**I declare that this report represents my own work. It has not been submitted by others or in another form for an assessment. Sources used have been cited and acknowledged.**

# Quality Analysis of Different Olive Oils: Preprocessing and preparing for classification using the fluorescence and chemical data of olive oils.

**Jason Wang**

**S4134626**

**S4134626@student.rmit.edu.au**

**MC267 Master of Data Science**

**RMIT University**

**17/05/2025**

# Table of Contents

# Abstract

This report aims to prepare a dataset for classifying olive oil into 3 different categories, these being Extra Virgin, Virgin and Lampante. The dataset consists of 24 different olive oil samples, each being tested 20 times using 2 different types of ultraviolet light. It also includes chemical measurements like acidity and peroxide value, as well as fluorescence spectra. The goal is to prepare the dataset for classification modelling that will help predict the 'Quality' label for unseen data. The proposed modelling is classification, more specifically, k-nearest neighbours (KNN) and Decision Tree (DT). Grid search was used to find optimal parameters. When using chemical properties, KNN produced the best score using parameters k = 6, uniform, p=2 compared to DT's max_depth = 2. KNN achieved a mean accuracy of 96% whereas DT reached a mean accuracy of 92%. When using spectra, KNN accuracy was 85%, DT accuracy was 65%.

KNN using chemical properties is the preferred classification model regarding future olive oil label testings.


# Introduction

Olive oil is a health food because of its monosaturated fats and antioxidants. Assessing the quality of olive oil is crucial both commercially and health wise. The traditional method of assessing olive oil is often time consuming and expensive, such as chemical lab analysis. However, new and emerging techniques analyses olive oils in a more automated way using machine learning through fluorescence spectroscopy.

One study describes fluorescence spectroscopy, in conjunction with 1D convolutional neural networks, an "approach [that] gives exceptional results for quality determination through the extraction of the relevant physicochemical parameters." (Venturini et al., 2023). This is useful because indicators like acidity, peroxide value and UV absorbance are important in classifying olive oils into Extra virgin, virgin or lampante.

Another study by Bavali et al. (2025) showcases that this technique requires minimal sample preparation, as well as not destructive to the olive oil. This is important because samples are destroyed when they are chemically altered through traditional chemical testing, and typically a sample should not be destroyed so that it can be reused for testings.

In this project, a dataset with fluorescence spectra of 24 olive oils and their chemical parameters are explored.

# Goal

The goal of this project is to prepare and explore the dataset which contains fluorescence spectra and chemical properties of olive oil samples. The aim is to build a machine learning model, more specifically, a classification model. The model will experiment using chemical features, and spectra. 2 different methods will be used, kNN and Decision Tree. This model will help predict the 'Quality' label for future unseen data.

# Methodology

Libraries used to analyse the datasets were pandas, numpy, seaborn, matplotlib and a variety of sklearn libraries. Dark counts correspond to the intensity of the spectra, measured by the spectrometer without any light. This is subtracted from the primary dataset. In essence:  True fluorescence = data – dark counts.

The datasets were firstly inspected such as understanding counts, column and their types and missing values. The 10 columns were Samples, Repetition, Led, Data, Quality, FAEES, K232, K270, Acidity and Peroxide Index. Within the 240 samples of lampante, 80 had no chemical properties, the remaining having either acidity or both. To handle these missing values, linear regression was used. The lampante samples that had both acidity and peroxide index values were a good basis on how to predict the rows that had a missing peroxide index but

had an acidity value. Afterwards, chemical features of FAEES and K270 were predicted using the predictors of acidity and peroxide index. K232 was calculated using the means of EXTRA and VIRGIN due to EXTRA having predominantly higher values, even though suggested research from Venturini, Sperti, et al. (2023) states that K232 should be lower with better quality.

The goal is to classify unseen data into a label. The label chosen is quality, hoping to use either the chemical properties or spectra to determine whether a certain olive sample is Extra, Virgin or Lampante. The dataset was split into 80/20 training to testing. The exploration of parameters and model comparison were done with a 5-fold stratified K-fold. This is crucial because the dataset has imbalanced amount of classes, with less lampante but more extra and virgin, so this ensures that there is an even amount. A consistent random state was used for reproducibility.

To visualise data, a variety of scatterplots, histograms and boxplots using seaborn and matplotlib were used.

The chemical properties that were now populated, as well as spectra, were all used for feature engineering. The columns were used. For spectra, new features were engineered, mean and max intensity, standard deviation, peak wavelength and area under curve.

2 algorithms were selected, KNN and DT. The KNN parameter used was k=6, weight = uniform, p=2. This was discovered through the use of gridsearch. Scikit Learn (2012) states that gridsearch exhaustively considers all parameter combinations. By giving it parameters, it finds the best combination of parameters to use. DT was found through the loop, where max_depth beyond 2 was not increasing the score.

# Results

**Samples**
By using the method unique (unique()), the Samples column returned 24 unique samples. Value counts returned all the samples having 40 each. This is concurrent with the 960 rows.

**Repetition**
Like samples, the unique samples is found. Repetition is the ‘n’th repetition. Since there are 20 repetitions. Each repetition from number 1 to 20 will occur 48 times. 24 for each sample, using 2 different UV LED’s.

**Led**
Value counts (value_counts()) show 480 for each sample, which is concurrent with the dataset. 2 LED’s, 960 in total.

**Quality**
By using value counts, it shows that extra has 400 rows, virgin has 320 and lampante has 240. This means that there are more observations of the higher grade olive oils. This is why 5-fold stratified K-fold was later used, since there is an imbalance in observations.

