# Supply Chain Analytics Project

## About Dataset

Understanding the dataset thoroughly is a fundamental and critical step before proceeding with any further analysis, whether it's exploratory analysis, predictive modeling, or any other data-driven task. Here's why a deep understanding of the dataset is of utmost importance:

1. **Data Quality Assurance**: By exploring the dataset thoroughly, you can identify issues like missing values, outliers, or inconsistencies. Addressing these issues early ensures that your subsequent analysis and models are based on reliable and accurate data.

2. **Model Performance**: Your familiarity with the data allows you to select appropriate modeling techniques that align with the dataset's characteristics. Different models have different assumptions, and choosing the right one improves predictive accuracy.

3. **Data Preprocessing**: Effective data preprocessing, including scaling, normalization, and handling categorical variables, is crucial. Without understanding the data, you might apply preprocessing steps that don't align with the data's nature.

4. **Insight Generation**: Thorough exploration often reveals hidden insights, trends, and patterns that can guide your analysis and provide valuable insights to stakeholders.

5. **Domain Knowledge Incorporation**: Deepening your understanding of the data helps you incorporate domain knowledge, which is essential for informed decision-making and relevant model interpretation.

6. **Avoiding Biases**: Understanding the data context helps you recognize potential biases that might impact your analysis and modeling results.

7. **Resource Allocation**: Exploring the data allows you to allocate resources efficiently, focusing on the most relevant attributes and avoiding unnecessary efforts on less informative features.

8. **Effective Communication**: A solid understanding of the dataset facilitates effective communication with team members, stakeholders, and collaborators.

9. **Minimizing Risks**: Inadequate understanding of the data can lead to inaccurate analyses, faulty models, and misinformed decisions, which can have significant consequences.

Data is the foundation of any data-driven project. Taking the time to explore and comprehend the data empowers you to make informed choices, increase the effectiveness of your analysis, and ultimately achieve more meaningful and accurate results.

The dataset is of a particular supply chain network.
These 4 echelons of this supply chain are taken into account : **Supplier, Distributor, Manufacturer and Retailer**.
Risk is associated with each echelon.
Risk Index is calculated for each echelon.

Risk Index for Supplier echelon can be formulated as:

$$RI_{supplier} = \sum_{i=1}^{n} as_{ij}.bs_{ij}.(1-(1-\Pi_{j=1}^{m} P(S_{ij})))$$

where,
i: i-th supplier
j: j-th demand
n: number of suppliers
m: number of demands

Where $as_{ij}$ is the consequence to the supply chain if the i-th supplier fails,
$bs_{ij}$ is the percentage of value added to the product by the i-th supplier,
$P(S_{ij})$ denotes the marginal probability that the i-th supplier fails for j-th demand.

Similarly,

$$RI_{distributor} = ad_{risk_i}.bm_i.(1-(1-P(M_j)))$$
$$RI_{manufacturer} = am_{risk_i}.bm_i.(1-(1-P(M_j)))$$
$$RI_{retailer} = ar_{risk_i}.br_i.(1-(1-P(R_j)))$$

The risk fluctuation subjected to the supply chain network is simulated by a sine-wave generator. This adds a dynamic and time-varying aspect to the dataset, enabling the study of how Risk Index values and other attributes change over time.
In real life also some Risk is associated with each echelon which we don't know in advance. Actual risk index can be calculated only after happening. In our dataset Risk Indices and total cost are calculated and recorded at different time stamps.

**SCM Stability Category**:
The SCM stability category is a discrete classification assigned to different time periods in the dataset. It categorizes the stability of the supply chain based on observed characteristics or

metrics. The categories likely range from lower stability (higher risk, higher uncertainty) to higher stability (lower risk, more predictability).
0: This could represent the category with the lowest level of supply chain stability. It might indicate situations where the supply chain is highly volatile or prone to disruptions.

1: This might indicate a slightly better level of stability compared to category 0 but still involves some degree of risk or uncertainty.

2: This could represent an intermediate level of supply chain stability. It might indicate that the supply chain is becoming more reliable and predictable.

3: This value would indicate a higher level of stability than the previous categories, suggesting a relatively stable supply chain with fewer disruptions.

