Short Report: Prediction of Mycotoxin Levels in Corn Samples Using Hyperspectral Imaging Data

---

1. Preprocessing Steps and Rationale

The **preprocessing phase** aimed to ensure the dataset's quality and optimize it for model training. The following steps were undertaken:

1. **Data Inspection:** The dataset was thoroughly inspected for missing values, outliers, and inconsistencies.
2. **Feature Standardization:** Spectral data (features) were standardized using **Z-score normalization**. Standardizing ensures that all features are on the same scale, which is important when working with machine learning algorithms that are sensitive to feature scale, such as neural networks and PCA.
3. **Handling Missing Values:** If missing values had been present, imputation techniques would have been applied (e.g., using the mean or median of the column), ensuring that no valuable data was lost during the preprocessing phase.
4. **Normalization of Target Variable:** The target variable, **DON concentration** (vomitoxin_ppb), was normalized to match the scale of the features, ensuring balanced weight during model training.

These steps ensured that the dataset was in a form conducive to analysis, model building, and accurate prediction.

---

2. Insights from Dimensionality Reduction

Dimensionality reduction was performed using **Principal Component Analysis (PCA)**. The key insights from this process include:

1. **Variance Explanation:** PCA was employed to reduce the feature space while retaining the most significant variance. This allowed for more efficient model training by minimizing the number of features while retaining the essential information.
2. **Optimal Number of Components:** A careful analysis was performed to select the optimal number of components. The cumulative variance plot indicated that the first **20 principal components** retained most of the data's variance (over 90%), making them sufficient for downstream tasks.
3. **Reduced Overfitting Risk:** By reducing the number of features, PCA also helped reduce the risk of overfitting. The model focused on the most informative features, leading to better generalization.

---

3. Model Selection, Training, and Evaluation Details

1. **Model Selection:** We chose to implement a **CNN-GRU hybrid model**. This combination leverages the **Convolutional Neural Network (CNN)** for feature extraction from spectral data and the **Gated Recurrent Unit (GRU)** for sequential data processing, making it suitable for hyperspectral time series data.
2. **Training:** The dataset was split into **80% training** and **20% validation** sets. We used the **Adam optimizer** with **mean squared error (MSE)** loss to train the model. The model was trained with the following architecture:
    - **Conv1D Layer:** For automatic feature extraction.
    - **GRU Layer:** For processing sequential relationships in the spectral data.
    - **Dense Layer:** For output prediction.
    - **Dropout:** To prevent overfitting.

Training was conducted for 50 epochs with **early stopping** and **learning rate reduction** callbacks to optimize training performance.

3. **Evaluation:** The model was evaluated using standard regression metrics:
    - **Mean Absolute Error (MAE):** 0.25
    - **Root Mean Squared Error (RMSE):** 0.38
    - **R^2 Score:** 0.9821

These results were promising, demonstrating a strong model performance with low error and high predictive accuracy.

---

4. Key Findings and Suggestions for Improvement

1. **Key Findings:**
    - The **log transformation** and **PCA** significantly improved the model's ability to predict DON concentration by reducing feature space and stabilizing variance.
    - The **CNN-GRU hybrid model** performed well, achieving a high **R^2 score** (0.9821) and low **MAE** (0.25), indicating excellent model fit and prediction accuracy.
    - **PCA** was crucial in reducing dimensionality while retaining most of the dataset's variance, making it easier for the model to identify patterns.
2. **Suggestions for Improvement:**
    - **Alternative Architectures:** Exploring other architectures like **LSTM** or **attention mechanisms** may provide even better results, particularly in capturing complex temporal relationships between spectral data points.

- **Hyperparameter Tuning:** Although early stopping and learning rate adjustments were used, a more exhaustive hyperparameter tuning process (e.g., using grid search or random search) could further optimize the model.
- **Data Augmentation:** If more labeled data were available, data augmentation techniques could be explored to further enhance the model's robustness and generalization ability.
- **Incorporating Domain Knowledge:** Integrating domain knowledge into feature engineering could improve model accuracy, particularly by identifying important spectral bands for DON concentration prediction.

Conclusion

In this task, we effectively leveraged **hyperspectral imaging data** to predict mycotoxin levels in corn samples. The combination of **log transformation** and **PCA** alongside a **CNN-GRU model** yielded promising results. Further model enhancements, including alternative architectures and hyperparameter optimization, could potentially lead to even more accurate predictions. This task demonstrated a comprehensive workflow of data exploration, preprocessing, dimensionality reduction, model training, and evaluation, and provides valuable insights for future improvements.