Bank Note Authentication using KNN and Random Forest

Shreya Chetan Pawaskar (Group No:25)

Abstract:

In the era of widespread internet accessibility and advancements in printing technology, the traditional use of physical currency in the form of bank notes has led to a significant challenge: the ease of producing counterfeit banknotes that are almost indistinguishable from legitimate ones. To address this concern, our project aims to leverage the power of machine learning algorithms, specifically k-nearest neighbors (KNN) and Random Forest, to classify bank notes as either legitimate or counterfeit. By extracting data from images of bank notes and analyzing different features, our project aims to develop a reliable and accurate classification system. We rigorously evaluate the performance using various metrics. This project explores the application of machine learning techniques, to address the critical task of banknote authentication for financial institutions and businesses.

1. Introduction:

The circulation of counterfeit banknotes poses a severe threat to the stability and integrity of the financial system. With the advancement of printing technology, it has become increasingly difficult to distinguish counterfeit banknotes from genuine ones through manual inspection alone. Counterfeit currency can lead to significant economic losses for businesses, financial institutions, and individuals, undermining trust in the monetary system. In this project, we aim to address this critical issue by using machine learning to authenticate banknotes based on their physical attributes. Through rigorous evaluation and comparison, we aim to identify the most effective approach for banknote authentication.

2. Methodology:

2.1 Dataset and External Libraries:

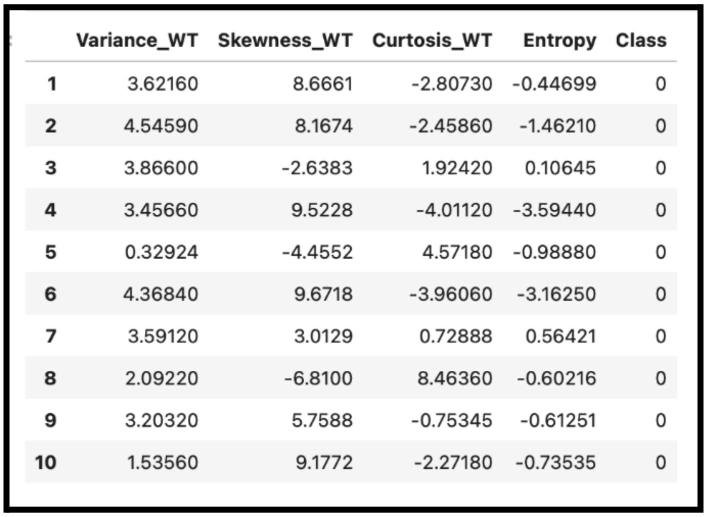
The project makes use of the banknote dataset sourced from the UCI Machine Learning Repository. To streamline data analysis and modeling processes, we harness several external libraries. These include pandas for data manipulation, numpy for performing numerical operations, matplotlib and seaborn for creating data visualizations, and scikit-learn for implementing algorithms and conducting thorough model evaluation.

2.1 Libraries Used

2.2 Importing the dataset and understanding it:

We'll begin by importing the dataset and examining its structure. The dataset contains four independent variables: variance, skewness, kurtosis, and entropy, along with a target variable indicating whether a banknote is authentic or counterfeit. It consists of 5 rows and 1372 columns. Among these columns, there are four continuous data features and one class variable. Here's a brief overview of the variables:

1. Variance: Reflects the amount of difference or variability in the data.
2. Skewness: Indicates asymmetry in the distribution of data.
3. Kurtosis: Describes the shape of the distribution's peak.
4. Entropy: Entropy is the measure of disorder or uncertainty.

