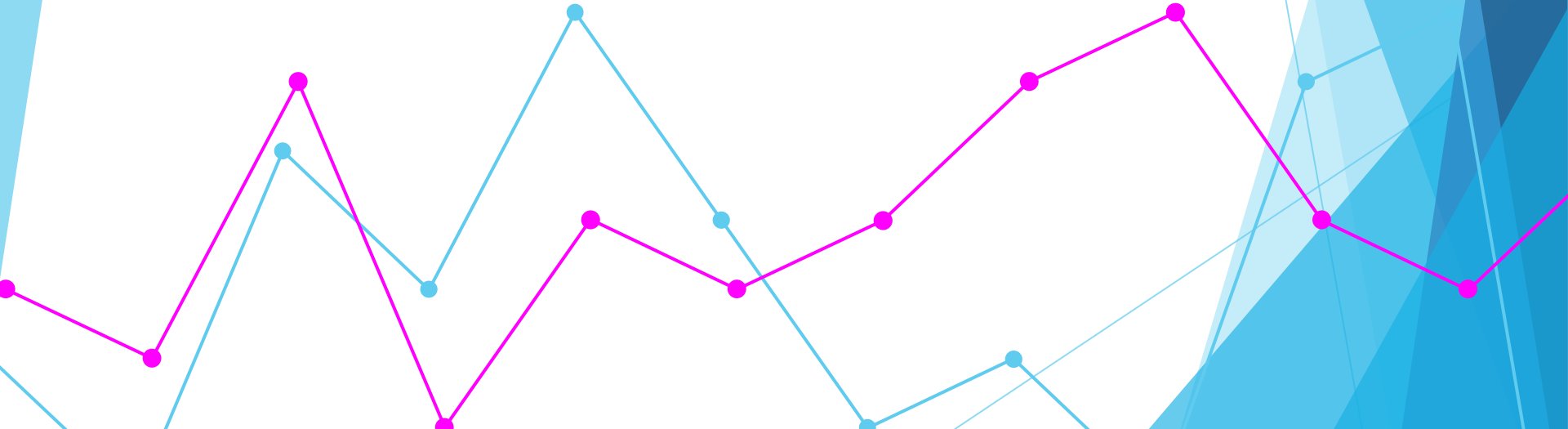


# Wine Data Set

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?**

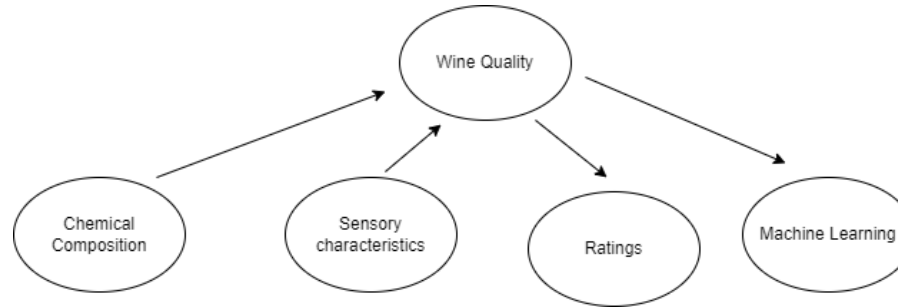
Experimental data is collected from various and varied regions throughout New Zealand that produce Pinot noir wine.

### ► **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:** Chromatography, 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*
  - $\lambda_{ij}$ : *factor loading coefficient*
  - $\varepsilon_{ij}$ : *random error*
  - $a$  *intercept*
  - $\eta$ : *latent variable*
- \* (*i equals to number of constructs, j the value of each construct*)

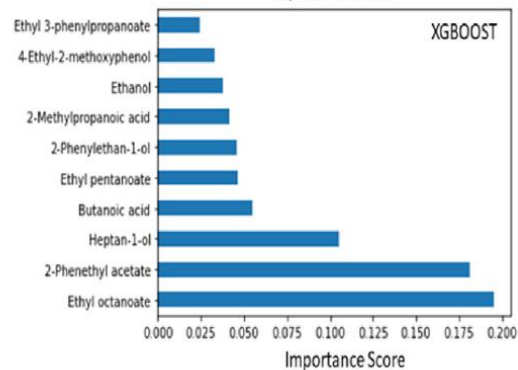
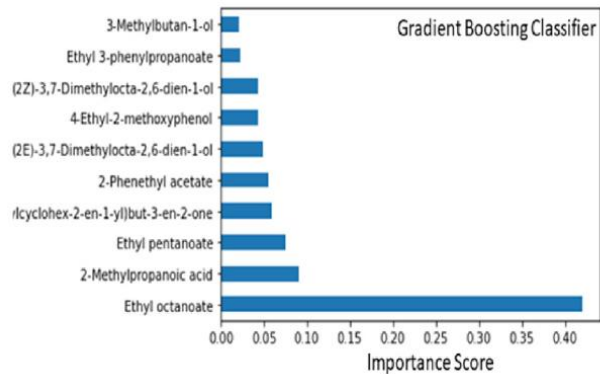
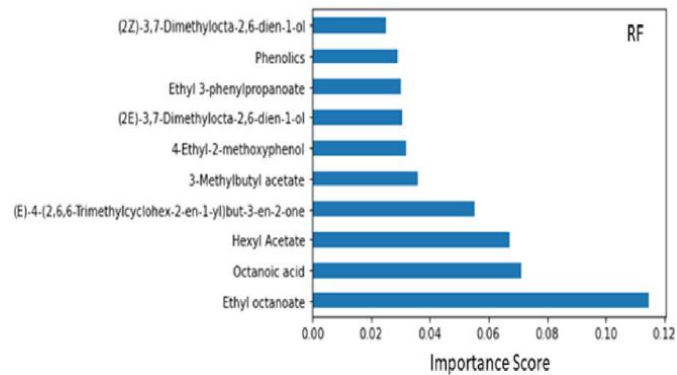
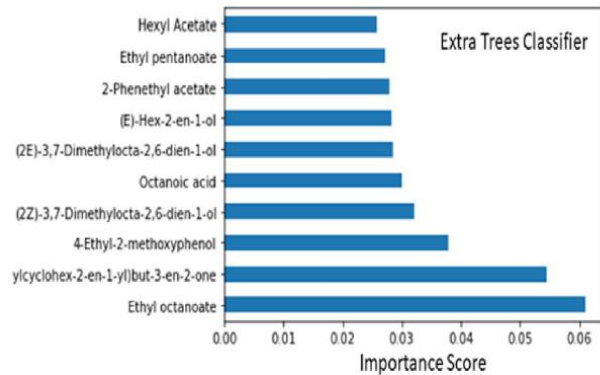


