UNVEILING THE FUTURE OF CARBON EMISSIONS TRADING: A MACHINE LEARNING AND NEURAL NETWORK PERSPECTIVE ON REGIONAL MARKETS.

Sindhu Nagesha - 017419987
Prayag Nikul Purani - 017416737
Syed Faraaz Ahmed - 017428619
Sai Vivek Chunduri- 017435301

DATA 270 Group 7

OVERVIEW

- ABSTRACT
- INTRODUCTION
- PROJECT BACKGROUND
- PROJECT REQUIREMENTS
- LITERATURE AND TECHNOLOGY SURVEY
- PROJECT MANAGEMENT PLAN & RESOURCE REQUIREMENTS
- DATA ENGINEERING
- MODEL DEPLOYMENT
- APPLICATION RESULTS

ABSTRACT

ABSTRACT

- Problem Statement: Investigate the dynamics of carbon emissions and trade across the globe to enhance climate mitigation strategies.
- Project Objective: Develop an integrated platform to forecast carbon credit needs by incorporating diverse sectoral and weather data.
- Algorithmic Approach: Employ CRISP-DM methodology, Implement various machine learning models, and ultimately select Hybrid CNN-ANN, Hybrid CNN-IBFA, SVR, and LSTM to analyze unified data, resulting in high predictive accuracy.
- Evaluation Metrics and Model Efficiency: Utilize R-squared values to assess model performance, with CNN-ANN, LSTM, CNN-IBFA, and SVR achieving 0.9997, 0.892, 0.9338 and 0.875 respectively.
- Application and Future Scope: Provide insights for policy formulation and suggest enhancing data granularity to improve predictive accuracy and support future research.

INTRODUCTION

PROJECT BACKGROUND

MOTIVATION

The motivation for this project stems from the pressing need to address carbon emissions' impact on the environment and the economy. By consolidating and analyzing data from diverse sources, including regional carbon emission markets and weather patterns, the project seeks to provide insights crucial for informed policy-making. Utilizing cutting-edge AI techniques such as machine learning and neural networks, the project aims to anticipate emissions trends and optimize carbon credit transactions. Ultimately, the project's goal is to contribute to a more sustainable future by leveraging technology to understand and mitigate the effects of carbon emissions.

NEEDS AND IMPORTANCE

Since the 1950s, the world has experienced significant modernization in what's often referred to as the "atomic age of science and technology," leading to new jobs, laws, and experiments. However, innovations now carry a significant stigma due to carbon emissions. The 2019 pandemic caused a dramatic 5.4% drop in carbon emissions in 2020, significantly impacting the carbon credit market. Post-pandemic lifestyles have a profound effect on carbon emissions, impacting the environment in various ways. Carbon dioxide gas emitted from vehicles, industries, and other sources is a major contributor.

TARGETED PROBLEM

Modernization brings a carbon emissions stigma, necessitating innovative carbon credit management. Pandemic-induced emission drops highlight the need for resilient carbon credit strategies. Paris Agreement targets demand substantial emissions reductions for climate resilience. Machine learning's carbon footprint raises sustainability concerns in model training. Integration of diverse data sources and advanced modeling improves carbon credit prediction accuracy.

DELIVERABLES

The project aims to consolidate data on carbon emissions and trade in global markets, integrating weather data to understand their impact on both weather and the economy. Utilizing Machine Learning and Neural Network techniques, it seeks to anticipate emissions and credit transactions, ultimately offering insights for policy through analysis of refined models.

PROJECT REQUIREMENTS

Functional Requirements:

- Utilize Python for coding, with TensorFlow and PyTorch for machine learning models.
- Employ AWS for cloud computing resources.
- Use command-line interfaces for script execution and data pipeline management.
- Develop APIs for integration with external systems. Ensure compliance with third-party licensing, preferring open-source options.
- Maintain adherence to budget and data privacy regulations, aiming for timely project milestones.

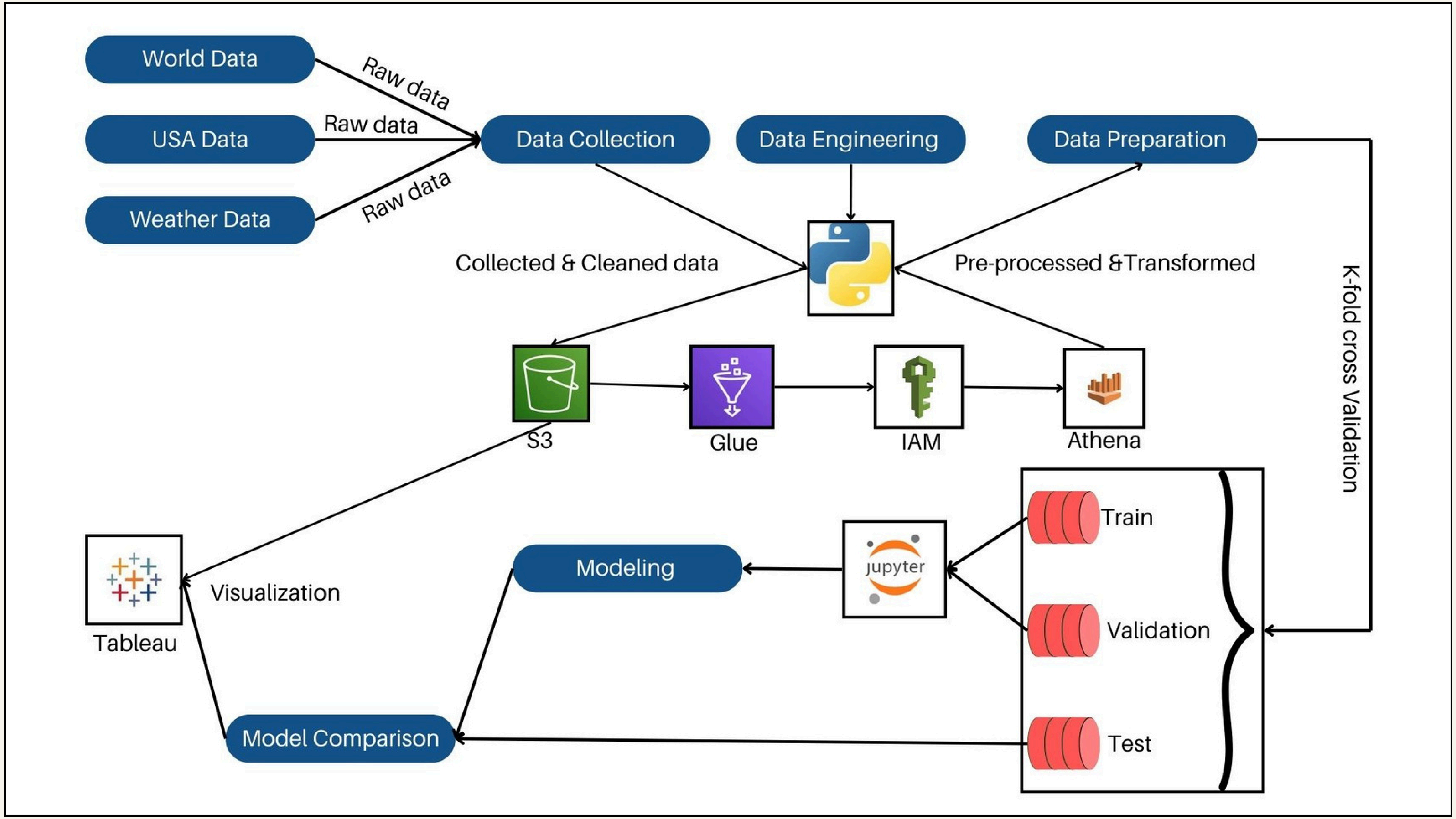
AI Requirements:

- Apply machine learning and deep learning models, including CNNs, for complex data analysis.
- Use evaluation metrics like accuracy, precision, recall, F1-score, and MSE.
- Focus on models that are computationally efficient, scalable, and interpretable.
- Select models that best meet project goals while considering resource constraints.

Data Requirements:

- Carbon data includes emissions per energy source, per capita, and economic output, sourced from the U.S. EIA.
- Weather data from The National Weather Service API covers temperature, wind speed, and direction, relevant to carbon emissions' environmental impact.
- Global carbon emissions data from the World Bank features details like country, year, population, GDP, and emissions from various sources.

PROJECT WORKFLOW



LITERATURE SURVEY

Authors	Title	Models	Result
Lu et al. (2020)	Carbon trading volume and price forecasting in China using multiple machine learning models	CEEMDAN-RBFNN CEMDAN-GWO-KNEA	Accuracy - 98.40% & 97.89%
Mao & Yu (2024)	A hybrid forecasting approach for china's national carbon emission allowance prices with balanced accuracy and interpretability	FS-CEEMDAN-VMD-GWO-LightGBM	MAE:69.11%, RMSE:69.24%, MAPE:3.60 %, and R2:8.75%.
Mardani et al. (2020)	A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques.	ANFIS-ANN SVD-SOM-ANN:	MAE = 0.065& 0.104
Wang & Ye (2017)	Forecasting Chinese carbon emissions from fossil energy consumption using non-linear grey multivariable models	NGM(1, N)	Accuracy - 99.35%
AcheamPong & Boateng (2019)	Modelling carbon emission intensity: Application of artificial neural network.	ANN	R2 - 0.8721

TECHNOLOGY SURVEY

Authors	Title	Models	Result
Boateng et al. (2020)	Predicting building-related carbon emissions: A test of machine learning models.	Random forest, KNN, XGBoost, DT, AdaBoost, SVR	R-square -- 99.8%, 99.87%, 99.77%, 99.63%,99.56%, 97.67%
Li et al. (2022)	Forecasting carbon price in China: a multimodel comparison	LSTM, MLP, RNN	MAE:0.617, MSE:0.957, RMSE:0.978
Wu et al. (2015)	Modelling and forecasting CO2 emissions in the BRICS (Brazil, Russia, India, China, and South Africa) countries using a novel multi-variable grey model.	GOM (1, N)	MAPE 4.56
Zhu et al. (2023)	Prediction of Carbon Emission Right Price Based on XGBoost Algorithm.	XGBoost	RMSE:48.42, MAE:5.15, MAPE:0.02
Lei & Yang (2020)	Influencing factors analysis and forecasting of residential energy-related CO2 emissions utilizing optimized support vector machine.	ICSO-SVM, PSO-SVM, GA-SVM,	MAPE(%):2.37%, 2.77%, 3.76%, 4.17%

PROJECT MANAGEMENT PLAN
AND
RESOURCE REQUIREMENTS

METHODOLOGY

CRISP-DM methodology is used for data mining projects, comprising six key phases.

1. Business Understanding involves translating business goals into data mining problems

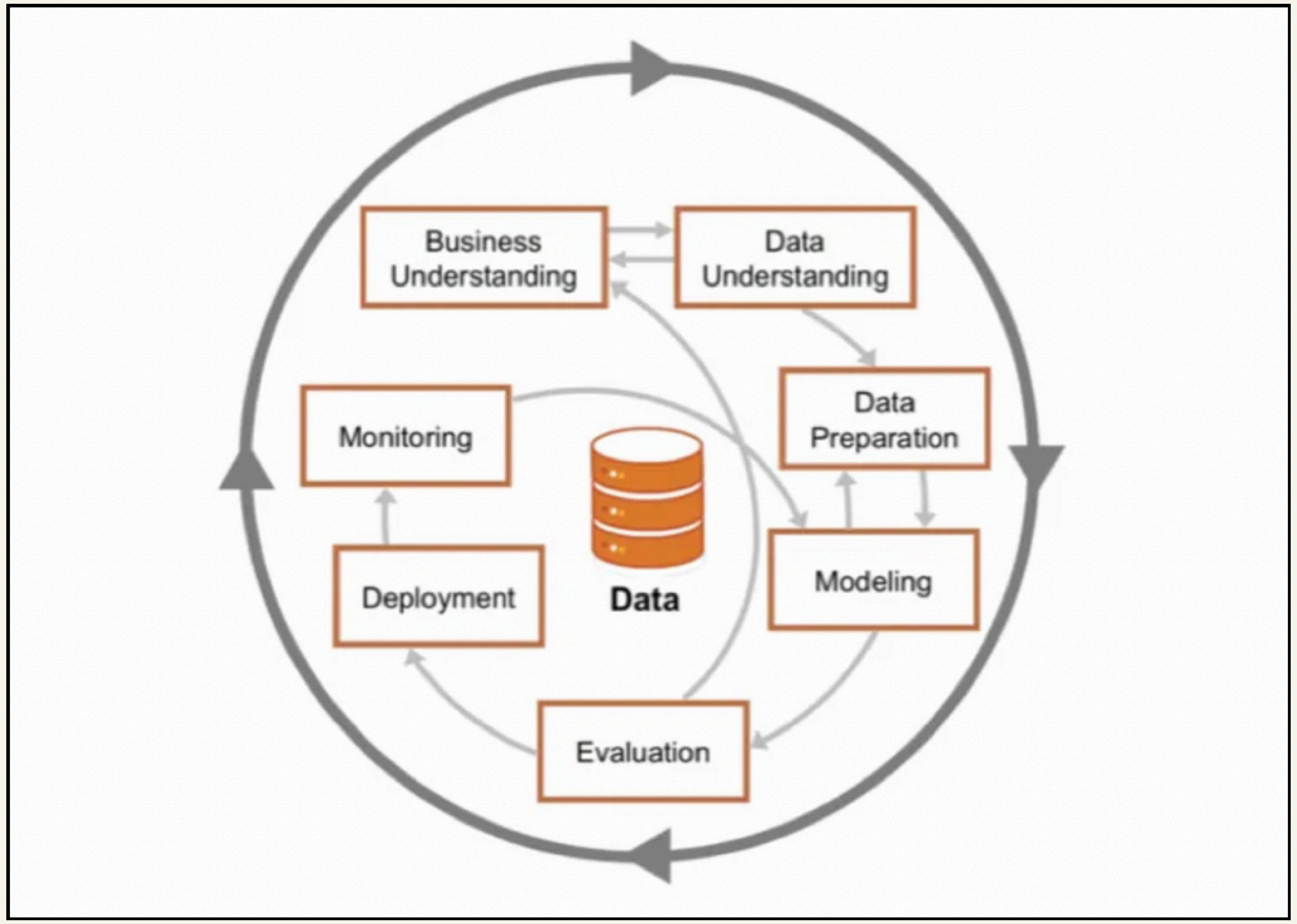
2. Data Understanding focuses on exploring and evaluating data quality and characteristics

