**Group Assignment 1**

*"* *Innovation Uncorked: VinifyTech's Journey to Transforming Winemaking"*

**Machine Learning**

**Spring 23/34**

**Nuno André da Silva**

**Qiwei Han**

**TA: Guilherme Ramos**

**TA: Alessandro Gambetti**

Hendrik Künnemann - 57995

Luc Marcel Pellinger - 58611

Moritz Lind - 61230

Tomás Shoemaker Simão Gonçalves - 47016

Zeno Eule - 59093

Content

[1. Executive Summary 3](#_Toc163038594)

[2. Introduction 3](#_Toc163038595)

[3. Data Description 3](#_Toc163038596)

[4. Methods 4](#_Toc163038597)

[a. Data Cleaning and Exploratory Data Analysis (EDA) 4](#_Toc163038598)

[b. Feature Engineering and Feature Importance 4](#_Toc163038599)

[c. Modeling 4](#_Toc163038600)

[5. Results 5](#_Toc163038601)

[a. Data Cleaning and Exploratory Data Analysis (EDA) 5](#_Toc163038602)

[b. Feature Engineering and Feature Importance 6](#_Toc163038603)

[c. Modeling 6](#_Toc163038604)

[6. Recommendations 7](#_Toc163038605)

# Executive Summary

VinifyTech embarked on a transformative journey to decode the intricate relationship between wine composition and quality, aiming to revolutionize winemaking through data-driven insights. Through collaboration with vineyards and wineries, we curated comprehensive datasets for red and white wines, meticulously characterized by experts. Our analysis encompassed data cleaning, exploratory data analysis (EDA), feature engineering, and modeling, leveraging machine learning algorithms to predict wine quality accurately.

In the data cleaning and EDA phase, we ensured dataset completeness by standardizing column names, addressing missing values, and removing duplicates. Insights from EDA revealed key chemical attributes and their correlations with wine quality, guiding feature engineering efforts. Feature groups were crafted to capture acidity, sugar content, sulfur dioxide levels, and more, enhancing predictive power and interpretability. Feature selection strategies and dimensionality reduction via PCA were employed to streamline analysis and maintain model robustness.

Various machine learning models, including Decision Trees, Random Forests, Stochastic Gradient Descent, and Support Vector Classifiers, were trained and evaluated for red and white wines. Each model exhibited distinct performance metrics across different quality categories, with SGD achieving high accuracy for red wine and RF showing superiority for white wine classification.

Based on F1-scores and interpretability, Decision Trees are recommended for red wine classification, while Random Forests are preferred for white wine. VinifyTech's predictive modeling can further evolve through hyperparameter tuning, feature refinement, and ensemble techniques, enhancing accuracy and driving innovation in winemaking. By integrating domain knowledge, prioritizing user feedback, and investing in infrastructure and talent, VinifyTech is poised to unlock the full potential of data-driven winemaking, shaping a future of excellence and innovation in the industry.

# Introduction

VinifyTech stands at the forefront of a transformative wave in the winemaking industry, tackling the perennial challenge of harmonizing tradition with the evolving landscape of high-quality wine production. Stemming from a visionary insight that wine embodies both artistry and scientific precision, our journey began with a resolute mission: to decode the intricate relationship between wine composition and its qualitative essence. Guided by our CEO and lead data scientist, we embarked on an ambitious quest to scrutinize wine data comprehensively, unraveling the cryptic interplay of chemical variables and quality ratings.

Armed with sophisticated machine learning algorithms, including regression, decision trees, and random forests, our endeavor transcends mere data analysis - it's an immersive exploration into the essence of winemaking. By delving into the minutiae of acidity levels, sugar content, and sulfur dioxide concentrations, we're not only crunching numbers; we're forging a profound understanding of the alchemy that defines every bottle of wine. The journey undertaken by VinifyTech is not just about technological innovation; it embodies a profound commitment to enhancing the very fabric of winemaking. Our objective is clear: to empower winemakers with actionable insights and cutting-edge tools that propel them beyond conventional boundaries, ensuring each vintage surpasses expectations. Through this relentless pursuit of excellence, we envision a paradigm shift in the industry—a future where every bottle embodies the pinnacle of quality and distinction, marking a new chapter in the legacy of winemaking.

