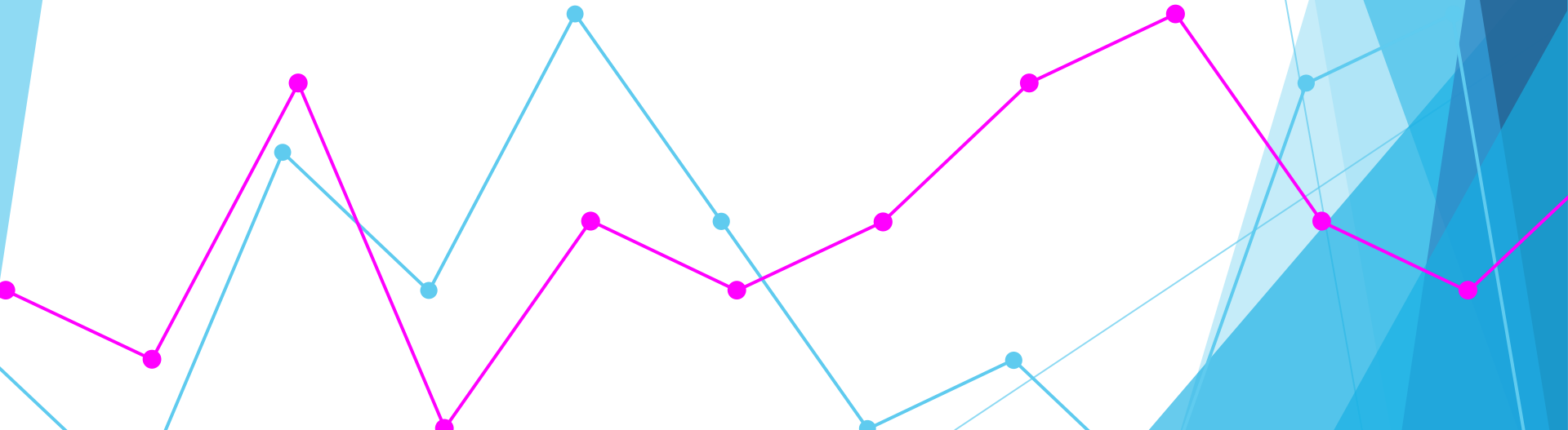
Wine Data Set

Kelepiri Zoi (78)
Vasilogamvros Evangelos (100)



Wine Data Set

Article Study and Presentation

- Six “W” questions for learning analytics
- Constructs & measurement model
- Predictive modelling

Data Analysis

- Descriptive statistics
- Machine Learning Models

Purpose of Research (1/2)

- This research aims to predict wine quality using synthetic data and experimental data from New Zealand's diverse regions. 18 Pinot noir wine samples with 54 characteristics were used, with 1381 samples generated using the SMOTE method. Six samples were retained for model testing. The quality of New Zealand Pinot noir wines is crucial in wineries worldwide.

Purpose of Research (2/2)

- The study compared four feature selection approaches for predicting wine quality using key variables (Extra Trees Classifier, RF, Gradient Boosting Classifier, XGBOOST). Seven machine learning algorithms were trained, with AdaBoost showing 100% accuracy without feature selection and Random Forest showing increased performance with key variables.

Six W Questions (1/3)

- **What we are measuring?**

Wine quality Pinot noir from New Zealand. 18 samples(7 were physicochemical and 47 chemical characteristics)

- **How we are measuring?**

Synthetic data using the SMOTE method (Synthetic Minority Over-sampling Technique). Synthetic data were generated from 12 original samples.

Six W Questions (2/3)

- **Why is knowledge important to us?**

Wine quality is an important issue in the wine industry. This research aims to predict the quality of Pinot noir by selecting the most suitable characteristics using machine learning, based on synthetic data and experimental data from different regions of New Zealand.

- **Who is the analytic for?**

The analysis is aimed at those interested in the wine industry, in particular those who want to predict the quality of Pinot noir using machine learning. The article also mentions that there are experts who worked on predicting the quality of wine.

Six W Questions (3/3)

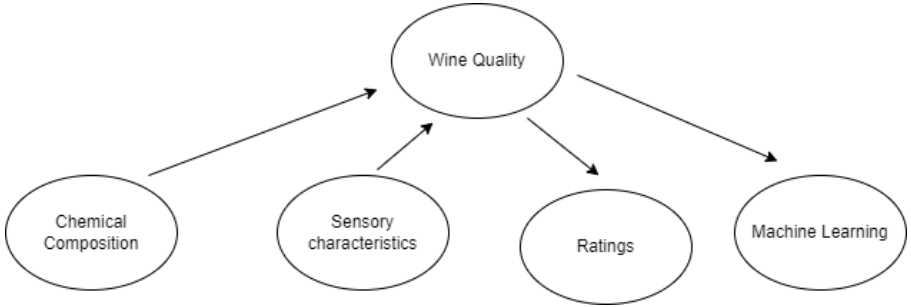
- **Where does the data collection happen?**

Experimental data is collected from various and varied regions throughout New Zealand that produce Pinot noir wine.

- **When does the data collection & feedback occur?**

No specific information is provided about when the data is collected or the feedback is given.

Measurement Model (1/4)



- **Constructs**

  Wine quality

- **Latent variable**

  The key latent variable on this scenario is the wine quality.

- **The indicators? (text and graphics)**

  • Chemical makeup (casual)

  • Sensory qualities (casual)

  • Opinions of experts (effect)

  • Machine Learning Models (effect)

Measurement Model - Constructs Type (2/4)

- Chemical composition both continuous and discrete according (the flavor,4-ethyl-2-methoxyphenol)
- Sensory characteristics both continuous and discrete according (softness, black glass)
- Assessments from experts are continuous
- Machine learning models results are continuous
- The wine quality is continuous (rating from low to high)

Measurement Model - Instruments (3/4)

- **Chemical Composition:** Chromotographty, ph – meter, spectrophotometry
  - Reliability → calibration, precision testing, and consistency
  - Validity → accuracy of these instruments in measuring specific components.
- **Sensory Evaluations:** Experts assessments with specific techniques
  - Reliability → rigorous training, calibration, and consistency checks
  - Validity →sensory descriptors, reference standards, and statistical analyses.
- **Experts evaluation:** use rating scales or scoring sheets
  - Reliability → training and calibration exercises
  - Validity →sensory and chemical analyses.
- **Machine Learning:** SVM, RF, and AdaBoost predict wine quality based on input features
  - Reliability → accuracy, precision, recall, and F1 score
  - Validity →involves models accurately reflecting data patterns

Measurement Model – Math Model (4/4)

