UNIVERSITAS INDONESIA

Faktor Determinasi Kualitas Anggur Merah

dengan Menggunakan Pemodelan Logit dan Probit

Paper ini Disusun untuk Memenuhi Tugas Akhir

Mata Kuliah Ekonometrika Cross Section dan Panel Data

LUDY HASBY AULIA

2106716912

FAKULTAS EKONOMI DAN BISNIS

PROGRAM STUDI ILMU EKONOMI

DEPOK

2023

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR DAN LAMPIRAN

STATEMENT OF AUTHORSHIP

"Saya yang bertandatangan dibawah ini menyatakan bahwa tugas terlampir adalah murni hasil pekerjaan saya sendiri. Tidak ada pekerjaan orang lain yang saya gunakan tanpa menyebutkan sumbernya.

Materi ini tidak/belum pernah disajikan/digunakan sebagai bahan untuk tugas pada mata ajaran lain kecuali saya menyatakan dengan jelas bahwa saya menyatakan menggunakannya.

Saya memahami bahwa tugas yang saya kumpulkan ini dapat diperbanyak dan atau dikomunikasikan untuk tujuan mendeteksi adanya plagiarisme."

Nama	: Ludy Hasby Aulia
NPM	: 2106716912
Tandatangan	: 
Mata Kuliah	: Ekonometrika Cross Section dan Panel Data
Judul Makalah	: Faktor Determinasi Kualitas Anggur Merah
Tanggal	: 21 Juni 2023
Dosen	: Nurkholis, S.E., M.S.E.

Faktor Determinasi Kualitas Anggur Merah

dengan Menggunakan Pemodelan Logit dan Probit

"Quality is never an accident; it is always the result of high intention, sincere effort, intelligent direction, and skillful execution." - *William A. Foster*

Abstrak- Sertifikasi kualitas produk diyakini merupakan kunci untuk meningkatkan penjualan. Hal itu dikarenakan kualitas produk merupakan suatu hal terpenting bagi keberlanjutan bisnis yaitu dengan memuaskan konsumen. Karena manusia memiliki preferensinya masing-masing, pengujian kualitas suatu produk menjadi sesuatu yang mahal dan tidak efisien pada sebelumnya. Penelitian ini penulis tujukan untuk menganalisis faktor yang berpengaruh terhadap kualitas anggur merah dengan membandingkan pendekatan model logit dan probit. Penelitian ini menggunakan dataset yang disediakan oleh UCI *Machine Learning Repository* dengan sampel anggur merah Vinho Verde dari Portugal Utara. Penelitian ini akan menggunakan bantuan program statistik Stata untuk pengolahan dan analisis data. Untuk menganalisis faktor yang berpengaruh pada kualitas anggur merah yang merupakan hasil dari data sensori, penelitian melibatkan sebelas fitur dengan masing-masing terdiri atas 1599 baris yang disediakan berdasar uji fisikokimia (*physicochemical tests*), terdiri atas fixed acidity (x_1), volatile acidity (x_2), citric acid (x_3), residual sugar (x_4), chlorides (x_5), free sulfur dioxide (x_6), total sulfur dioxide (x_7), density (x_8), ph (x_9), sulphates (x_{10}), dan alcohol (x_{11}). Penelitian ini menggunakan beberapa tahapan analisis, yaitu analisis deskriptif/univariate, melakukan *data preprocessing,* pemodelan, dan evaluasi model. Analisis deskriptif digunakan untuk mendapat gambaran karakteristik data, kemudian dilakukan treatment untuk mendapat hasil estimasi yang lebih baik melalui *data preprocessing*. Hasil penelitian mendapatkan kesimpulan bahwa variabel independen terbukti mempengaruhi variabel dependen dengan empat variabel yang signifikan setelah melalui uji Wald, yaitu (-) jumlah asam / *volatile acidity* (x_2), (-) total sulfur dioksida (S02) (x_7), (+) aditif anggur (sulphates) (x_{10}), dan (+) persen kandungan alkohol (x_{11}). Selanjutnya, Setelah melakukan beberapa uji evaluasi model, antara lain LR test, APER (*Apparent Error Rate)*, McFadden R^2 , AIC, dan BIC didapatkan kesimpulan penulis bahwa model probit pada kasus penentuan faktor determinasi kualitas anggur merah dinilai lebih baik daripada logit walaupun dengan mempertimbangkan kompleksitas model logit mendapat nilai BIC dan AIC yang lebih kecil.