# Data Description

In our relentless pursuit of excellence, VinifyTech collaborated with local vineyards and wineries, leveraging their expertise to collect meticulously curated datasets. Through partnerships with vintners who share our vision for innovation, we gained access to a diverse range of white and red wine samples, each meticulously characterized by trained sommeliers and experts. This collaborative approach ensured the authenticity and richness of our data while fostering a sense of community in our quest to redefine winemaking through data-driven insights.

Aligned with VinifyTech's mission to bridge tradition and innovation in winemaking, we curated two comprehensive datasets focusing on white and red wines. These datasets encapsulate essential parameters crucial for understanding the intricate relationship between wine composition and quality. Each wine sample underwent physicochemical tests and sensory evaluations, with experts ranking each wine on a scale from 0 to 10, providing a nuanced spectrum of quality assessments.

Containing features such as acidity levels, sugar content, sulfur dioxide concentrations, and quality ratings, these datasets serve as the cornerstone for our machine learning initiative. Through meticulous curation and rigorous analysis, we aim to unravel hidden correlations between these chemical components and perceived wine quality. With these datasets, VinifyTech embarks on a journey to revolutionize winemaking by leveraging data-driven insights to craft superior quality wines that resonate with the modern consumer palate.

# Methods

## Data Cleaning and Exploratory Data Analysis (EDA)

The initial step in our study involved preparing the dataset for analysis through data cleaning procedures. We began by standardizing column names to ensure consistency and facilitate data access. This process involved renaming columns with spaces to ensure uniformity across the dataset. Following column renaming, we conducted a thorough check for missing values within the dataset. Any missing values found were appropriately addressed through imputation or removal, depending on the data's nature and extent of missingness. Additionally, we identified and handled duplicate records within the dataset to ensure data uniqueness and avoid potential biases.

Exploratory Data Analysis (EDA) served as a critical precursor to our analysis, providing insights into the dataset's characteristics and variable relationships. We initiated the EDA process by generating summary statistics for all variables, offering insights into central tendencies, dispersions, and distributions of the data. We also utilized histograms to visualize the distribution of each variable, enabling us to assess data spread and shape intuitively. This visual exploration allowed us to identify potential patterns and anomalies within the dataset.

To streamline our analysis, we grouped dataset features into three categories based on their characteristics: Acid Content Features, Taste Profiles Features, and Preservation Characteristics Features. Each feature grouping underwent detailed analysis to uncover insights specific to its domain. We employed a combination of histograms, boxplots, and violin plots to analyze feature distributions within red and white wines separately. These visualizations facilitated effective comparison of feature characteristics across different wine types and quality ratings.

To compare red and white wines in terms of quality ratings and feature distributions, we created binary color features and categorized wine quality ratings into low, medium, and high categories. Through visual comparisons using histograms, boxplots, and violin plots, we evaluated differences in quality ratings and feature characteristics between red and white wines.

## Feature Engineering and Feature Importance

We conducted feature engineering to boost predictive power and interpretability. Total acidity was computed by summing fixed, volatile, and citric acid content. Average acidity was derived from their mean. Ratios like acid ratio, acidity to alcohol ratio, and pH to acidity ratio were calculated to explore acidity's relationships. Sugar-related features assessed sweetness-acidity balance, including sugar to alcohol rate and sugar to acidity rate. Sulfur dioxide features evaluated preservation, e.g., SO2 ratio and sulfates to SO2 rate. An ageability score was computed using weighted averages of acidity, total sulfur dioxide, alcohol, and residual sugar. We performed a correlation matrix to identify highly correlated features, removing those above a 0.75 correlation threshold to address multicollinearity. Next, we shuffled the dataset and standardized numerical features using the Scikit-learn library's StandardScaler object. Principal Component Analysis (PCA) was applied to reduce dimensionality while retaining 90% of variance, determining the optimal number of principal components through a cumulative explained variance plot. Finally, we created a new dataframe containing PCA components for subsequent analysis.

## Modeling

In our exploration of wine quality assessment, we undertake a systematic approach using various machine learning models. From Decision Trees to Random Forests, each model offers unique perspectives on wine characteristics, aiming to refine predictive accuracy and deepen our understanding of quality evaluation dynamics. Join us on this journey to unlock hidden insights and merge tradition with innovation in winemaking.

