Wine Quality

Sindre Moene, 15.11.24

# **DESCRIBE THE PROBLEM**

## **SCOPE**

**Scope**  
The purpose of this project is to predict wine quality based on chemical properties using machine learning. This can help wineries and retailers improve their processes and offer consistent quality to customers. Instead of relying on manual tasting or basic chemical tests, this solution will provide an automated and reliable way to determine quality.

**Business objective**  
The project’s main goal is to create a tool that predicts wine quality accurately, saving time and improving decision-making for producers and retailers.

**Existing solutions**  
Currently, wine quality is often assessed by expert tasters or lab tests. While effective, these methods can be subjective, time-consuming, or expensive. A machine learning model can offer faster and more consistent results.

**Metrics**  
The performance of the solution will be evaluated based on accuracy and F1-score to ensure it performs well in real-world scenarios.

**System overview**  
The project will include a Flask-based web app where users can input wine data to get instant predictions. The app will integrate the machine learning model with a user-friendly interface.

**Key stakeholders**

* Wineries: To standardize wine quality evaluation
* Retailers: To categorize and market products better
* Consumers: To easily identify quality wines

**Project timeline**

* Week 1: Data cleaning, exploratory analysis, and initial feature engineering
* Week 2: Train and evaluate baseline models (e.g., Logistic Regression, Decision Tree)
* Week 3: Optimize the model and select the best-performing algorithm
* Week 4: Develop the Flask web application, test the system, and deploy it

**Required resources**

* Team: Just me
* Computational power: Local system or cloud-based if necessary
* Data: Provided wine dataset

## **METRICS**

**Success criteria**  
The model should achieve a minimum F1-score of 0.8 to be considered successful.

**Metrics to track**

* Accuracy and F1-score to gauge prediction performance
* Response time to ensure the app is fast and user-friendly

**Connection to business goals**  
The app must deliver reliable and fast predictions to improve productivity and ease of use for its stakeholders.

# **DATA**

**Dataset overview**  
The dataset includes chemical properties of red and white wine, such as acidity, residual sugar, and alcohol, alongside quality ratings.

**Source**  
The dataset is publicly available and already labeled with quality scores.

**Data preparation**

* Check for missing values and remove or impute them.
* Normalize features like sugar content and acidity to avoid scaling issues.
* Identify and remove outliers to improve model accuracy.

**Ethical considerations**  
Since the data doesn’t include personal information, there are no privacy concerns.

# **MODELING**

**Approach**  
I’ll start with simple models like Logistic Regression to set a baseline. More advanced models, such as Random Forest or Gradient Boosting, will then be trained and evaluated.

**Evaluation process**  
I’ll use cross-validation to ensure the models are robust. By analyzing mistakes and focusing on feature importance, I’ll refine the model for better results.

# **DEPLOYMENT**

**Deployment plan**  
The model will be integrated into a Flask app. Users will input wine data into the app, which will return a quality prediction instantly.

**Monitoring**  
The app will include logging to monitor usage and prediction accuracy. If the app shows a decline in accuracy, I will update the model with fresh data.

# **REFERENCES**

* Data: <https://archive.ics.uci.edu/dataset/186/wine+quality>
* For better understanding: <https://rpubs.com/cw14/stat-272>, <https://www.analyticsvidhya.com/blog/2021/04/wine-quality-prediction-using-machine-learning/>

# **Reflection:**

For this project, I initially planned to create a Norwegian house price estimator, using datasets from Statistisk sentralbyrå (SSB) to predict property prices based on various economic and demographic factors. The concept was to gather housing-related data, such as average income levels, population density, property type, and other regional indicators, to create a robust model for estimating house prices across different areas of Norway. However, this task quickly became challenging due to inconsistencies within the datasets. The data from SSB had varying formats, column names, and levels of granularity, making it extremely difficult to merge and create a unified dataset. Many columns with similar information were labeled differently, and crucial data was often missing or inconsistently structured across files. Merging and aligning these datasets to form a cohesive and usable dataset took an unexpectedly large amount of time and effort, leading me to shift my focus to a more manageable project. This experience highlighted the importance of data consistency and the complexities involved in real-world data integration.