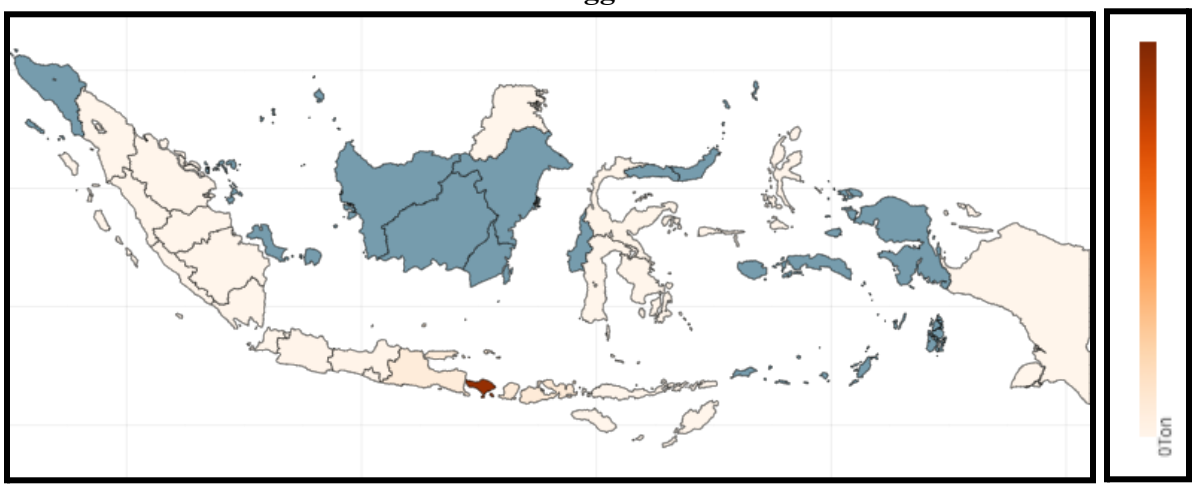
Kata kunci: anggur merah, kualitas, logit, probit, faktor determinasi

I. PENDAHULUAN

Ekonometrika meruapakan cabang ilmu ekonomi yang menggabungkan teori ekonomi dan metode statistik untuk memahami dan menganalisis hubungan kausal antara variabel-variabel yang saling terkait. Dalam ekonometrika, statistika digunakan untuk mengolah dan menganalisis data dalam rangka mendapat informasi yang relevan dan dapat dipercaya tentang hubungan antara variabel-variabel ekonomi. Beberapa metode statistika yang mampu menangani data dengan variabel respon bersifat kualitatif adalah model logit dan probit. Salah satu contoh kasus data kualitatif adalah status kerja, dimana terdapat dua kategori status kerja yaitu bekerja dan tidak bekerja (Ayu, 2019).

Anggur merah merupakan anggur hasil fermentasi yang berasal dari pigmen natural pada kulit anggur yang berwarna gelap (BRC, 2023). Berdasar The Business Research Company, CAGR (*Compound Annual Growth Rate)* 2022-2023 dari anggur merah ini sebesar 6,3% dimana di tahun 2022 *market size* secara global sebesar $102,97 Miliar dan $109,5 Miliar di tahun 2023. Pada pasar anggur merah, Amerika Utara dan Eropa merupakan pemain besar dimana besar pasar mereka dari global masing-masing 40% dan 32% di tahun 2022 menurut Red Wine Market Outlook Fact.MR's 2022-2023. Di Indonesia sendiri, walaupun sebagai negara tropis, anggur tetap bisa berkembang. Menurut BPS (Badan Pusat Statistik) 2021, produksi anggur mencapai 12.164 ton dengan CAGR sebesar 2,2% dibanding 2020. Seperti dijelaskan pada Gambar 1.1, Bali merupakan provinsi dengan hasil produksi anggur terbesar, yaitu 10.234 ton (84% dari pasar produksi Indonesia).

