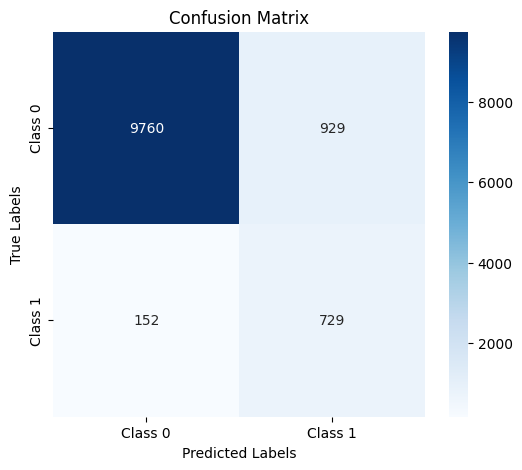
|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Project Report  Isra Mansoor /Muhammad Abdullah / ML Project/ 17/12/2024 |  |
|  |  |  |  |
|  | | | |

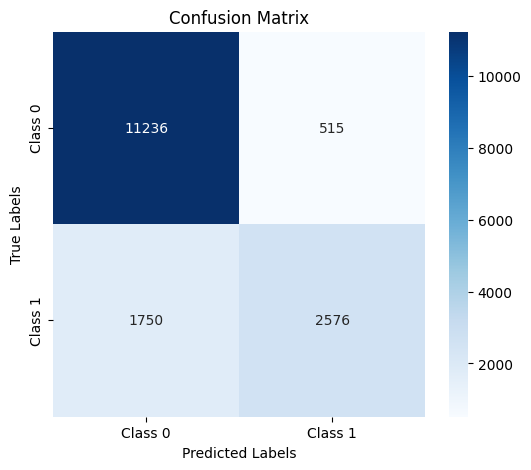
|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | |
|  |  | TABLE OF CONTENTS  |  |  | | --- | --- | | Contents……………………………………………Page# | | | Introduction | 3 | | Data Analysis | 3 | | Data Preprocessing | 6 | | Datasets | 9 | | Supervised Learning Algos | 12 | | Unsupervised Techniques | 29 | |  |

Github Repository:  
  
https://github.com/isramansoor9/404520\_Isra-Mansoor

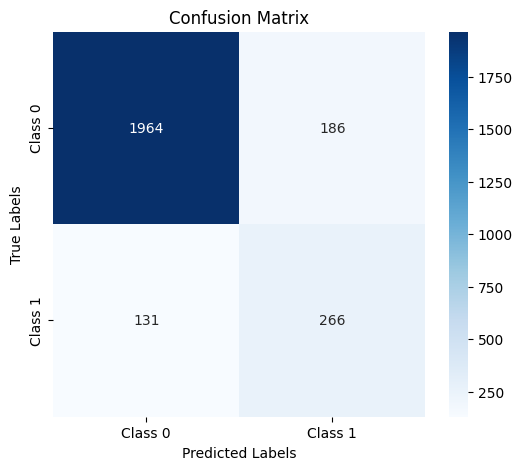
|  |  |  |
| --- | --- | --- |
| INTRODUCTION |  | |
| Agriculture forms the backbone of many economies, and effective crop classification is crucial for optimizing resource allocation, enhancing agricultural management, and guiding policy formulation. This project focuses on classifying two vital crop types—rice and cotton—using their spectral characteristics as captured by the Normalized Difference Vegetation Index (NDVI). The NDVI, a widely recognized vegetation health indicator, is calculated using Near-Infrared (NIR) and Red light reflectance. Its ability to differentiate vegetation health and growth patterns makes it a powerful tool for crop classification.  The study leverages the NDVI’s capacity to track the unique growth cycles of rice and cotton over a growing season. Rice typically exhibits a later peak NDVI in September, reflecting its extended growth period, while cotton shows an earlier NDVI peak around July to August. These distinct temporal patterns allow for precise identification and classification of the two crops, which is critical for improving yield predictions, resource distribution, and decision-making in the agricultural sector.  To achieve this, a time-series dataset containing NDVI measurements for rice and cotton crops across three consecutive years (2021–2023) is analyzed. The dataset includes NDVI values recorded twice a month over six months, encapsulating the full growing season.  We applied both Supervised Machine Learning Algorithms and Unsupervised Machine Learning Algorithms to check their performances. For cross validation first we trained on 2021,2022 and tested in 2023, trained on 2022,2023 and tested on 2021 and trained on 2021,2023 and tested on 2022 data.  Dataset Analysis  The dataset used in this project consists of NDVI time-series values collected for rice and cotton crops over three consecutive years—2021, 2022, and 2023. The dataset was highly imbalanced with different number of samples for both crops across each year.  The table below shows the distribution of data for each year:   |  |  |  |  | | --- | --- | --- | --- | | Year | 2021 | 2022 | 2023 | | Cotton | 2884 | 12412 | 11778 | | Rice | 420 | 4688 | 920 |   The dataset includes **12 NDVI values per instance**, representing bi-monthly observations over a six-month growing season.  The Average NDVI values generally range between **0 and 0.85**, indicating typical vegetation health measurements.  Average Values of Rice 2023  Average Values of Cotton 2021    Kernel density plots were used to analyze NDVI distributions across all datasets. In the **Cotton 2022** dataset (see side figure), features NDVI02 and NDVI10 show a rise in density for values between **0.0 and 0.2**, reflecting distinct growth stages. Similarly different trends, with varying densities, were observed in other datasets, highlighting the unique spectral patterns of rice and cotton crops. These visualizations were crucial for understanding crop dynamics and guiding data preprocessing for classification.  To evaluate skewness across datasets, histograms were plotted. In the **Rice 2022** dataset (see figure below), NDVI features showed varying skewness: **NDVI01–NDVI07** displayed positive skewness, while **NDVI08–NDVI12** exhibited negative skewness. This indicates differences in data distribution within the dataset.  Additionally, other datasets showed diverse skewness patterns, with some features being symmetrically distributed. These variations highlight the need for targeted preprocessing, as skewness can impact the performance of machine learning models.    Boxplots were generated to analyze the presence of outliers across all datasets. In the **Cotton 2023** dataset (see figure), features such as **NDVI06**, **NDVI04**, and **NDVI12** were observed to have no outliers, while others exhibited varying degrees of outlier presence.  Similar mixed trends were noted in other rice and cotton datasets: some features consistently lacked outliers, while others displayed outliers in certain years but not in others. These variations emphasize the need for dataset-specific outlier handling to ensure data quality.  Violin plots were generated to visualize the distribution of NDVI features across datasets. These plots combine boxplot characteristics with a kernel density estimate, providing a detailed view of data distribution and spread.  In the **Rice 2023** dataset (see figure below), the features exhibited varied trends: some showed symmetrical distributions, while others displayed skewed or multimodal patterns. This highlights the diversity in NDVI feature behavior, even within a single dataset.  The datasets were analyzed to identify duplicate rows and missing values. While no missing values were detected in any dataset, duplicate rows were found in some, with **Cotton 2021** showing the highest count (537 duplicates) and others like **Rice 2023** having none. Addressing these duplicates is essential to maintain data quality and improve the reliability of analysis.   |  |  |  |  | | --- | --- | --- | --- | | Year | 2021 | 2022 | 2023 | | Cotton Missing Values | 0 | 0 | 0 | | Cotton Duplicates | 537 | 64 | 1 | | Rice Missing Values | 0 | 0 | 0 | | Rice Duplicates | 0 | 10 | 0 |   Data Preprocessing  Data preprocessing is a crucial step to improve dataset quality and ensure consistency for effective machine learning. Given the insights from the data visualization section, it was clear that preprocessing was essential to address issues like outliers, duplicates, and class imbalance. Techniques such as scaling and outlier removal were applied to refine the datasets, while augmentation and SMOTE were used to balance the data, resulting in a more robust foundation for classification.  In the preprocessing implementation, numerical features were scaled using MinMaxScaler to standardize the NDVI values, and missing values were imputed with median values where necessary. Duplicate rows were removed to avoid redundancy, and outliers were handled using the Interquartile Range (Z-Score) method. Finally, label columns were added to distinguish between crop types, and datasets for each year were combined for streamlined analysis. These steps ensured the datasets were clean, balanced, and ready for machine learning workflows.  Sliding window data augmentation ...Data augmentation was applied to enhance the dataset and address class imbalance using a sliding window approach. This technique generates new samples by calculating the rolling average within a defined window size, ensuring that temporal relationships in the time-series data are preserved. The sliding\_window\_augment function was used to augment the datasets by iterating through and creating new samples until the target size was reached. For cotton, the target size matched the largest dataset, while for rice, it was set to approximately 40% of that, resulting in balanced datasets for more effective model training.    Sliding Window Diagram    Average Values of Cotton 2021  Average Values of Rice 2023  After preprocessing, the NDVI values for both rice and cotton datasets were scaled to fall between 0 and 1, ensuring consistency across all features. This scaling step standardizes the data, making it suitable for machine learning models, as it eliminates biases that could arise from different feature ranges. The visualizations of rice 2023 and cotton 2021 (See above figures) post-scaling demonstrate how the NDVI values are now distributed within this normalized range, offering a clearer view of crop health and growth patterns for further analysis and classification tasks.  After applying the preprocessing steps, including scaling, the density distributions of NDVI features for both rice and cotton datasets have shown notable changes. For instance, the density values of NDVI01 and NDVI10 now exhibit peaks between the 0.0 and 0.2 range with densities of 12 and 14, respectively. This shift can be attributed to the data augmentation techniques, such as sliding window augmentation, which introduced more instances of lower NDVI values. This process helps balance the dataset by generating new samples, leading to a higher concentration of data points within these value ranges. These changes reflect how preprocessing, and augmentation help modify the distribution and enhance model training.  Histograms were plotted again to evaluate the skewness across datasets. In the case of Rice 2022 (see below figure), the NDVI features now show a more balanced distribution with reduced skewness. Previously, some features exhibited positive or negative skewness, indicating uneven data distributions, but after preprocessing, the data appears more symmetrical and uniform. This trend is  observed across all other datasets as well. These improvements in data quality emphasize the effectiveness of preprocessing and ensure better suitability for machine learning models, reducing potential biases.  Boxplots were generated to analyze the presence of outliers across all datasets, and outliers were reduced using the IQR method. As seen in the figure, the removal of outliers has improved the distribution of the data. This trend is consistent across the remaining datasets as well, where outliers were effectively reduced, ensuring cleaner and more reliable data for analysis. The features also showed symmetrical distribution now.  Datasets  The datasets were labeled to distinguish between cotton and rice, with a label of '0' for cotton and '1' for rice. The datasets for each year (2021, 2022, and 2023) were then concatenated to combine both cotton and rice data into a single dataset for each year. These combined datasets—**year\_2021\_copy**, **year\_2022\_copy**, and **year\_2023\_copy**—were only preprocessed with MinMax scaling, duplicate removal, and outlier handling. These "copy" datasets are intended to be used for testing, and thus were kept in their original form without any augmentation. We also scaled the preprocessed and combined year datasets using the MinMax scaler again.  On the other hand, the remaining datasets were augmented to increase the sample size for training. This augmentation process ensures that we have a more balanced dataset, and further, Synthetic Minority Oversampling Technique (SMOTE) will be applied to address class imbalance between cotton and rice. These augmented datasets will be used for training the machine learning models, ensuring that the training data is enriched and better representative of the overall data distribution.  Visualizations of these datasets, including line plots of average NDVI values, density plots, histograms, box plots, and violin plots, were generated. These plots highlight the uniform distributions and improved data quality achieved through preprocessing, with similar trends observed.  During feature analysis, we visualized the correlation matrices of the yearly datasets (2021, 2022, and 2023) to identify highly correlated features. A notable observation was that **NDVI03, NDVI10, NDVI12** consistently showed a high correlation (e.g., 0.85 in the 2023 dataset). The heatmaps clearly highlighted this, and visual evidence will be provided for Year 2023.  We then calculated the Variance Inflation Factor (VIF) values for all features across datasets to quantify multicollinearity. **NDVI03** exhibited significantly high VIF values, further confirming its redundancy in the feature set.  To assess its impact, we tested models with and without **NDVI03**. Results indicated better performance after removing the feature, reinforcing our dec ision to drop it. Consequently, **NDVI03** was removed from all datasets, including both training (augmented) and testing dat asets, to enhance model performance and ensure a robust analysis.  Initially, we examined the sample distribution within each dataset. After augmentation, the cotton datasets for all three years contained **9400 samples**, while rice datasets had **2865 samples**. These values were determined by the maximum sample size across the three years after preprocessing—**9400** for cotton and **2865** for rice. The smaller years were augmented to match these values. Despite augmentation, the datasets remained imbalanced, necessitating further adjustments.  To address this imbalance, we applied SMOTE (Synthetic Minority Oversampling Technique) to increase the rice samples to **6580**, which is approximately 70% of the cotton sample size. This ratio was chosen as it struck a balance between addressing imbalance and avoiding the generation of excessive synthetic data, ensuring better generalization for supervised models.  The distribution of the label feature before SMOTE was visualized for each year, highlighting the imbalance. After applying SMOTE, the datasets achieved a more balanced structure, optimizing them for training supervised learning models.    Number of Samples after SMOTE  Number of Samples before SMOTE  Evaluation Metrix for Supervised Machine Learning Models  In our project, we used the following metrics to evaluate model performance, keeping in mind the dataset's imbala nce, where rice samples are fewer than cotton samples:  Accuracy: Provides an overall measure of performance but can be misleading due to the imbalance in the dataset.  Precision, Recall, and F1-Score: Recall is crucial to ensure rice samples (minority class) are correctly identified, while F1-score balances recall and precision, making it ideal for imbalanced data.  Confusion Matrix: Helps visualize the model's errors and understand misclassification patterns between cotton and rice.  ROC Curve and AUC: Essential for assessing the model's ability to distinguish between cotton and rice, especially under imbalanced conditions.  These metrics together ensure a comprehensive evaluation, balancing overall performance with sensitivity to minority class predictions.  Training Testing Strategy  Cross-Validation Combinations:   * Train on Year 1 and Year 2, Test on Year 3 * Train on Year 1 and Year 3, Test on Year 2 * Train on Year 2 and Year 3, Test on Year 1   By iteratively training on two years and testing on the remaining year, the model's performance is validated on unseen data, reflecting its ability to handle diverse growing conditions over time.  Evaluation Metrix for Unsupervised Learning Approach  In the unsupervised learning section, we evaluated clustering using the following methods:  PCA Visualization: Reduced data to two components for scatter plots comparing predicted clusters with true labels, revealing clustering patterns.  Confusion Matrix: Assessed alignment between predicted clusters and ground truth labels to evaluate meaningful groupings.  Silhouette Analysis: Measured clustering quality with average silhouette scores and visualized distribution to identify cluster separability.  Cluster Purity: Calculated to quantify how closely clusters match the true labels, providing a straightforward measure of clustering accuracy.  These techniques provided a comprehensive evaluation of the clustering algorithm's performance and interpretability.  Supervised Learning Algorithms **XGBoost (Extreme Gradient Boosting)**  |  |  |  |  | | --- | --- | --- | --- | | Year | Train (2021,22) Test 2023 | Train (2021,23) Test 2022 | Train (2023,22) Test 2021 | | Accuracy | 90.90 | 85.81 | 86.21 | | ROC Curve Area | 93 | 88 | 85 | | F1-Score (Cotton) | 95 | 91 | 92 | | F1-Score (Rice) | 58 | 69 | 60 | | Recall (Cotton) | 92 | 96 | 90 | | Recall (Rice) | 83 | 58 | 65 | | Precision (Cotton) | 98 | 86 | 93 | | Precision (Rice) | 45 | 84 | 55 |  **Training on year\_2021 and year\_2022, Testing on year\_2023**  **Training on year\_2021 and year\_2023, Testing on year\_2022** | |  |
| **Training on year\_2023 and year\_2022, Testing on year\_2021** | |  |
| **XGBoost (GRID SEARCH)**  |  |  |  |  | | --- | --- | --- | --- | | Year | Train (2021,22) Test 2023 | Train (2021,23) Test 2022 | Train (2023,22) Test 2021 | | Best Parameters | learning\_rate: 0.4, max\_depth: 7, n\_estimators: 175 | learning\_rate: 0.4, max\_depth: 8, n\_estimators: 175 | learning\_rate: 0.4, max\_depth: 8, n\_estimators: 175 | | Accuracy | 90.65 | 85.91 | 89.55 | | ROC Curve Area | 93 | 89 | 87 | | F1-Score (Cotton) | 95 | 91 | 93 | | F1-Score (Rice) | 57 | 69 | 63 | | Recall (Cotton) | 91 | 96 | 91 | | Recall (Rice) | 83 | 60 | 67 | | Precision (Cotton) | 98 | 87 | 94 | | Precision (Rice) | 44 | 83 | 59 |  **Training on year\_2021 and year\_2022, Testing on year\_2023** | |  |



