**Clustering wine using Machine learning models in python**

**Course:** Machine Learning

**Team:**

Rohitha Sai Alla – 700734780

Ramyasree Kuppam -700734078

Pradeep Reddy Malgari – 700741163

Namratha Ravula – 700740300

**Git link:**  https://github.com/RohithaSaiML/ML-Project

**Abstract:** Wine informatics is a relatively young subject of data science that analyzes massive wine databases. Many recent studies in wine informatics have focused on supervised learning. Wine, weather, region, and price forecasting. Using a dataset unique to the Bordeaux area, this investigation examines the effectiveness of K-means clustering as an unsupervised learning technique, specifically focusing on the ideal K search and filtering procedure in order to locate representative wines within each cluster (Xavier, 2020). Wine merchants may use this data to narrow their offerings to the best possible choices, wine experts can use the data to study wines in a certain area, vintage, and price range, and wine drinkers can use the data to find new wines to try. The same method may be used in other studies to compare and contrast wine from various regions/countries, vintages, and turning points. Leading wine distributor Champagne Company specializes in bringing premium and super-premium wines, champagnes, and spirits into the East African region. The company's major objective is to increase accessibility to the most expensive wine labels at a reasonable price. The business aims to meet the rising local need for specialized goods and the expanding tourism market, where visitors will not accept less. It is hoped that this article will pave the way for future research in unsupervised learning techniques for the field of Wine informatics studies of learning.

## **Keywords:** K-Means, Supervised learning, Semi Supervised learning, Unsupervised learning, Birch Clustering, DBScan clustering.

**I.INTRODUCTION**

The area of data science is gaining a lot of traction as a result of the proliferation of data produced all over the internet these days. This has contributed to the development in prominence of the data science sub field within the discipline of computer engineering. It is considered to be one of the most competitive industries of the 21st century. Each single sector is going to make an effort to modify its work processes. The practice of data science is based on the data-driven methodology, which assists organizations and businesses of all kinds in better comprehending the requirements of their respective customers. At this point in history, the era of the tech revolution, data has emerged as one of the most valuable aspects of data technology. According to the way in which they conduct their operations, many businesses amass various kinds of data. For instance, the data for a retail business may include the kind of goods they offer to their customers as well as the items themselves and how they sell them to those customers. In a similar vein, it may be the consumer's interest in viewing when it comes to Netflix. The goal of data science, as well as the goal of decision-making, is to obtain the most recent information from data that might otherwise go unnoticed. Additionally, the intent of deciding is to determine how they are anticipating the features to come from customer engagement strategy and reviews, as well as how they will use Mining in sentiment classification to make more informed business decisions.

In recent years, there has been a rise in demand for wine, which has resulted in a boost in wine consumption. As a result of rising consumer demand, the wine industry is exploring different ways to produce high-quality wines at lower expense. Depending on the climatic circumstances and the producer's goals, several varieties of wine may be created. It has become extremely relevant over the decades to classify multiple kinds of wines in order to ensure quality. This is due to the fact that the chemical structure of the majority of wines is comparable to that of one another, and that different types of wines have varying concentration. Whereas it has not been able to classify effectively based on wind qualities owing to a lack of techniques of production, thanks to the advent of a machine learning technique, feature choices can now be extracted, and wines may be categorized according to their qualities. This is owing to the fact that not only is it feasible, but it is also feasible to comprehend the significance of the chemical analysis variables in wines, which might be overlooked due to the cost-effectiveness of doing so. In furthermore, the framework of this white paper is centered on doing an investigation into the reasons behind the characteristics of wine that makes it appealing to consumers, and it does so by studying the aforementioned criteria.