**Data**

Means were calculated by stacking the arrays. By averaging all the indexes of the same quality oil, a mean was reached for the 3 different qualities. overall mean was also calculated. What can be gathered from this that from the higher qualities, there is a small peak at index 100, which is absent in lampante. The large peak is at index 650. Extra reaches an intensity of 10,000 whereas virgin is just short of it. Lampante reaches 4000.

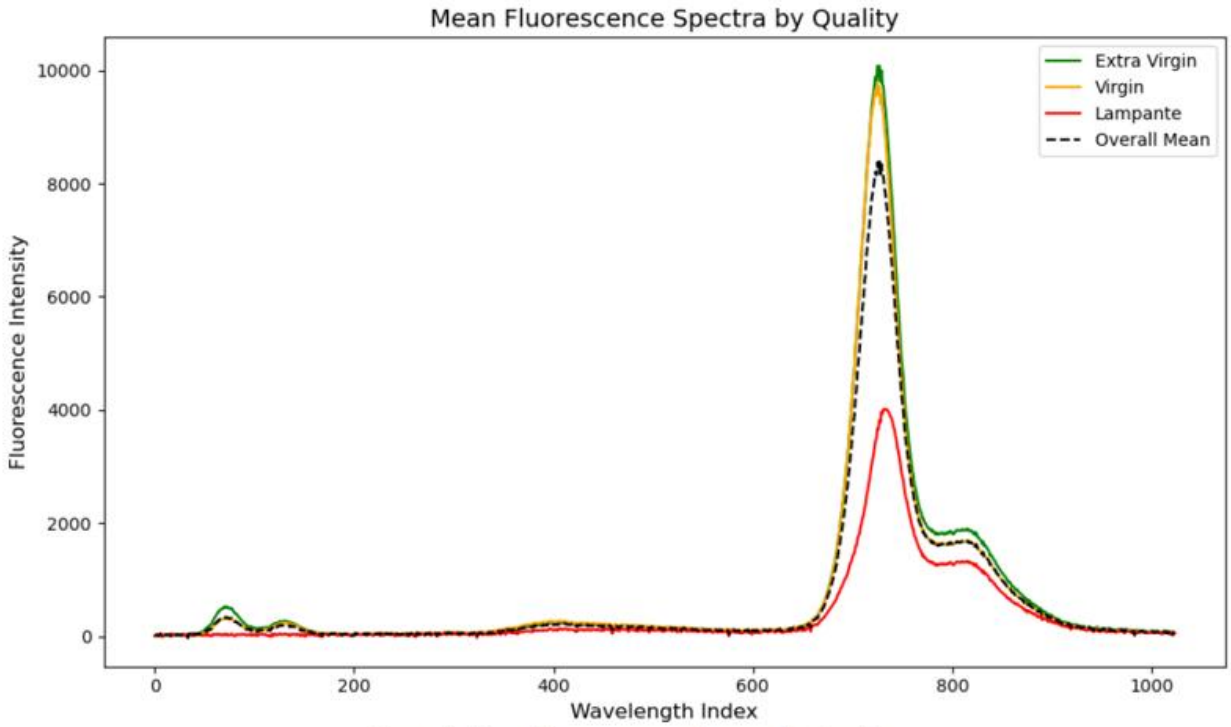


*Figure 1. Mean Fluorescence Spectra by Quality*

**FAEES**

FAEES is fatty acid ethyl esters. The increased amount of FAEES in olive oil, the lower the quality grade. It is expected that extra and virgin have lower quantities compared to lampante.