## **Training on year\_2021 and year\_2023, Testing on year\_2022**



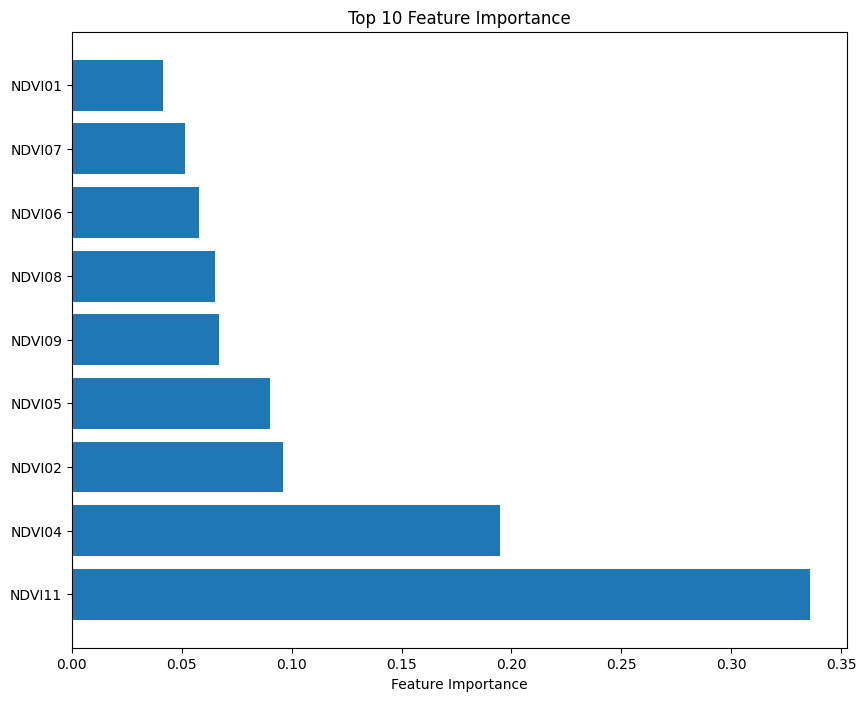
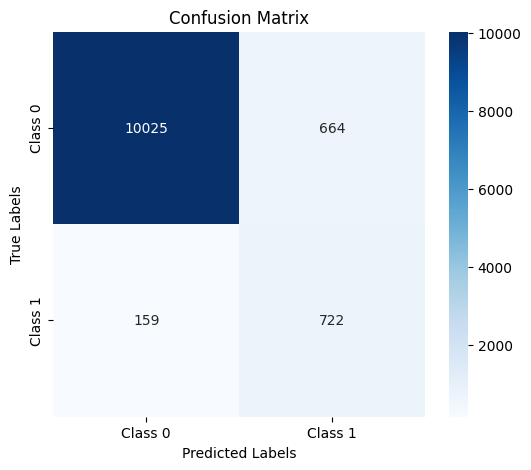
## **Training on year\_2022 and year\_2023, Testing on year\_2021**



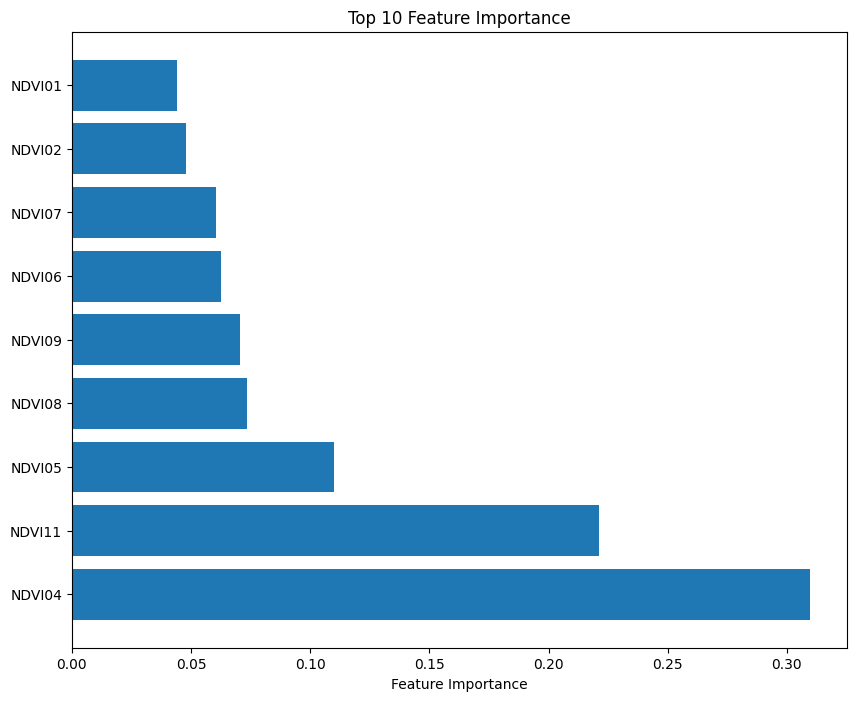
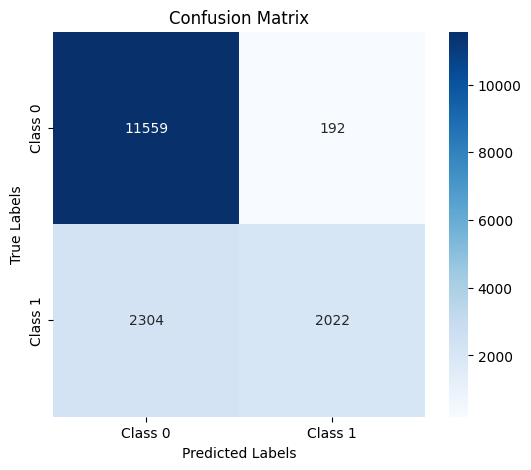
## **Random Forest**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Train (2021,22) Test 2023 | Train (2021,23) Test 2022 | Train (2023,22) Test 2021 |
| Accuracy | 92.89 | 84.47 | 89.48 |
| ROC Curve Area | 94 | 91 | 91 |
| F1-Score (Cotton) | 96 | 90 | 94 |
| F1-Score (Rice) | 64 | 62 | 66 |
| Recall (Cotton) | 94 | 98 | 94 |
| Recall (Rice) | 82 | 47 | 66 |
| Precision (Cotton) | 98 | 83 | 94 |
| Precision (Rice) | 52 | 91 | 66 |