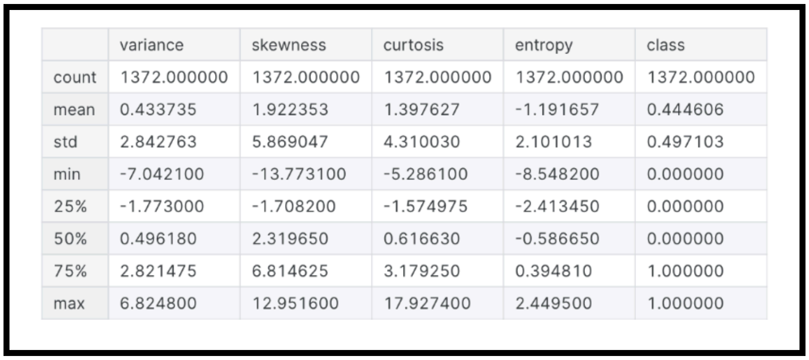


	Variance_WT	Skewness_WT	Curtosis_WT	Entropy	Class
1	3.62160	8.6661	-2.80730	-0.44699	0
2	4.54590	8.1674	-2.45860	-1.46210	0
3	3.86600	-2.6383	1.92420	0.10645	0
4	3.45660	9.5228	-4.01120	-3.59440	0
5	0.32924	-4.4552	4.57180	-0.98880	0
6	4.36840	9.6718	-3.96060	-3.16250	0
7	3.59120	3.0129	0.72888	0.56421	0
8	2.09220	-6.8100	8.46360	-0.60216	0
9	3.20320	5.7588	-0.75345	-0.61251	0
10	1.53560	9.1772	-2.27180	-0.73535	0

2.2 Dataset

2.3 Data Preprocessing:

Before proceeding with modeling, we conduct crucial data preprocessing tasks. This includes checking for missing values & identifying potential outliers within the dataset. Any missing data is addressed, and the Interquartile Range (IQR) method is utilized to detect and eliminate outliers. The provided table outlines the description of each variable, allowing us to analyze the mean and distribution of the data. The mean of the class variable is close to 0.5, indicating a balanced dataset.

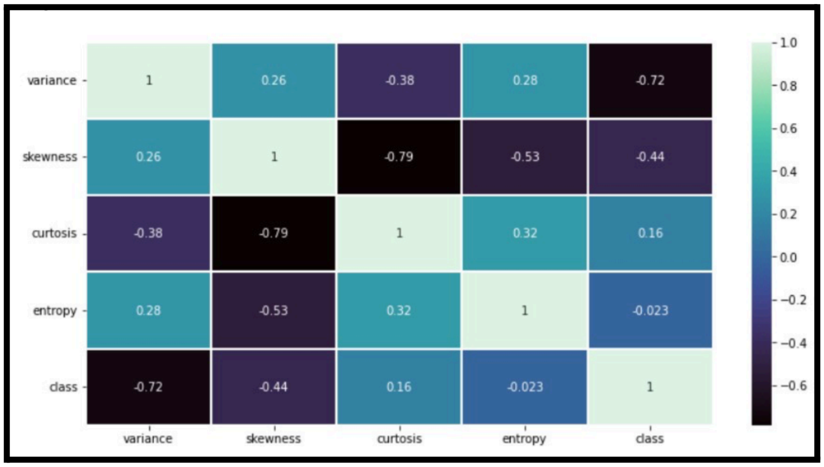


	variance	skewness	curtosis	entropy	class
count	1372.000000	1372.000000	1372.000000	1372.000000	1372.000000
mean	0.433735	1.922353	1.397627	-1.191657	0.444606
std	2.842763	5.869047	4.310030	2.101013	0.497103
min	-7.042100	-13.773100	-5.286100	-8.548200	0.000000
25%	-1.773000	-1.708200	-1.574975	-2.413450	0.000000
50%	0.496180	2.319650	0.616630	-0.586650	0.000000
75%	2.821475	6.814625	3.179250	0.394810	1.000000
max	6.824800	12.951600	17.927400	2.449500	1.000000

2.3 Dataset Description

2.4 Correlation Analysis:

To understand the connections between the variables, we create a correlation matrix. This matrix is then presented visually through a heatmap, allowing us to pinpoint any significant correlations among the variables that could impact the modeling phase. Notably, we observe a strong negative correlation between kurtosis and skewness, while a moderately high correlation exists between kurtosis and entropy.

