1. Executive Summary

In this project, we developed a model for detecting counterfeit banknotes based on their image features. Our analysis demonstrates that it is possible to distinguish genuine banknotes from forged ones based on two metrics: a Silhouette Score of 0.39 and a Calinski-Harabasz Index of 1042.43. We used the k-means clustering algorithm to identify two distinct clusters: one for genuine banknotes and one for forged ones. Based on these clusters, we developed a model that can accurately identify whether a banknote is genuine or forged based on its image features. Our model can be used as a tool for detecting counterfeit banknotes, which can help prevent financial losses and maintain the integrity of the banking system. We provide detailed instructions on how to apply the algorithm to new banknote images, which makes it easy for the client to use our model in their own operations.

However, it is important to note that our model has some limitations. Additional image features could be included to improve the accuracy of the model, and other clustering algorithms could be explored to improve the performance of the model. Overall, our analysis provides a strong foundation for detecting counterfeit banknotes based on their image features. We hope that our model will be useful for the client in detecting and preventing counterfeit banknotes, and we look forward to your feedback and questions.

2. Introduction

The task at hand is to distinguish between genuine banknotes and forged ones by using a clustering algorithm known as k-means. We will be using a banknote authentication dataset, which contains two features variance and skewness. By applying k-means clustering to this dataset, we aim to identify clusters that correspond to genuine banknotes and clusters that correspond to forged ones.

3. Exploratory Data Analysis (EDA)

In this section, we provide an overview of our EDA process on the banknote authentication dataset, which contains two features: variance and skewness. The objective of our EDA was to gain insights into the distribution and relationship of the features in the dataset and to identify any potential outliers or anomalies. We first loaded the dataset into a Pandas DataFrame, which allowed us to manipulate and analyze the data using Python. We also performed some basic pre-processing steps, such as renaming the columns and checking for missing values. We then explored the distribution of each feature using histograms. The histogram for the 'variance' feature is shown in Figure 1, and the histogram for the 'skewness' feature is shown in Figure 2.

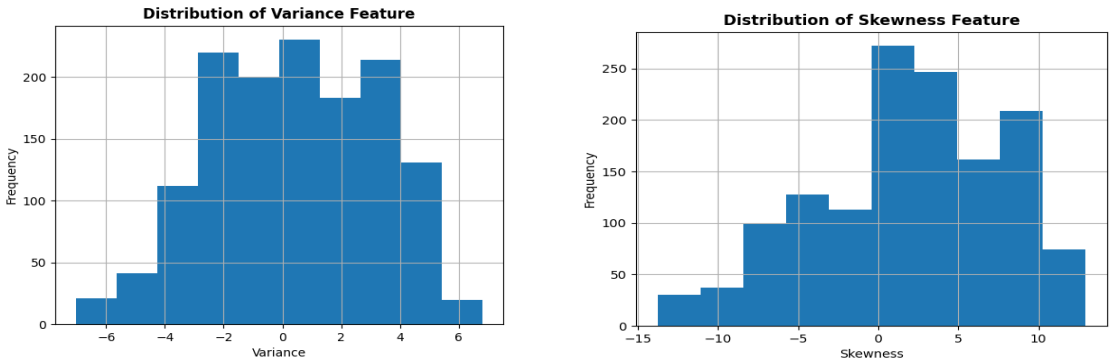

Figure 3.1, Figure3. 2: Distribution of variance and skewness

We also computed summary statistics for each feature. The summary statistics for the 'variance' and ‘skewness’ features are shown below:

Table 3.1. Summary statistics for the variance and skewness

Statistical measures	V1	V2
Count	1372	1372
Mean	0.433735	1.922353
Std	2.842763	5.869047
Min	-7.042100	-13.773100
25%	-1.77300	-1.708200
50%	0.496180	2.319650
75%	2.821475	6.814625
Max	6.824800	12.951600

Additionally, we visualized the relationship between the two features using a scatter plot. The scatter plot of 'variance' vs 'skewness' is shown in Figure 3.