3. Data Preparation ensures data is clean and suitable for modeling

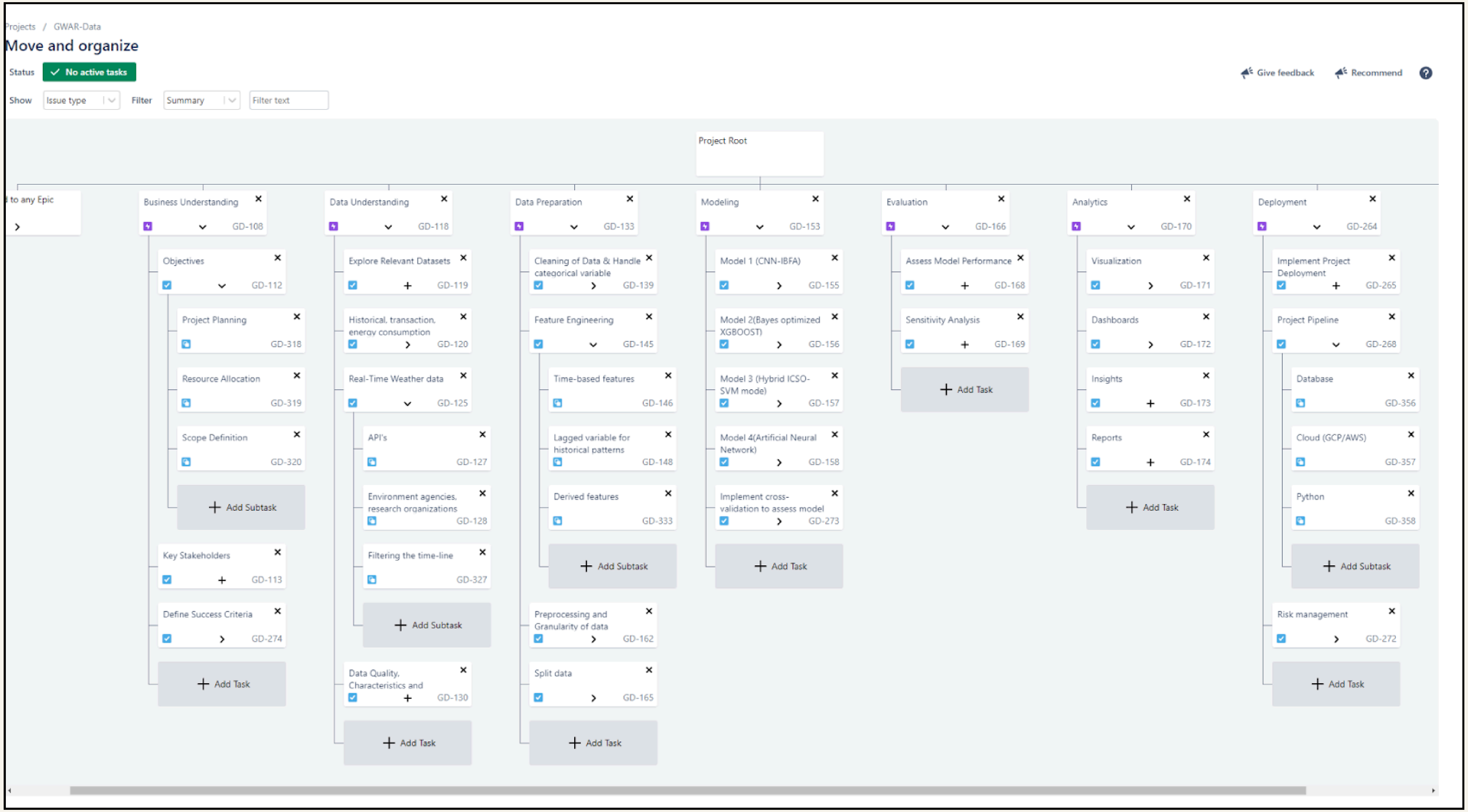
4. Modelling includes selecting and building data mining models

5. Evaluation assesses model performance

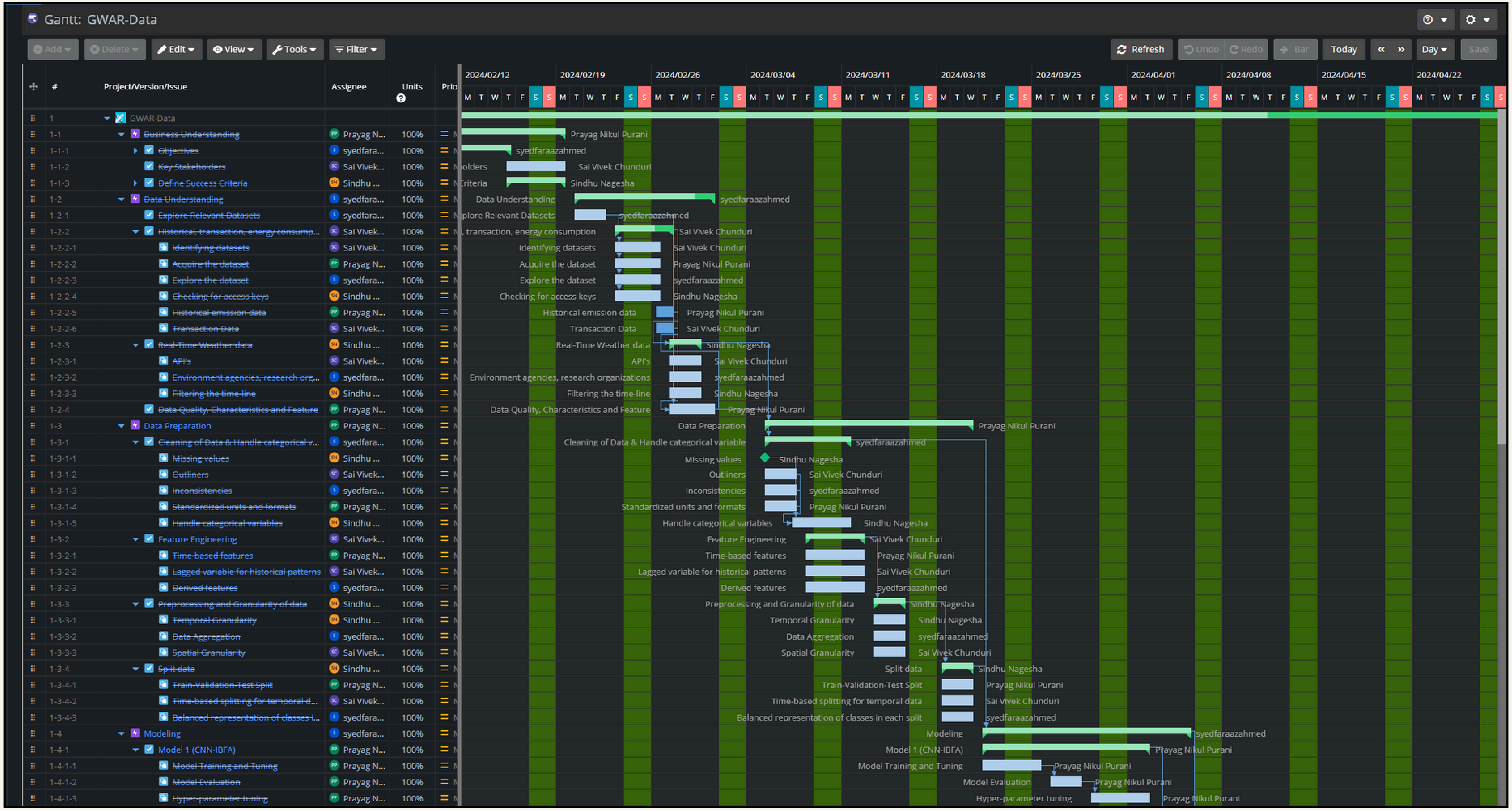
6. Deployment integrates it into the business environment



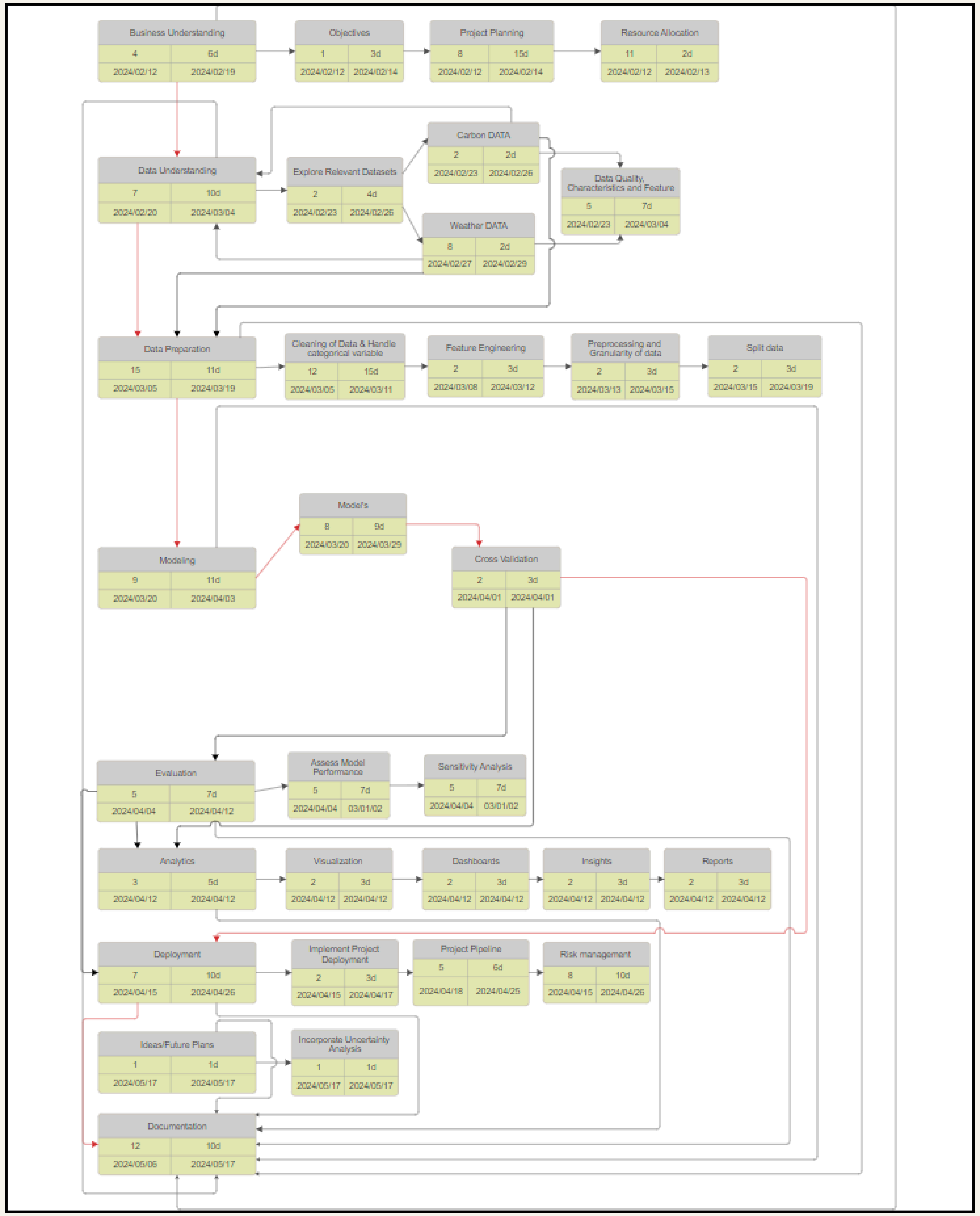
PROJECT ORGANIZATION PLAN (WBS)



PROJECT ORGANIZATION PLAN (GANTT CHART)



PROJECT ORGANIZATION PLAN (PERT CHART)



PERT CHART contains in total of 32 tasks.

Critical Path: The red arrowed line

"4 → 7 → 15 → 9 → 8 → 2 → 12",

takes upto maximum of 72 days.

PROJECT RESOURCE REQUIREMENT

- **HARDWARE REQUIREMENT**

Specification	Cost	Duration	Justification
Laptop	$2,800	Feb 5 - May 20	As we are a team of four and to work parallel we will be requiring the 4 laptops.
GPU	$2,000	Feb 5 - May 20	We will be performing queries that require parallel processing and GPUs will have a great impact on the performance of input queries.
5888 CUDA Core GPU	$600	Feb 5 - May 20	16 GB dedicated memory for ML and DL models

- **SOFTWARE REQUIREMENT**

Specification	Cost	Duration	Justification
Pytorch	Free	Feb 5 - May 20	It offers an adaptable framework for deep learning model construction and training.
TensorFlow & Keras	Free	Feb 5 - May 20	Support for implementing complex machine learning algorithms and models, essential for the project's objectives.
NLTK	Free	Feb 5 - May 20	Packages for parsing, tokenization, categorization, stemming, tagging, and semantic reasoning are included with NLTK
Pandas	Free	Feb 5 - May 20	Time-series data, it is a vital tool since it offers simple-to-use data structures and methods for handling structured data.
Numpy	Free	Feb 5 - May 20	A core module for scientific computing in Python matrices
spaCY	Free	Feb 5 - May 20	It provides a range of tools and pre-trained algorithms for handling and interpreting text data.
OpenCV	Free	Feb 5 - May 20	OpenCV, short for Open Source Computer Vision Library, is a software library utilized for computer vision and machine learning tasks.
Sci-kit learn	Free	Feb 5 - May 20	A potent open-source machine learning package for Python is called sci-kit-learn.
AWS (S3, Glue, Athena, R	Free	Feb 5 - May 20	AWS services will be used for ETL processes and storing data in S3 buckets.

- **TOOLS & LICENSE REQUIREMENT**

Specification	Cost	Duration	Justification
Jupyter Notebook	Free	Feb 5 - May 20	Jupyter Notebook provides an interactive environment for conducting data exploration, analysis, and model prototyping.
Power BI	$85/month	Feb 5 - May 20	These tools facilitate the communication of insights derived from data analysis, enhancing decision-making processes.
Google Colab	Free	Feb 5 - May 20	This will need to share code work with teammates
JIRA	Free	Feb 5 - May 20	This tool is a project management tool that will help to make a roadmap for the project.
Github	Free	Feb 5 - May 20	Version control tool to make the project publicly available and so can be used by others to learn and make improvements on the exciting model
Zoom	Free	Feb 5 - May 20	To connect with the team members and to plan for the task and the coordination management.
Grammary	30/month	Feb 5 - May 20	To make the reports free from all the grammatical errors and to check
Google Doc	Free	Feb 5 - May 20	To make documents simultaneously in a group.

DATA ENGINEERING

DATA COLLECTION

1. World data: This dataset originated from the World Bank website. It contains information on 245 countries, with approximately 80 different features and 50,000 records. The size of the data file is around 4 MB.

2. US Data: The U.S. Energy Information Administration (EIA) provides the time-series data, which comprises four datasets with dimensions of 59 rows and 57 columns each. Each file's size falls approximately between 50-70 kB.

3. Weather Data: This dataset acquired through an API from the National Weather Service, consists of 6,000 entries with 6 attributes. When stored in an Excel file format, the size of the dataset is around 6 MB.

RAW DATA

World Data

	country	year	iso_code	population	gdp	cement_co2	cement_co2_per_capita	co2	co2_growth_abs	co2_growth_prct	...	share_global_other_co2	sh
0	Afghanistan	1850	AFG	3752993.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
1	Afghanistan	1851	AFG	3767956.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
2	Afghanistan	1852	AFG	3783940.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
3	Afghanistan	1853	AFG	3800954.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
4	Afghanistan	1854	AFG	3818038.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	

5 rows × 79 columns

USA Data

t1

State	1970	1971	1972	1973	1974	1975	1976	1977	1978	...	2016	2017	2018	2019	2020	2021	Percent	Absolute	Percent.1	Absolute.1
Alabama	102.646851	98.461114	104.932504	109.563135	108.777543	107.779346	108.089155	111.683852	106.629516	...	113.986483	108.594556	112.355761	106.254617	98.431803	108.392103	0.055971	5.745252	0.101190	9.960300
Alaska	11.348910	12.636423	13.420588	12.490564	12.779110	14.524477	15.969357	17.950301	19.482875	...	33.405683	33.731095	34.515445	34.276830	35.977743	38.872526	2.425221	27.523616	0.080460	2.894783
Arizona	24.906189	26.998731	30.179241	34.448720	36.737072	38.221132	43.777153	50.500706	49.284884	...	90.860970	90.480865	94.099855	92.555687	80.153897	83.024267	2.333479	58.118078	0.035811	2.870370
Arkansas	36.178889	35.091287	37.189109	40.829866	39.112512	36.365524	38.857529	41.649053	42.418586	...	62.127003	64.176360	70.785991	65.073337	54.749619	62.024941	0.714396	25.846052	0.132884	7.275322
California	294.372200	305.833367	312.722261	329.285277	304.471018	311.485516	326.902403	354.482753	345.243293	...	353.372145	356.532043	358.605130	358.266355	303.815453	324.039053	0.100780	29.666853	0.066565	20.223600
Colorado	43.017559	43.591690	47.467615	51.067649	50.480507	51.801656	55.134964	58.314615	58.444211	...	88.428665	88.778171	90.066247	91.747156	79.911265	85.381781	0.984812	42.364222	0.068457	5.470516
Connecticut	47.831994	45.894756	47.215575	48.576991	45.425744	41.688093	43.432917	43.051435	43.969455	...	34.096250	33.860382	37.386598	36.635352	33.769097	36.573257	-0.235381	-11.258737	0.083039	2.804160
Delaware	16.072307	15.929520	16.013802	17.243519	16.657406	15.543109	16.157030	16.139793	16.183481	...	14.420754	13.768762	13.991987	13.593983	12.476870	12.966634	-0.193231	-3.105673	0.039254	0.489763

