microorganisms that exhibit high degradative abilities. This selection process involves screening various microbial strains from contaminated environments and laboratory culture collections to identify those with the highest degradation potential. Advanced molecular techniques, such as metagenomics and 16S rRNA sequencing, are used to characterize and select the most promising strains.

2. **Pollutant Profiling:**
   The project involves the identification and profiling of organic pollutants present in contaminated soil and water samples. This profiling helps in understanding the type and concentration of pollutants, which in turn determines the best microbial candidates for degradation. Analytical techniques such as gas chromatography-mass spectrometry (GC-MS) and high-performance liquid chromatography (HPLC) are used to accurately identify and quantify pollutants.

3. **Bioreactor Optimization:**
   The system includes a bioreactor setup that allows for controlled experiments to optimize environmental conditions, such as temperature, pH, nutrient levels, and oxygen supply, which are crucial for maximizing microbial degradation efficiency. The bioreactors are equipped with sensors and control systems to maintain optimal conditions and collect data in real-time. Different bioreactor configurations, such as batch, fed-batch, and continuous flow, are evaluated to determine the most effective setup for pollutant degradation.

4. **Enzyme Analysis and Enhancement:**
   The project examines the enzymes involved in the degradation pathways of selected microbes. Enzyme activity is analysed, and strategies are implemented to enhance their catalytic efficiency through genetic engineering or cofactor addition. Enzyme assays are conducted to measure the activity of key enzymes, and techniques such as site-directed mutagenesis and directed evolution are used to improve enzyme performance. The goal is to boost the rate of pollutant breakdown and expand the range of compounds that can be degraded.

5. **Field Application and Monitoring:**
   The optimized microbial consortia are tested in real-world scenarios. Soil and water samples are inoculated with the selected microbial strains, and the degradation progress is monitored through regular sampling and pollutant quantification using chromatographic and spectroscopic techniques. Portable field kits and remote sensing technologies are employed for on-site monitoring, providing real-time feedback on pollutant levels and microbial activity.

6. **Data Collection and Machine Learning Integration:**
   Data regarding environmental conditions, pollutant concentrations, and microbial activity are collected over time. A machine learning model is used to predict optimal conditions for microbial activity and guide adjustments in real-time to improve degradation outcomes. The machine learning models are trained using historical data to identify patterns and predict the most effective operational parameters. Techniques such as regression analysis, clustering, and neural networks are employed to optimize the degradation process dynamically.

7. **Community Interactions:**
   The project takes into account microbial community interactions, as different microbial species may work synergistically to degrade complex pollutants. The system models these interactions to develop a balanced microbial consortium capable of handling a diverse range of pollutants. Co-culture experiments are conducted to evaluate the interactions between different microbial species, and microbial community dynamics are monitored using techniques such as fluorescence in situ hybridization (FISH) and next-generation sequencing (NGS).

8. **Toxicity Assessment:**
   The toxicity of both pollutants and their degradation by-products is continuously assessed to ensure that the microbial degradation process results in non-toxic end-products, making the treated soil and water safe for the environment. Ecotoxicological assays, such as tests on aquatic organisms and plants, are conducted to evaluate the impact of the degradation process. Additionally, advanced chemical analysis is used to identify any harmful by-products formed during degradation.

9. **Scalable Implementation:**
   Strategies are being developed to scale the process from laboratory to field-level applications, with the goal of ensuring the degradation process is effective across different environmental conditions and pollutant concentrations. Pilot-scale field trials are conducted to evaluate the scalability of the microbial degradation approach. Factors such as microbial inoculum delivery methods, nutrient amendments, and environmental variability are considered to ensure successful large-scale implementation.

10. **Nutrient and Electron Donor Supplementation:**
    To enhance microbial activity, nutrient supplementation is provided when necessary. This includes adding sources of nitrogen, phosphorus, and trace elements to support microbial growth. Additionally, electron donors or acceptors are supplied to stimulate the specific metabolic pathways involved in pollutant degradation, depending on whether the process is aerobic or anaerobic.

11. **Genetic Engineering and Synthetic Biology:**

Genetic engineering tools are used to enhance the pollutant-degrading capabilities of selected microbial strains. Synthetic biology approaches are applied to construct engineered microbial consortia with tailored properties, such as improved degradation rates and resistance to toxic pollutants. Plasmid-based expression systems and CRISPR-Cas9 technology are employed to introduce and regulate genes involved in pollutant degradation.

12. **Bio stimulation and Bioaugmentation:**
    The project utilizes bio stimulation and bioaugmentation strategies to enhance pollutant degradation. Bio stimulation involves modifying environmental conditions to stimulate the growth of indigenous pollutant-degrading microbes, while bioaugmentation involves adding selected microbial strains to contaminated sites to boost degradation rates. These strategies are tailored based on the pollutant type and site characteristics.

13. **Mathematical Modelling and Process Simulation:**
    Mathematical models are developed to simulate the microbial degradation process under various environmental conditions. These models help in understanding the kinetics of pollutant degradation and predicting the outcomes of field applications. Process simulation tools are used to optimize operational parameters and design large-scale bioremediation systems.

The project integrates microbiology, environmental engineering, data science, and synthetic biology to create a comprehensive system that can effectively and efficiently degrade organic pollutants, making it a promising approach to address environmental contamination challenges.

3. Observation and Additional Functionalities

Observation

Pollution from organic chemicals such as pesticides, herbicides, and industrial waste is a significant threat to soil and water quality, affecting both the environment and human health. Traditional remediation methods are often slow, costly, or inefficient. The existing approaches may also fail to address complex and diverse contamination scenarios. There is a need for a more adaptive and efficient solution that can handle different pollutants while ensuring minimal environmental disruption. The goal is to overcome these challenges by integrating AI with bioremediation, which will enhance the degradation process and provide continuous monitoring and control.

Possible Additional Functionality

Data Collection and Machine Learning Integration:

Data regarding environmental conditions, pollutant concentrations, and microbial activity are collected over time. A machine learning model is used to predict optimal conditions for microbial activity and guide adjustments in real-time to improve degradation outcomes. The machine learning models are trained using historical data to identify patterns and predict the most effective operational parameters. Techniques such as regression analysis, clustering, and neural networks are employed to optimize the degradation process dynamically.

