LAPORAN ANALISIS KLASIFIKASI: Memprediksi Kualitas White Wine

Disusun oleh : M. Bayu Chandra Adi

Nim : A11.2022.14666

Audiens: Chief Data Officer (CDO)

Tanggal: 11 November 2025

1. Tujuan Utama Analisis

Tujuan utama dari analisis ini adalah untuk **memprediksi** dan **menginterpretasi** faktor-faktor yang menentukan kualitas *white wine*. Analisis ini berfokus pada dua hal:

- **Prediksi (Manfaat Bisnis):** Membangun model klasifikasi yang mampu memprediksi kategori kualitas wine (Rendah, Sedang, Tinggi) berdasarkan 11 fitur kimia-fisiknya. Ini dapat digunakan untuk mengotomatisasi proses *Quality Control* (QC), membantu standardisasi, dan memberikan peringatan dini untuk *batch* yang berpotensi berkualitas buruk.
- **Interpretasi (Manfaat Wawasan):** Mengidentifikasi faktor kimia apa (misalnya, alkohol, keasaman, gula) yang menjadi pendorong utama kualitas wine. Wawasan ini sangat berharga bagi tim produksi dan *wine-maker* untuk menyesuaikan proses demi mencapai kualitas yang diinginkan.

2. Deskripsi Singkat Data

- Dataset: winequality-white.csv
- Ukuran: 4.898 sampel (baris) dan 12 kolom.
- Variabel Target (y): quality (skor numerik asli dari 3 hingga 9).
- Variabel Fitur (X): 11 fitur kimia-fisik, termasuk alcohol, volatile acidity, residual sugar, density, dan pH

3. Ringkasan Eksplorasi Data (EDA) dan Pembersihan

Analisis eksplorasi awal mengungkapkan dua temuan kunci:

1. Kualitas Data: Data sangat bersih dan berkualitas tinggi. Tidak ada nilai yang hilang (*missing values*) di seluruh 4.898 sampel, sehingga tidak diperlukan imputasi.
2. Ketidakseimbangan Kelas (Flaw): Variabel target quality sangat tidak seimbang. Mayoritas wine memiliki skor '6' (2.198 sampel) dan '5' (1.457 sampel), sementara skor ekstrem seperti '3' (20 sampel) dan '9' (5 sampel) hampir tidak ada.

Tindakan yang Diambil (Feature Engineering): Untuk membuat model yang lebih stabil dan bermakna secara bisnis, saya mengubah target 7 kelas ini menjadi 3 kategori yang lebih seimbang:

- 'Rendah' (Skor 3, 4, 5) - 1.640 sampel
- 'Sedang' (Skor 6) - 2.198 sampel
- 'Tinggi' (Skor 7, 8, 9) - 1.060 sampel

Semua 11 fitur numerik juga telah di-scaling (distandarisasi) untuk memastikan model seperti Regresi Logistik bekerja secara optimal.

4. Ringkasan Pelatihan Model

Tiga model klasifikasi yang berbeda sifatnya telah dilatih dan diuji menggunakan *test-split* (30%) yang sama untuk perbandingan yang adil.

- Regresi Logistik (Baseline): Model ini berfungsi sebagai *baseline* yang sederhana dan mudah diinterpretasi. Model ini mencapai akurasi keseluruhan 57.55%. Performanya paling lemah, terutama dalam memprediksi kelas 'Tinggi' (Recall hanya 0.37).
- Decision Tree: Model non-linear ini menunjukkan peningkatan signifikan, mencapai akurasi 63.47%. Model ini lebih seimbang, dengan F1-score 0.67 (Rendah), 0.63 (Sedang), dan 0.59 (Tinggi).
- Random Forest (Ensemble): Model *ensemble* ini memberikan performa terbaik secara keseluruhan dengan akurasi 71.97%. Yang terpenting, model ini paling seimbang dalam memprediksi ketiga kelas (F1-score 0.73, 0.72, 0.70).

