COL 761

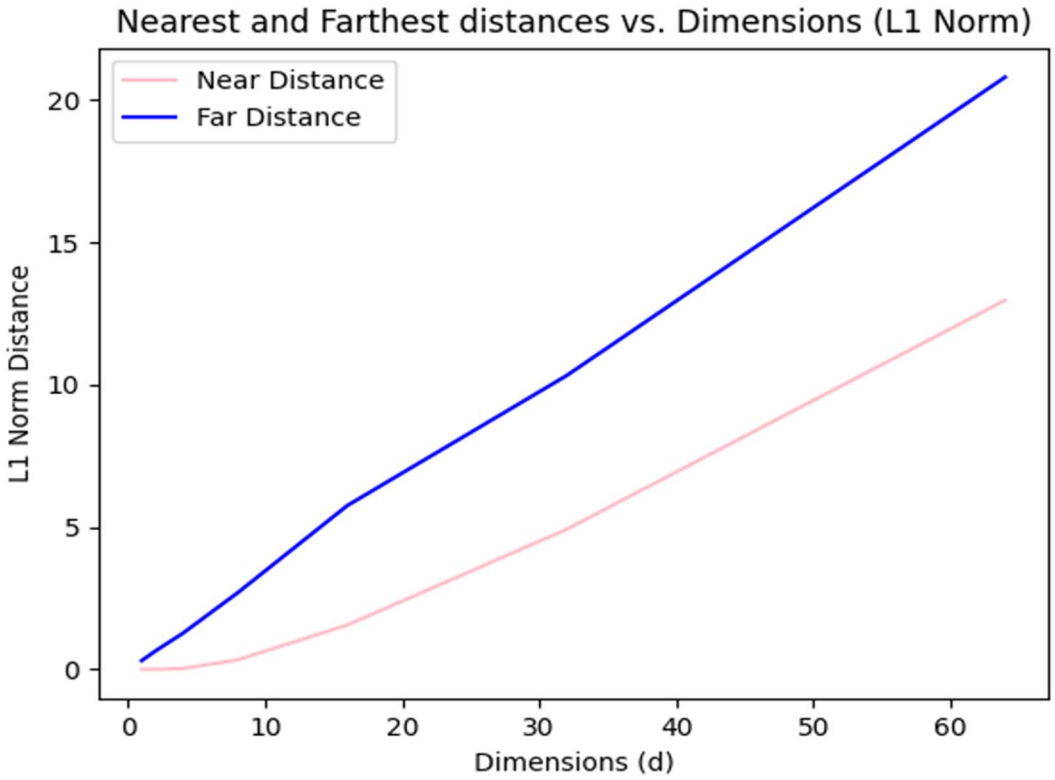
DATA MINING

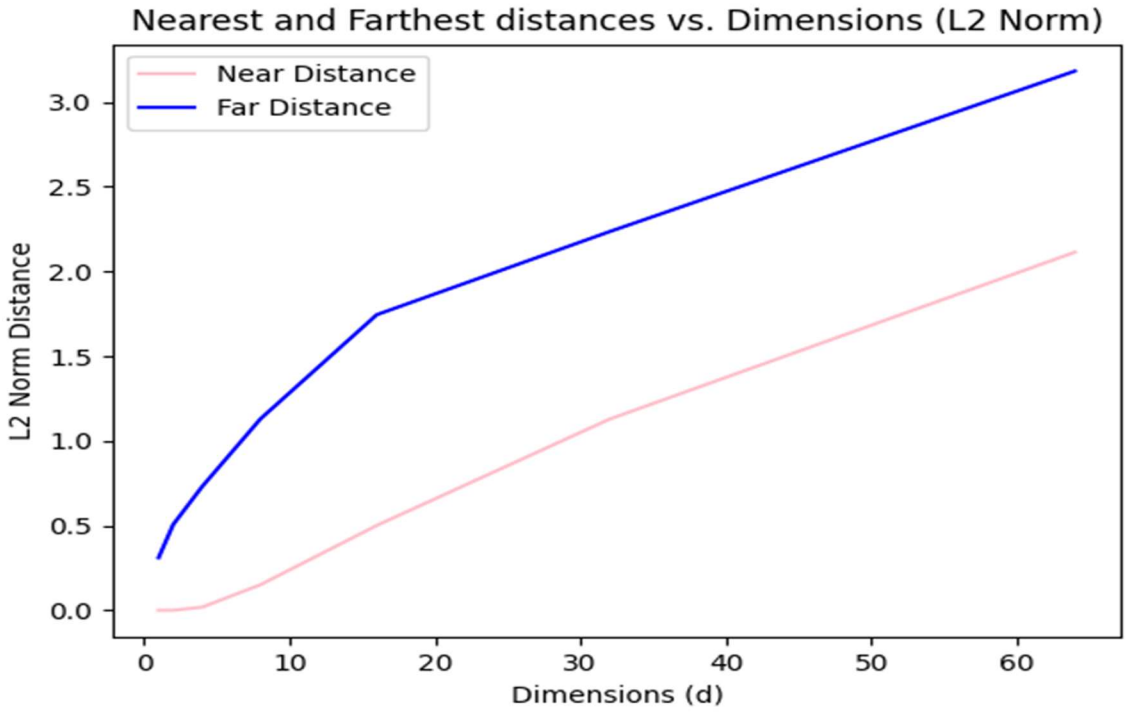
Assignment – 3

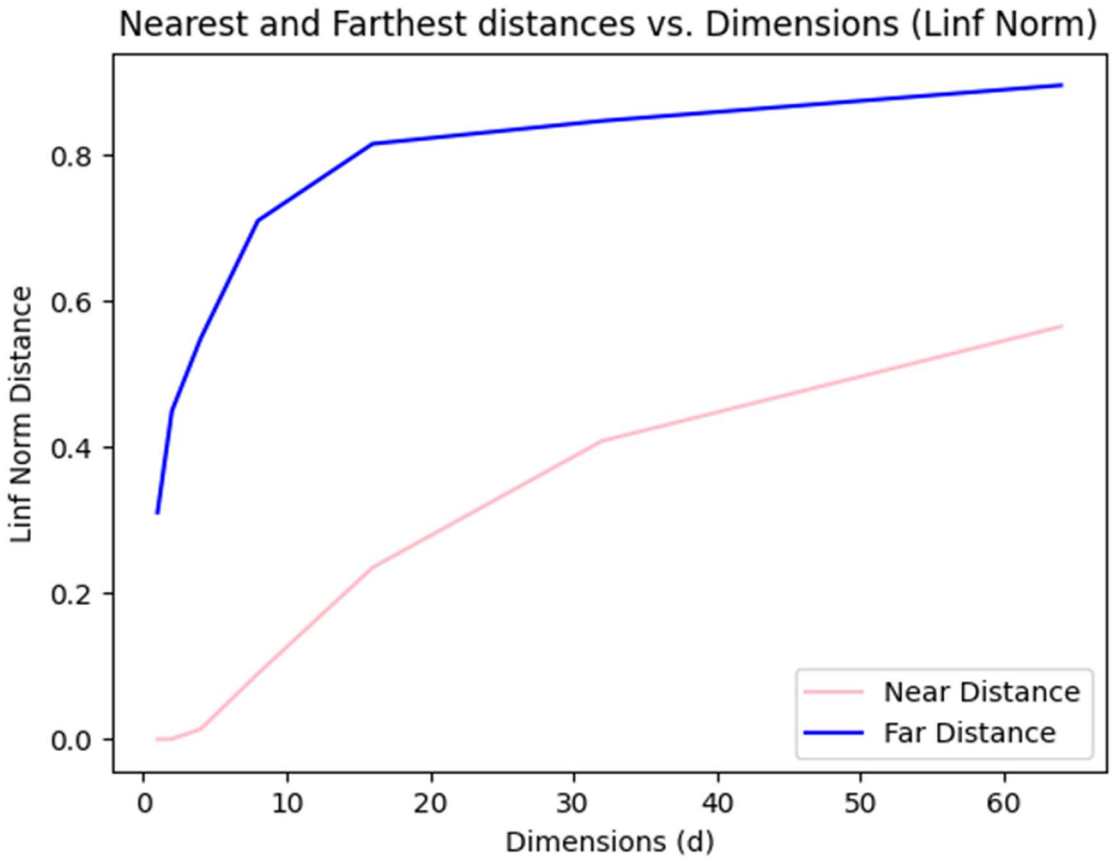
TEAM MEMBERS:

1. Banoth Chethan Naik - 2020CS10333
2. Perugu Krishna Chaitanya Yadav - 2020EE10521
3. Devarakonda Ronith Kumar - 2020EE10486