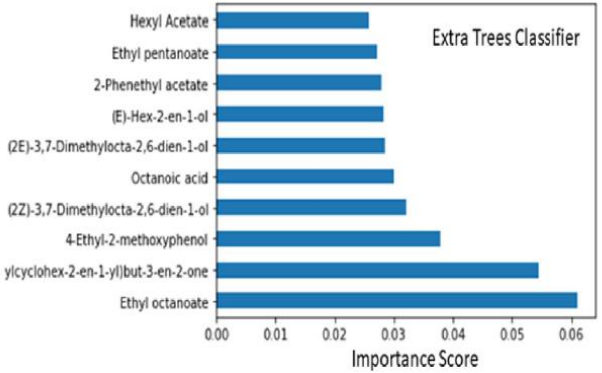
- **Math Model:** $x_{ij}=\lambda_{ij}\eta+\varepsilon_{ij} + a$
- *x_{ij}: indicator*
- *λ_{ij}: factor loading coefficient*
- *ε_{ij}: random error*
- *a intercept*
- *η: latent variable*

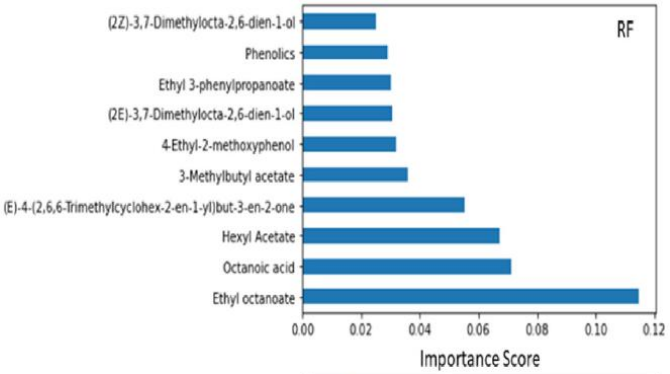
* *(i equals to number of constructs, j the value of each construct)*

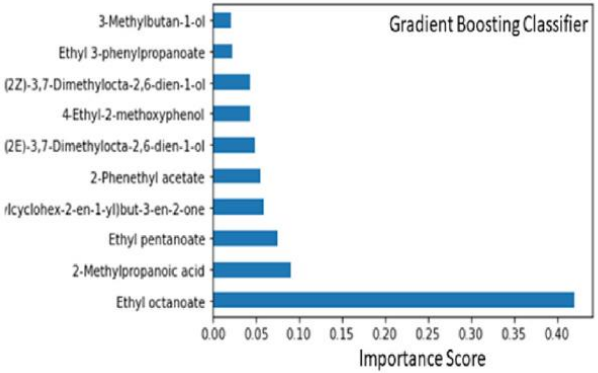
Predictive Model

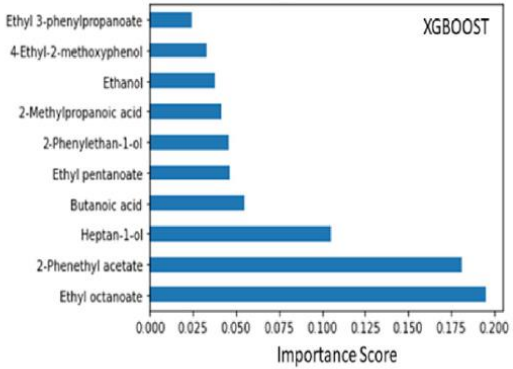
- Predictive modeling is a commonly used statistical technique to predict future behavior.
- Classification is a process related to categorization, the process in which ideas and objects are recognized, differentiated and understood.
- Classification is the grouping of related facts into classes. It may also refer to a process which brings together like things and separates unlike things.

Features selection

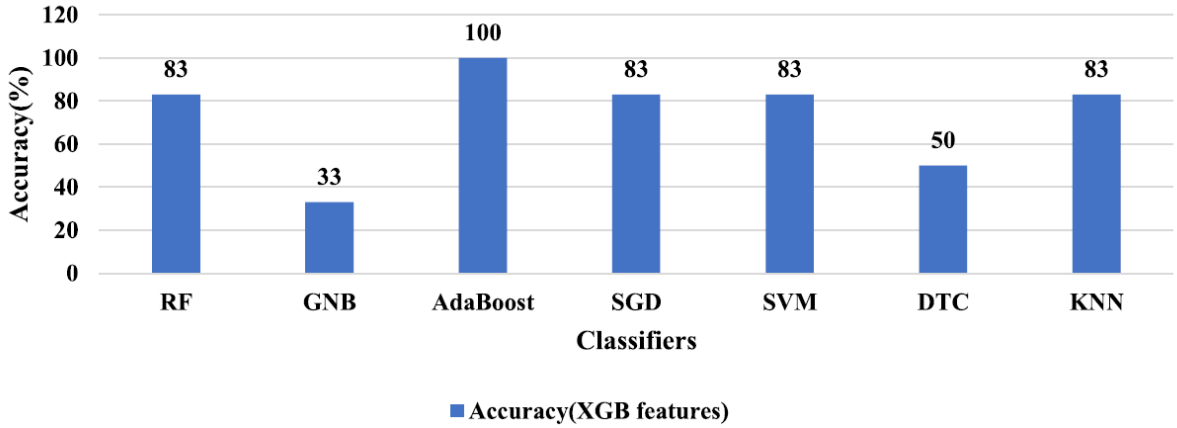


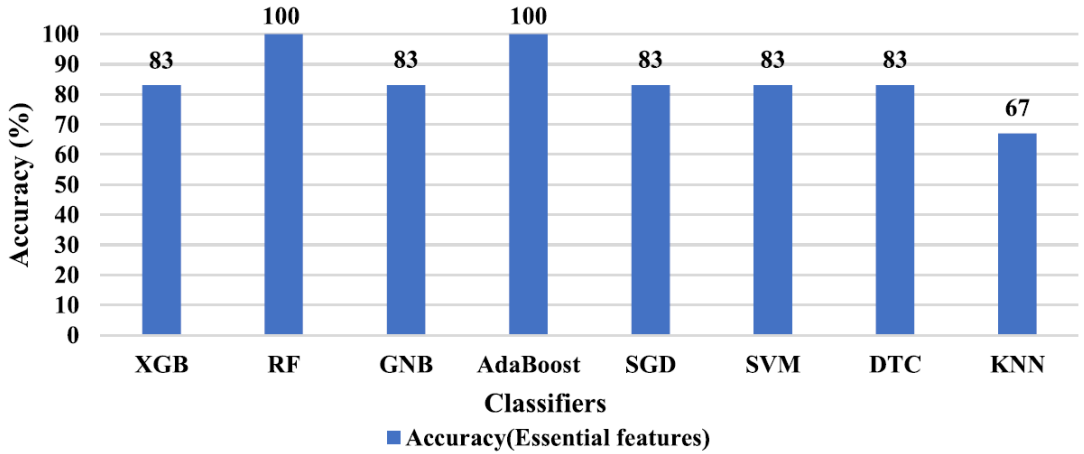






Accuracy


120
100
80
60
40
20
0
Accuracy(%)
83
33
100
83
83
50
83
RF
GNB
AdaBoost
SGD
SVM
DTC
KNN
Classifiers
Accuracy(XGB features)