Additional Machine Learning Functionalities:

- **Reinforcement Learning**: Implementing reinforcement learning to dynamically adjust operational parameters, such as nutrient supply, temperature, and pH, based on real-time feedback to maximize degradation efficiency.
- **Transfer Learning**: Leveraging pre-trained models from similar pollutant degradation projects to reduce the need for extensive data collection and speed up model training for new pollutant types.
- **Predictive Analytics**: Using machine learning to forecast potential issues, such as microbial population decline or pollutant rebound, allowing for preemptive corrective actions.
- **Explainable AI (XAI)**: Incorporating XAI to better understand how the model makes decisions, enabling researchers to fine-tune parameters more effectively and improve trust in the model's predictions.
- **Anomaly Detection**: Developing models to detect anomalies in microbial activity or environmental conditions, allowing for quick identification and mitigation of issues that could hinder the degradation process.
- **Graph-Based Machine Learning**: Applying graph-based models to understand and optimize microbial community interactions, predicting synergistic or antagonistic relationships among microbes to design more effective consortia.
- **Deep Learning for Image Analysis**: Utilizing deep learning for analyzing images of microbial cultures, biofilms, and soil samples to monitor microbial health and pollutant degradation visually.

4. Solution Overview

The proposed solution involves using AI-driven models to enhance the natural process of bioremediation. By integrating AI with microbial technologies, we can achieve efficient pollutant degradation in various environments. Sensors will collect data on key environmental parameters (e.g., pollutant levels, pH, temperature, and moisture), which will be fed into AI algorithms to predict pollutant behaviour and optimize conditions for microbial activity. The system will adaptively respond to changes in the environment, ensuring continuous and efficient bioremediation.

Data Collection and Management

Data collection is carried out through IoT sensors that gather information on soil and water quality, including pH, temperature, moisture, and pollutant concentrations. The data is transmitted to a central cloud-based database for storage and analysis. Data management practices ensure that the data remains accurate, secure, and accessible for analytics and decision-making.

Generate the Raw data with 10,000 Samples

Code

```
import pandas as pd
import numpy as np

# Generating a sample dataset with 10,000 rows and 7 variables after preprocessing
np.random.seed(42)

# Variables:
# 1. pH Level (range: 4.0 - 9.0)
# 2. Temperature (Celsius, range: 5 - 40)
# 3. Moisture Content (% range: 10 - 100)
# 4. Pollutant Concentration (mg/L, range: 0 - 500)
# 5. Microbial Population Density (CFU/mL, range: 10^3 - 10^9)
# 6. Nutrient Level (mg/L, range: 0 - 100)
# 7. Predicted Pollutant Degradation Efficiency (% range: 0 - 100)

# Creating data dictionary
data = {
    "pH_Level": np.round(np.random.uniform(4.0, 9.0, 10000), 2),
    "Temperature_C": np.round(np.random.uniform(5, 40, 10000), 1),
    "Moisture_Content_%": np.round(np.random.uniform(10, 100, 10000), 1),
    "Pollutant_Concentration_mg_L": np.round(np.random.uniform(0, 500, 10000), 2),
    "Microbial_Population_Density_CFU_mL": np.round(np.random.uniform(1e3, 1e9, 10000)),
    "Nutrient_Level_mg_L": np.round(np.random.uniform(0, 100, 10000), 1),
    "Predicted_Degradation_Efficiency_%": np.round(np.random.uniform(0, 100, 10000), 1),
}

# Creating DataFrame
df = pd.DataFrame(data)
```

```
# Displaying the first few rows of the dataframe
print(df.head())


# Optionally, visualize the data using histograms
import matplotlib.pyplot as plt


# Plotting histograms for each variable
df.hist(bins=50, figsize=(15, 10))
plt.tight_layout()
plt.show()
```

Output

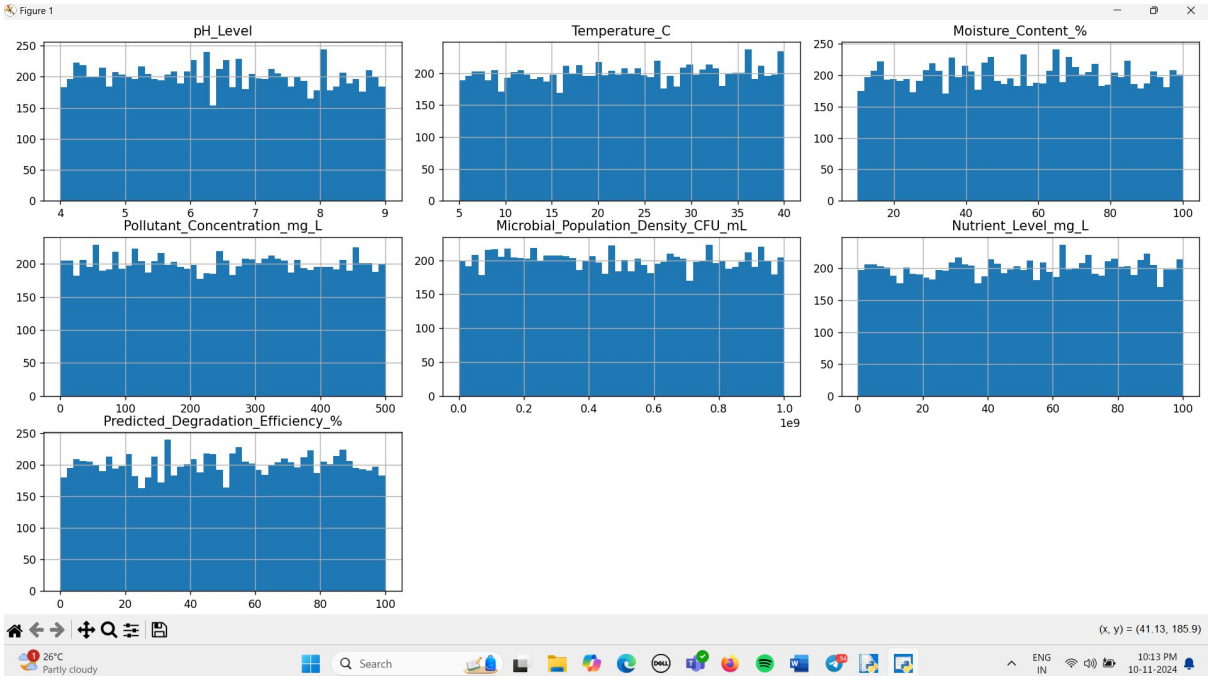
This code will display only first few rows of the data as shown below

```
>>>
========== RESTART: D:/NiraiAnbu/Machine-Learn-Preprocess-raw-data-generate.py ==========
   pH_Level  ...  Predicted_Degradation_Efficiency_%
0      5.87  ...                                74.2
1      8.75  ...                                88.1
2      7.66  ...                                46.3
3      6.99  ...                                28.9
4      4.78  ...                                31.9

[5 rows x 7 columns]
```