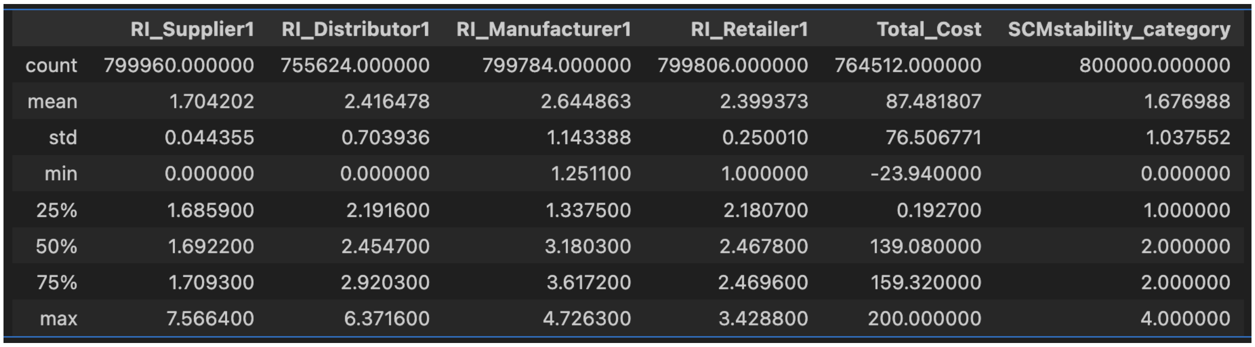
4: This value might represent the highest category of supply chain stability. It could indicate a very stable and well-managed supply chain with minimal risks and disruptions.

# Objectives

1. Exploratory Analysis of dataset
2. Predictive Modeling: Multi-class classification of Risk Indices and Total Cost record

# Exploratory Analysis

**Summary**:



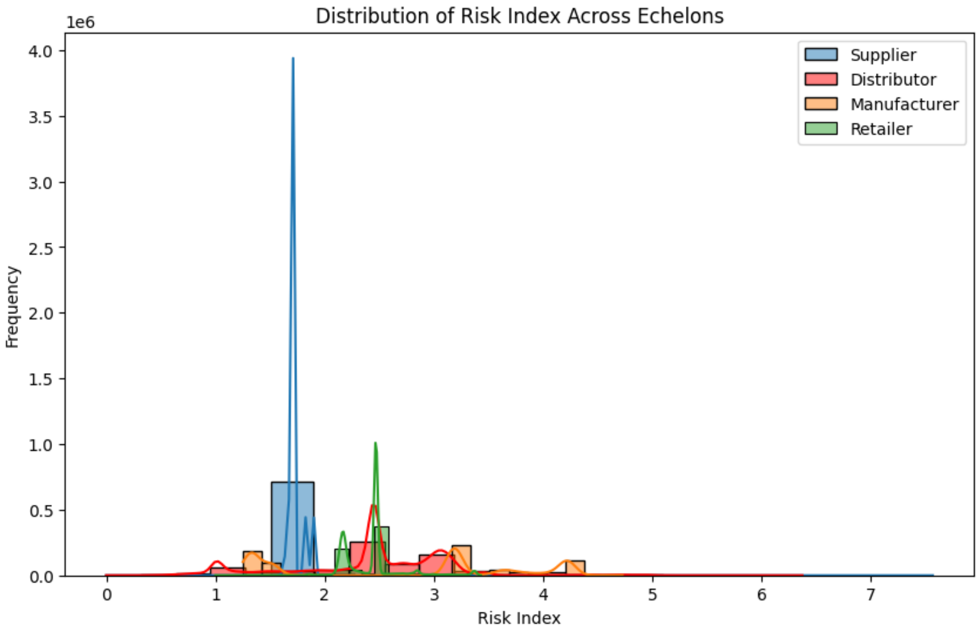
| | RI_Supplier1 | RI_Distributor1 | RI_Manufacturer1 | RI_Retailer1 | Total_Cost | SCMstability_category |
|---|---|---|---|---|---|---|
| count | 799960.000000 | 755624.000000 | 799784.000000 | 799806.000000 | 764512.000000 | 800000.000000 |
| mean | 1.704202 | 2.416478 | 2.644863 | 2.399373 | 87.481807 | 1.676988 |
| std | 0.044355 | 0.703936 | 1.143388 | 0.250010 | 76.506771 | 1.037552 |
| min | 0.000000 | 0.000000 | 1.251100 | 1.000000 | -23.940000 | 0.000000 |
| 25% | 1.685900 | 2.191600 | 1.337500 | 2.180700 | 0.192700 | 1.000000 |
| 50% | 1.692200 | 2.454700 | 3.180300 | 2.467800 | 139.080000 | 2.000000 |
| 75% | 1.709300 | 2.920300 | 3.617200 | 2.469600 | 159.320000 | 2.000000 |
| max | 7.566400 | 6.371600 | 4.726300 | 3.428800 | 200.000000 | 4.000000 |

