

## Data Collection and Preprocessing Phase

Date	June 2024
Team ID	739964
Project Title	EcoForecast: AI-powered prediction of carbon monoxide levels
Maximum Marks	6 Marks

### Preparation Template

The images will be preprocessed by resizing, normalizing, augmenting, denoising, adjusting contrast, detecting edges, converting color space, cropping, batch normalizing, and whitening data. These steps will enhance data quality, promote model generalization, and improve convergence during neural network training, ensuring robust and efficient performance across various computer vision tasks.

Section	Description
Data Overview	There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc. In this project we have used .csv data.
Data Preparation	These are the general steps of pre-processing the data before using it for machine learning
Handling missing values	We use Handling missing values For checking the null values
Handling categorical data	As we can see our dataset has categorical data we must convert the categorical data to integer encoding or binary encoding
Handling Outliers in Data	With the help of boxplot, outliers are visualized. And here we are going to find upper bound and lower bound of numerical features with some mathematical formula.
<b>Data Preparation</b>	

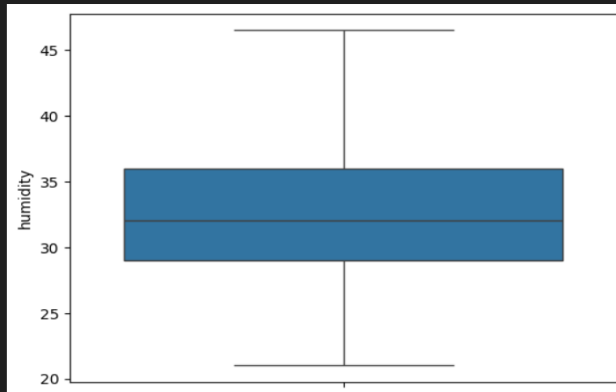
Collect the dataset	Please refer to the link given below to download the dataset. <a href="#">data set link</a>																																																																																				
Importing the libraries	<pre>• #importing the libraries required import pandas as pd import numpy as np import matplotlib.pyplot as plt import seaborn as sns  #importing warnings import warnings warnings.filterwarnings('ignore')</pre>																																																																																				
Loading Data	We use the code Data =pd.read_csv('arduino_data.csv') For reading the dataset																																																																																				
Handling missing values	<pre>#checking the null values data.isnull().sum()  ✓ 0.0s  timestamp    0 temp         0 humidity     0 ppm          0 dtype: int64</pre>																																																																																				
Handling Timestamp values	<pre>Click to add a breakpoint pd.to_datetime(data['timestamp'])  # Extracting year, month, and day data['Year'] = data['timestamp'].dt.year data['Month'] = data['timestamp'].dt.month data['Day'] = data['timestamp'].dt.day data['Hour'] = data['timestamp'].dt.hour  # Display the modified DataFrame print(data[ ['timestamp', 'Year', 'Month', 'Day', 'Hour', 'temp']])</pre> <table><thead><tr><th></th><th>timestamp</th><th>Year</th><th>Month</th><th>Day</th><th>Hour</th><th>temp</th></tr></thead><tbody><tr><td>0</td><td>2023-06-09 10:46:48+05:30</td><td>2023</td><td>6</td><td>9</td><td>10</td><td>38</td></tr><tr><td>1</td><td>2023-06-09 10:47:49+05:30</td><td>2023</td><td>6</td><td>9</td><td>10</td><td>38</td></tr><tr><td>2</td><td>2023-06-09 10:48:49+05:30</td><td>2023</td><td>6</td><td>9</td><td>10</td><td>38</td></tr><tr><td>3</td><td>2023-06-09 10:49:50+05:30</td><td>2023</td><td>6</td><td>9</td><td>10</td><td>38</td></tr><tr><td>4</td><td>2023-06-09 10:50:50+05:30</td><td>2023</td><td>6</td><td>9</td><td>10</td><td>38</td></tr><tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr><tr><td>10303</td><td>2023-06-17 02:51:13+05:30</td><td>2023</td><td>6</td><td>17</td><td>2</td><td>28</td></tr><tr><td>10304</td><td>2023-06-17 02:52:13+05:30</td><td>2023</td><td>6</td><td>17</td><td>2</td><td>28</td></tr><tr><td>10305</td><td>2023-06-17 02:53:14+05:30</td><td>2023</td><td>6</td><td>17</td><td>2</td><td>28</td></tr><tr><td>10306</td><td>2023-06-17 02:54:14+05:30</td><td>2023</td><td>6</td><td>17</td><td>2</td><td>28</td></tr><tr><td>10307</td><td>2023-06-17 02:55:15+05:30</td><td>2023</td><td>6</td><td>17</td><td>2</td><td>28</td></tr></tbody></table>		timestamp	Year	Month	Day	Hour	temp	0	2023-06-09 10:46:48+05:30	2023	6	9	10	38	1	2023-06-09 10:47:49+05:30	2023	6	9	10	38	2	2023-06-09 10:48:49+05:30	2023	6	9	10	38	3	2023-06-09 10:49:50+05:30	2023	6	9	10	38	4	2023-06-09 10:50:50+05:30	2023	6	9	10	38	...	...	...	...	...	...	...	10303	2023-06-17 02:51:13+05:30	2023	6	17	2	28	10304	2023-06-17 02:52:13+05:30	2023	6	17	2	28	10305	2023-06-17 02:53:14+05:30	2023	6	17	2	28	10306	2023-06-17 02:54:14+05:30	2023	6	17	2	28	10307	2023-06-17 02:55:15+05:30	2023	6	17	2	28
	timestamp	Year	Month	Day	Hour	temp																																																																															
0	2023-06-09 10:46:48+05:30	2023	6	9	10	38																																																																															
1	2023-06-09 10:47:49+05:30	2023	6	9	10	38																																																																															
2	2023-06-09 10:48:49+05:30	2023	6	9	10	38																																																																															
3	2023-06-09 10:49:50+05:30	2023	6	9	10	38																																																																															
4	2023-06-09 10:50:50+05:30	2023	6	9	10	38																																																																															
...	...	...	...	...	...	...																																																																															
10303	2023-06-17 02:51:13+05:30	2023	6	17	2	28																																																																															
10304	2023-06-17 02:52:13+05:30	2023	6	17	2	28																																																																															
10305	2023-06-17 02:53:14+05:30	2023	6	17	2	28																																																																															
10306	2023-06-17 02:54:14+05:30	2023	6	17	2	28																																																																															
10307	2023-06-17 02:55:15+05:30	2023	6	17	2	28																																																																															

## Handling Outliers

```
data['humidity']=np.where(data['humidity']>46.5,46.5, data ['humidity'])
sns.boxplot(data['humidity'])
```

✓ 0.1s

<Axes: ylabel='humidity'>



```
data['ppm'] = np.where(data['ppm'] > 78.975,78.975, data['ppm'])
sns.boxplot(data['ppm'])
```

<Axes: ylabel='ppm'>