Gambar 1.1 Distribusi Produksi Anggur di Indonesia



Sumber: Badan Pusat Statistik, diolah oleh DataIndonesia.id

Kualitas anggur adalah hal yang sangat penting bagi konsumen sebagaimana industri manufaktur, Dengan sertifikasi kualitas produk, industri anggur mengalami kenaikan pada penjualan mereka (Kothawade, 2021). Sebelumnya, pengujian kualitas produk akan dilakukan pada akhir dari produksi sehingga akan memakan waktu dan sumber daya manusia yang mahal untuk membuat penilaian. Karena setiap manusia mempunyai preferensinya masing-masing, identifikasi kualitas anggur merupakan tugas yang menantang. Ada beberapa metode untuk memprediksi kualitas anggur merah, seperti dijelaskan sebelumnya, probit dan logit dapat membantu untuk membuat prediksi faktor apa sajakah yang signifikan dalam menentukan kualitas anggur merah.

Penelitian mengenai faktor determinasi kualitas anggur merah pernah dilakukan dengan berbagai pemodelan. Pemodelan tersebut diantaranya dengan Decision Tree, Random Forest, ANN (Artificial Neural Network), SVM (Support Vector Machine), Logistic Regression, XGBoost yang menggunakan *supervised learning models* dengan *library* yang disediakan pada bahasa pemrograman Python. penelitian tersebut umumnya terfokus pada evaluasi model terbaik pada klasifikasi anggur merah. Kesimpulan penelitian tersebut berbeda-beda sesuai dengan *treatment* pada data yang dilakukan. Tampaknya, penelitian dengan menggunakan model Probit dengan kasus tersebut belum pernah dilakukan.

Berdasarkan uraian tersebut, maka penulis ingin melakukan penelitian yakni mencari tahu faktor apa sajakah yang signifikan mempengaruhi kualitas anggur merah dan membandingkan model logit dan probit dengan menggunakan program statistik stata pada studi kasus tersebut. Penelitian ini akan menggunakan dataset yang disediakan pada repositori UCI Machine Learning (Cortez et al., 2009) dengan sampel anggur merah Vinho Verde, dari Portugal Utara. Dataset ini terdiri atas sebelas fitur input dan satu fitur output. Sebelas fitur input tersebut berdasarkan uji fisikokimia (*physicochemical tests*) dan fitur output nya berdasarkan data sensori yang berskala 11, kualitas 0 sampai 10 (0 berarti sangat buruk dan 10 berarti sangat baik).

II. Tinjauan Pustaka

A. Tabel Kontingensi

Tabel kontingensi digunakan untuk menguji hubungan dua variabel kategorik. Semakin banyak kategori dari variabel, sampel yang dibutuhkan akan semakin banyak. Hal itu dikarenakan tabel kontingensi memiliki syarat homogen atau setiap entri harus berupa objek

yang sama. Selain itu ada syarat *mutually exclusive,* yaitu saling melengkapi antara level satu dengan level lain dan syarat mutually exhaustive, yaitu dari level terkecil dekomposisinya lengkap.

Tabel 1. Tabel Kontingensi rxc

Row	Column			
	1	2	..	c
1	n_{11}	n_{11}	..	n_{11}
2	n_{11}	n_{11}	..	n_{11}
:	:	:	:	:
r	n_{11}	n_{11}	..	n_{11}

B. Generalized Linear Model

GLM digunakan untuk memodelkan hubungan variabel dependen yang berdistribusi tidak normal dengan variabel independen. GLM menggabungkan tiga komponen utama.

1. Komponen random

   menentukan distribusi probabilitas dari variabel dependen/respon. Seperti distribusi binomial untuk regresi logistik biner (STAT 504, …)

2. Komponen sistematik

   menentukan variabel independen pada model (x_1, x_2, …., x_3).

3. Fungsi Link

   menentukan hubungan antara komponen random dengan komponen sistematik. Ini akan menunjukkan bagaimana nilai yang diharapkan dari dependen berhubungan dengna kombinasi linear variabel independen. Sebagai contoh $\eta = g(E(Yi)$ pada regresi klasik, $\eta = log(\frac{\pi}{1-\pi})$ = logit(π) pada regresi logistik. Dalam GLM dengan fungsi link logit untuk variabel respon yang berdistribusi binomial, terdapat beberapa asumsi yang harus dipenuhi, antara lain (Hardin, 2007).

- Bentuk fungsi link yang tepat, umumnya fungsi link yang digunakan adalah logit
- Nilai p berada dalam rentang ($0 \leq p \leq 1$)
- Nilai η berada dalam rentang ($-\infty < \eta < \infty$), artinya nilai kombinasi linear prediktor mengambil nilai positif atau negatif yang tidak terbatas.