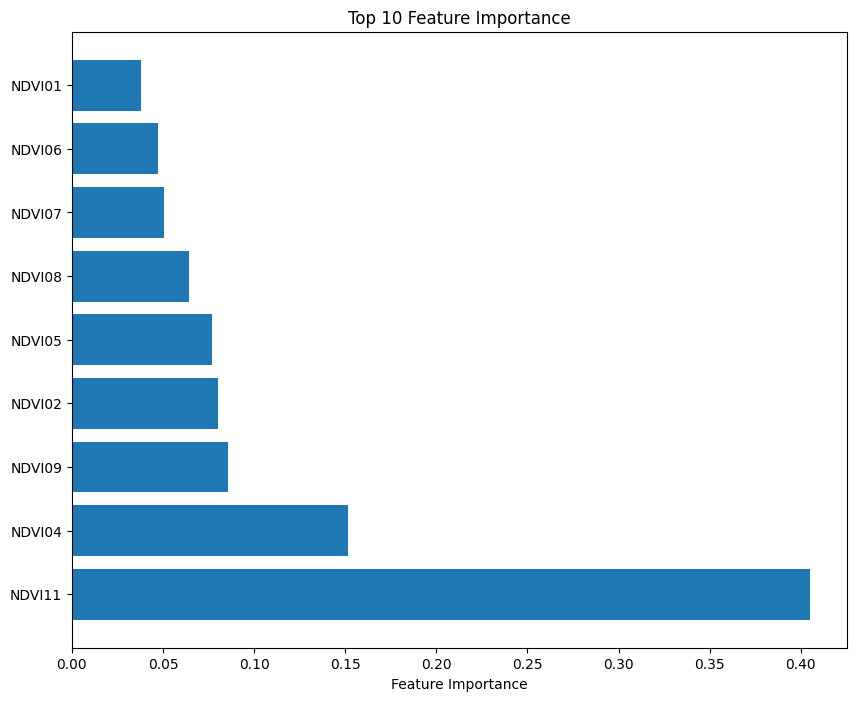
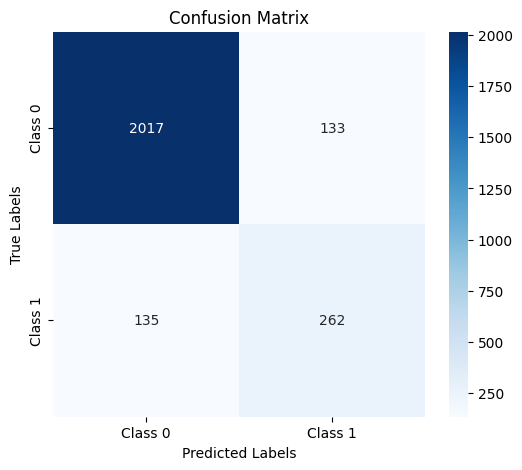
## **Training on year\_2021 and year\_2022, Testing on year\_2023**



## **Training on year\_2021 and year\_2023, Testing on year\_2022**



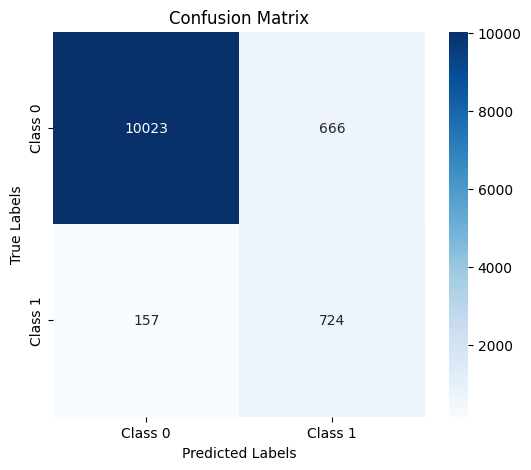
## **Training on year\_2022 and year\_2023, Testing on year\_2021**



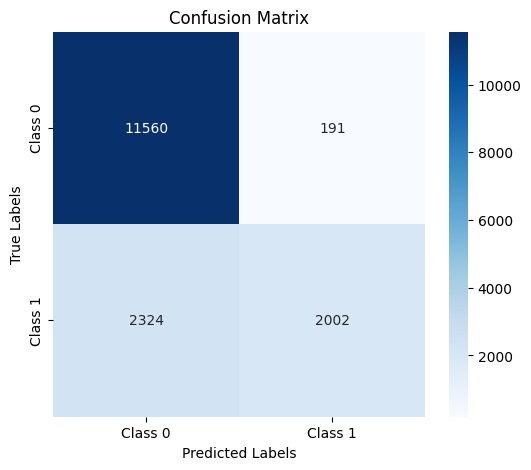
## **Random Search (Grid Search)**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Train (2021,22) Test 2023 | Train (2022,23) Test 2021 | Train (2023,21) Test 2022 |
| Best Parameters | 'max\_depth': 30, 'min\_samples\_leaf': 1, 'n\_estimators': 150 | max\_depth': 30, 'min\_samples\_leaf': 1, 'n\_estimators': 50 | 'max\_depth': 30, 'min\_samples\_leaf': 1, 'n\_estimators': 100 |
| Accuracy | 92.89 | 84.36 | 89.51 |
| ROC Curve Area | 94 | 91 | 91 |
| F1-Score (Cotton) | 96 | 90 | 94 |
| F1-Score (Rice) | 64 | 61 | 66 |
| Recall (Cotton) | 94 | 98 | 94 |
| Recall (Rice) | 82 | 46 | 65 |
| Precision (Cotton) | 98 | 83 | 94 |
| Precision (Rice) | 52 | 91 | 67 |

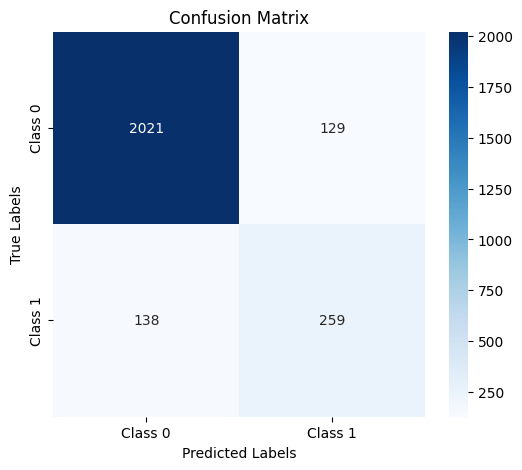
## **Training on year\_2021 and year\_2022, Testing on year\_2023**



## **Training on year\_2022 and year\_2023, Testing on year\_2022**



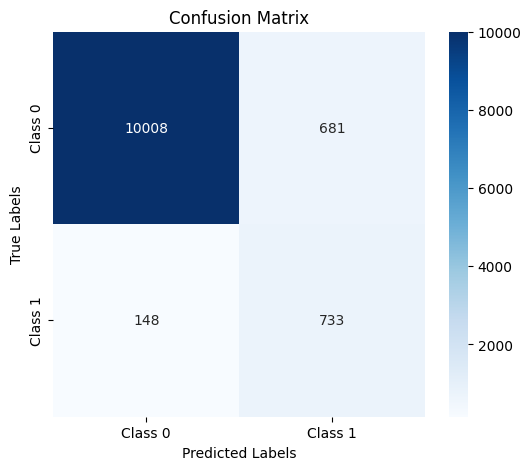
## **Training on year\_2022 and year\_2023, Testing on year\_2021**



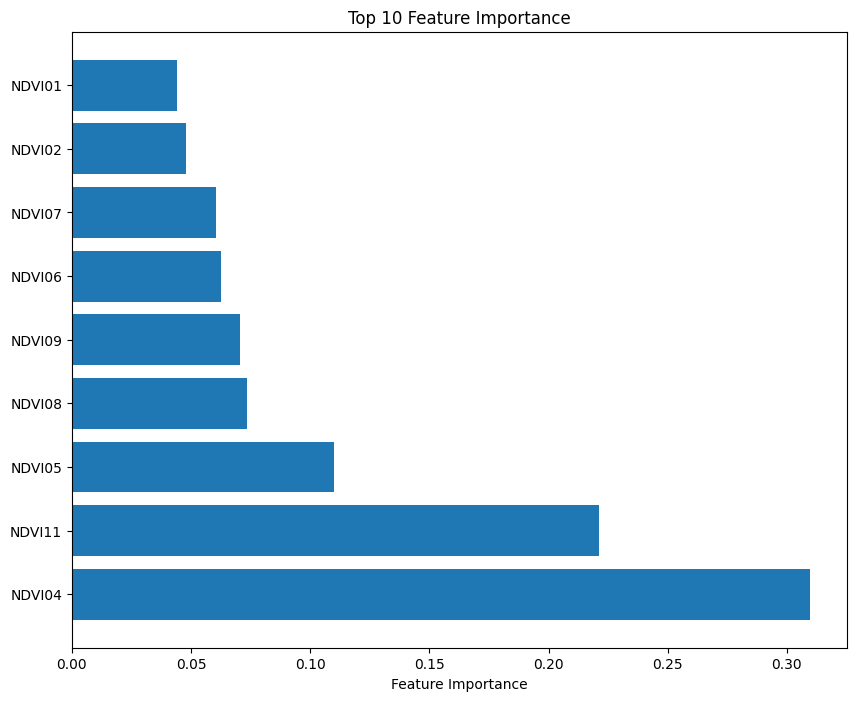
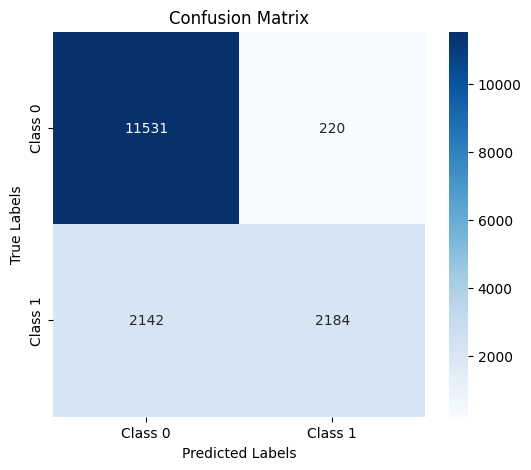
## **Bagging**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Train (2021,22) Test 2023 | Train (2021,23) Test 2022 | Train (2023,22) Test 2021 |
| Accuracy | 92.83 | 85.30 | 89.40 |
| ROC Curve Area | 95 | 92 | 92 |
| F1-Score (Cotton) | 96 | 91 | 94 |
| F1-Score (Rice) | 64 | 65 | 66 |
| Recall (Cotton) | 94 | 98 | 94 |
| Recall (Rice) | 83 | 50 | 66 |
| Precision (Cotton) | 99 | 84 | 94 |
| Precision (Rice) | 52 | 91 | 66 |

