Fruits Freshness Classification

based on comprehensive image features
using Random Forest and
XG Boost Classifier Model

Presented by: Kah Hou

Applied Data Science (Cohort Jan 2024)

Problem Statements

- Difficult to identify freshness of fruits in large volume
- Food poisoning and allergic reaction
- Bad customer satisfaction leads to bad reputation

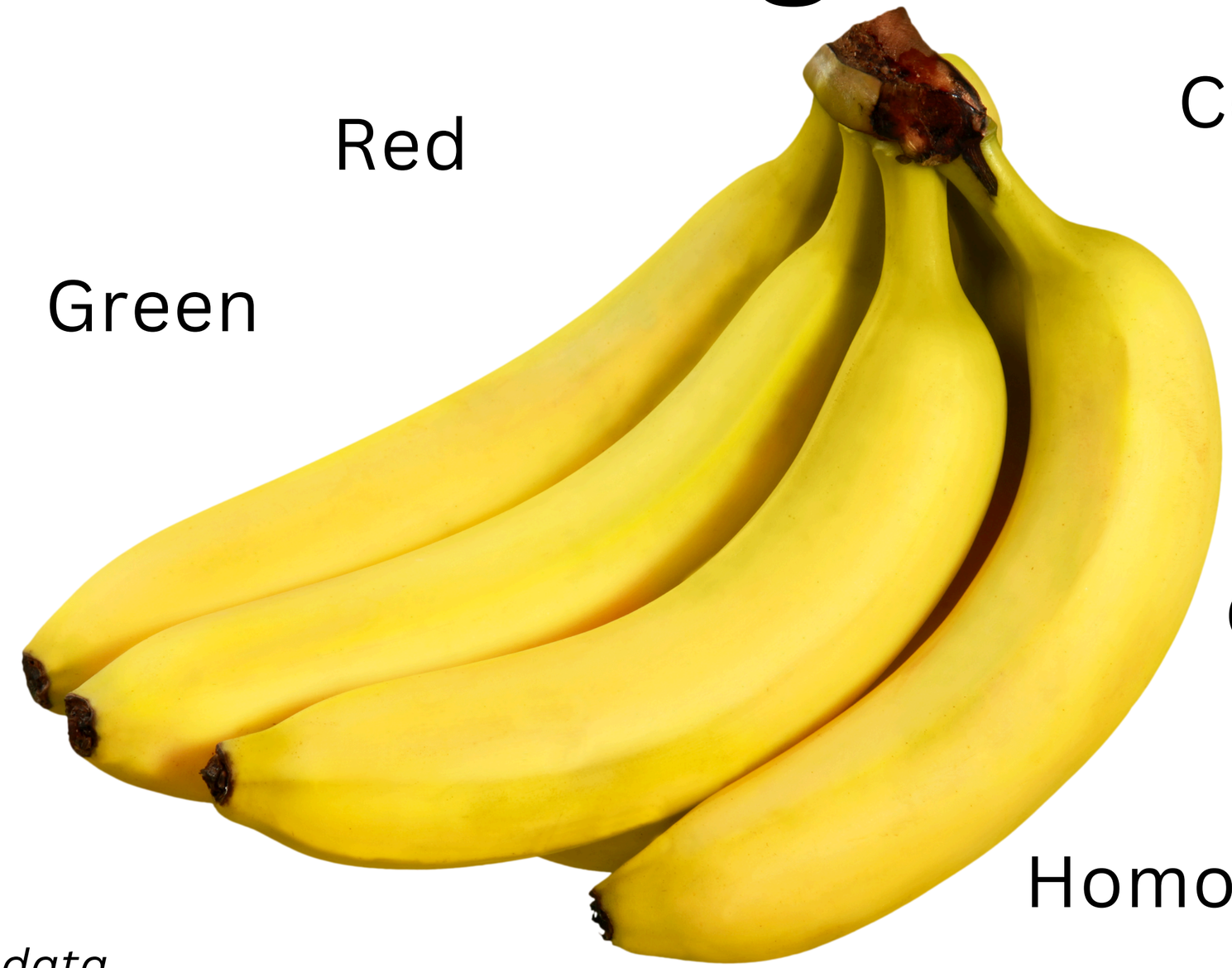
References:
[1] https://www.mdpi.com/1424-8220/22/21/8192
[2] https://www.coldhubs.com/coldhubnews/2023/5/11/risk-in-consuming-rotten-fruits-and-vegetables#:~:text=Consuming%20rotten%20fruits%20and%20vegetables%20can%20lead%20to%20food%20poisoning,that%20can%20make%20one%20sick.
[3] https://www.the-sun.com/money/9652675/aldi-customer-frustrated-refund/

Understanding Dataset

Red

Green

Blue



Contrast

Energy

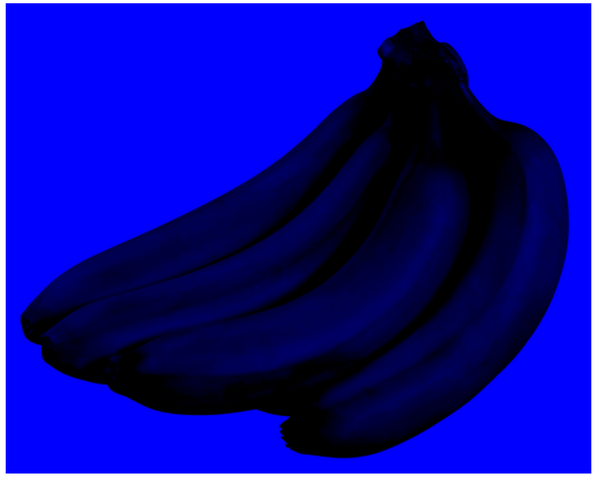
Correlation

Homogeneity

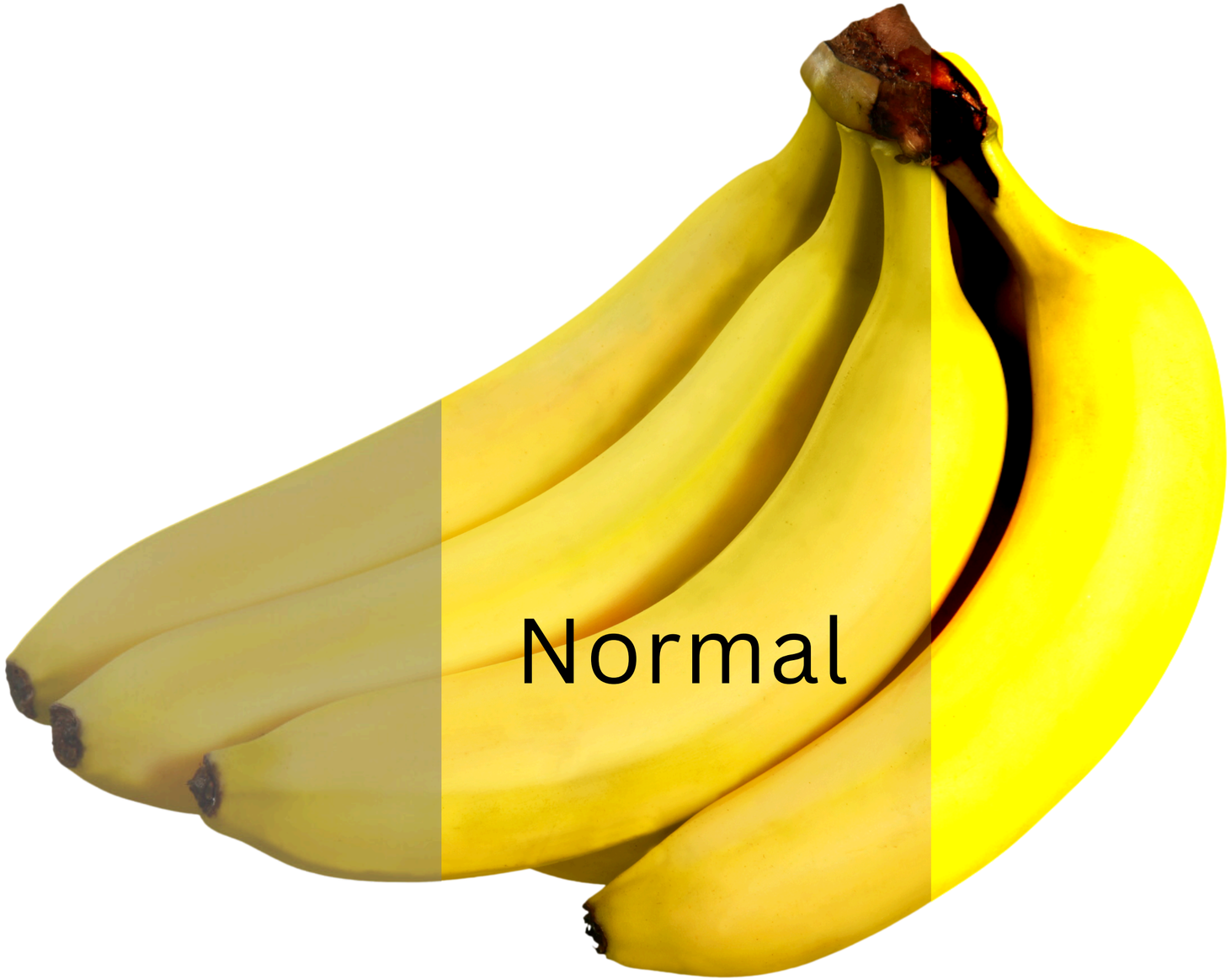
Fresh or Not?

