Contents

Figures

Tables

Objective

The Python code developed in this study performs several tasks including data preprocessing and visualization using Pandas, Seaborn, Matplotlib, and Scikit-learn libraries. In the following sections, each of the mentioned tasks is described.

Predict outcomes using three different approaches: -

1.Neural network with back-propagation implemented (BP).

2. Neural network with back-propagation using free software (BP-F).

3. Multiple linear regression using free software (MLR-F).

Dataset:

1. Turbine dataset

2. Synthetic dataset

3. Boston House Price :- This is Sklearn dataset which is available in sklearn datasets.

https://scikit-learn.org/1.0/modules/generated/sklearn.datasets.load_boston.html

- ✓ The dataset contains 506 entries.
- ✓ The features vary significantly in their ranges and scales. For example, CRIM (crime rate) has a mean of approximately 3.61 but a maximum value of about 88.98, indicating potential outliers or a wide variation in crime rates across towns.
- ✓ The CHAS variable, which is a dummy variable, has values 0 or 1, indicating whether the tract bounds the Charles River.
- ✓ The mean number of rooms (RM) is around 6.28, with a minimum of 3.56 and a maximum of 8.78.
- ✓ The median value of homes (MEDV) has a mean of approximately $22,533 with a wide range from $5,000 to $50,000.
- ✓ There are no missing values in any of the columns, which is beneficial for analysis.

Loading and Displaying Data:

In the first step, the 'A1-turbine.txt' and 'A1-synthetic.txt' files are loaded using their paths and Pandas DataFrame. Then the columns' names are corrected if needed and the first few rows are displayed as shown in

Table 1 and Table 2 for A1-turbine and A1-synthetic datasets, respectively.

Table 1: The display of the first few rows of the A1-turbine dataset

#	Column	Count	Non-Null	Dtype
0	Height over Sea	450	non-null	float64
1	Fall 1	450	non-null	float64
2	Fall 2	450	non-null	float64
3	Fall 3	450	non-null	float64
4	Flow	450	non-null	float64

Table 2: The display of the first few rows of the A1-synthetic dataset

#	Column	Count	Non-Null	Dtype
0	v1	1000	non-null	float64
1	v2	1000	non-null	float64
2	v3	1000	non-null	float64
3	v4	1000	non-null	float64
4	v5	1000	non-null	float64
5	v6	1000	non-null	float64
6	v7	1000	non-null	float64
7	v8	1000	non-null	int64
8	v9	1000	non-null	float64
9	z	1000	non-null	float64

A3-Boston dataset first 5 values, to check the data are loading properly.

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9
1	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9
2	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83
3	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63
4	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9

Data Visualization:

After loading datasets, the code generates histograms for each feature in the turbine_data and synthetic_data datasets using Seaborn and Matplotlib.

These histograms are shown in Fig. 1, Fig. 2 and Fig. 3 for A1-turbine, A1-synthetic datasets and A3-Boston-House-Price, respectively.