Model	Akurasi	F1-Score (Rendah)	F1-Score (Sedang)	F1-Score (Tinggi)
Regresi Logistik	57.55%	0.58	0.61	0.47
Decision Tree	63.47%	0.67	0.63	0.59
Random Forest	**71.97%**	**0.73**	0.72	0.70

5. Rekomendasi Model

Saya merekomendasikan Random Forest sebagai model akhir.

Model ini memberikan keseimbangan terbaik antara akurasi prediktif dan keandalan di semua kelas. Dengan akurasi 72%, model ini secara signifikan mengungguli *baseline* dan memberikan F1-score yang konsisten di atas 0.70 untuk ketiga kategori. Untuk tujuan QC, memiliki model yang andal dalam mengidentifikasi *semua* kelas (termasuk 'Rendah' dan 'Tinggi') lebih penting daripada akurasi mentah, dan Random Forest adalah pemenangnya di sini.

6. Temuan Kunci & Wawasan (dari Random Forest)

Analisis *feature importance* dari model Random Forest mengungkap faktor-faktor pendorong utama kualitas wine:

1. alcohol (Alkohol): Sejauh ini, ini adalah prediktor terpenting. Kandungan alkohol memiliki korelasi positif yang kuat dengan kualitas yang lebih tinggi.
2. density (Kepadatan): Faktor terpenting kedua. Kepadatan seringkali terkait dengan kandungan gula dan alkohol.
3. volatile acidity (Keasaman Menguap): Faktor penting ketiga, yang sering dianggap sebagai indikator negatif (cacat wine jika terlalu tinggi).
4. free sulfur dioxide (Sulfur Dioksida Bebas): Berperan sebagai pengawet.
5. residual sugar (Sisa Gula): Menunjukkan tingkat kemanisan.

Wawasan Kunci untuk Stakeholder: Temuan ini memberikan wawasan yang dapat ditindaklanjuti: alcohol, density, dan volatile acidity adalah tiga fitur teratas yang harus difokuskan oleh tim produksi. Yang menarik, fitur seperti pH dan citric acid (keasaman sitrat) ternyata memiliki dampak yang lebih kecil dibandingkan yang diperkirakan banyak orang.

7. Saran dan Langkah Selanjutnya (Flaws & Action Plan)

Meskipun model Random Forest kami adalah yang terbaik, akurasi 72% menunjukkan masih ada ruang untuk perbaikan.

- Kelemahan (Flaw): Model ini masih "ragu-ragu". Meskipun F1-score-nya seimbang, *recall* untuk kelas 'Tinggi' (0.66) dan 'Rendah' (0.70) masih lebih rendah daripada kelas 'Sedang' (0.76). Ini berarti model ini masih cenderung keliru memprediksi kelas minoritas sebagai kelas mayoritas 'Sedang'. Ini disebabkan oleh sisa *class imbalance* (kelas 'Sedang' masih 33% lebih besar dari kelas 'Rendah').
- Rencana Aksi / Langkah Selanjutnya:
    1. Menyeimbangkan Data (Oversampling): Langkah selanjutnya yang paling logis adalah menerapkan teknik *oversampling* seperti SMOTE (*Synthetic Minority Over-sampling Technique*) pada data latih. Ini akan membuat "sampel sintetis" untuk kelas 'Rendah' dan 'Tinggi' agar jumlahnya setara dengan kelas 'Sedang', yang kemungkinan besar akan meningkatkan *recall* untuk kedua kelas tersebut.
    2. Model yang Lebih Canggih: Jika SMOTE tidak cukup, kita dapat mencoba model *ensemble* yang lebih canggih seperti XGBoost (*Gradient Boosting*), yang seringkali sangat baik dalam menangani dataset yang tidak seimbang dan dapat meningkatkan akurasi beberapa poin persentase lagi.