10k rows of data

Red, Green, Blue (RGB)



Contrast

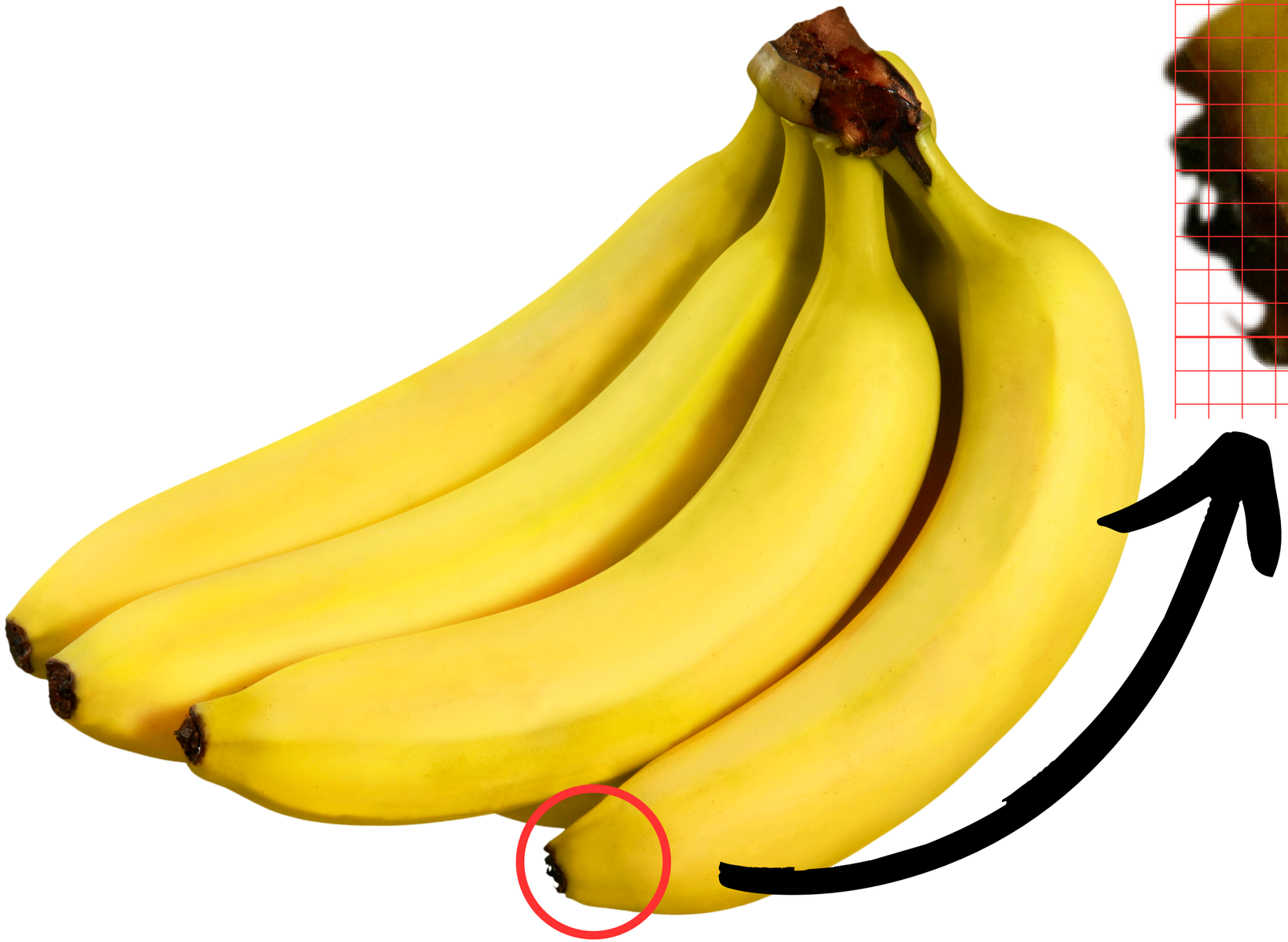


Low

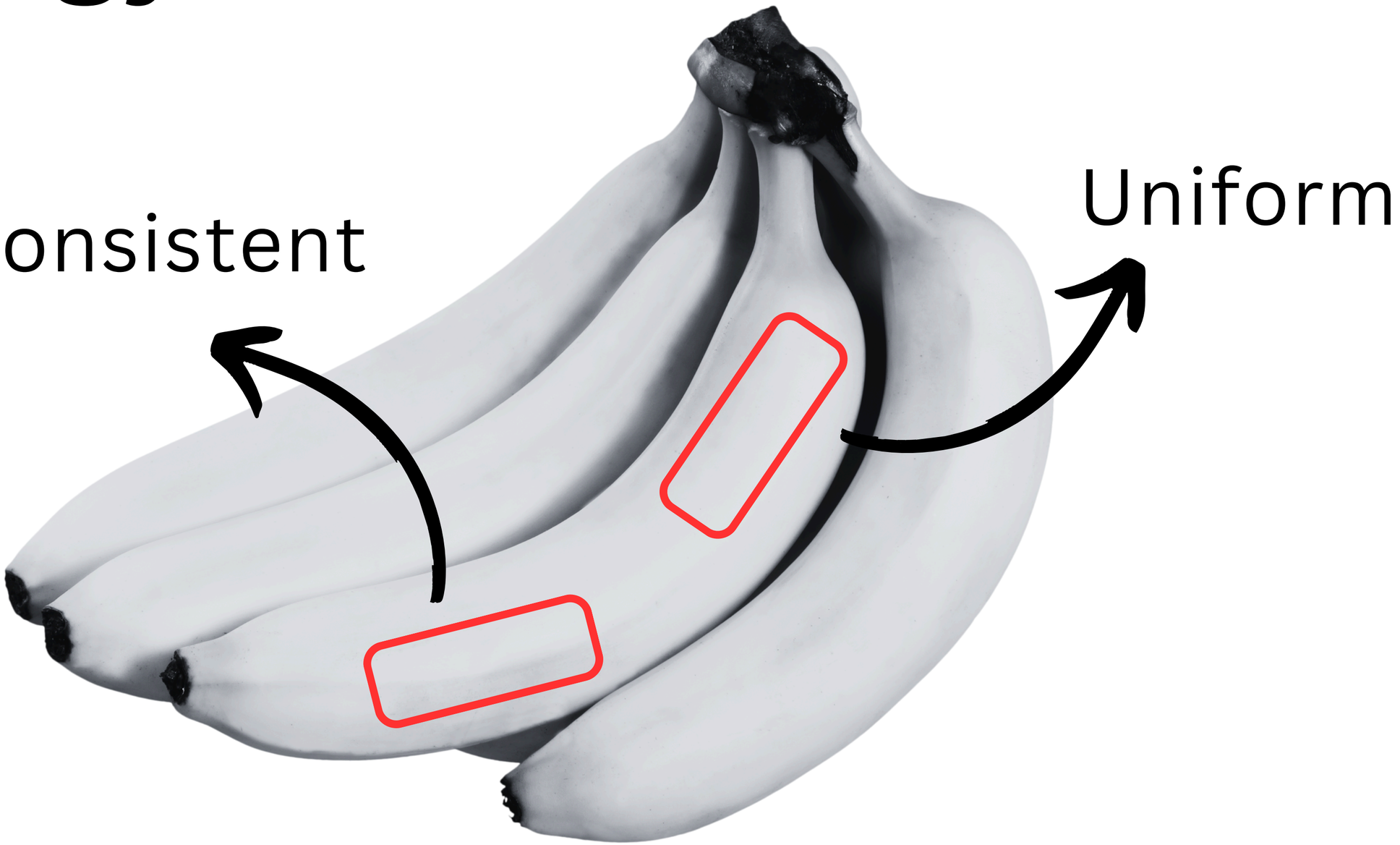
Normal

High

Pixel Correlation

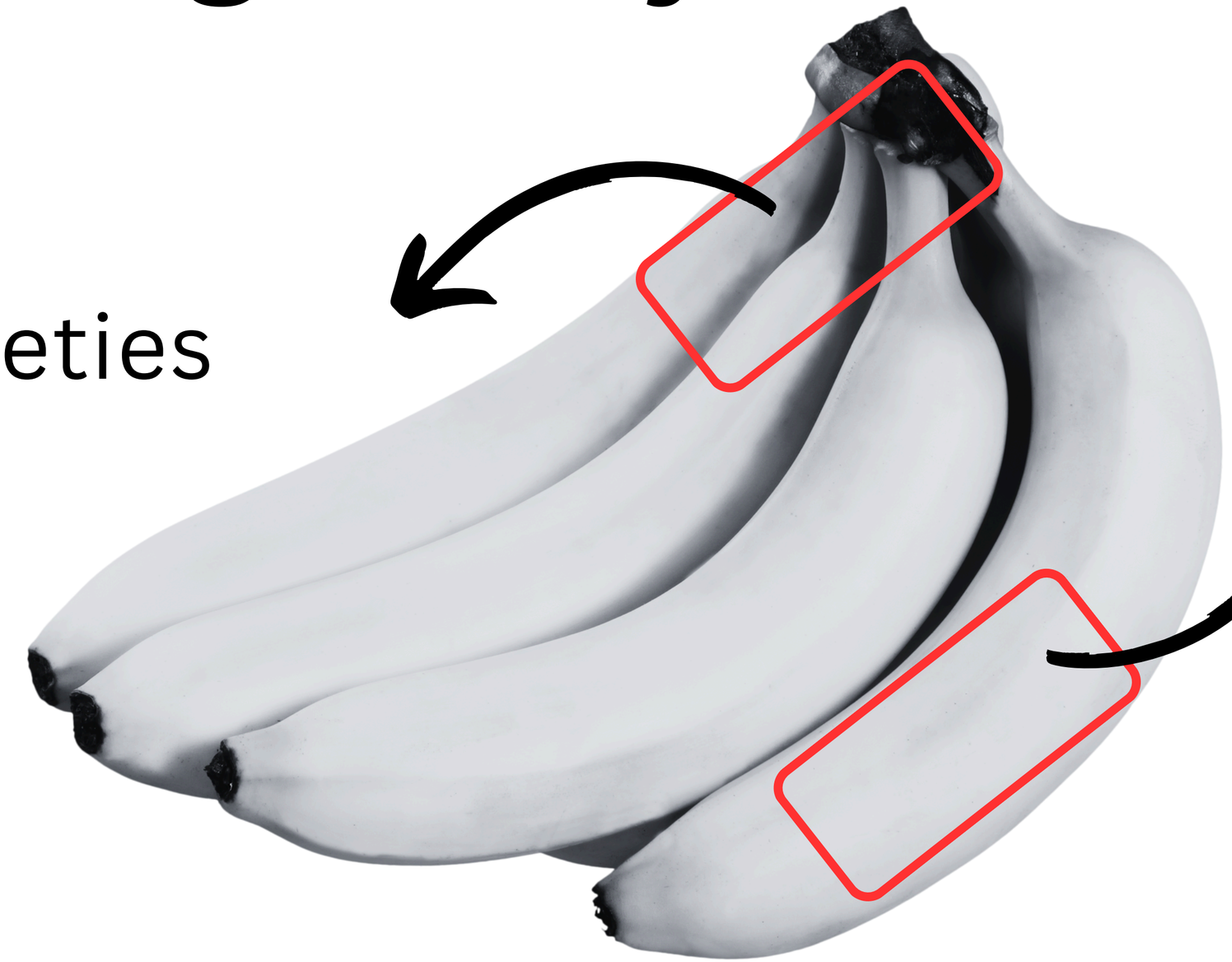


Energy (Smoothness)