Weather Data

	country	year	iso_code	population	gdp	cement_co2	cement_co2_per_capita	co2	co2_growth_abs	co2_growth_prct	...	share_global_other_co2	sha
0	Afghanistan	1850	AFG	3752993.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
1	Afghanistan	1851	AFG	3767956.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
2	Afghanistan	1852	AFG	3783940.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
3	Afghanistan	1853	AFG	3800954.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
4	Afghanistan	1854	AFG	3818038.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	

5 rows × 79 columns

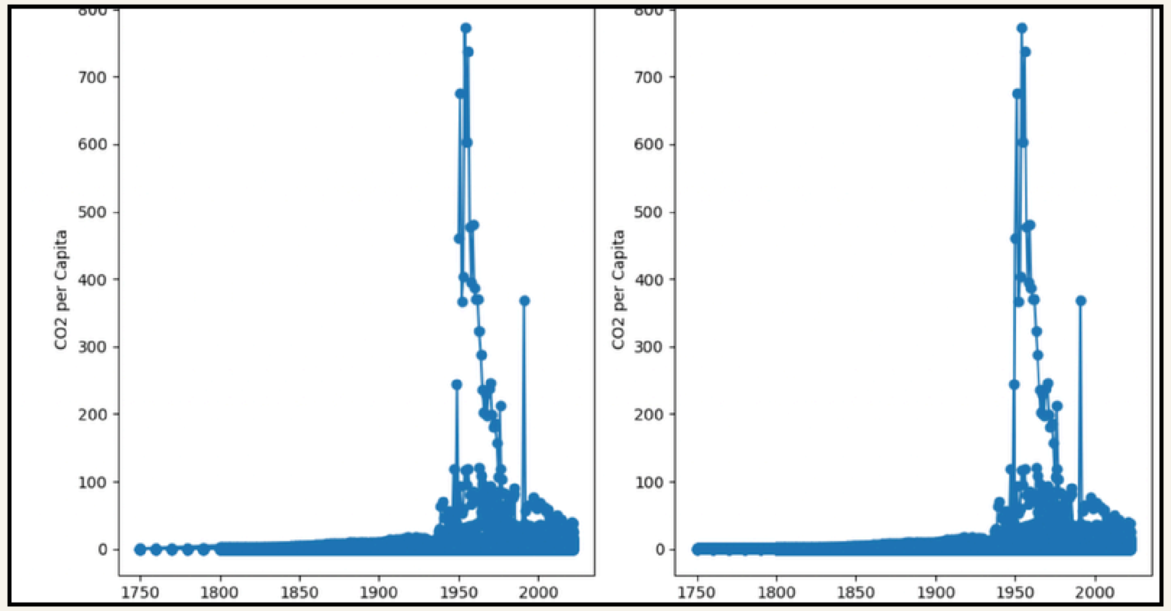
DATA PROCESS

	Raw	Pre-processing	Transformation	Preparation
World Data	World Bank Data 50k records & 76 features	Cleaning (handling null values, Type 3 noise, outlines, data scrubbing, and data auditing) 39717 records & 76 features	Data Aggregation, Data Regularization, Standardization, L2 & Dropout, Data Reduction 39717 records & 40 features	K-fold cross-validation method Train - 38447 records Test - 4805 records Validation - 4805 records
USA Data	4 total sheets, 3 sheets- 52 records & 57 features and other sheet of 52 records x 30 featrures.	Cleaning (Removal of unnecessary data, Duplicates & missing values, Transpose rows & columns, Rename the columns, Combine all 4 files) 182 records & 52 features	Data Smoothing & Data Normalization 182 records & 52 features	Train- 137 records & 52 features Test- 24 records & 52 features Validation- 20 records & 52 features
Weather Data	National Weather Service 6k records & 6 features	Data Cleaning (handling null values and timeframe selection) 3669 records & 4 features	Data discretization 3669 records & 4 features	N.A 3669 records & 4 features

DATA PRE-PROCESSING WORLD DATA

World Data → Cleaning

Timeframe (Noise Type 3)
Data scrubbing + Data auditing
Duplicates & missing values
Removal of unnecessary data



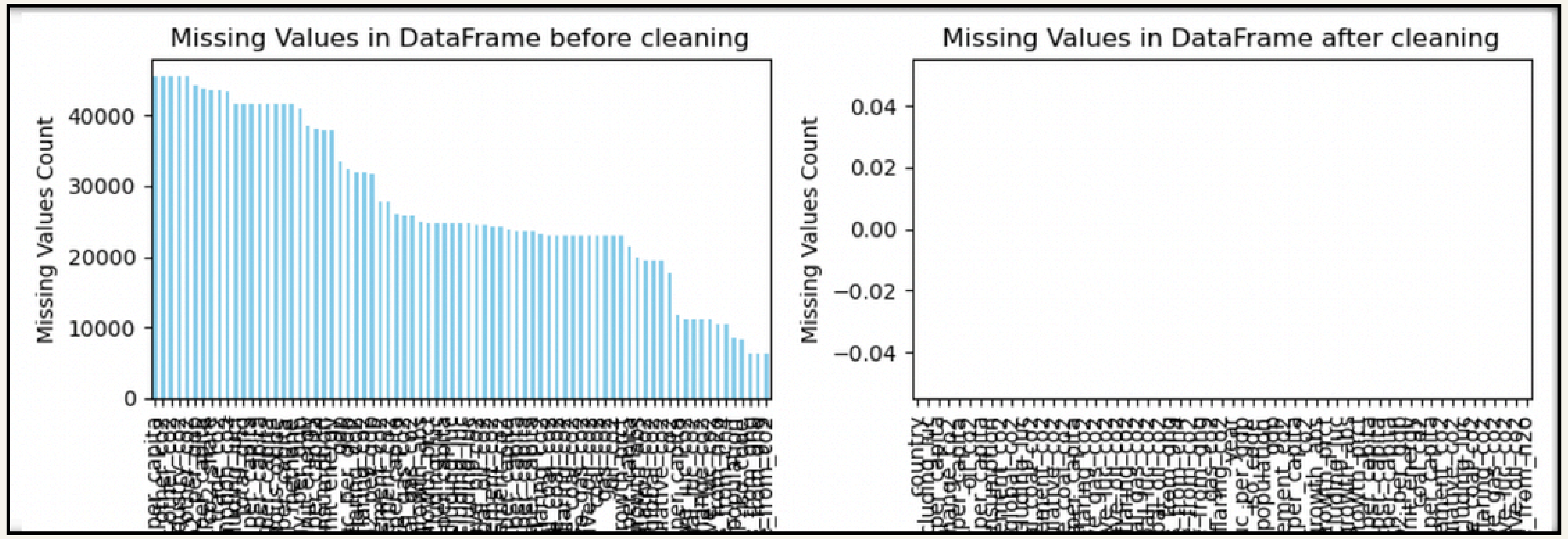
Duplicates & missing values

```
df_filtered = df[df['year'] >= 1970]
```

Timeframe (Noise Type 3)

```
# List of countries you want to select
#Data scrubbing & Data auditing
countries = ['Africa','Asia','Asia (excl. China and India)','Australia','Canada','China','Europe','France','Germany'
       'Iraq','Italy','Japan','Kuwait','New Zealand','North America','North Korea','Pakistan','Qatar','Russia','Saud
```

Data scrubbing + Data auditing



Removal of unnecessary data

```
---  ------                                          --------------  -----
 0   year                                            39717 non-null  int64
 1   population                                      39717 non-null  float64
 2   gdp                                             39717 non-null  float64
 3   cement_co2                                      39717 non-null  float64
 4   cement_co2_per_capita                           39717 non-null  float64
 5   co2                                             39717 non-null  float64
 6   co2_growth_abs                                  39717 non-null  float64
 7   co2_growth_prct                                 39717 non-null  float64
 8   co2_including_luc                               39717 non-null  float64
 9   co2_including_luc_growth_abs                    39717 non-null  float64
 10  co2_including_luc_growth_prct                   39717 non-null  float64
 11  co2_including_luc_per_capita                    39717 non-null  float64
 12  co2_including_luc_per_gdp                       39717 non-null  float64
 13  co2_including_luc_per_unit_energy               39717 non-null  float64
 14  co2_per_capita                                  39717 non-null  float64
 15  co2_per_gdp                                     39717 non-null  float64
```

Final clean data

DATA PRE-PROCESSING USA DATA

USA Data → **Cleaning**

4 files

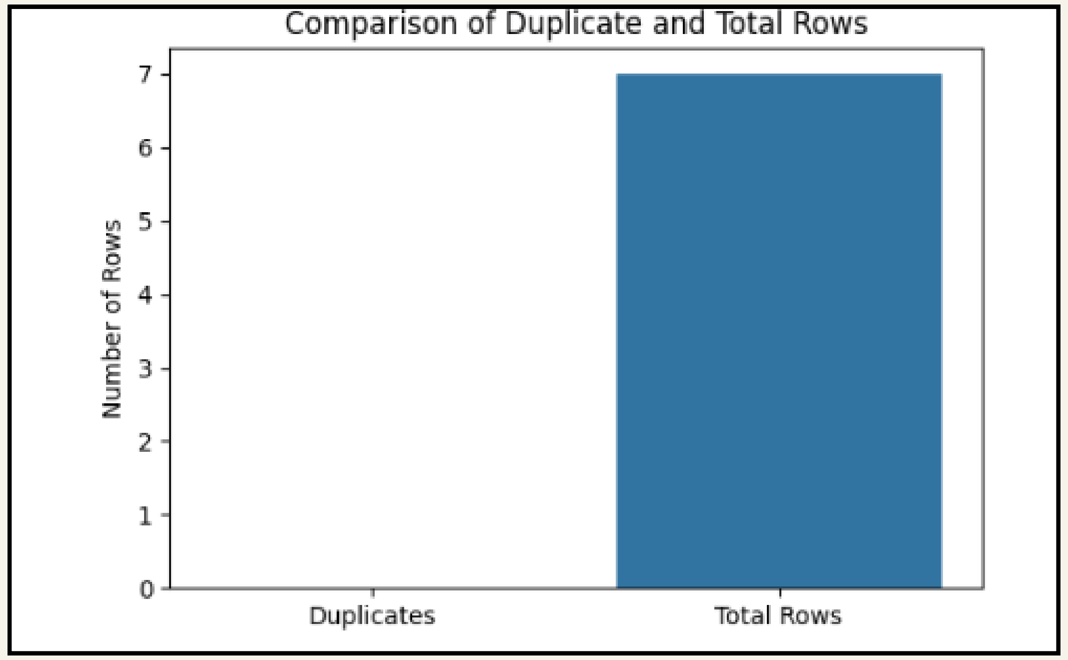
Removal of unnecessary data
Duplicates & missing values
Transpose rows & column
Rename the columns
Combine all 4 files

```
Number of duplicate rows for Table1: 0
Empty DataFrame
Columns: [State, CO2_Emissions_1970, CO2_Emissions_1971, CO2_Emissions_1972, CO2_Emissions_1973, CO2_Emissions_1974, CO2_Emissions_1975, CO2_Emissions_1976,
Index: []

[0 rows x 57 columns]
```

checking for duplicates

Year	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	District of Columbia
PerCapita_CO2Emissions_1970	29.744089	37.33194	13.898543	18.726133	14.71346	19.351129	15.729034	29.169341	18.044736
PerCapita_CO2Emissions_1971	28.155881	39.98868	14.239837	17.79477	15.031621	18.920004	14.993386	28.19384	15.836108
PerCapita_CO2Emissions_1972	29.650326	41.421568	15.029503	18.419589	15.191754	19.737054	15.384677	27.947299	15.236913
PerCapita_CO2Emissions_1973	30.604228	37.735844	16.218795	19.82995	15.77868	20.459795	15.833439	29.833077	16.210966
PerCapita_CO2Emissions_1974	29.999322	37.475395	16.525898	18.616141	14.379476	19.866394	14.777405	28.670234	13.717723
PerCapita_CO2Emissions_1975	29.295827	38.628928	16.726973	16.835891	14.462137	20.031576	13.521924	26.478891	11.159409
PerCapita_CO2Emissions_1976	28.939533	39.823833	18.660338	17.906695	14.902553	20.947935	14.087874	27.384796	11.113122
PerCapita_CO2Emissions_1977	29.545993	44.541691	20.825033	18.854257	15.859107	21.63005	13.950562	27.263164	11.702886
PerCapita_CO2Emissions_1978	27.626074	48.105863	19.596375	18.911541	15.118379	21.121869	14.220393	27.199127	11.203367
PerCapita_CO2Emissions_1979	28.054604	43.357908	21.298971	17.721678	15.560132	20.67202	13.570624	28.70124	9.300423
PerCapita_CO2Emissions_1980	27.473647	42.920295	19.310375	16.385129	14.470489	20.238976	12.896822	29.394011	8.228573
PerCapita_CO2Emissions_1981	26.473861	40.775932	21.313077	18.724187	13.761752	19.416225	11.325476	27.420983	7.365017
PerCapita_CO2Emissions_1982	23.166184	53.146802	20.233271	18.621636	12.037846	19.263574	11.29168	24.538069	6.176019

Final Preprocessed US Dataset after Combining



DATA PRE-PROCESSING WEATHER DATA

Weather Data → **Cleaning**

Unstructured
data

Extracting data
Spliting


```
   [-86.77161380000001, 32.817508700000005],
   [-86.7984859, 32.8188788]]]},
 'properties': {'updated': '2024-04-22T21:11:21+00:00',
  'units': 'us',
  'forecastGenerator': 'BaselineForecastGenerator',
  'generatedAt': '2024-04-22T22:19:52+00:00',
  'updateTime': '2024-04-22T21:11:21+00:00',
  'validTimes': '2024-04-22T15:00:00+00:00/P7DT22H',
  'elevation': {'unitCode': 'wmoUnit:m', 'value': 116.1288},
  'periods': [{'number': 1,
    'name': 'This Afternoon',
    'startTime': '2024-04-22T17:00:00-05:00',
    'endTime': '2024-04-22T18:00:00-05:00',
    'isDaytime': True,
    'temperature': 68,
    'temperatureUnit': 'F',
    'temperatureTrend': None,
    'probabilityOfPrecipitation': {'unitCode': 'wmoUnit:percent',
     'value': None},
```


API Call

```
In [6]: response = requests.get('http://api.weather.gov/points/32.806671,-86.791130').json()
        t = response["properties"]
        t
```


```
Out[6]: {'@id': 'https://api.weather.gov/points/32.8067,-86.7911',
 '@type': 'wx:Point',
 'cwa': 'BMX',
 'forecastOffice': 'https://api.weather.gov/offices/BMX',
 'gridId': 'BMX',
 'gridX': 62,
 'gridY': 53,
 'forecast': 'https://api.weather.gov/gridpoints/BMX/62,53/forecast',
 'forecastHourly': 'https://api.weather.gov/gridpoints/BMX/62,53/forecast/hourly',
 'forecastGridData': 'https://api.weather.gov/gridpoints/BMX/62,53',
 'observationStations': 'https://api.weather.gov/gridpoints/BMX/62,53/stations',
 'relativeLocation': {'type': 'Feature',
  'geometry': {'type': 'Point', 'coordinates': [-86.876651, 32.788282]},
  'properties': {'city': 'Maplesville',
   'state': 'AL',
   'distance': {'unitCode': 'wmoUnit:m', 'value': 8254.5041247304},
   'bearing': {'unitCode': 'wmoUnit:degree_(angle)', 'value': 75}}},
 'forecastZone': 'https://api.weather.gov/zones/forecast/ALZ035',
 'county': 'https://api.weather.gov/zones/county/ALC021',
 'fireWeatherZone': 'https://api.weather.gov/zones/fire/ALZ035',
 'timeZone': 'America/Chicago',
 'radarStation': 'KBMX'}
```


Unstructured Data

Out[4]:

click to scroll output; double click to hide

	name	startTime	endTime	isDaytime	temperature	windSpeed	windDirection	shortForecast	detailedForecast
0	Afternoon	2024-04-22T17:00:00-05:00	2024-04-22T18:00:00-05:00	True	68	low	N	Sunny	Sunny, with a high near 68. North wind around ...
1	Tonight	2024-04-22T18:00:00-05:00	2024-04-23T06:00:00-05:00	False	40	low	N	Clear	Clear, with a low around 40. North wind 0 to 5...
2	Tuesday	2024-04-23T06:00:00-05:00	2024-04-23T18:00:00-05:00	True	75	low	S	Sunny	Sunny, with a high near 75. South wind 0 to 10...
3	Tuesday Night	2024-04-23T18:00:00-05:00	2024-04-24T06:00:00-05:00	False	51	low	S	Clear	Clear, with a low around 51. South wind 0 to 1...
4	Wednesday	2024-04-24T06:00:00-05:00	2024-04-24T18:00:00-05:00	True	78	low	SW	Mostly Sunny	Mostly sunny, with a high near 78. Southwest w...

