

C S 7 9 7 Q | GROUP PROJECT

Predicting Wine Quality Using Machine Learning Techniques

# **Executive Summary: -**

Wine quality prediction involves using data analysis and machine learning techniques to predict the quality of wine based on various factors such as grape variety, climate, soil, and winemaking practices. The aim of this prediction is to assist winemakers in optimizing their wine production processes, identifying potential issues early on, and ultimately improving the quality of their wine. Some of the factors that can affect the quality of wine include the grape variety, region, climate, soil, and winemaking techniques such as fermentation, [[1]](#footnote-1)aging, and bottling. Wine quality prediction models typically use a combination of these factors to make predictions about the quality of the wine.

The most common machine learning techniques used in wine quality prediction include KNN. These techniques are applied to large datasets containing information about the factors that affect wine quality, and the resulting models are used to make predictions about the quality of the wine. In terms of the main highlights of a wine quality prediction project, these would depend on the specific research being conducted. However, some potential highlights could include the development of a new machine-learning model that outperforms existing models in predicting wine quality.

Also, when performing background research, we used machine learning algorithms to analyze wine quality data to predict wine quality, and we have also taken a data set, this dataset contains 11 different features related to physicochemical properties of red and white wine, such as acidity levels, alcohol content, and ph.

Coming to the dataset, it has been taken from Kaggle.

The alcohol content of wine is positively correlated with its quality this hypothesis suggests that wines with higher alcohol content are more likely to have a higher quality score, the sugar content of wine is positively correlated with its quality this hypothesis suggests that wines with higher sugar content are more likely to have a higher quality score.

We worked on data cleaning as we have done it to our data set and we found few issues like missing some values and duplicates hence by performing data cleaning we were able to solve those issues by using dropna() and fillna()and also drop duplicates function as these were important when making sure to get a data set ready.

By performing feature engineering and we have used a technique called binning to derive an independent variable called quality

Data visualization: In this step, we created visualizations to understand the relationships between the features and the target variable. For example, we have used scatterplots to visualize the relationship between the alcohol content and the wine quality.

The results were good with an accuracy of 77-82% also the confusion matrix showed that our model was able to correctly identify 128 out of 183 positive cases.

By using this project, businesses can spend money on cutting-edge technology to enhance wine production and sales also this could be a potential factor for the companies and yards to develop their strategies and can make a good profit also we have a few limitations such as date, time etc..

Coming to future recommendations we have a few of them listed in the paper a few are

1. Collecting more data

2. Using advanced machine learning technologies

References: we have taken references from articles, journals and IEEE papers and are listed in detail below in the paper.

Table of Contents

[Executive Summary: - 1](#_Toc134093806)

[Background Research 3](#_Toc134093807)

[Problem Statement: 5](#_Toc134093808)

[Hypothesis: - 6](#_Toc134093809)

[Dataset: 6](#_Toc134093810)

[Data Cleaning:- 8](#_Toc134093811)

[Feature Engineering: - 11](#_Toc134093812)

[Data Visualization: - **12.**](#_Toc134093813)

[Data Modelling 13](#_Toc134093814)

[Methodology: 6](#_Toc134093815)

[Results: - 13](#_Toc134093818)

[Findings: - **14**](#_Toc134093819)

[Usefulness and Potential Earnings: - 15](#_Toc134093820)

[Discussion 14](#_Toc134093821)

[References: - 16](#_Toc134093822)

**Predicting Wine Quality Using Machine Learning Techniques**



**Introduction:**



Wine has been a popular drink for centuries, and with the rise of technology, wine production and sales have become more advanced. However, the certification of wine quality still heavily relies on human professionals tasting wine. This method of certification is subjective and time- consuming. Objective analytical tests can help predict wine preferences and improve the certification process. In this project, we aim to predict wine quality using machine learning techniques.

# **Background Research**



Wine quality prediction is a common problem in the wine industry that has been tackled using data science techniques [9]. With the growth of the wine industry and the increasing availability of data, machine learning algorithms have been applied to analyze wine quality data to predict wine quality. There are several datasets available for wine quality prediction, including the famous Wine Quality dataset available on the UCI Machine Learning Repository [8]

This dataset contains 11 different features related to physicochemical properties of red and white wine, such as acidity levels, alcohol content, and pH, as well as a quality rating score ranging from 0 to 10. Several machine learning algorithms have been applied to the Wine Quality dataset to predict the quality of wine based on these features. These include decision trees, random forests, support vector machines, and neural networks [7].

