**MULTIMEDIA UNIVERSITY OF KENYA**

**FACULTY OF SCIENCE AND TECHNOLOGY**

**MATHEMATICS DEPARTMENT**

**MULTIVARIATE METHODS**

**MULTIVARIATE ANALYSIS OF BANANA FLOUR PHYSICOCHEMICAL PROPERTIES USING PRINCIPAL COMPONENT ANALYSIS**

**SMA 2437**

**MARTIN KOOME SCT-254-039/2021**

**APRIL 2025**

## **ABSTRACT**

This study applies Principal Component Analysis (PCA) to a dataset of banana flour chemical and physical components to reduce dimensionality and enhance interpretability for classification modeling. PCA identifies the most influential features, thereby simplifying the dataset without significant loss of information. The analysis explores the variance structure, visualizes data clusters, and evaluates how PCA can assist in distinguishing between four types of banana flour groups. A new sample is projected onto the PCA space to demonstrate prediction and visualization of unknown samples. The study concludes by using Support Vector Classifier (SVC) to assess classification performance with and without PCA transformation, with minimal accuracy trade-offs observed.

## **BACKGROUND OF THE STUDY**

Banana flour can be derived from various parts of the banana at different ripening stages: green pulp, green peel, ripe pulp, and ripe peel. Each group has unique physicochemical characteristics, which affect its suitability in food processing and nutrition. While component analysis of banana flour has been conducted in nutritional research, its integration with PCA for group classification remains underexplored.

PCA enables transformation of correlated variables into uncorrelated principal components, making patterns and clustering more apparent. This method is particularly useful in food chemistry, where datasets often have high collinearity among nutrients, minerals, and chemical compounds.

## **PROBLEM STATEMENT**

The high dimensionality of banana flour data introduces noise and complexity, impeding the use of traditional analytical methods. Visual or statistical analysis becomes less effective as the number of variables increases. PCA serves to distill the dataset into fewer dimensions, preserving essential variance that informs classification, especially when training machine learning models.

## **OBJECTIVES**

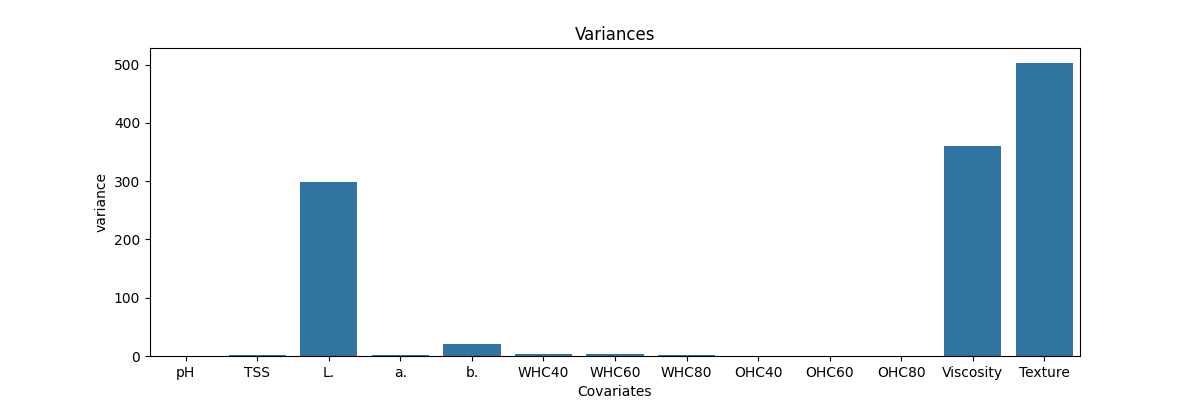
The main objective is to apply PCA for dimensionality reduction and investigate its effectiveness in classifying banana flour samples based on chemical attributes.