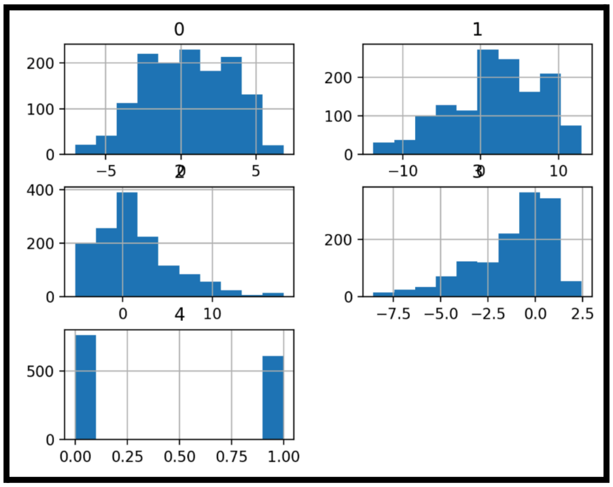


2.4 Correlation Matrix

3. Results and Analysis:

3.1 Understanding the data:

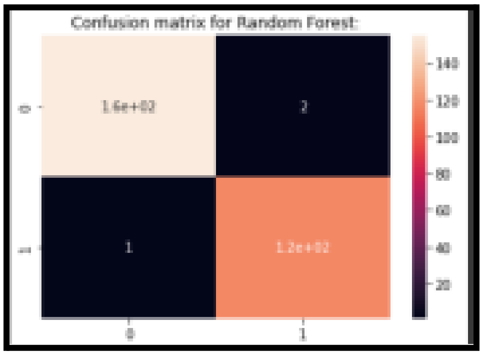
We visualize the distribution of the variables in the dataset using histograms and boxplots. We observe that the variables are normally distributed and do not have significant outliers.



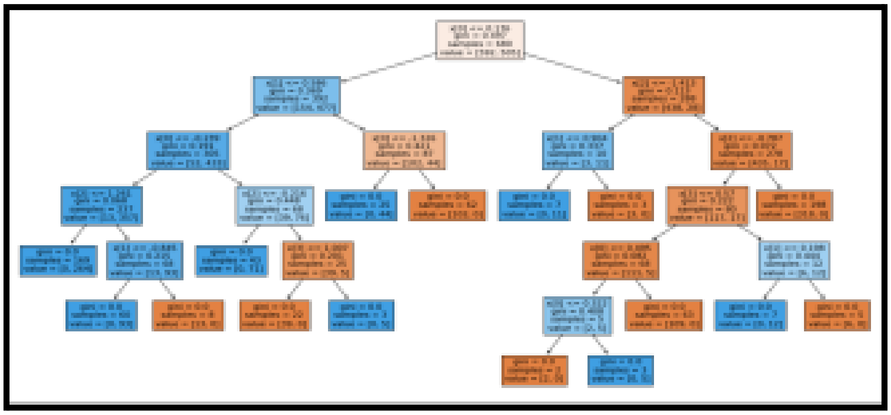
3.2 Plots & Conclusions from plot

3.2.1 Random Forest:

We apply the Random Forest algorithm to classify the banknotes. The model's performance is assessed using various metrics, including the confusion matrix, accuracy, precision, recall, and F1-score. The obtained accuracy of 98.91% signifies the model's ability to effectively classify banknotes as either authentic or counterfeit.