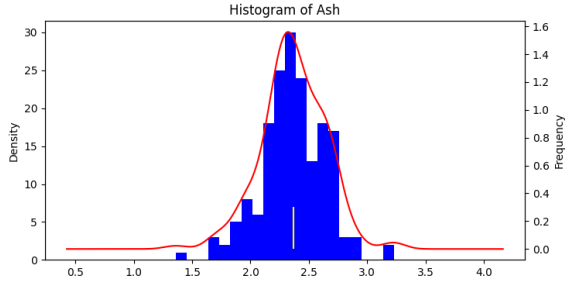

100
90
80
70
60
50
40
30
20
10
0
Accuracy (%)
83
100
83
100
83
83
83
67
XGB
RF
GNB
AdaBoost
SGD
SVM
DTC
KNN
Classifiers
Accuracy(Essential features)

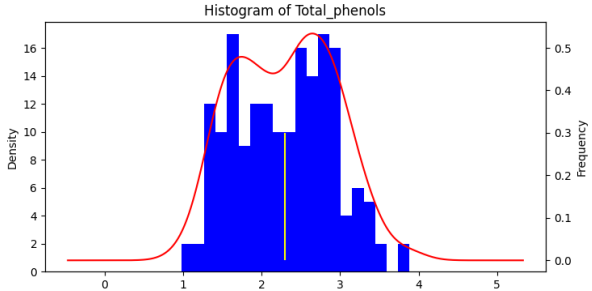
Variables in DataSet

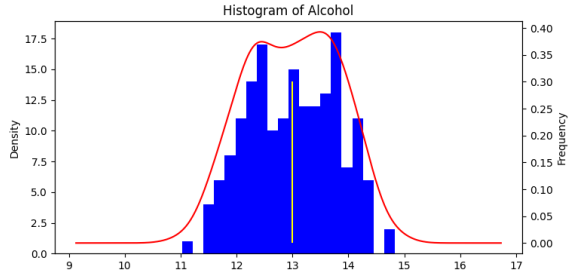
Name	Content	Type
Class	Κατηγορική (ποιοτική)	Τάξη
Alcohol	Συνεχής (ποσοτική)	Αλκοόλ
Malicacid	Συνεχής (ποσοτική)	Μηλοξίνη
Ash	Συνεχής (ποσοτική)	Στάχτη
Alcality of Ash	Συνεχής (ποσοτική)	Αλκαλικότητα Στάχτης
Magnesium	ποσοτική	Μαγνήσιο
Total phenols	Συνεχής (ποσοτική)	Συνολικές φαινόλες
Flavanoids	Συνεχής (ποσοτική)	Φλαβανοειδή
Non-flavanoid phenols	Συνεχής (ποσοτική)	Μη φλαβανοειδείς φαινόλες
Proanthocyannis	Συνεχής (ποσοτική)	Προανθοκυανίνες
Color - intensity	Συνεχής (ποσοτική)	Ένταση χρώματος
Hue	Συνεχής (ποσοτική)	Απόχρωση
Diluted wines	Συνεχής (ποσοτική)	Αραιωμένο κρασί
Proline	ποσοτική	Προλίνη

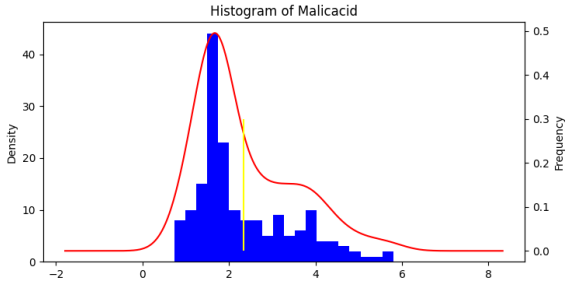
Descriptive Statistics

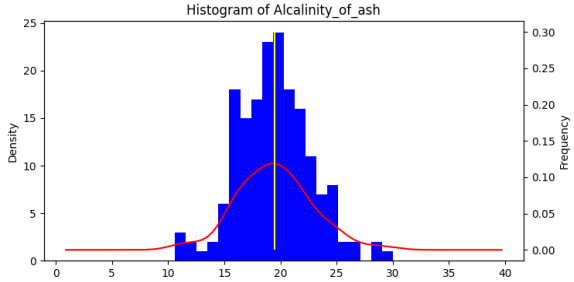
Variable	count	mean	std	min	25%	50%	75%	max
Class	178,00	1,94	0,77	1,00	1,00	2,00	3,00	3,00
Alcohol	178,00	13,00	0,81	11,03	12,36	13,05	13,67	14,83
Malicacid	178,00	2,36	1,11	0,74	1,60	1,85	3,08	5,8
Ash	178,00	2,36	0,27	1,36	2,21	2,36	2,55	3,23
Alcality of Ash	178,00	19,49	3,33	10,6	17,2	19,5	21,5	30,00
Magnesium	178,00	99,74	14,28	70,00	88,00	98,00	107,00	162,00
Total phenols	178,00	2,29	0,62	0,98	1,74	2,35	2,80	3,88
Flavanoids	178,00	2,09	0,99	0,34	1,20	2,135	2,875	5,080
Non-flavanoid phenols	178,00	0,36	0,12	0,13	0,27	0,34	0,43	0,66
Proanthocyannis	178,00	1,59	0,57	0,41	1,25	1,55	1,95	3,58
Color - intensity	178,00	5,05	2,32	1,28	3,22	4,69	6,2	13,0
Hue	178,00	0,95	0,22	0,48	0,78	0,96	1,12	1,71
Diluted wines	178,00	2,61	0,7	1,27	1,93	2,78	3,17	4,00
Proline	178,00	746,9	314,9	278,0	500,5	673,5	985,0	1680,0