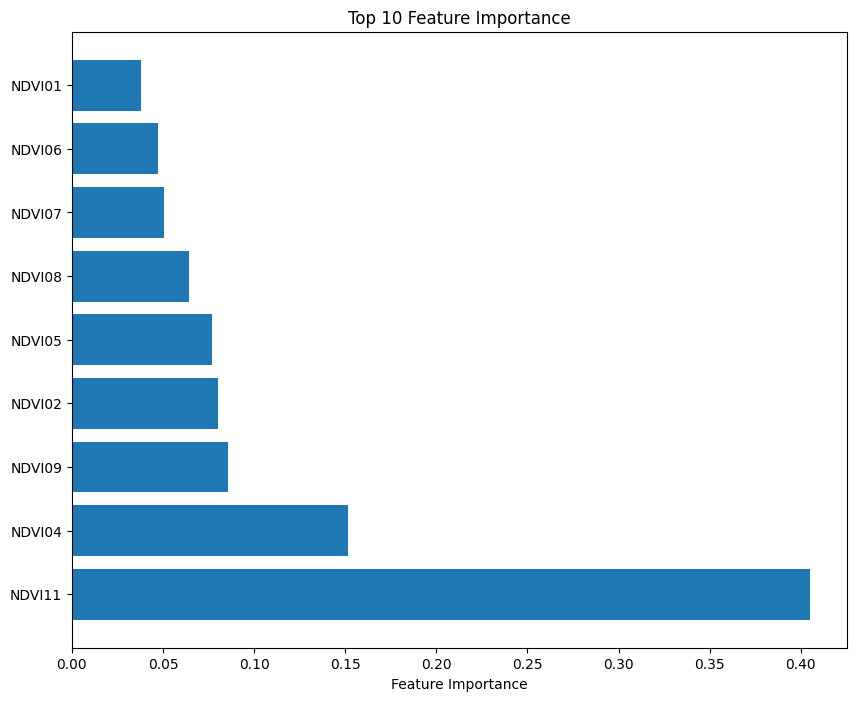
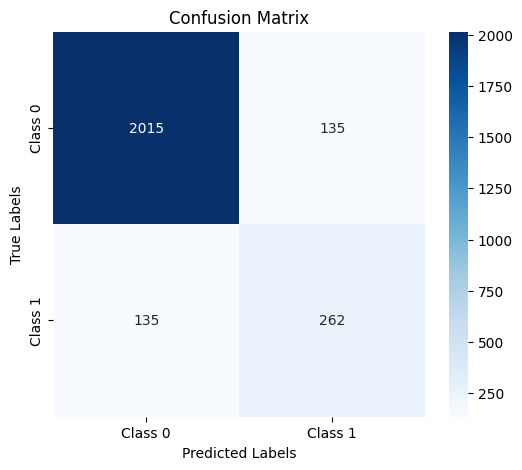
## **Training on year\_2021 and year\_2022, Testing on year\_2023**

## **Training on year\_2021 and year\_2023, Testing on year\_2022**



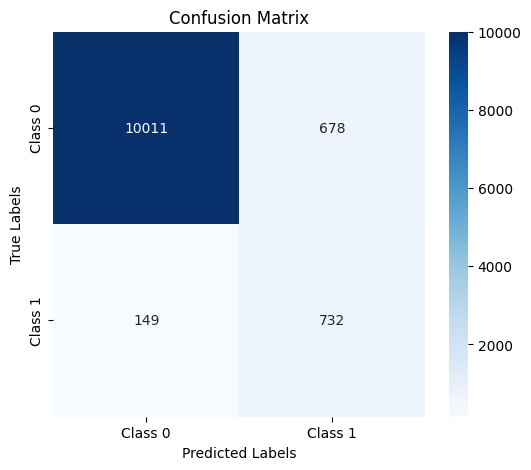
## **Training on year\_2022 and year\_2023, Testing on year\_2021**



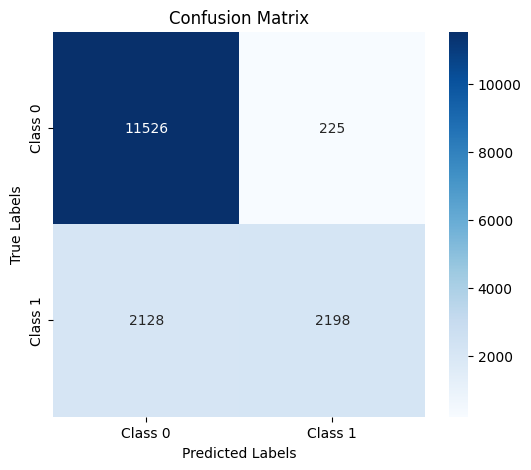
## **Random Search (Grid Search)**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Train (2021,22) Test 2023 | Train (2022,23) Test 2021 | Train (2023,21) Test 2022 |
| Best Parameters | 'estimator': RandomForestClassifier(max\_depth=30, n\_estimators=10, random\_state=42), 'n\_estimators': 50 | 'estimator': RandomForestClassifier(max\_depth=30, n\_estimators=10, random\_state=42), 'n\_estimators': 50 | 'estimator': RandomForestClassifier(max\_depth=30, n\_estimators=10, random\_state=42), 'n\_estimators': 50 |
| Accuracy | 92.85 | 85.36 | 89.36 |
| ROC Curve Area | 94 | 92 | 92 |
| F1-Score (Cotton) | 96 | 91 | 94 |
| F1-Score (Rice) | 64 | 65 | 66 |
| Recall (Cotton) | 94 | 98 | 94 |
| Recall (Rice) | 83 | 51 | 65 |
| Precision (Cotton) | 99 | 84 | 94 |
| Precision (Rice) | 52 | 91 | 66 |

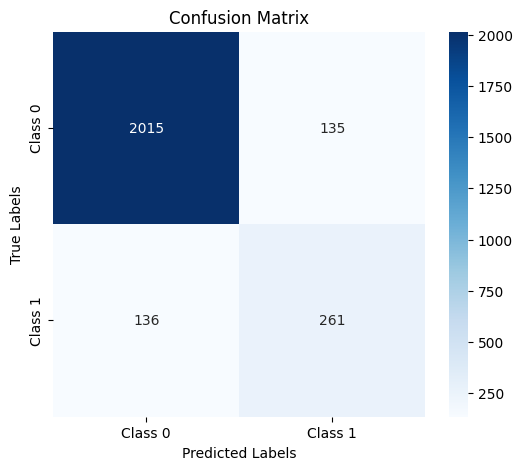
## **Training on year\_2021 and year\_2022, Testing on year\_2023**



## **Training on year\_2022 and year\_2023, Testing on year\_2022**



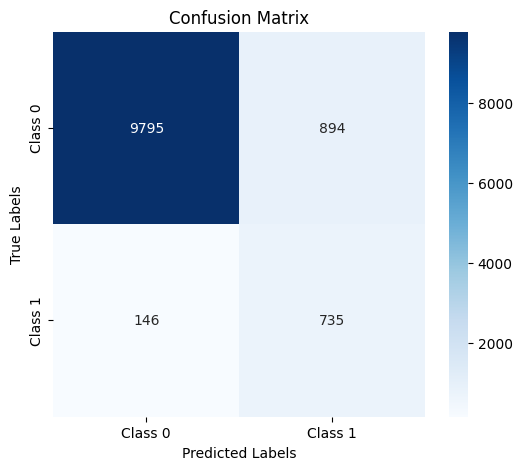
## **Training on year\_2022 and year\_2023, Testing on year\_2021**



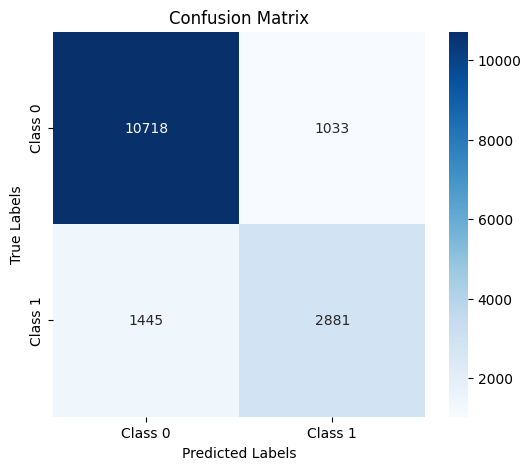
## **SVM**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Train (2021,22) Test 2023 | Train (2021,23) Test 2022 | Train (2023,22) Test 2021 |
| Accuracy | 91.01 | 84.59 | 87.08 |
| ROC Curve Area | 94 | 88 | 85 |
| F1-Score (Cotton) | 95 | 90 | 92 |
| F1-Score (Rice) | 59 | 70 | 60 |
| Recall (Cotton) | 92 | 91 | 92 |
| Recall (Rice) | 83 | 67 | 63 |
| Precision (Cotton) | 99 | 88 | 93 |
| Precision (Rice) | 45 | 74 | 58 |

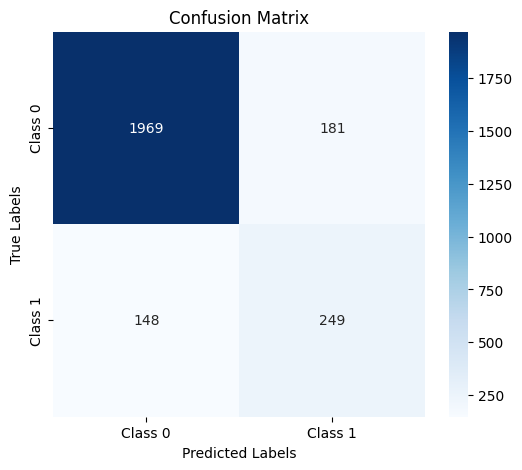
## **Training on year\_2021 and year\_2022, Testing on year\_2023**



## **Training on year\_2021 and year\_2023, Testing on year\_2022**



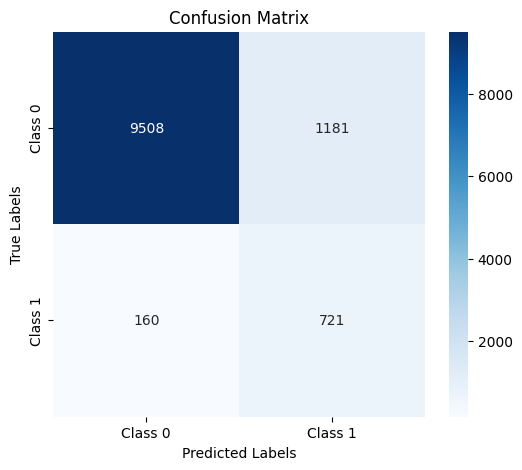
## **Training on year\_2022 and year\_2023, Testing on year\_2021**