# Predictive Model

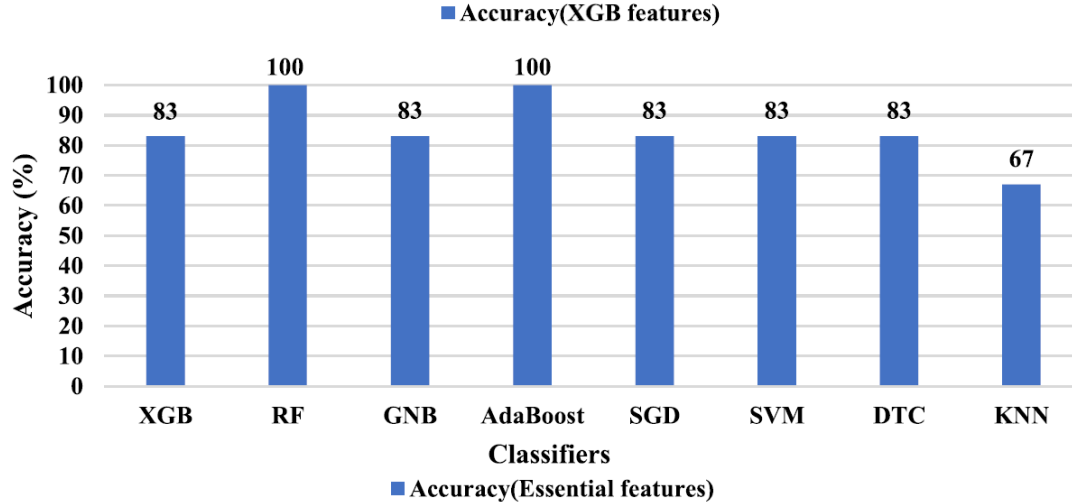
- ▶ Predictive modeling is a commonly used statistical technique to predict future behavior.
- ▶ Classification is a process related to categorization, the process in which ideas and objects are recognized, differentiated and understood.
- ▶ Classification is the grouping of related facts into classes.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