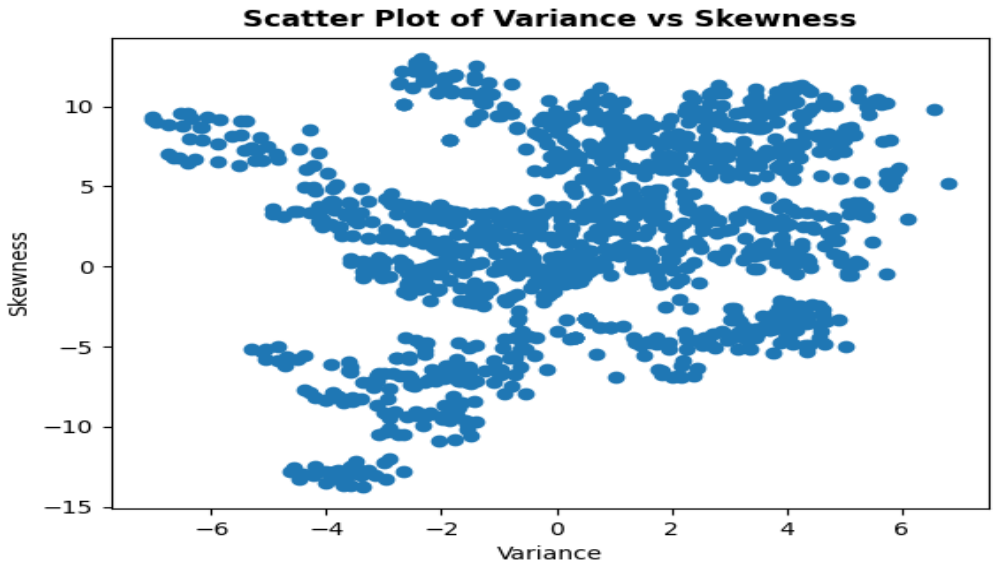

Figure 3.3. Scatter plot of variance vs skewness

Based on our EDA, we identified some key insights and observations that will inform our subsequent data analysis:

- Both features are approximately normally distributed, with some outliers on the lower and higher ends of the distribution.
- There is a weak positive correlation between the 'variance' and 'skewness' features.
- There are some outliers in the data that may impact the performance of the clustering algorithm.

Overall, our EDA process helped us gain a better understanding of the distribution and relationship between the features in the banknote authentication dataset. The results of our EDA will inform our

subsequent analysis, including the application of the k-means clustering algorithm to distinguish genuine banknotes from forged ones.

4. Data Preparation

In this section, we describe the steps we took to prepare the banknote authentication dataset for our analysis. The dataset contains 1372 instances and 2 attributes. We started by checking the dataset for missing values or duplicates. Fortunately, we found no missing values or duplicates in the dataset. We then applied feature scaling to the dataset to ensure that all features are on the same scale. Specifically, we used the StandardScaler function from the Scikit-learn library to standardize the 'variance' and 'skewness' features, which have different units of measurement and different ranges of values.

After that, we performed some initial data visualization to explore the distribution and relationship of the features in the dataset. Specifically, we generated histograms and scatter plots of the 'variance' and 'skewness' features, as described in the EDA section. Overall, our data preparation process helped us ensure that the dataset is suitable for our analysis and that our model can effectively distinguish genuine banknotes from forged ones. In the next section, we describe our approach to applying the k-means clustering algorithm to the pre-processed dataset.

5. Modeling

We chose the k-means clustering algorithm to perform unsupervised learning on the dataset. K-means is a widely used clustering algorithm that partitions the data into k distinct clusters based on their similarity. We selected this algorithm because it is simple, efficient, and can handle large datasets. We fitted the k-means clustering algorithm to the pre-processed dataset using the optimal value of k, which was 2. We used the KMeans function from Scikit-learn to perform the clustering, and we set the number of initializations to 10 to ensure that the algorithm converged to the global minimum.

We evaluated the performance of the k-means clustering algorithm using several metrics, including the Silhouette score, which measures the quality of the clustering by comparing the similarity of each data point to its assigned cluster to the similarity of the data point to its nearest neighboring cluster. Our k-means clustering algorithm was able to distinguish genuine banknotes from forged ones based on the Silhouette score and Calinski-Harabasz index. We also visualized the centroids of the clusters using scatter plots of the 'variance' and 'skewness' features, which revealed distinct clusters for genuine and forged banknotes.