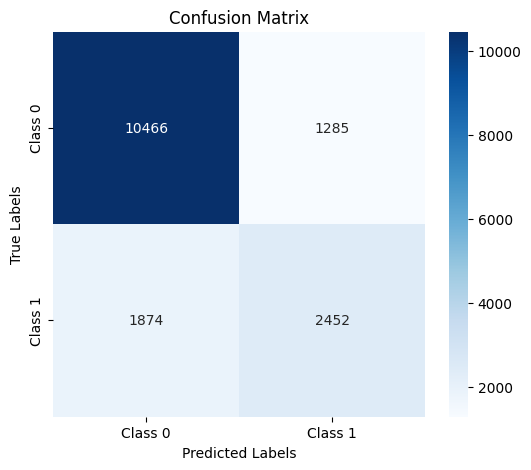
## **SVM (Grid Search)**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Train (2021,22) Test 2023 | Train (2022,23) Test 2021 | Train (2023,21) Test 2022 |
| Best Parameters | 'C': 100, 'gamma': 1, 'kernel': 'rbf' | 'C': 100, 'gamma': 1, 'kernel': 'rbf' | 'C': 100, 'gamma': 1, 'kernel': 'rbf' |
| Accuracy | 88.40 | 80.35 | 82.49 |
| ROC Curve Area | 91 | 78 | 75 |
| F1-Score (Cotton) | 93 | 87 | 89 |
| F1-Score (Rice) | 52 | 61 | 52 |
| Recall (Cotton) | 89 | 89 | 87 |
| Recall (Rice) | 82 | 57 | 60 |
| Precision (Cotton) | 98 | 85 | 92 |
| Precision (Rice) | 38 | 66 | 45 |

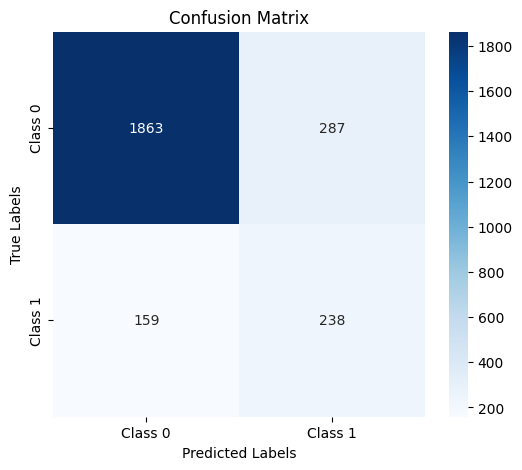
## **Training on year\_2021 and year\_2022, Testing on year\_2023**



## **Training on year\_2022 and year\_2023, Testing on year\_2022**



## **Training on year\_2022 and year\_2023, Testing on year\_2021**



## **Analysis of models**

XGBoost generally offers strong performance with high accuracy (85.81%–90.90%) and ROC curve areas (85–93). It excels in predicting Cotton with high precision (93–98), but struggles with Rice, where precision is much lower (44–84). Its F1-Score for Cotton is consistently high (91–95), while Rice shows significant variation (57–69). Grid search optimization improves accuracy slightly, especially for Cotton, but does not drastically change the model's performance.

Random Forest delivers comparable accuracy (84.47%–92.89%) and ROC curve areas (91–94). It performs well with Cotton in terms of precision (83–98) and F1-Score (90–96), but its Rice performance fluctuates, with precision ranging from 52 to 91. Grid search tuning, with parameters like max\_depth of 30, results in minor improvements but doesn't drastically alter overall performance.

Bagging maintains consistent accuracy (85.30%–92.83%) and ROC curve areas (92–95), showing strong precision for Cotton (94–99). However, its performance with Rice remains modest, with precision ranging from 52 to 66. The model's F1-Score for Cotton is strong (90–96), while Rice's remains lower. Grid search does not significantly change its results but helps fine-tune the model.

SVM shows the lowest accuracy (80.35%–91.01%) and ROC curve areas (75–91) among the four models. Its precision for Cotton is high (85–98), but for Rice, it struggles with much lower precision (38–74). The F1-Score for Cotton is fairly consistent (87–93), but for Rice, it varies significantly (52–61). Grid search improves the model slightly with parameters like 'C' set to 100 and 'gamma' at 1, but it still lags behind other models in performance.

Based on the performance metrics, **XGBoost** appears to be the best model overall, especially when focusing on Cotton classification. It consistently achieves high accuracy (85.81%–90.90%), a strong ROC curve area (85–93), and excellent precision and F1-Score for Cotton (93–98 for precision and 91–95 for F1-Score). While its performance on Rice is not as strong as on Cotton, it still offers reasonable results compared to the other models, especially in terms of recall and precision. The ability to fine-tune hyperparameters through grid search also enhances its performance, making it more adaptable to different datasets.

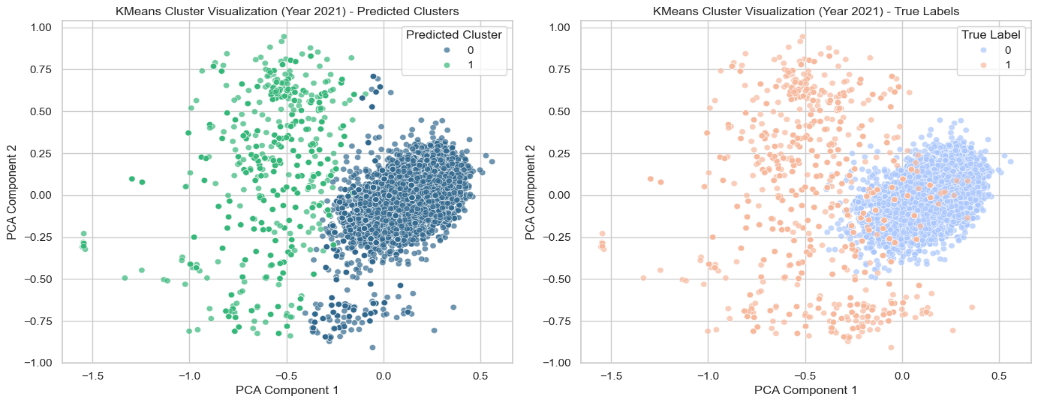
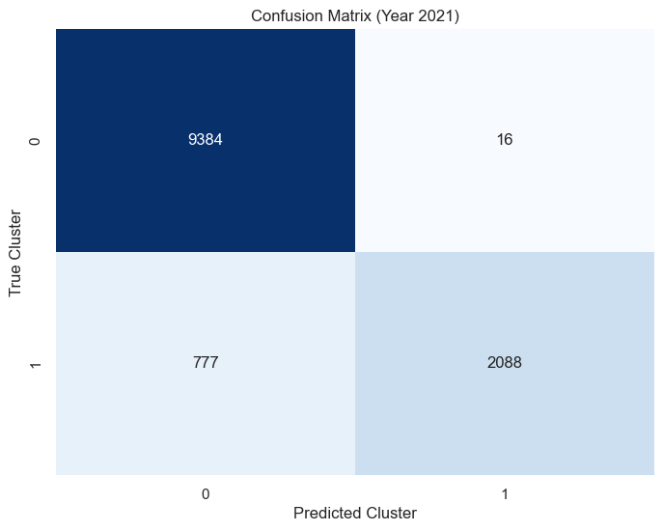
**Random Forest** and **Bagging** are also strong contenders, with stable accuracy and good precision for Cotton, but XGBoost outperforms them in terms of overall accuracy and F1-Score for Cotton, which is crucial in many real-world applications.

**SVM**, on the other hand, lags behind in accuracy, ROC curve area, and performance on Rice, making it less effective for this particular task. Therefore, XGBoost stands out as the best model for its balance of high performance, adaptability, and fine-tuning potential.

Unsupervised Learning Approach

K-means Clustering

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Number of Cluster | **2** | **2** | **2** |
| Cluster Purity | **93.53** | **0.9337** | **0.9668** |
| Sillhouette Score | **0.447** | **0.2501** | **0.3765** |

Year 2021

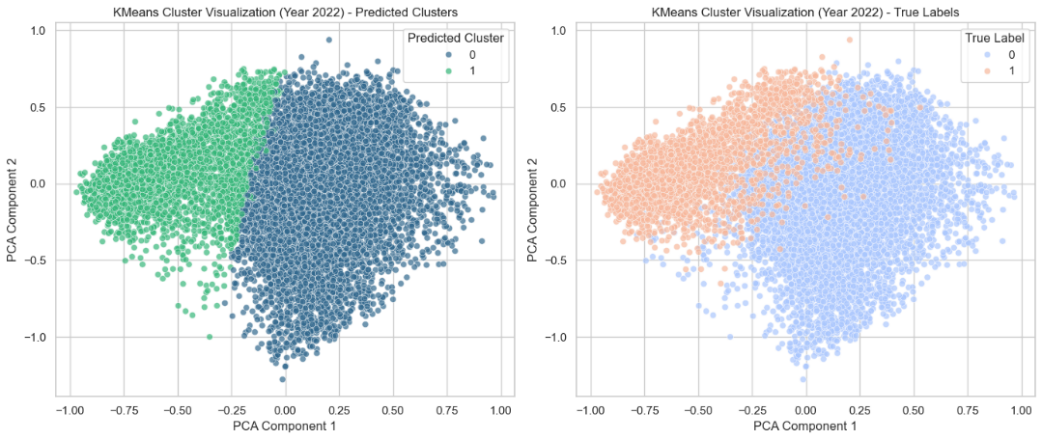
The predicted clusters (left plot) align well with the true labels (right plot), demonstrating good separability. Clusters are distinct with minimal overlap, indicating clear grouping in the dataset.

**Cluster Purity**: High at **93.53%**, suggesting the majority of points in each cluster share the same label.

**Silhouette Score**: **0.447**, indicating moderately well-defined clusters.

The model performed effectively on 2021, capturing the inherent structure of the data.

A blue squares with white text

Description automatically generatedYear 2022

Predicted clusters still align with true labels, but there appears to be more overlap compared to 2021.The separability between clusters is less distinct.

**Cluster Purity**: Still high at **93.37%,** but slightly lower than 2021, showing slight degradation in clustering quality.

**Silhouette Score: 0.2501**, indicating weaker cluster cohesion and separation compared to the previous year.

Conclusion: Clustering performance was slightly less robust, possibly due to increased data complexity or noise.

A screenshot of a graph

Description automatically generatedA blue squares with white text

Description automatically generatedYear 2023

The predicted clusters closely align with true labels, with improved cluster separability compared to 2022.The distinct separation seen in the PCA plot indicates better-defined clusters.