**Decision Tree (DT):**

We employed the Decision Tree classifier for its simplicity and interpretability, creating tree-like structures based on feature values to capture nonlinear relationships and significant predictors of wine quality. Hyperparameters like maximum depth and minimum samples per leaf were fine-tuned to optimize performance without overfitting.

**Random Forest (RF):**

To enhance accuracy and reduce overfitting, we utilized the Random Forest Classifier, aggregating predictions from multiple decision trees trained on subsets of the dataset. By fine-tuning parameters such as minimum samples per leaf and the number of trees, we balanced complexity and accuracy effectively.

**Stochastic Gradient Descent (SGD):**

Incorporating the Stochastic Gradient Descent (SGD) Classifier facilitated efficient handling of large-scale datasets and sparse data. This model optimizes a linear model using iterative updates based on small subsets of training data, ensuring computational efficiency and scalability. Hyperparameters were fine-tuned to enhance prediction accuracy.

**Support Vector Classifier (SVC):**

The Support Vector Classifier (SVC) was integrated for its effectiveness in handling high-dimensional data and complex decision boundaries. By aiming to find an optimal hyperplane that maximizes the margin between quality classes, the SVC enhances generalization and robustness. Optimized hyperparameters contribute to more accurate predictions of wine quality.

# Results

## Data Cleaning and Exploratory Data Analysis (EDA)

**Red Wine**

In the data cleaning and exploratory data analysis (EDA) phase for red wine, we conducted a thorough examination of the dataset to gain insights into its characteristics and feature relationships. Initially, we addressed variable names containing spaces by renaming columns for consistency and ease of manipulation. With 1599 rows and 12 columns, the dataset provided substantial data for analysis, devoid of missing values, ensuring data completeness. Additionally, removing 240 duplicate rows ensured data uniqueness, mitigating potential biases or inaccuracies.

Analyzing the dataset revealed key insights into wine attributes and quality ratings. The average quality rating across 1,359 wine samples was computed to be 5.62, suggesting a focus on mid-range wines. Variability in chemical attributes like fixed acidity, volatile acidity, and alcohol content underscored the diverse compositions present. Correlation analysis highlighted the positive correlation between alcohol content and quality, while volatile acidity showed a significant negative correlation with quality, providing valuable guidance for winemakers.

To deepen our understanding, features were grouped into categories reflecting acid content, taste profiles, and preservation characteristics. This allowed for focused analysis, revealing insights into their collective impact on wine quality. For instance, pH levels exhibited a bell-shaped distribution, with higher-quality wines showing a subtle shift towards lower pH, reflecting the importance of acidity levels. Similarly, alcohol content demonstrated a clear upward trend with quality, while residual sugar showed minimal variation.

These observations inform our approach to machine learning model development, guiding our efforts to accurately predict wine quality and revolutionize winemaking practices. By comprehensively analyzing the dataset, VinifyTech is poised to leverage machine learning algorithms effectively in enhancing wine production processes.

**White Wine**

In our analysis of white wine data, consisting of 4898 rows and 12 columns, we undertook a meticulous exploration to understand the intricate relationships between attributes and their influence on wine quality, mirroring our approach with red wine data. The absence of missing values and removal of 937 duplicate rows ensured data completeness and uniqueness, laying a solid foundation for analysis. Initial insights revealed an average wine quality rating of 5.85, indicating a focus on mid-range wines with slightly higher ratings than red wines.

Key chemical attributes displayed significant variability, hinting at the complex factors shaping wine quality. Exploring correlations highlighted the strong positive correlation between alcohol content and quality, while density exhibited a notable negative correlation. Relationships between pH, fixed acidity, and citric acid underscored fundamental aspects of wine chemistry.

Grouping features facilitated a focused analysis, revealing nuanced insights into their impact on wine quality. For instance, volatile acidity trends towards lower levels in mid-quality wines, aligning with its correlation with quality. Similarly, sulphates showed a trend towards higher median values with increasing wine quality.

These observations provide valuable guidance for optimizing production processes and enhancing wine quality. By understanding these dynamics, VinifyTech is well-positioned to leverage machine learning algorithms accurately and drive innovation in the winemaking industry.

**Red Wine vs. White Wine Comparison**

When comparing red and white wines, we uncover key insights into their chemical composition, quality distribution, and the relationship between attributes and wine quality.