The purity of wines is among the most important challenges facing the wine business today. Whether the individual who is defining quality is an expertise or not, quality can be determined by the person doing the defining. Because they have such a profound grasp of how wine is made, especially its chemical make-up, experts may provide a perspective on wine quality that is distinctive from that of the general public. On either hand, those who are not knowledgeable about wine are more prone to talk about its quality in terms of its price, appearance, and origin. The aroma and flavor characteristics of the wine play a significant part in determining the quality of the product (Water house et al., 2016). The flavor, smell, color, and other qualities of the wine are all determined by the wine's chemical composition (Sousa et al., 2014; Water house et al., 2016). The kind of grape, the conditions of the climate, the microbial strains that are present during fermenting, and the viticulture practices all have an impact on the chemical makeup. The capacity of modern digital technology to resolve mathematical equations has contributed to the rise in popularity of machine learning by making it possible to improve the effectiveness of the process. One of the contributing factors to the rise in popularity of learning algorithms is the proliferation of high-quality resources with which to experiment. This suggests that it is capable of providing accurate interpretations, which may assist in the process of making crucial decisions. As a consequence of its increasing applicability in a wide variety of industries, the volume of research that is being carried out in the field has fast increased, leading to the development of new sub specialties. Because of the quick pace at which methodologies in such fields are both developing and becoming extinct, it is possible that some of the study that has been published, even within the last ten years, may be rendered irrelevant by current practices.

**II.MOTIVATION**

Nevertheless, there are a few challenges connected with carrying out this study endeavor. In this research, the most major challenge we are working to solve is the small size of the sample we were able to collect. In the same way that it is exceedingly difficult and costly to get massive volumes of data in other types of experimental study, viticulture presents similar challenges. As a result of this reason, in order to tackle this issue, we developed synthetic data that had features that were similar to the features of the actual data. The risk of sensitive information being compromised is another challenge that must be overcome.

**III.MAIN CONTRIBUTIONS AND OBJECTIVES**

1)The fundamental objective of this study is to forecast the quality of wine by utilizing machine learning methods that take into account both physiochemical and chemical aspects of the wine.

2)In order to tackle this problem, we used 12 samples to create synthetic occurrences, and we saved six examples for the purpose of testing the models. 3)Third, we took on the challenge of dealing with a high number of variables (54 in this research), and we used a variety of different feature selection strategies to handle the difficulty we were facing. We contrasted the information to 54 other traits, as well as the 10 most important aspects, and the six most important features. illustrating the workflow architecture that was used in this research

4)Analyse the data in terms of various features responsible for Wine Clustering.

5)Finding the most suited machine learning model for the correct classification of clustering.

**IV.RELATED WORK**

## **Problem Statement:**

The task of a Data Scientist working is to perform wine segmentation using the given data set. The data set contains the findings of a chemical examination of three different varieties of wine produced in the same area of the United States.

Upon modeling, we are required to perform statistical analysis on the clusters, highlighting key findings, and making recommendations on business decisions based on your findings. Also, we want to build a classification machine learning model to carry out the predictions. This will include data processing to make the data set variables ready for machine learning algorithms during the designing of the data models to improve the accuracy of the model predictions. The project will be implemented in the Jupyter Notebook environment to generate the final notebook file.

Over the course of the last several years, a number of different methods to machine learning as well as methodology selections for feature extraction have been tried out on a wine database in an attempt to enhance it more useful. In order to categorize the quality degree of a wine, Er and Atasoy came up with a method that was based on support vector machines classifiers. It makes use of three different methods of categorization. An algorithm to forecast the grade or state of the vine was developed by our team after analyzing feedback received from customers.