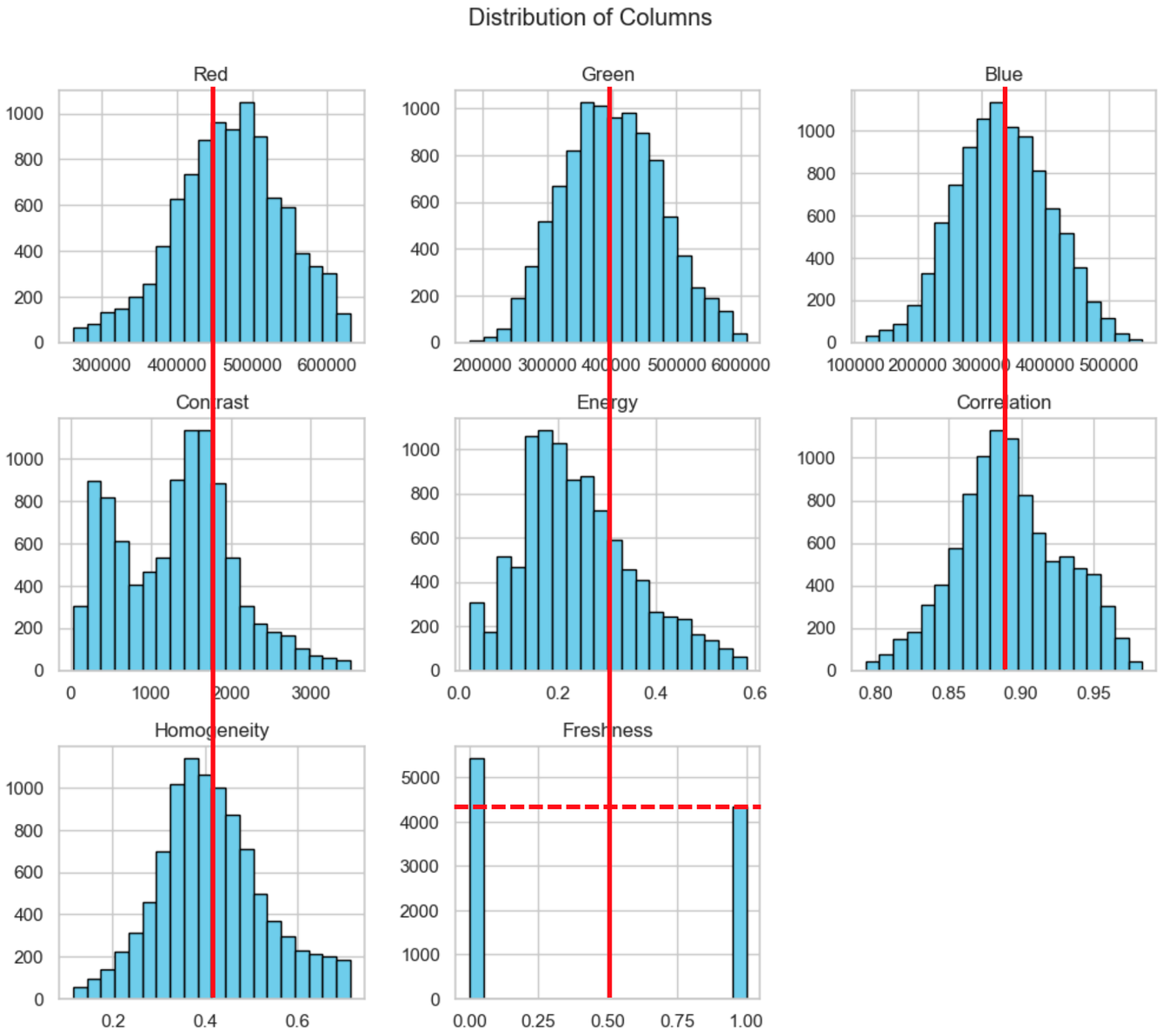
Homogeneity

Varieties

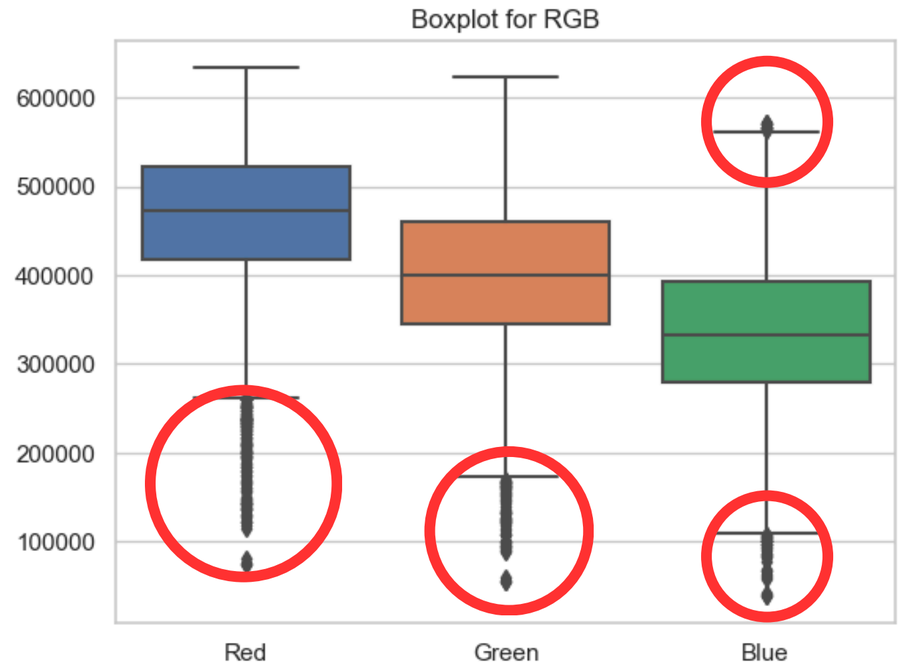


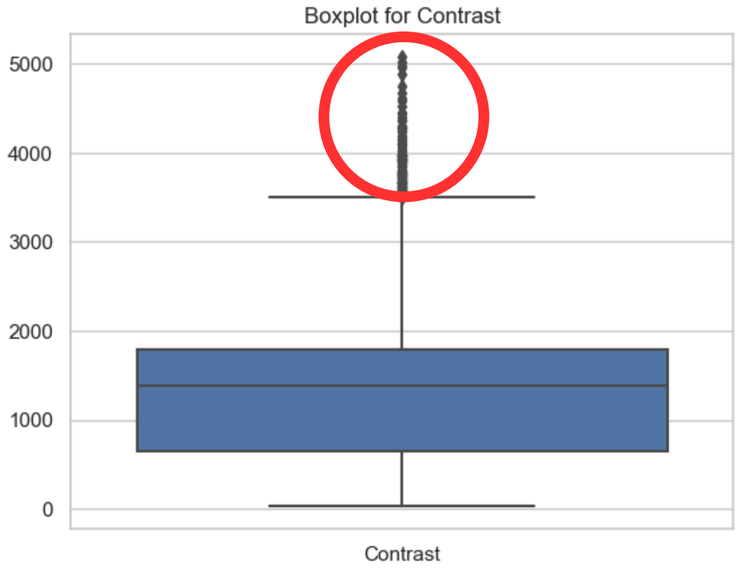