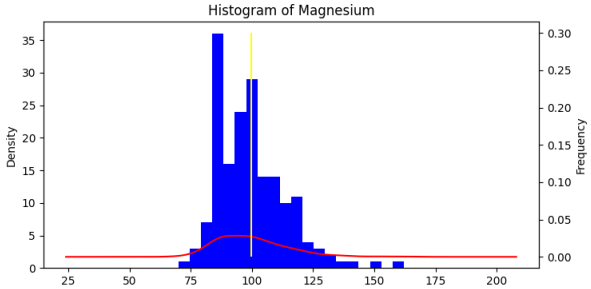

Histogram of Ash
Density
Frequency

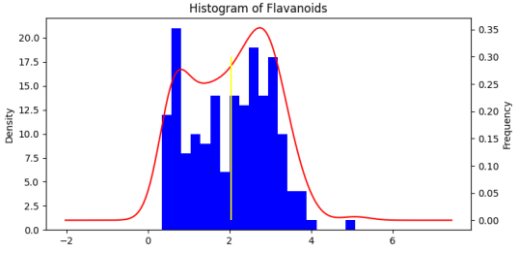

Histogram of Total_phenols
Density
Frequency

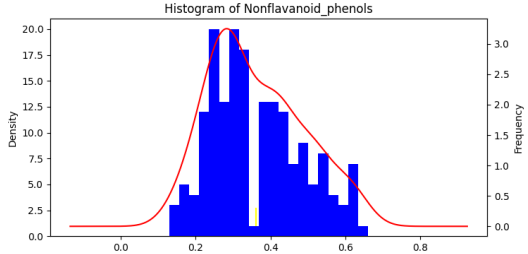

Histogram of Alcohol
Density
Frequency

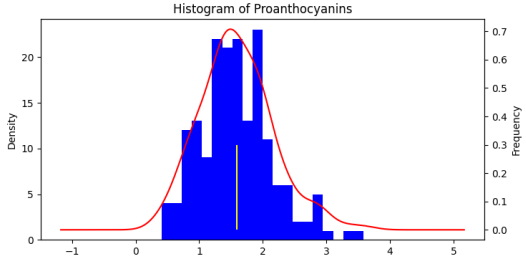

Histogram of Malicacid
Density
Frequency

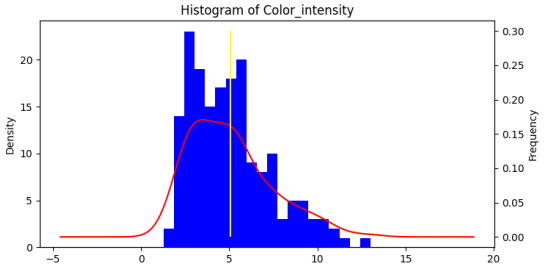

Histogram of Alcalinity_of_ash
Density
Frequency

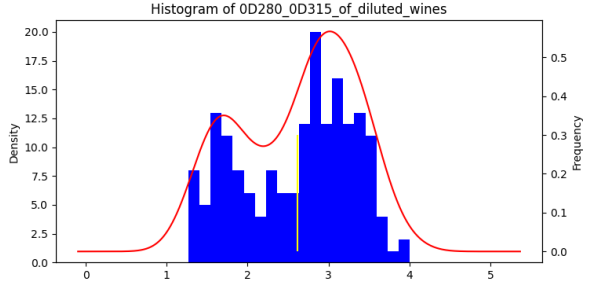

Histogram of Magnesium
Density
Frequency

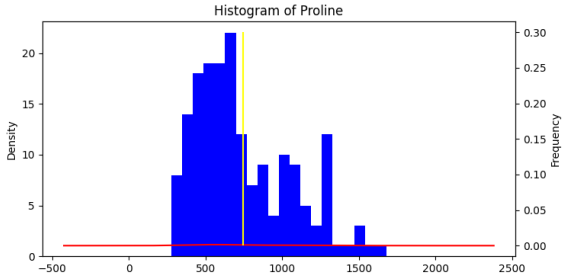

Histogram of Flavanoids
Density
Frequency

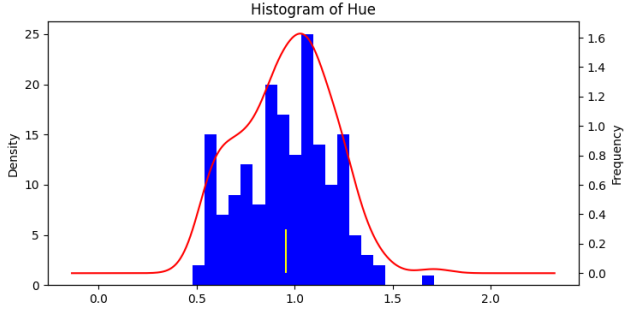

Histogram of Nonflavanoid_phenols
Density
Frequency

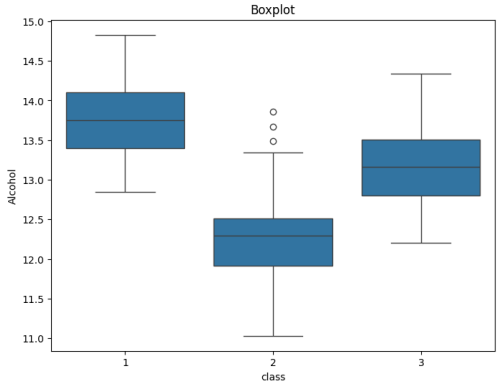

Histogram of Proanthocyanins
Density
Frequency

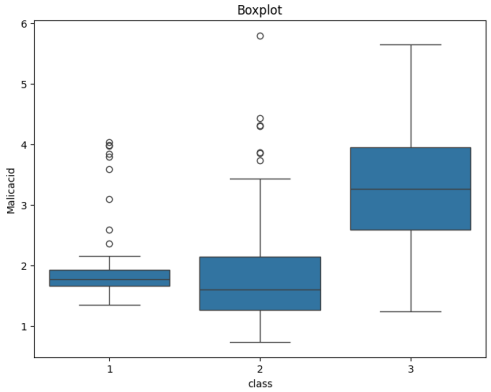

Histogram of Color_intensity
Density
Frequency

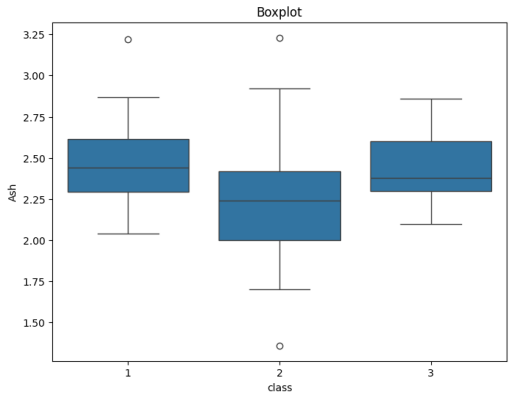

Histogram of 0D280_0D315_of_diluted_wines
Density
Frequency

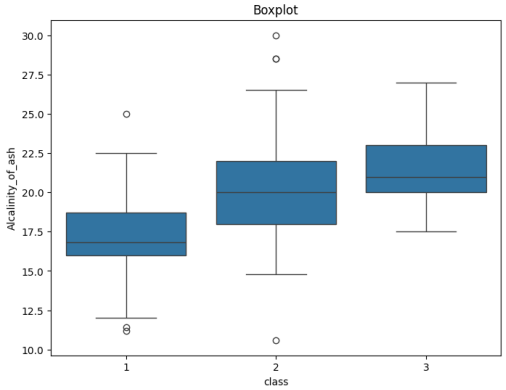

Histogram of Proline
Density
Frequency

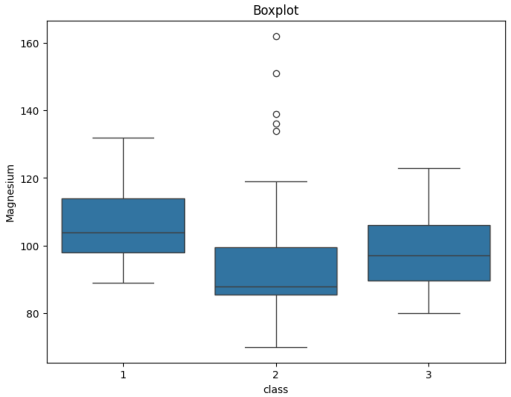

Histogram of Hue
Density
Frequency

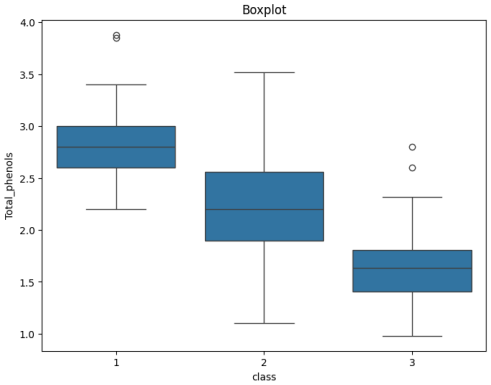

Boxplot
Alcohol
class

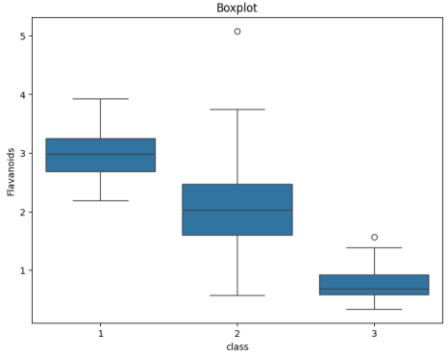