**Missing Values**:

| | |
|---|---|
| Timestamp | 0 |
| RI_Supplier1 | 40 |
| RI_Distributor1 | 44376 |
| RI_Manufacturer1 | 216 |
| RI_Retailer1 | 194 |

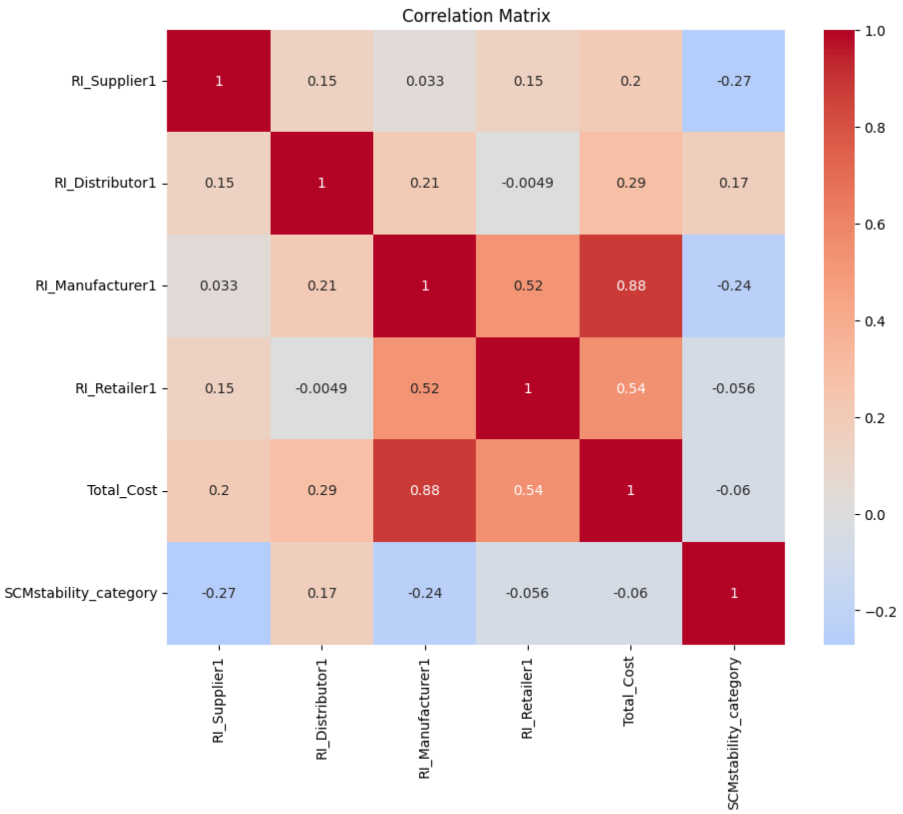
```
Total_Cost               35488
SCMstability_category    0
```

Total data points before cleaning: 8,00,000
Total data points after cleaning: 7,32,331