The dataset comprises 6497 rows and 13 columns, with white wines constituting 74.45% of samples and red wines 25.55%. While the dataset is balanced in terms of missing values and duplicates, there's an imbalance in wine color distribution, with white wines outnumbering red wines by 200%.

Binning quality ratings into 'low', 'medium', and 'high' categories clarifies wine quality trends. Both red and white wines predominantly fall into the ‘medium’ category (red: ~82%, white: 75%), with fewer samples in the ‘high’ (red: ~13%, white: 21%) and ‘low’ categories (red: ~5%, white: ~4%), suggesting a perception of generally higher quality for white wines.

Analysis of chemical composition reveals that white wines tend to have slightly lower fixed acidity compared to red wines, with higher volatile acidity levels in red wines. White wines exhibit a wider distribution of residual sugar levels, reflecting diverse winemaking styles, while red wines are consistently produced in a drier style.

In terms of taste-related features, alcohol content doesn't vary significantly between red and white wines, but chloride levels are slightly higher in red wines. Preservation-related features show higher levels of sulfates and sulfur dioxide in white wines, indicating their greater susceptibility to oxidation compared to red wines.

Analyzing quality ratings within feature groupings reveals that wines with higher quality ratings tend to have lower volatile acidity levels and slightly higher citric acid levels. They also exhibit slightly higher alcohol content and lower chloride levels. Preservation-related features show that higher quality wines have slightly higher sulphate levels and lower density levels, suggesting better preservation characteristics and a lighter mouthfeel.

In summary, while red and white wines differ in chemical composition and quality distribution, both exhibit similar trends in taste and preservation-related features affecting wine quality. Understanding these nuances is essential for optimizing production processes and enhancing overall wine quality.

## Feature Engineering and Feature Importance

In feature engineering, we've meticulously crafted various groups to capture complex relationships between chemical components and wine quality. Acid Balance features focus on acidity equilibrium crucial for taste and stability, while Sugar Related features explore sweetness interactions revealing insights into wine body and flavor. Sulfur Dioxide features illuminate wine preservation, crucial for shelf life, while Salt and Alcohol Balance features delve into salinity's influence on alcohol perception, impacting flavor balance. Overall Balance and Quality Indicators serve as proxies for wine character relative to alcohol strength. Interaction Terms uncover nuanced component relationships, revealing intricate dynamics.

Following feature engineering, selection is paramount. We've used a correlation matrix to identify highly correlated features, eliminating redundancy to focus on foundational aspects directly influencing wine quality. By removing highly correlated features (abs(correlation) > 0.7), like totalAcidity and acidAlcoholRatio, we maintain a robust yet simplified dataset, free of multicollinearity issues. Retaining original measures such as fixedAcidity ensures our model remains grounded in measurable wine properties, aiding effective and interpretable predictive modeling. Additionally, standardization and categorical variable conversion ensure seamless model integration.

Moreover, dimensionality reduction via PCA captures most dataset variance with fewer components, streamlining analysis while preserving information. Compressing data into nine principal components, retaining 90% variance, enhances computational efficiency without sacrificing predictive accuracy. However, this entails some interpretability loss as principal components represent linear feature combinations. Nevertheless, our strategies empower effective utilization of wine data richness, facilitating robust and insightful modeling outcomes.

## Modeling

**Decision Tree (DT)**

Red Wine: The Decision Tree classifier achieved a cross-validation accuracy of 0.58. It exhibited moderate performance across different quality classes, with varying precision, recall, and F1-score metrics. Class 0 (low-quality) showed the highest precision and recall, indicating a relatively strong ability to identify instances of low-quality wine. However, class 2 (high-quality) had lower values, suggesting challenges in accurately classifying high-quality wine instances. Overall, the model achieved an accuracy of 0.58, implying moderate predictive performance.

White Wine: For the white wine dataset, the Decision Tree classifier attained a cross-validation accuracy of 0.54. It demonstrated slightly lower performance compared to red wine, with class 1 (medium-quality) exhibiting the best precision and recall values. However, similar to the red wine dataset, class 2 (high-quality) showed relatively lower precision and recall scores, indicating potential difficulties in accurately identifying high-quality instances. Further optimization may be necessary to improve its ability to distinguish between different quality classes.