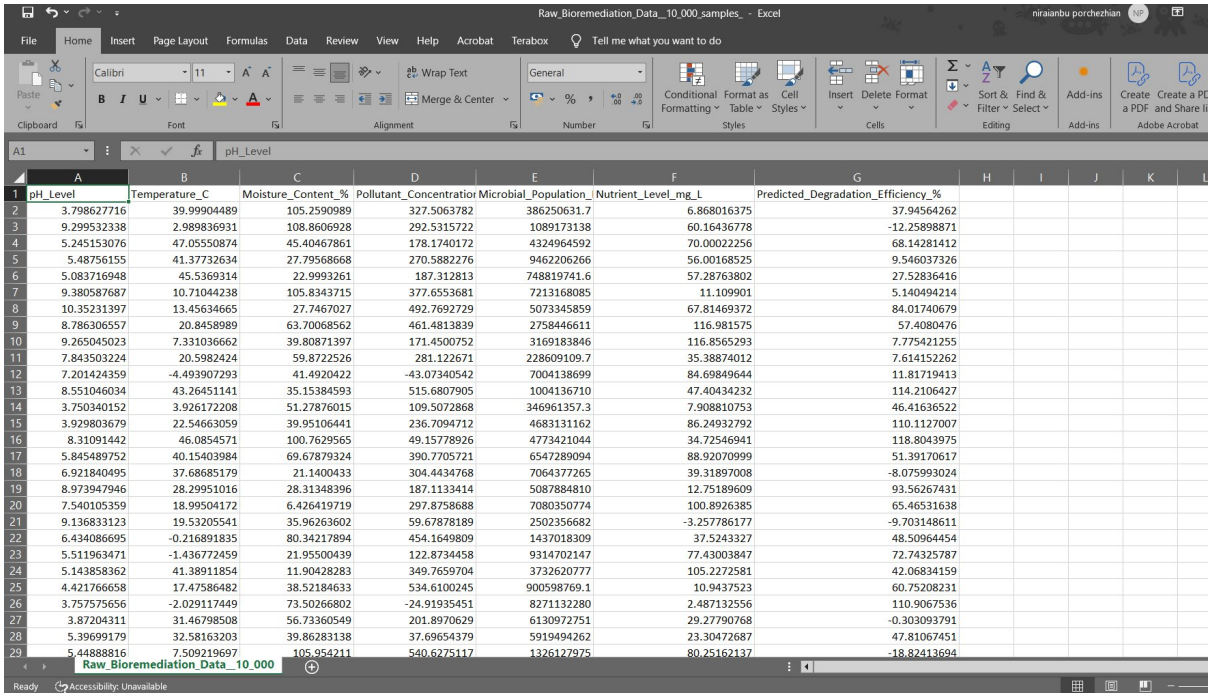
Processed output would have 10,000 samples as shown below.

Sample Bioremediation Data (10,000 Samples)

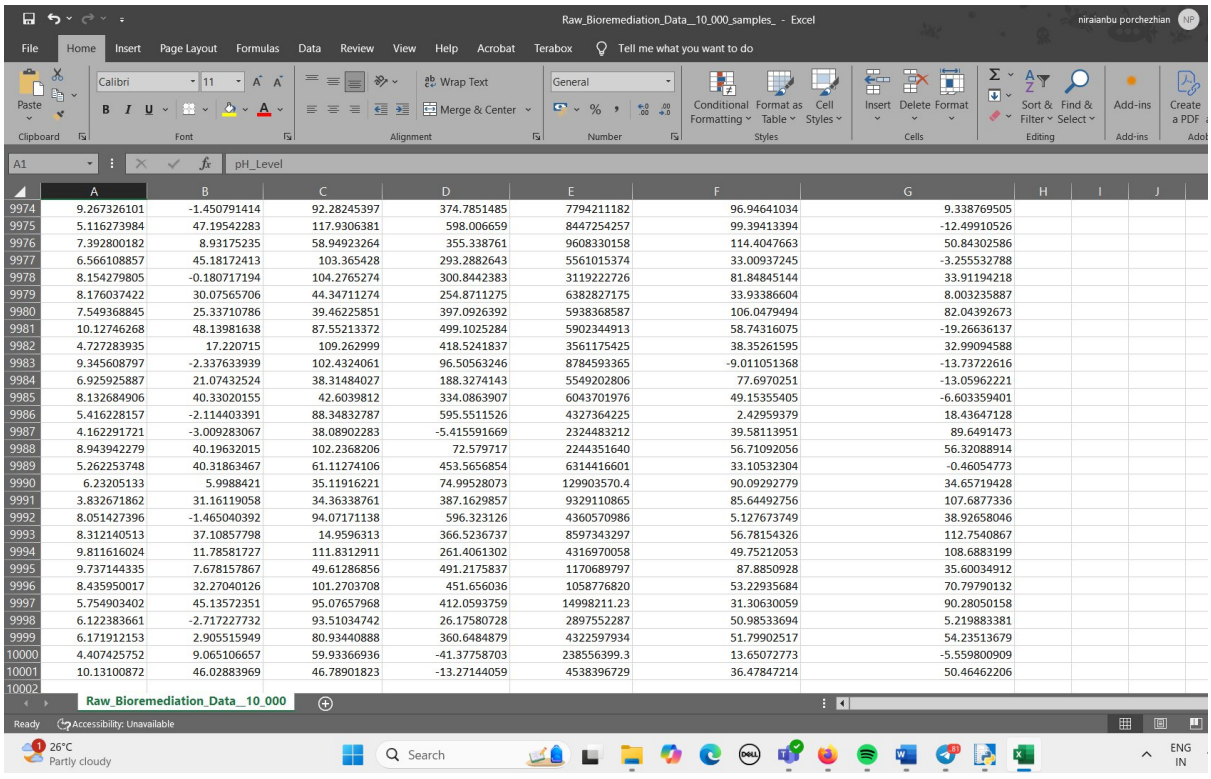
	pH_Level	Temperature_C	Moisture_Content_%	Pollutant_Concentratio	Microbial_Population_D	Nutrient_Level_mg_L	Predicted_Degradation
1	5.87	18.1	75.7	319.07	298912742.0	84.7	74.2
2	8.75	16.7	26.6	229.65	94818680.0	49.5	88.1
3	7.66	11.2	41.2	482.25	126360099.0	19.5	46.3
4	6.99	26.3	69.7	109.49	180671948.0	73.7	28.9
5	4.78	21.7	53.4	293.93	203654131.0	41.9	31.9
6	4.78	35.3	76.5	350.11	242262567.0	59.5	69.7
7	4.29	6.1	96.5	412.78	255460774.0	10.7	56.8
8	8.33	27.5	20.5	203.49	455716890.0	63.2	48.6
9	7.01	31.7	73.9	343.46	509573681.0	37.4	20.3
10	7.54	31.6	30.7	151.6	308879899.0	33.4	87.5
11	4.1	36.0	47.3	217.74	907296513.0	98.0	78.1
12	8.85	30.5	13.0	447.81	553140170.0	43.2	14.1
13	8.16	37.5	22.2	474.95	787477141.0	67.8	36.7
14	5.06	16.6	38.8	359.41	924237908.0	62.8	54.2
15	4.91	22.6	40.8	187.18	163135951.0	10.4	45.4
16	4.92	5.5	91.0	410.03	427875056.0	74.8	39.8
17	5.52	5.2	76.8	12.93	433210969.0	32.3	71.9
18	6.62	13.4	97.5	51.37	354472896.0	79.9	5.5
19	6.16	8.5	63.9	330.69	8906858.0	60.8	47.1
20	5.46	14.1	31.8	145.62	815972292.0	31.7	26.2
21	7.06	11.2	39.5	475.69	111560210.0	35.6	8.5
22	4.7	6.0	38.8	218.03	995480602.0	73.2	46.5
23	5.46	36.8	39.4	214.68	772958719.0	55.0	17.1