Fungsi link yang memenuhi asumsi-asumsi itu dalam GLM dengan variabel dependen berdistribusi binomial, beberapa diantaranya:

1. Fungsi logistik (logit), $\eta = log(\frac{p}{1-p})$ dan probabilitas logit, $p = \frac{e^{\eta}}{1+e^{\eta}}$.
2. Fungsi probit, dengan menggunakan invers fungsi invers kumulatif normal standar, $\eta = \Phi^{-1}(p)$ dan probabilitas probit, $p = \Phi(\eta)$

C. Metode Maximum Likelihood Estimation (MLE)

Metode MLE umumnya digunakan pada estimasi parameter model yang distribusinya diketahui dengan variabel dependennya (Y) diasumsikan data biner dengan distribusi bernoulli (Mccullagh, 1983). MLE untuk θ, yang dinotasikan $\hat{e}_{MLE}$ merupakan θ yang memaksimumkan likelihoodnya (Greene, 2005). Jika dari fungsi likelihood tersebut diperoleh bentuk yang tidak *closed form*, estimasi maksimum likelihood akan menggunakan metode numerik iterasi Newton Raphson (Mccullagh, 1983).

D. Uji Signifikansi Parameter

Setelah melakukan estimasi parameter, dilakukan uji signifikansi parameter untuk menentukan variabel yang memiliki pengaruh signifikan terhadap variabel respon. Uji signifikansi parameter terdiri dari dua tahap, yaitu uji signifikansi secara serentak dan uji signifikansi secara parsial (Ratnasari & Putri, 2015).

1. Uji signifikansi parameter secara serentak digunakan untuk mengetahui apakah secara bersama-sama variabel-variabel memiliki pengaruh signifikan terhadap variabel respon. Hipotesis yang diuji adalah:
   - H0: $\beta1 = \beta2 = ... = \beta n = 0$ (Tidak ada variabel yang berpengaruh secara signifikan)

- H1: Minimal ada satu βj ≠ 0 (Setidaknya satu variabel berpengaruh secara signifikan)

Statistik uji yang digunakan adalah likelihood ratio test (LRT) (LR chi2)

LR = -2ln($\frac{L(m1)}{L(m2)}$), dengan m1 adalah model reduksi dan m2 adalah model penuh.

Daerah kritis : Tolak H0 jika LR > $\chi^2_{(\propto,v)}$.

2. Uji signifikansi parameter secara parsial digunakan untuk mengetahui apakah setiap variabel memiliki pengaruh signifikan terhadap variabel respon secara individu. Hipotesis yang diuji untuk masing-masing variabel adalah:
   - H0: βj = 0 (Variabel j tidak memiliki pengaruh signifikan)
   - H1: βj ≠ 0 (Variabel j memiliki pengaruh signifikan)

   Statistik uji yang digunakan adalah statistik W,

   W = $\frac{\widehat{\beta}_i}{SE(\widehat{\beta}_i)}$ dan $SE(\widehat{\beta}_i) = \sqrt{(\sigma^2(\widehat{\beta}_i))}$

   Daerah kritis : Tolak H0 jika $|W| > Z_{\alpha/2}$.

E. *Apparent Error Rate* (APER)

Prosedur klasifikasi digunakan untuk mengevaluasi tingkat keakuratan suatu fungsi klasifikasi dengan melihat peluang kesalahan klasifikasi. Salah satu ukuran yang umum digunakan untuk mengukur tingkat kesalahan adalah Apparent Error Rate (APER). APER menggambarkan proporsi sampel yang salah diklasifikasikan oleh fungsi klasifikasi (Agresti, 2002). APER dapat dihitung dengan

APER(%) = $\frac{FP + FN}{TP + TN + FP + FN}$, sebaliknya 1 -APER(%) adalah ukuran ketepatan klasifikasi (%).

F. Pemilihan Model Terbaik

Pemilihan model terbaik dalam kasus tertentu dapat dilakukan dengan menggunakan metode AIC (Akaike Information Criterion) dan R-Square McFadden (Ratnasari & Putri, 2015).

Kedua metode ini memberikan indikator yang berguna dalam membandingkan kualitas model dan membantu dalam memilih model yang paling cocok untuk data yang ada.