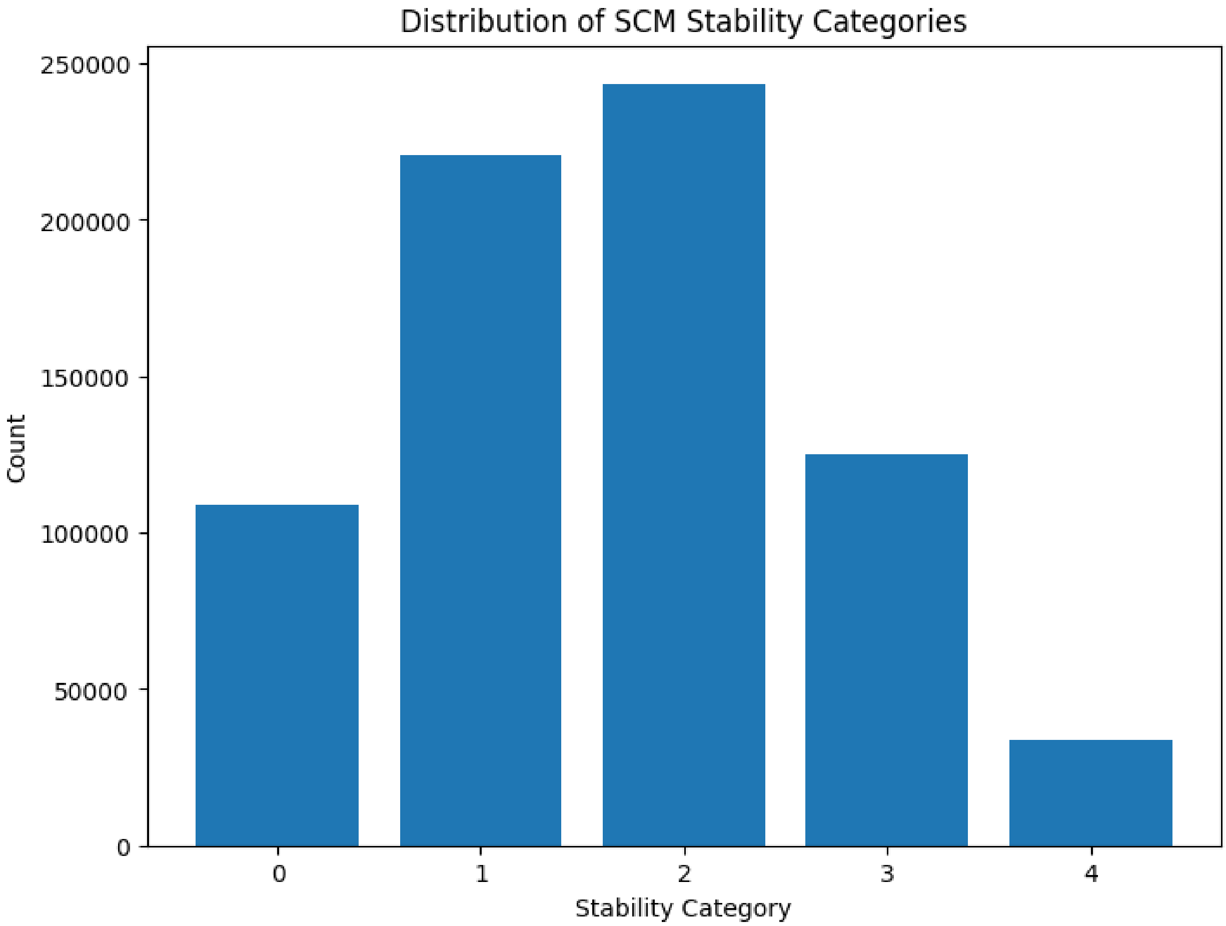
**Distribution graph’s Interpretations:**

- Most of the RI values of Supplier echelon b/w 1 and 2
- Spread of RI values of the Distributor echelon look more wide as compared to others.
- More risk is associated with the Distributor and Manufacturer echelon as they are spread more in rightward.



**Correlation graph's Interpretations**:

- RI_Supplier has highest cross correlation with Total_Cost
- RI_Distributor has highest cross correlation with Total_Cost
- RI_Manufactures has highest cross correlation with Total_Cost
- RI_Retailer has highest cross correlation with Total_Cost
- Total_Cost has highest cross correlation with RI_Manufacturer
- SCMstability_category has highest cross correlation with RI_Distributor



**Bar graph's Interpretations:**

- Most of the time this Supply chain network is found in Stability category 2.
- Very rarely the Supply Chain Network is in the most stable category (4).