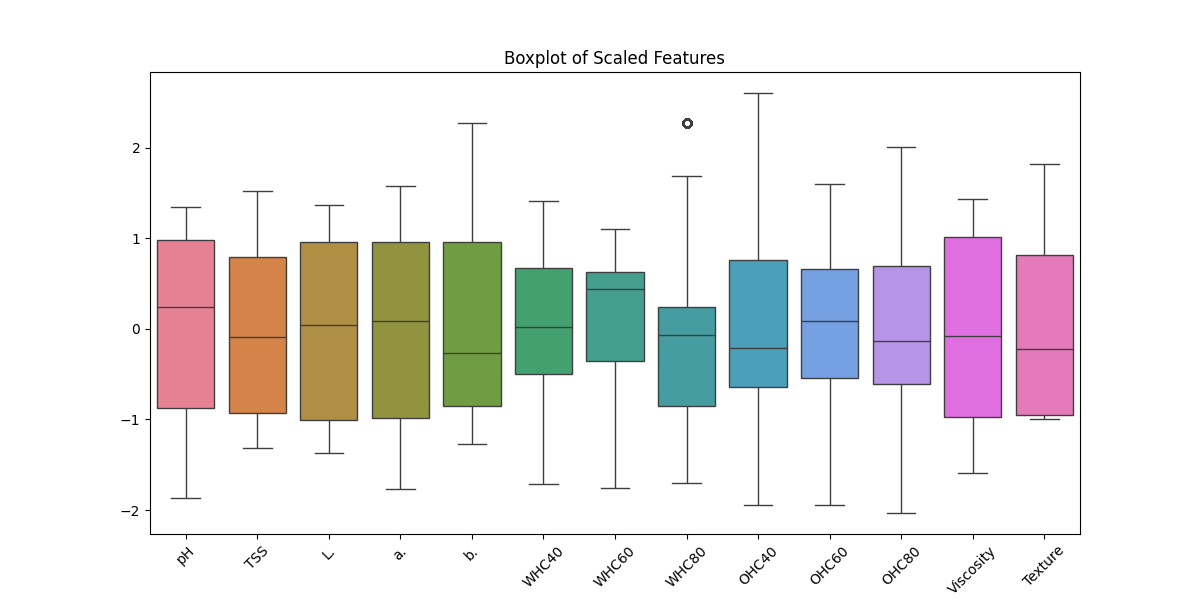
**Specific Objectives**

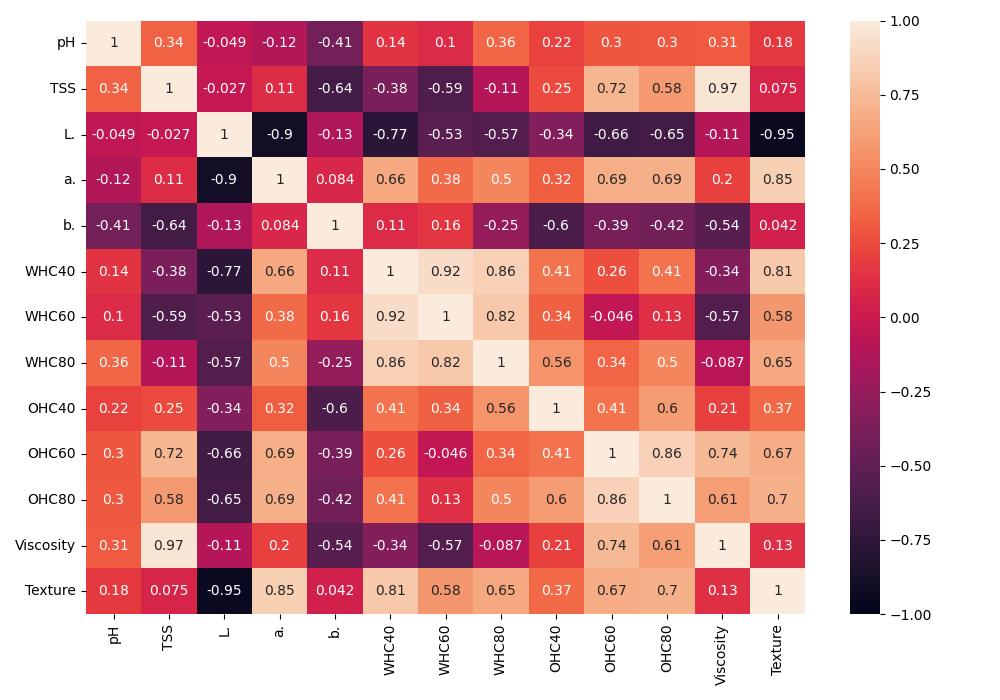
* To compute and visualize the variance of each component
* To generate a PCA biplot to explore group clustering
* To project new unidentified samples onto the PCA space
* To train an SVC model and assess classification accuracy
* To compare the model performance with and without PCA transformation