Downloaded Excel File – data output with 10,000 samples



	A	B	C	D	E	F	G
1	pH_Level	Temperature_C	Moisture_Content_%	Pollutant_Concentration	Microbial_Population_	Nutrient_Level_mg_L	Predicted_Degradation_Efficiency_%
2	3.798627716	39.99904489	105.2590989	327.5063782	386250631.7	6.868016375	37.94564262
3	9.299532338	2.989836931	108.8606928	292.5315722	1089173138	60.16436778	-12.25898871
4	5.245153076	47.05550874	45.40467861	178.1740172	4324964592	70.00022256	68.14281412
5	5.48756155	41.37732634	27.79568668	270.5882276	9462206266	56.00168525	9.546037326
6	5.083716948	45.5369314	22.9993261	187.312813	748819741.6	57.28763802	27.52836416
7	9.380587687	10.71044238	105.8343715	377.6553681	7213168085	11.109901	5.140494214
8	10.35231397	13.45634665	27.7467027	492.7692729	5073345859	67.81469372	84.01740679
9	8.786306557	20.8458989	63.70068562	461.4813839	2758446611	116.981575	57.4080476
10	9.265045023	7.331036662	39.80871397	171.4500752	3169183846	116.8565293	7.775421255
11	7.843503224	20.5982424	59.8722526	281.122671	228609109.7	35.38874012	7.614152262
12	7.201424359	-4.493907293	41.4920422	-43.07340542	7004138699	84.69849644	11.81719413
13	8.551046034	43.26451141	35.15384593	515.6807905	1004136710	47.40434232	114.2106427
14	3.750340152	3.926172208	51.27876015	109.5072868	346961357.3	7.908810753	46.41636522
15	3.929803679	22.54663059	39.95106441	236.7094712	4683131162	86.24932792	110.1127007
16	8.31091442	46.0854571	100.7629565	49.15778926	4773421044	34.72546941	118.8043975
17	5.845489752	40.15403984	69.67879324	390.7705721	6547289094	88.92070999	51.39170617
18	6.921840495	37.68685179	21.1400433	304.4434768	7064377265	39.31897008	-8.075993024
19	8.973947946	28.29951016	28.31348396	187.1133414	5087884810	12.75189609	93.56267431
20	7.540105359	18.99504172	6.426419719	297.8758688	7080350774	100.8926385	65.46531638
21	9.136833123	19.53205541	35.96263602	59.67878189	2502356682	-3.257786177	-9.703148611
22	6.434086695	-0.216891835	80.34217894	454.1649809	1437018309	37.5243327	48.50964454
23	5.511963471	-1.436772459	21.95500439	122.8734458	9314702147	77.43003847	72.74325787
24	5.143858362	41.38911854	11.90428283	349.7659704	3732620777	105.2272581	42.06834159
25	4.421766658	17.47586482	38.52184633	534.6100245	900598769.1	10.9437523	60.75208231
26	3.757575656	-2.029117449	73.50266802	-24.91935451	8271132280	2.487132556	110.9067536
27	3.87204311	31.46798508	56.73360549	201.8970629	6130972751	29.27790768	-0.303093791
28	5.39699179	32.58163203	39.86283138	37.69654379	5919494262	23.30472687	47.81067451
29	5.44888816	7.509219697	105.954211	540.6275117	1326127975	80.25162137	-18.82413694



	A	B	C	D	E	F	G
9974	9.267326101	-1.450791414	92.28245397	374.7851485	7794211182	96.94641034	9.338769505
9975	5.116273984	47.19542283	117.9306381	598.006659	8447254257	99.39413394	-12.49910526
9976	7.392800182	8.93175235	58.94923264	355.338761	9608330158	114.4047663	50.84302586
9977	6.566108857	45.18172413	103.365428	293.2882643	5561015374	33.00937245	-3.255532788
9978	8.154279805	-0.180717194	104.2765274	300.8442383	3119222726	81.84845144	33.91194218
9979	8.176037422	30.07565706	44.34711274	254.8711275	6382827175	33.93386604	8.003235887
9980	7.549368845	25.33710786	39.46225851	397.0926392	5938368587	106.0479494	82.04392673
9981	10.12746268	48.13981638	87.55213372	499.1025284	5902344913	58.74316075	-19.26636137
9982	4.727283935	17.220715	109.262999	418.5241837	3561175425	38.35261595	32.99094588
9983	9.345608797	-2.337633939	102.4324061	96.50563246	8784593365	-9.011051368	-13.73722616
9984	6.925925887	21.07432524	38.31484027	188.3274143	5549202806	77.6970251	-13.05962221
9985	8.132684906	40.33020155	42.6039812	334.0863907	6043701976	49.15355405	-6.603359401
9986	5.416228157	-2.114403391	88.34832787	595.5511526	4327364225	2.42959379	18.43647128
9987	4.162291721	-3.009283067	38.08902283	-5.415591669	2324483212	39.58113951	89.6491473
9988	8.943942279	40.19632015	102.2368206	72.579717	2244351640	56.71092056	56.32088914
9989	5.262253748	40.31863467	61.11274106	453.5656854	6314416601	33.10532304	-0.46054773
9990	6.23205133	5.9988421	35.11916221	74.99528073	129903570.4	90.09292779	34.65719428
9991	3.832671862	31.16119058	34.36338761	387.1629857	9329110865	85.64492756	107.6877336
9992	8.051427396	-1.465040392	94.07171138	596.323126	4360570986	5.127673749	38.92658046
9993	8.312140513	37.10857798	14.9596313	366.5236737	8597343297	56.78154326	112.7540867
9994	9.811616024	11.78581727	111.8312911	261.4061302	4316970058	49.75212053	108.6883199
9995	9.737144335	7.678157867	49.61286856	491.2175837	1170689797	87.8850928	35.60034912
9996	8.435950017	32.27040126	101.2703708	451.656036	1058776820	53.22935684	70.79790132
9997	5.754903402	45.13572351	95.07657968	412.0593759	14998211.23	31.30630059	90.28050158
9998	6.122383661	-2.717227732	93.51034742	26.17580728	2897552287	50.98533694	5.219883381
9999	6.171912153	2.905515949	80.93440888	360.6484879	4322597934	51.79902517	54.23513679
10000	4.407425752	9.065106657	59.93366936	-41.37758703	238556399.3	13.65072773	-5.559800909
10001	10.13100872	46.02883969	46.78901823	-13.27144059	4538396729	36.47847214	50.46462206
10002							