# Modeling

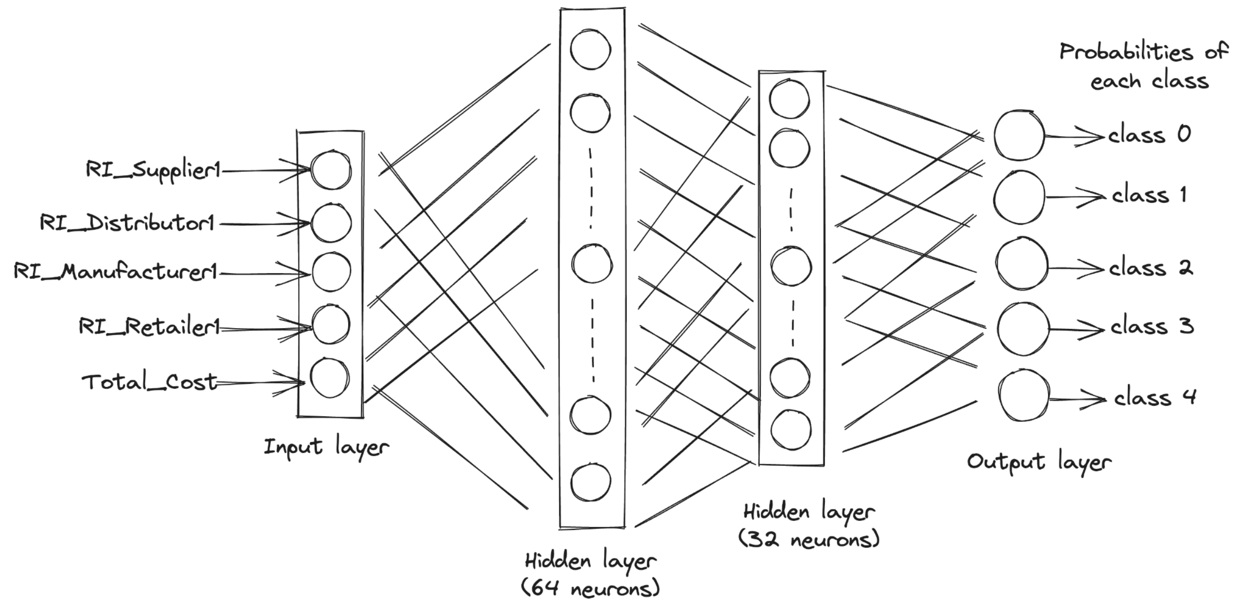
Task type: **Multi-class classification**
Type of model selected: **Feed Forward Neural Network (FFNN)**

**Input Features:**
**RI_Supplier1**: Risk Index value associated with the supplier echelon.
**RI_Distributor1**: Risk Index value associated with the distributor echelon.
**RI_Manufacturer1**: Risk Index value associated with the manufacturer echelon.
**RI_Retailer1**: Risk Index value associated with the retailer echelon.
**Total_Cost**: Total cost associated with the supply chain at that timestamp.

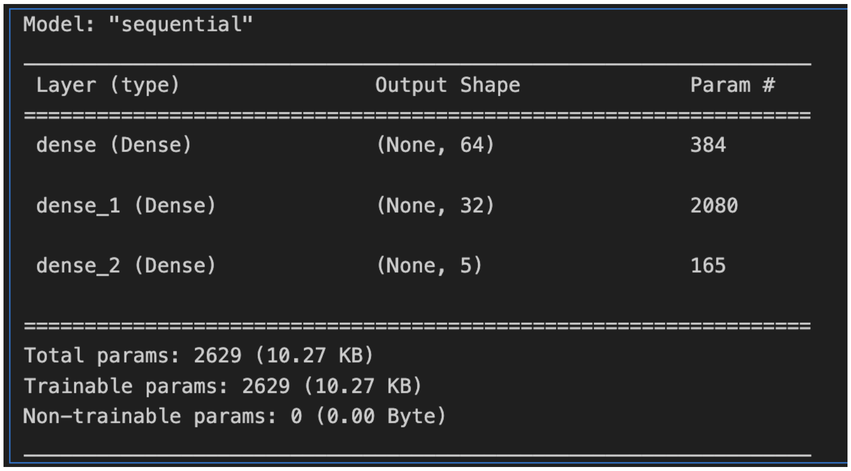
**Output label / class**:

**SCMstability_category**: Categorical classification representing the stability category of the supply chain at that timestamp.

**Neural Network Drawing:**



**Model Summary:**



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                384

 dense_1 (Dense)             (None, 32)                2080

 dense_2 (Dense)             (None, 5)                 165

=================================================================
Total params: 2629 (10.27 KB)
Trainable params: 2629 (10.27 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

**Justification of Model's Architecture choices:**

- Input layer is used to feed input features values to the network
- Two hidden layers are kept in the architecture. Large number of hidden layers can lead to overfitting as the model becomes very complex in that case.
- Number of neurons hidden are like automatically engineered features by a neural network.