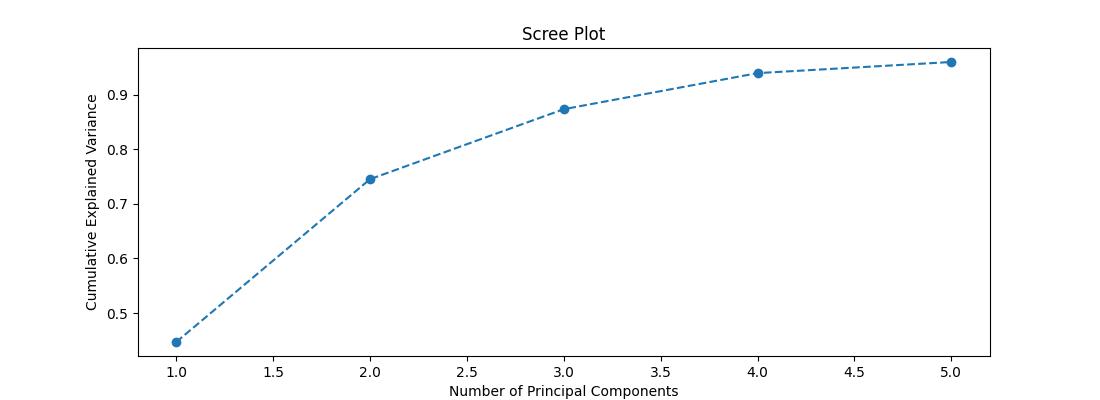
## **METHODOLOGY**

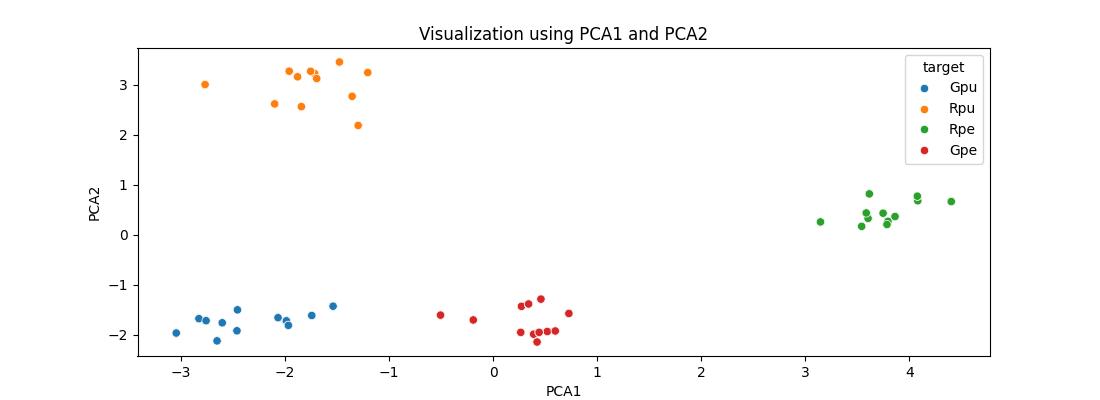
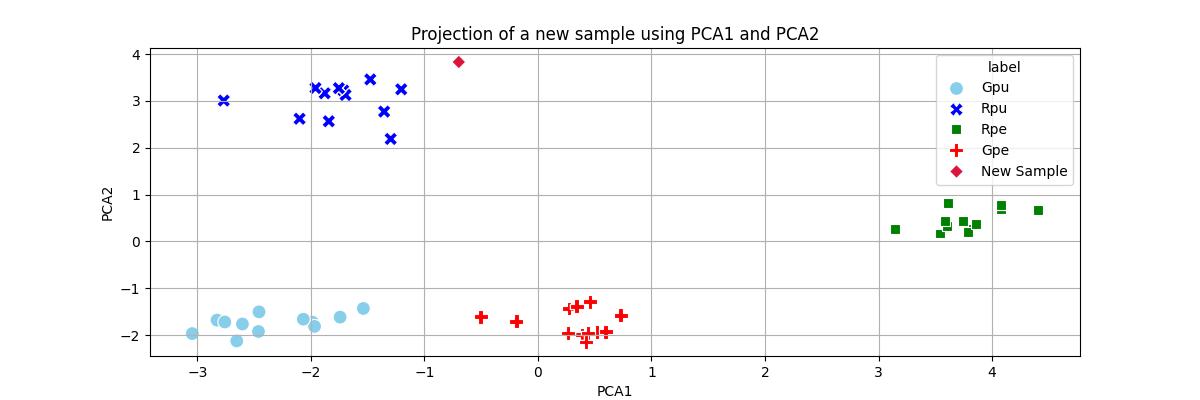
The dataset consists of 13 quantitative features and a categorical variable defining the flour group. Analysis was conducted using Python libraries including NumPy, pandas, matplotlib, seaborn, and scikit-learn.