# Features selection



# Accuracy



# Variables in DataSet

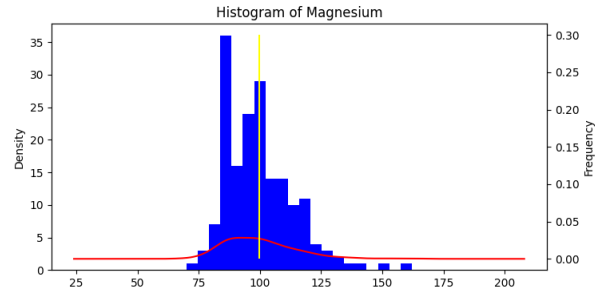
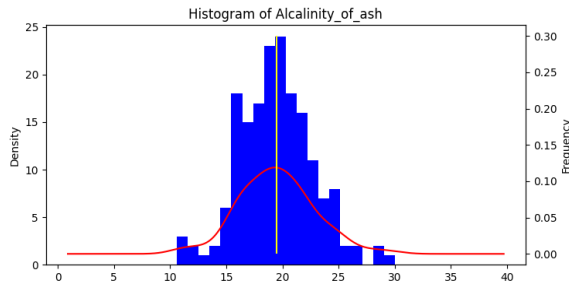
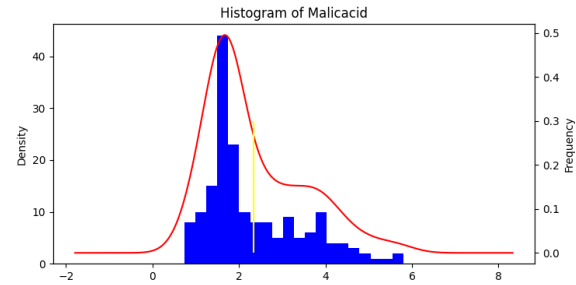
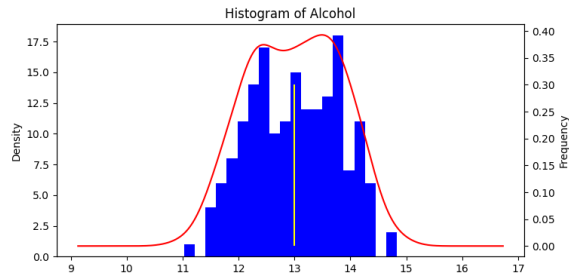
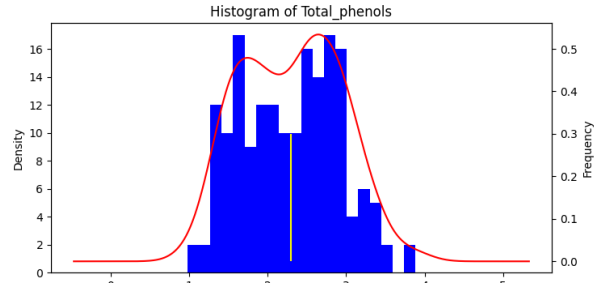
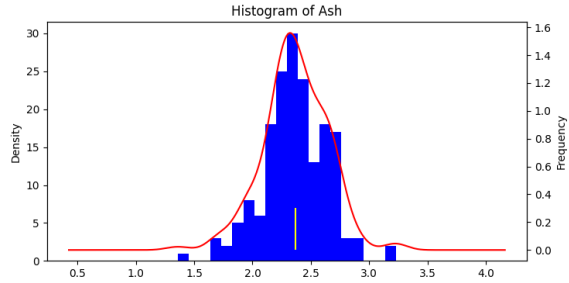
Name	Content	Type
<b>Class</b>	Κατηγορική (ποιοτική)	Τάξη
<b>Alcohol</b>	Συνεχής (ποσοτική)	Αλκοόλ
<b>Malicacid</b>	Συνεχής (ποσοτική)	Μηλοξίνη
<b>Ash</b>	Συνεχής (ποσοτική)	Στάχτη
<b>Alcality of Ash</b>	Συνεχής (ποσοτική)	Αλκαλικότητα Στάχτης
<b>Magnesium</b>	ποσοτική	Μαγνήσιο
<b>Total phenols</b>	Συνεχής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

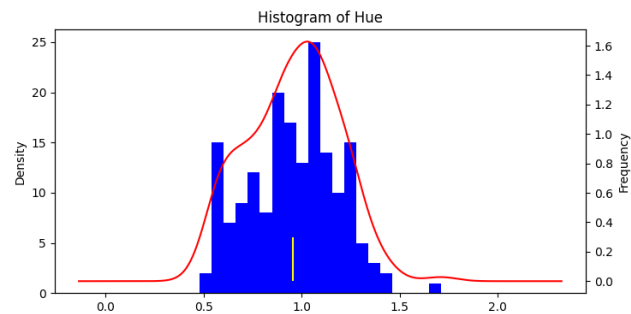
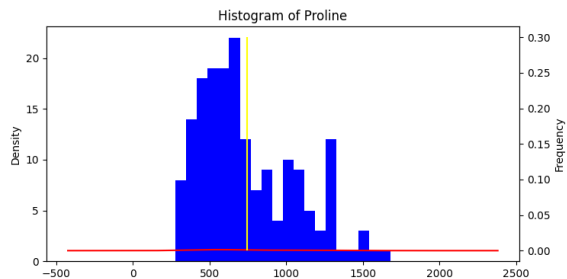
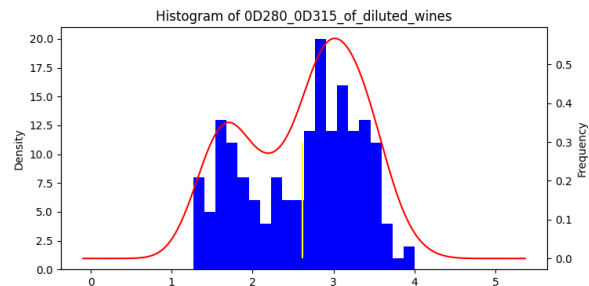
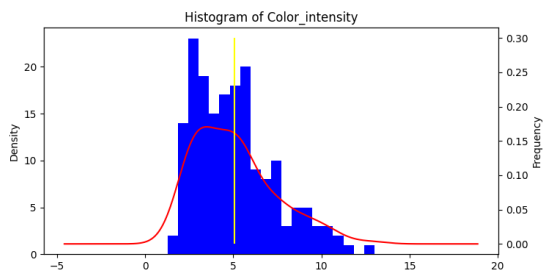
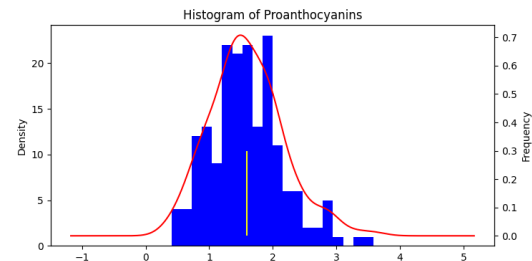
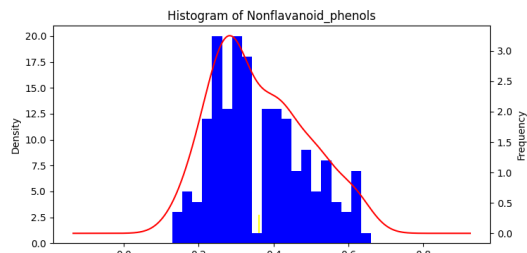
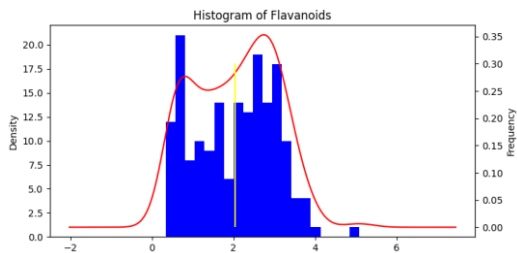