1. AIC (Akaike Information Criterion)

   AIC adalah suatu metode yang digunakan untuk membandingkan model berdasarkan kualitas fit model dan kompleksitasnya. AIC menggabungkan dua komponen yaitu maksimum fungsi likelihood yang dihasilkan oleh model dan jumlah parameter yang digunakan dalam model. Tujuannya adalah untuk memilih model yang memiliki keseimbangan yang baik antara penjelasan data dan kompleksitas model. Model dengan nilai AIC yang lebih rendah dianggap lebih baik.

2. R-Square McFadden

   R-Square McFadden adalah ukuran relatif yang digunakan untuk mengukur kualitas penjelasan model logit atau probit. R-Square McFadden dihitung dengan,

   R-Square Mcfadden = 1- ln($\frac{L1}{L0}$)

   Nilai R-Square McFadden berkisar antara 0 hingga 1, dimana semakin tinggi nilainya, semakin baik model dalam menjelaskan variasi dalam data. Namun, perlu diingat bahwa R-Square McFadden tidak dapat langsung dibandingkan antara model yang menggunakan fungsi link yang berbeda (misalnya logit dan probit) karena memiliki skala yang berbeda.

III. METODOLOGI PENELITIAN

A. Sumber Data

Penelitian ini akan menggunakan dataset yang disediakan pada repositori UCI Machine Learning (Cortez et al., 2009) dengan sampel anggur merah Vinho Verde, dari Portugal Utara. Dataset ini terdiri atas sebelas fitur input dan satu fitur output. Sebelas fitur input tersebut berdasarkan uji fisikokimia (physicochemical tests) dan fitur output nya berdasarkan data sensori yang berskala 11, kualitas 0 sampai 10 (0 berarti sangat buruk dan 10 berarti sangat baik).

B. Variabel penelitian

Penelitian ini menggunakan dua jenis variabel, yaitu variabel respon dan variabel prediktor. Variabel respon, dalam hal ini, adalah kualitas anggur merah (*quality*). Pada awalnya, variabel ini memiliki tingkatan kategorik, namun untuk keperluan analisis, akan diubah menjadi variabel biner dengan nilai 0 dan 1. Hal tersebut didukung dengan hasil pengujian uji coba penulis dengan estimasi *ordered* dan binary, dan dihasilkan hasil estimasi yang lebih baik pada binary.

Sementara itu, variabel-variabel prediktor digunakan untuk mengetahui faktor-faktor yang berpengaruh terhadap kualitas anggur merah. Variabel prediktor ini disediakan oleh UCI Machine Learning (Cortez et al., 2009) dan terdiri dari berbagai fitur fisikokimia yang terkait dengan anggur merah Vinho Verde. Fitur-fitur ini memberikan informasi yang relevan untuk mengidentifikasi faktor-faktor yang mempengaruhi kualitas anggur merah.

Tabel 2. Variabel Penelitian

Variabel	Kategori
Fixed acidity (x_1)	Kebanyakan asam yang terlibat dengan anggur atau tetap atau tidak mudah menguap (tidak mudah menguap)
	Jenis data *float* / desimal (4,6-15,9)
Volatile acidity (x_2)	Jumlah asam asetat dalam anggur, yang pada kadar terlalu tinggi dapat menyebabkan rasa cuka yang tidak enak
	Jenis data *float* / desimal (0,12-1,58)
Citric acid (x_3)	Ditemukan dalam jumlah kecil, asam sitrat dapat menambah 'kesegaran' dan rasa anggur
	Jenis data *float* / desimal (0-1)
Residual sugar (x_4)	Jumlah gula yang tersisa setelah fermentasi berhenti, jarang ditemukan anggur dengan berat kurang dari 1 gram/liter dan
	Jenis data *float* / desimal (0,9- 15,5)
Chlorides (x_5)	Jumlah garam dalam anggur
	Jenis data *float* / desimal (0,01 - 0,61)
Free sulfur dioxide	Bentuk bebas SO2 ada dalam kesetimbangan antara molekul SO2 (sebagai gas terlarut) dan ion bisulfit; itu mencegah