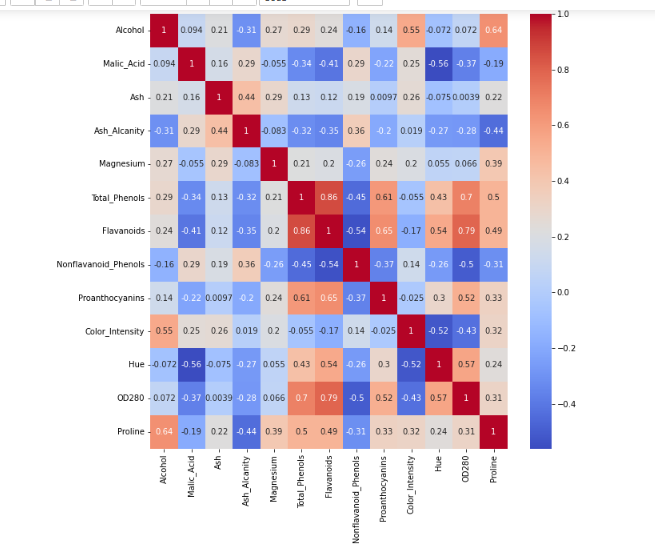
The third method for selecting features that were to be offered was based on a technique called hierarchical grouping as well as association rule. During the process of producing wine in warehousing, Applasammy presented an algorithm to forecast the quality of the wine. Beltran presented a method for the classification of wine data sets that was derived from genetic algorithms and fragrance chromatograms. ion of wine data sets that was derived from genetic algorithms and fragrance chromatograms. In order to categorize wines as per their specific characteristics and categories, Thakkar suggested using an algorithm that was formulated on the analytically layer clustering method. In order to provide clients with recommendations for the top wine items, an algorithm was proposed that is centered on a central clustering approach and makes use of a svm classifier as well as a randomized forest prediction method. They employed both white wine and red wine databases for their research project. They employed methods such as random forests, support vector machine and naive Bayes for the forecast, which was done by Kumar et al. (2020), who were interested in predicting the grade of red wine based on its many different characteristics (Kumar et al., 2020). They have determined the metrics, which includes things like precision, recollection, f1-score, accuracy, specificity, and categorization error.

**V.PROPOSED FRAMEWORK**

The Spyder notebook, Python version 3.7, 8 gigabytes of random-access memory (RAM), and Intel(R) Core (TM) i5-7200U CPU were used for the whole of the research presented here. This study expanded upon the foundation laid by earlier investigations on wine production by including participants' impressions of red wine's complexity (Parr et al., 2020). Through this investigation, the researchers were able to identify important aspects as well as the link that exists between perceptions of quality complexity. In order to gain a deeper comprehension of the intrinsic, physiological wine elements that affect wine experts' perceptions of service quality, sophistication, and varietal representatives in Pinot noir wines, a research project involving delicious taste was carried out with the participation of wine experts.

**Feature analysis using the correlation matrix:**

Here we used the python module SK learn to calculate the correlation between the columns and also other calculations such as the Variance inflation factor to select the most influencing features for the wine prediction.



1. **Means Clustering:**

K-means algorithm generally works on the centroid of clusters i.e. based on the number of clusters, for each clusters we have to find the middle point of data which equally separates the data .We have to repeat this for every cluster and find the best optimal solution. In this to find the best optimal solution we worked on different number of clusters and we found that the cluster with value 3 separates the data in more efficient way.

**VI.DATA DESCRIPTION**

The raw data only contains 18 samples, which is much too little for an analysis to be performed using ML algorithms. We used the Synthetic Minority over Sampling Method in order to generate sufficient samples for the development of machine learning systems (SMOTE). In order to prepare our dataset for analysis using SMOTE, we first divided it in half. In order to prevent data breach, the remaining seven samples from one dataset of 12 were put aside and used as a test dataset for the following assessment of machine learning techniques. This dataset was cast aside while the original dataset of samples collected was used to generate synthetic observations. Typically, while trying to balance data, statisticians will utilize the SMOTE approach, which involves establishing minority class samples in order to compare them to the classification model. For instance, if a dataset comprises 1000 items and 600 of those observations are red wine and 400 of those samples are white wine, then red wine would be the dominant class and white wine would be the minority category. In contrast, the circumstances surrounding our research are quite unique: we were given a total of 18 samples, each of which consisted of Pinot noir wines. Hence in order to generate synthetic data, we had to make a few assertions, including the following:

# Check dataset shape

df.shape()

# check data types

df.dtypes()

# dropping duplicates, if any

df.drop\_duplicates(inplace = True)

df.shape

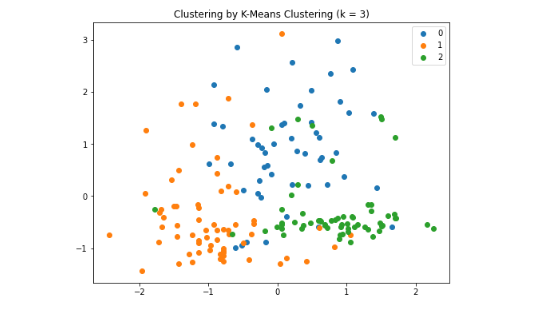
# check for missing data

df.isna().sum()