Similar

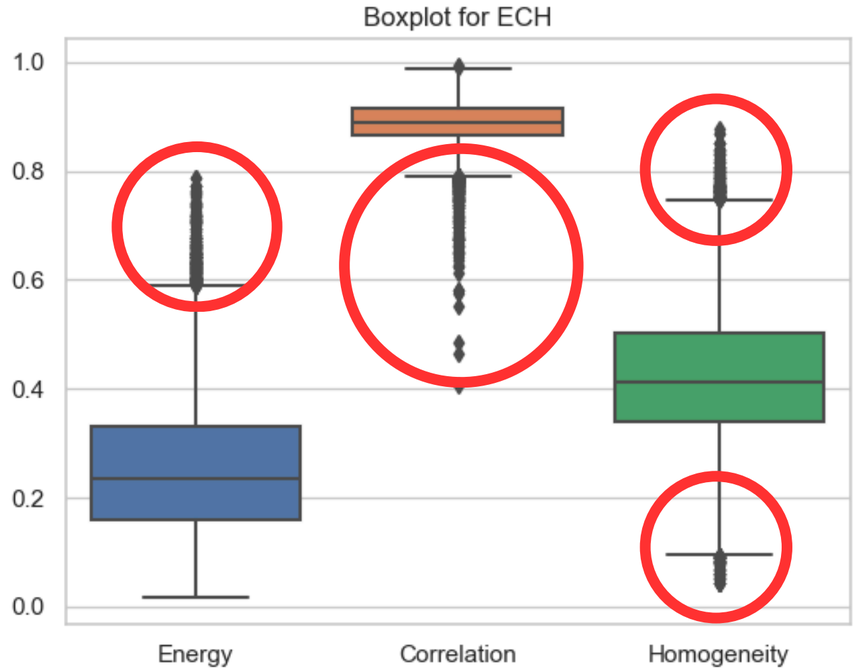
Data Distribution (Overall)



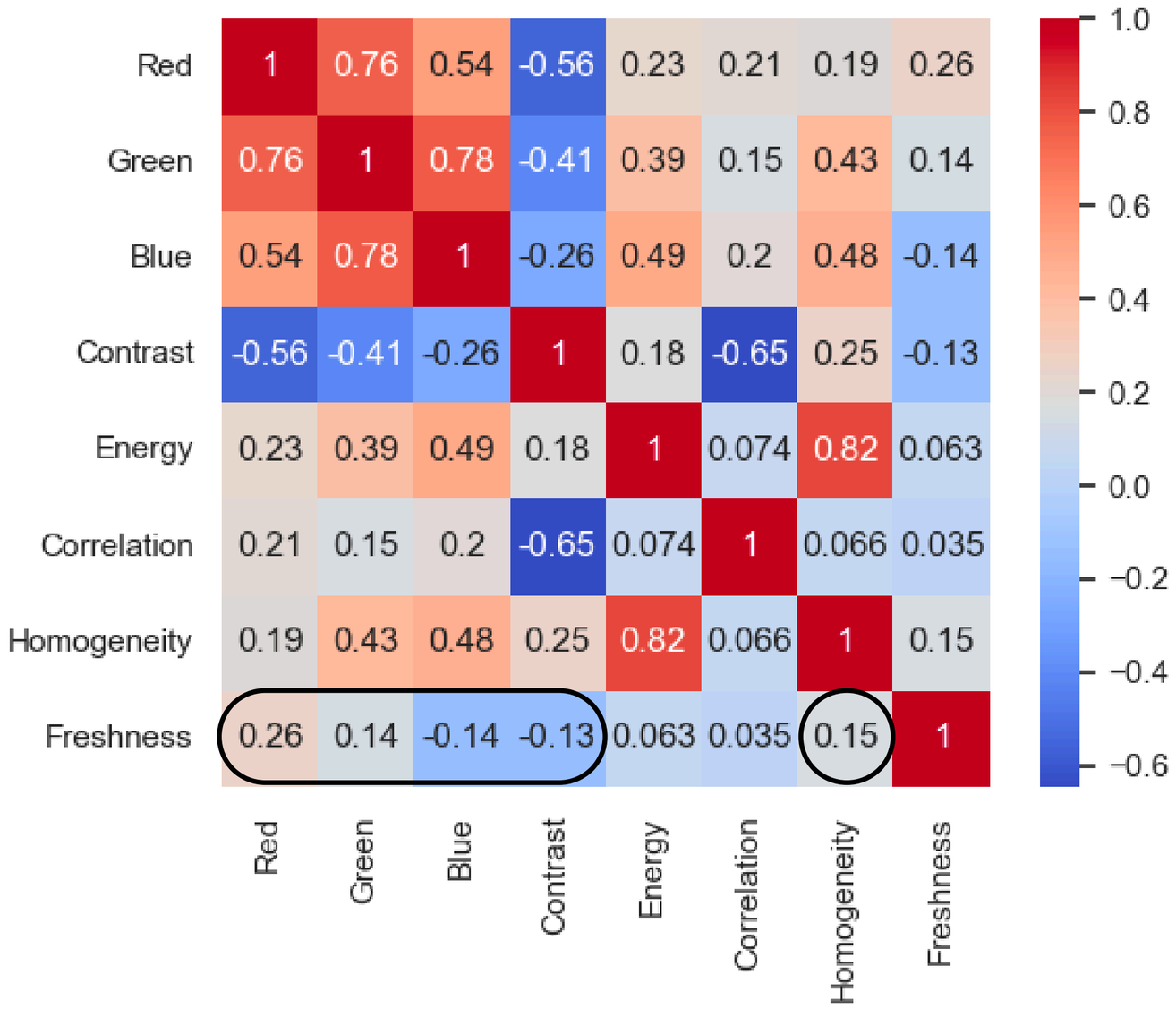
Outliers (Box Plot)