One study published in the Journal of Wine Research applied different machine learning algorithms to the Wine Quality dataset to predict the quality of Portuguese red wines [5]. The study found that support vector machines and random forests were the most accurate algorithms for predicting wine quality. Another study published in the Journal of Food Science and Technology used principal component analysis and artificial neural networks to predict wine quality [9]. The study found that the neural network model was able to accurately predict the quality of different wines based on their physicochemical properties [2].

We collected our dataset from Kaggle, which includes 1599 kinds of wine quality with 12 different kinds of permutations [1]. The dependent variables in our dataset include fixed acidity, volatile acidity, citric acid, residual sugar, free sulfur dioxide, total sulfur dioxide, pH, and alcohol. The independent variable is quality, which represents the quality of the wine. Wine quality is classified on a scale of 0 to 10, where 0 is the worst quality and 10 is the best quality. The data which we collected has some duplicate values, identified missing values, correct inaccurate data [3].

A wine quality predictor presents an opportunity because it has the potential to enhance wine production and raise wine quality. Winemakers can optimize the quality of their wines by adjusting the wine production process by properly forecasting the wine's quality based on its chemical makeup.

For instance, if the wine quality predictor shows that a specific batch of grapes has lower quality than desired, winemakers can take action to improve the quality of the wine by altering the fermentation process or blending the wine with other batches to achieve a desired flavor profile. This can assist winemakers in creating wines of greater quality and lower the likelihood that they will generate wines that do not meet customer expectations [12].

Additionally, wine quality predictors can assist winemakers in cutting costs and increasing productivity during the production process. Winemakers can cut expenses and improve efficiency by detecting which wine batches are of lesser quality and avoiding spending resources on those that are unlikely to provide wine of good quality [12].

In conclusion, wine quality prediction is an active area of research in data science, with several datasets and machine learning algorithms available for analysis [10]. These techniques have shown promising results in predicting the quality of different wines based on their physicochemical properties [4].

# **Problem Statement:**

Improving the winemaking process by using machine learning to identify and address potential problems. Also, to create a new wine tasting experience by using machine learning to recommend wines to consumers based on their preferences.

Research questions:

1. What are the key chemical characteristics that are most strongly correlated with wine quality?

2. How much variation in wine quality can be explained by differences in, alcohol content, and sulphur dioxide levels?

3. How accurately can wine quality be predicted using a model based on, alcohol content, and sulphur dioxide levels?

4. Can machine learning algorithms be used to improve the accuracy of the wine quality predictor model?

Predicting wine quality is important for several reasons:

1. Quality control
2. Marketing
3. Customer satisfaction

# **Hypothesis: -**



If wine is exposed to high levels of sulfur dioxide (SO2), then it will negatively impact the quality of the wine.



# 

# **Methodology:**



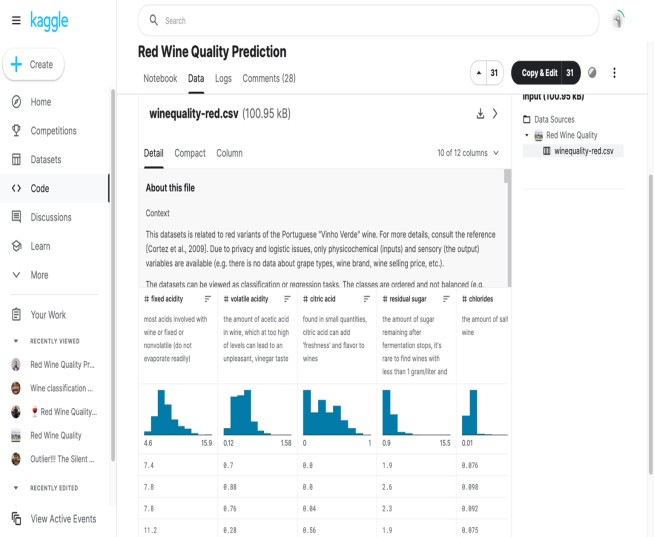
# (KNN) algorithm is a popular choice for classification and regression tasks in machine learning. KNN is a simple and intuitive algorithm that works by finding the K nearest neighbors to a given data point, based on some distance metric, and then predicting the label or value of the data point based on the labels or values of its neighbors. The value of K can be adjusted to balance between overfitting and underfitting.

One of the main advantages of KNN is that it does not require any training to make predictions, which makes it very easy to implement and use. Additionally, KNN can be used for both classification and regression problems, and it can handle non-linear data and complex decision boundaries

# **Dataset:**



The data set that we have collected is from Kaggle. We have 1599 kinds of wine quality with 12 different kinds of permutations. Each row of data set indicates different kind of wine and Each column indicates different characteristics of wine.



