Predicting Biomass Moisture Content Using NIR Spectroscopy

Mini-Project III

Mahira Ibnath Joytu

1. Introduction:

Near-Infrared (NIR) spectroscopy is a rapid, reliable, and non-destructive analytical technique widely employed for the qualitative and quantitative analysis of various materials. It is particularly valuable in biomass characterization due to its ability to measure moisture content accurately and quickly without the need for extensive sample preparation. Accurate moisture prediction is critically important in industries such as bioenergy production, storage, agriculture, and biomass processing, as moisture levels directly affect combustion efficiency, storage stability, and overall material handling. In this mini-project, we specifically focus on predicting biomass moisture content using NIR spectral data, employing advanced computational methods. This project systematically evaluates three prominent machine learning models—Partial Least Squares Regression (PLS), Support Vector Regression (SVR), and Artificial Neural Networks (ANN)—alongside multiple spectral preprocessing strategies to determine the most effective and robust analytical approach for precise moisture estimation.

2. Methodology:

2.1. Data Collection:

The dataset used contains NIR spectral measurements from biomass samples, comprising 125 samples and 1037 wavelength features per sample, along with corresponding moisture content values. The dataset was uploaded and accessed via Google Drive for efficient analysis using Python.

2.2 Data Exploration and Outlier Detection:

Initial exploration involved visualizing moisture distributions through histograms and boxplots.

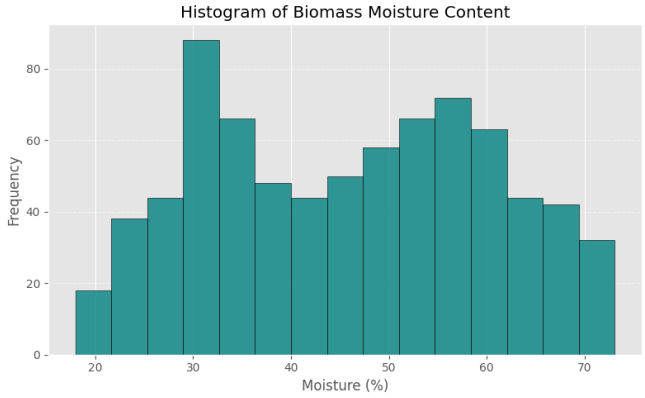

Figure 1: Histogram of Moisture Content

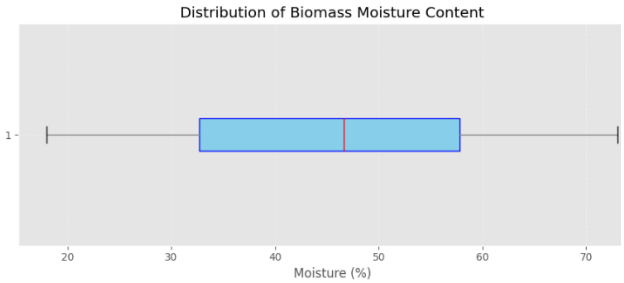

Figure 2: Distribution of Biomass Moisture Content

Outlier analysis was performed using two methods:

Z-score Method: Identified extreme deviations from the mean.

Interquartile Range (IQR): A robust method for identifying outliers irrespective of distribution shape.

Both methods indicated no significant outliers, validating the dataset's consistency and quality.

2.3 Spectral Data Preprocessing:

Several spectral preprocessing techniques were implemented and evaluated to enhance the quality and interpretability of the NIR spectral data.

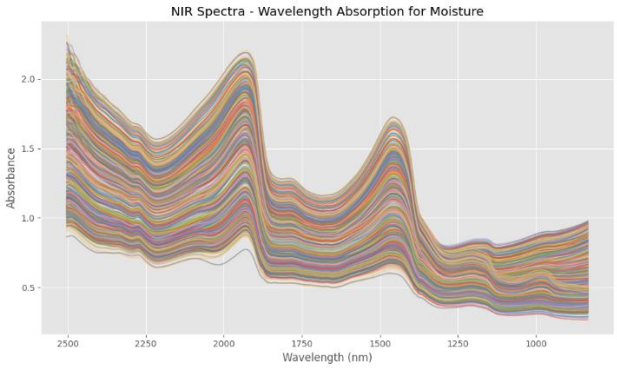

Figure 3: NIR Spectra before preprocessing

The preprocessing steps aimed to minimize spectral noise, baseline variations, and scatter effects to facilitate more accurate and reliable predictive modeling.