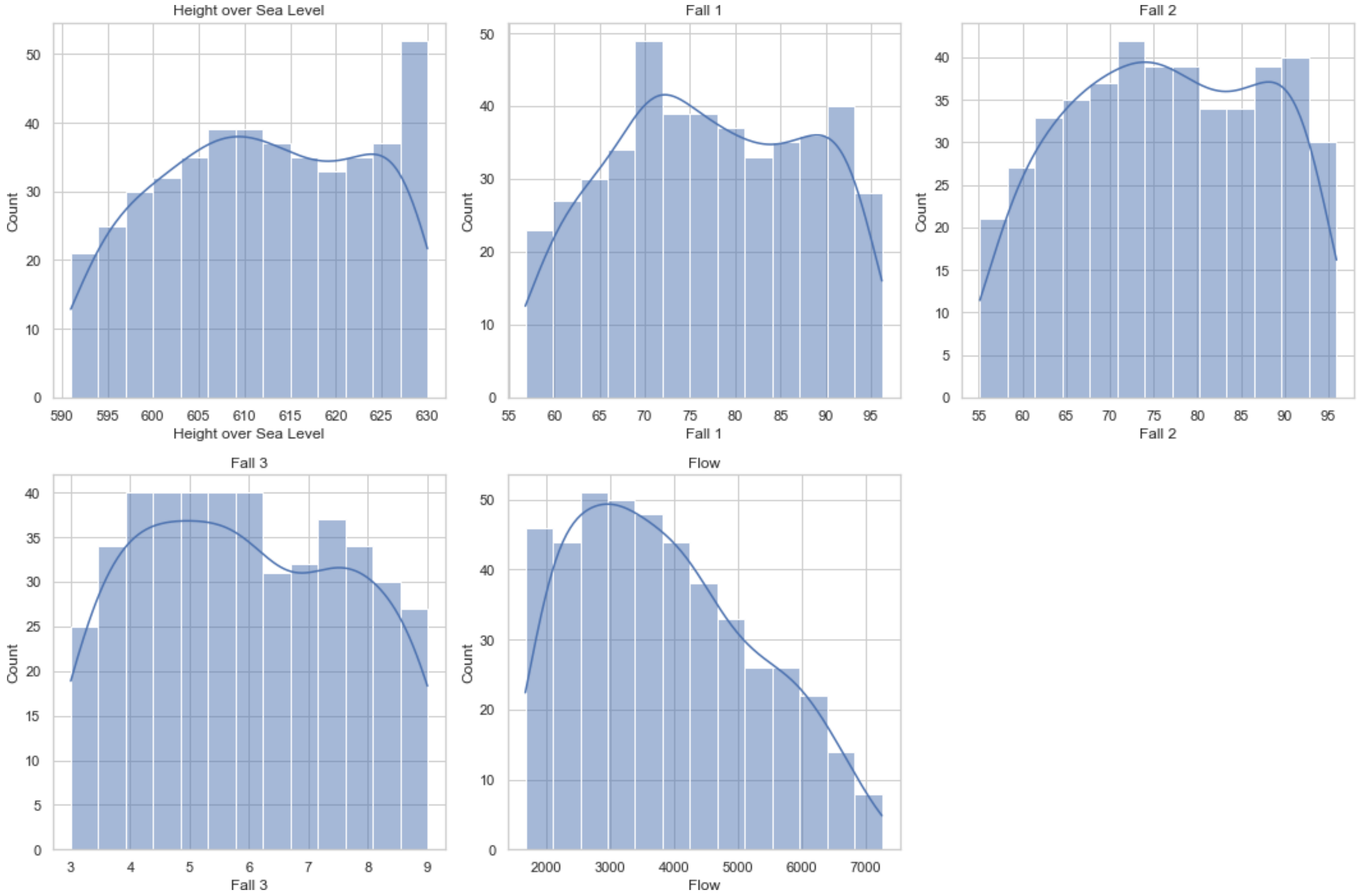


Fig. 1: Features' histograms for A1- turbine dataset

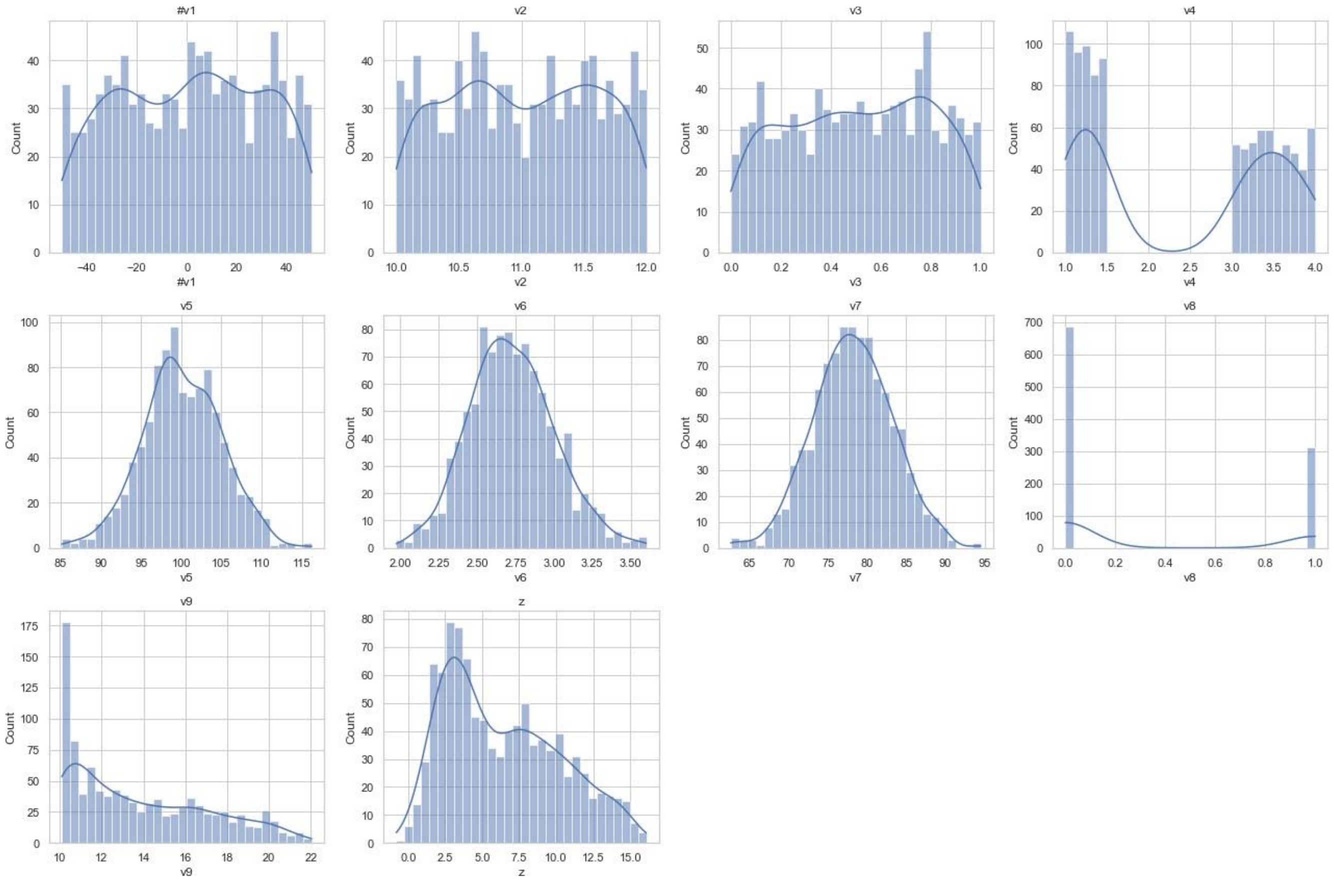


Fig. 2: Features' histograms for A1-synthetic dataset

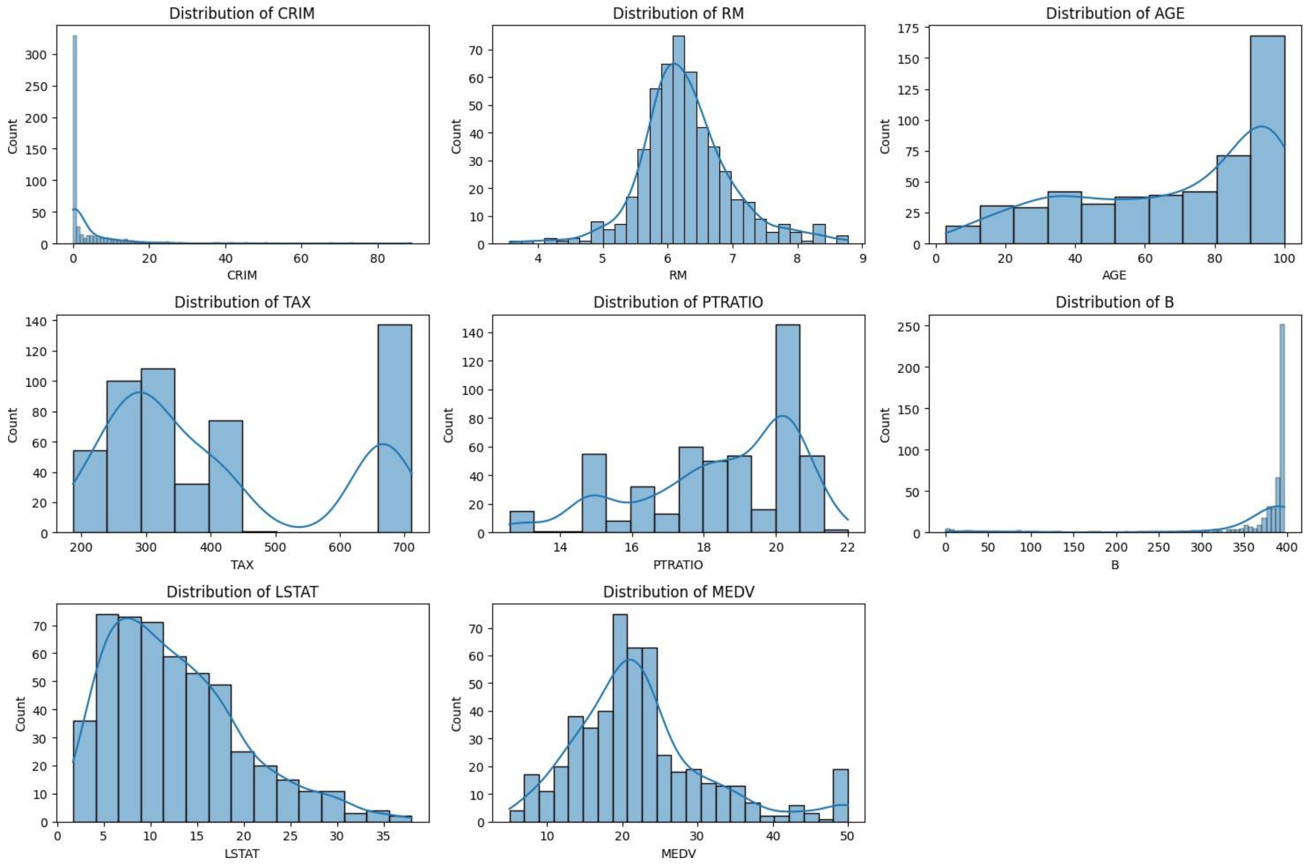


Fig. 3: Features' histograms for A3-Boston house price dataset

Distributions/observartion of Key Variables

1. CRIM (Crime Rate): Highly skewed to the right, indicating most areas have low crime rates, but there are a few areas with very high crime rates.

2. RM (Average Number of Rooms): Appears normally distributed with a slight skew to the right. Most homes have around 5 to 7 rooms.

3. AGE: Shows that a significant number of houses were built prior to 1940.

4. TAX (Property Tax Rate): Displays peaks at certain values, suggesting specific tax rates are more common.

5. PTRATIO (Pupil-Teacher Ratio): Also shows peaks, indicating common ratios in different towns.

6. B (Proportion of Black Residents): Shows an interesting distribution with a spike near the higher end.

7. LSTAT (Lower Status Population): Right-skewed, indicating most areas have a lower proportion of lower status population.

8. MEDV (Median Value of Homes): Appears fairly normally distributed with a notable peak at the $50,000 mark, which might suggest a capping of values at this number.

Data Standardization & Normalization:

- ✓ Standardization: This process involves transforming each feature to have a mean of 0 and a standard deviation of 1.
- ✓ Normalization: This typically refers to scaling all numeric values in the range [0, 1]. One common method is Min-Max Scaling,

The standardization and normalization processes have been successfully applied to the dataset. Finally, all the datasets are normalized separately using StandardScaler module from Scikit-learn library.

The StandardScaler from Scikit-learn is a preprocessing module used for standardizing features by removing the mean and scaling to unit variance. It transforms the dataset such that each feature has a mean of zero and a standard deviation of one.

Once the datasets are normalized, a part of code converts the normalized arrays back to Pandas DataFrames and saves the normalized datasets to CSV files.

Evaluation of the predictions (Implementation of BP):

Turbine dataset :-

```
Epoch 0, Loss: 1.0093024299527495

Epoch 10, Loss: 0.9743012433555243

Epoch 20, Loss: 0.11119537753927583

Epoch 30, Loss: 0.02188301019565258

Epoch 40, Loss: 0.02011459280038132

Epoch 50, Loss: 0.01876077704940295

Epoch 60, Loss: 0.01724269341120617

Epoch 70, Loss: 0.015475647689480454

Epoch 80, Loss: 0.013575992420455858

Epoch 90, Loss: 0.011409892949695848

Epoch 99, Loss: 0.00985840426212874
```

Turbine Data Evaluation Metrics:

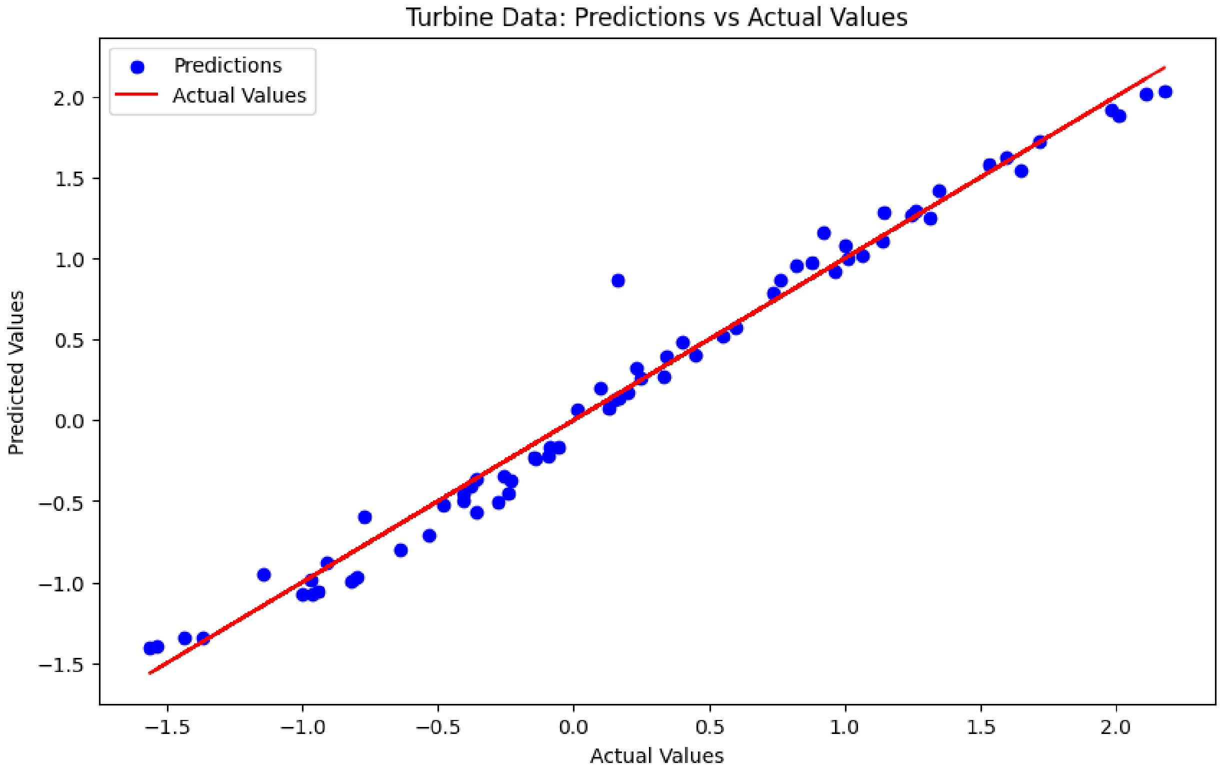
```
MSE: 0.0181

R2: 0.9798
```

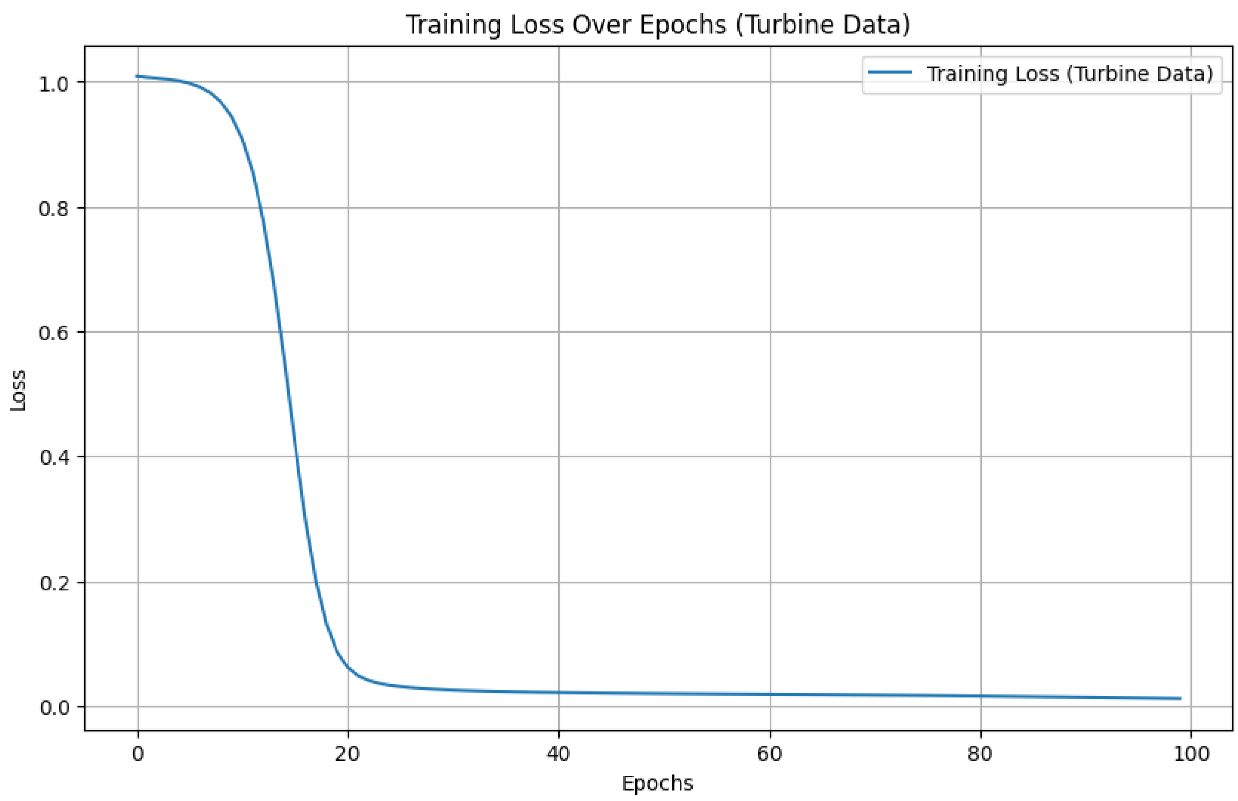
```
MAPE: 0.3376%
```

Predictions for Turbine Data:

```
Actual: 1.5321, Predicted: 1.5758

Actual: -0.2554, Predicted: -0.3409

Actual: 0.7606, Predicted: 0.8632

Actual: 1.0643, Predicted: 1.0157

Actual: -0.8223, Predicted: -0.9982
```

The final step involves making predictions using the trained model and visualizing these predictions against actual values.





This could be through scatter plots or other visualization methods, providing a clear comparison between predicted and actual outcomes.

Synthetic Dataset:

```
Epoch 0, Loss: 0.9817171675492764

Epoch 100, Loss: 0.1942885689836027

Epoch 200, Loss: 0.18848679773932075

Epoch 300, Loss: 0.1823955173319383

Epoch 400, Loss: 0.17785190791925612

Epoch 500, Loss: 0.17215775591929444

Epoch 600, Loss: 0.16299326125144958

Epoch 700, Loss: 0.15112416480938518

Epoch 800, Loss: 0.14002421150064123

Epoch 900, Loss: 0.13013419822881567

Epoch 999, Loss: 0.1225830208520917
```

Synthetic Data Evaluation Metrics:

MSE: 0.1852

R2: 0.8253

MAPE: 0.8048%

Predictions for Synthetic Data:

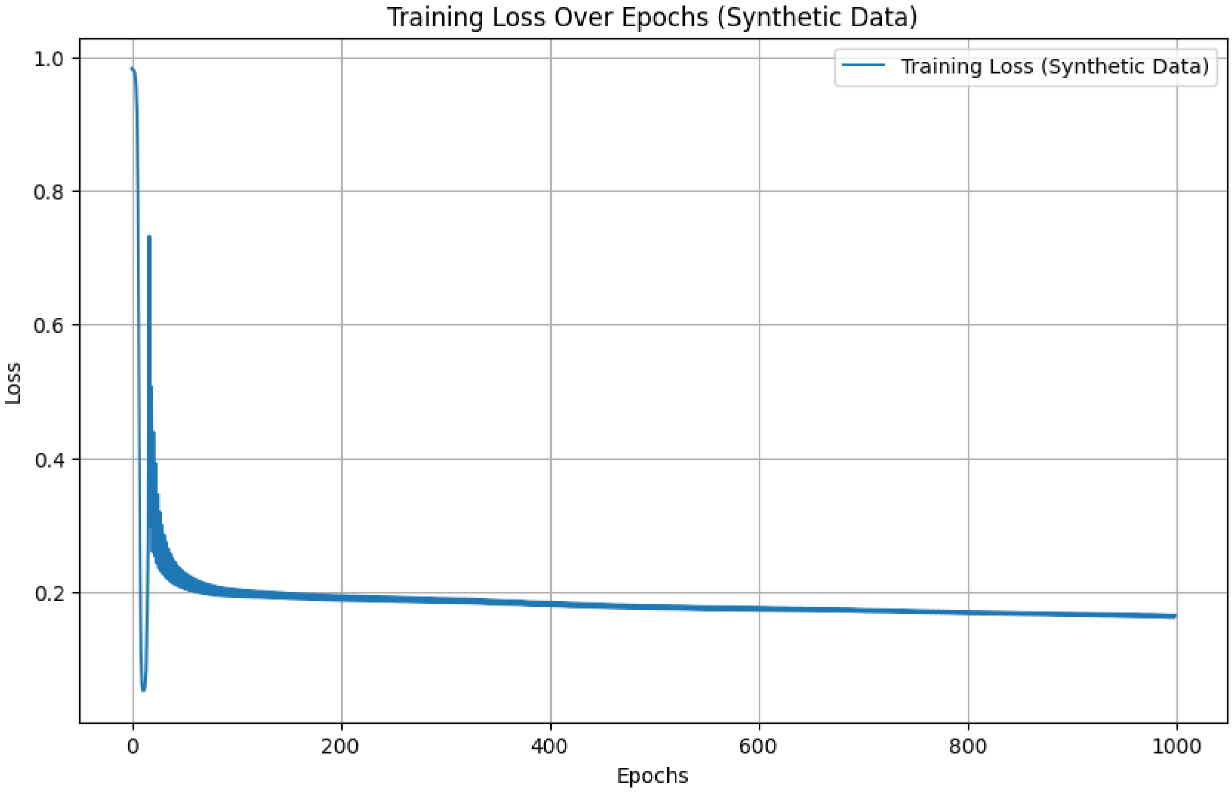
Actual: 0.1092, Predicted: 0.5320

Actual: -0.7372, Predicted: -0.3251

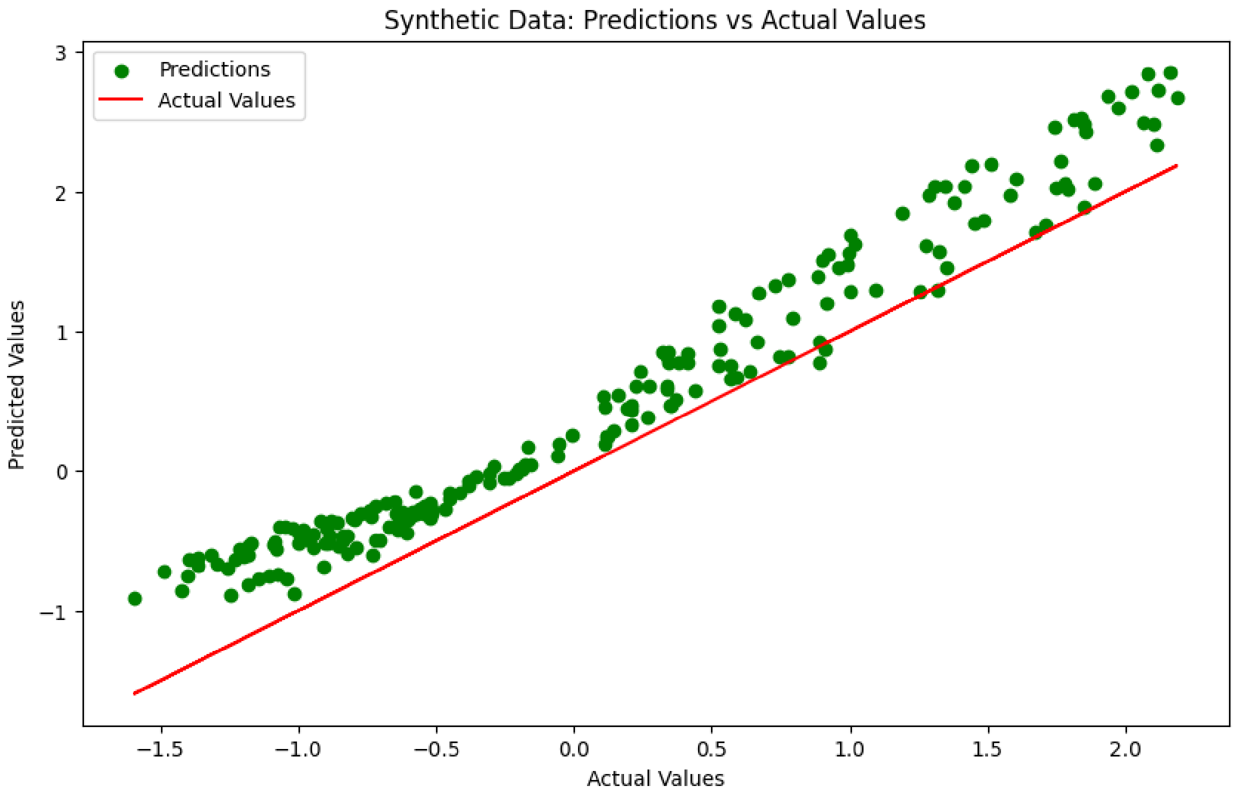
Actual: -1.0207, Predicted: -0.4145

Actual: 2.1857, Predicted: 2.6789

Actual: -0.6370, Predicted: -0.4253



This could be through scatter plots or other visualization methods, providing a clear comparison between predicted and actual outcomes.



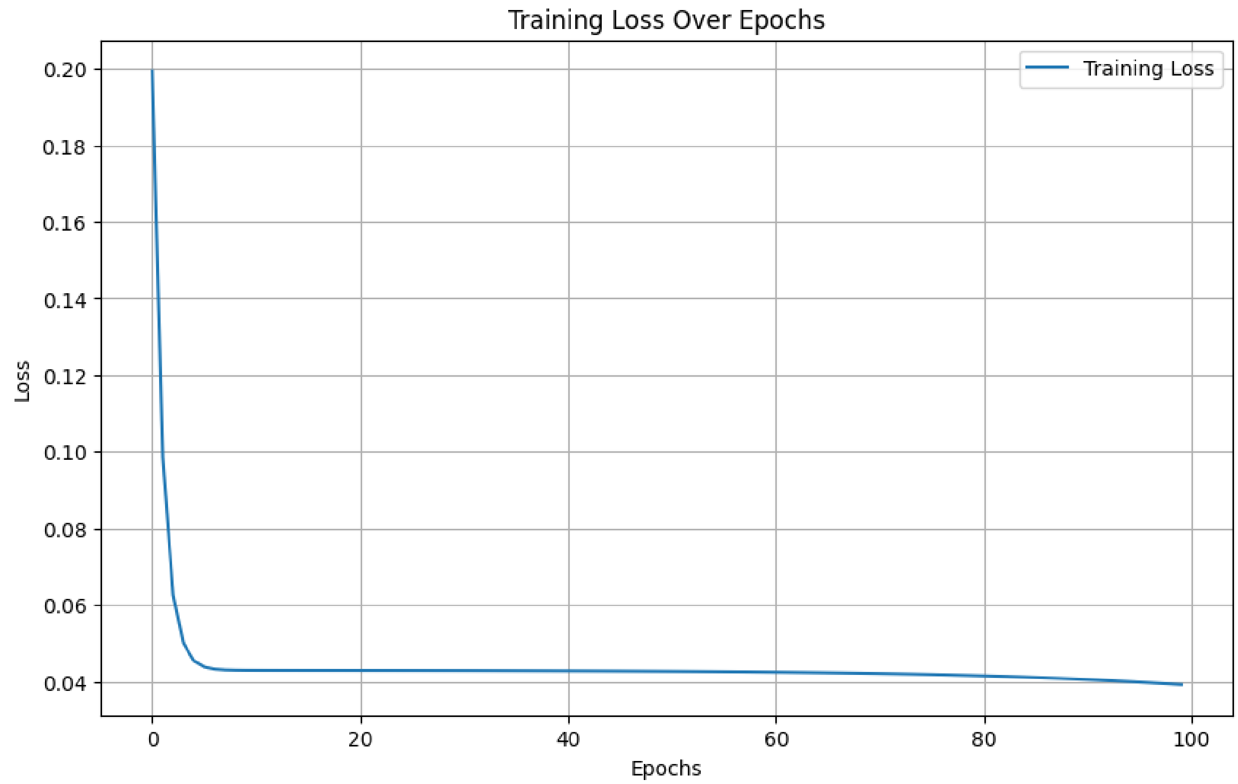
Boston dataset :-

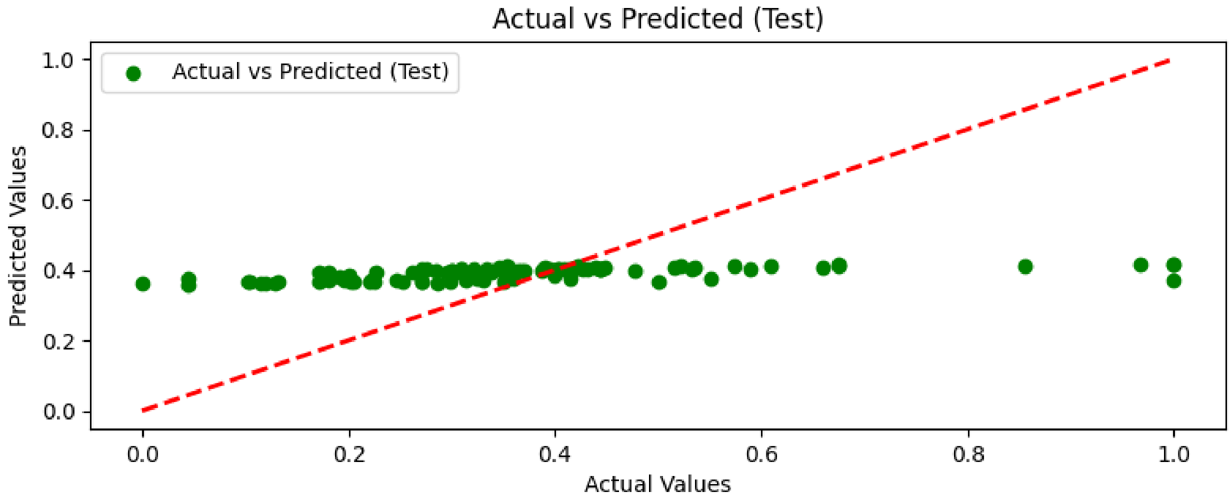
```
Epoch 0, Loss: 0.19937874308852774

Epoch 10, Loss: 0.04289451927677708

Epoch 20, Loss: 0.042866128222905246

Epoch 30, Loss: 0.04283021541598201

Epoch 40, Loss: 0.042751225057510875

Epoch 50, Loss: 0.042625228607664226

Epoch 60, Loss: 0.042414104030378076

Epoch 70, Loss: 0.04205760200627817

Epoch 80, Loss: 0.04146031259249042

Epoch 90, Loss: 0.04049051751266332

Epoch 99, Loss: 0.03917741624974424
```