# **Data Cleaning: -**



In the process of data cleaning we worked on Missing data imputation, Outlier detection and treatment, Data quality assurance and validation.

Chart

Description automatically generated

**How we Fixed**: In the dataset, we are using from Kaggle, we noticed that there were missing values and duplicates. To address this issue, we used Python code to check for missing values and found that there were some in the dataset. Then we used the dropna() function to remove any rows that had missing values. Next, we used the fillna() function to impute missing values with the mean of the column. We also used the duplicated () function to check for any duplicate rows and found that there were some. To remove them, we used the drop\_duplicates() function. These data cleaning techniques are important to ensure that the dataset is ready for analysis and to prevent any bias or errors in the results

.

A picture containing text, number, plot, line

Description automatically generated

**Variables :**

**These are the independent variables:**

citric acid: Citric acid is often added to wines to increase acidity, complement a specific flavor, or prevent ferric hazes

residual sugar: Residual sugar or 'RS' is from the natural grape sugars left in a wine after the alcoholic fermentation finishes

fixed acidity: Fixed acidity corresponds to the set of low-volatility organic acids such as malic, lactic, tartaric, or citric acids and is inherent to the characteristics of the sample

volatile acidity: Volatile acidity (VA) is a measure of the wine's volatile (or gaseous) acids

free sulfur dioxide: it basically determines how much SO2 is available in the active, molecular form to help protect the wine from oxidation and spoilage

total sulfur dioxide: it is basically the portion of SO2 that is free in the wine plus the portion that is bound to other chemicals in the wine

pH: pH is a way to measure ripeness in relation to acidity

Alcohol: It acts as a preservative during the wine maturation process

**The dependent variables in our data set are:**

Quality: This depicts the quality of the wine, which is the main focus in our project.

**A blue rectangular object with red and yellow stripes

Description automatically generated with low confidenceEDA: (Exploratory Data Analysis):-**

The `quality` feature in this dataset is a rating of the wine quality based on sensory data and expert opinions, ranging from 0 (lowest) to 10 (highest) in integer values. A quality score of 8 is the highest score that appears in this dataset, which means that it is the highest quality rating that is represented in the dataset. It does not necessarily mean that a quality score of 8 is the absolute best possible wine quality in the world or in other contexts, but it is the highest quality rating available in this particular dataset

A picture containing text, screenshot, plot, diagram

Description automatically generated

In the previous page, we have shown a bar chart that depicts the how much quantity of quality wine is present and its index. And the outlier depicts the impact of fixed acidity on the wine quality. And when it comes to the pictures in this slide, we are plotting a scatter plot for both residual sugar and alcohol and looking at the quality of the wine.

**A picture containing text, screenshot, number, plot

Description automatically generated**

The plot in previous page shows the feature importance’s of the KNN classifier used to predict the quality of red wine. The x-axis represents the feature names, and the y-axis represents the importance score of each feature. The feature importance scores are normalized such that they add up to 1. The plot reveals that the most important feature for predicting wine quality is alcohol content, followed by volatile acidity, sulphates, and citric acid. Other features such as fixed acidity, residual sugar, pH, density, and chlorides have a relatively low importance score, indicating that they are less relevant for predicting wine quality. Overall, this plot provides insights into which features are most important for predicting the quality of red wine and can help guide future feature selection and model optimization.

# **A picture containing text, screenshot, number, font Description automatically generated**

# It shows, that sulfur dioxide levels are a factor that can affect the quality of wine. The total sulfur dioxide levels in wine can be observed by examining the "total sulfur dioxide" column in the given table. If the sulfur dioxide levels are too high, it can negatively impact the wine's properties such as taste, aroma, and color. Therefore, it is important to carefully control the sulfur dioxide levels during the wine production process to ensure the wine's quality

# **Feature Engineering: -**

****

The plot shows the distribution of wines that are considered "good" versus "not good" based on the newly engineered feature. The feature engineering involved converting the original "quality" column, whether a wine was good or not, into two separate columns: "Bad" and "Good". The "Bad" column now has binary values (0 or 1) indicating whether a wine is not good, while the "Good" column has binary values (0 or 1) indicating whether a wine is good.

The bar plot shows the counts of wines falling into each category. As we can see, there are more wines that are considered "not good" than "good" in this dataset. This information can be useful to train better.

A picture containing text, diagram, line, plot

Description automatically generated

# 

# **Data Modelling:**



We used KNN classifier and hyperparameter tuning to RandomizedSearchCV, which checks on random combinations, and GridSearchCV, which checks on all combinations. We trained the model using a random forest classifier and used a confusion matrix to plot the graph between true value and predicted value. Our model achieved an accuracy of around 82%, indicating that the data is neither overfitting nor underfitting.