* **Exploratory Analysis**: Variance plots were created to identify dominant variables and the need for scaling. We observe that viscosity, texture and L covariates have high variance which would lead to high variations when PCA is applied to the data.
* **Data Preprocessing**: The group column was encoded and all features standardized to ensure a uniform scale. I used a standard scaler module in python which removes the mean and scales the features to unit variance. The scaled features had a mean of 0 and a standard deviation of 1 which depicts a Gaussian Distribution.



* **Correlation Analysis**: A correlation matrix was generated and visualized using a heatmap to assess linear relationships. We have independent covariates that have a very high correlation with each other which may have a negative effect on the classification model perfomance.
* **Principal Component Analysis**: PCA was applied using scikit-learn, and the proportion of explained variance was plotted to determine the optimal number of components. The optimal number of principal components used in this analysis are 5 as shown in the scree plot.



* **Biplot Construction**: A biplot was generated using the first two components and loaded vectors for visualization.
* **New Sample Projection**: A synthetic sample was projected onto the PCA space to simulate prediction and classification of an unknown banana flour sample.
* **Oversampling**: Each group was oversampled equally using resample() to balance the dataset and improve classifier learning.
* **Model Training**: A Support Vector Classifier (SVC) was trained using both raw and PCA-transformed data. Performance was compared using accuracy scores.

## **RESULTS AND DISCUSSION**

### **Explained Variance**

Initial variance analysis revealed substantial disparities in feature magnitudes, affirming the necessity of scaling prior to PCA. Without scaling, variables with larger numeric ranges would dominate the principal components, potentially skewing results. By applying standardization (mean = 0, variance = 1), the contribution of all features was normalized.

The first two principal components explained approximately 72.5% of the total variance, indicating that a significant portion of the dataset's information could be retained in just two dimensions. This dimensionality reduction not only simplifies the feature space but also allows more efficient downstream modeling and visualization.

### **PCA Visualization**

The PCA scatterplot revealed visible and well-separated clusters corresponding to the four banana flour groups. Each group showed unique distribution patterns across PC1 and PC2:

* **Green pulp and green peel** appeared closely related but formed separate clusters implying to have closely similar physicochemical composition.
* **Ripe pulp** was distributed further along PC2, suggesting distinct chemical composition.
* **Ripe peel** flour being furthest along PC1 implies it has the most distinct physicochemical profile compared to the other flour types based on the variables contributing to that component. In PCA terms, it contributes most to the variance captured by PC1, highlighting its distinctiveness.

### **New Sample Projection**

A new synthetic point was generated using mean-centered values close to the observed mean of ripe pulp samples. After PCA transformation, the point projected near the ripe pulp cluster, confirming its compositional similarity.

This projection validates PCA’s use in practical scenarios such as predicting or classifying unknown samples in quality control environments. Such application is common in chemometrics and functional food research, where PCA aids in sample discrimination and authenticity verification.

### **Classification Performance**

* **With Raw Data**: Accuracy = 98%
* **With PCA-transformed Data**: Accuracy = 97%

While classification with raw features slightly outperformed PCA-reduced data, the reduction in dimensionality from 13 to 5 features resulted in:

* Faster model training time
* Lower overfitting risk
* Improved data visualization and feature interpretability

The Support Vector Classifier, known for its effectiveness in small-to-medium datasets, performed robustly in both cases. However, PCA offered added value in terms of data exploration and feature compression, which is critical when scaling up to larger datasets or deploying models in resource-constrained environments.

PCA trade-offs are typical in real-world modeling: a slight loss in predictive power is often acceptable for the gain in interpretability, computational efficiency, and robustness.

## **CONCLUSION**

PCA proved to be a valuable tool in reducing the dimensionality of banana flour data, enhancing visualization and preparing the data for classification. It helped uncover structure among samples and allowed new data points to be projected for predictive classification. While PCA did not significantly improve accuracy over raw data, it offered a more interpretable representation and reduced feature complexity.

## **REFERENCES**

1. Koome, M. (2025). *Banana flour PCA classification* (v1.0) [Multivariate Methods]. GitHub.

<https://github.com/KoomeMartin/Multivariate-Methods-> (Full project Documentation).

1. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101
2. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202
3. Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303–304. https://doi.org/10.1038/nbt0308-303