(x_6)	Jenis data integer (1-72)
Total sulfur dioxide (x_7)	Jumlah S02 bentuk bebas dan terikat; dalam konsentrasi rendah, SO2 sebagian besar tidak terdeteksi dalam anggur, tetapi pada konsentrasi SO2 bebas lebih dari 50 ppm, SO2 menjadi jelas di hidung dan rasa anggur
	Jenis data integer (6-289)
Density (x_8)	Kerapatan air mendekati kerapatan air tergantung pada persen alkohol dan kandungan gula
	Jenis data *float* / desimal (0,99-1)
pH (x_9)	Menjelaskan seberapa asam atau basa anggur dalam skala dari 0 (sangat asam) hingga 14 (sangat basa); kebanyakan anggur antara 3-4 pada skala pH
	Jenis data *float* / desimal (2,74-4,01)
Sulphates (x_{10})	Aditif anggur yang dapat berkontribusi terhadap kadar gas sulfur dioksida (S02), dan yang bertindak sebagai antimikroba
	Jenis data *float* / desimal (0,33-2)
Alcohol (x_{11})	Persen kandungan alkohol anggur
	Jenis data *float* / desimal (8,4-14,9)
Quality (y)	Variabel output (berdasarkan data sensorik, nilai antara 0 dan 10)
	Jenis data integer, *range* data pada (3-8)

C. Langkah-langkah Penelitian

Untuk menyelesaikan permasalahan yang ada, penelitian ini dilakukan dengan mengikuti langkah-langkah berikut.

1. Analisis deskriptif/ analisis univariate, analisis yang akan meliputi visualisasi data untuk mendapatkan gambaran/karakteristik data yang akan diolah. Hal ini berupa gambaran persebaran atau distribusi data, korelasi antar variabel, deskripsi data, dan sebagainya.
2. Pra Pemrosesan data (*Data preprocessing*), data mentah yang didapat akan mendapatkan *treatment* denan cara menghilangkan beberapa permasalahan yang dapat mengganggu

saat pemodelan/pemrosesan data sehingga menghasilkan hasil estimasi yang baik. Tahapan ini akan meliputi pengecekan *outlier,* penghilangan data terpencil, standarisasi data sehingga menyerupai distribusi normal, dan melakukan klasifikasi untuk mendapatkan variabel respon yang diinginkan.

3. Dilakukan perbandingan antara model logit dan probit. Langkah analisis yang dilakukan meliputi.
    - Estimasi parameter
    - Uji signifikansi parameter secara serentak untuk mengetahui apakah terdapat variabel prediktor yang signifikan berpengaruh terhadap variabel respon.
    - Uji signifikansi parameter secara parsial untuk mengetahui apakah terdapat variabel prediktor yang signifikan berpengaruh secara individu.
    - Uji kesesuaian model untuk mengevaluasi sejauh mana model cocok dengan data yang ada.
    - Perhitungan ketepatan klasifikasi untuk mengevaluasi sejauh mana model dapat mengklasifikasikan dengan tepat.

IV. ANALISIS DAN PEMBAHASAN

A. Analisis Deskriptif / Univariate

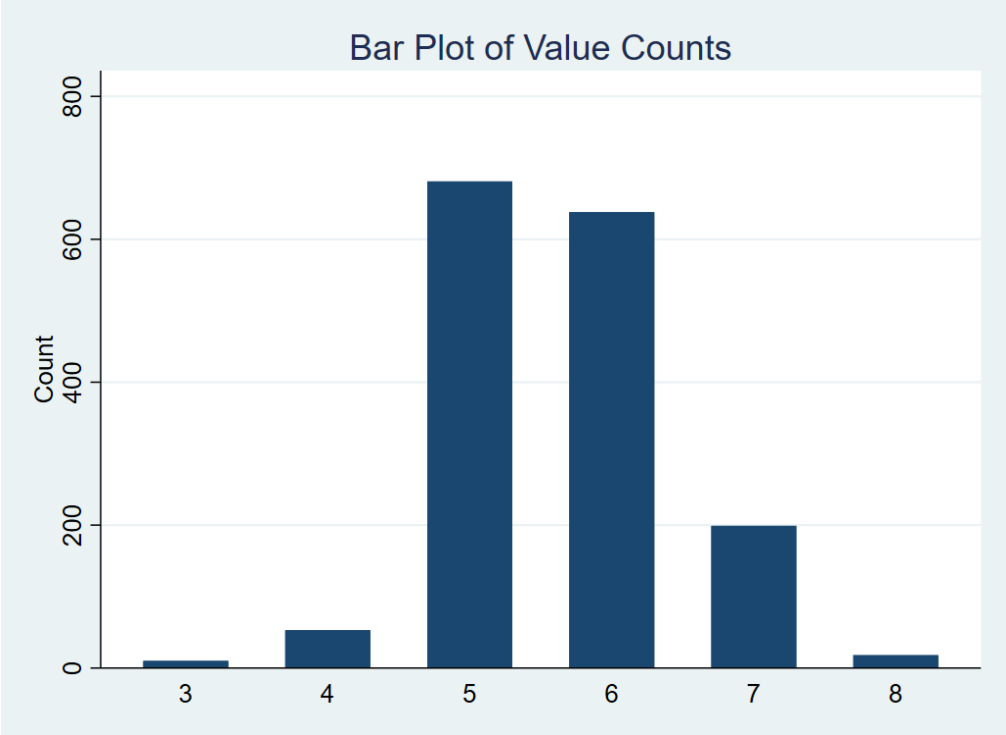
Sebelum melakukan analisis lebih lanjut, Penulis perlu mengetahui ringkasan dari kumpulan data hasil pengukuran sedemikian rupa sehingga penulis dapat mendapatkan informasi yang berguna. Analisis univariate merupakan jenis analisa yang dilakukan untuk menganalisis tiap variabel dengan menggunakan statistik, tabel, grafik (Reyvan, 2021). Analisis univariate ini merupakan tahap awal untuk memahami dataset yang ada. Analisis deskriptif atau univariate ini tidak dapat memberikan kesimpulan pada kita walaupun analisis yang telah dilakukan mencapai sebuah kesimpulan berdasar hipotesis apapun yang telah kita gunakan (Gomes, 2021).