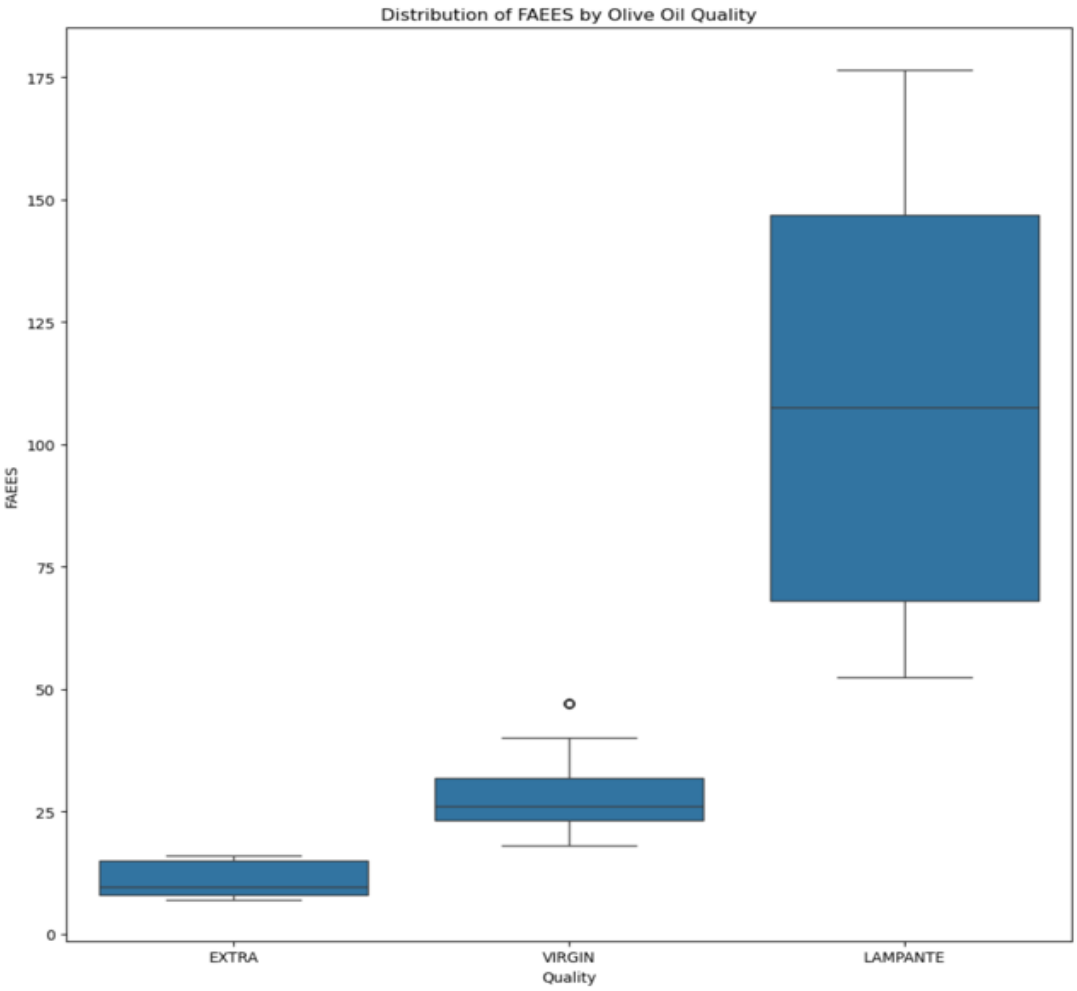


*Figure 2. Distribution of FAEES by Quality*

**K232**

María Isabel Sánchez-Rodríguez et al. (2023) suggest that lower K232 values is synonymous with a better grade olive oil. A higher K232 means the oil absorbed more light at 232nm. This shows more oxidation in the oil. This is similar with K270. It is to be noted that all the lampante values for K270 and FAEES were imputed using linear regression. However, K232 was engineered through the median of extra and virgin, then inputted into lampante. This was 1.63. This is because when inspecting the data, the values giving 1.40 were virgin, and 1.91 showing extra. This is against the hypothesis. If linear regression was to be used, it would suggest that lampante would have K232 values between 0.5-1.0, which is factually incorrect.

| K232 | |
|---|---|
| 1.63 | 480 |
| 1.54 | 120 |
| 1.44 | 80 |
| 1.40 | 40 |
| 1.48 | 40 |
| 1.55 | 40 |
| 1.64 | 40 |
| 1.66 | 40 |
| 1.74 | 40 |
| 1.91 | 40 |

*Figure 3. K232 values*

**K270**

Like K232, a lower value indicates a better quality. However, there was no discrepancy between the hypothesised values of extra and virgin. The lower numbers primarily corresponded to extra. Middle numbers were mainly virgin. Linear regression computed the missing values for lampante based off this.

| K270 | |
|---|---|
| 0.12 | 200 |
| 0.13 | 200 |
| 0.11 | 80 |
| 0.15 | 80 |
| 0.16 | 80 |
| 0.18 | 80 |
| 0.14 | 80 |
| 0.17 | 80 |
| 0.21 | 40 |
| 0.19 | 40 |