Boxplot
Malicacid
class

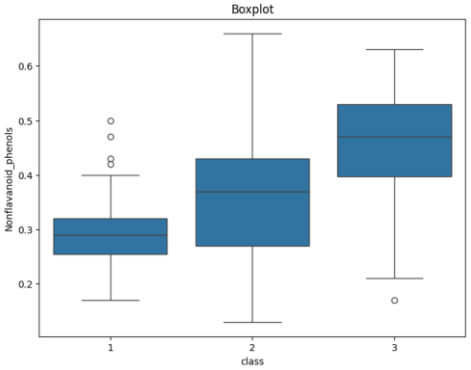

Boxplot
Ash
class

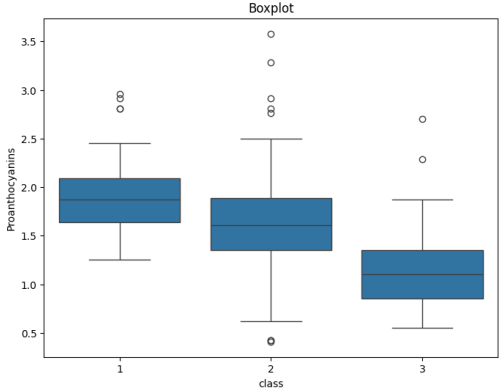

Boxplot
Alcalinity_of_ash
class

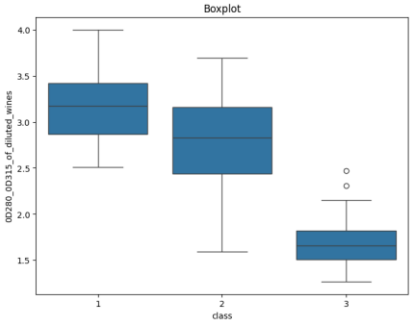

Boxplot
Magnesium
class

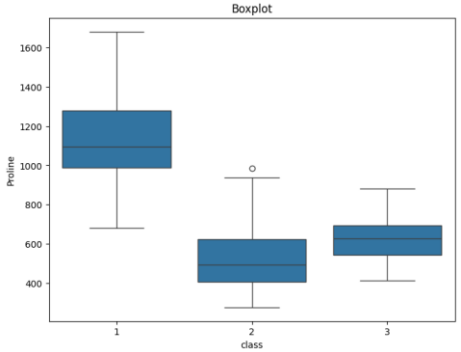

Boxplot
Total_phenols
class

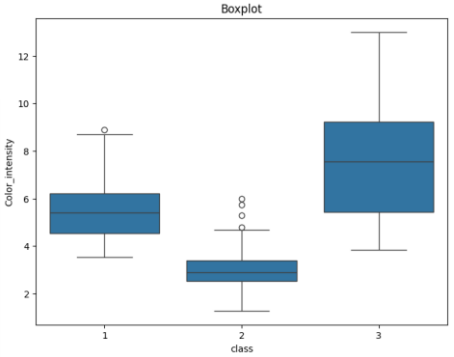

Boxplot
Flavanoids
class

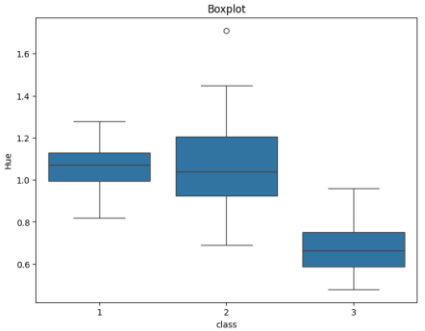

Boxplot
Nonflavanoid_phenols
class


Boxplot
Proanthocyanins
class

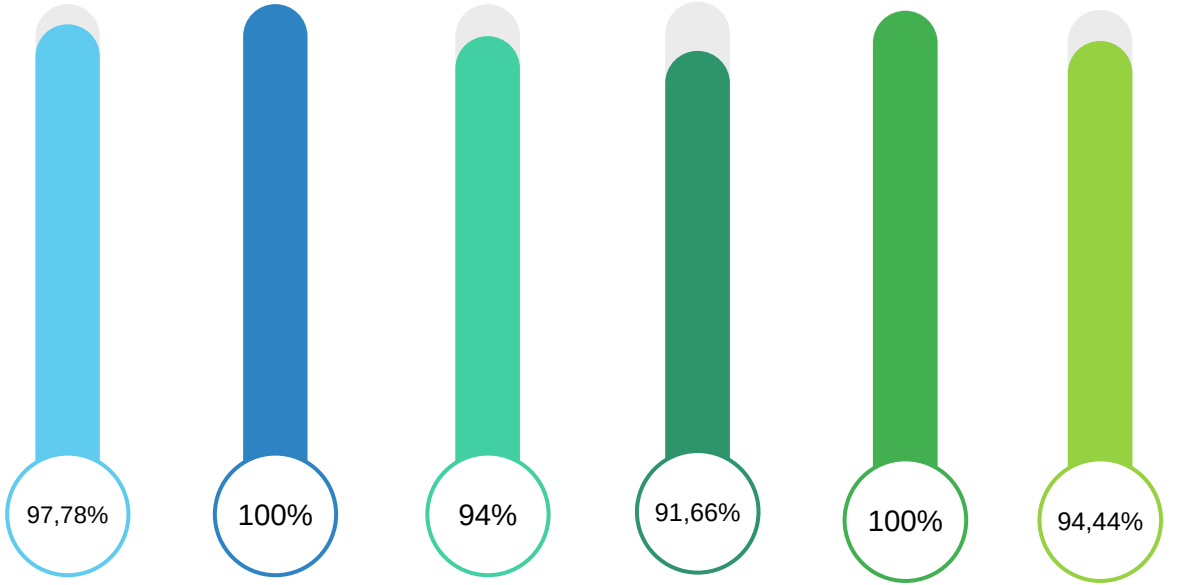

Boxplot
OD280_OD315_of_diluted_wines
class


Boxplot
Proline
class


Boxplot
Color_intensity
class


Boxplot
Hue
class

Machine Learning Model – all features



Logistic Regression

Random Forest

KNN

Decision Tree

Gradient Boosting

AdaBoost

Machine Learning Model – all features

MODEL	ACCURACY	PRECISION			RECALL			F1-SCORE		
LR	0,9778	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
		1,00	1,00	0,92	1,00	0,94	1,00	1,00	0,97	0,96
RF	1,00	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
KNN	0,94	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
		0,93	1,00	0,89	1,00	0,86	1,00	0,97	0,92	0,94
Gradient Boosting	1,00	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