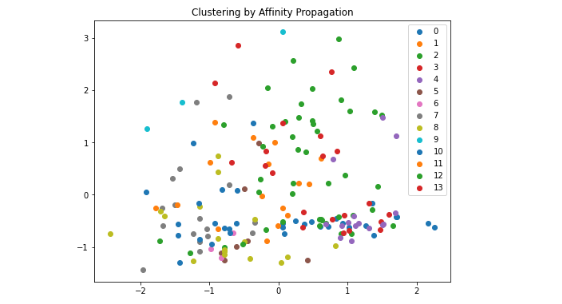
**VII.RESULTS/ANALYSIS**

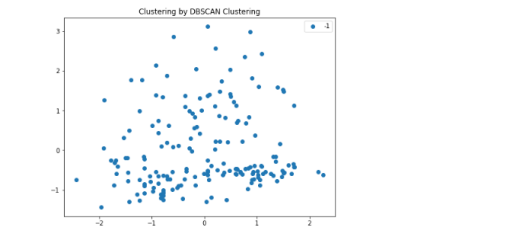
The findings show that K-means And Birch Clustering are the superior machine learning method for predicting wine quality. Simultaneously, K-means and Birch Clustering may make more accurate predictions by considering just some of the variables.

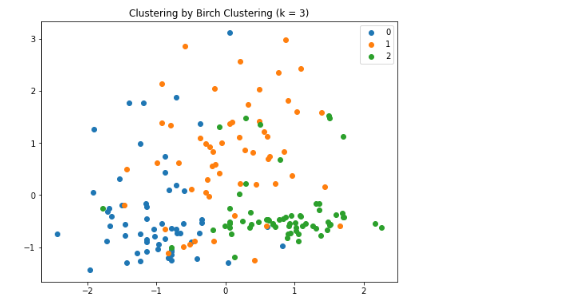
The purpose of this research is to use machine learning methods to predict a wine's quality score. Our team performed a number of measures to cleanse and organize the data in anticipation of a machine learning study to help us get there. Due to a lack of raw data, creating synthetic data was an important part of this research. Machine learning models cannot be trained or tested effectively with such a little data set. As a result, we employed an SMOTE method (described in Section 2.2) on a subset of 12 samples and put the rest to use for testing. In this research, we included numerous feature-related situations to boost the effectiveness of the models. Models of machine learning were evaluated both with and without feature extraction and critical variables. By training and testing our model in a variety of scenarios, we were able to determine that the K-Means and Birch classifier produced the most accurate predictions of wine quality. We also showed the importance of feature selection by showing how it improves classifier performance. We also demonstrated that critical variables chosen using any of the four feature selection methods improved performance of the models. Companies spend millions studying how to better forecast wine quality thanks to the importance of quality accreditation in the wine industry. Recent years have seen a rise in the use of machine learning methods to the study of wine. Researchers showed that the SMOTE algorithm combined with machine learning approaches could accurately categorize 4898 samples of Portuguese white wine and make predictions on their quality on the basis of "high," "normal," and "bad" categories. In another research, data mining techniques are used to glean insights from the same underlying raw wine data collection. Oenological theory is used to verify the study's findings. The authors discovered that volatile acidity degrades sensory properties and proposed that a happy medium should be reached between wine's richness and its freshness to provide optimal drinking pleasure (Cortez et al., 2021). In addition, significant factors influencing wine quality are identified using a Sequential Backward Elimination (RFE) technique, and performance measures are derived using non-linear decision tree classifications (Gupta,2018).



This is the visualization of Kmeans Clustering using the k value as 3.