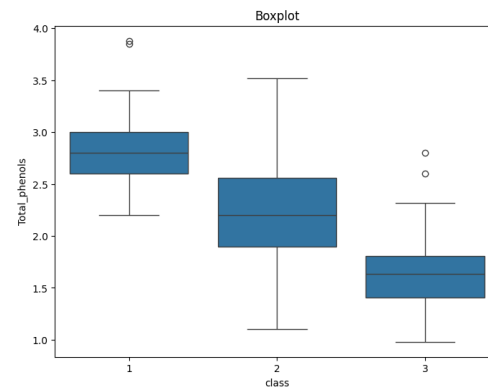
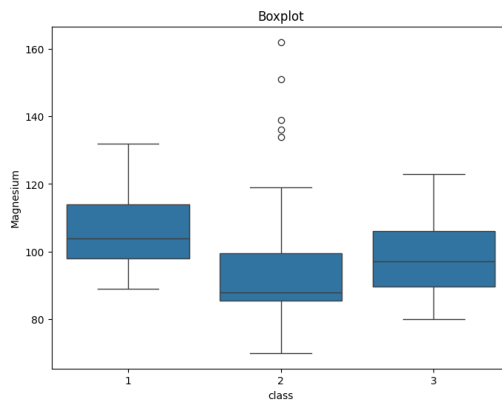
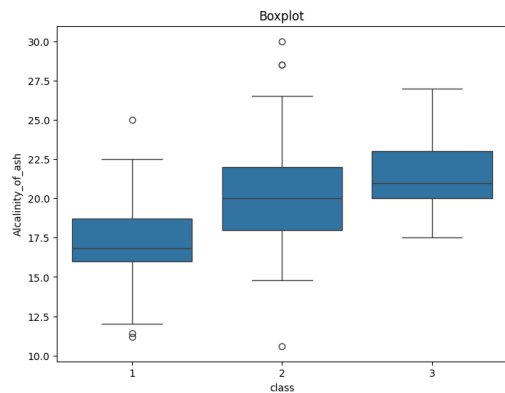
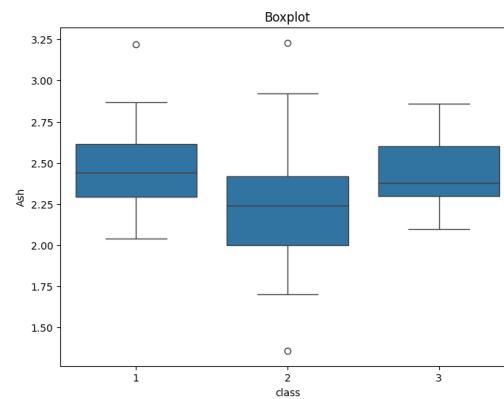
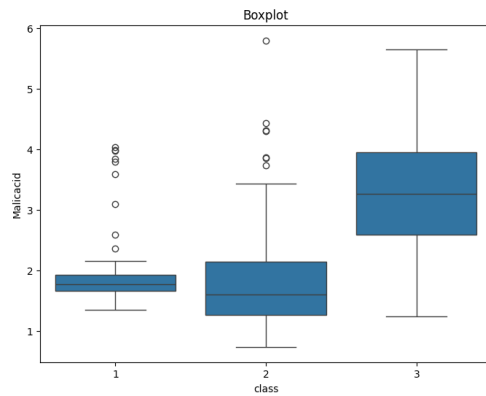
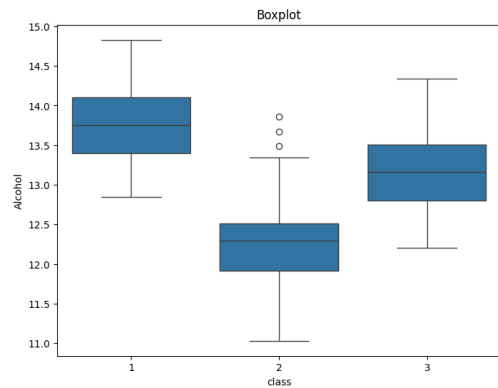
# Descriptive Statistics