*Figure 4. K270 values*

**Acidity**

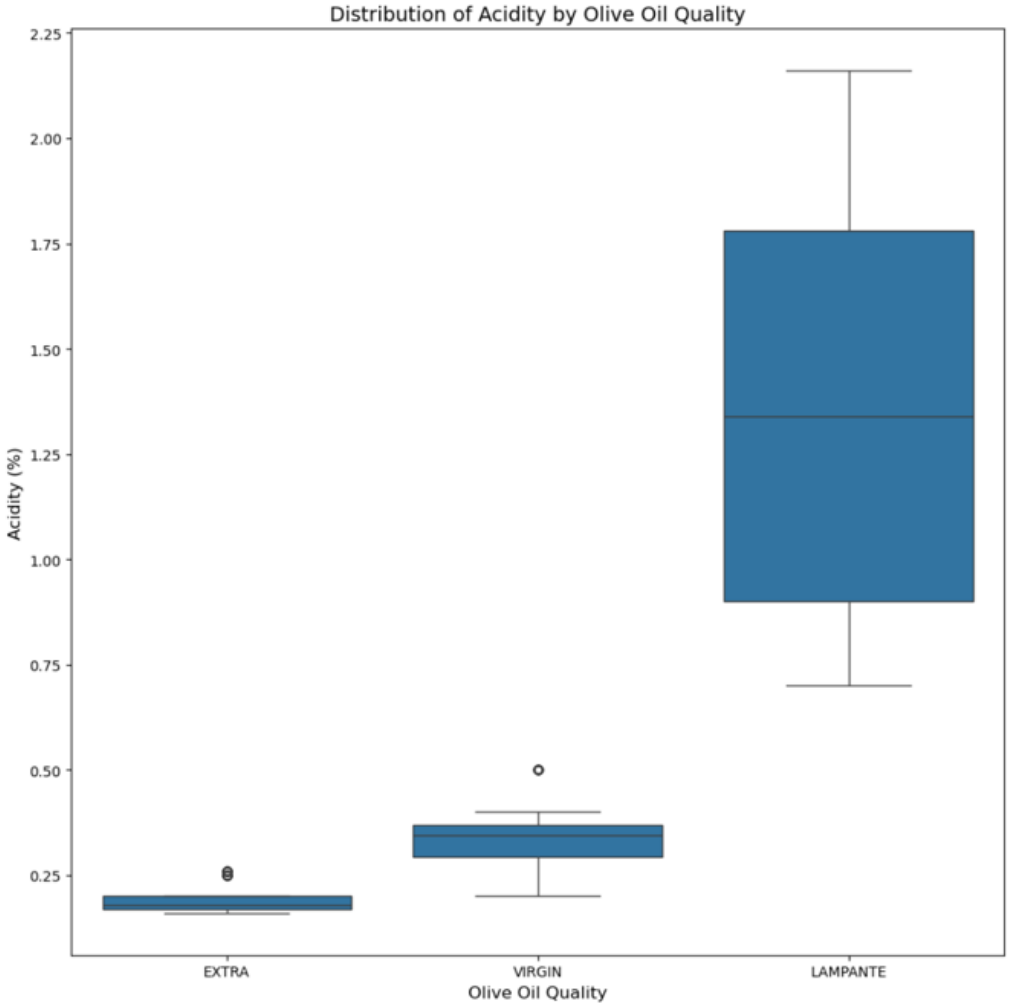


*Figure 5. Acidity by Olive Oil Quality*

Since most values of acidity were present, an appropriate graph can be displayed. The higher the acidity, the worse quality.

**Peroxide Index**

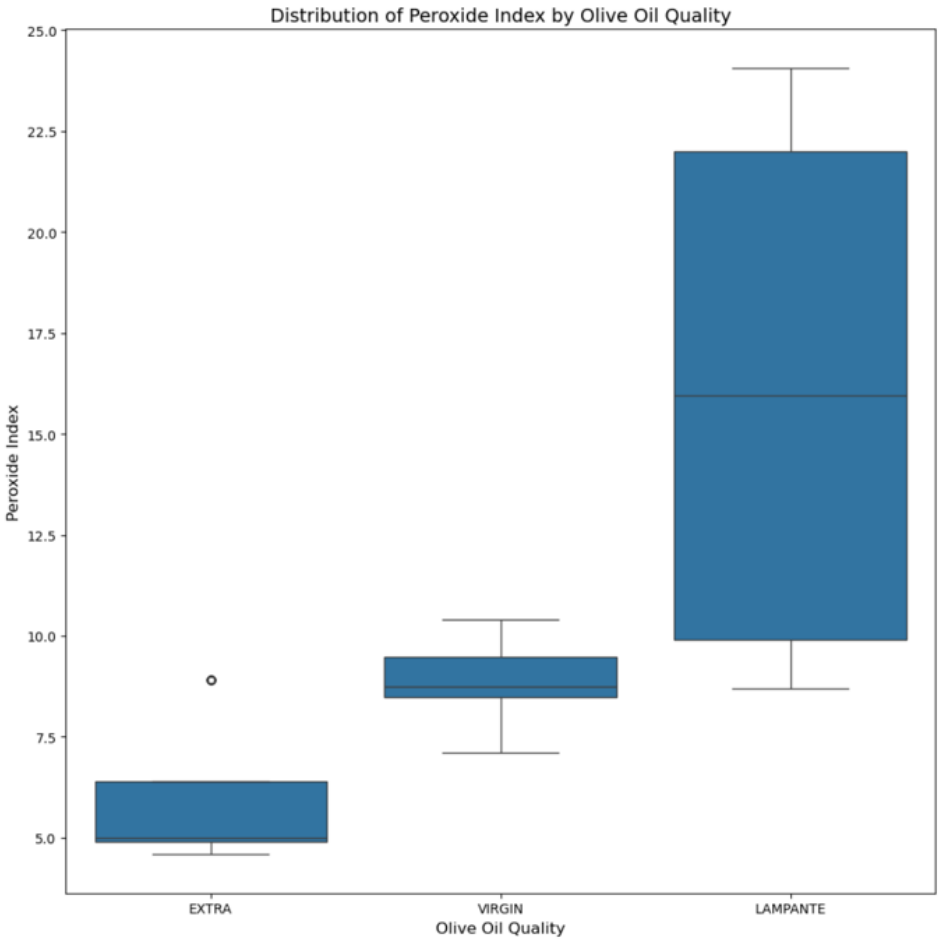


*Figure 6. Peroxide Index by Olive Oil Quality*

As peroxide index increases, the lower the quality of olive oil.

## Pairs Comparison

### FAEES vs Quality

As olive oil degrades, the amount of FAEES accumulate. This is because FAEES are the breakdown products of triglycerides reacting with ethanol (Venturini, Michela Sperti, et al., 2023). This usually signifies poor handling or storage. This suggests that lampante will have the highest amount of FAEES, extra virgin with the least, and virgin in between. The boxplot shows a clear increase in FAEES between the quality levels, supporting the hypothesis.

### Acidity vs Quality

The acidity increases as the quality worsens. This is because acidity relates to triglyceride hydrolysis. Thus, FAEES and acidity have a positive correlation, because they are a product of one another. It is expected that as quality drops, acidity rises. The boxplot confirms the trend, where lampante have higher acidity whereas extra shows the lowest.

### Peroxide vs Quality

The peroxide index will increase as quality drops. This is because peroxide formation is a sign of oxidation of the unsaturated fatty acids (Venturini, Michela Sperti, et al., 2023). Oxidation is from poor handling and storage. Thus, premium oils have low peroxide values because of strong quality control. Boxplot supports the hypothesis, where lampante has the highest peroxide values, and extra the least.