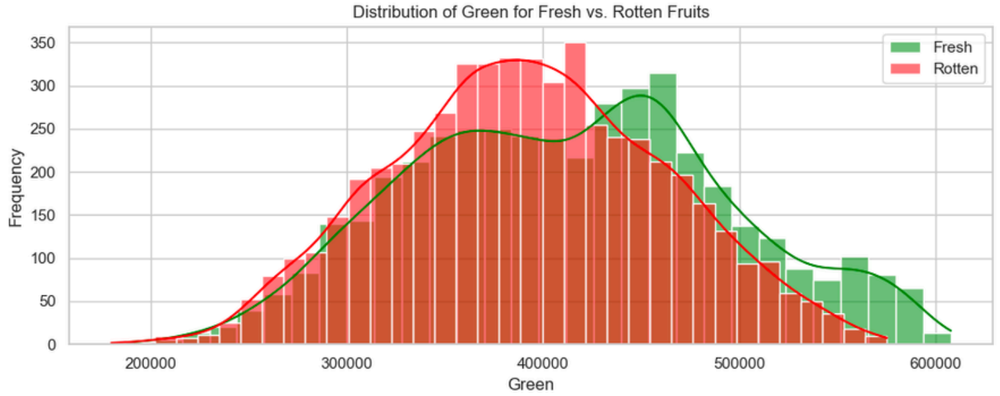


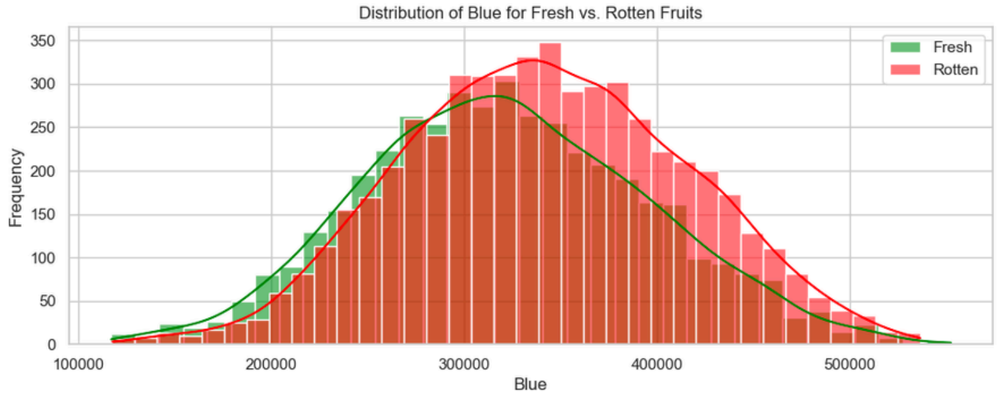


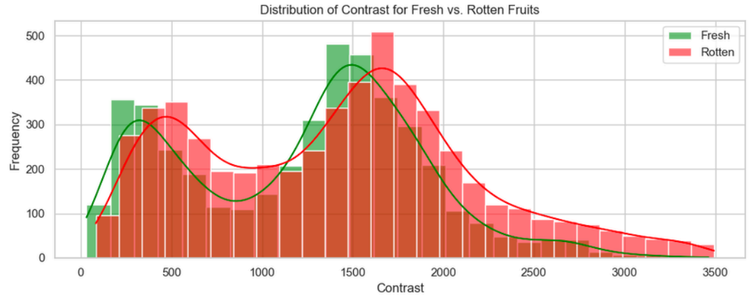
Correlation Matrix

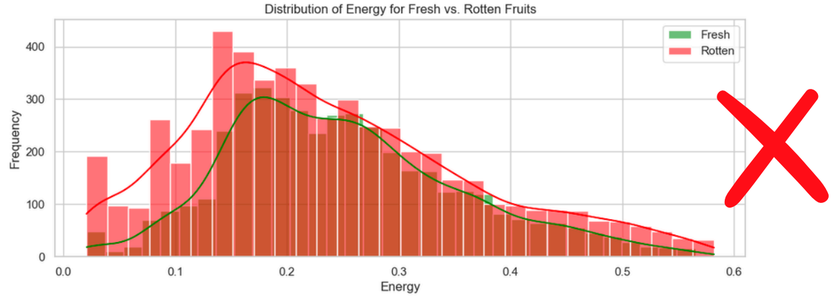


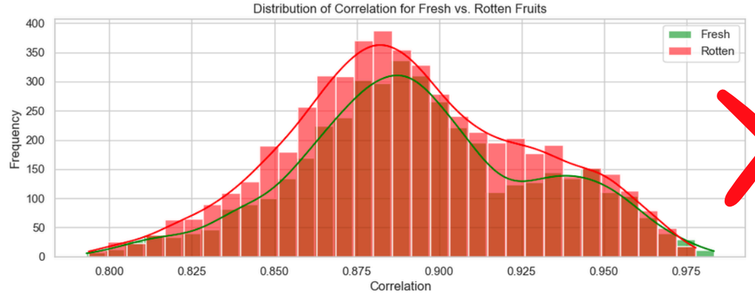
Data Distribution (Versus)