Q1. Uniformly Distributed Points in High-Dimensional Spaces







- **Curse of Dimensionality:**
  As the number of dimensions increases, data points become more spread out, leading to data sparsity.
- **Distance Measures:**
  L1 Distance (Manhattan Distance): Measures the absolute sum of distances in each dimension. With more dimensions, L1 distance increases due to the cumulative effect of absolute differences.
  Reasons Behind Increase with Dimensions: As the number of dimensions increases, the cumulative effect of absolute differences along each dimension contributes to a rapid increase in L1 distance. Each dimension adds to the overall distance, making L1 particularly sensitive to the sparsity introduced by the

curse of dimensionality. Outliers in any dimension have a pronounced impact on the total L1 distance.

Linf Distance (Chebyshev Distance): Represents the maximum distance in any dimension. In higher dimensions, this distance becomes nearly constant since it only considers the maximum difference along a single dimension.

Reasons Behind Constancy with Dimensions: In higher dimensions, the overall spread of data points increases. However, Linf distance is primarily influenced by the maximum difference along a single dimension. This means that, despite the increased overall spread, the maximum difference becomes a dominating factor, leading to a more constant distance measure.

L2 Distance (Euclidean Distance): Falls between L1 and Linf distances. It is less affected by dimensionality compared to L1, providing a balance between the extremes.

Reasons Behind Balanced Response: L2 distance offers a balanced response to dimensionality changes. The squared sum of differences provides a middle ground, distributing the impact more evenly across dimensions. This results in a less exaggerated response to extreme values in individual dimensions compared to L1 distance.

- **Observation from Plots:**

  Visualizations of the data reveal that L1 distance increases with dimensionality, while Linf distance remains constant, and L2 distance offers a middle ground. This aligns with the intuition derived from the curse of dimensionality, where increased dimensions lead to more uniform distances between data points.

Q2 Graph Classification and Regression Using Graph Neural Networks

Task – 1: Classification

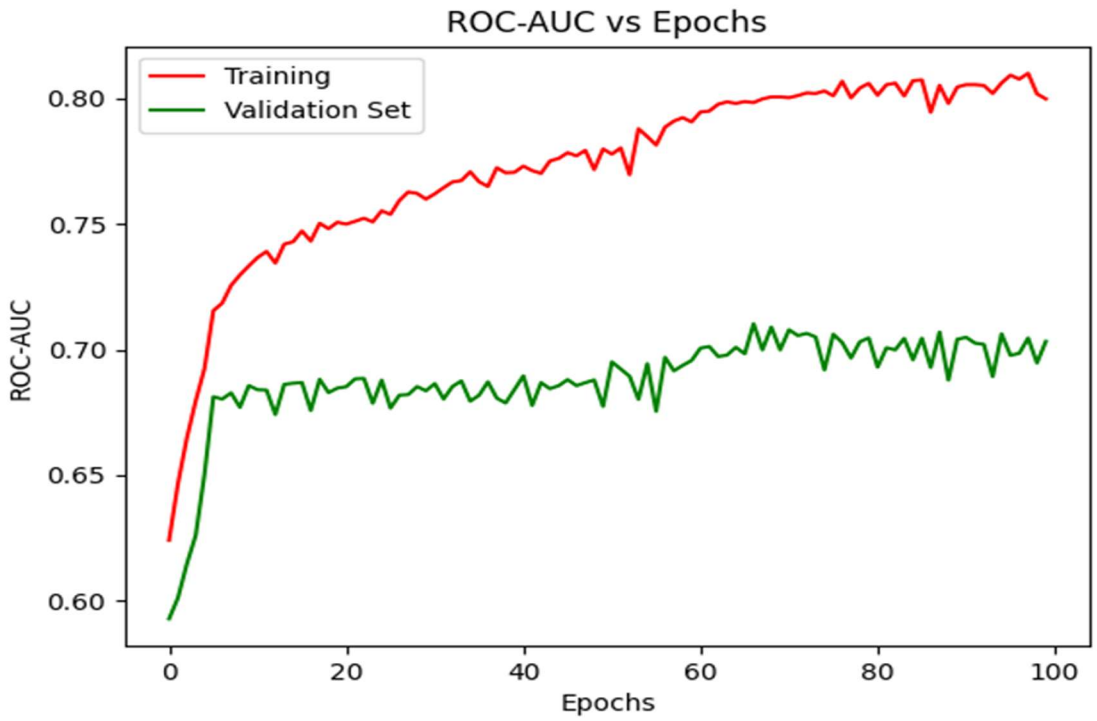
Model Design and Training:

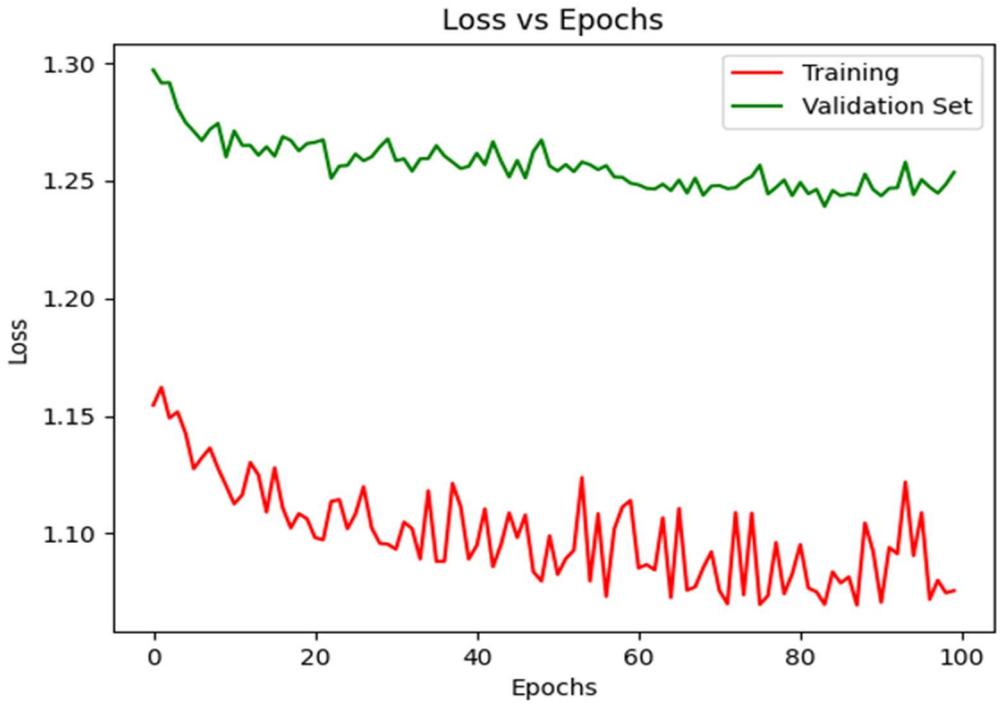
- Utilizing a provided encoder, we encode nodes into a 16-dimensional space and edges into a 2-dimensional space.
- The model architecture incorporates three Graph Attention Network (GAT) layers, each with a hidden dimension of 32.
- Node embeddings from all three layers are concatenated, creating a comprehensive representation for each node.
- The graph's embedding is generated by applying mean pooling to the node embeddings.
- A linear layer is then applied to produce a real value, and its sigmoid represents the probability of belonging to class 1.
- Binary Cross-Entropy (BCE) Loss serves as the loss function, and optimization is carried out using the Adam Optimizer with a learning rate of 0.01.
- Addressing the dataset's bias toward class 0 graphs, a weighted loss function is employed to enhance learning effectiveness.
- Training focuses on minimizing BCE Loss, and the model saving criterion is based on achieving the best validation ROC-AUC score.

Comparisons With Baselined Implementations:

ROC-AUC	Random Model	Logistic Regression Model	Implemented GNN Model
Training	0.52	0.59	0.79
Validation	0.47	0.54	0.68

Training and Validation learning curves:



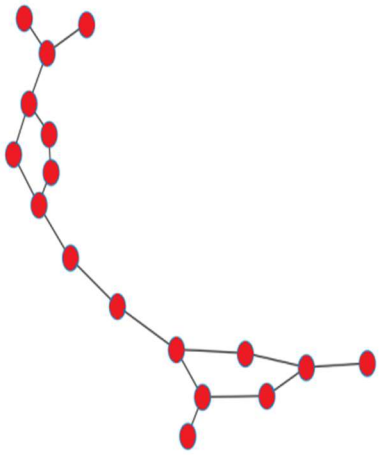


- ROC-AUC scores for the GNN model in both training and validation phases are notably higher (0.79 and 0.68, respectively) compared to Logistic Regression (0.52 and 0.47) and the Random Model (0.59 and 0.54).