**Cluster Purity**: Improved to **96.68%**, reflecting a higher proportion of correctly grouped points.

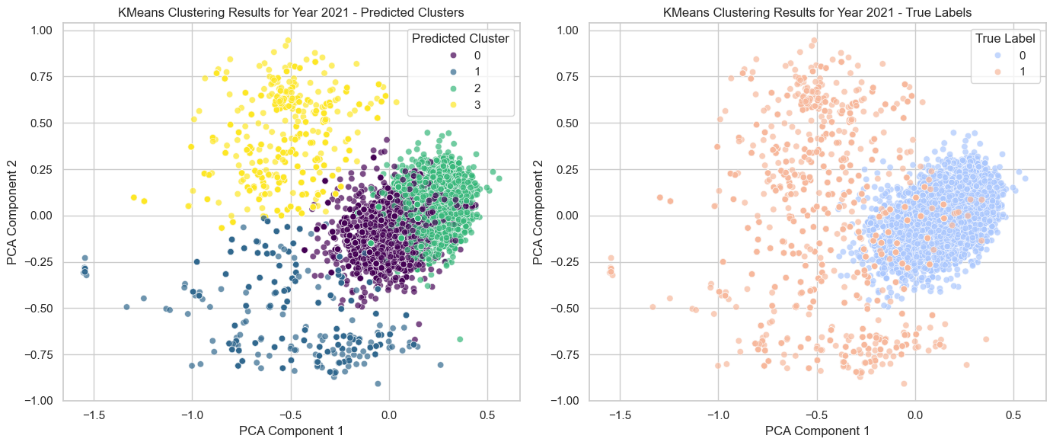
**Silhouette Score**: **0.3765**, indicating improved cohesion and separation compared to 2022.

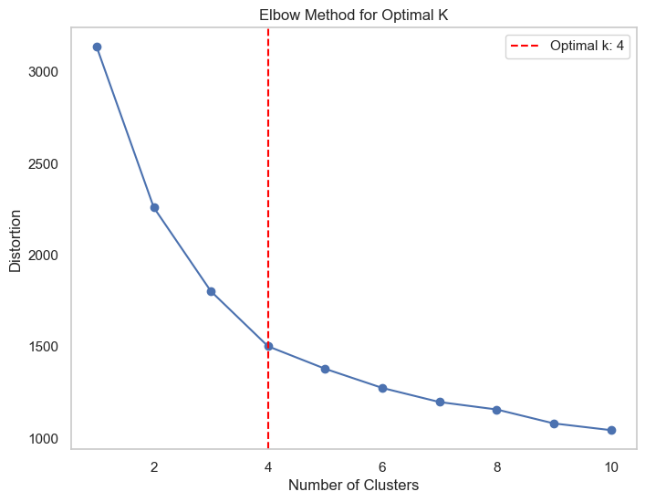
Clustering quality improved significantly in 2023, likely due to refinements in the dataset or model adjustments.

K-means (Grid Search Clustering)

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Number of Cluster | **4** | **3** | **3** |
| Cluster Purity | **0.9699** | **0.9181** | **0.9613** |
| Sillhouette Score | **0.2420** | **0.2536** | **0.2157** |

A screenshot of a graph

Description automatically generatedYear 2021

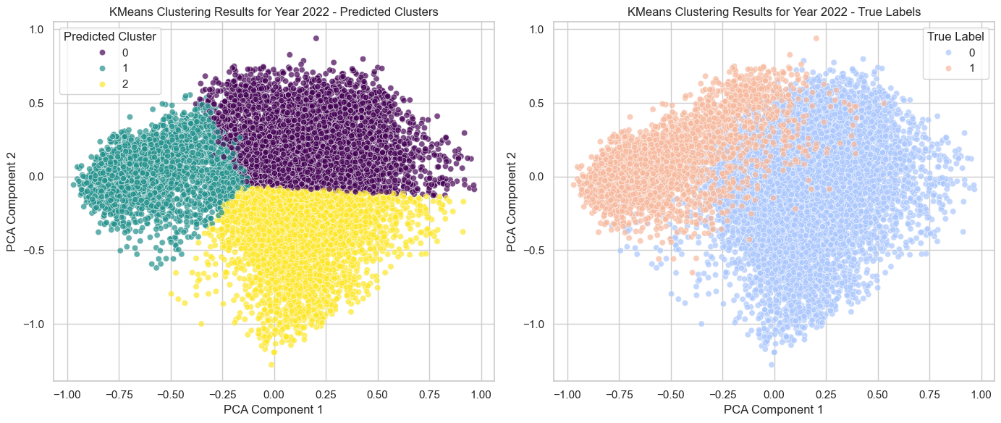


For the 2021 dataset, the optimal number of clusters was determined to be **4** using the elbow method. This clustering likely reflects subcategories or distinct varieties within the rice and wheat samples.

**Cluster Purity**: **0.9699**, indicating strong alignment between clusters and the true labels, showcasing the model's accuracy.

**Silhouette Score**: **0.2420**, suggesting moderate cohesion and separation between clusters. The slightly lower score might result from the higher number of clusters introducing some overlap.

A blue squares with numbers

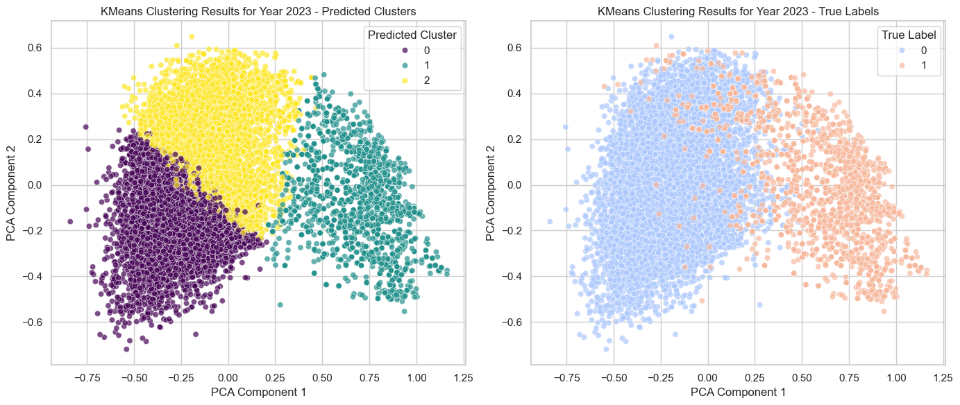
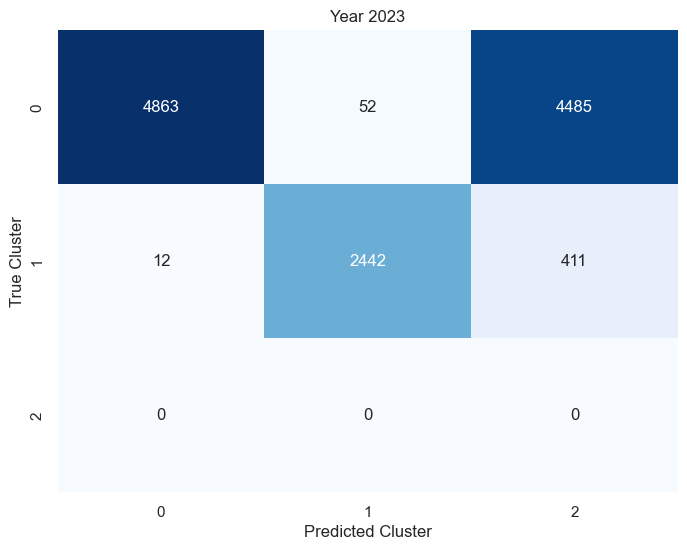
Description automatically generatedYear 2022

A graph with a line

Description automatically generatedIn 2022, the elbow method identified the optimal clusters as **3**, possibly due to reduced variability or overlaps in the dataset.

**Cluster Purity: 0.9181**, indicating good performance but slightly lower than in 2021, likely due to fewer clusters. **Silhouette Score: 0.2536**, a modest improvement over 2021, reflecting slightly better-defined cluster boundaries. The model successfully captured the dominant structure in data while grouping similar categories together.

A graph with a blue line

Description automatically generatedYear 2023

For the 2023 dataset, the optimal number of clusters remained **3**, consistent with 2022, indicating stability in the dataset's structure.

**Cluster Purity**: **0.9613**, demonstrating strong alignment with the true labels and an improvement over 2022. **Silhouette Score**: **0.2157**, slightly lower than previous years, suggesting some increased complexity or overlap among data points.

Hierarchical Clustering

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Number of Cluster | **2** | **2** | **2** |
| Cluster Purity | **0.9755** | **0.9060** | **0.9838** |
| Sillhouette Score | **0.437** | **0.2419** | **0.3676** |

## A blue squares with white text Description automatically generatedA screenshot of a graph Description automatically generated**Year 2021**

The predicted clusters (left plot) align well with the true labels (right plot), showing good separability. The clusters are distinct with minimal overlap, suggesting effective grouping in the dataset.

**Cluster Purity**: High at 97.55%, indicating that most points in each cluster belong to the same class. **Silhouette Score**: 0.437, suggesting moderately well-defined clusters, with some room for improvement in terms of cohesion and separation.  
The model performed well for 2021 dataset, with clear clusters and high purity, though further refinement could help in enhancing cohesion.

## **Year 2022**

The predicted clusters still align with the true labels, but there is more overlap compared to 2021. The separability between clusters is less distinct, indicating a slight decline in clustering quality.

**Cluster Purity**: Still high at 90.60%, but noticeably lower than 2021, reflecting a slight degradation in clustering accuracy. **Silhouette Score**: 0.2419, indicating weaker cohesion and separation between clusters than in 2021.Clustering performance was less robust in 2022, possibly due to changes in data characteristics, such as increased complexity or noise.

## **Year 2023**

The predicted clusters align well with the true labels, with improved separability compared to 2022. There is minimal overlap between the clusters, showing clearer grouping in the dataset.

**Cluster Purity**: Very high at 98.38%, suggesting the majority of points within each cluster are correctly grouped with the same class. **Silhouette Score**: 0.3676, showing a slight improvement over 2022, indicating better cohesion and separation, but still with some room for improvement.The model's performance in 2023 showed significant improvement in clustering quality, with higher purity and better separation compared to 2022, although further optimization could help achieve even clearer clusters.

## **Hierarchical Clustering (Grid Search Clustering)**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Number of Cluster | **3** | **2** | **2** |
| Cluster Purity | **0.8818** | **0.9060** | **0.9838** |
| Sillhouette Score | **0.4582** | **0.2419** | **0.3676** |

## A blue and white box with numbers Description automatically generated**Year 2021**

A graph with colored lines

Description automatically generated

The predicted clusters align reasonably well with the true labels, but there are some discrepancies. The separability between clusters is moderate, and there is visible overlap in the predicted clusters. **Number of Clusters**: 3, which is higher than expected for the dataset (Rice and Cotton data). This indicates that the algorithm may have detected subcategories within the two mains. **Cluster Purity**: 88.18%, indicating a moderate level of accuracy in grouping the data points, but not perfect. This suggests that the three clusters captured a general structure, but some points were misclassified. **Silhouette Score**: 0.4582, reflecting moderate cohesion and separation between the clusters. It indicates that while the clusters are somewhat distinct, the separability could be improved for clearer clustering.

## A graph with numbers and lines Description automatically generatedA blue squares with white text Description automatically generatedA screenshot of a graph Description automatically generated**Year 2022**

Even we used Dendrogram to find optimal number of clusters, we got the same result as we got in simple Hierarchical Clustering for 2022.

A blue squares with black numbers

Description automatically generated

## A graph with lines and numbers Description automatically generated**Year 2023**

Even we used Dendrogram to find optimal number of clusters, we got the same result as we got in simple Hierarchical Clustering for 2023.