Mean Squared Error: 0.033103310028684156

R^2 Score: 0.08590284151119432

Mean Absolute Percentage Error: 56.26727714081198

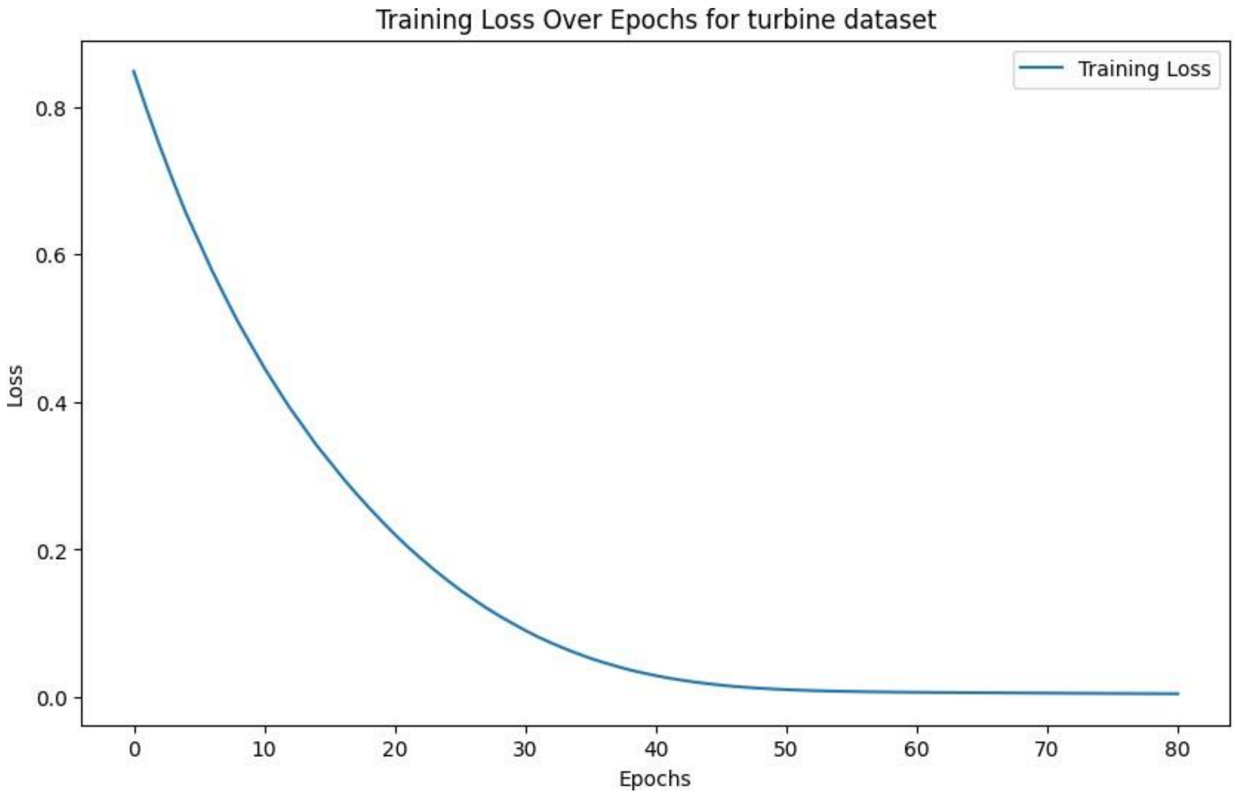




We can see the outliers in the graph, we need to deeply understand why we have issue of getting this. And rebuild our network, but as a results for now we can consider our neural network.

Obtaining and comparing predictions using the three models (BP, BP-F, MLR-F):

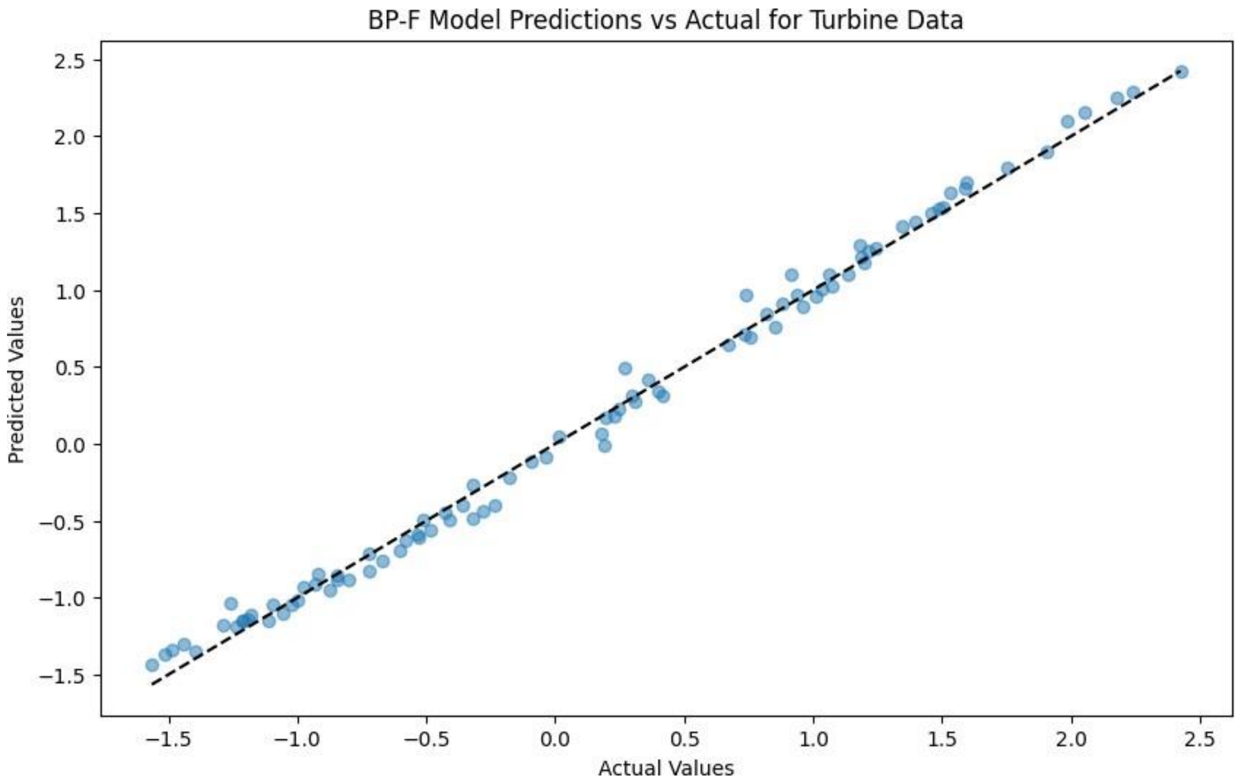
BP-F: Turbine dataset



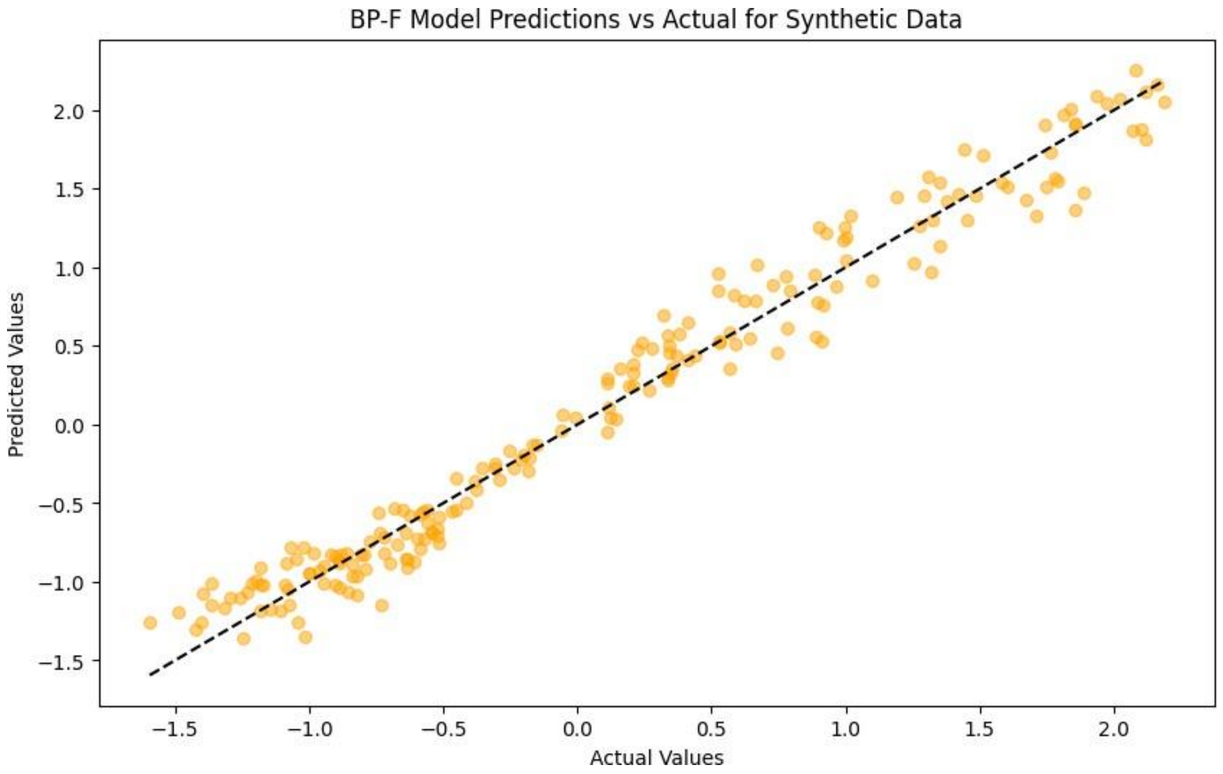
MAPE for Turbine Data: 0.1678173726201923