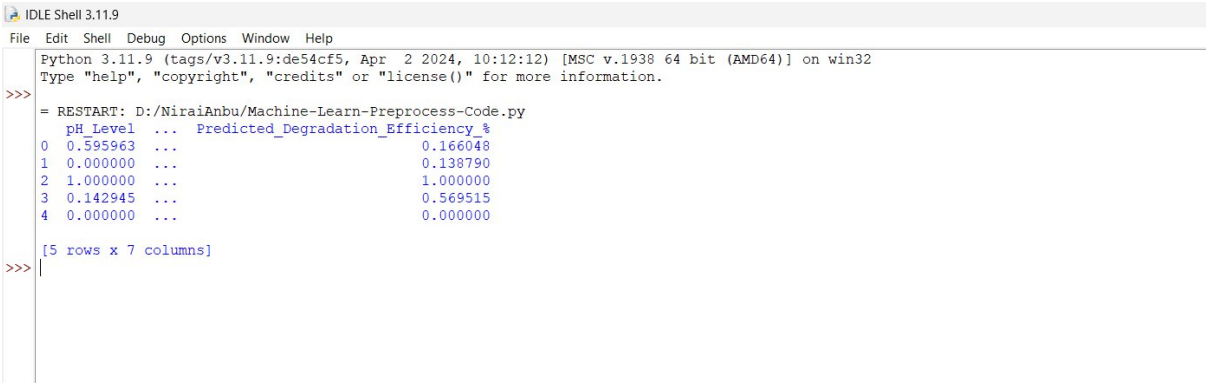
Preprocess data

```
# Preprocessing steps for the raw data

# Step 1: Handling Outliers

# - pH should be between 4.0 and 9.0

# - Temperature should be between 5 and 40 Celsius

# - Moisture Content should be between 10% and 100%

# - Pollutant Concentration should be >= 0 mg/L

# - Microbial Population Density capped to realistic range of 10^3 to 10^9 CFU/mL

# - Nutrient Level should be between 0 and 100 mg/L

# - Predicted Degradation Efficiency should be between 0 and 100%


processed_df = raw_df.copy()


# pH Level Clipping

processed_df['pH_Level'] = processed_df['pH_Level'].clip(4.0, 9.0)


# Temperature Clipping

processed_df['Temperature_C'] = processed_df['Temperature_C'].clip(5, 40)


# Moisture Content Clipping

processed_df['Moisture_Content_%'] = processed_df['Moisture_Content_%'].clip(10, 100)


# Pollutant Concentration - Removing Negative Values

processed_df['Pollutant_Concentration_mg_L'] =
processed_df['Pollutant_Concentration_mg_L'].clip(lower=0)


# Microbial Population Density Clipping

processed_df['Microbial_Population_Density_CFU_mL'] =
processed_df['Microbial_Population_Density_CFU_mL'].clip(1e3, 1e9)


# Nutrient Level Clipping

processed_df['Nutrient_Level_mg_L'] = processed_df['Nutrient_Level_mg_L'].clip(0, 100)
```

```
# Predicted Degradation Efficiency Clipping
processed_df['Predicted_Degradation_Efficiency_%'] =
processed_df['Predicted_Degradation_Efficiency_%'].clip(0, 100)


# Step 2: Handling Missing Values (if any)
# - Filling missing values with median values of each column
processed_df.fillna(processed_df.median(), inplace=True)


# Step 3: Standardizing/Scaling (Optional, depending on model requirements)
# - Here we normalize data for consistency in analysis if required for modeling
from sklearn.preprocessing import MinMaxScaler


scaler = MinMaxScaler()
scaled_columns = ['pH_Level', 'Temperature_C', 'Moisture_Content_%',
        'Pollutant_Concentration_mg_L', 'Microbial_Population_Density_CFU_mL',
        'Nutrient_Level_mg_L', 'Predicted_Degradation_Efficiency_%']


processed_df[scaled_columns] = scaler.fit_transform(processed_df[scaled_columns])


# Displaying the preprocessed dataframe to the user
tools.display_dataframe_to_user(name="Preprocessed Bioremediation Data (10,000 samples)",
dataframe=processed_df)
```