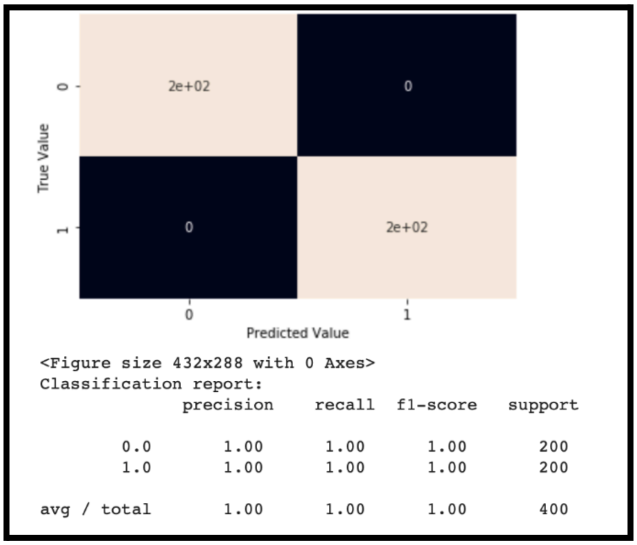


Decision Tree:



3.2.2 KNN:

We also apply the KNN algorithm to classify the banknotes. We evaluate the performance of the model using the same metrics as the Random Forest algorithm. The model achieves an accuracy of 100%, which is higher than the Random Forest algorithm.



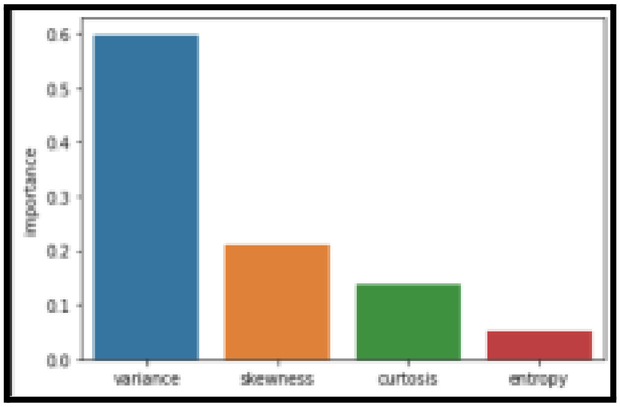
```
<Figure size 432x288 with 0 Axes>
Classification report:
              precision    recall  f1-score   support

        0.0       1.00      1.00      1.00       200
        1.0       1.00      1.00      1.00       200

avg / total       1.00      1.00      1.00       400
```

3.2.3 Final Observation Table:

We compare the performance of the Random Forest and KNN algorithms using the observation table. The KNN algorithm outperforms the Random Forest algorithm in terms of accuracy, precision, recall, and F1-score metrics. The following table shows all the features and their importance in terms of information gain



Observation Table

Method	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Random Forest	0.99	0.98	0.99	0.99
KNN	1.0	1.0	1.0	1.0

4. Conclusion:

In this study, we investigated the use of ML algorithms to authenticate banknotes based on their physical characteristics. Specifically, we applied the Random Forest and KNN for classification purposes. Upon thorough evaluation, we observed that the KNN algorithm outperformed Random Forest across various performance metrics such as accuracy, precision, recall, and F1-score. This superiority of KNN suggests its effectiveness in capturing the underlying patterns and relationships within the dataset, thereby accurately distinguishing between authentic and counterfeit banknotes.

The findings strongly suggest the efficacy of machine learning in banknote authentication. By analyzing features, these algorithms demonstrate a high accuracy in detecting counterfeit banknotes, a critical aspect for upholding the integrity of financial systems. The implications of this study are significant for financial institutions and businesses. Implementing ML-based counterfeit detection systems can improve security measures. Accurate banknote authentication can prevent financial losses, uphold consumer trust, and contribute to the stability of the monetary system. In essence, this research underscores the potential of ML techniques in addressing real-world challenges.

5. References:

1. C. Kumar and A . K. Dudyala, "Banknote authentication using deci-sion tree rules and machine learning techniques," 2015 International Conference on Advances in Computer Engineering and Applications, 2015, pp. 310-314, doi: 10.1109/ICACEA.2015.7164721.
2. https://archive.ics.uci.edu/ml/datasets/banknote+authentication
3. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d1
4. https://machinelearningprojects.net/bank-note-authentication

5. https://github.com/sbsreedh/Banknote-Authentication-Dataset-KNN-Analysis/blob/master/Banknote_authentication_Dataset.ipynb

Contributions:

I have done this project solo.