Final selected data

DATA TRANSFORMATION – WORLD DATA

World Data → Data Aggregation → Data Standization → Data Regularization → Data Reduction

Data Aggregation

Country-wise split
Year-wise split
Combine & detailed split

Data Standization

Mean = 0 & Std = 1

Data Regularization

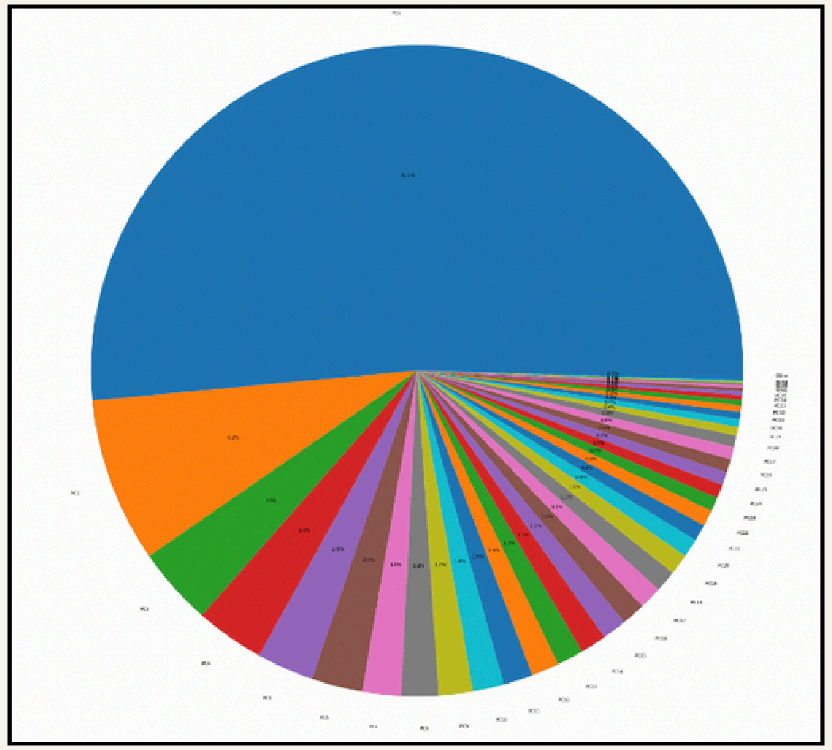
L2 regularization
Dropout

Data Reduction

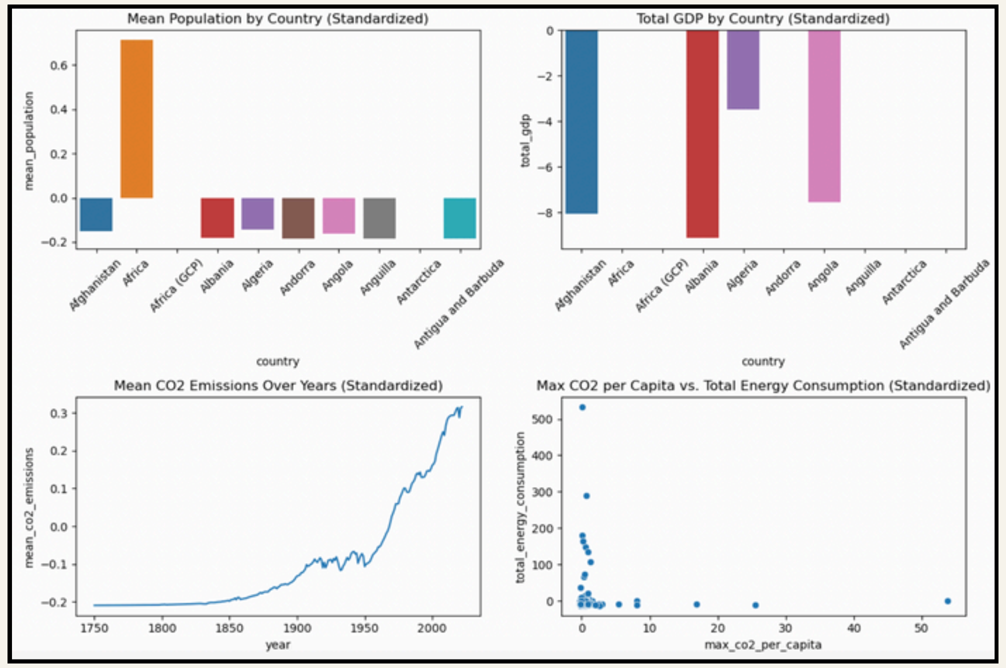
PCA
SVD

```
In [70]: def custom_aggregation(group):
             return pd.Series({
                 'mean_population': group['population'].mean(),
                 'total_gdp': group['gdp'].sum(),
                 'mean_co2_emissions': group['co2'].mean(),
                 'max_co2_per_capita': group['co2_per_capita'].max(),
                 'min_co2_per_capita': group['co2_per_capita'].min(),
                 'total_energy_consumption': group['primary_energy_consumption'].sum()
                 # Add more aggregations as needed
             })
```

Function for data aggregation



Variance for features



Data after data standization

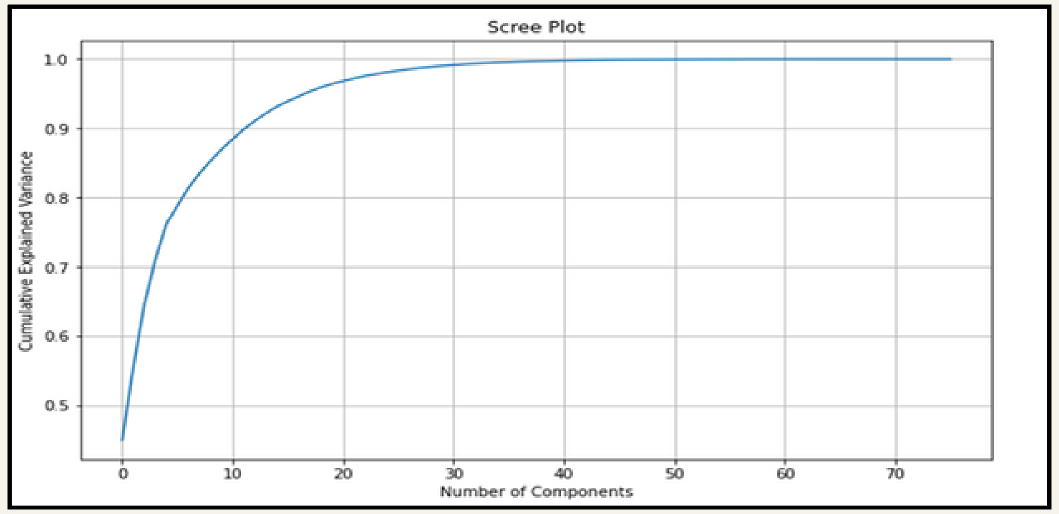
Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	4,160
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 64)	4,160
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 1)	65

Total params: 8,385 (32.75 KB)

Trainable params: 8,385 (32.75 KB)

L2 regularization Dropout



Scree Plot

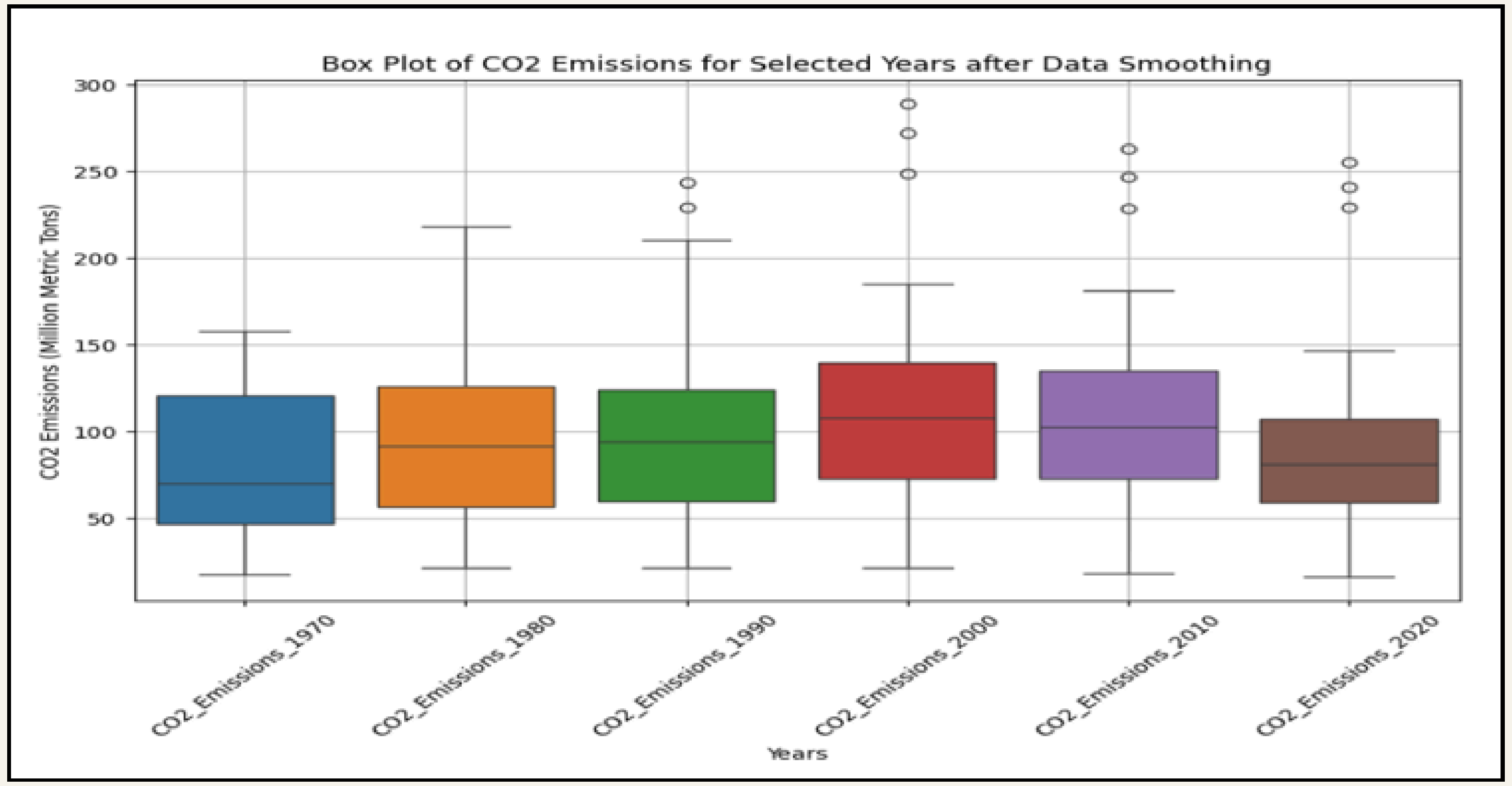
DATA TRANSFORMATION - US DATA

USA Data → **Data Smoothing** → **Data Normalization**

4 files

moving average method

Range [0,1]



Box Plot after performing Data Smoothing

State	Year	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	of Columbia	...	South Dakota	Tennessee	Texas	Utah	Vermont	Virginia
0	CO2_Emissions_1970	0.022852	0.001367	0.004558	0.007211	0.067970	0.008820	0.009953	0.002479	0.001907	...	0.000898	0.017029	0.083158	0.004437	0.000000	0.019081
1	CO2_Emissions_1971	0.021647	0.001663	0.005007	0.006892	0.069934	0.008871	0.009407	0.002430	0.001486	...	0.000886	0.016494	0.086221	0.004661	0.000000	0.019077
2	CO2_Emissions_1972	0.021966	0.001694	0.005406	0.006959	0.067995	0.009236	0.009180	0.002269	0.001226	...	0.000865	0.017991	0.086921	0.004432	0.000000	0.018366
3	CO2_Emissions_1973	0.021998	0.001393	0.006054	0.007409	0.068637	0.009582	0.009053	0.002402	0.001257	...	0.000747	0.019681	0.091007	0.004737	0.000000	0.017754
4	CO2_Emissions_1974	0.022788	0.001646	0.006922	0.007445	0.065688	0.009949	0.008836	0.002500	0.001000	...	0.000849	0.018718	0.092952	0.005353	0.000000	0.017548
5	CO2_Emissions_1975	0.023231	0.002118	0.007483	0.007063	0.069351	0.010557	0.008268	0.002348	0.000615	...	0.001295	0.019136	0.090897	0.005846	0.000000	0.017206
6	CO2_Emissions_1976	0.021614	0.002155	0.008090	0.007040	0.068512	0.010513	0.008016	0.002195	0.000388	...	0.001410	0.020512	0.088818	0.005312	0.000000	0.017124
7	CO2_Emissions_1977	0.021940	0.002511	0.009258	0.007424	0.072267	0.010878	0.007714	0.002135	0.000433	...	0.001340	0.020028	0.094373	0.005417	0.000000	0.016905
8	CO2_Emissions_1978	0.020722	0.002812	0.008937	0.007526	0.069763	0.010819	0.007844	0.002134	0.000339	...	0.001460	0.020227	0.100025	0.005817	0.000000	0.016223
9	CO2_Emissions_1979	0.021560	0.002460	0.010312	0.007084	0.072419	0.010870	0.007443	0.002380	0.000140	...	0.001489	0.019572	0.102919	0.006360	0.000000	0.017086
10	CO2_Emissions_1980	0.021577	0.002687	0.010155	0.006922	0.071508	0.011419	0.007478	0.002710	0.000134	...	0.001469	0.020241	0.107733	0.006594	0.000000	0.016263
11	CO2_Emissions_1981	0.021450	0.002717	0.011974	0.008311	0.071243	0.011527	0.006691	0.002566	0.000048	...	0.001391	0.020515	0.109571	0.006328	0.000000	0.015953
12	CO2_Emissions_1982	0.019801	0.004527	0.012404	0.008813	0.067178	0.012521	0.007155	0.002426	0.000258	...	0.001673	0.019181	0.109800	0.006420	0.000000	0.016128
13	CO2_Emissions_1983	0.019550	0.004918	0.011375	0.009761	0.065096	0.012061	0.006943	0.003032	0.000140	...	0.001370	0.020667	0.110619	0.006552	0.000000	0.016656
14	CO2_Emissions_1984	0.019551	0.005146	0.011647	0.008702	0.067537	0.012171	0.007176	0.002928	0.000058	...	0.001270	0.020411	0.109251	0.006713	0.000000	0.016975
15	CO2_Emissions_1985	0.021051	0.005268	0.012240	0.009648	0.068629	0.012338	0.007198	0.002744	0.000000	...	0.001370	0.021391	0.110423	0.006983	0.000096	0.016877
16	CO2_Emissions_1986	0.020982	0.005720	0.011124	0.009762	0.065499	0.012090	0.007487	0.002627	0.000032	...	0.001218	0.021637	0.109330	0.006596	0.000000	0.017852
17	CO2_Emissions_1987	0.020751	0.005205	0.010805	0.008846	0.069575	0.011861	0.007303	0.002717	0.000000	...	0.000927	0.020738	0.107446	0.008314	0.000094	0.018676
18	CO2_Emissions_1988	0.020072	0.005081	0.010970	0.009243	0.068149	0.011809	0.007493	0.002731	0.000000	...	0.001320	0.020374	0.108725	0.009380	0.000165	0.018486