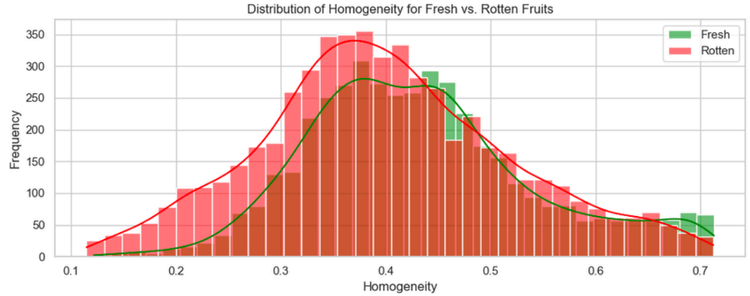




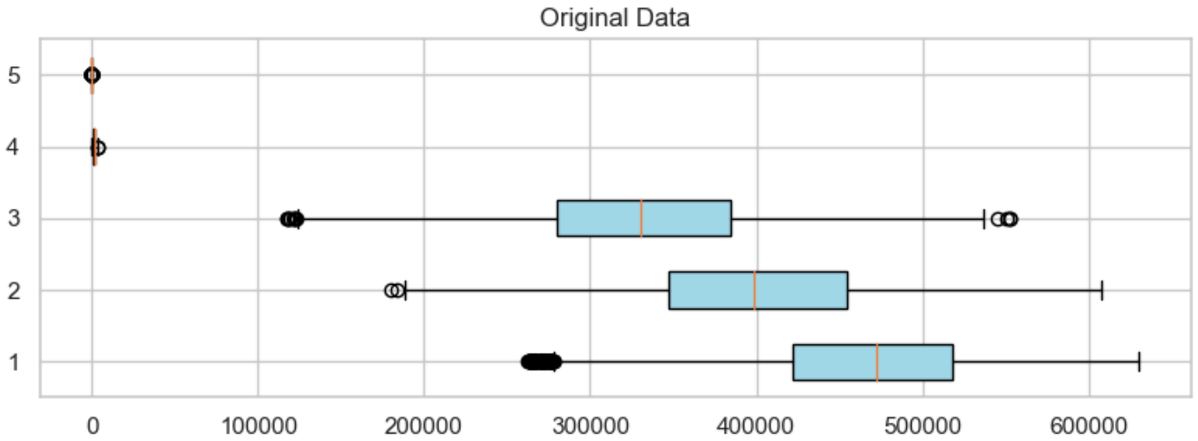


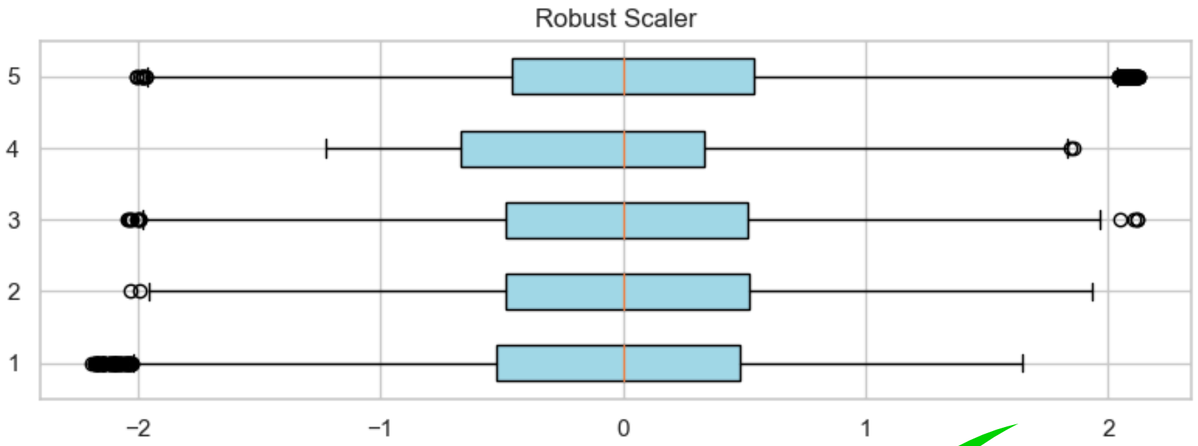


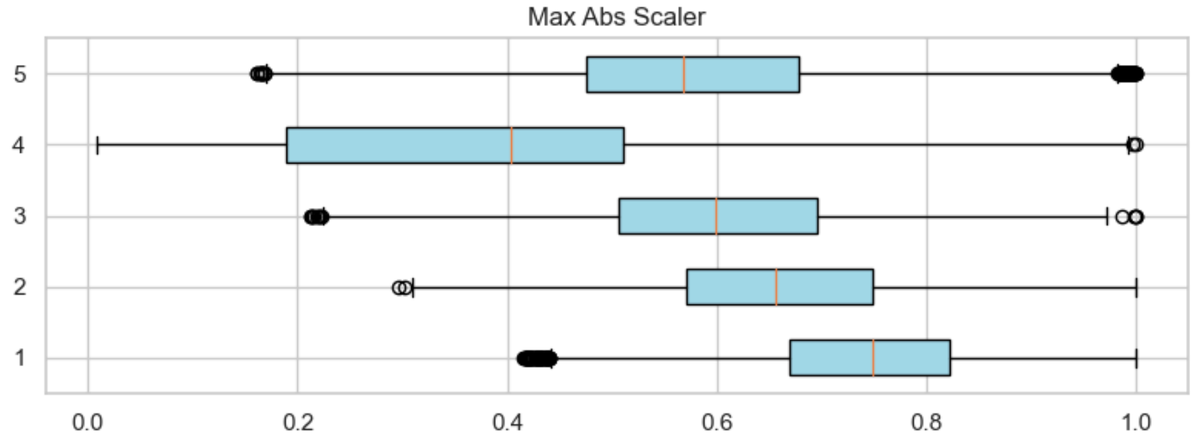


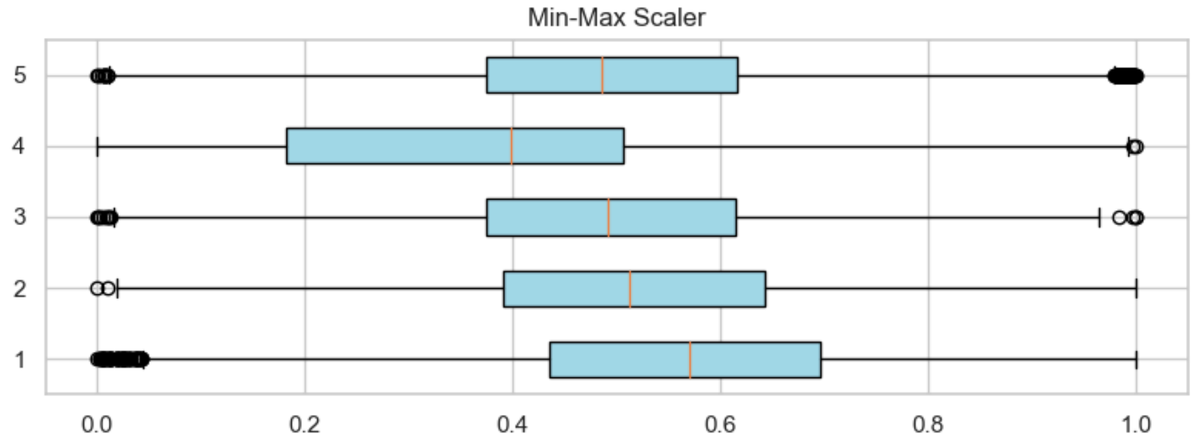


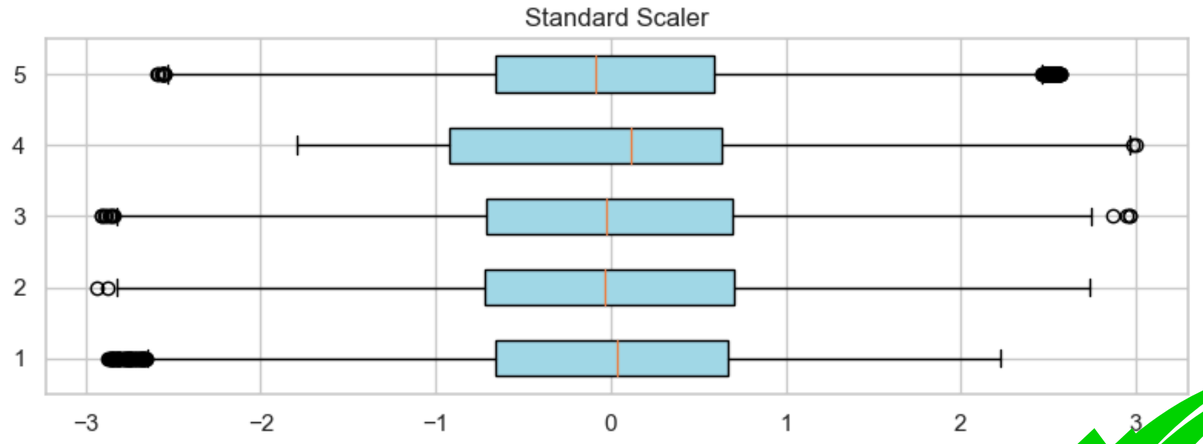
Data Scaling











Classification Performance Metrics

Accuracy

$$\frac{\text{True Positives}+\text{True Negatives}}{\text{Total Instances}}$$

Precision

$$\frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$$

Recall (Sensitivity)

$$\frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$$

Matthews Correlation Coefficient (MCC)

$$\text{MCC} = \frac{(\text{TP} \backslash\text{times}\text{TN}) - (\text{FP} \backslash\text{times}\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$$

**Covers 4 confusion matrix

Kappa Coefficient (Cohen's Kappa)

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

**Observed accuracy to expected accuracy

F1 Score

$$: 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision}+\text{Sensitivity}}$$

Classification Performance on Train Dataset

	Model	Accuracy	Precision	Senstivity	F1 Score	MCC Score	Kappa Coeff
0	AdaBoost	75.48546137615149	75.12617461709034	76.36380413385825	75.6858310714984	51.052052357826014	50.969840609939446
1	XGBoost	90.16314137203437	89.71703535512196	90.79191272965876	90.21351739753547	80.40349648550983	80.32590553723433
2	Naive Bayes	65.37296073284958	64.7919044291292	67.39173228346456	66.01154551828718	30.824124064967943	30.744519545851016
3	KNN	89.86166435077966	88.81664829027555	91.316437007874	89.98761372177042	79.8752535618295	79.72270489286548
4	Random Forest	91.54037362976685	91.7492087057908	91.76201607611547	91.52793694569417	83.26089987101402	83.18549968458122
5	Decision Tree	86.41145592095104	86.72524078859796	85.80790682414698	86.35653483094515	72.8106564010814	73.16385459108209
6	SVM	85.49297514281302	85.96243453940791	84.94156003937009	85.39495363566962	71.0763040058981	70.98558190785963
7	Logistic Regression	73.71494004425915	73.94028016863263	73.37331856955379	73.60645739684533	47.491609728247056	47.42882443766669

