Machine Learning Assignment Report

1. Objective

To develop a machine learning model to predict mycotoxin levels in corn samples.

2. Data Preprocessing

- **Dataset**: A hyperspectral dataset containing spectral reflectance values and vomitoxin concentration levels.

- **Missing Values**: Handled by replacing missing feature values with the median.

- **Feature Scaling**: Standardization using StandardScaler to normalize spectral data.

- **Train-Test Split**: 80-20 split for training and evaluation.

3. Exploratory Data Analysis (EDA)

- **Spectral Reflectance Trends**: Average reflectance plotted across wavelength bands showed variation, indicating key distinguishing features.

- **Heatmap Analysis**: A heatmap of the first 50 samples revealed patterns of spectral reflectance variations, highlighting important wavelengths.

4. Dimensionality Reduction

- **Principal Component Analysis (PCA)**: Reduced dimensions to two principal components, preserving significant variance. PCA visualization indicated potential groupings in the data.

- **t-SNE Projection**: Used for non-linear dimensionality reduction, revealing natural clusters among samples based on spectral signatures.

5. Model Training

- **Deep Learning Model**: A neural network with:

    o 128-node hidden layer (ReLU activation function)

    o 64-node hidden layer (ReLU activation function)

    o Output layer for regression

- **Training**: Model trained for 50 epochs using Adam optimizer and Mean Squared Error (MSE) loss function.

6. Model Evaluation

- **Performance Metrics**:

    ➢ **Mean Absolute Error (MAE)**: Quantifies average prediction error.

    ➢ **Root Mean Squared Error (RMSE)**: Measures the average magnitude of errors.

    ➢ **R-squared Score (R^2)**: Indicates how well the model explains variance in vomitoxin levels.

- **Outcomes**:
  - If $R^2 < 0.5$, the model's predictive ability is limited, suggesting the need for feature engineering or additional data.

- **Actual vs. Predicted Values**:
  - A scatter plot comparison showed how well the model predicted vomitoxin levels, with deviations indicating potential gaps in learned patterns.

7. Conclusion

- The model successfully predicted vomitoxin concentration with moderate accuracy.

- Spectral data provided valuable features, but further improvements could include additional feature extraction techniques or alternative models.