- This highlights the superior predictive performance of the GNN model, demonstrating its effectiveness in capturing complex relationships within the graph data compared to the baseline models.

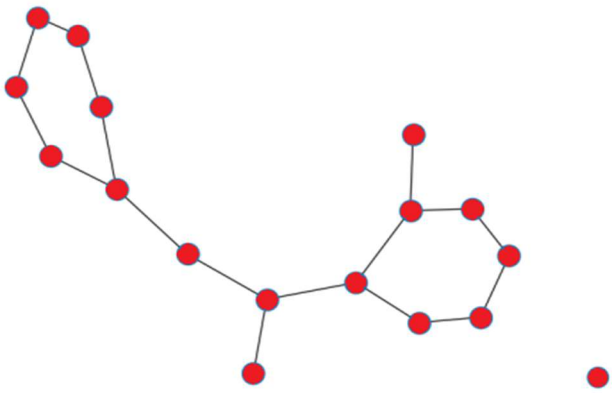
- Our approach demonstrates superior performance compared to the other baseline methods. Examining the learning curves reveals a consistent upward trend in ROC-AUC as the number of epochs increases.

Graph Visualization with NetworkX: Identifying Model Misclassifications and Poor Performance



Pred : 1 Real : 0

Pred : 0 Real : 1



Analysis: The graphs depicted above exhibit numerous clusters of nodes, with limited or absent connections between them. Our model faces challenges in accurately predicting the true class for these specific graph configurations.

Task – 2: Regression

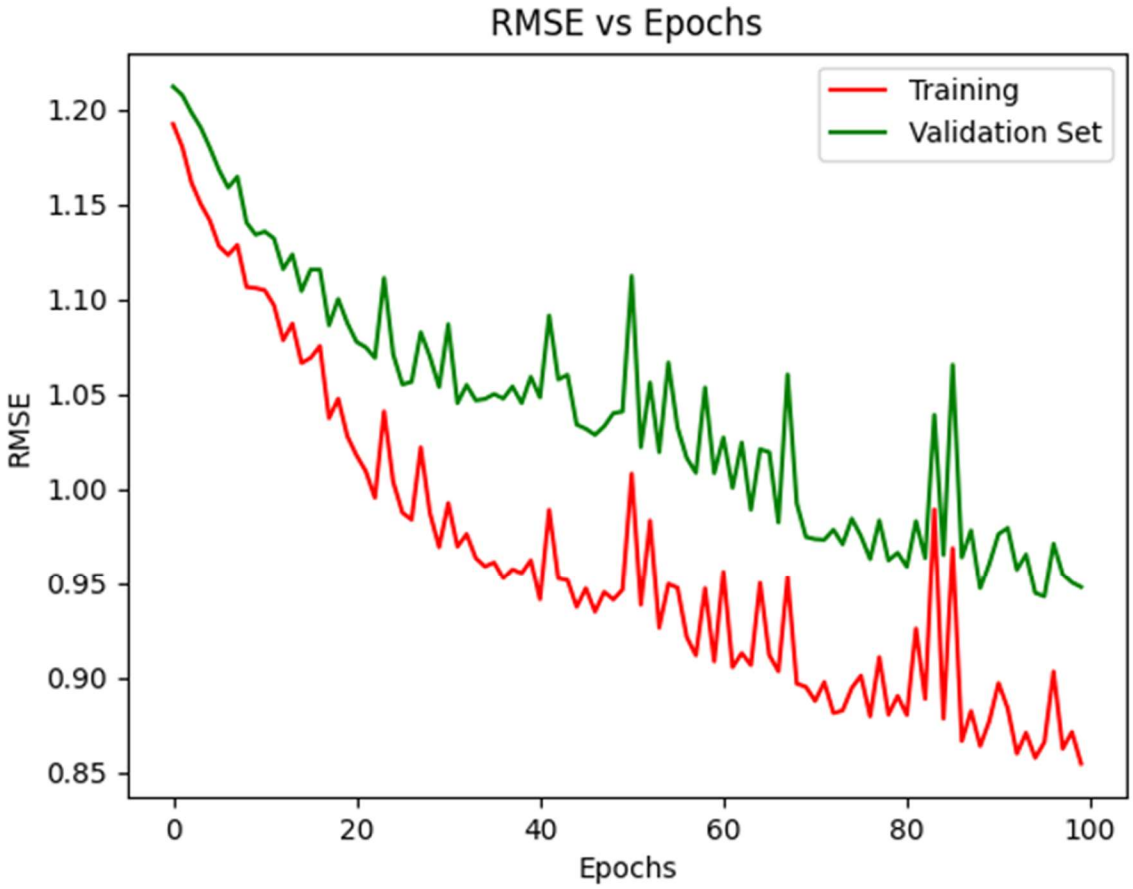
Model Design and Training:

- Nodes and edges undergo encoding using a provided encoder, with nodes represented in a 16-dimensional space and edges in a 2-dimensional space.
- The model structure integrates three Graph Attention Network (GAT) layers, each featuring a hidden dimension of 32.
- Node embeddings from all three layers are combined, creating a comprehensive representation for each node.
- Graph embedding is derived by applying mean pooling to the node embeddings.
- To produce predictions, a linear layer is applied at the conclusion of the process.
- Mean Squared Error (MSE) Loss is employed as the loss function, and optimization is carried out using the Adam Optimizer with a learning rate of 0.01.
- The model is saved based on achieving the optimal Root Mean Squared Error (RMSE) score during validation.

Comparisons With Baselined Implementations:

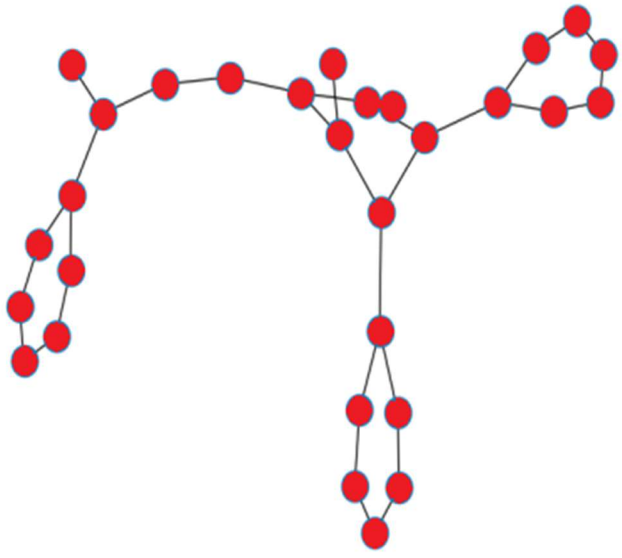
RMSE	Random Model	Linear Regression Model	Implemented GNN Model
Training	2.15	1.16	0.86
Validation	2.39	1.28	0.97

Training and Validation learning curves:

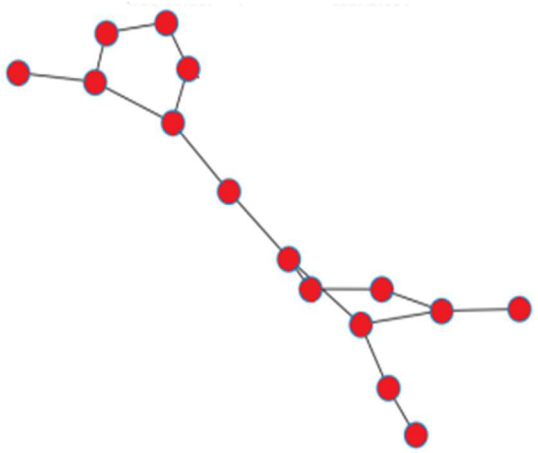


- In terms of RMSE, the GNN Model outperforms both Linear Regression and the Random Model during training, achieving significantly lower scores (0.86 compared to 1.16 and 2.15, respectively).
- This trend is consistent in the validation phase, with the GNN Model showcasing superior performance with an RMSE of 0.97 compared to Linear Regression (1.28) and the Random Model (2.39).
- The substantial reduction in RMSE underscores the effectiveness of the GNN Model in capturing intricate relationships within the data, leading to more accurate predictions compared to the baseline models.

Graph Visualization with NetworkX: Identifying Model Misclassifications and Poor Performance



Pred : -1.8245006 Real : 1.2017736



Pred : 1.58 Real : 3.19

Analysis: In the depicted graphs, numerous clusters of nodes are present, some with sparse or non-existent connections. Our model encounters challenges in accurately predicting the true class for these specific graph configurations.