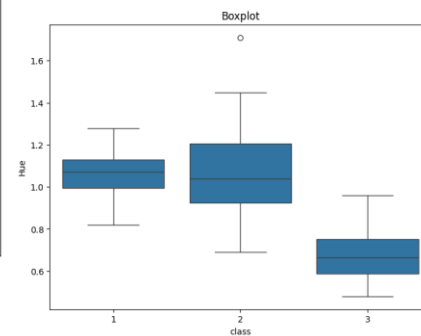
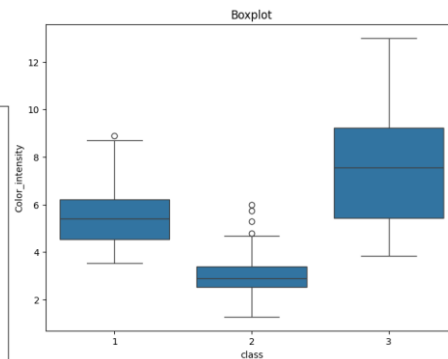
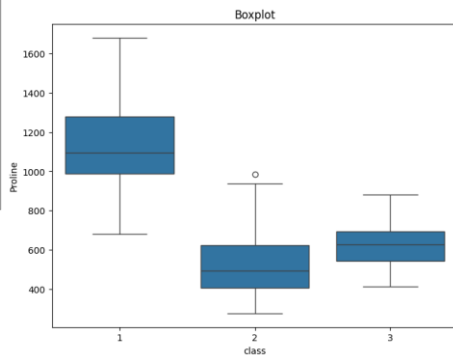
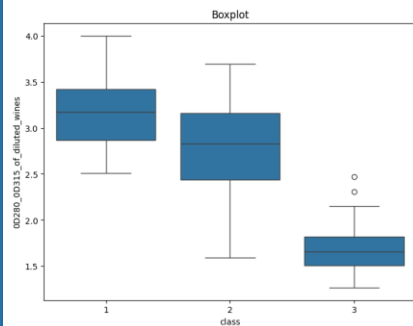
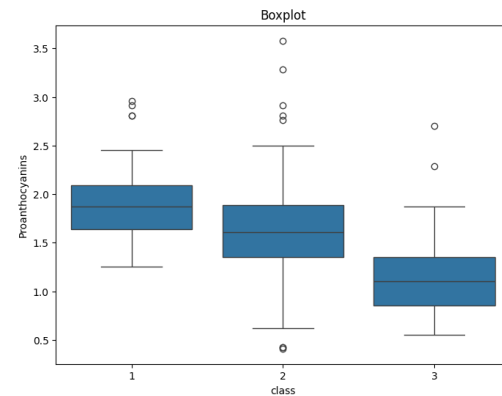
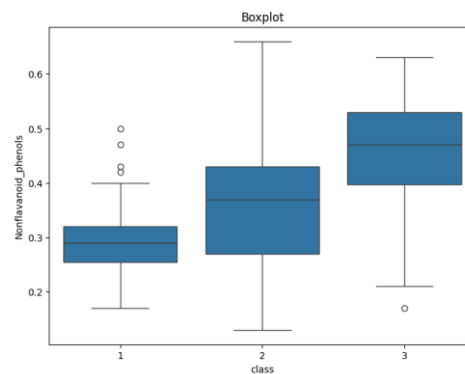
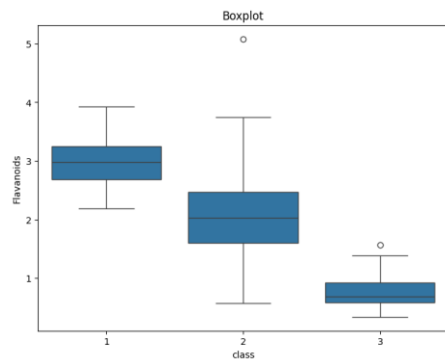
Variable	count	mean	std	min	25%	50%	75%	max
Class	178,00	1,94	0,77	1,00	1,00	2,00	3,00	3,00
Alcohol	178,00	13,00	0,81	11,03	12,36	13,05	13,67	14,83
Malicacid	178,00	2,36	1,11	0,74	1,60	1,85	3,08	5,8
Ash	178,00	2,36	0,27	1,36	2,21	2,36	2,55	3,23
Alcality of Ash	178,00	19,49	3,33	10,6	17,2	19,5	21,5	30,00
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
Proanthocyanis	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