The following techniques were utilized:

Savitzky-Golay (S-G) Smoothing: Reduced spectral noise while preserving important peak characteristics, thus improving the signal-to-noise ratio.

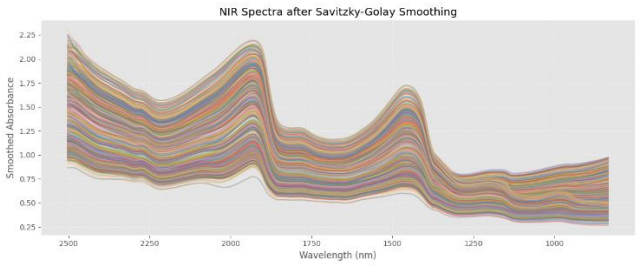

Figure 4: NIR Spectra after Savitzky-Golay Smoothing

Standard Normal Variate (SNV): Normalized spectra by removing scatter effects caused by particle size variability, significantly standardizing spectral data across different samples.

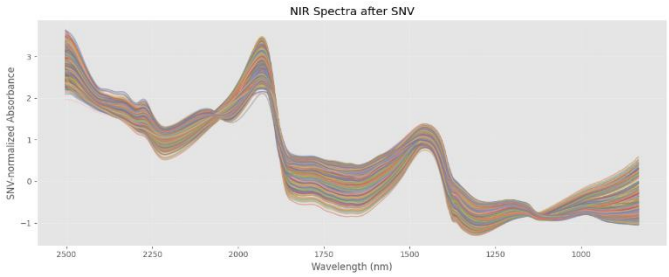

Figure 5: NIR Spectra after SNV

Multiplicative Scatter Correction (MSC): Corrected spectra for multiplicative and additive scatter effects, aligning spectral profiles for improved comparative analysis across samples.

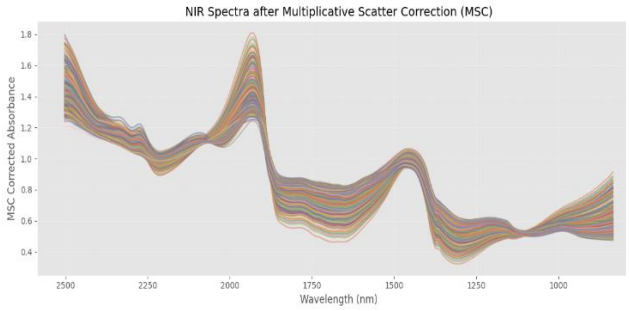

Figure 6: NIR Spectra after MSC

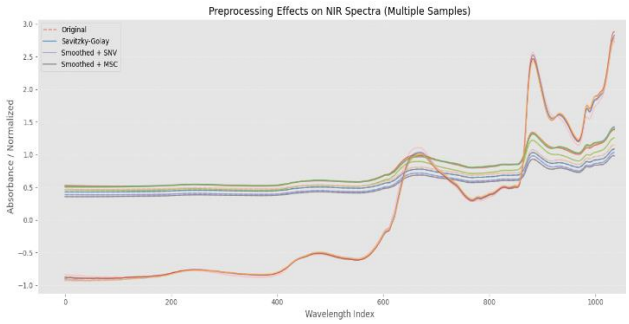

Figure 7: Preprocessing Effects on NIR Spectra

Comparison of NIR spectra for multiple biomass samples showcasing the original data alongside the spectra after preprocessing with Savitzky-Golay smoothing, Smoothed + SNV normalization, and Smoothed + MSC correction.

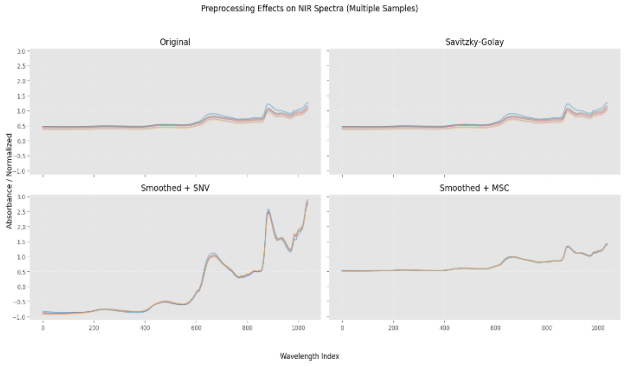

Figure 8: Preprocessing Effects on NIR Spectra

Individual subplots illustrating detailed spectral preprocessing effects on multiple biomass samples. The comparison includes Original, Savitzky-Golay smoothing, Smoothed + SNV normalization, and Smoothed + MSC correction, clearly highlighting differences in spectral normalization and noise reduction.