### Mean Fluorescence vs Quality

Strong fluorescence bands in extra virgin olive oil are related to carotenoids and pheophytins. These both decline as olive oil deteriorate (Venturini et al., 2024). Thus, the greater the quality oil, the more carotenoids and pheophytins, leading to greater fluorescence.

The boxplot depict that Extra has a max and min, as well as a greater median, and less spread compared to virgin. This is expected. However, virgin is quite close to extra, many values are the same or close to extra. This shows that the difference is better control in handling, because the spread is greater in virgin. Lampante has much lower values, and a much larger spread.

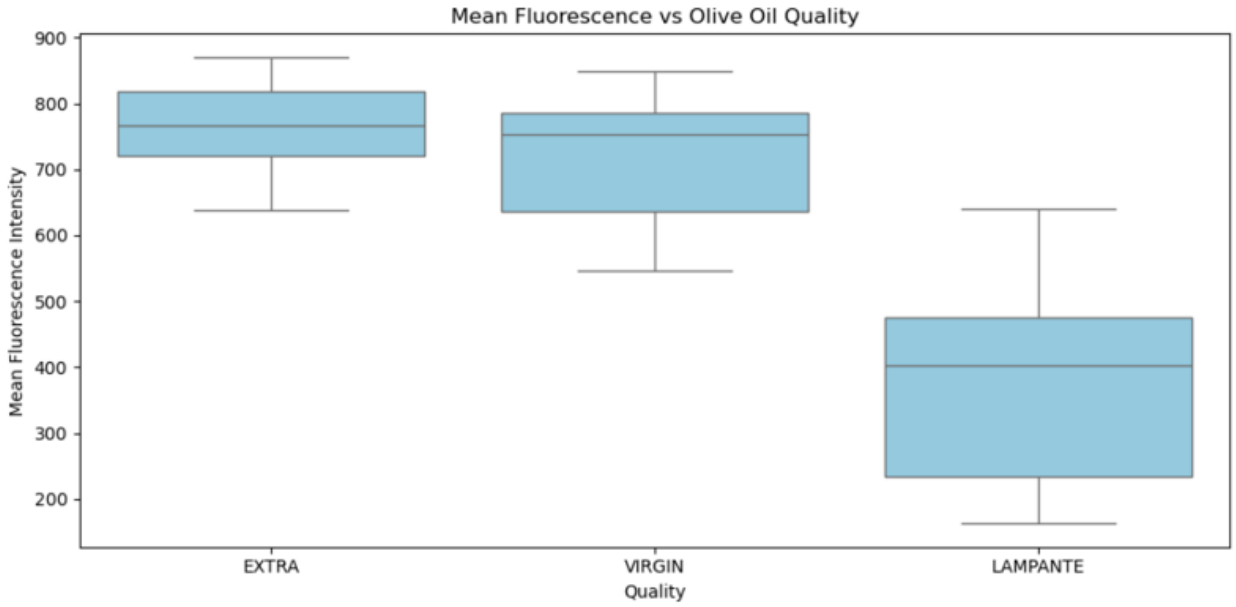


*Figure7. Mean Fluorescence by Olive Oil Quality*

**K232 vs K270**

There is a positive correlation between both, as they are similar measures. As an oil oxidises, peroxides and trienes accumulate (Chen et al., 2022). These absorb the 232 and 270 wavelength respectively. So, the more of these properties, the higher absorption, leading to a greater number, signifying worse quality. The scatterplot shows a positive trend, proving that these two metrics move with oxidation.

**FAEES vs Acidity**

These are positively correlated because both show the degradation of olive oils through hydrolysis. This is because acidity relates to triglyceride hydrolysis. FAEES and acidity are both increased when more hydrolysis occurs. In figure 8, there is a positive correlation, where increased FAEES equates to increased acidity.

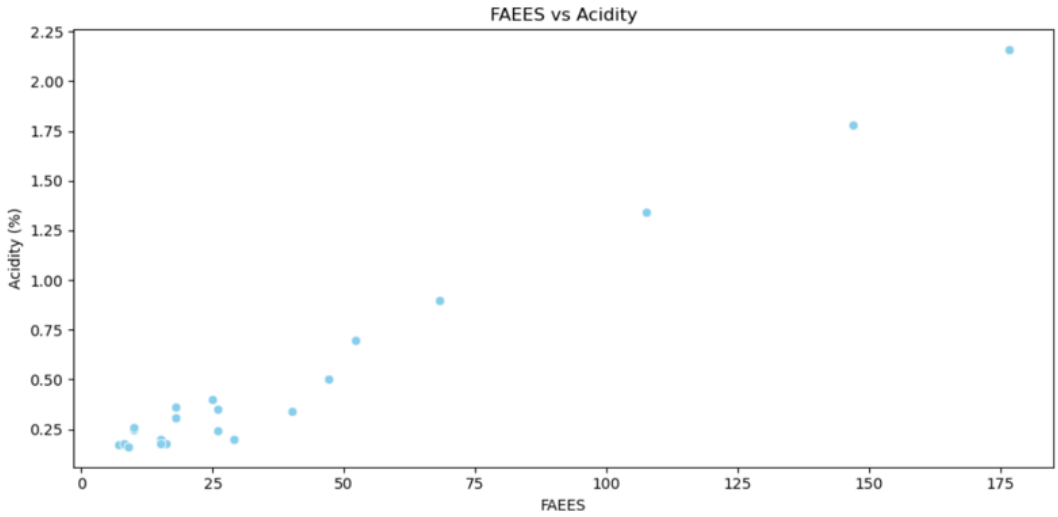


*Figure 8. FAEES against Acidity*

**K232 vs FAEES**

Higher primary oxidation released higher FAEES concentration. This is a positive correlation. This is because oxidation leads to fatty acids breaking down into dienes (Chen et al., 2022). This absorbs UV at wavelength 232nm. The scatterplot shows a mild but positive trend. However, variability is large.