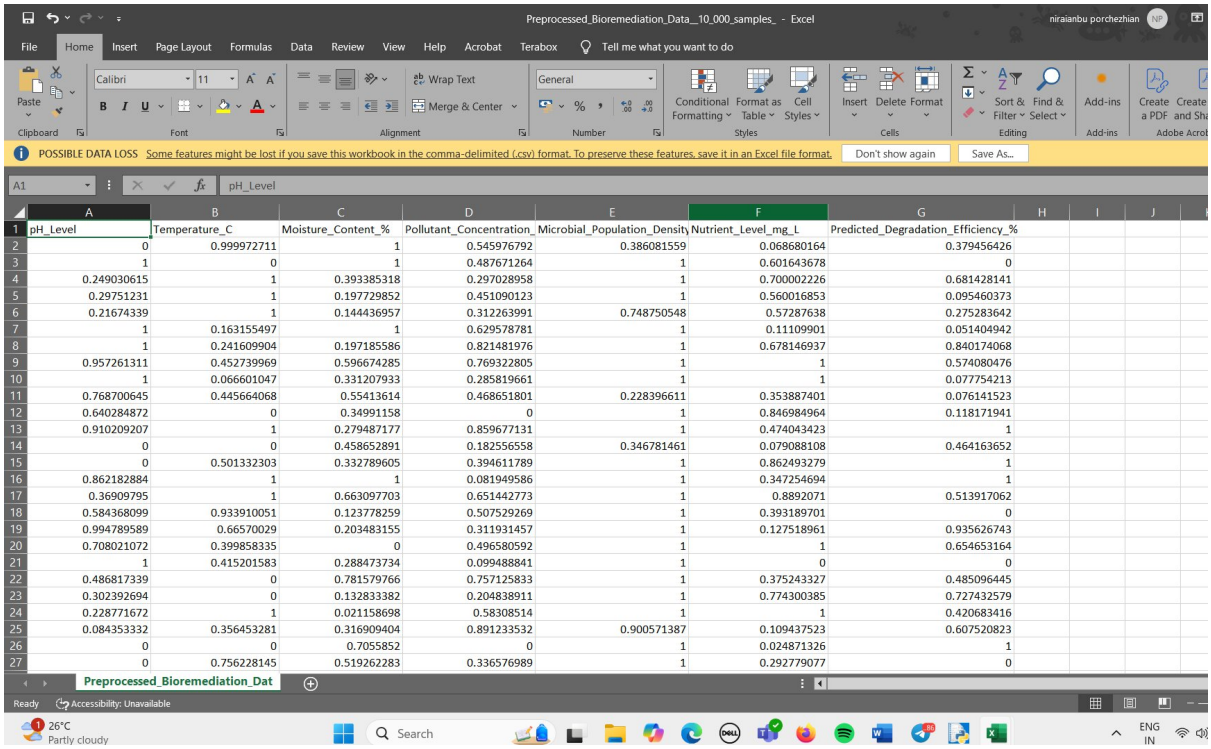
Output

Output Will have few rows and columns only as per standard statement display. It has generated 10,000 samples with 7 variables after preprocessing.



Preprocessed Bioremediation Data (10,000 Samples)

	pH_Level	Temperature_C	Moisture_Content_%	Pollutant_Concentration_mg_L	Microbial_Population_Density_CFU_mL	Nutrient_Level_mg_L	Predicted_Degradation_Efficiency_%
1	0.0	0.9999727110199035	1.0	0.5459767919581643	0.38608155912419134	0.06868016374765354	0.3794564262231618
2	1.0	0.0	1.0	0.4876712637249444	1.0000000000000002	0.6016436777704328	0.0
3	0.2490306151631707	1.0	0.3933853178961428	0.2970289582524824	1.0000000000000002	0.7000022256149948	0.6814281412186319
4	0.29751231009821644	1.0	0.19772985202372761	0.45109012308915514	1.0000000000000002	0.5600168525053472	0.09546037326181918
5	0.2167433896884119	1.0	0.14443695669879292	0.31226399104861113	0.7487505477536154	0.5728763802358047	0.2752836415900105
6	1.0	0.16315549657554868	1.0	0.6295787808448713	1.0000000000000002	0.11109901003329598	0.05140494214025011
7	1.0	0.24160990437299978	0.1971855855828808	0.821481976230446	1.0000000000000002	0.6781469371520256	0.8401740678641778
8	0.9572613113020496	0.45273996856947124	0.5966742846255791	0.7693228050219721	1.0000000000000002	1.0	0.5740804760480899
9	1.0	0.06660104747853332	0.33120793299372875	0.28581966110868184	1.0000000000000002	1.0	0.07775421254521181
10	0.7687006447270728	0.4456640684584145	0.5541361399686455	0.468651801207089 44	0.22839661083727578	0.3538874012428512	0.07614152261590192
11	0.64028487175507	0.0	0.3499115800544924	0.0	1.0000000000000002	0.84698496438507 2	0.11817194131430939
12	0.9102092067067211	1.0	0.2794871769773729	0.8596771311700702	1.0000000000000002	0.47404342319296816	1.0
13	0.0	0.0	0.4586528905944563	0.18255655804273688	0.3467814614569173	0.07908810753077031	0.46416365216052624
14	0.0	0.5013323025704239	0.33278960451469763	0.3946117886612915	1.0000000000000002	0.8624932791572498	1.0
15	0.862182884006657	1.0	1.0	0.08194958591933948	1.0000000000000002	0.3472546941254946	1.0
16	0.3690979504218812	1.0	0.6630977026865632	0.6514427733947031	1.0000000000000002	0.8892070998995913	0.5139170616575636
17	0.5843680990275846	0.9339100512250809	0.12377825894148016	0.5075292692774913	1.0000000000000002	0.39318970076400844	0.0
18	0.994789589123874	0.665700290341535	0.2034831550840251	0.311931457397654 43	1.0000000000000002	0.1275189608929299	0.9356267430536264
19	0.7080210717697075	0.39985833496107853	0.0	0.49658059214513034	1.0000000000000002	1.0	0.654653163836569
20	1.0	0.41520158306071014	0.2884737335118445	0.09948884068348926	1.0000000000000002	0.0	0.0
21	0.486817339038901	0.0	0.781579766054407	0.7571258325880516	1.0000000000000002	0.37524332703544383	0.48509644542584895
22	0.3023926941601627	0.0	0.13283338210030704	0.20483891074394064	1.0000000000000002	0.7743003846785905	0.7274325787131386
23	0.22877167243741647	1.0	0.0211586980561594	0.583085140229205	1.0000000000000002	1.0	0.42068341593719993



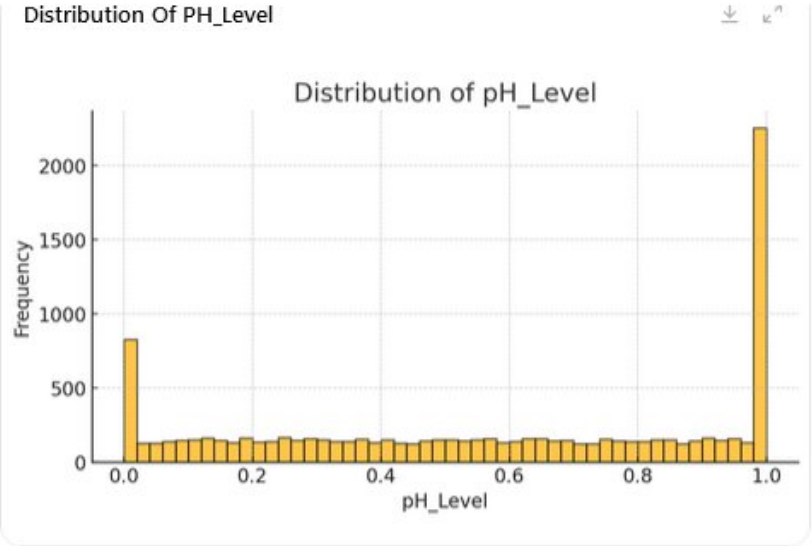
Preprocessed_Bioremediation_Data_10_000_samples_ - Excel