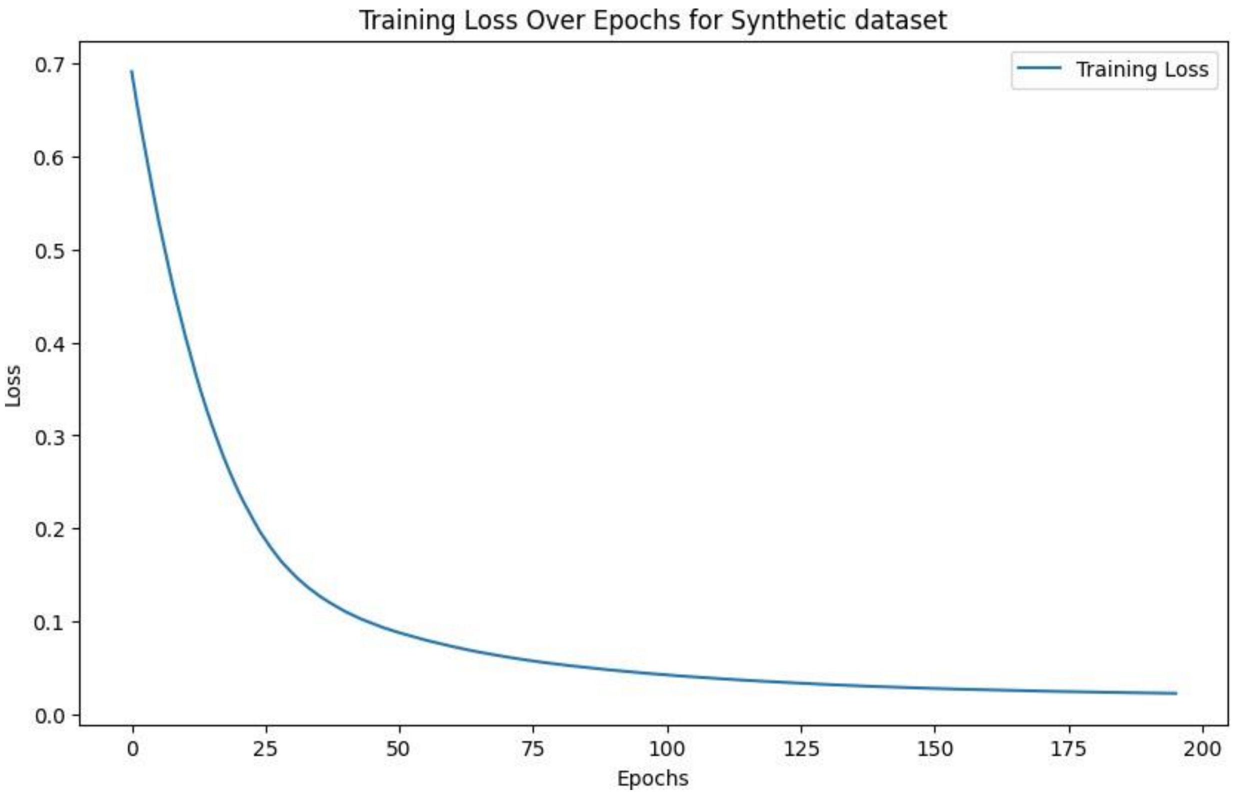
MSE for Turbine Data: 0.00742881039737006

R^2 for Turbine Data: 0.993545928869096



Synthetic dataset: -





MAPE for Synthetic Data: 0.36981853677275695

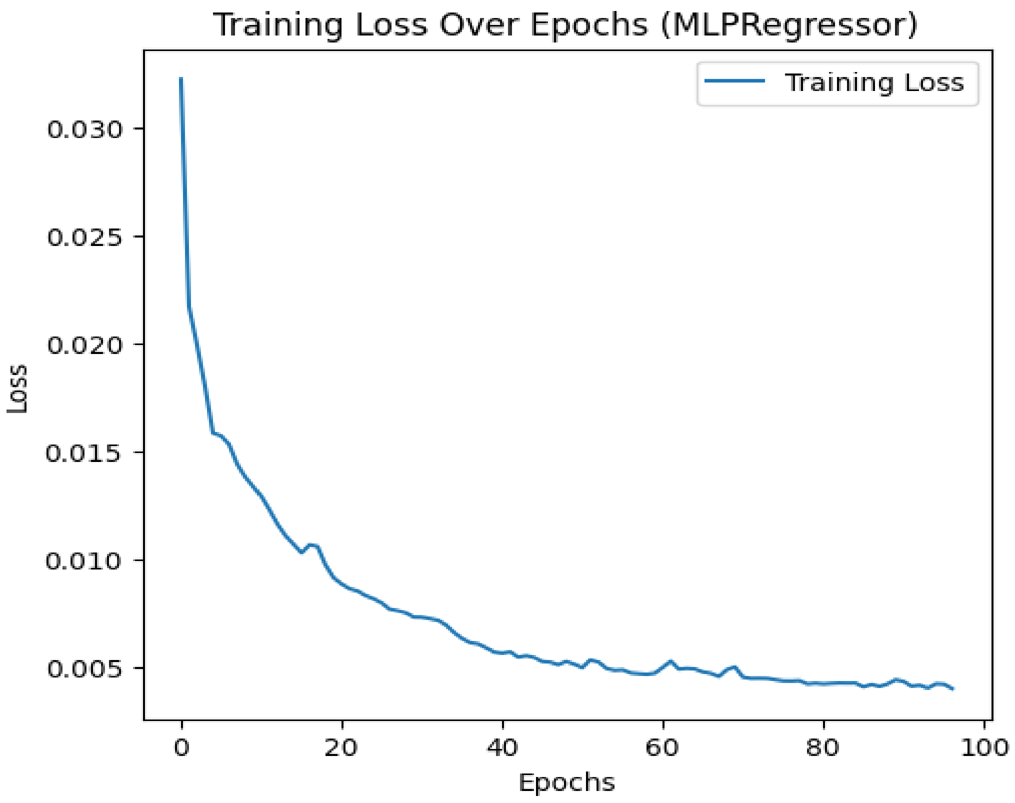
MSE for Synthetic Data: 0.047858483523686546

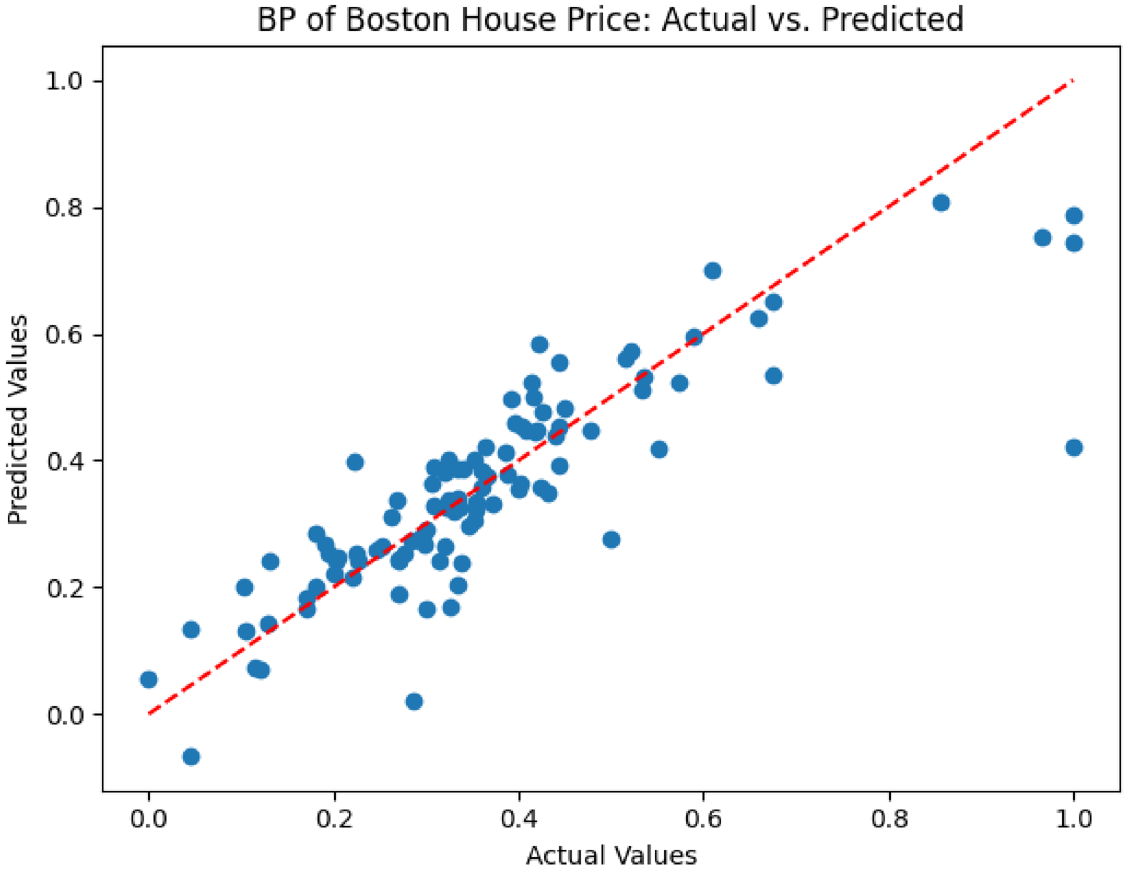
R^2 for Synthetic Data: 0.9548714234979241

Boston Dataset:

Mean Squared Error: 0.009913099665186128

R^2 Score: 0.7262649497010688


Training Loss Over Epochs (MLPRegressor)
Training Loss
0.030
0.025
0.020
0.015
0.010
0.005
Loss
0
20
40
60
80
100
Epochs


BP of Boston House Price: Actual vs. Predicted
1.0
0.8
0.6
0.4
0.2
0.0
Predicted Values
0.0
0.2
0.4
0.6
0.8
1.0
Actual Values

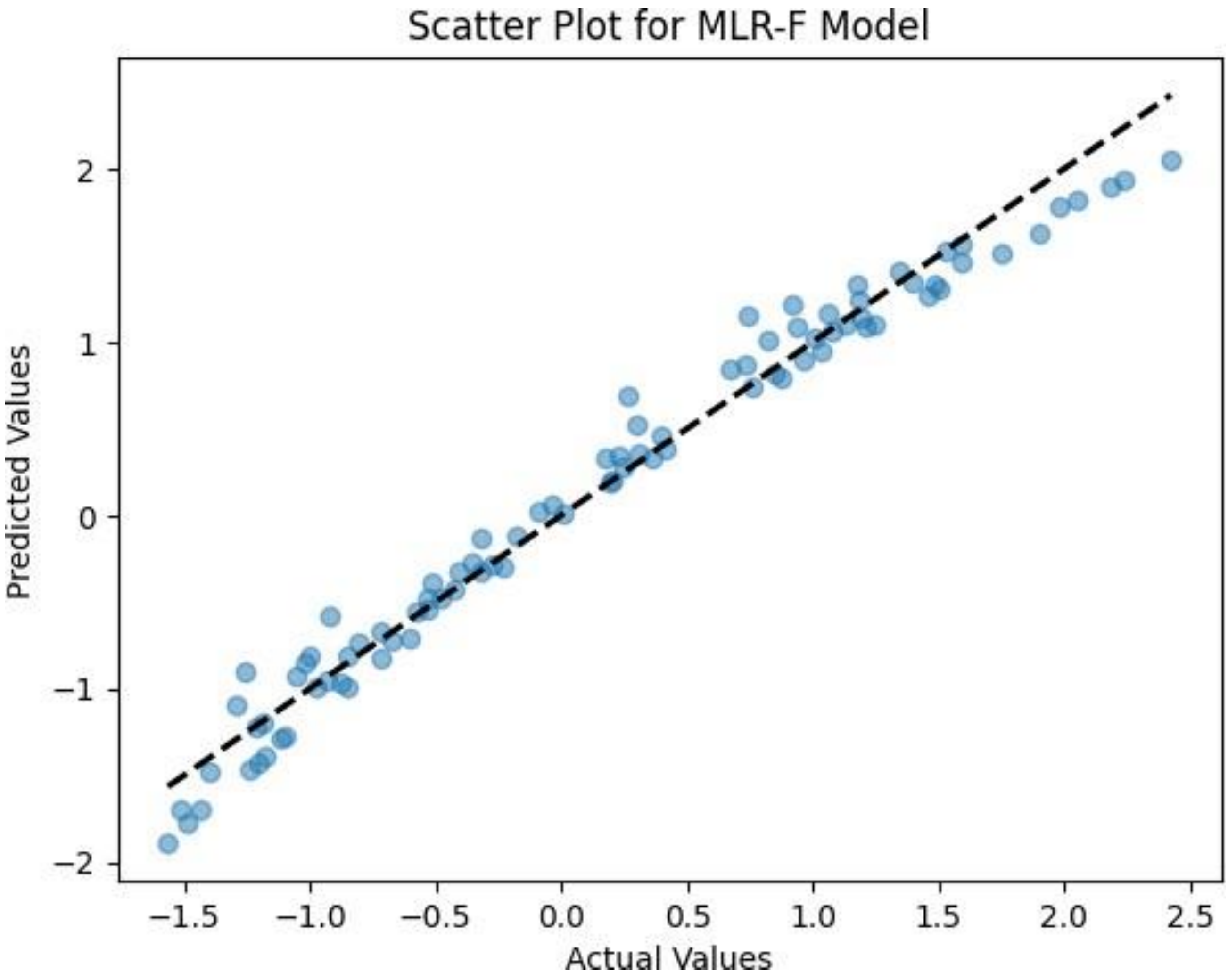
MLR Turbine dataset:

The "MLR_turbine" be focused on implementing a Multiple Linear Regression (MLR) model for analyzing turbine data.