DataFrame after performing Data Normalization

DATA TRANSFORMATION – WEATHER DATA

Weather Data → **Data Standization** → **Data Discretization**

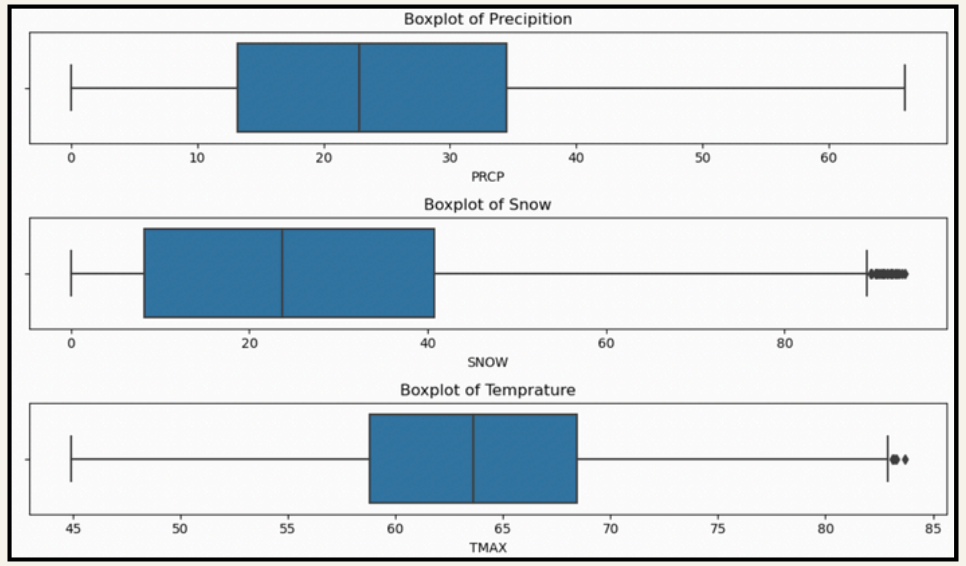
Unstructured data

Mean = 0 & Std = 1

wind speed
Relative Humidity


```
[-86.77161380000001, 32.817508700000005],
[-86.7984859, 32.8188788]]]},
'properties': {'updated': '2024-04-22T21:11:21+00:00',
 'units': 'us',
 'forecastGenerator': 'BaselineForecastGenerator',
 'generatedAt': '2024-04-22T22:19:52+00:00',
 'updateTime': '2024-04-22T21:11:21+00:00',
 'validTimes': '2024-04-22T15:00:00+00:00/P7DT22H',
 'elevation': {'unitCode': 'wmoUnit:m', 'value': 116.1288},
 'periods': [{'number': 1,
   'name': 'This Afternoon',
   'startTime': '2024-04-22T17:00:00-05:00',
   'endTime': '2024-04-22T18:00:00-05:00',
   'isDaytime': True,
   'temperature': 68,
   'temperatureUnit': 'F',
   'temperatureTrend': None,
   'probabilityOfPrecipitation': {'unitCode': 'wmoUnit:percent',
    'value': None},
   'dewpoint': {'unitCode': 'wmoUnit:degC', 'value': 2.2222222222222223},
```


Unstructured data



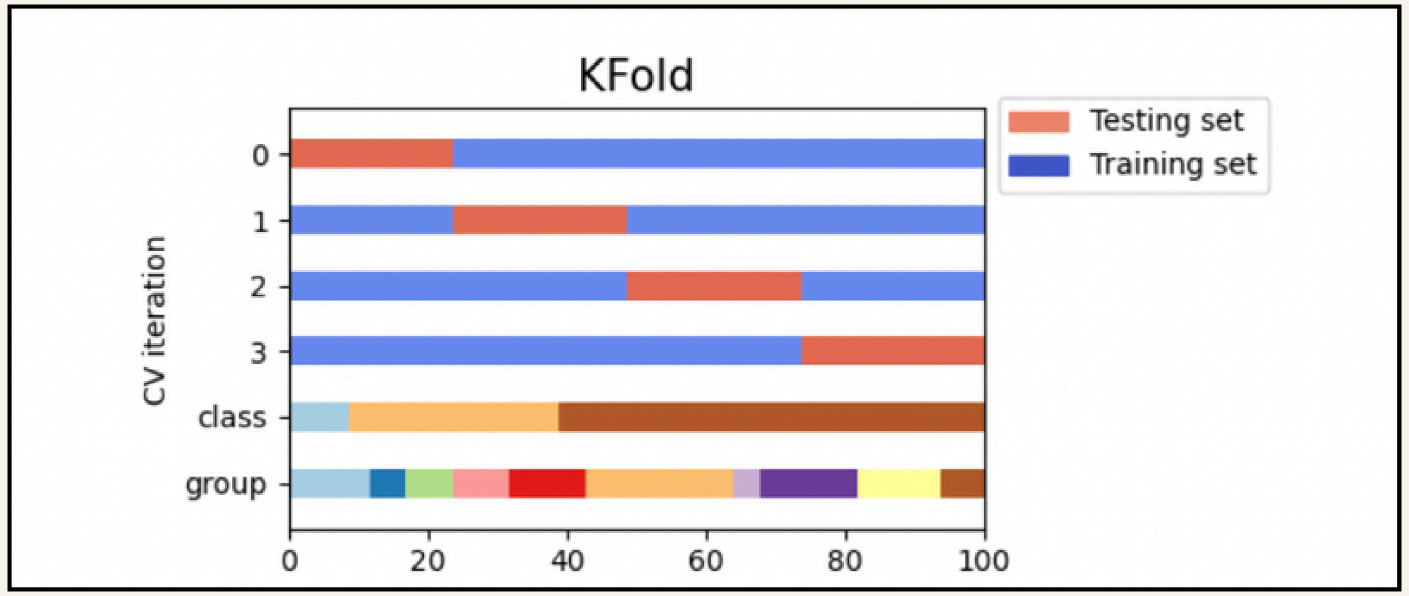
Data Standization

name	startTime	endTime	isDaytime	temperature	windSpeed	windDirection	shortForecast	detailedFore
This Afternoon	2024-05-06T14:00:00-07:00	2024-05-06T18:00:00-07:00	True	72	mediun	NNW	Sunny	Sunny, with a high near 72. N northwest
Tonight	2024-05-06T18:00:00-07:00	2024-05-07T06:00:00-07:00	False	46	low	NNW	Mostly Clear	Mostly clear. Low around 46, temperatur
Tuesday	2024-05-07T06:00:00-07:00	2024-05-07T18:00:00-07:00	True	76	mediun	NNW	Sunny	Sunny, with a high near 76. N northwest
Tuesday Night	2024-05-07T18:00:00-07:00	2024-05-08T06:00:00-07:00	False	49	mediun	NW	Mostly Clear	Mostly clear, with a low around Northwes
Wednesday	2024-05-08T06:00:00-07:00	2024-05-08T18:00:00-07:00	True	80	low	NW	Sunny	Sunny, with a high near Northwest wind 5

Data Discretization Results

DATA PREPARATION

World Data
K-fold cross validation
k=5




```
Fold 5:
  Train count: 38447
  Validation count: 4805
  Test count: 4806
```

US DATA
Based on years

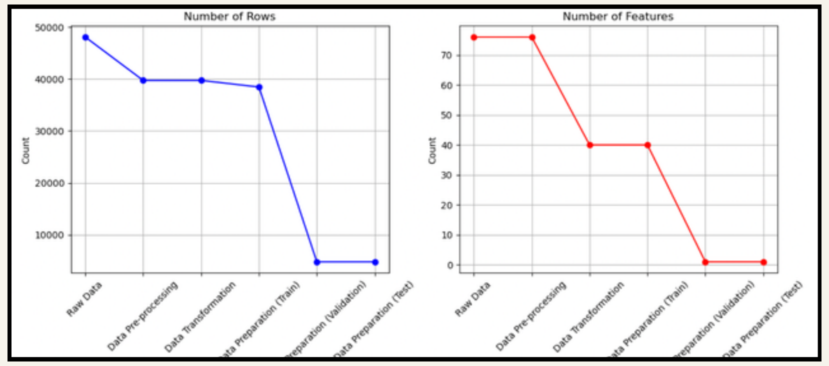
```
Training Data Shape:
(137, 52)

Validation Data Shape:
(20, 52)

Test Data Shape:
(24, 52)
```

Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	District of Columbia	...	South Dakota	Tennessee	Texas	Utah	Vermont	Virginia	Washington	West Virginia
0.001367	0.004558	0.007211	0.067970	0.008820	0.009953	0.002479	0.001907	...	0.000898	0.017029	0.083158	0.004437	0.000000	0.019081	0.009231	0.016811
0.661033	0.415963	0.463792	0.509437	0.669604	0.729537	0.844926	0.898430	...	0.272811	0.721032	0.449581	0.737408	0.596554	0.832878	0.000000	1.000000
0.578143	0.039114	0.150161	0.057859	0.164537	0.081220	0.390382	0.134487	...	0.041506	0.174564	0.454182	0.245758	0.005048	0.147083	0.021103	0.730603
0.001663	0.005007	0.006892	0.069934	0.008871	0.009407	0.002430	0.001486	...	0.000886	0.016494	0.086221	0.004661	0.000000	0.019077	0.009461	0.016885
0.668860	0.426215	0.475300	0.512043	0.655435	0.638635	0.827102	0.857719	...	0.223600	0.687147	0.445764	0.736976	0.585109	0.805189	0.000000	1.000000
0.646242	0.049559	0.131938	0.067907	0.158013	0.067021	0.372918	0.086549	...	0.040758	0.160244	0.476134	0.256518	0.000000	0.145976	0.029575	0.740277
0.579489	0.051270	0.119120	0.054517	0.145488	0.058378	0.309811	0.055421	...	0.036595	0.175621	0.428152	0.205053	0.000000	0.118396	0.054044	0.741294
0.645681	0.417768	0.476800	0.507131	0.651617	0.601244	0.789026	0.830657	...	0.222071	0.688398	0.427575	0.698643	0.536699	0.767861	0.000000	1.000000
0.001694	0.005406	0.006959	0.067995	0.009236	0.009180	0.002269	0.001226	...	0.000865	0.017991	0.086921	0.004432	0.000000	0.018366	0.010307	0.018439
0.001393	0.006054	0.007409	0.068637	0.009582	0.009053	0.002402	0.001257	...	0.000747	0.019681	0.091007	0.004737	0.000000	0.017754	0.010929	0.019459

DATA STATISTICS

Process	Method	Records
Data Collection	Raw Data	48058 rows & 76 features
Data Pre-processing	Handled noise, missing values, data scrubbing, and data auditing	39717 rows & 76 features
Data Transformation	Data Aggregation	39717 rows & 76 features
Data Transformation	Data Standardization	39717 rows & 76 features
Data Transformation	L2 regularization	39717 rows & 76 features
Data Transformation	Dropout	39717 rows & 76 features
Data Transformation	PCA	39717 rows & 40 features
Data Transformation	SVD	39717 rows & 40 features
Data Preparation	K-fold cross-validation (Train data)	38447 rows & 40 features
Data Preparation	K-fold cross-validation (Validation data)	4805 rows & 1 feature
Data Preparation	K-fold cross-validation (Test data)	4806 rows & 1 feature

World Data

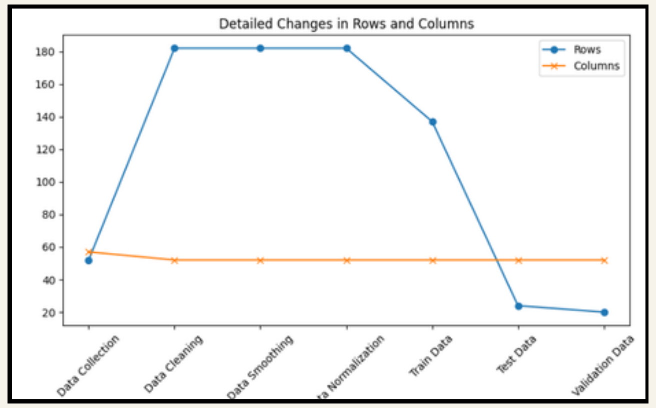


Weather Data

Process	Method	Records
Data Collection	Extracting and splitting	5868 records
Data Cleaning	Handling null values and selection of time frame	3668 records
Data Transformation	Data Standardization	3668 records
Data Transformation	Data discretization	3668 records

Process	Method	Records
Data Collection	Raw Data	
	Sheet -1	52 rows x 57 columns
	Sheet -4	52 rows x 57 columns
	Sheet -6	52 rows x 57 columns
	Sheet -7	52 rows x 30 columns
Data Cleaning	Remove unnecessary rows and columns, check for duplicate and missing values, rename columns, transpose rows and columns, and merge all four sheets into a single dataset	182 rows x 52 columns
Data Transformation	Data Smoothing	182 rows x 52 columns
Data Transformation	Data Normalization	182 rows x 52 columns
Data Preparation	Train data	137 rows x 52 columns
Data Preparation	Test data	24 rows x 52 columns
Data Preparation	Validation data	20 rows x 52 columns