Seperti yang dijelaskan pada bab sebelumnya, penelitian ini mempunyai 12 fitur/kolom dengan menggunakan fitur output yang didapatkan berdasar data sensori, yaitu qualitas anggur merah sebagai variabel dependen dan sebelas fitur lainnya akan menjadi variabel independen, yang meliputi fixed acidity (x_1), volatile acidity (x_2), citric acid (x_3), residual sugar (x_4), chlorides (x_5), free sulfur dioxide (x_6), total sulfur dioxide (x_7), density (x_8), ph (x_9), sulphates

(x_{10}), dan alcohol (x_{11}). Selain itu, dataset ini memiliki 1599 baris sehingga total keseluruhan observasi sebesar 19.188 data.

Variabel dependen *quality* atau kualitas anggur merah terdiri atas 11 kelas dengan *range* 5 dan terlihat memiliki distribusi simetris dengan modus di kelas lima (5), mean (5,6), dan median (6) (Gambar 2).

Gambar 2. Barplot Kualitas Anggur Merah



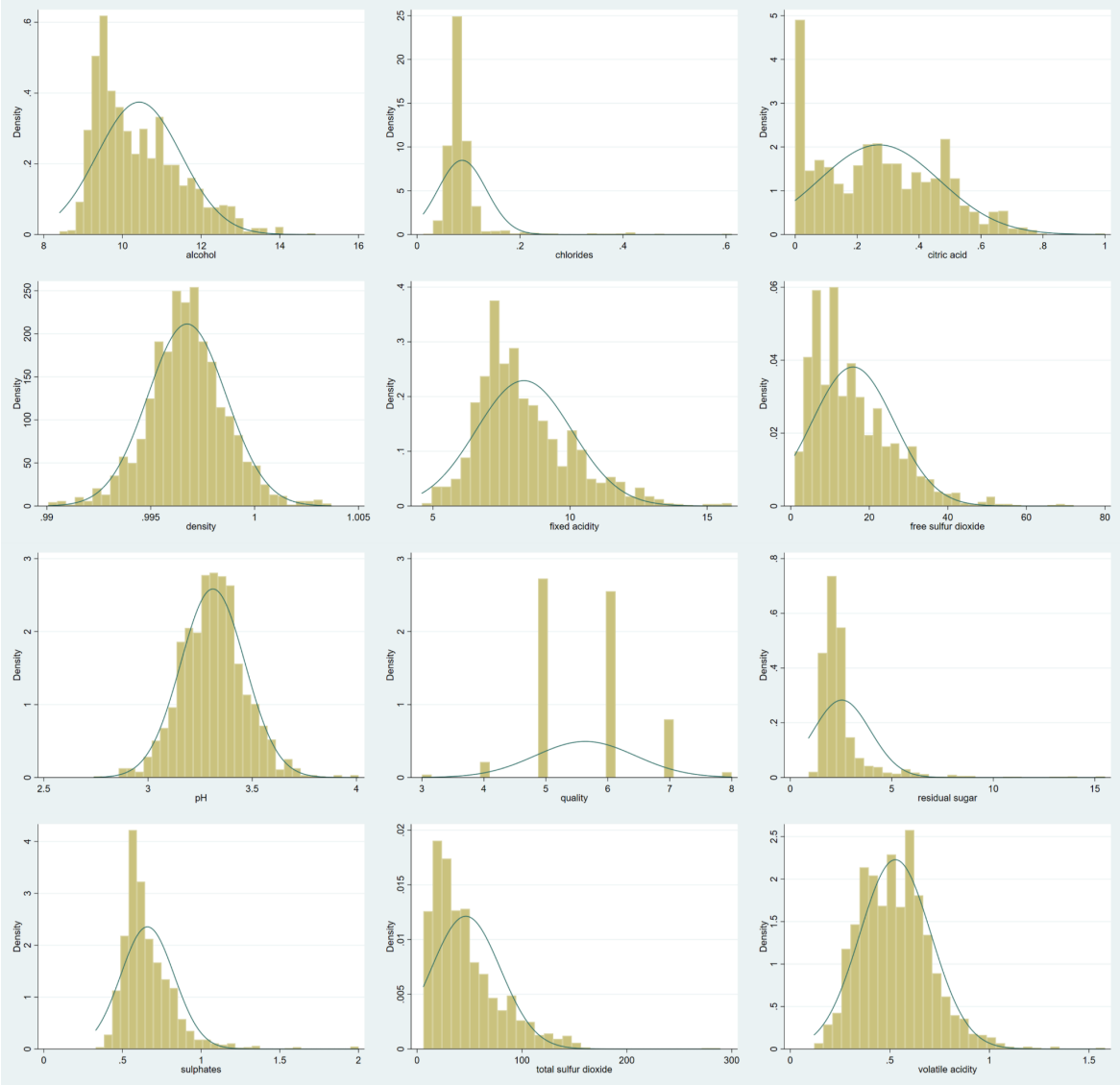
Sumber : UCI Machine Learning Repository, diolah

variabel fixed acidity memiliki range 11,3 dengan rata-rata sebesar 8.31, hal itu menunjukkan data terlihat memiliki penyebaran yang tidak normal terdiavasi dari 11,3/2 = 6,65. Rata-rata, standar deviasi, nilai range (maximum-minimum) selengkapnya dapat dilihat pada tabel 3. Untuk memperoleh gambaran yang lebih jelas penulis akan menggunakan histogram untuk membantu memahami gambaran penyebaran data.

Tabel 3. Deskripsi Statistik Observasi Original