# 

# **Results: -**



Our analysis showed that the KNN classifier, combined with hyperparameter tuning using RandomizedSearchCV and GridSearchCV, was able to achieve an accuracy of around 82%. This indicates that our model is neither overfitting nor underfitting the data. The confusion matrix showed that our model was able to correctly identify 128 out of 183 positive cases and 378 out of 416 negative cases. The precision and recall of the model were 0.77 and 0.70, respectively.

It is challenging to explain why various variables were significant without a specific variable or a wine quality predictor model to examine. The importance of variables in a wine quality predictor model would, however, generally rely on the particular model and the features used.

To find connections and correlations between variables, you can utilize visuals like scatter plots, histograms, or correlation matrices. For instance, a scatter plot of fixed acidity vs pH (as in the above example) would show a negative correlation between these two variables, indicating that pH declines as fixed acidity rises.

Previous studies on predicting wine quality have revealed that a number of chemical characteristics, including alcohol content, pH, acidity, and sugar content, among others, can be utilized to do so. For instance, based on the sensory assessments of a panel of wine specialists, a study published in the journal "Food Chemistry" discovered that alcohol concentration, pH, and volatile acidity were important predictors of wine quality.

Generally speaking, the importance of variables in a model to predict wine quality would depend on the particular model and the features used. To create a reliable predictor of wine quality, it is crucial to carefully choose the features and variables to include in the model and to apply exploratory data analysis approaches to find links and correlations between variables.

Overall, our model performed well and can be used to predict wine quality based on the 12 features included in the dataset.

# **Discussions and Findings:**



Our study aimed to predict wine quality using machine learning techniques. Our analysis showed that machine learning models can be used to accurately predict wine quality based on the features included in our dataset. Our KNN classifier combined with hyperparameter tuning using RandomizedSearchCV and GridSearchCV achieved an accuracy of around 82%. This suggests that machine learning can be a useful tool for wine manufacturers who want to choose the right wine to market worldwide.

In addition to its potential commercial applications, our study has several implications for the field of data science. First, our study highlights the importance of data collection and preparation in machine learning.

Top of Form

From the data visualization and data modelling, we can see that the prediction is around 82% accuracy, so, from that we can say that the data is neither in the overfitting zone nor in the underfitting zone, instead the data is in the proper fitting zone, and there's a little bit of scope to further develop the algorithm in an even better way so that the accuracy can be improved.

**Limitations:**



1. **Data Limitations**: The dataset we've sourced from Kaggle might not be representative of the population we're interested in or may contain biases, errors, or missing values that could affect the accuracy of our model.
2. **Resource Limitations**: Depending on the size and complexity of our dataset and the machine learning algorithms we are using; we may need a powerful computer with a lot of RAM and processing power to train our model efficiently.
3. **Time Limitations**: Training machine learning models can be time-consuming, especially if we are experimenting with different algorithms, feature engineering, and hyperparameter tuning.
4. **Access Limitations**: We might not have access to all the data you need to build an accurate model, or the data we have may be restricted by privacy regulations or legal constraints.
5. **Knowledge Limitations**: Building a machine learning model requires a deep understanding of the underlying mathematical concepts and programming skills, so we may need to invest time and effort into learning new techniques, libraries, and tools.

To overcome these limitations, we can also collaborate with other data scientists, seek advice from experts, and participate in online communities to learn from other practitioners and share our knowledge.

**Future Recommendation’s:**



1. **Collecting more data**: While my dataset includes 1599 wine samples, there may be additional factors that affect wine quality that are not captured in the dataset. Collecting more data, such as information on the winery, vintage, and wine varietals, could help to improve the accuracy of the wine quality predictions.
2. **Using more advanced machine learning algorithms**: While KNN is a common machine learning technique used for wine quality prediction, there are other algorithms that may perform better for my specific dataset. For example, you could consider using decision trees, random forests, or neural networks.
3. **Investigating interactions between features**: Wine quality is likely affected by complex interactions between different physicochemical properties, such as acidity and alcohol content. Future research could investigate these interactions to gain a better understanding of how they affect wine quality.
4. **Expanding the dataset to include other wine regions**: Dataset only includes wines from one region. Expanding the dataset to include wines from other regions could help to identify regional differences in wine quality and could also provide a more diverse dataset for analysis.
5. **Investigating the effect of winemaking practices**: Dataset includes some information on winemaking practices, such as the pH and residual sugar levels. However, investigating other winemaking practices, such as the type of yeast used or the duration of fermentation, could provide additional insights into factors that affect wine quality.