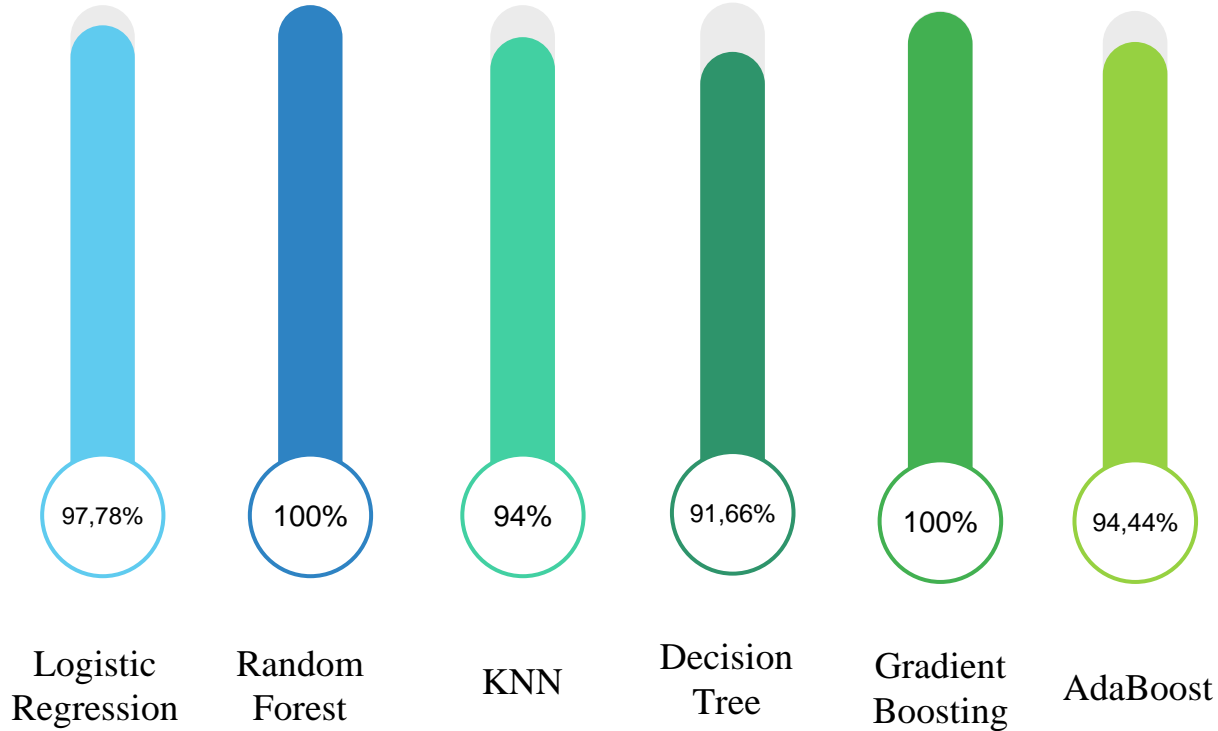








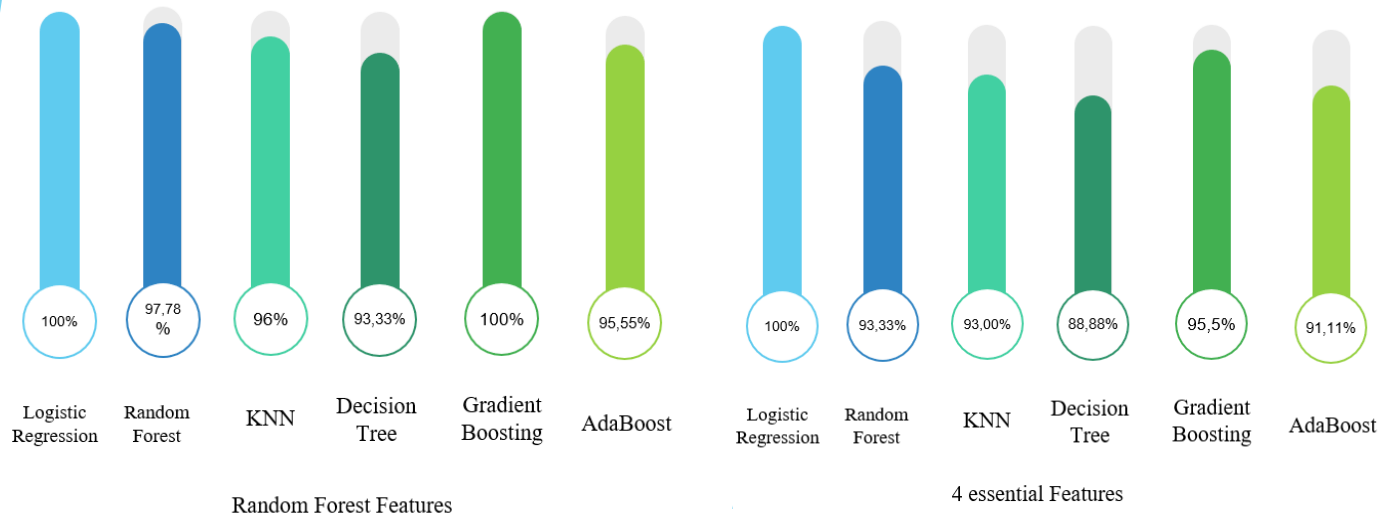
# Machine Learning Model – all features



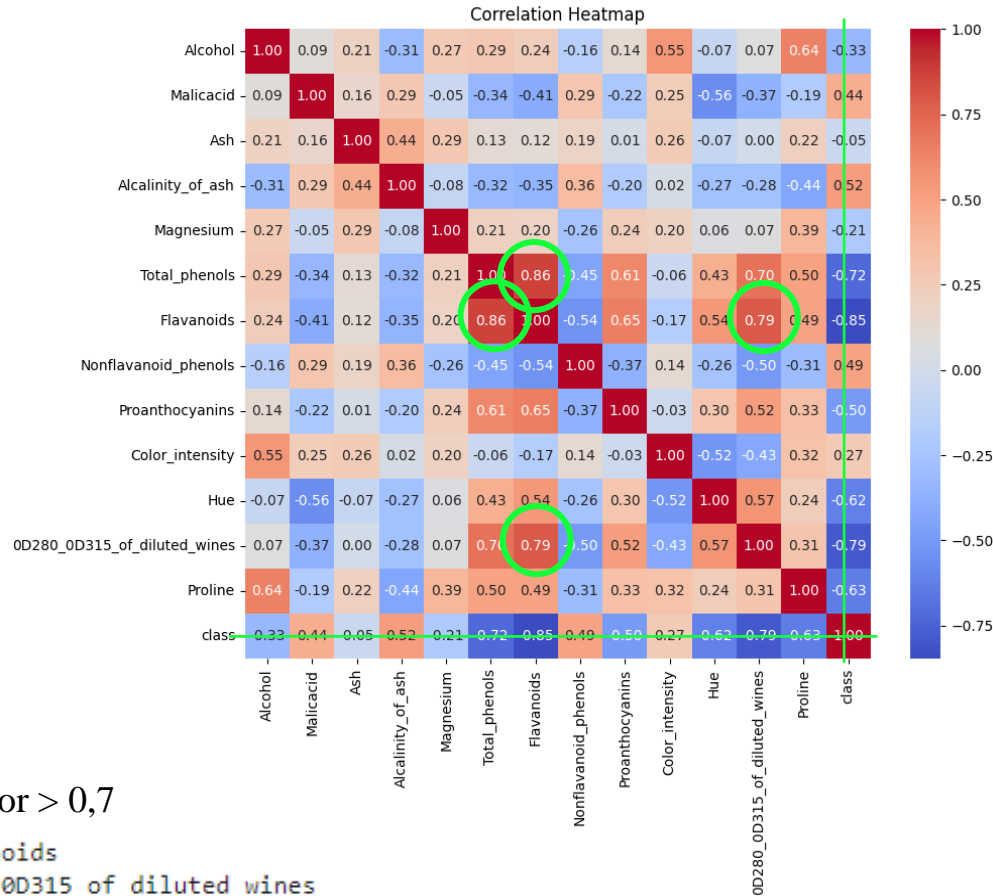
## Top 5 variables

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

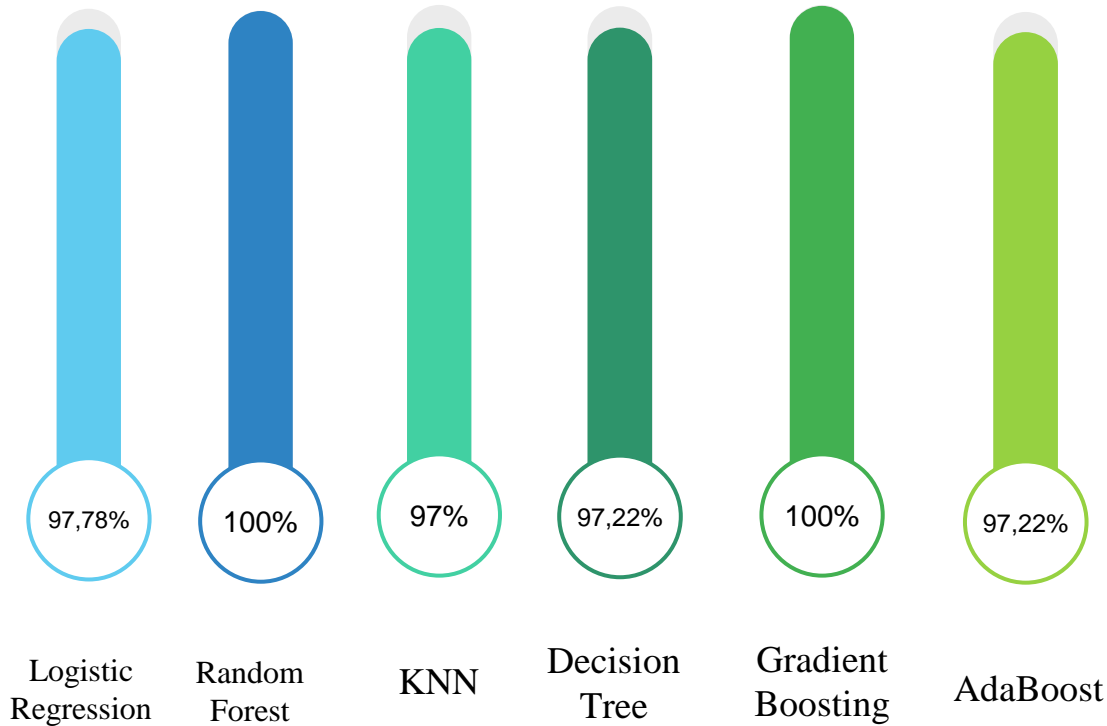


# Correlation Analysis





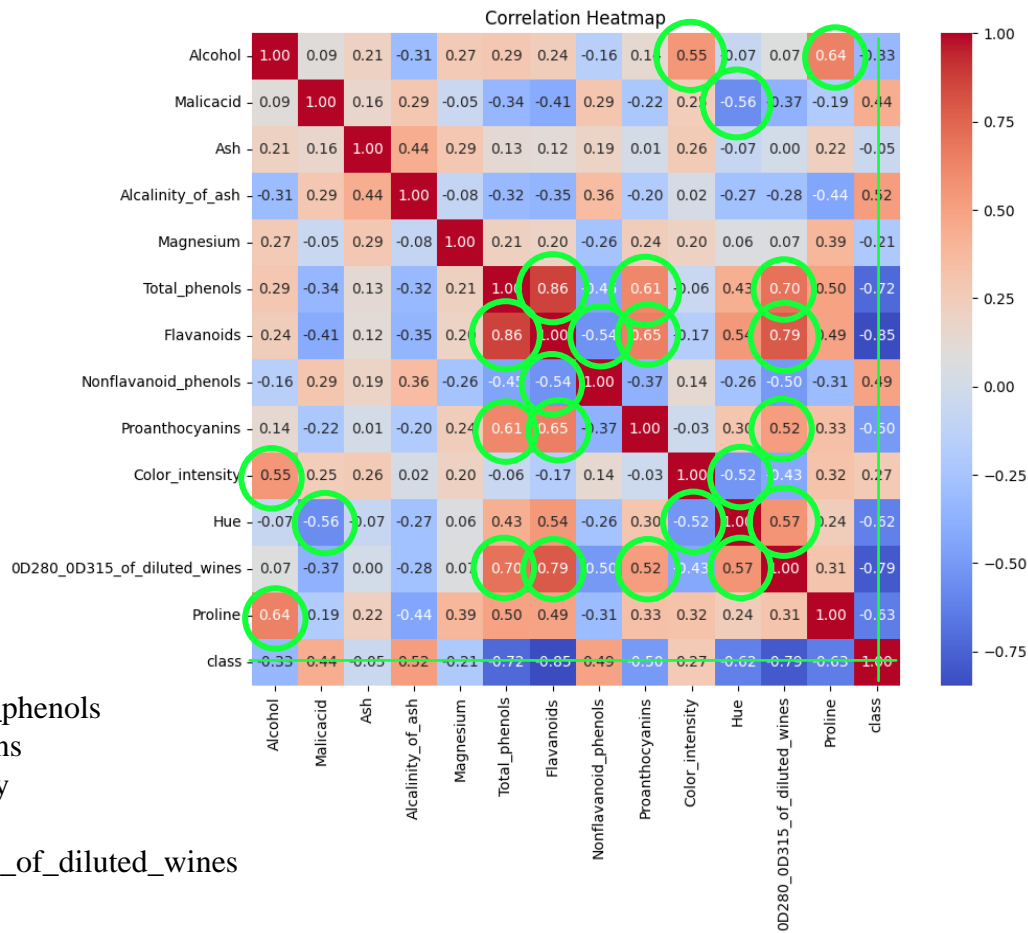