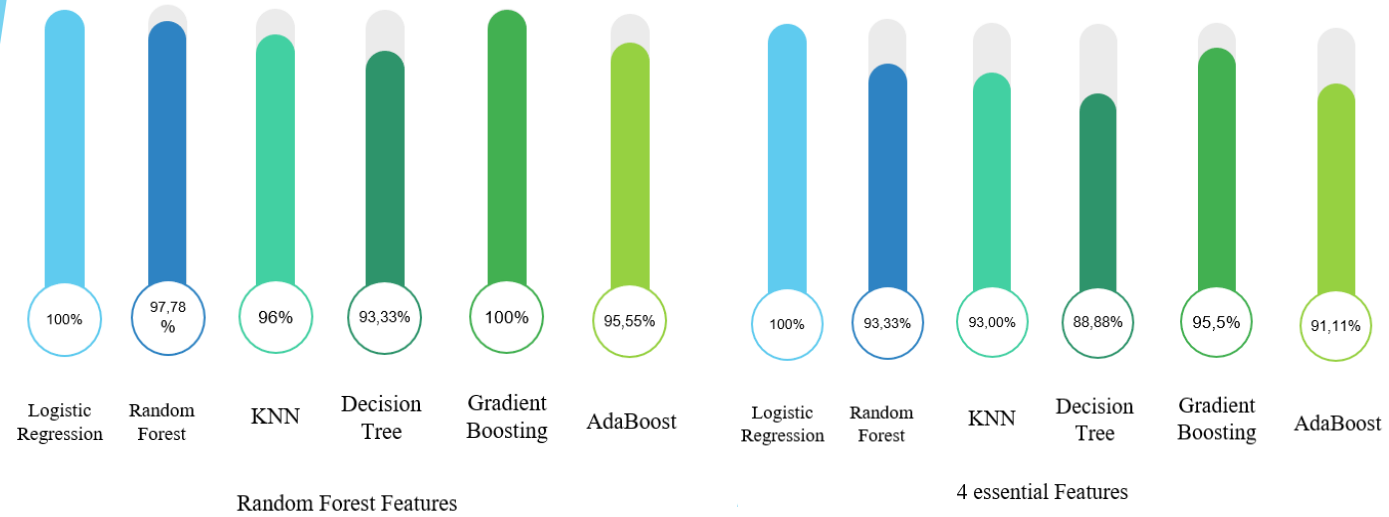
Από τα παραπάνω αποτελέσματα παρατηρούμε ότι ο Gradient Boosting και ο RF έχουν τη μέγιστη απόδοση τόσο στο accuracy, όσο και στο precision και στο recall. Αυτό σημαίνει ότι προβλέπονται σωστά όλα τα κρασιά με βάση την πραγματική κατηγορία στην οποία ανήκουν (1 → υψηλή ποιότητα, 2 → μέτρια, 3 → χαμηλή).

Στο Logistic Regression έχουμε υψηλό accuracy, ενώ το precision και το recall είναι 1. Η επιλογή του **precision** θα βοηθήσει να διασφαλίσουμε ότι τα κρασιά που προβλέπουμε ως υψηλής ποιότητας είναι πραγματικά υψηλής ποιότητας, αποφεύγοντας την κατάταξή τους ως χαμηλής ποιότητας. Αυτή η πληροφορία για την ποιότητα του κρασιού είναι αρκετά σημαντική για το εμπόριο των κρασιών και την εμπορική τους αξία, για αυτό το λόγο πρέπει να αποφύγουμε τη λανθασμένη πρόβλεψη ενός κρασιού ως χαμηλής ποιότητας.

Top 5 variables

Random Forest	Gradient Boosting Classifier
Flavanoids	Proline
Color_Sensity	Color_Sensity
Alcohol	Diluted wines
Proline	Flavanoids
Diluted wines	Hue

Random Forest Variables VS 4 essential



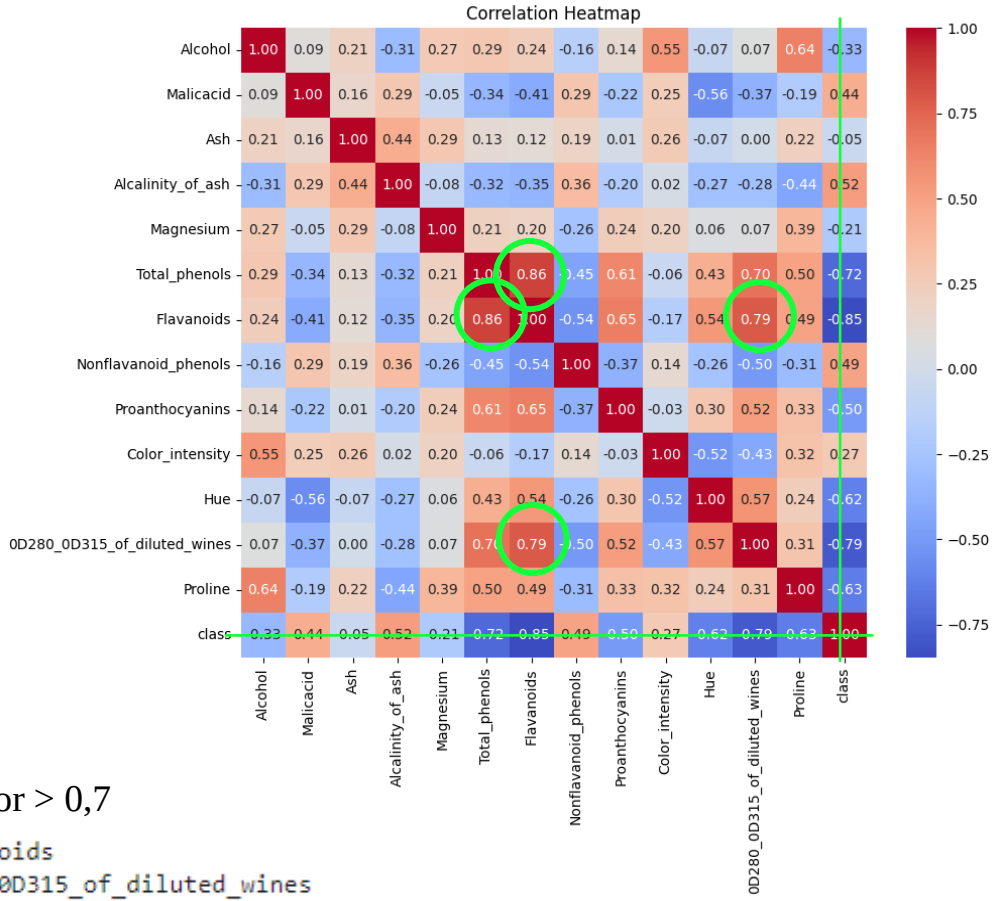
MODEL(5F)	ACCURACY	PRECISION			RECALL			F1-SCORE		
LR	1,00	1	2	3	1	2	3	1	2	3
		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
RF	0,9778	1	2	3	1	2	3	1	2	3
		1,00	0,95	1,00	1,00	1,00	0,92	1,00	0,97	0,96
KNN	0,96	1	2	3	1	2	3	1	2	3
		0,94	1,00	0,92	1,00	0,89	1,00	0,97	0,94	0,96
Gradient Boosting	1,00	1	2	3	1	2	3	1	2	3
		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