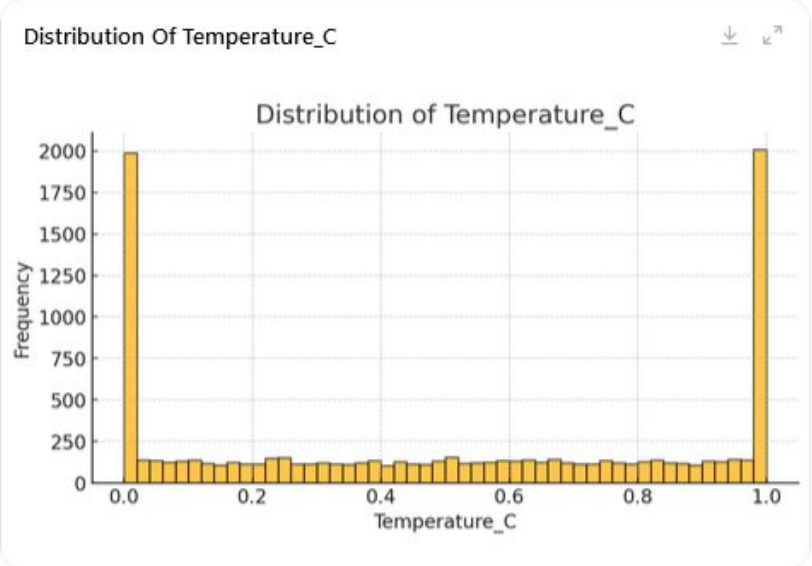
	A	B	C	D	E	F	G
1	pH_Level	Temperature_C	Moisture_Content_%	Pollutant_Concentration_	Microbial_Population_Density	Nutrient_Level_mg_L	Predicted_Degradation_Efficiency_%
2	0	0.999972711	1	0.545976792	0.386081559	0.068680164	0.379456426
3	1	0	1	0.487671264	1	0.601643678	0
4	0.249030615	1	0.393385318	0.297028958	1	0.700002226	0.681428141
5	0.29751231	1	0.197729852	0.451090123	1	0.560016853	0.095460373
6	0.21674339	1	0.144436957	0.312263991	0.748750548	0.57287638	0.275283642
7	1	0.163155497	1	0.629578781	1	0.11109901	0.051404942
8	1	0.241609904	0.197185586	0.821481976	1	0.678146937	0.840174068
9	0.957261311	0.452739969	0.596674285	0.769322805	1	1	0.574080476
10	1	0.066601047	0.331207933	0.285819661	1	1	0.077754213
11	0.768700645	0.445664068	0.55413614	0.468651801	0.228396611	0.353887401	0.076141523
12	0.640284872	0	0.34991158	0	1	0.846984964	0.118171941
13	0.910209207	1	0.279487177	0.859677131	1	0.474043423	1
14	0	0	0.458652891	0.182556558	0.346781461	0.079088108	0.464163652
15	0	0.501332303	0.332789605	0.394611789	1	0.862493279	1
16	0.862182884	1	1	0.081949586	1	0.347254694	1
17	0.36909795	1	0.663097703	0.651442773	1	0.8892071	0.513917062
18	0.584368099	0.933910051	0.123778259	0.507529269	1	0.393189701	0
19	0.994789589	0.66570029	0.203483155	0.311931457	1	0.127518961	0.935626743
20	0.708021072	0.399858335	0	0.496580592	1	1	0.654653164
21	1	0.415201583	0.288473734	0.099488841	1	0	0
22	0.486817339	0	0.781579766	0.757125833	1	0.375243327	0.485096445
23	0.302392694	0	0.132833382	0.204838911	1	0.774300385	0.727432579
24	0.228771672	1	0.021158698	0.58308514	1	1	0.420683416
25	0.084353332	0.356453281	0.316909404	0.891233532	0.900571387	0.109437523	0.607520823
26	0	0	0.7055852	0	1	0.024871326	1
27	0	0.756228145	0.519262283	0.336576989	1	0.292779077	0

Plot Data distribution for each variable

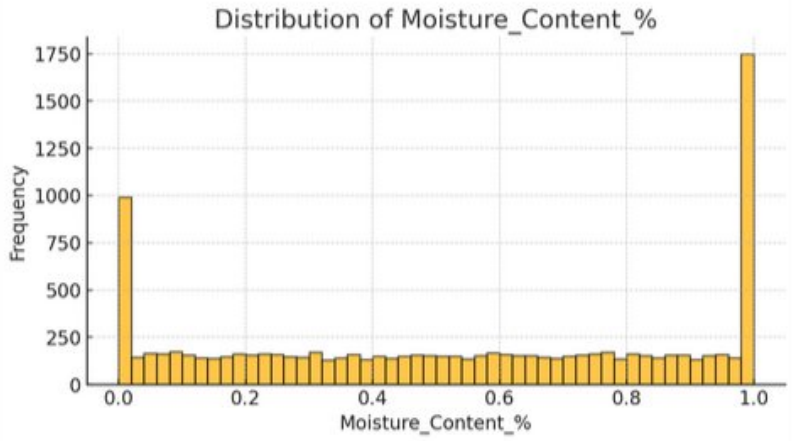
```
import matplotlib.pyplot as plt

# Plotting the data distribution for each variable in the preprocessed dataset

variables = processed_df.columns


# Creating a histogram for each variable

for var in variables:
    plt.figure(figsize=(8, 4))
    plt.hist(processed_df[var], bins=50, alpha=0.7, edgecolor='black')
    plt.title(f'Distribution of {var}')
    plt.xlabel(var)
    plt.ylabel('Frequency')
    plt.grid(axis='y', linestyle='--', alpha=0.7)
    plt.show()
```

Output

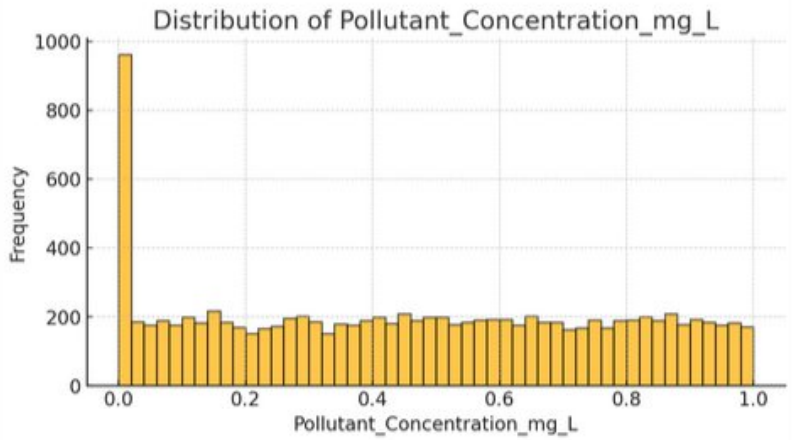

Distribution Of PH_Level
Distribution of pH_Level
Frequency
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
pH_Level