- Mapping of 5 features to 64 features initially to detect patterns in the data better. In the second hidden layer neurons are reduced to half to control complexity of the model which affects training time too.
- RELU: This works like a drop-out layer (randomly some neurons are dropped during backpropagation). Most common way to reduce overfitting is to apply ReLU in all hidden layers.
- Last layer is the output which gives the prediction vector. Probability of data point to belong in each class is given.

**For Training:**

- Softmax activation action is used in case of multi-class classification.
- **Loss Function:** Categorical Cross Entropy (as it is multi-class classification task)
- **Optimiser:** Adam (experimentally found that it works great)
- **Learning rate**: 0.001 (default learning rate of Adam optimiser in Keras)
- Batch size: 25, Number of Epochs: 20

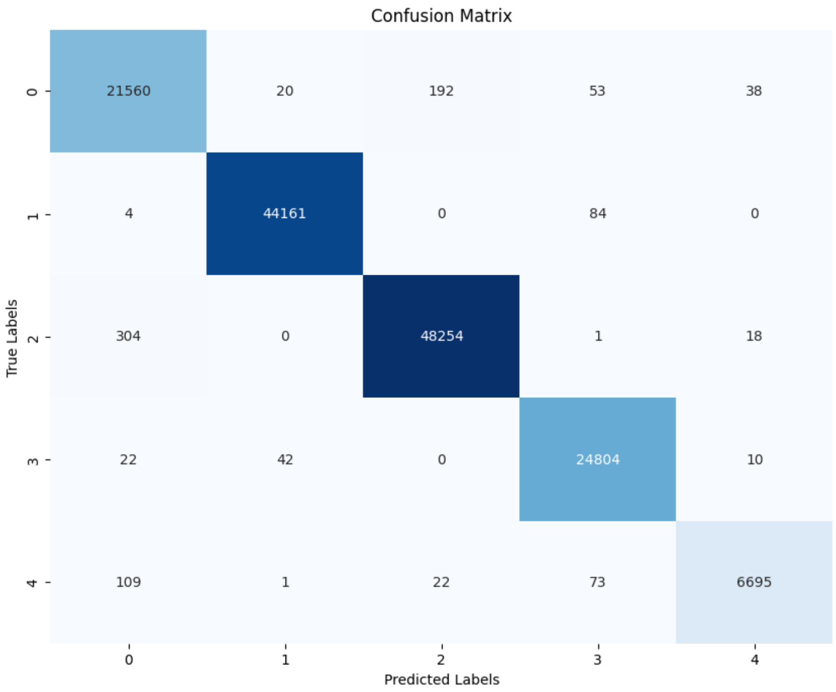
**Last Few Epochs of Training:**

```
Epoch 18/20
18748/18748 [==============================] - 11s 590us/step - loss: 0.0235 -
accuracy: 0.9933 - precision: 0.9934 - recall: 0.9932 - f1_score: 0.9901 - val_loss: 0.0238 -
val_accuracy: 0.9935 - val_precision: 0.9936 - val_recall: 0.9934 - val_f1_score: 0.9907

Epoch 19/20
18748/18748 [==============================] - 11s 579us/step - loss: 0.0234 -
accuracy: 0.9934 - precision: 0.9935 - recall: 0.9933 - f1_score: 0.9904 - val_loss: 0.0229 -
val_accuracy: 0.9939 - val_precision: 0.9940 - val_recall: 0.9938 - val_f1_score: 0.9914

Epoch 20/20
18748/18748 [==============================] - 11s 577us/step - loss: 0.0227 -
accuracy: 0.9935 - precision: 0.9936 - recall: 0.9934 - f1_score: 0.9906 - val_loss: 0.0251 -
val_accuracy: 0.9932 - val_precision: 0.9933 - val_recall: 0.9932 - val_f1_score: 0.9900
```