MODEL(4F)	ACCURACY	PRECISION			RECALL			F1-SCORE		
LR	1,00	1	2	3	1	2	3	1	2	3
		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
RF	0,9333	1	2	3	1	2	3	1	2	3
		0,88	0,94	1,00	1,00	0,89	0,92	0,94	0,91	0,96
KNN	0,93	1	2	3	1	2	3	1	2	3
		0,88	1,00	0,92	1,00	0,83	1,00	0,94	0,91	0,96

- Το LR και στις δύο περιπτώσεις έχει την απόλυτη απόδοση. Ωστόσο ο RF παρουσιάζει μια μείωση στο accuracy, όπως και το precision για την κατηγορία 1, ενώ το recall παρέμεινε στο ίδιο επίπεδο.

- Αυτό υποδηλώνει ότι με τον αλγόριθμο RF, ίσως έχουμε περισσότερα false positives για την κατηγορία 1 (προβλέψεις που δεν αντιστοιχούν σε υψηλή ποιότητα κρασιού), και όχι τόσα false negatives (κρασιά υψηλής ποιότητας που δεν προβλέπονται σωστά).

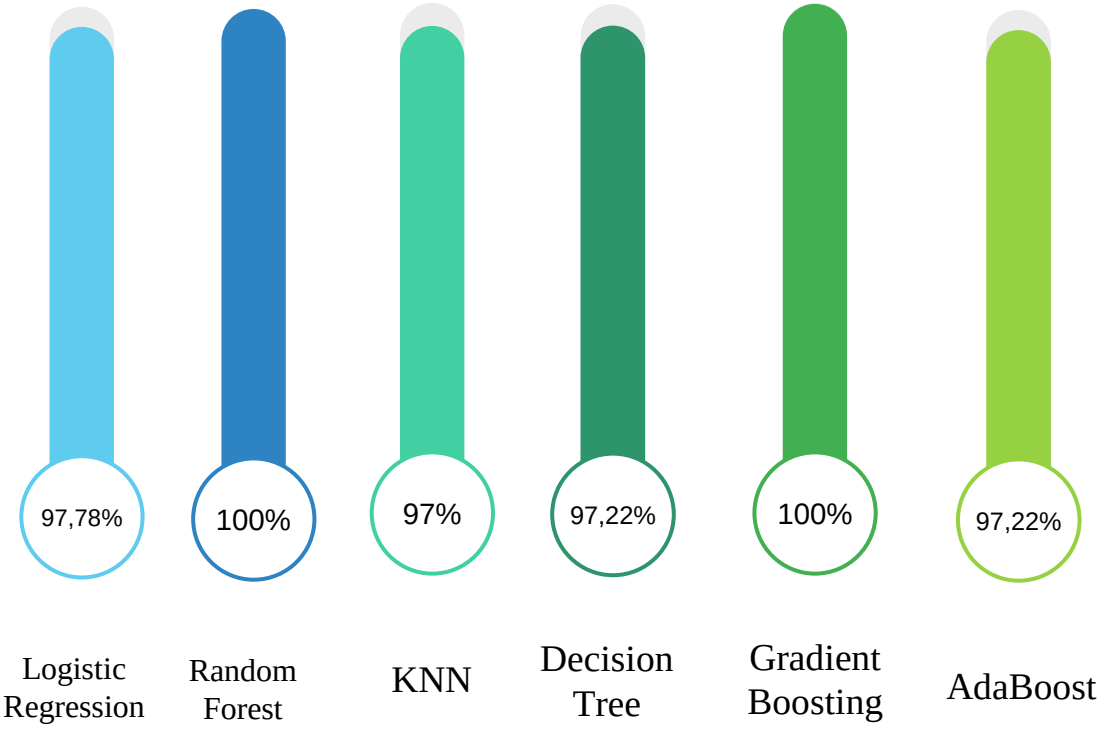
Correlation Analysis



For cor > 0,7

Flavanoids
0D280_0D315_of_diluted_wines

Machine Learning Models – Accuracy



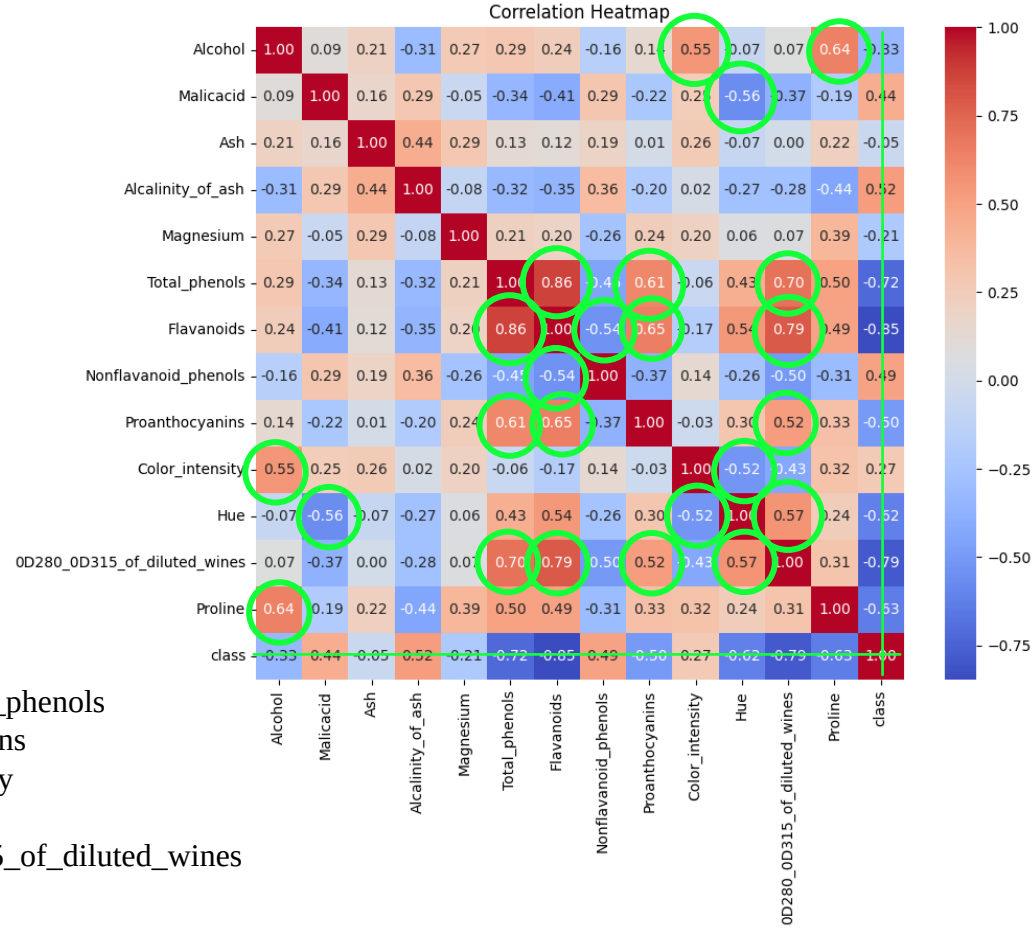
Drop columns: Flavanoids, 0D280_0D315_of_diluted_wines