8 models evaluated

Classification Performance on Test Dataset

	Model	Accuracy	Precision	Senstivity	F1 Score	MCC Score	Kappa Coeff
0	AdaBoost	75.0853242320819	71.18768328445748	74.2354740061162	72.67964071856288	49.836720621706675	49.79975290654009
1	XGBoost	89.72696245733789	89.06128782001551	87.76758409785933	88.40970350404312	79.19278566381003	79.185950298146
2	Naive Bayes	62.32081911262799	56.986301369863014	63.608562691131496	60.115606936416185	24.747727627440323	24.614065180102905
3	KNN	87.37201365187714	85.10479041916167	86.92660550458714	86.00605143721634	74.51729928801278	74.50342713826899
4	Random Forest	90.0	89.80392156862746	87.53822629969419	88.6566008517228	79.73891241731465	79.71810555212993
5	Decision Tree	85.05119453924914	83.77329192546584	82.49235474006116	83.12788906009246	69.71686987690813	69.71020129944921
6	SVM	84.43686006825939	82.97213622291022	81.9571865443425	82.46153846153847	68.47895010536217	68.47476055465998
7	Logistic Regression	72.42320819112628	68.65671641791045	70.33639143730886	69.48640483383686	44.34821716044135	44.337440045142486

Regression Performance on Test Dataset

	Model	MSE	RMSE	MAE
0	AdaBoost	0.2491467576791808 7	0.4991460284117072 4	0.2491467576791808 7
1	XGBoost	0.1027303754266211 5	0.320515795908128 6	0.1027303754266211 5
2	Naive Bayes	0.3767918088737201 6	0.6138336980597596	0.3767918088737201 6
3	KNN	0.1262798634812286 6	0.3553587813481308	0.1262798634812286 6
4	Random Forest	0.1	0.3162277660168379 4	0.1
5	Decision Tree	0.1494880546075085 2	0.3866368510728232 5	0.1494880546075085 2
6	SVM	0.1556313993174061 3	0.3945014566733640 6	0.1556313993174061 3
7	Logistic Regression	0.2757679180887372	0.5251360948256529	0.2757679180887372

Random Forest with Hypertuning Parameters

'n_estimators': [100, **200**, 300, 400, 500],

'max_depth': [**None**, 10, 20, 30, 40],

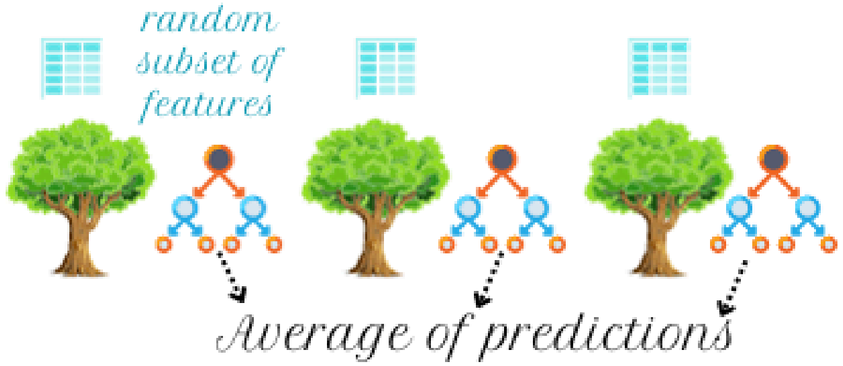
'min_samples_split': [**2**, 5, 10, 15, 20],

'min_samples_leaf': [**1**, 2, 4, 6, 8],

'max_features': ['auto', '**sqrt**', 'log2', 0.2, 0.4]

Random Forest

Parallel training



6 hours of training !!!

XG Boost with Hypertuning Parameters

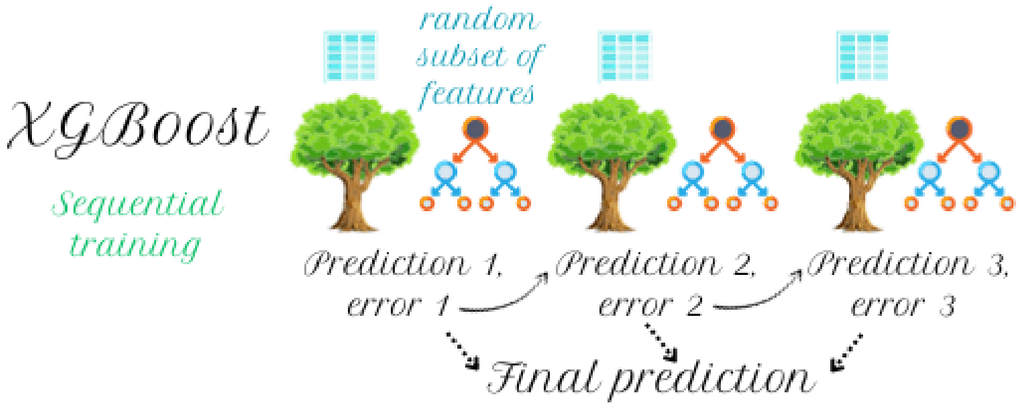
'n_estimators': [100, 200, **300**, 400, 500],

'max_depth': [3, 5, 7, 10, **15**],

'learning_rate': [0.01, 0.05, **0.1**, 0.2, 0.3],

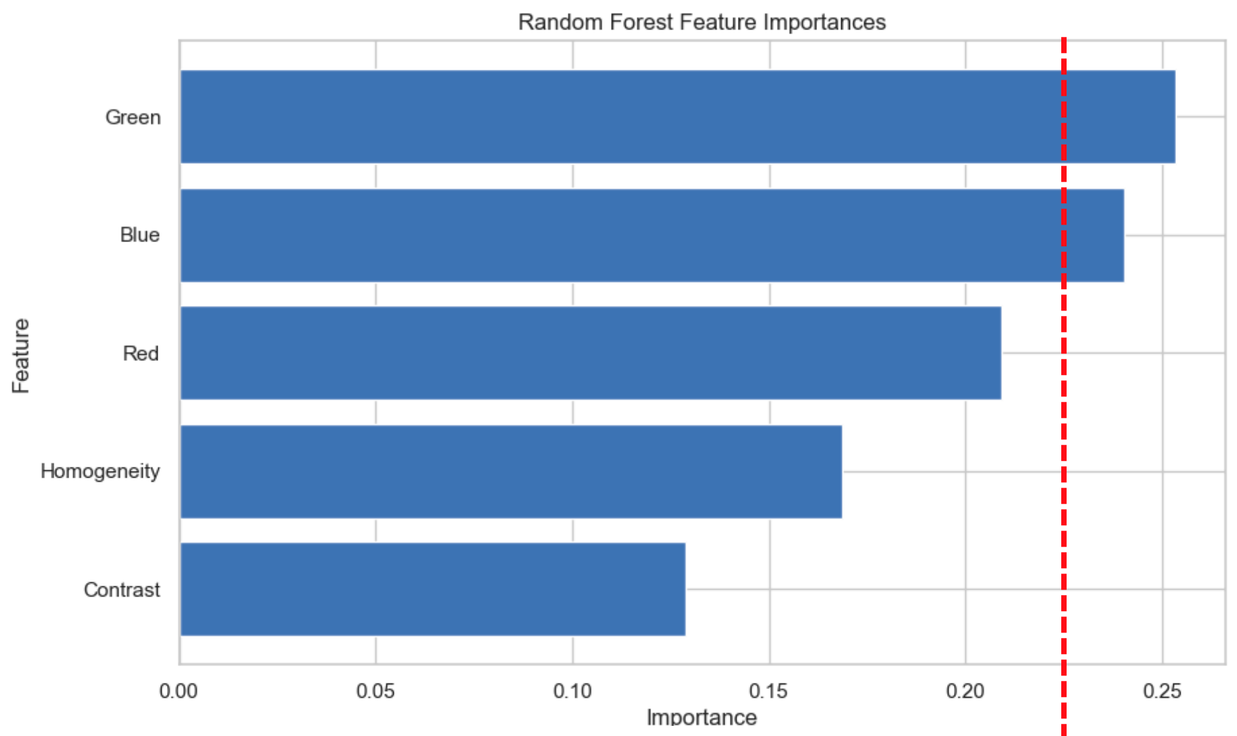
'subsample': [0.7, **0.8**, 0.85, 0.9, 1.0],