### K270 vs Acidity

Further oxidation in oils increase in K270 readings and acidity. A positive correlation. This is because when peroxide breaks down, carbonyl compounds are formed. This absorbs UV at wavelength 270m.
Figure 9 shows that as acidity increases, K270 does moderately increase, showing a positive trend.

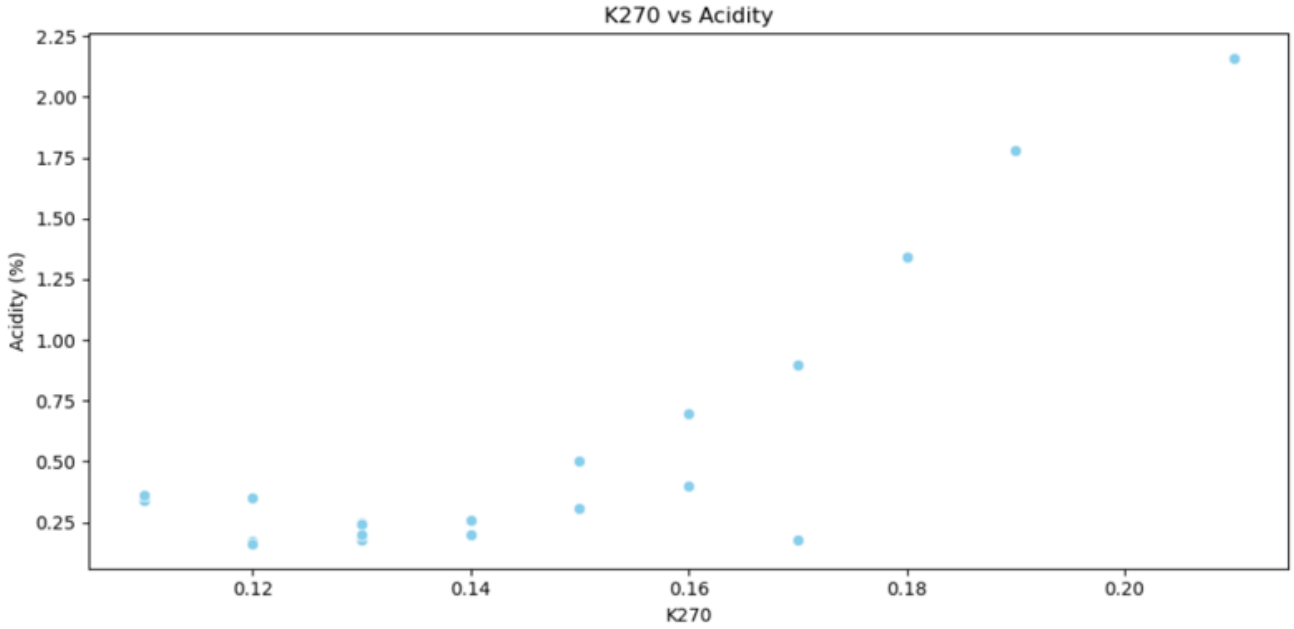


*Figure 9. K270 against Acidity.*

### Acidity vs Peroxide Index

Acidity and peroxide index have a positive relationship. This is because they both reflect hydrolysis and oxidation processes. Bad handling and storage lead to increased peroxides and fatty acids, which in turn lead to increased acidity and peroxide levels. Figure 10 depicts that increasing acidity increases peroxide levels. This is a strong positive correlation.