Machine Learning Models – Accuracy

MODEL	ACCURACY	PRECISION			RECALL			F1-SCORE		
LR	0,9778	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
		0,94	1,00	1,00	1,00	0,94	1,00	0,97	0,97	1,00
RF	1,0	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
KNN	0,97	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
		1,00	1,00	0,89	1,00	0,93	1,00	1,00	0,96	0,94
Gradient Boosting	1,00	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
		1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Το γεγονός ότι το recall είναι 1.00 στο LR σημαίνει ότι το μοντέλο ανιχνεύει όλα τα πραγματικά θετικά παραδείγματα της κατηγορίας 1 (υψηλή ποιότητα κρασιού). Το 94% των προβλέψεων για υψηλή ποιότητα κρασιού είναι ακριβές. Παρατηρούμε μια μείωση στο precision σε σχέση με το πλήθος των μεταβλητών του dataset που χρησιμοποιούμε. Το μοντέλο που αναπτύχθηκε με όλες τις μεταβλητές είχε precision 1.00. Αυτό σημαίνει ότι ο LR με λιγότερες μεταβλητές χάνει κάποιες προβλέψεις υψηλής ποιότητας κρασιού.

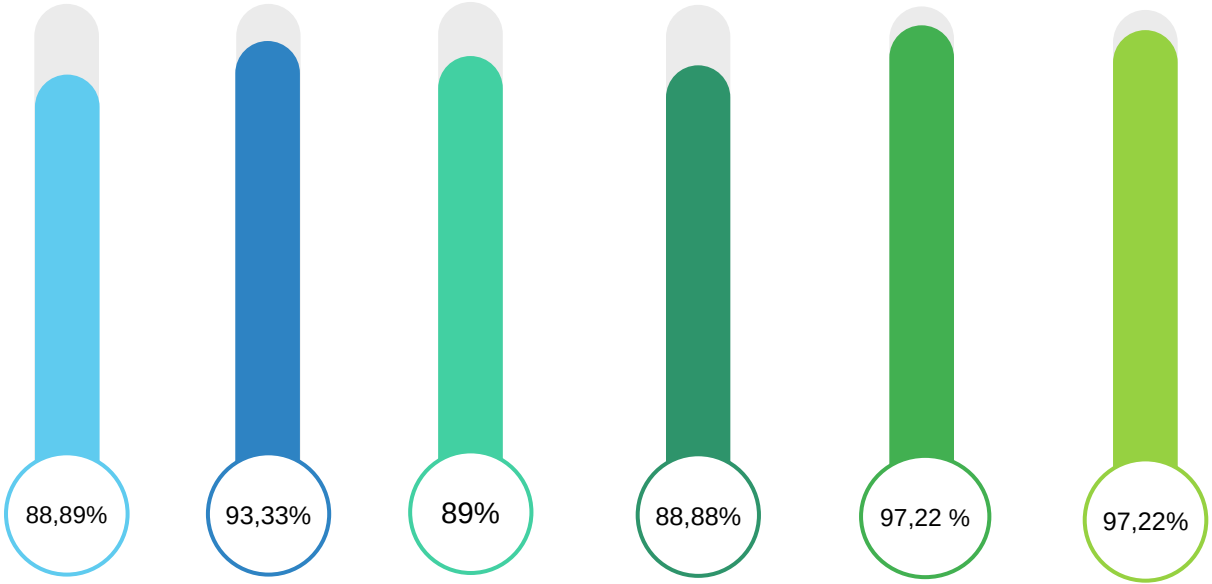
Correlation Analysis



For cor > 0,5

- Flavanoids
- Nonflavanoid_phenols
- Proanthocyanins
- Color_intensity
- Hue
- 0D280_0D315_of_diluted_wines
- Proline

Machine Learning Models – Accuracy



Logistic Regression | Random Forest | KNN | Decision Tree | Gradient Boosting | AdaBoost

Drop columns: Flavanoids, Nonflavanoid_phenols, Proanthocyanins, Color_intensity, Hue, 0D280_0D315_of_diluted_wines, Proline

Machine Learning Models – Accuracy

MODEL	ACCURACY	PRECISION			RECALL			F1-SCORE		
LR	0,889	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
		0,82	1,00	0,83	0,93	0,89	0,83	0,87	0,94	0,83
RF	0,9333	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
		0,83	1,00	1,00	1,00	0,94	0,83	0,91	0,97	0,91
KNN	0,89	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
		0,92	1,00	0,73	0,86	0,86	1,00	0,89	0,92	0,84

Ο RF με μεταβλητές corr > 0.7 είχε precision και accuracy 1.00. Με τη χρήση των μεταβλητών με corr > 0.5 μειώθηκε και το accuracy και το precison που σημαίνει ότι το μοντέλο έχει την τάση να προβλέπει λιγότερο υψηλής ποιότητας κρασιά, αλλά αυτά που προβλέπει είναι πιο πιθανό να είναι πράγματι υψηλής ποιότητας (1.00 recall).

Machine Learning Models – Accuracy

Algorithms	Df	Df (cor > 0.7)	Df (cor > 0.5)
Logistic Regression	0.9778	0.9778	0.890
Random Forest	1.00	1.00	0.933
KNN	0.94	0.97	0.89
Decision Trees	0.916	0.972	0.888
Gradient Boosting	1.00	1.00	0.972
AdaBoost	0.944	0.972	0.972

Neural Network – Accuracy

Df	Df (cor > 0.7)	Df (cor > 0.5)
0.3333	0.3333	0.2888

Thank you!