Predicting Wine Quality with Spark MLlib

Group Info
Group 5
Arnav Chawla
Guadalupe Ramirez Lara

Project Objectives
This project analyzes how physicochemical attributes influence wine quality for both red and white wines. We used Spark MLlib to perform regression modeling, explore trends, and discover correlations. We aimed to identify influential features (e.g., alcohol, sulphates, density) that impact quality scores. We visualized patterns to reveal useful insights for wine producers and consumers.

Dataset Metadata

Property	Red Wine	White Wine
Rows	1,599	4,898
Columns	11 + 1 (quality)	11 + 1 (quality)
Target	`quality` (int)	`quality` (int)
Source	UCI ML Repo	UCI ML Repo
Alcohol Mean	10.42%	10.51%
Quality Range	3 to 8	3 to 9
Format	CSV	CSV

Summary of Findings

Using Spark MLlib, we trained linear regression models for red, white, and combined wine datasets.
- Alcohol showed the strongest positive correlation with quality across both datasets ($r \approx 0.44–0.47$).
- Density and volatile acidity negatively impacted quality scores.
- The regression model explained 35% of the variance ($R^2 \approx 0.35$) in red wine and ~29% in white.
- These results indicate that a few chemical attributes heavily influence perceived wine quality.

Tools & Commands Used

Tool/Command	Purpose
awk, head, wc	Extract row/column counts and means
summary_stats.sh	Shell script for alcohol stats
PySpark (MLlib)	Regression analysis with VectorAssembler, LinearRegression()
wine_spark_analysis.py	Runs analysis for red, white, or both
seaborn, matplotlib	Visualizations (heatmap, boxplot)

Visual Plots

1. Correlation Heatmap (Combined Wine Data)
   - Shows positive/negative relationships between all variables.

2. Boxplot of Alcohol vs. Quality
   - Reveals trend that higher alcohol content generally predicts higher wine ratings.