USA Data

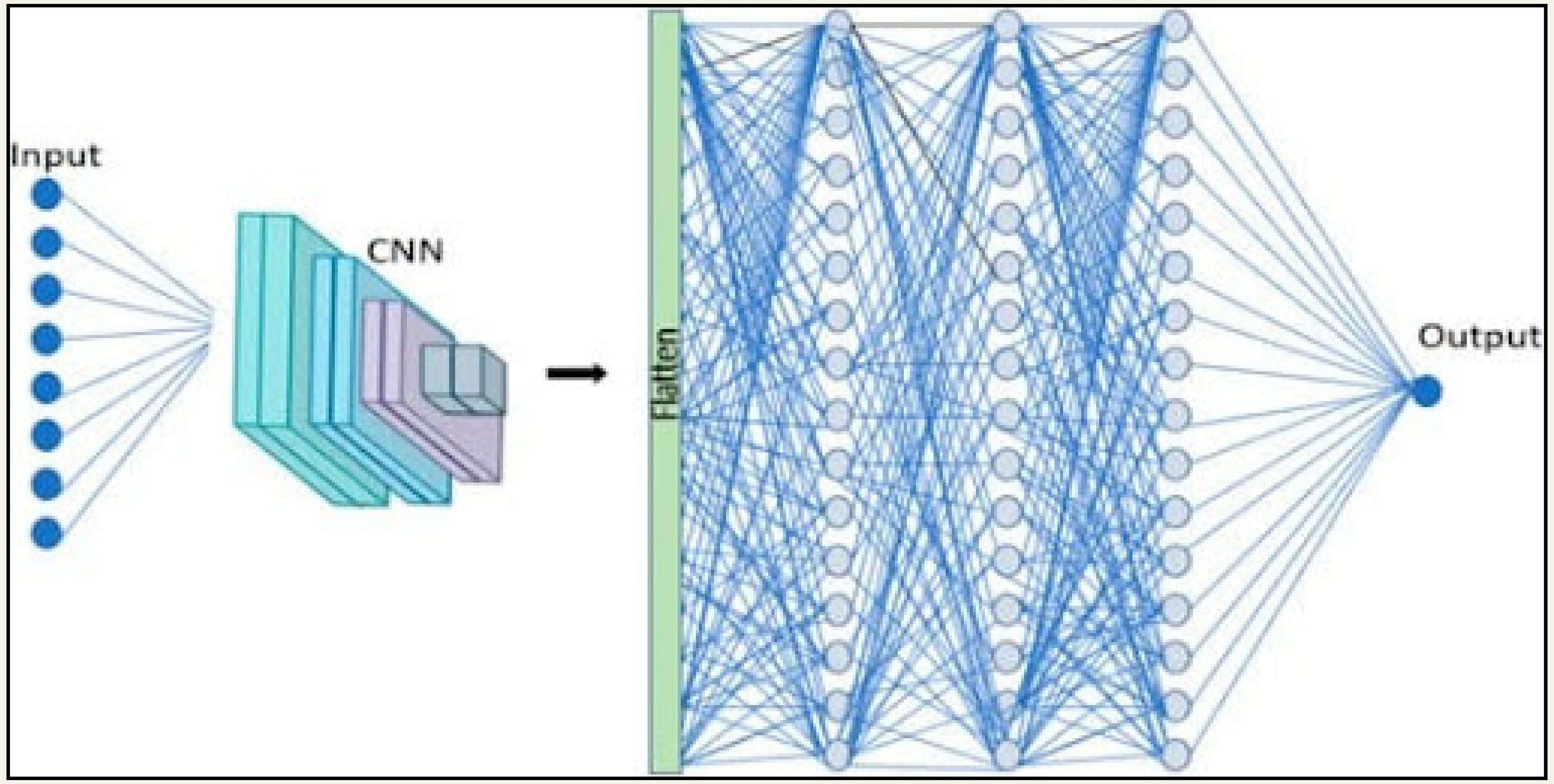


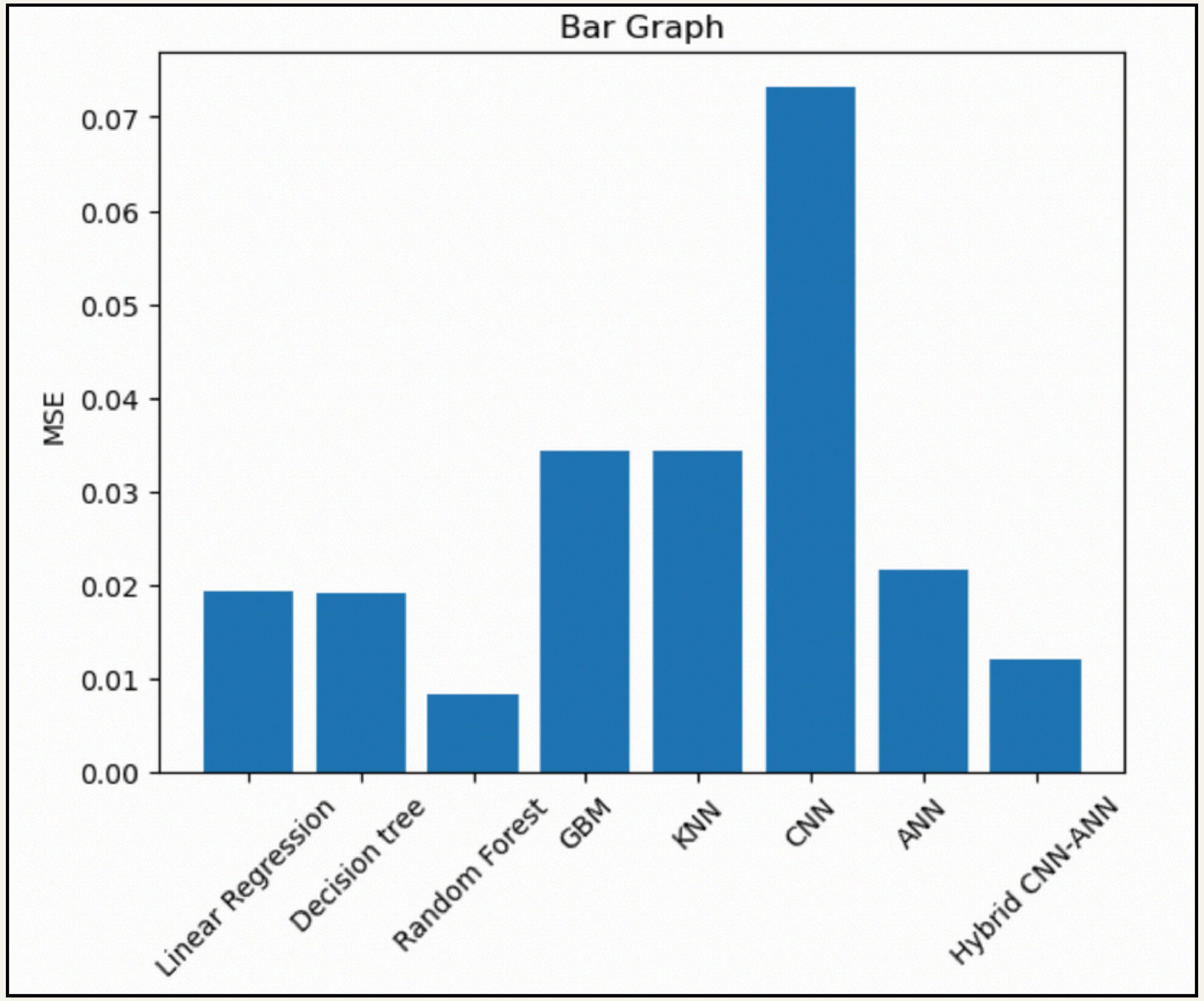
MODEL DEPLOYMENT

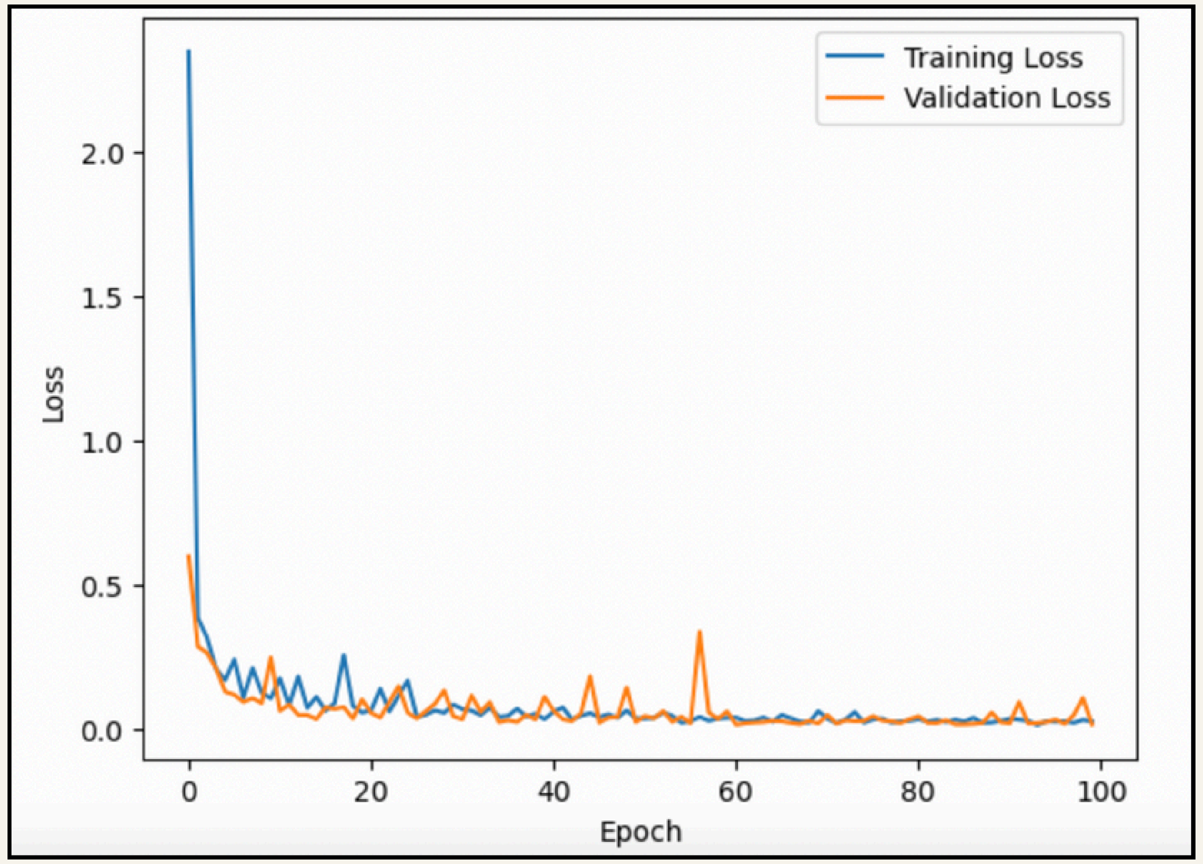
MODEL SELECTION, UPDATES, JUSTIFICATION AND COMPARISION

- The initial phase involved evaluating a range of models, including KNN, XGBoost, Decision Tree, Random Forest, CNN, ANN, Linear Regression, Hybrid CNN-ANN, Hybrid CNN-IBFA, SVR, and LSTM.
- Ultimately, the selection was finalized with four models: Hybrid CNN-ANN, Hybrid CNN-IBFA, SVR, and LSTM.
- Each model was chosen for its effectiveness in handling different types of data, which is crucial for making accurate predictions.
- For instance, the Hybrid CNN models are used to process complex, detail-rich data, and the LSTM for data that involves time-related changes.
- Throughout the project, the models have been continuously updated based on their performance.
- This often involves tweaking model settings or focusing more on one model over others to better align with the project's evolving needs.
- For example, the Hybrid CNN-ANN has shown outstanding accuracy, evidenced by a very high R-square value of 0.99962.
- Although the Hybrid CNN-IBFA has a higher error rate (MSE), it still performs robustly with an R-square of 0.9338.
- The SVR excels in handling noisy data, with an MSE of 0.0768 and an R-square of 0.875.
- The LSTM is particularly adept at accurately predicting future trends, boasting the lowest error rate (MSE) of 0.0038, although its R-square is moderately good at 0.88920.
- Specific modifications have been made to each model to better suit the project, such as adapting the LSTM to manage longer data sequences without losing accuracy, and ensuring the Hybrid CNN-IBFA can process larger datasets without a decline in performance.

HYBRID CNN-ANN



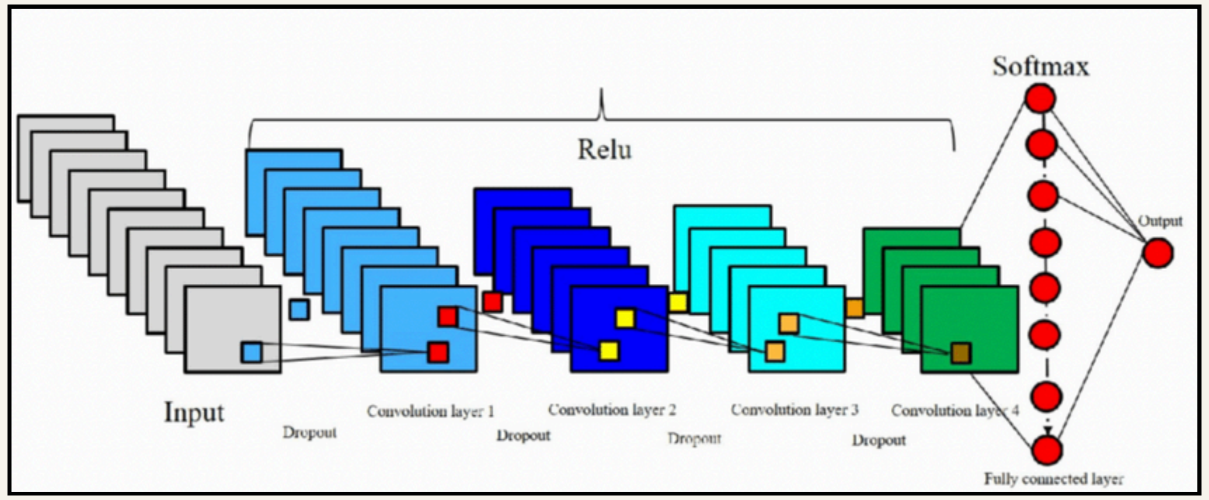




Epochs: 100
Batch Size: 32
Optimizer: adam
Activation = ReLU
Learning Rate: 2e-5
Train Accuracy: 93.07 %
CV Accuracy: 92.00%

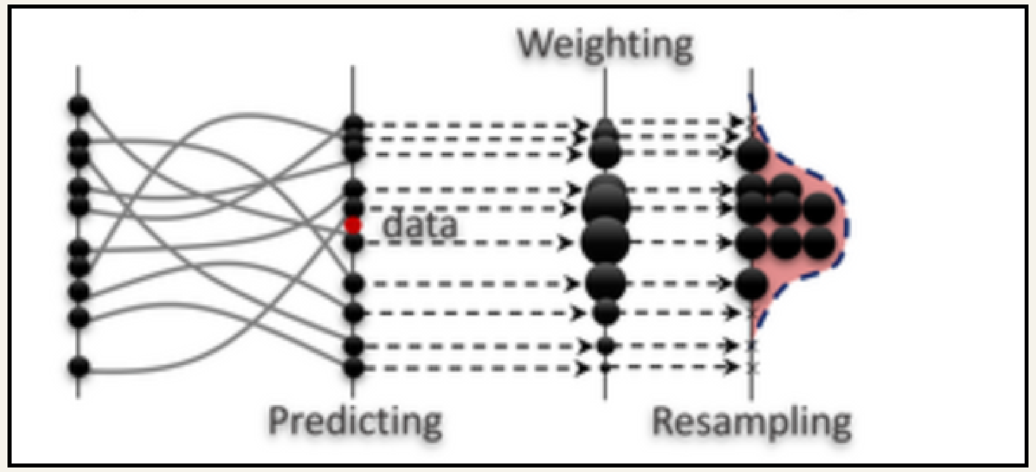
Test Evaluation Matrix:
MSE: 0.0121
MAE: 0.0279
R-squared: 0.9996