**Random Forest (RF)**

Red Wine: The Random Forest classifier achieved a cross-validation accuracy of 0.63, indicating moderate predictive performance. While class 0 (low-quality) demonstrated high precision and recall, class 2 (high-quality) exhibited lower values, suggesting challenges in accurately classifying high-quality instances. The overall accuracy of the model was 0.60, reflecting its ability to make relatively accurate predictions across different quality categories.

White Wine: For white wine, the Random Forest classifier obtained a cross-validation accuracy of 0.61. Similar to red wine, it showed varying precision, recall, and F1-score values across different quality classes, indicating room for improvement in accurately identifying high-quality instances. Despite this, the model demonstrated moderate predictive performance with an overall accuracy of 0.60.

**Stochastic Gradient Descent (SGD)**

Red Wine: The SGD classifier yielded a cross-validation accuracy of 0.80 for red wine, demonstrating robust performance. It exhibited strong precision and recall for class 0 (low-quality) but faced challenges in accurately classifying instances of high-quality wine (class 2). The overall accuracy of the model was 0.59, indicating its ability to make accurate predictions across different quality categories.

White Wine: For white wine, the SGD classifier achieved a high cross-validation accuracy of 0.93. Despite this, it showed varying precision and recall values across different quality classes, suggesting potential areas for improvement, particularly in correctly classifying instances of high-quality wine. Nevertheless, the model demonstrated strong predictive performance, with an overall accuracy of 0.53.

**Support Vector Classifier (SVC)**

Red Wine: The SVC attained a cross-validation accuracy of 0.65 for red wine, indicating its effectiveness in generalizing to unseen data. It demonstrated commendable precision and recall for class 0 (low-quality) and class 2 (high-quality), with overall accuracy of 0.62. The pipeline comprising standard scaling and the SVC classifier facilitated efficient model training and enhanced predictive performance.

White Wine: For white wine, the SVC achieved a cross-validation accuracy of 0.61. While it showed good precision and recall for class 0 (low-quality) and class 1 (medium-quality), there were challenges in accurately classifying instances of high-quality wine (class 2). Nevertheless, the model demonstrated satisfactory predictive performance, with an overall accuracy of 0.61. Further optimization may be necessary to enhance its ability to accurately predict the quality of white wine based on chemical attributes.

# Recommendations

Based on the analysis of various machine learning models applied to both red and white wine datasets, the choice of model should prioritize the F1-score, as it provides a balanced measure of precision and recall. Considering this criterion and the performance of different models:

**For red wine:**

* The Decision Tree (DT) classifier demonstrates the highest F1-score for classifying low-quality red wine (F1-score: 0.68), indicating robust performance in correctly identifying instances of this category.
* Additionally, the DT classifier maintains competitive F1-scores for medium and high-quality red wine (F1-score: 0.51 and 0.42, respectively).
* Moreover, the Decision Tree model offers good explainability, making it easier to interpret the factors contributing to wine quality predictions.
* Therefore, for red wine, the Decision Tree (DT) classifier is recommended due to its higher F1-scores and good explainability.

**For white wine:**

* The Random Forest (RF) classifier demonstrates the highest F1-score for classifying medium-quality white wine (F1-score: 0.61), indicating robust performance in correctly identifying instances of this category.
* Moreover, the RF classifier achieves competitive F1-scores for low and high-quality white wine (F1-score: 0.67 and 0.45, respectively).
* Although Random Forest models may offer slightly less explainability compared to Decision Trees, their superior predictive performance makes them a suitable choice for white wine classification.

Considering the overall performance and F1-scores across quality categories, the Random Forest (RF) classifier is recommended for white wine classification, while the Decision Tree (DT) classifier is preferred for red wine, based on their respective F1-scores and the interpretability of the Decision Tree model.

VinifyTech can enhance its predictive modeling by refining hyperparameters, feature selection, and ensemble techniques to improve accuracy, especially in classifying high-quality wine instances. Techniques like partial dependence plots and SHAP values can enhance model interpretability, providing deeper insights into wine quality predictors. Integrating domain knowledge from winemakers can enrich feature engineering and deepen understanding of wine characteristics. Continuous data collection, monitoring, and validation are crucial for model relevance and robustness. Investing in infrastructure and talent supports data analytics initiatives, fostering innovation and collaboration. Prioritizing user feedback and tailoring solutions can drive adoption and loyalty, enhancing the wine experience. By embracing these strategies, VinifyTech can unlock data-driven winemaking's full potential, driving innovation and excellence in the industry.