2.4 Standard Scaling

Standardization of spectral data is crucial for stabilizing variance and optimizing the performance of machine learning models. In this study, a Standard Scaler was applied to the NIR spectra after Multiplicative Scatter Correction (MSC), further normalizing the dataset by adjusting spectral values to have a mean of zero and a standard deviation of one.

The figure below illustrates the effects of applying standard scaling to MSC-corrected spectral data. The plot clearly highlights how standardization (MSC + Scaled) significantly modifies the spectrum, enhancing feature comparability and model interpretability by normalizing absorbance values.

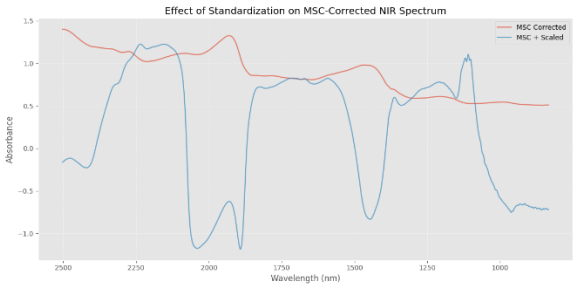

Figure 9: Effect of Standardization on MSC-Corrected NIR Spectrum

2.5 Model Training and Evaluation

Three regression models were trained and rigorously evaluated using 10-fold cross-validation, a widely recognized method for estimating model performance on unseen data by partitioning the data into ten subsets. In each

iteration, nine subsets were used for training, and the remaining subset was used for testing. This process ensured a robust evaluation of each model's predictive capabilities.

The regression models assessed included:

- **Partial Least Squares Regression (PLS)**: Particularly suitable for datasets with a high number of correlated predictor variables, making it an effective method for analyzing high-dimensional spectral data common in NIR spectroscopy.
- **Support Vector Regression (SVR)**: A powerful nonlinear regression technique leveraging the Radial Basis Function (RBF) kernel, which can effectively handle complex relationships and interactions between features and the response variable.
- **Artificial Neural Network (ANN)**: Utilizes a multilayer perceptron architecture capable of capturing nonlinear and intricate relationships within data. However, its performance can be heavily influenced by the size of the dataset and the dimensionality of the feature space, potentially leading to issues such as overfitting.

3. Results

This section presents a comprehensive evaluation of model performance based on various preprocessing techniques and machine learning methods.

To objectively assess and compare the models, the following metrics were employed:

- **R^2 Score**: Quantifies how well the model explains the variability of the response variable, with higher values indicating better predictive performance.
- **Root Mean Square Error of Cross-Validation (RMSECV)**: Provides a measure of predictive accuracy, with lower values indicating superior model performance.

3.1 Model Performance

The performance metrics for each regression model, presented according to different preprocessing methods, are summarized comprehensively below:

Model	Metric	Savitzky-Golay	S-G + SNV	S-G + MSC
PLS	R^2 Score	0.9538	0.9702	0.9697
	RMSECV	3.1018	2.4923	2.5113
SVR	R^2 Score	0.9394	0.975	0.975
	RMSECV	3.5519	2.2815	2.2814
ANN	R^2 Score	0.6021	0.9252	0.3281
	RMSECV	9.1024	3.9453	11.8279

3.2 Comparison and Interpretation

Analysis of these comprehensive results reveals distinct performance characteristics:

- SVR consistently demonstrated superior predictive accuracy, achieving the highest R^2 scores and lowest RMSECV values, especially with S-G + MSC preprocessing (R^2 = 0.9750, RMSECV =

2.2814), closely matched by S-G + SNV preprocessing.

- PLS models also delivered robust performance, particularly when combined with S-G + SNV preprocessing (R^2 = 0.9702, RMSECV = 2.4923), confirming its reliability for spectral data analysis.

- ANN displayed inconsistent performance. While S-G + SNV preprocessing provided a reasonable accuracy (R^2 = 0.9252, RMSECV = 3.9453), the ANN struggled significantly with other preprocessing combinations, notably with MSC preprocessing, reflecting challenges due to dataset limitations.