CNN-IBFA HYBRID MODEL



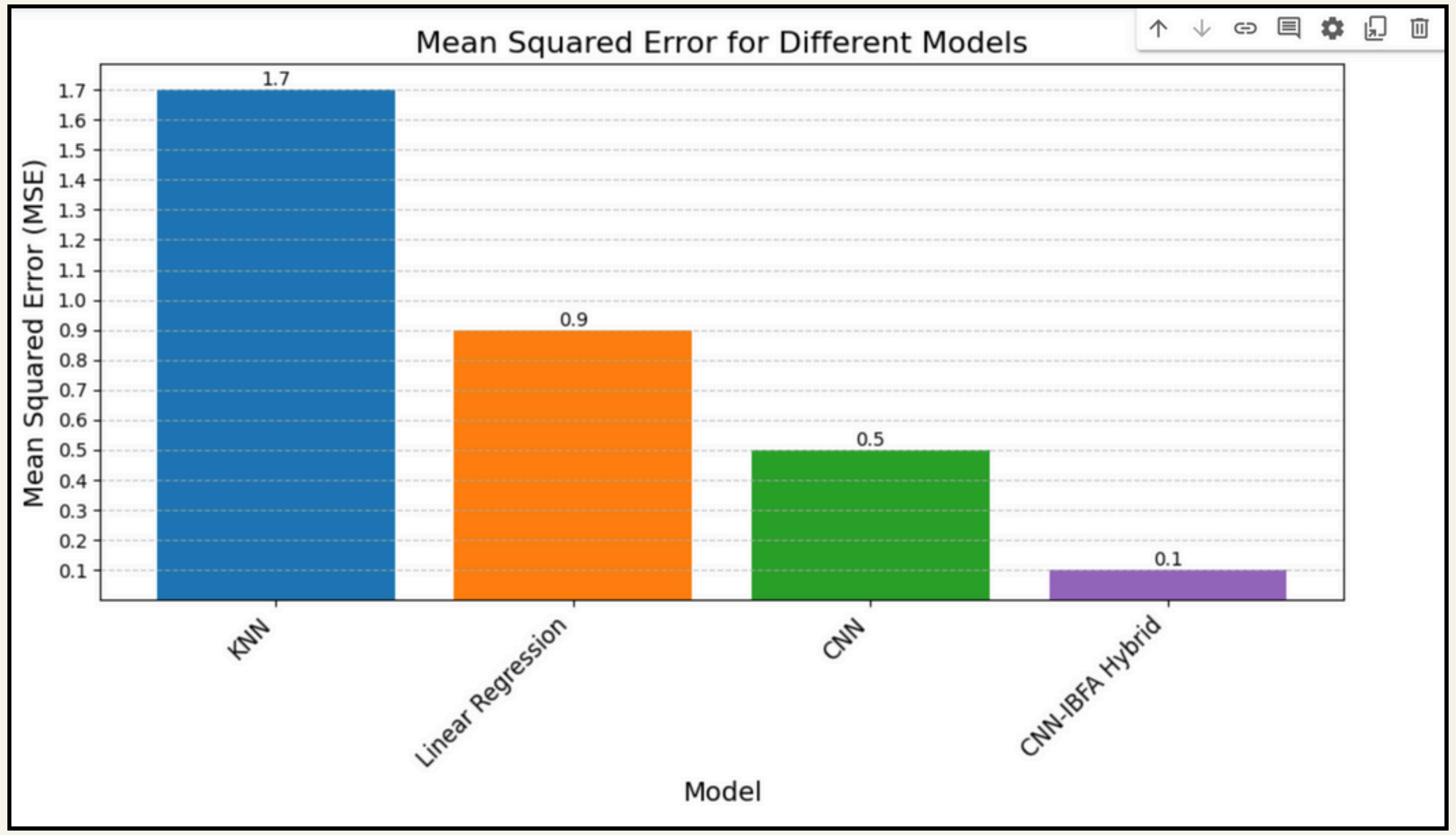
Convolutional neural network

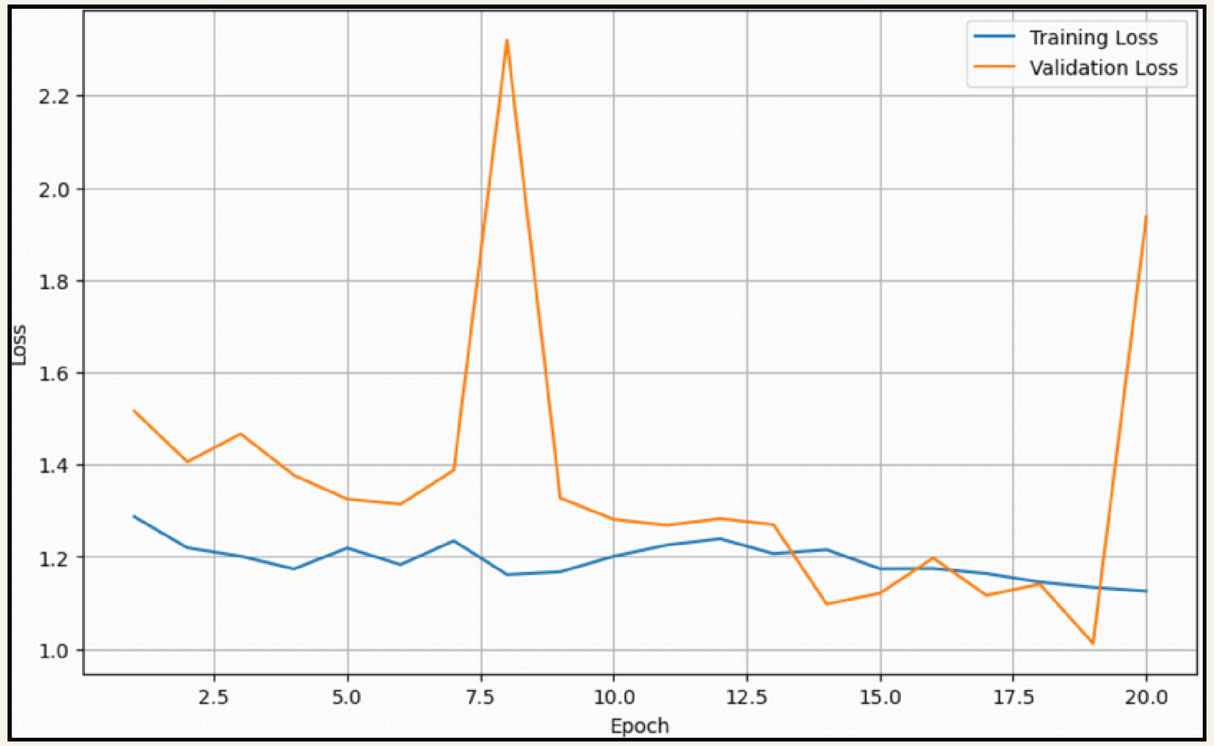
+



Iterative Bayesian filtering and adaptive

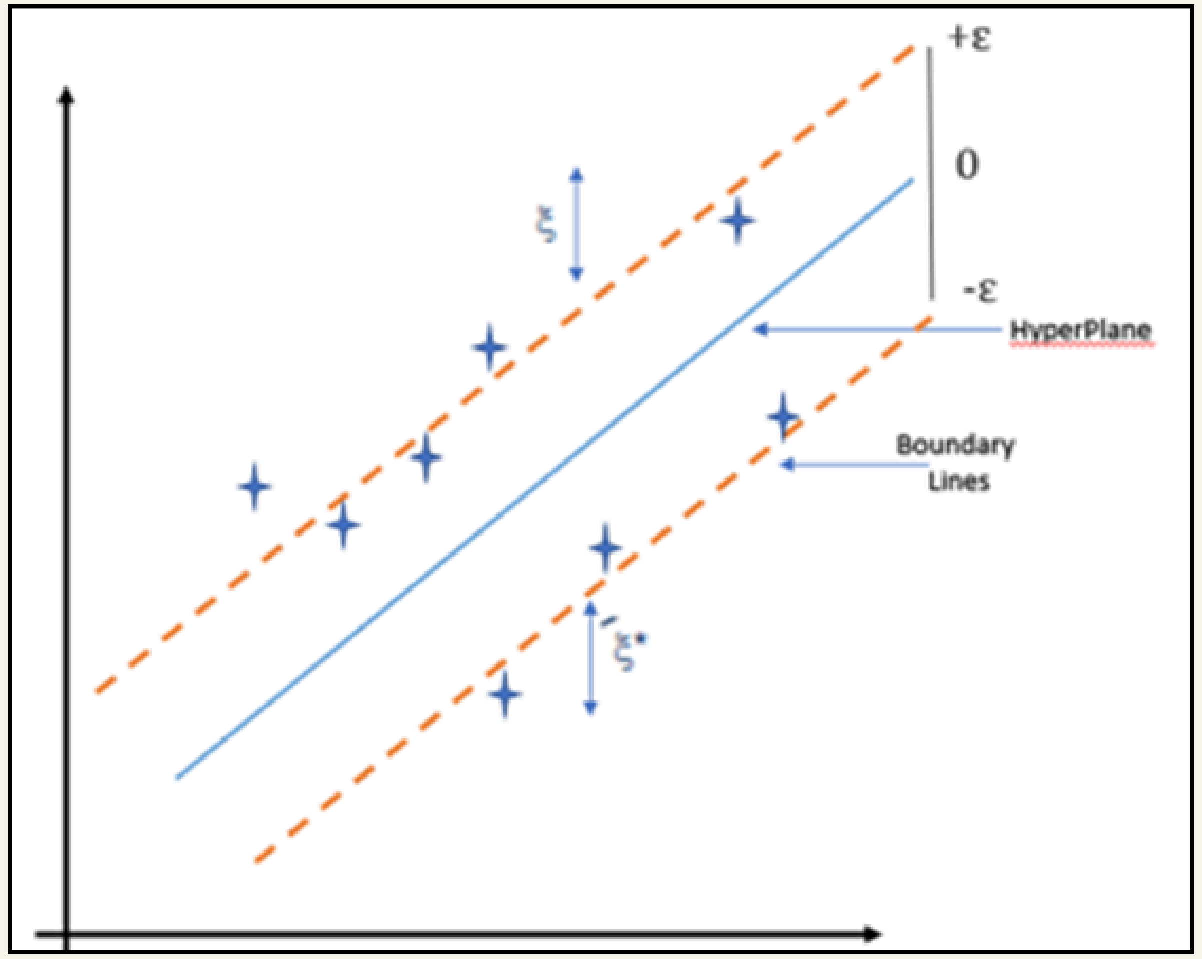
Epochs: 20
Optimizer: Adam
Activation: ReLu
Learning Rate: 4.3e-4.
Training Accuracy: 99 %
CV accuracy: 94%

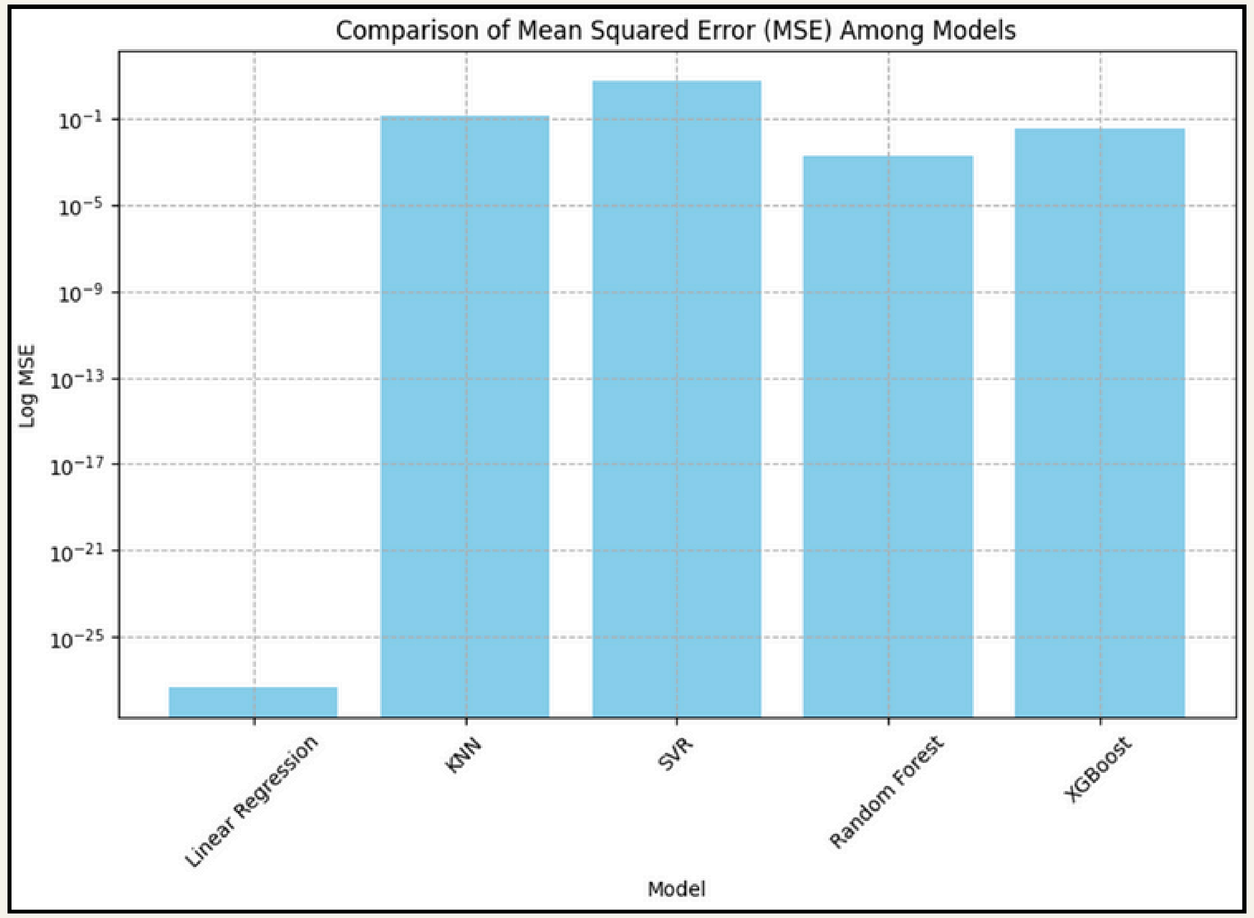




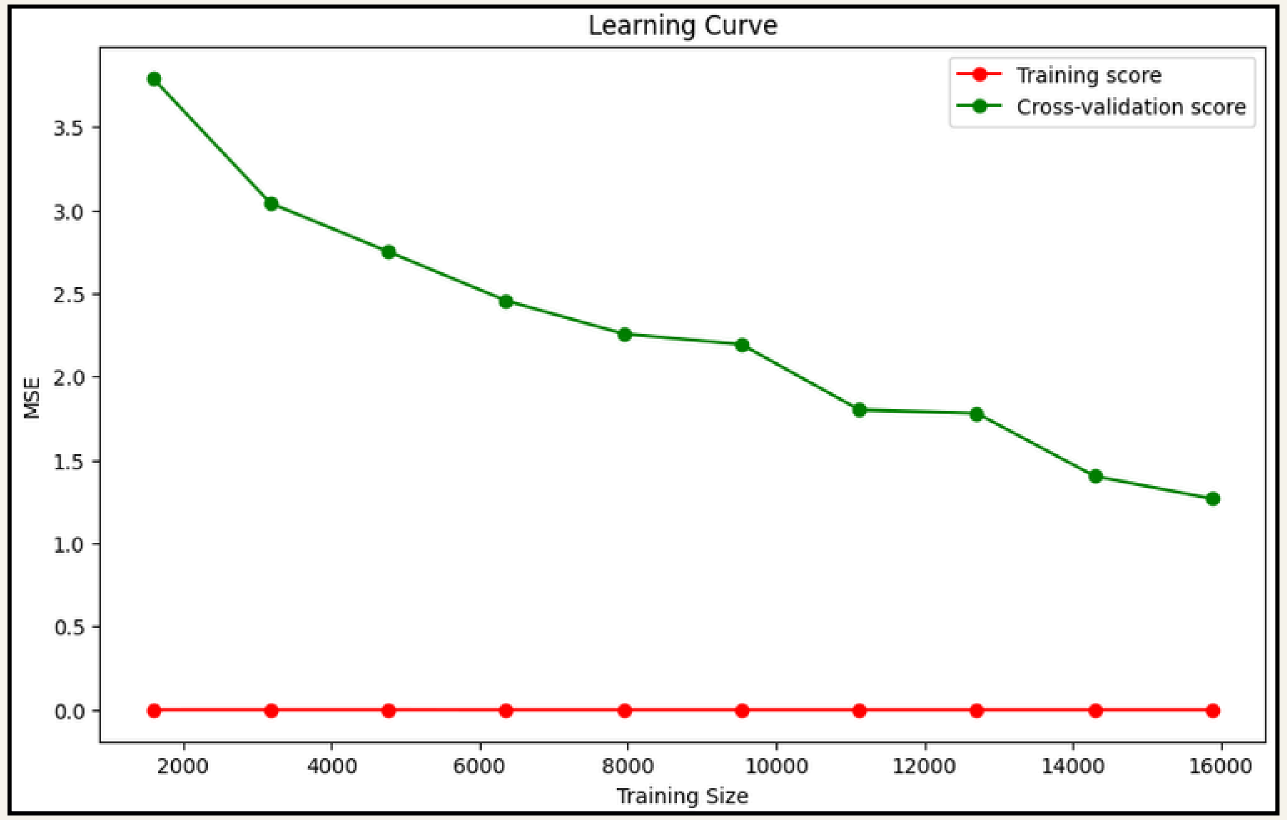
MSE: 0.11418
MAE: 0.0182
R-squared: 0.9338

SUPPORT VECTOR REGRESSION (SVR)





Kernel Type: Linear
Regularization Parameter (C) : 100
Epsilon (ε) : 0.02
Gamma (γ): 0.05