Distribution Of Temperature_C
Distribution of Temperature_C
Frequency
2000
1750
1500
1250
1000
750
500
250
0
0.0
0.2
0.4
0.6
0.8
1.0
Temperature_C

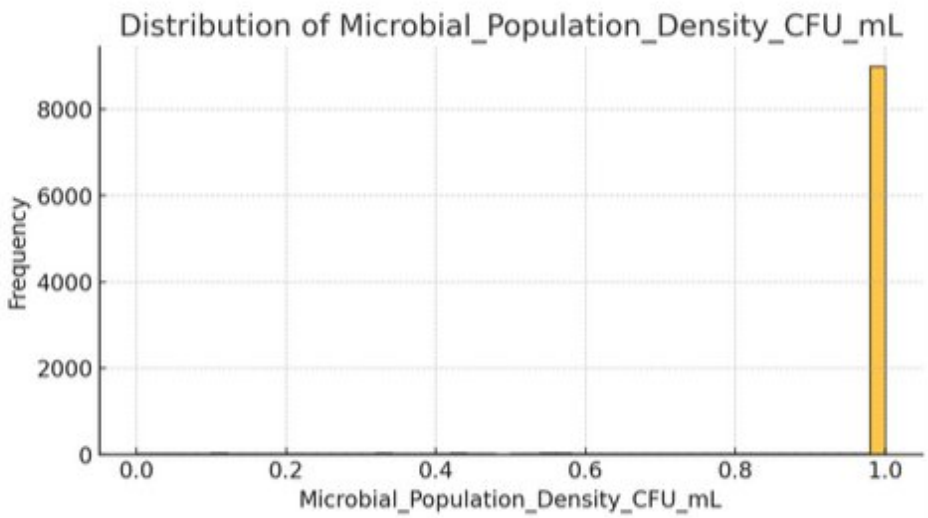
Distribution Of Moisture_Content_%



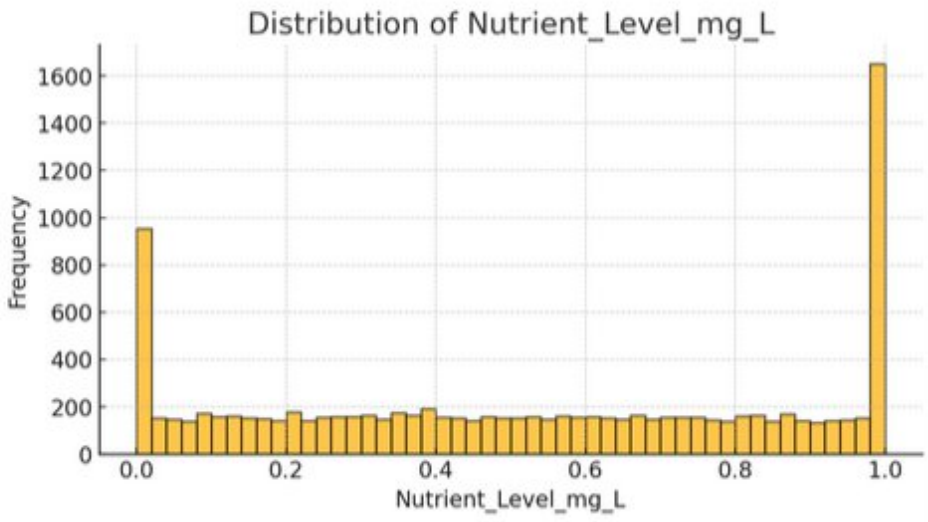
Distribution Of Pollutant_Concentration_mg_L

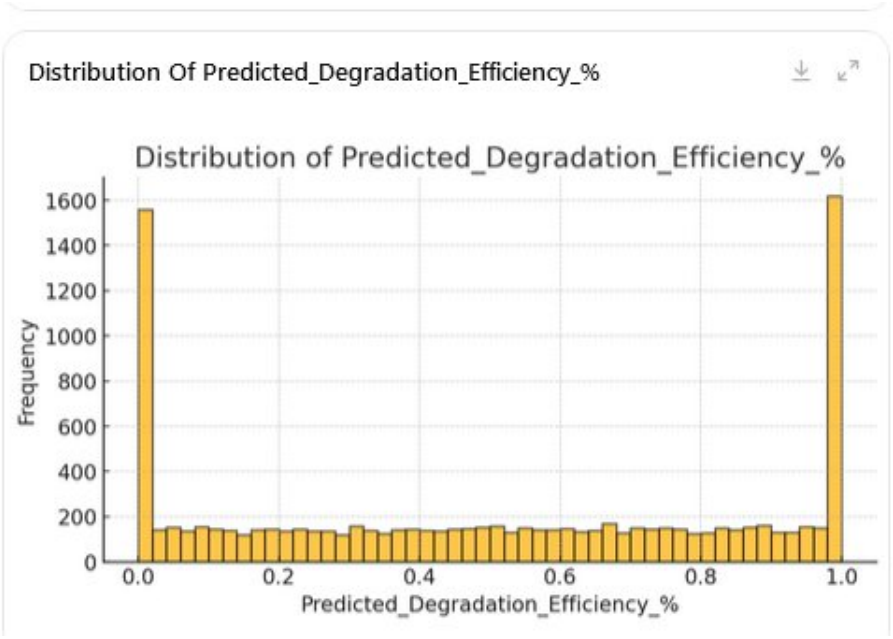


Distribution Of Microbial_Population_Density_CFU_mL



Distribution Of Nutrient_Level_mg_L





5. Conclusion

The AI-driven bioremediation project has the potential to significantly reduce organic pollutant levels in contaminated soil and water, contributing to environmental sustainability and public health. The integration of AI, IoT, and microbial technologies ensures a scalable and adaptive solution that can address various pollution challenges. This project serves as a model for leveraging advanced technologies to solve complex environmental problems and promote sustainable practices.

6. References

1. **"Microbial Degradation of Organic Pollutants: A Review of Techniques and Advances in Environmental Remediation"**
   Journal of Environmental Management, Volume 315, 2023, Article 115235.
   This comprehensive review discusses various microbial degradation techniques and recent advancements in environmental remediation.
2. **"Machine Learning Integration in Bioremediation: Optimizing Microbial Degradation of Pollutants"**
   Biotechnology Advances, Volume 60, 2022, Article 108076.
   This paper explores the integration of machine learning approaches to enhance microbial degradation processes in bioremediation efforts.
3. **"Advances in Genetic Engineering of Microbes for Pollutant Degradation"**
   Current Opinion in Biotechnology, Volume 72, 2021, Pages 145-152.
   This article reviews recent progress in genetically engineering microorganisms to improve their efficiency in degrading environmental pollutants.

Github link : https://github.com/Niraianbu/project