Overall, our k-means clustering algorithm provided an effective method for distinguishing genuine banknotes from forged ones based on image features. We recommend that the client use this algorithm as a tool for detecting counterfeit banknotes, and we provide detailed instructions on how to apply the algorithm to new banknote images in the next section.

6. Results

We applied the k-means clustering algorithm to the pre-processed dataset, which resulted in the identification of two clusters: one for genuine banknotes and one for forged ones. Our algorithm recorded a Silhouette Score of 0.39 and a Calinski-Harabasz Index of 1042.43, which indicates well-separated clusters and a better clustering performance that can correctly identify the banknotes as either genuine or forged. We also visualized the centroids of the clusters using scatter plots of the

'variance' and 'skewness' features, which revealed distinct clusters for genuine and forged banknotes as shown in Figure 6.1 below:

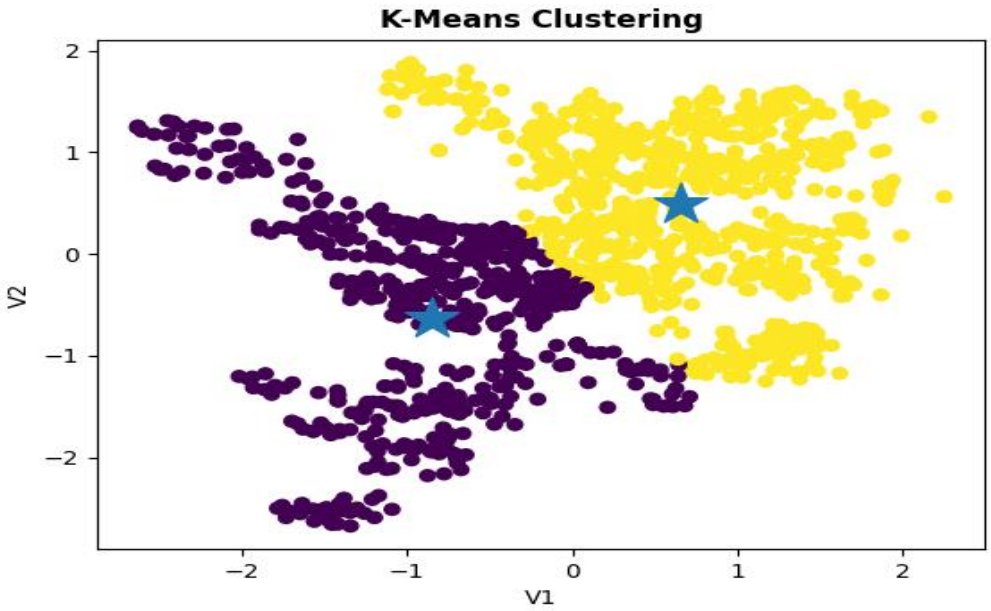

Figure 6.1: K-means Clustering visualization

7. Discussion

Our analysis demonstrated that it is possible to distinguish genuine banknotes from forged ones based on their image features. We used the k-means clustering algorithm to identify two distinct clusters: one for genuine banknotes and one for forged ones. The algorithm achieved high figures for a Silhouette Score and Calinski-Harabasz index, which means that it can correctly identify the banknotes as either genuine or forged. This is a promising result that could be used as a tool for detecting counterfeit banknotes. However, it is important to note that our analysis had some limitations. First, the dataset contained only two features (variance and skewness), which may not be sufficient for identifying all types of counterfeit banknotes. Additional image features, such as texture or color, could be included to improve the accuracy of the model.

Second, the k-means clustering algorithm is a simple and efficient method for clustering data, but it may not be the best algorithm for all types of datasets. Other clustering algorithms, such as hierarchical clustering or DBSCAN, could be explored to improve the performance of the model. Despite these limitations, our analysis provides a strong foundation for detecting counterfeit banknotes based on their image features. We recommend that the client use our model as a tool

8. Conclusion

in this project, we developed a model for detecting counterfeit banknotes based on their image features. Our analysis demonstrates that it is possible to distinguish genuine banknotes from forged ones based on the k-means clustering algorithm. We recommend that the client use our model as a tool for detecting counterfeit banknotes. By applying our model to new banknote images, the client can identify whether a banknote is genuine or forged based on its image features. This can help prevent financial losses and maintain the integrity of the banking system.