# Machine Learning Models – Accuracy



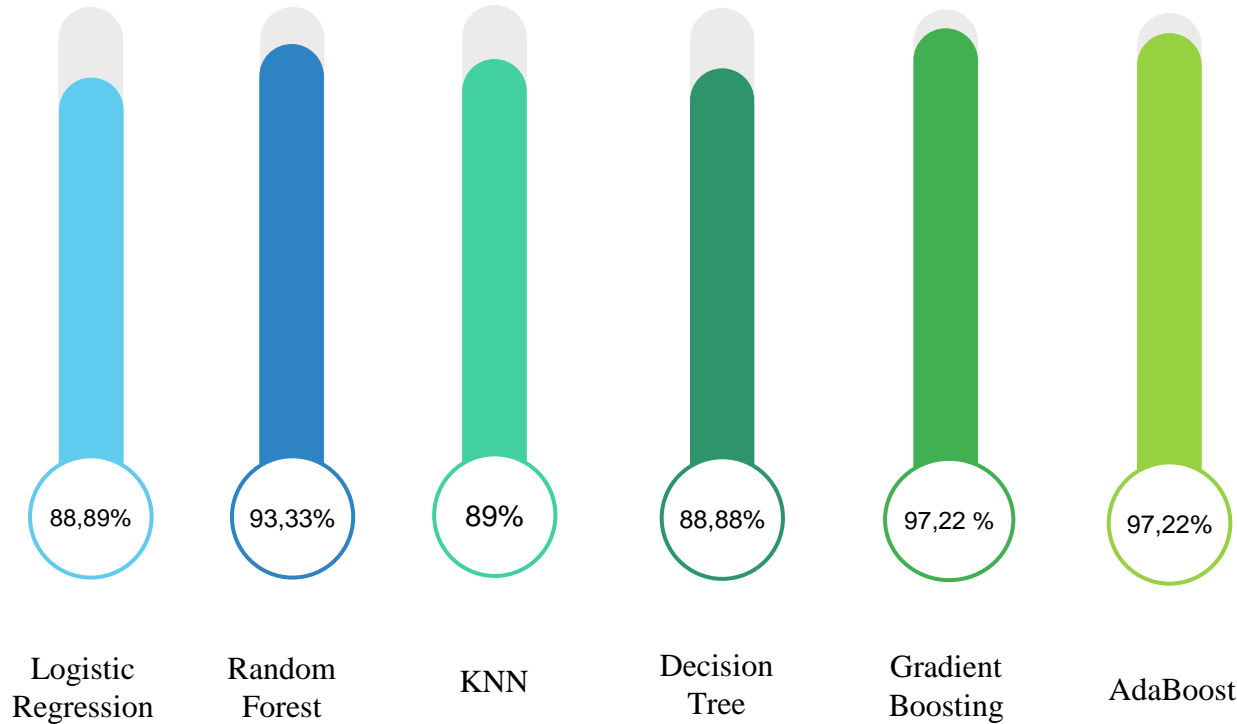
Drop columns: Flavanoids, 0D280\_0D315\_of\_diluted\_wines



# Correlation Analysis



# Machine Learning Models – Accuracy



Drop columns: Flavanoids, Nonflavanoid\_phenols, Proanthocyanins,  
Color\_intensity, Hue, OD280\_OD315\_of\_diluted\_wines, Proline



# 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

<b>Df</b>	<b>Df ( cor &gt; 0.7)</b>	<b>Df (cor &gt; 0.5)</b>
0.3333	0.3333	0.2888



The background features abstract, overlapping geometric shapes in various shades of blue, ranging from light sky blue to deep navy blue. These shapes are primarily located on the right side of the slide, creating a modern, layered effect.

# Thank you!