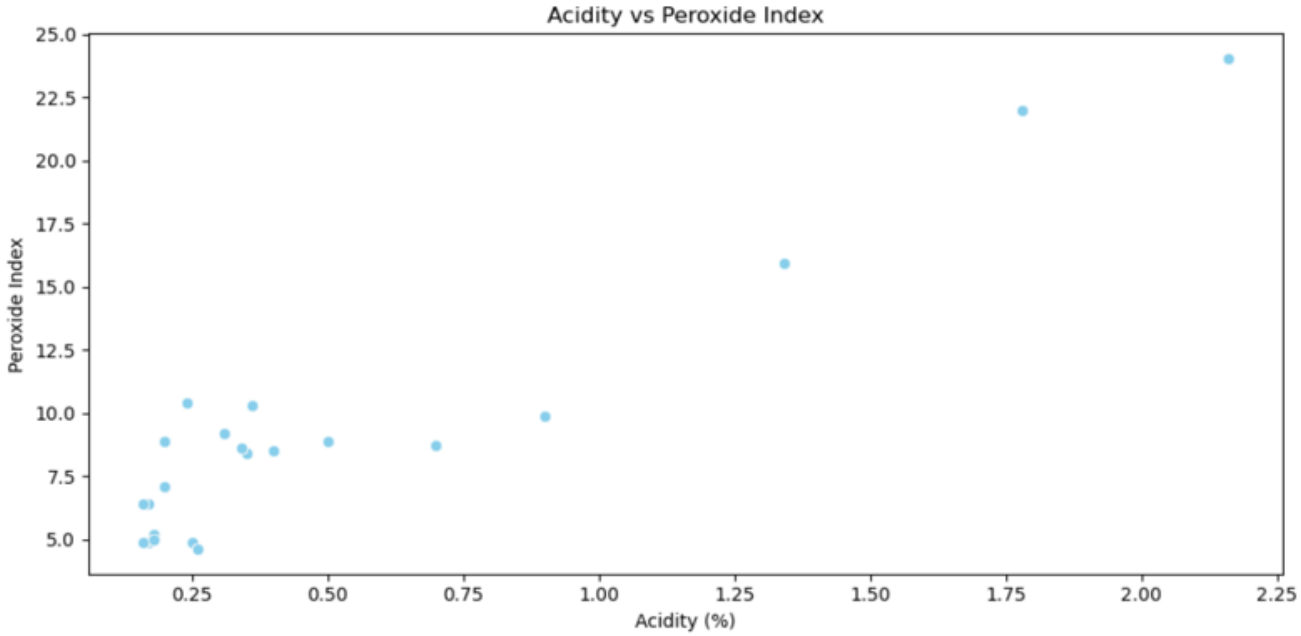


*Figure 10. Acidity against Peroxide index.*

**K232 vs LED Type**

Having different excitation wavelengths may lead to slightly different K232 measurements. However, under ideal calibrations, there should be no difference. This is because K232 should be independent of LED choice, provided instruments are calibrated well. Figure 11 shows that there is no difference at all, proving that the experiment had undergo ideal conditions.

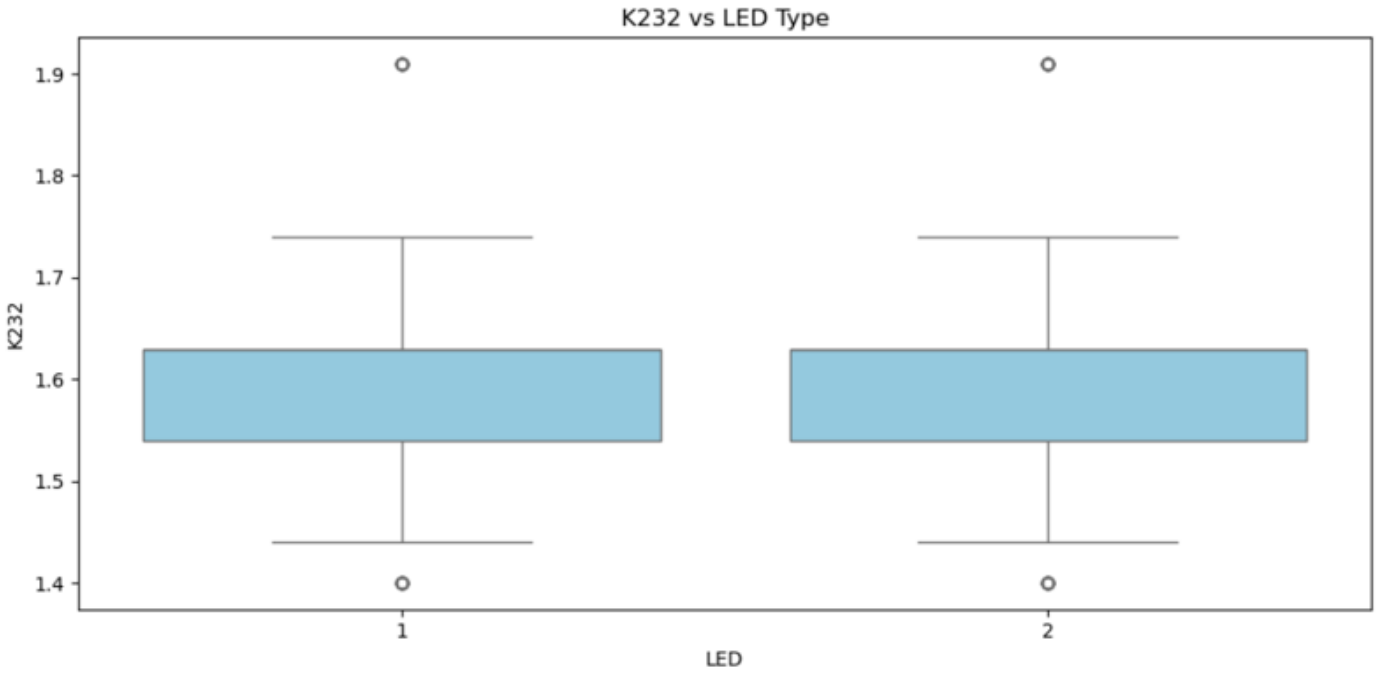


*Figure 11. K232 against LED types.*

# Discussion

The spectral measurements support the hypothesis that extra virgin and virgin olive oil fluoresce much greater than lampante. This is because of the presence of carotenoids, pheophytins and chlorophyll. These are all natural compounds to olive oil, showing that higher grade olive oil possesses more natural compounds and less oxidised compounds. Thus, fluorescence is a great indicator on oil freshness as well as pigment content, where a brighter fluoresce signifies a higher-grade olive oil.

The chemical properties of FAEES, K232, K270, Acidity and Peroxide Index are **ALL** positively correlated with each other. The reason being is that they are all triggered through the same process. Heat, light or oxygen are all triggers for olive oil degradation, where they can promote hydrolysis of lipids, or oxidation of fatty acids. Exposure to any of these rapidly degrade the quality of the olive oil. The process is that the trigger occurs, leading to hydrolysis of the fatty acids, leading to increased FAEES. Acidity and FAEES are increasing rapidly at the beginning of the trigger and continue to rise. Oxidation of lipids then occur, resulting in increased peroxide values and increased absorption at 232nm. A secondary oxidation occurs where these trienes are formed and absorb 270nm.