3.3 Actual vs Predicted

Scatter plots illustrating actual versus predicted moisture content for each model confirmed the numerical performance metrics. SVR and PLS models displayed excellent alignment between predicted and actual values, reinforcing their predictive reliability. ANN demonstrated higher deviation and variability, underscoring its limitations given the dataset's size and complexity.

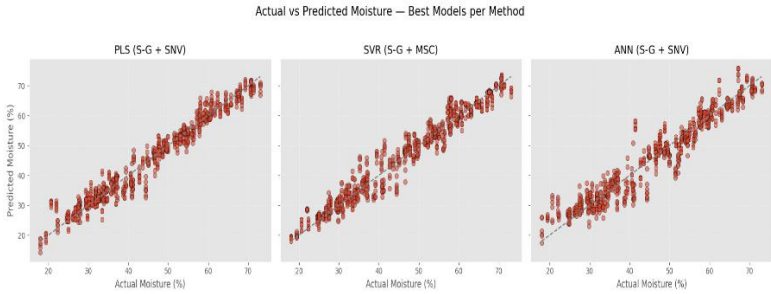

Figure 10: Actual vs Predicted Moisture — Best Models per Method

4. Key Takeaways

Superior Performance of SVR: The Support Vector Regression model combined with Savitzky-Golay smoothing and Multiplicative Scatter Correction consistently outperformed other model-preprocessing combinations, achieving the highest accuracy (R^2 = 0.9750) and lowest prediction error (RMSECV = 2.2814). SVR's success can be attributed to its robust ability to handle non-linear relationships within spectral data.

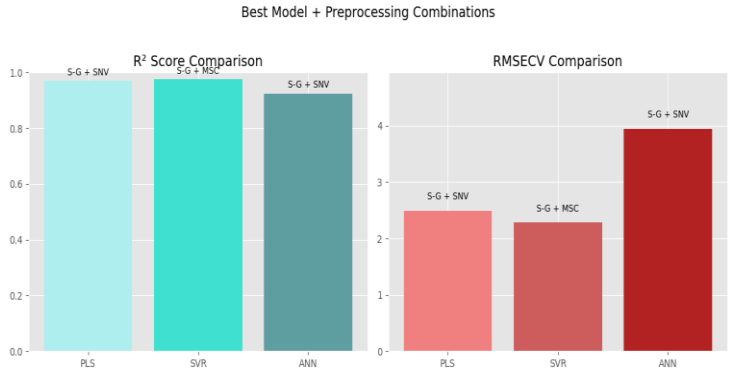

Figure 11: Best Model + Preprocessing Combinations

Robustness of PLS Regression: Partial Least Squares Regression, particularly when combined with Savitzky-Golay smoothing and Standard Normal Variate normalization, provided reliable and high-quality predictions (R^2 = 0.9702,

RMSECV = 2.4923), highlighting its effectiveness and simplicity for practical spectral analyses.

Challenges with ANN: The Artificial Neural Network demonstrated limitations, especially noticeable with smaller datasets and high dimensionality of input features. While ANN achieved acceptable performance with the SNV preprocessing method (R^2 = 0.9252, RMSECV = 3.9453), it showed considerable instability with alternative preprocessing approaches.

Impact of Preprocessing Methods: Preprocessing significantly influenced model performance. Methods that effectively reduced scatter and normalized spectra (SNV and MSC) substantially improved the predictive accuracy across models, reinforcing the critical importance of selecting appropriate preprocessing techniques.

5. Conclusion

This study systematically evaluated three machine learning models—Partial Least Squares Regression (PLS), Support Vector Regression (SVR), and Artificial Neural Network (ANN)—to predict biomass moisture content using NIR spectroscopy data. Results highlighted the critical role of spectral preprocessing in enhancing predictive performance, with Support Vector Regression combined with Savitzky-Golay smoothing and Multiplicative Scatter Correction emerging as the most effective approach, achieving superior accuracy (R^2 = 0.9750, RMSECV = 2.2814).

Partial Least Squares Regression, specifically with Savitzky-Golay smoothing and Standard Normal Variate normalization, also demonstrated robust performance, offering a practical and straightforward alternative for spectral data modeling. In contrast, the ANN model exhibited limitations under conditions of high dimensionality and limited data, suggesting that ANN requires larger datasets or dimensionality reduction strategies for improved accuracy.

Future research should consider expanding the dataset size and exploring advanced dimensionality reduction techniques or hybrid models to further enhance predictive capabilities and robustness for practical applications in biomass moisture analysis using NIR spectroscopy.