Machine Learning Model for Mycotoxin Level Prediction

Data Preprocessing

The dataset underwent several preprocessing steps to enhance model performance:

- **Handling Missing Values**: Missing data points were imputed using median values to preserve the dataset's distribution.
- **Feature Scaling**: Standardization (z-score normalization) was applied to ensure uniform feature importance.
- **Categorical Encoding**: One-hot encoding was used for categorical variables to facilitate model compatibility.
- **Outlier Detection**: Z-score and IQR-based filtering helped remove extreme outliers that could affect model accuracy.

Dimensionality Reduction

- **Principal Component Analysis (PCA)** was employed to reduce dimensionality while retaining maximum variance.
- PCA results showed that the first few components explained a significant portion of variance, allowing dimensionality reduction without significant information loss.
- **Feature Importance Analysis** indicated that specific variables had a higher influence on predictions, which guided feature selection.

Model Selection and Training

- Various models were tested, including **Logistic Regression, Random Forest, XGBoost, and Neural Networks**.
- Hyperparameter tuning was conducted using **Grid Search and Random Search**.
- **Cross-validation (k-fold)** was applied to mitigate overfitting and improve generalization.
- XGBoost outperformed other models in accuracy, but Logistic Regression was considered due to its interpretability.

Model Evaluation

- **Metrics Used**: Accuracy, Precision, Recall, F1-score, and ROC-AUC were used to assess model performance.
- **Best Performing Model**: XGBoost achieved the highest accuracy, but simpler models like Logistic Regression provided interpretable results with minimal loss in accuracy.

Key Findings and Recommendations

- **Feature engineering played a crucial role in improving model accuracy**.
- **Dimensionality reduction (PCA) helped streamline computation without significant performance loss**.
- **Future Improvements**:
  - Incorporate more advanced feature selection techniques.
  - Experiment with deep learning models for enhanced accuracy.
  - Use ensemble methods to combine model strengths and further optimize predictions.