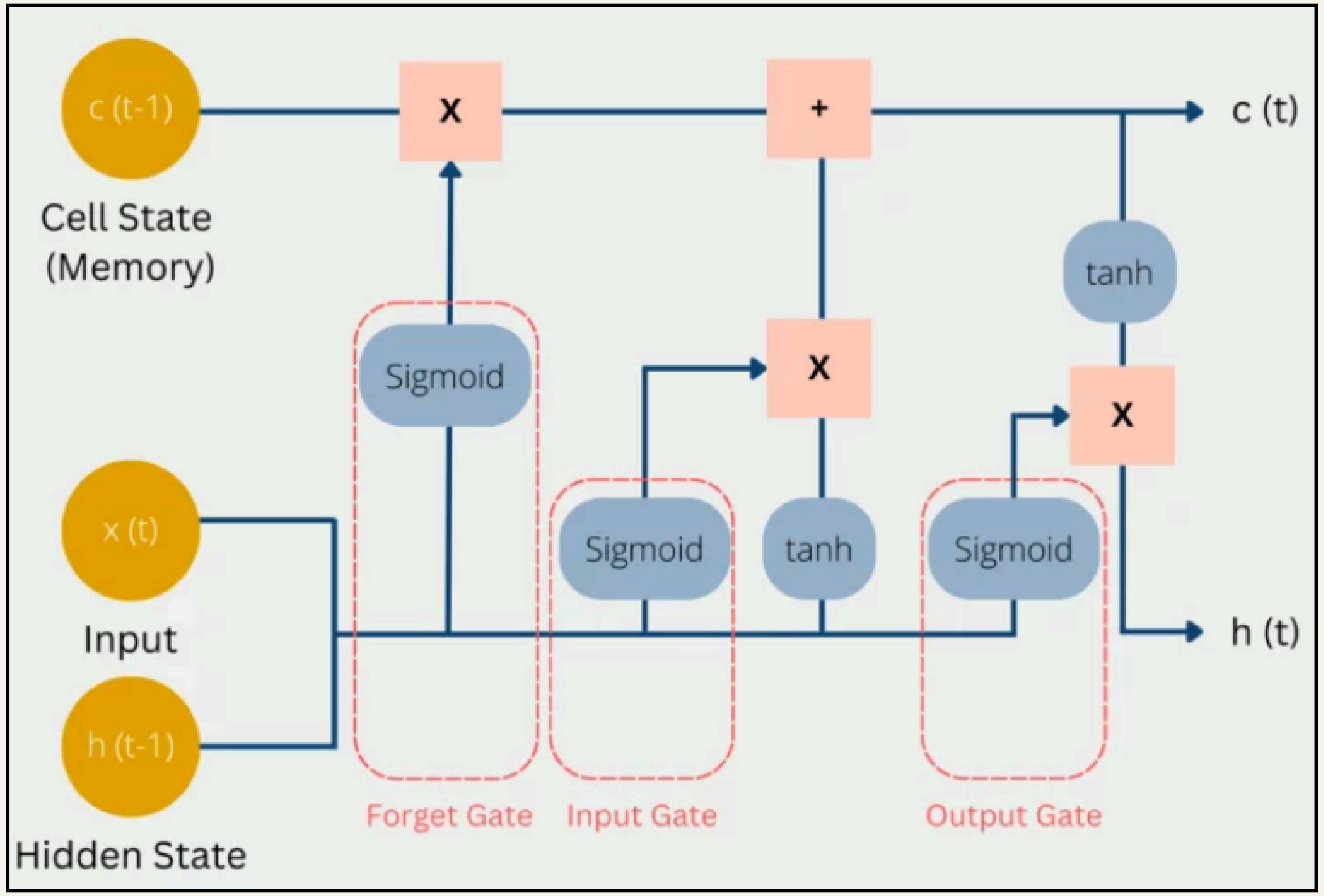
Test Evaluation Metrics:

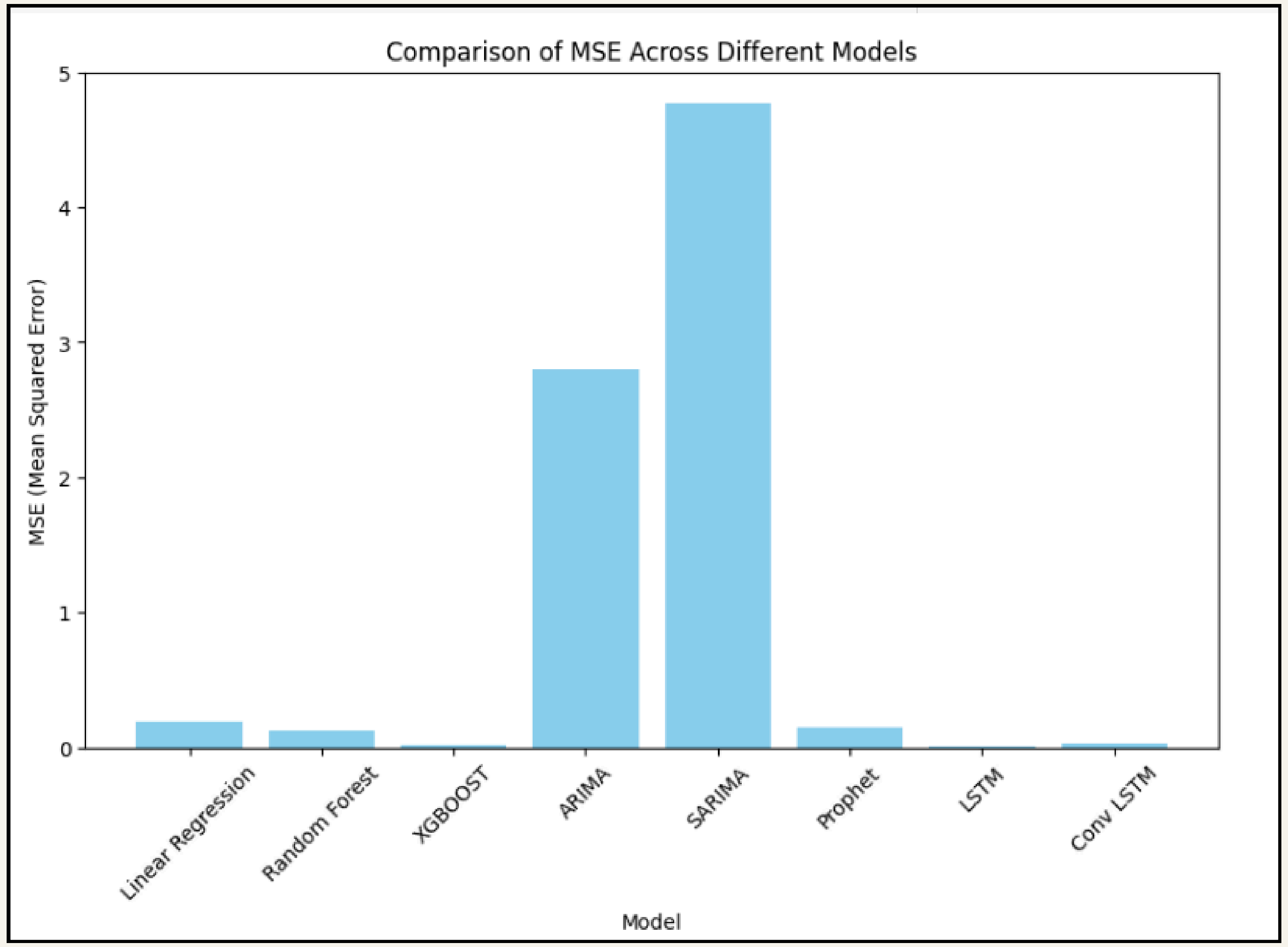
MSE: 0.0768

MAE: 0.00949

R-Squared: 0.875

LONG SHORT-TERM MEMORY(LSTM)





HyperParamater Tuning - Using GridSearch CV:
Units : 50
Activation Function: 'relu'
Optimizer: Adam with a learning_rate of 0.01
Epochs: 150
Batch Size: 32
Verbosity: 1

Test Evaluation Metrics:
MSE: 0.0038
MAE: 0.035411
R-Squared: 0.88920

MODEL COMPARISION

Model	Hybrid CNN-ANN	Hybrid CNN-IBFA	SVR	LSTM
Description	To leverage spatial feature extraction capabilities while accommodating the flexibility of neural networks.	Identify and extract meaningful spatial patterns, textures, and structures. To handle diverse and intricate spatial patterns.	To minimize significant errors while being robust to minor deviations, handling both linear and non linear relationship depending on the chosen kernel function.	A type of recurrent neural network capable of learning order dependence in sequence prediction and time series problems.
Advantages	Effective feature extraction, Flexibility versatility, and Hierarchical feature learning	High performance, Robustness to Variability, Scalability.	Effective in High-Dimensional Spaces, Versatile with Different Kernel Functions, Handles Non-Linear Data	Excel at modeling long-term dependencies in complex time-series, overcoming vanishing gradients for accurate forecasting.
Disadvantages	Computational complexity, Data requirements, and Interpretability	Computational Intensity and Algorithmic Complexity, Overtraining	Complexity in Choosing the Right Hyperparameters, Computationally Intensive and Difficulty with Noisy Data	Computationally costly, slow to train on big datasets, and prone to overfitting noisy data.
Targeted Problem	Spatial data analysis and Complex pattern recognition	Identifying intricate patterns within datasets using sophisticated algorithms and techniques.	Finding a hyperplane that best fits the data points within a specified margin of tolerance.	Capture model complexity and variable-length time-series patterns over long ranges.
MSE MAE R-square	0.0121 0.027928 0.99962	0.11418 0.0182 0.9338	0.0768 0.00949 0.875	0.0038 0.035411 0.88920

APPLICATION RESULTS

localhost:5000

Environmental Data

Cost to Offset Top 4 Emitters' Carbon Footprint

	China	India	Russia	United States
2023	$146,497,400,000.00	$567,674,200.00	$14,755,060,000.00	$1,141,095,000.00
2024	$156,366,400,000.00	$608,121,200.00	$13,182,630,000.00	$1,168,953,000.00
2025	$163,360,300,000.00	$643,958,300.00	$10,889,510,000.00	$1,167,791,000.00
2026	$173,154,600,000.00	$688,613,300.00	$8,029,967,000.00	$1,164,584,000.00

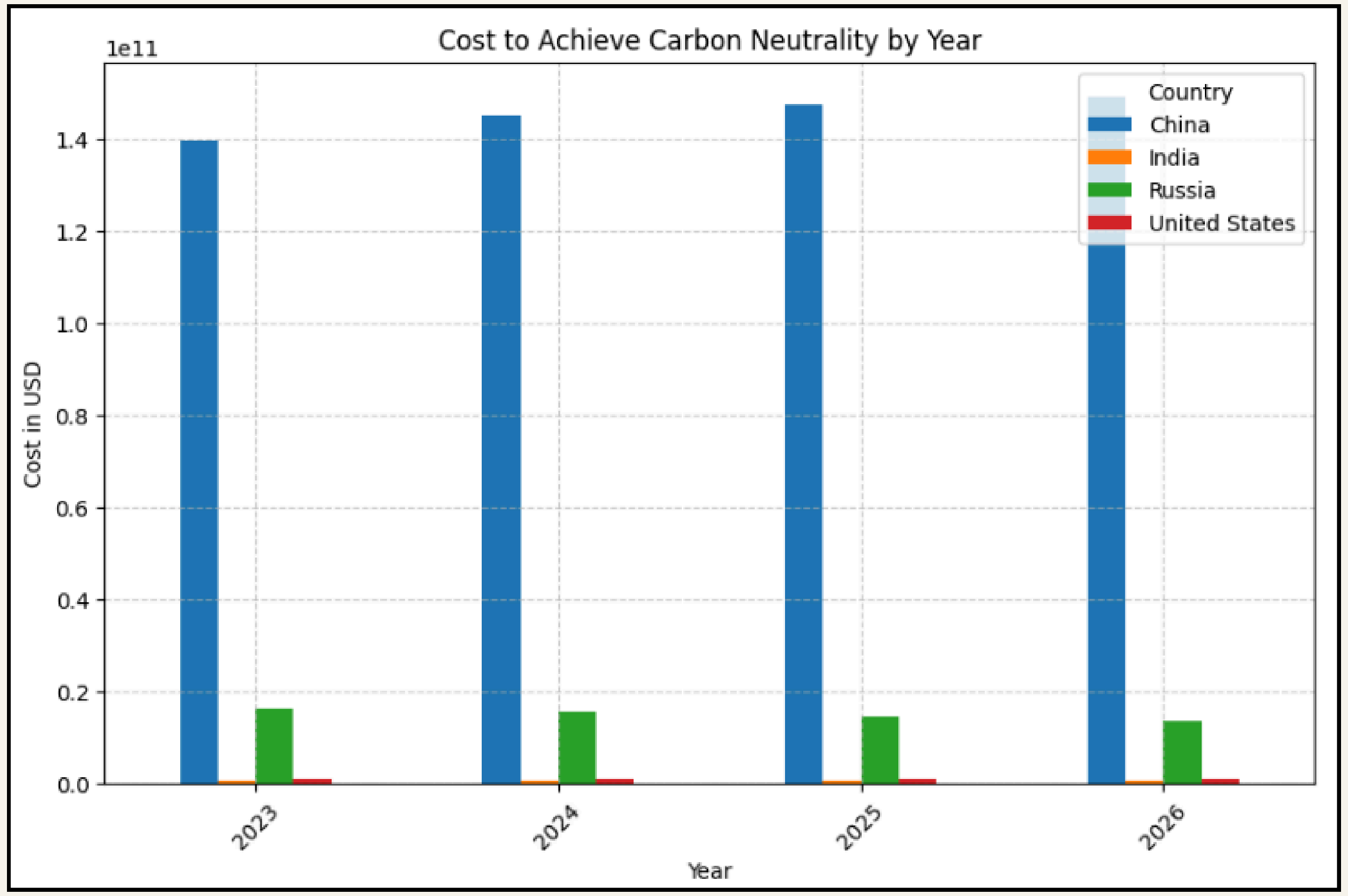
Carbon Emissions of Top 4 Countires

Price as of May 5

	China	India	Russia	United States
2023	12067.552856	2723.127324	1350.011139	5239.351029
2024	12531.528857	2877.580555	1293.267057	5276.914026
2025	12726.294565	2954.769346	1209.748493	5212.758325
2026	12887.386095	3014.155150	1130.689892	5132.366530

Carbon Emissions and their offset cost

Comparision of Offset Cost



CONCLUSION

The application of advanced models including Hybrid CNN-ANN, Hybrid CNN-IBFA, SVR, and LSTM has significantly bolstered our efforts in managing carbon emissions and accurately forecasting carbon credit needs. The Hybrid CNN-ANN model utilizes spatial and temporal data for precise carbon credit predictions, crucial for effective carbon trading strategies. The Hybrid CNN-IBFA model further enhances accuracy through optimization algorithms. Additionally, SVR provides robust regression capabilities for continuous data predictions, while LSTM excels in capturing long-term dependencies in time-series data, essential for tracking emission trends. Together, these models form a comprehensive toolkit that adapts dynamically to environmental changes, supporting sustainable policy decisions and promoting effective climate action. This approach marks a substantial progression in optimizing carbon credit systems and improving emission management.

REFERENCES

- Abdi, H. (2007). Singular value decomposition (SVD) and generalized singular value decomposition. Encyclopedia of measurement and statistics, 907(912), 44.
- Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102-4107.
- Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012, April). The'K'in K-fold Cross Validation. In ESANN (Vol. 102, pp. 441-446).
- Camizuli, E., & Carranza, E. J. (2018). Exploratory Data Analysis (EDA). The Encyclopedia of Archaeological Sciences, 1–7. https://doi.org/10.1002/9781119188230.saseas0271
- Cortes, C., Mohri, M., & Rostamizadeh, A. (2012). L2 regularization for learning kernels. arXiv preprint arXiv:1205.2653.
- Dasu, T., & Johnson, T. (2003). Exploratory data mining and data cleaning. John Wiley & Sons.
- Dimitrova, E. S., Licona, M. P. V., McGee, J., & Laubenbacher, R. (2010). Discretization of time series data. Journal of Computational Biology, 17(6), 853-868.
- Fink, E. L. (2009). The FAQs on data transformation. Communication Monographs, 76(4), 379-397.
- Gal, M. S., & Rubinfeld, D. L. (2019). Data standardization. NYUL Rev., 94, 737.
- Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. Psychophysiology, 48(12), 1711-1725.
- Jankowski, N. (2000). Data regularization. Neural Networks and Soft Computing, Zakopane, Poland, 209-214.
- Jin, R., Breitbart, Y., & Muoh, C. (2009). Data discretization unification. Knowledge and Information Systems, 19, 1-29.

THANK YOU!