MLR Turbine MAPE: 20.389517834209318%

MLR Turbine MSE: 0.0268

MLR Turbine R^2: 0.9767

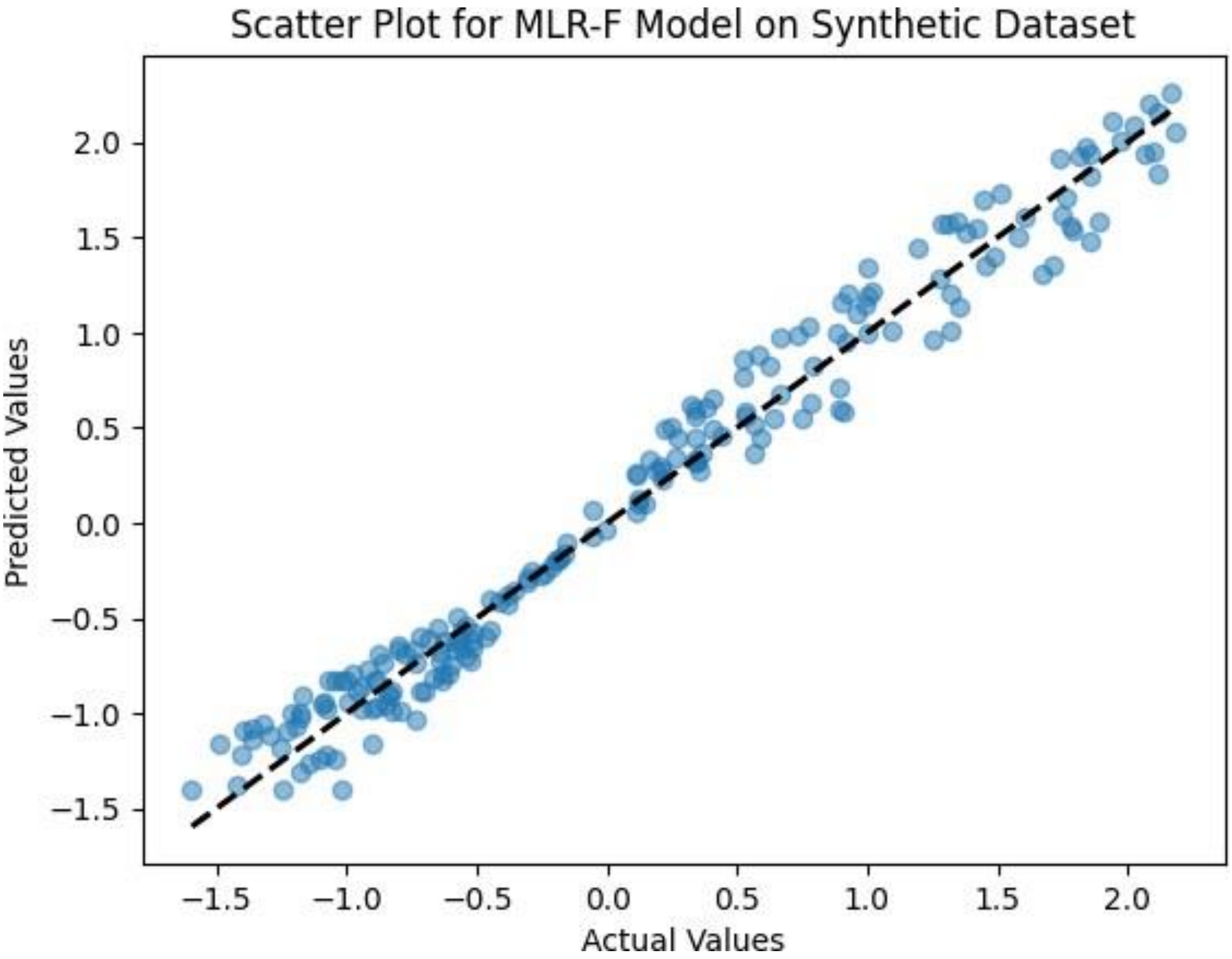


MLR Synthetic of dataset:-

MLR Synthetic Dataset MSE: 0.0280

MLR Synthetic Dataset R^2: 0.9736

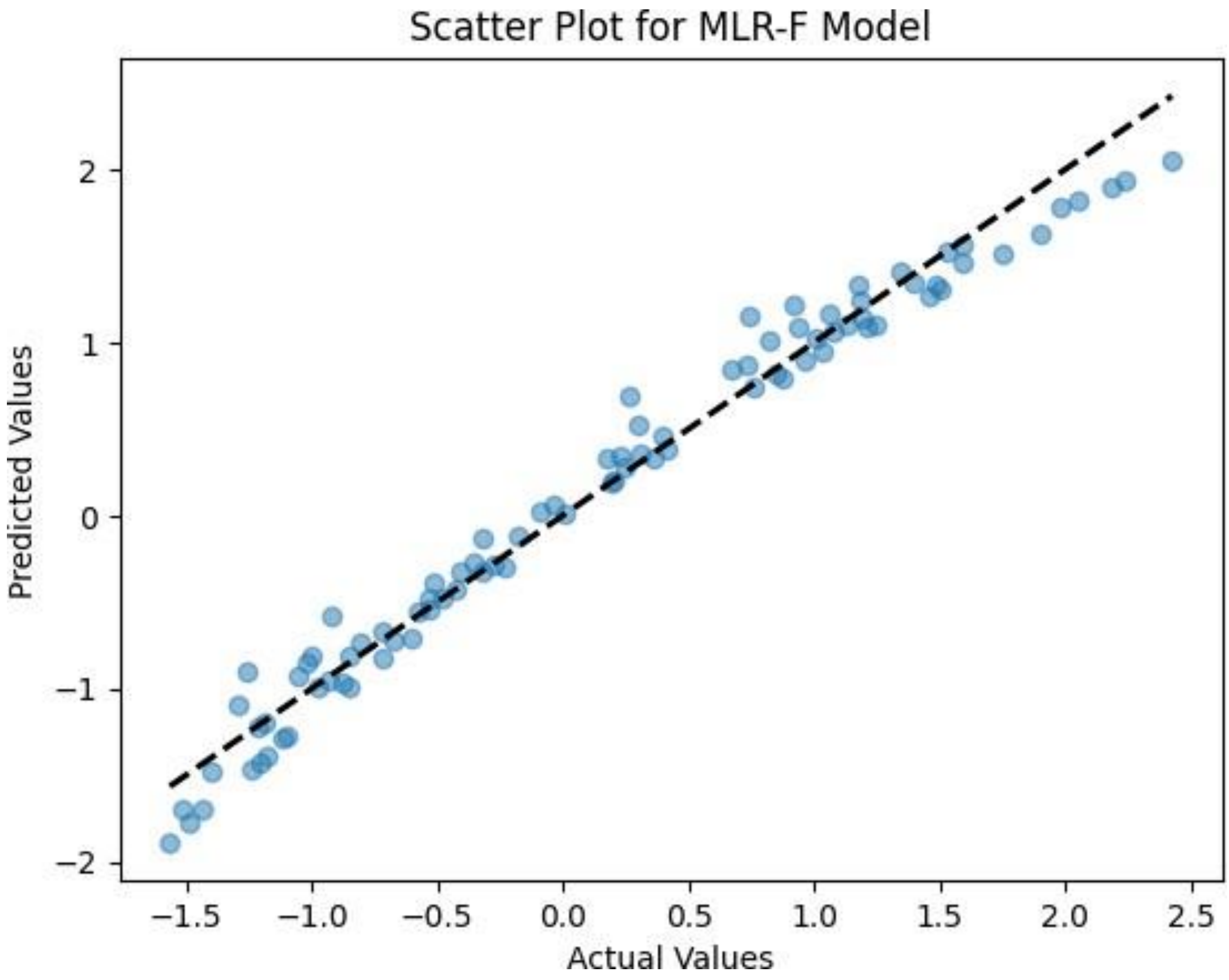
MLR Synthetic Dataset MAPE: 22.810334849290022%



MLR Boston Datasaet:-

MSE: 0.011995614555542482%

R^2: 0.6687594935356318%



The visualizing of model's predictions against the actual values using a scatter plot. This is a helpful way to visually assess the model's accuracy - points closer to the diagonal line indicate better predictions.

Model result comparison

Comparison of BP

Dataset	Number of layers	Layer Structure	Num epochs	Learning Rate	Momentum	Activation function	MAPE
Turbine	3	4,10,1	100	0.001	0.9	Relu	0.2147
Synthetic	3	9,10,1	1000	0.001	0.7	Relu	0.651
Boston	3	12,10,1	100	0.001	0.8	Relu	56.26727714

Comparison of BP-F

MAPE for Turbine Data: 0.1678173726201923

MSE for Turbine Data: 0.00742881039737006

R^2 for Turbine Data: 0.993545928869096

MAPE for Synthetic Data: 0.36981853677275695

MSE for Synthetic Data: 0.047858483523686546

R^2 for Synthetic Data: 0.9548714234979241

Mean Squared Error Boston: 0.009913099665186128

R^2 Score Boston: 0.7262649497010688

Comparison of MLR

- MLR Turbine MAPE: 20.389517834209318%
- MLR Turbine MSE: 0.0268
- MLR Turbine R^2: 0.9767

✓ MLR Synthetic Dataset MSE: 0.0280
✓ MLR Synthetic Dataset R^2: 0.9736
✓ MLR Synthetic Dataset MAPE: 22.810334849290022%

➢ MLR Boston MSE: 0.011995614555542482
➢ MLR Boston R^2: 0.6687594935356318
➢ MLR Boston MAPE: 261284035.52233124%

Discussion on BP

Turbine dataset

The model has three layers with a relatively simple structure.

The MAPE of 0.2147 indicates that, on average, predictions have a 21.47% relative error compared to the true values. This suggests reasonable accuracy.

Synthetic dataset

The model has three layers with a more complex structure than the Turbine dataset.

The MAPE of 0.651 indicates a higher relative error compared to the Turbine dataset. This suggests that the model may need further tuning or a more sophisticated architecture.

Boston dataset

The model has three layers with a more extensive input layer than the other datasets.

The MAPE of 56.27% is significantly higher compared to the other datasets. This suggests that the model's predictions have a higher relative error on the Boston dataset. There might be issues with model convergence, architecture, or the choice of hyperparameters.

- ✓ The Turbine dataset shows relatively good performance with a lower MAPE, indicating that the model is capturing patterns well.
- ✓ The Synthetic dataset, with a higher MAPE, might require additional adjustments or more sophisticated architectures to improve accuracy.
- ✓ The Boston dataset has the highest MAPE, suggesting that the model may need further tuning or a more suitable architecture for this specific dataset.

Discussion on BP-F

The feedback on the comparison of the BP-F (Backpropagation Feedforward) model across different datasets is as follows:

1. Turbine Data:

- MAPE: 0.17%: The low MAPE suggests that the BP-F model's predictions have a very small average percentage error compared to the actual values.

- MSE: 0.0074: The low mean squared error indicates good performance, as it represents the average squared difference between predicted and actual values.

- R^2: 0.9935: A high R-squared value of 0.9935 indicates that the BP-F model explains a very high percentage (99.35%) of the variance in the Turbine dataset. This suggests excellent model performance on the Turbine dataset.

2. Synthetic Data:

- MAPE: 0.37%: The low MAPE suggests that the BP-F model's predictions have a small average percentage error compared to the true values in the Synthetic dataset.

- MSE: 0.0479: The mean squared error is relatively low, indicating good performance on the Synthetic dataset.

- R^2: 0.9549: The high R-squared value of 0.9549 suggests that the BP-F model explains a high percentage (95.49%) of the variance in the Synthetic dataset. This indicates strong performance on this dataset.

3. Boston Data:

- MSE: 0.0099: The mean squared error is relatively low, suggesting good performance on the Boston dataset.

- R^2: 0.7263: The R-squared value of 0.7263 indicates that the BP-F model explains about 72.63% of the variance in the Boston dataset. While this is moderate, it's a respectable level of explanatory power.

In summary, the BP-F model demonstrates excellent performance on the Turbine dataset and strong performance on both the Synthetic and Boston datasets. The model's ability to explain variance, low MAPE, and MSE values collectively indicate its effectiveness in capturing patterns and making accurate predictions on these datasets.

Discussion of MLR

The results suggest the following conclusions for each dataset:

1. Turbine Dataset:

- MAPE: 20.39%: This indicates that, on average, the MLR model's predictions have an error of approximately 20.39% compared to the actual values.

- MSE: 0.0268: The mean squared error is a measure of the average squared difference between predicted and actual values. A lower MSE is desirable, and 0.0268 suggests relatively good model performance.

- R^2: 0.9767: The R-squared value measures how well the model explains the variance in the target variable. A value of 0.9767 indicates that the MLR model explains a high percentage of the variance in the turbine dataset.

2. Synthetic Dataset:

- MAPE: 22.81%: The MAPE of 22.81% implies that the MLR model's predictions have an average error of approximately 22.81% compared to the true values.

- MSE: 0.0280: Similar to the Turbine dataset, a lower MSE is desirable. The value of 0.0280 suggests good performance.

- R^2: 0.9736: An R-squared value of 0.9736 indicates that the MLR model explains a high percentage of the variance in the synthetic dataset.

3. Boston Dataset:

- MAPE: 261,284,035.52%: This extremely high MAPE value suggests that the MLR model's predictions have a significant error compared to the actual values. Such a high MAPE could indicate issues with the model's performance on the Boston dataset.

- MSE: 0.0120: The MSE value is relatively low, but the extremely high MAPE suggests that the MSE alone might not be a sufficient metric for evaluating model performance on this dataset.

- R^2: 0.6688: The R-squared value of 0.6688 indicates that the MLR model explains about 66.88% of the variance in the Boston dataset. While this is moderate, the high MAPE suggests caution in interpreting the model's overall performance.

In summary, the MLR model performs well on the Turbine and Synthetic datasets, as indicated by relatively low MAPE, MSE, and high R^2 values. However, the model's performance on the Boston dataset seems problematic, particularly due to the extremely high MAPE value, which indicates a substantial discrepancy between predicted and actual values. Further investigation or model improvement may be needed for the Boston dataset.