**VIII.REFERENCES**

1. Trivedi, A., & Sehrawat, R. (2018, July). Wine quality detection through machine learning algorithms. In 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE) (pp. 1756-1760). IEEE.
2. Luna, R., Araya, M., Caris, J., & Cuevas-Valenzuela, J. (2020, November). A Digital Platform for the Management of Grapes and Wine Quality in the Winery. In 2020 39th International Conference of the Chilean Computer Science Society (SCCC) (pp. 1-7). IEEE.
3. Aich, S., Al-Absi, A. A., Hui, K. L., Lee, J. T., & Sain, M. (2018, February). A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques. In 2018 20th International conference on advanced communication technology (ICACT) (pp. 139-143). IEEE.
4. Waterhouse, A. L., Sacks, G. L., & Jeffery, D. W. (2016). Understandingwinechemistry.Cap.31.Grape genetics, chemistry, and breeding. In Understanding Wine Chemistry (pp. 2–5).
5. Sousa, E. C., Uchôa-Thomaz, A. M. A., Carioca, J. O. B., de Morais, S. M., de Lima, A., Martins, C. G., Alexandrino, C. D., Ferreira, P. A. T., Rodrigues, A. L. M., Rodrigues, S. P., Silva, J. do N., & Rodrigues, L. L. (2014). Chemical composition and bioactive compounds of grape pomace (Vitis vinifera L.), Benitaka variety, grown in the semiarid region of Northeast Brazil. Food Science and Technology, 34(1), 135–142. IEEE.
6. Dahal, K. R., Dahal, J. N., Banjade, H., & Gaire, S. (2021). Prediction of wine quality using machine learning algorithms. Open Journal of Statistics, 11, 278–289. IEEE.
7. Sinha, Kumar, A. (2020). Wine Quality and Taste Classification Using Machine Learning Model. International Journal of Innovative Research in Applied Sciences and Engineering, 4, 715-21.IEEE
8. Kumar, S., Agrawal, K., & Mandan, N. (2020). Red wine quality prediction using machine learning techniques. In 2020 International Conference on Computer Communication and Informatics, ICCCI 2020.IEEE
9. Gupta, U., Patidar, Y., Agarwal, A., & Singh, K. P. (2020). Wine quality analysis using machine learning algorithms. Lecture Notes in Networks and Systems, 106, 11–18. IEEE
10. Lee..S.Park, J., & Kang, K. (2015). Assessing wine quality using a decision tree. In 1st IEEE International Symposium on Systems Engineering, ISSE 2015 - Proceedings (pp. 176–178).IEEE
11. Chen, Q., Ma, X., Yu, Y., Sun, Y., & Zhu, Z. (2022). Multi-objective evolutionary multi-tasking algorithm using cross-dimensional and prediction-based knowledge transfer. *Information Sciences*, *586*, 540–562
12. Dabbura, I. (2022, September 27). *K-means clustering: Algorithm, applications, evaluation methods, and drawbacks*. Medium. R
13. McCune, J., Riley, A., & Chen, B. (2021). Clustering in wineinformatics with attribute selection to increase uniqueness of clusters. *Fermentation*, *7*(1), 27.
14. Xavier. (2020, June 3). *Tutorial: Clustering wines with K-means*. Kaggle. Retrieved September 30, 2022

15)P.Appalasamy, A.Mustapha, N.D.Rizal, F.Johari, and A.F.Mansor.(2012). “Classification-based Data Mining Approach for Quality Control in Wine Production,” Journal of Applied Sciences, 12(6),

16) Yesim Er\*1 , Ayten Atasoy1.(2016). “The Classification of White Wine and Red Wine According to Their Physicochemical Qualities”,ISSN 2147-67992147- 6799,

17) Mahima Gupta, U., Patidar, Y., Agarwal, A., & Singh, K. P. (2020). Wine quality analysis using machine learning algorithms. Lecture Notes in Networks and Systems, 106, 11–18.

18)Friedman, .H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29, 1189-1232.Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29, 1189-1232.

19)Nguyen, D. (2020, November 25). Red wine quality prediction using regression modeling and machine learning. Medium.

20)Palmer J, Chen B (2018) Wine informatics : regression on the grade and price of wines through their sensory attributes