When comparing classification models using chemical properties, KNN achieved 96% accuracy, whereas DT output 88%. By using grid search's optimal parameters, the high accuracy suggests that these quality clusters tightly, and that Euclidean method effectively captures these clusters. In terms of spectra, both classifiers dropped in performance when using the engineered features. KNN dropped to 85%, and DT to 65%. This may be due to a large spread in lampante spectra introduced variation and noise, as well as missing several values in the spectra. The missing values were left as NaNs rather than inputting as 0 since 0 represents no fluorescence. However, this may have influenced both models to perform worse than using chemical properties.

Choosing between using chemical properties and spectra depends on the practical implications of future practices. Several factors that need to be accounted for are speed, cost, and simplicity. Using spectra offers rapid screening compared to chemical analysis taking a few days. Spectra also proves beneficial when it comes to cost. Besides the upfront cost of equipment, saving on worker cost and sample cost can influence a decision. Spectra is also more simple, where an instrument is needed to be deployed, whereas chemical analysis requires titrations and careful titrations. Nevertheless, chemical analysis is showing greater results.

The limitations to this project is that the sample diversity is not large enough. There are only 24 unique samples, each with 40 repetitions. Even with all these repetitions, many lampante observations were empty. It would be much better to condense the 20 repetitions into 1 mean, and do it again for the second LED, thus instead of 40 observations, there would be 2. Then the dataset is condensed and more samples can be tested for a more thorough approach.

# Conclusion

This study aims to model a machine learning program to classify olive oil into 3 categories of extra virgin, virgin and lampante. It uses data off the existing chemical properties and fluorescence spectra of each observation to do this. Missing values in the original dataset were handled through linear regression, as well as calculating median value of all values for K232. Instead of dropping empty rows, these rows now have computed values that use regression from extra and virgin to complete or dataset.

This was then programmed through 2 different classification models, KNN and DT. These two models would then be trained to predict the quality label, through both chemical properties and given spectra. When modelling using chemical properties, KNN gave an accuracy of 96%, whereas DT outputted 88%. KNN parameters used were k=6, weights=uniform, p=2. These parameters were found to be most effective through grid search. DT uses default parameters, and through manual looping found that max_depth=2 was the earliest stage to give the best result.

When modelling for spectra, by using features engineered like mean and max intensity, it was found that both KNN and DT performances had decreased in comparison to using chemical properties. KNN showed an accuracy of 85%, and DT with 65%. Moving forward, it can be said that chemical analysis would prove better.

However, it is to be noted that a large portion of the dataset has engineered data using linear regression. It may or may not reflect that chemical analysis is better, because the data is not 100% true. Another factor is that chemical analysis is costly and time consuming. Typical chemical analysis usually takes a few days, as well as the sample not being reusable. Using spectra may give a slightly worse score, however the positives of being cost and time efficient, as well as keeping the sample intact, may outweigh the benefits of chemical analysis.

The final verdict is that if accuracy is the greatest priority to a project without restraints, chemical analysis suits best. However, if time or costs are a concern, using the fluorescence spectra takes a few minutes, whilst still giving high accuracy.

# References


- Bavali, A., Rahmatpanahi, A., & Chegini, R. M. (2025). Quantitative analysis of the olive oil adulteration based on the assessment of multi-angle LIF spectral data using learning algorithms. *Journal of Food Composition and Analysis*, *144*, 107723. https://doi.org/10.1016/j.jfca.2025.107723
- Chen, S., Du, X., Zhao, W., Guo, P., Chen, H., Jiang, Y., & Wu, H. (2022). Olive oil classification with Laser-induced fluorescence (LIF) spectra using 1-dimensional convolutional neural network and dual convolution structure model. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *279*, 121418. https://doi.org/10.1016/j.saa.2022.121418
- María Isabel Sánchez-Rodríguez, Sánchez-López, E., Marinas, A., José María Caridad, & Francisco José Urbano. (2023). Enhancing the machine learning classification of olive oils by adding agro-climatic to NIR spectral information. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-3116942/v1
- *numpy.trapz — NumPy v1.21 Manual*. (2021). Numpy.org. https://numpy.org/doc/1.21/reference/generated/numpy.trapz.html
- Scikit Learn. (2012). *3.2. Tuning the hyper-parameters of an estimator — scikit-learn 0.22 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/grid_search.html
- Venturini, F., Fluri, S., Manas Mejari, Baumgartner, M., Piga, D., & Umberto Michelucci. (2024). Shedding light on the ageing of extra virgin olive oil: Probing the impact of temperature with fluorescence spectroscopy and machine learning techniques. *Lebensmittel-Wissenschaft + Technologie/Food Science & Technology*, *191*, 115679–115679. https://doi.org/10.1016/j.lwt.2023.115679
- Venturini, F., Michela Sperti, Umberto Michelucci, Gucciardi, A., & Deriu, M. A. (2023, January 10). *Dataset of Fluorescence Spectra and Chemical Parameters of Olive Oils*. Research Gate. https://doi.org/10.48550/arXiv.2301.04471
- Venturini, F., Sperti, M., Michelucci, U., Gucciardi, A., Martos, V. M., & Deriu, M. A. (2023). Extraction of physicochemical properties from the fluorescence spectrum with 1D convolutional neural networks: Application to olive oil. *Journal of Food Engineering*, *336*, 111198. https://doi.org/10.1016/j.jfoodeng.2022.111198