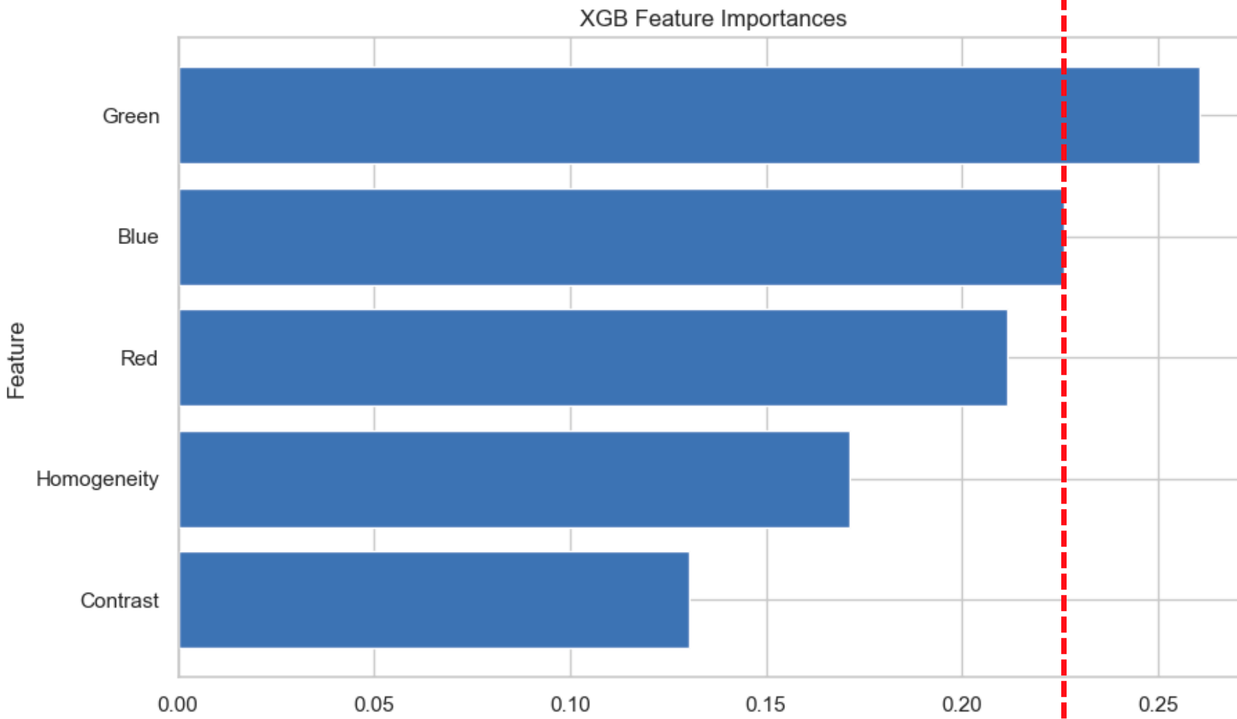
'gamma': [0, **0.1**, 0.2, 0.3, 0.5]



Only 1 hour of training !!!

Feature Importances





Classification Performance on Test Dataset

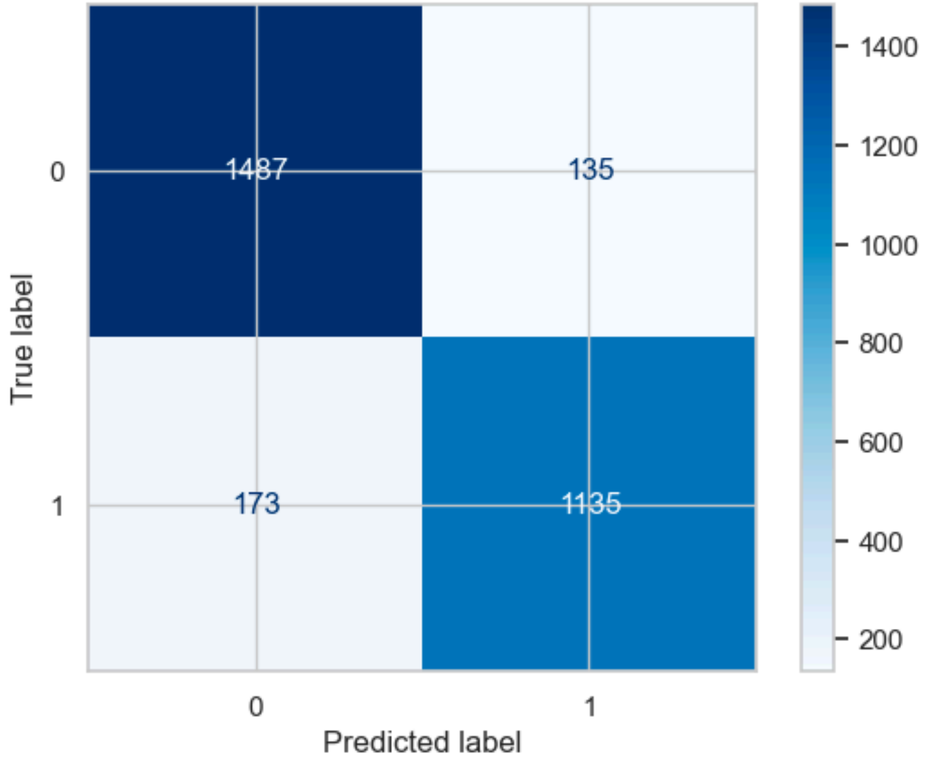
Model	Random Forest	XG Boost
Accuracy	0.8948	0.9058
Precision	0.8937	0.8975
Sensitivity (Recall)	0.8677	0.8906
F1	0.8805	0.8940
MCC	0.7869	0.8092
Kappa	0.7867	0.8092

Regression Performance on Test Dataset

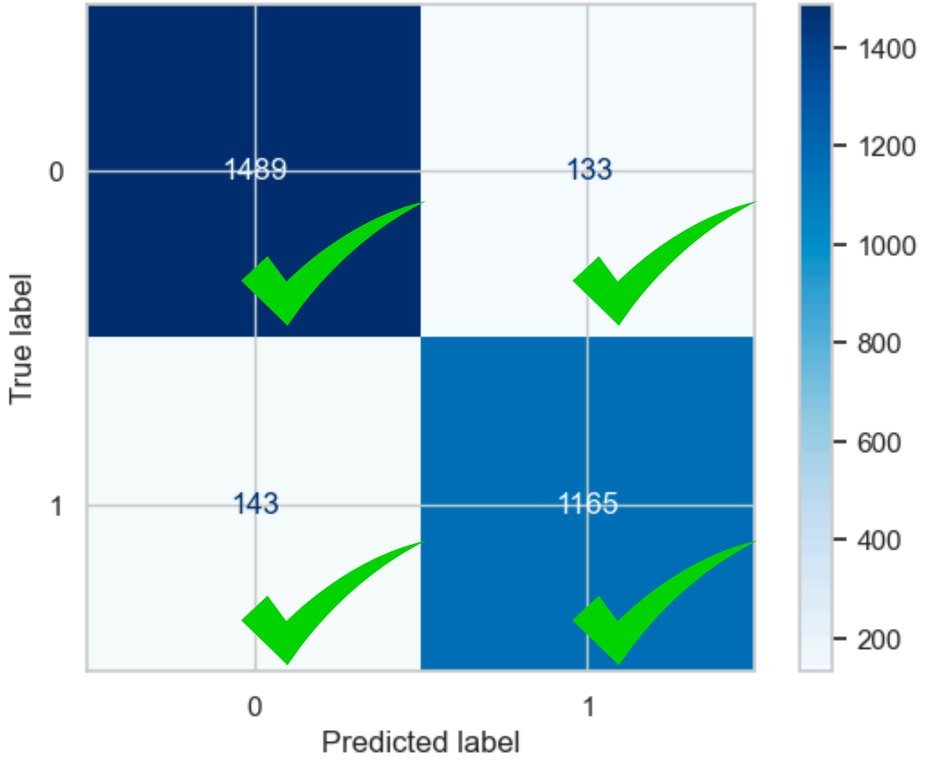
Model	Random Forest	XG Boost
MSE	0.1051	0.09419
RMSE	0.3242	0.3069
MAE	0.1051	0.0942

Confusion Matrix

Random Forest



XG Boost



Streamlit Dashboard Demo

Limitation & Improvements

Image properly masked before feature extraction

Require more image features

Identify the name and specific type of fruits

Interview expert aunties on how they differentiate fresh or rotten fruits

Use Neural Network models like GANs or CNN for real-time analysis

Target Audience & Industry

Fruit Inspection Manager

Government Inspector

Consumer

Streamlit Dashboard

Github Repo

THANK YOU