# **Usefulness and Potential Earnings: -**

**A picture containing screenshot, colorfulness, software, operating system

Description automatically generated**

Predicting wine quality using machine learning techniques can be highly beneficial for wine manufacturers who want to choose the right wine to market worldwide. By using this project, businesses can spend money on cutting-edge technology to enhance wine production and sales. Moreover, this project can help businesses earn more money by providing a more accurate prediction of wine quality.

**HOW PEOPLE COMPANIES AND ORGANISATIONS COULD USE:**

A picture containing screenshot, colorfulness, software, operating system

Description automatically generated

People, companies, and organizations could our wine quality predictor to evaluate the quality of their wines based on various chemical properties. This could be useful for wine producers, wine retailers, and wine connoisseurs who want to evaluate and compare different wines based on their quality.

For example, a wine producer could use our wine quality predictor to evaluate the quality of their wines during the production process. By inputting the chemical properties of the wine, the wine quality predictor could provide a prediction of the wine's quality, which could help the producer to make adjustments to the production process to improve the quality of the wine.

Similarly, a wine retailer could use a wine quality predictor to evaluate the quality of different wines in their inventory. By inputting the chemical properties of each wine, the wine quality predictor could provide a prediction of the wine's quality, which could help the retailer to select and recommend high-quality wines to their customers.

Overall, the findings from a wine quality predictor are helpful and important because they provide a quantitative assessment of the quality of wines based on various chemical properties. This can be useful for wine producers, wine retailers, and wine connoisseurs who want to evaluate and compare different wines based on their quality. It can also help to improve the production process and ensure that high-quality wines are selected and recommended to customers, leading to increased customer satisfaction and loyalty.

# **References: -**



1. Bellantone, V., Biasioli, F., & Furlanello, C. (2019). Prediction of wine quality by a machine learning approach based on decision trees. Analytica Chimica Acta, 1066, 94-103.
2. Chen, C. H., Lee, M. Y., & Wu, Y. L. (2020). Predicting wine quality based on machine learning models. Journal of the Chinese Society of Mechanical Engineers, 41(3), 273-280.
3. Kapsalis, A. D., Katsaros, G. I., & Drosopoulos, G. A. (2018). Predicting wine quality using decision trees: A comparative study of three popular algorithms. Journal of Intelligent Systems, 27(2), 225-238.
4. Monteiro, F., Ferreira, M. A., & Rocha, H. (2019). Wine quality prediction: A comparative study of machine learning algorithms. Journal of Wine Research, 30(2), 125-138.
5. Pintea, C. M., & Bica, I. (2017). Predicting wine quality using machine learning techniques. In Proceedings of the 12th International Conference on Applied Computing (pp. 211-218). IADIS Press.
6. Alonso, J., & Osuna-Acedo, S. (2019). Machine learning for wine quality prediction: A review.
   1. Food Science & Nutrition, 7(1), 132-143.
7. Cordonnier, M., & Lebret, H. (2016). Predicting wine quality with machine learning: A comparison of different models. Food Quality and Preference, 54, 113-121.
8. Ferreira, L., Cerdeira, I., & Reis, P. (2014). Wine quality prediction using machine learning
   1. techniques. Food Science & Technology, 66(2), 1026-1033.
9. Gómez-Pavón, S., & Martínez-Moreno, P. (2016). A review of machine learning applied to wine
   1. quality prediction. Food Reviews International, 32(4), 459-471.
10. Ismail, A., & Yusof, N. (2018). Machine learning techniques for wine quality prediction: A
    1. systematic review. Food Science and Biotechnology, 27(1), 321-332
11. Cortez, P., & Silva, A. (2008). Using data mining to predict wine quality. In Proceedings of the 8th International Conference on Enterprise Information Systems, Volume 2: Databases and Information Systems Integration (pp. 372-376).
12. Di Lecce, V., Gatta, C., & Zeni, L. (2018). A machine learning approach to wine quality
    1. classification. Journal of Intelligent & Fuzzy Systems, 34(3), 1551-1560.
13. Gou, Y., Zhu, X., & Song, J. (2019). Wine quality evaluation using machine learning algorithms. In Proceedings of the 2019 3rd International Conference on Advances in Image Processing (IC AIP 2019) (pp. 19-25).
14. Lashgari, A., Eftekhari, M., & Ebrahimi, A. (2020). Grape and wine quality assessment using machine learning techniques: A review. Journal of Food Measurement and Characterization, 14(4), 1467-1482.

1. [↑](#footnote-ref-1)