## Model Evaluation (on Test Data)

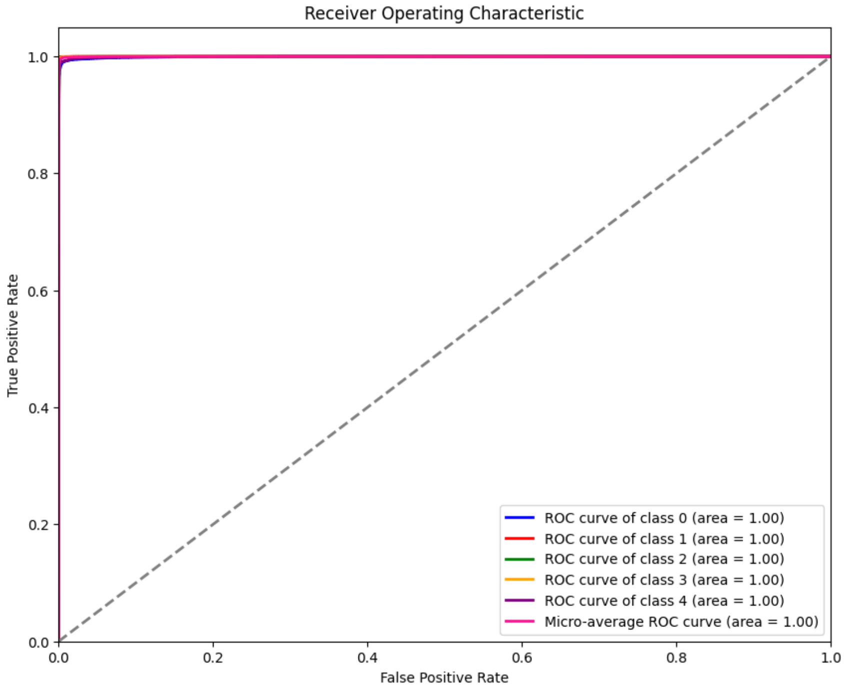
**Accuracy: 99.32 %** **Precision: 99.32 %** **Recall: 99.32 %** **F1-score: 99.32 %**

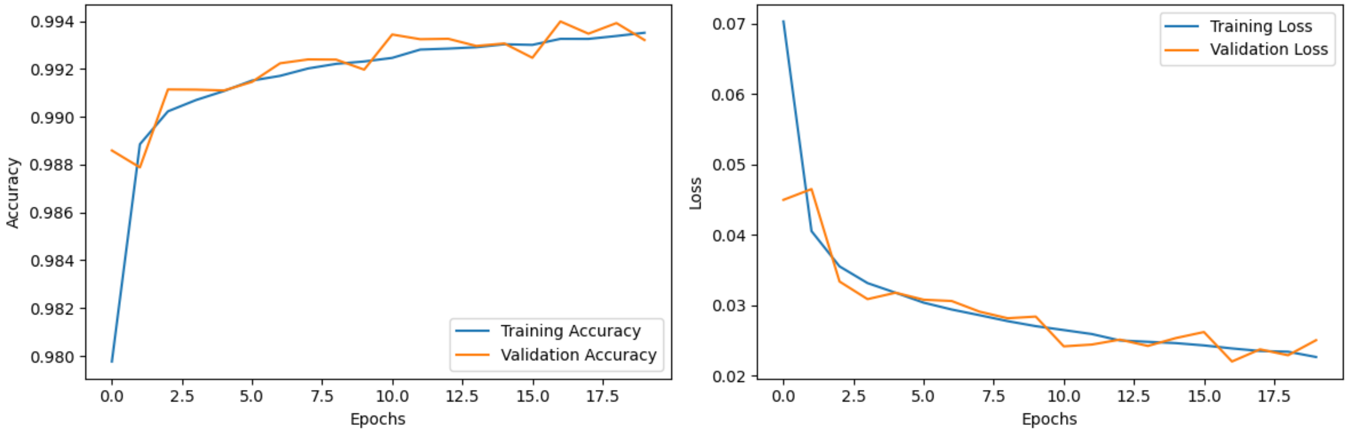
**Confusion Matrix:**



| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 21560 | 20 | 192 | 53 | 38 |
| 1 | 4 | 44161 | 0 | 84 | 0 |
| 2 | 304 | 0 | 48254 | 1 | 18 |
| 3 | 22 | 42 | 0 | 24804 | 10 |
| 4 | 109 | 1 | 22 | 73 | 6695 |

## ROC Curve:





# References

1. Banerjee, Heerok; Saparia, Grishma; Ganapathy, Velappa; Garg, Priyanshi; Shenbagaraman, V. M. (2019), “Time Series Dataset for Risk Assessment in Supply Chain Networks”, Mendeley Data, V2, doi: 10.17632/gystn6d3r4.2

2. Saparia, Grishma & Banerjee, Heerok & Garg, Priyanshi & Ganapathy, V. & V M, Shenbagaraman. (2019). Time-series Dataset for Risk Assessment in Multi-echelon Supply Chain Networks. 10.17632/gystn6d3r4.2".