## **DB SCAN**

**Eps = 0.27, min\_samples= 102**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Number of Cluster | **2 + (Outliers)** | **1** | **3 + (Outliers)** |
| Cluster Purity | **0.9822** | **0.0** | **0.9932** |
| Sillhouette Score | **0.4101** | **0.09** | **0.2001** |

## A blue squares with white text Description automatically generatedA screenshot of a graph Description automatically generated**Year 2021**

The predicted clusters show good alignment with the true labels, with the model identifying two main clusters and a group of outliers. The separability between clusters is decent, although the presence of outliers may affect the overall cohesion.

**Number of Clusters**: 2 main clusters plus outliers. This indicates that DBSCAN detected the two primary groups (Rice and Cotton) but also identified some data points as noise or outliers. **Cluster Purity**: 98.22%, showing that the majority of the points in each cluster belong to the same class. The high purity suggests that DBSCAN did a good job of separating the data points into meaningful clusters, even if some points were marked as outliers. **Silhouette Score**: 0.4101, indicating moderate cohesion and separation. The presence of outliers likely reduces the overall cohesion of the clusters, leading to a somewhat lower silhouette score.

## A screenshot of a graph Description automatically generated**Year 2022**

The predicted clustering did not align well with the true labels in 2022. DBSCAN identified only one cluster, with all the data points grouped together, but the low purity and silhouette score indicate poor clustering. Their was no fault as we can see that data points are not separated and DBSCAN forms clusters on this basis.

## A blue and white graph Description automatically generated**Year 2023**

DBSCAN identified three clusters in 2023, including outliers, with a relatively good fit to the true labels. However, the silhouette score is still lower, indicating that the clusters could be better defined.**Number of Clusters**: 3 clusters plus outliers. DBSCAN identified three clusters, which could correspond to subgroups within the Rice and Cotton data, but also labeled some points as outliers.

**Cluster Purity**: 99.32%, which is very high, suggesting that the clusters identified by DBSCAN are mostly accurate and align well with the true labels. The presence of outliers did not significantly reduce the purity.

**Silhouette Score**: 0.2001, which is still relatively low. This indicates that while DBSCAN was able to correctly identify the clusters, the cohesion and separation between them are not very strong.DBSCAN showed improvement in 2023, identifying three clusters with high purity.

## **DB SCAN (Grid Search Clustering)**

All the Models chose eps= 0.27 and min\_samples = 60

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Number of Cluster | **6 + (Outliers)** | **2 + (Outliers)** | **3 + (Outliers)** |
| Cluster Purity | **0.9714** | **0.7878** | **0.9910** |
| Sillhouette Score | **0.4563** | **0.2315** | **0.2946** |

## A screenshot of a graph Description automatically generated**Year 2021**

DBSCAN identified six clusters and outliers, capturing subgroups in the data but maintaining high accuracy. **Number of Clusters**: 6 + outliers. **Cluster Purity**: 97.14%, indicating strong alignment with true labels despite over-segmentation. **Silhouette Score**: 0.4563, reflecting moderate cohesion and separation. DBSCAN effectively clustered the data, though the high number of clusters suggests over-segmentation.

## A screenshot of a graph Description automatically generated**Year 2022**

A blue squares with white text

Description automatically generated

DBSCAN identified two clusters and outliers, but performance degraded compared to 2021. **Number of Clusters**: 2 + outliers. **Cluster Purity**: 78.78%, indicating reduced clustering accuracy. **Silhouette Score**: 0.2315, showing weak cohesion and separation.  
The model struggled with lower purity and cohesion, possibly due to noise or data complexity.

A graph with blue squares and numbers

Description automatically generated

## **Year 2023**

DBSCAN identified three clusters with outliers, achieving high accuracy but moderate separation. **Number of Clusters**: 3 + outliers. **Cluster Purity**: 99.10%, demonstrating excellent alignment with true labels. **Silhouette Score**: 0.2946, indicating some cohesion and separation issues.  
**Conclusion**: The model performed well, balancing accuracy and identifying meaningful clusters.

## **Gaussian Mixture Models**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Number of Cluster | **2** | **2** | **2** |
| Cluster Purity | **0.9865** | **0.7668** | **0.9857** |
| Sillhouette Score | **0.4080** | **0.1678** | **0.3668** |

## A blue squares with white text Description automatically generated**Year 2021**

GMM identified two clusters, showing strong alignment with true labels and moderate cohesion. **Number of Clusters**: 2. **Cluster Purity**: 98.65%, indicating excellent clustering accuracy. **Silhouette Score**: 0.4080, reflecting moderate cohesion and separation. GMM performed effectively, capturing the underlying structure with high purity.

## **Year 2022**

A blue squares with black numbers

Description automatically generated

GMM identified two clusters but struggled with reduced accuracy and weaker separation. **Number of Clusters**: 2. **Cluster Purity**: 76.68%, indicating moderate clustering accuracy, with some misclassifications. **Silhouette Score**: 0.1678, showing poor cohesion and separation. Clustering performance degraded, likely due to noise or data complexity.

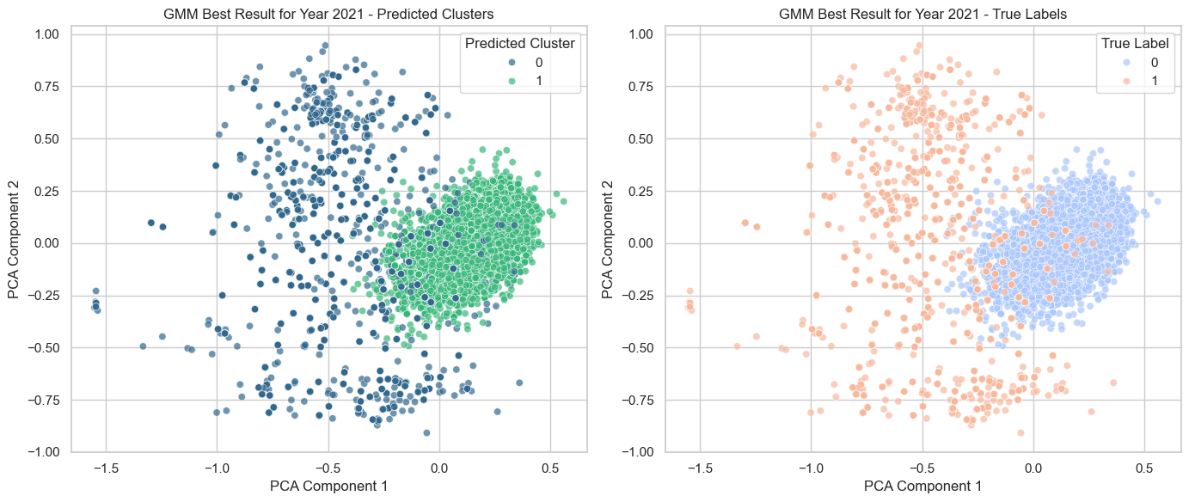
## A screenshot of a graph Description automatically generatedA blue squares with black numbers Description automatically generated**Year 2023**

GMM identified two clusters with high accuracy and moderate separation. **Number of Clusters**: 2. **Cluster Purity**: 98.57%, demonstrating strong clustering accuracy. **Silhouette Score**: 0.3668, indicating improved cohesion and separation compared to 2022. GMM performed well in 2023, achieving high purity with moderate cluster definition.

## **Gaussian Mixture Models (Grid Search Clustering)**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Parameters | **covariance\_type:full, init\_params:random, n\_components: 2** | **covariance\_type: diag, init\_params: kmeans, n\_components: 5** | **covariance\_type: full, init\_params: kmeans, n\_components: 5** |
| Number of Cluster | **2** | **4** | **4** |
| Cluster Purity | **0.9875** | **0.9730** | **0.9901** |
| Sillhouette Score | **0.4056** | **0.1480** | **0.07** |

## **Year 2021**



GMM identified two clusters with parameters optimized for full covariance and random initialization, showing excellent clustering accuracy.

**Parameters:** covariance\_type: full, init\_params: random, n\_components: 2. **Number of Clusters:** 2. **Cluster Purity: 98.75%,** indicating excellent clustering alignment. **Silhouette Score: 0.4056**, reflecting moderate cohesion and separation.  
Conclusion: GMM performed effectively with high accuracy and reasonable cluster definition.

## **Year 2021**

GMM identified four clusters, with optimized diagonal covariance and k-means initialization, achieving high purity but poor separation. **Parameters**: covariance\_type: diag, init\_params: kmeans, n\_components: 5. **Number of Clusters**: 4. **Cluster Purity**: 97.30%, indicating strong clustering accuracy despite additional clusters. **Silhouette Score**: 0.1480, showing poor cohesion and separation. GMM achieved good accuracy but struggled with cluster definition due to multiple groups and weak separability.

## **Year 2023**

A screenshot of a graph

Description automatically generated

GMM identified four clusters with full covariance and k-means initialization, achieving the highest purity but poor cohesion.

**Parameters**: covariance\_type: full, init\_params: kmeans, n\_components: 5. **Number of Clusters**: 4. **Cluster Purity**: 99.01%, reflecting excellent clustering accuracy. **Silhouette Score**: 0.07, indicating weak cohesion and separation among clusters. GMM achieved high accuracy but poor cluster definition, likely due to over-segmentation.

Analysis of Unsupervised Learning Approach without PCA

In the unsupervised learning approach without PCA, K-means, hierarchical clustering, DBSCAN, and Gaussian Mixture Models (GMM) each displayed distinct strengths and weaknesses. K-means performed exceptionally well in 2021, achieving high purity and clear separability between clusters. However, its performance declined in 2022 due to increased data complexity, although it showed some improvement in 2023. Hierarchical clustering remained consistent, with strong performance in 2021 and 2023, achieving high purity and minimal overlap in clusters. However, it struggled in 2022, with weaker cohesion and a slight decline in accuracy. DBSCAN performed well in 2021, detecting distinct clusters and outliers, achieving high purity and moderate cohesion. However, it faced challenges in 2022, where it failed to capture the true structure of the data, resulting in poor clustering. GMM displayed strong performance in 2021 with high accuracy, but its performance significantly degraded in 2022, likely due to noise and overlapping clusters. By 2023, GMM performed better, maintaining high purity but with moderate cohesion. The overall trend indicates that while K-means and hierarchical clustering showed strong and stable performance, DBSCAN and GMM faced more significant challenges, particularly in the more complex data from 2022.

**K-Means:** Best Silhouette (2021: 0.4478), High Purity (2023: 0.9668).

**Hierarchical:** Best Purity (2023: 0.9838), Consistent Silhouette (2021: 0.4377).

**DBSCAN:** Highest Purity (2023: 0.9932), Weak Silhouette (2022: 0.0907).

**Gaussian Mixture:** Strong Purity (2021: 0.9865), Low Silhouette (2022: 0.1678).

**Hierarchical** is the best due to its balance between clustering quality (silhouette) and accuracy (purity). DBSCAN is preferable only when the highest purity is essential, despite occasional structural weaknesses.

Unsupervised Learning Approach (Using PCA)

K-means Clustering

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Number of Cluster | **2** | **2** | **2** |
| Cluster Purity | **0.9324** | **0.7665** | **0.9710** |
| Sillhouette Score | **0.4350** | **0.2190** | **0.4482** |

## A blue squares with white text Description automatically generated**Year 2021**

K-means identified two clusters, showing good accuracy and moderate separation after dimensionality reduction with PCA. **Number of Clusters**: 2. **Cluster Purity**: 93.24%, indicating strong alignment with true labels. **Silhouette Score**: 0.4350, reflecting moderate cohesion and separation. The model effectively captured the structure of the data with good clustering performance.

## A blue squares with white text Description automatically generated**Year 2022**

K-means identified two clusters but struggled with reduced accuracy and weaker separation, likely due to increased complexity in the data. **Number of Clusters**: 2. **Cluster Purity**: 76.65%, indicating moderate clustering accuracy. **Silhouette Score**: 0.2190, showing poor cohesion and separation. The model's performance degraded in 2022, reflecting challenges in clustering the data effectively.

## A blue squares with numbers Description automatically generated**Year 2023**

K-means identified two clusters with improved accuracy and the highest silhouette score among the three years, indicating well-defined clusters. **Number of Clusters**: 2. **Cluster Purity**: 97.10%, demonstrating excellent clustering accuracy. **Silhouette Score**: 0.4482, reflecting good cohesion and separation. The model performed best in 2023, achieving high accuracy and clear cluster definition.

## **Hierarchical Clustering**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Number of Cluster | **2** | **2** | **2** |
| Cluster Purity | **0.9769** | **0.7665** | **0.9808** |
| Sillhouette Score | **0.4172** | **0.2421** | **0.4415** |

## A blue squares with black numbers Description automatically generated**Year 2021**

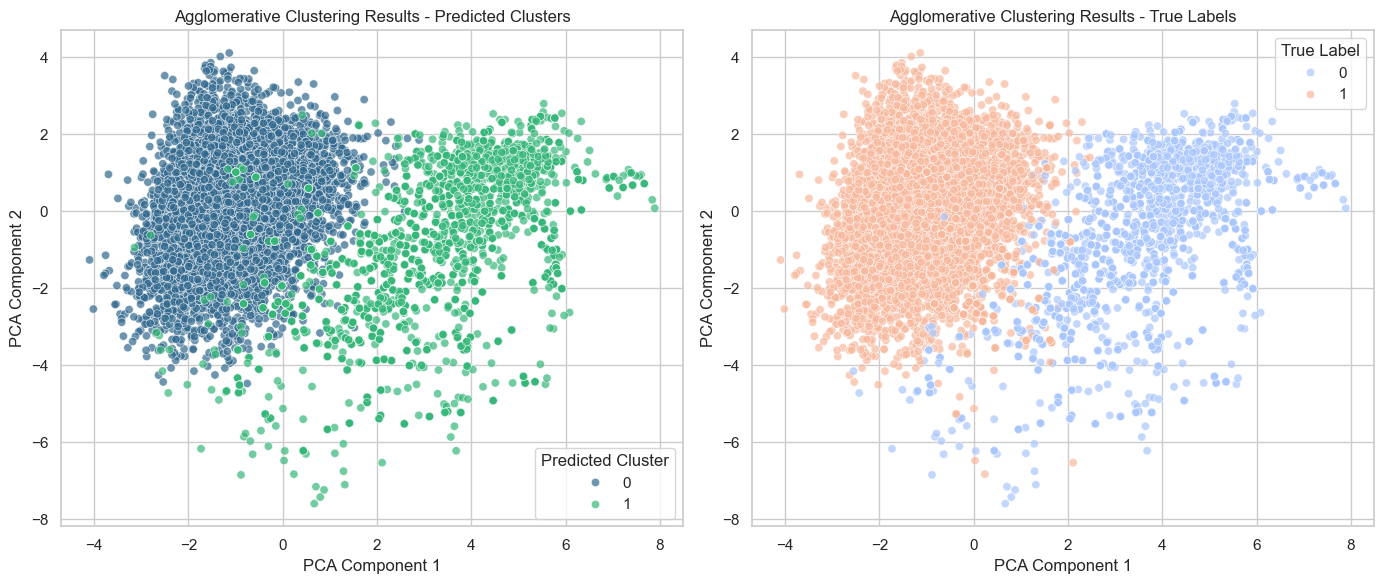
Hierarchical clustering identified two clusters with high accuracy and moderate separation. **Number of Clusters**: 2. **Cluster Purity**: 97.69%, indicating excellent clustering alignment. **Silhouette Score**: 0.4172, reflecting moderate cohesion and separation.The model effectively captured the data structure with high accuracy and reasonable cluster definition.

## A blue squares with white text Description automatically generatedA screenshot of a graph Description automatically generated**Year 2022**

Hierarchical clustering identified two clusters but showed reduced accuracy and weaker separation compared to 2021. **Number of Clusters**: 2. **Cluster Purity**: 76.65%, indicating moderate clustering accuracy. **Silhouette Score**: 0.2421, showing weak cohesion and separation. The model struggled with lower accuracy and weaker cluster definition, likely due to increased data complexity or noise.

## **Year 2023**

A blue squares with white text

Description automatically generated

Hierarchical clustering identified two clusters with strong accuracy and improved separation compared to 2022. **Number of Clusters**: 2. **Cluster Purity**: 98.08%, demonstrating excellent clustering accuracy. **Silhouette Score**: 0.4415, reflecting good cohesion and separation. The model performed well, achieving high accuracy and clear cluster definition.

## **DBSCAN**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Number of Cluster | **8 + (Outliers)** | **2 + (Outliers)** | **2 + (Outliers)** |
| Cluster Purity | **0.9698** | **0.8058** | **0.7958** |
| Sillhouette Score | **0.3015** | **0.2888** | **0.2425** |

## **Year 2021**

DBSCAN identified eight clusters with outliers, achieving high purity but moderate cohesion. **Number of Clusters**: 8 + outliers. **Cluster Purity**: 96.98%, indicating strong alignment with true labels. **Silhouette Score**: 0.3015, reflecting moderate cohesion and separation.  
**Conclusion**: The model effectively captured distinct subgroups but exhibited over-segmentation.

## **Year 2022**

DBSCAN identified two clusters with outliers, achieving moderate accuracy and slightly improved cohesion. **Number of Clusters**: 2 + outliers. **Cluster Purity**: 80.58%, indicating reasonable alignment with true labels. **Silhouette Score**: 0.2888, showing weak cohesion and separation. The model performed moderately well but struggled with weaker cluster definition compared to 2021.

## A blue squares with white text Description automatically generated**Year 2023**

DBSCAN identified two clusters with outliers, showing reduced accuracy and the weakest cohesion among the three years.

**Number of Clusters**: 2 + outliers. **Cluster Purity**: 79.58%, reflecting moderate clustering accuracy. **Silhouette Score**: 0.2425, indicating poor cohesion and separation. The model's performance degraded in 2023, with lower accuracy and weak cluster separability.

## **Gaussian Mixture Model**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | 2021 | 2022 | 2023 |
| Number of Cluster | **2** | **2** | **2** |
| Cluster Purity | **0.9832** | **0.8679** | **0.9830** |
| Sillhouette Score | **0.3974** | **0.2147** | **0.4380** |

## A blue squares with white text Description automatically generated**Year 2021**

GMM identified two clusters, showing excellent clustering accuracy and moderate cohesion. **Number of Clusters**: 2. **Cluster Purity**: 98.32%, indicating strong alignment with true labels. **Silhouette Score**: 0.3974, reflecting moderate cohesion and separation. The model effectively captured the structure of the data with high accuracy.

## A blue squares with white text Description automatically generatedA screenshot of a graph Description automatically generated**Year 2022**

GMM identified two clusters , achieving moderate accuracy but weaker separation compared to 2021. **Number of Clusters**: 2 . **Cluster Purity**: 86.79%, reflecting good but reduced clustering accuracy. **Silhouette Score**: 0.2147, indicating weak cohesion and separation. The model's performance degraded, likely due to noise or increased data complexity.

## **Year 2023**

A blue squares with white text

Description automatically generatedGMM identified two clusters, achieving high accuracy and improved separation compared to 2022. **Number of Clusters**: 2. **Cluster Purity**: 98.30%, demonstrating excellent clustering accuracy. **Silhouette Score**: 0.4380, reflecting good cohesion and separation. The model performed well, achieving high purity and well-defined clusters.

Analysis of Unsupervised Learning Approach with PCA

With PCA applied, dimensionality reduction helped improve the separation and clarity of clusters for most models in 2023, especially K-means and hierarchical clustering. K-means, initially facing some difficulty in 2022, saw improved cluster purity and cohesion in 2023. Similarly, hierarchical clustering benefited from PCA, enhancing cluster separability and accuracy. DBSCAN, however, struggled even with PCA, showing weaker performance in 2022 and 2023 compared to 2021, particularly with lower purity and cohesion. GMM also showed notable improvements in 2023, with higher cluster purity and better-defined boundaries, although it still faced challenges with weak separation in some cases. While PCA helped K-means and hierarchical clustering enhance their clustering quality, DBSCAN and GMM did not fully benefit from the dimensionality reduction, maintaining issues with cohesion and separation. This suggests that PCA was more beneficial for models like K-means and hierarchical clustering, which rely more on clear separability, while DBSCAN and GMM struggled with issues related to data density and overlap even after dimensionality reduction.

**K-Means:** High purity (2023: 0.9710), decent Silhouette (2023: 0.4482).

**Grid Search:** Good purity (2023: 0.9637), lower Silhouette (2022: 0.2774).

**Hierarchical:** Balanced, high purity (2023: 0.9808) and solid Silhouette (2023: 0.4415).

**DBSCAN:** Highest purity (2023: 0.9896), weak Silhouette (2023: 0.2425).

**Best Model:** Hierarchical Clustering stands out as the best, offering a strong balance between Silhouette score (2023: 0.4415) and purity (2023: 0.9808), ensuring both good cluster quality and accuracy.

Overall Comaprison

Overall, comparing the top-performing models, K-means and hierarchical clustering emerged as the most consistent performers across both with and without PCA. These models achieved high cluster purity and clear separability, particularly in 2021 and 2023, though K-means faced some challenges in 2022. Both models benefitted from PCA, with clear improvements in separability and cohesion. In contrast, DBSCAN and GMM showed more fluctuating results. DBSCAN performed well in 2021 but suffered in 2022 and 2023, particularly with poor cohesion and high sensitivity to noise. GMM achieved good results in 2021 but faced significant degradation in 2022, with only a slight recovery in 2023. While PCA improved K-means and hierarchical clustering, it had less of an impact on DBSCAN and GMM, highlighting that models with clearer cluster boundaries, like K-means and hierarchical clustering, are more robust to dimensionality reduction.

**With PCA** performs better than **without PCA**.

**Improved Purity**: K-Means and DBSCAN show higher purity scores with PCA (K-Means: 0.9710, DBSCAN: 0.9896) compared to without PCA.

**Stable Clustering**: Hierarchical maintains balanced performance with both high purity and solid Silhouette scores (with PCA: 0.4415, without PCA: 0.4377).

**Better Overall Performance**: With PCA, models like K-Means and DBSCAN perform better in terms of purity, while Hierarchical remains the most reliable model overall.