Variable	Obs	Mean	Std. Dev.	Min	Max
fixedacidity	1,599	8.319637	1.741096	4.6	15.9
citricacid	1,599	.2709756	.1948011	0	1
chlorides	1,599	.0874665	.0470653	.012	.611
totalsulfu~e	1,599	46.46779	32.89532	6	289
ph	1,599	3.311113	.1543865	2.74	4.01
alcohol	1,599	10.42298	1.065668	8.4	14.9
volatileac~y	1,599	.5278205	.1790597	.12	1.58
residualsu~r	1,599	2.538805	1.409928	.9	15.5
freesulfur~e	1,599	15.87492	10.46016	1	72
density	1,599	.9967467	.0018873	.99007	1.00369
sulphates	1,599	.6581488	.169507	.33	2
quality	1,599	5.636023	.8075694	3	8

Gambar 3. Histogram Distribusi Variabel



Sumber : UCI Machine Learning Repository, diolah

Dapat dilihat dari Gambar 3, bahwa variabel alcohol, chlorides, citric acid, fixed acidity, free sulfur dioxide, residual sugar, sulphates, total sulfur dioxide, dan volatile acidity memiliki distribusi yang tidak simetris, yaitu condong ke arah kanan (*positive skew*) yaitu kondisi dimana median < mean, observasi data terlihat berkumpul pada ekor kiri. Namun dapat dilihat bahwa variabel density, pH, dan quality terlihat memiliki distribusi yang simetris, dimana median sama dengan nilai mean. Adapun data-data yang memiliki distribusi yang tidak simetris sangat memungkinkan memerlukan *treatment* sehingga menghasilkan estimasi yang lebih baik pada model sebagaimana akan dilakukan pada tahap *preprocessing*.

Tabel 4. Matriks Korelasi antar Variabel

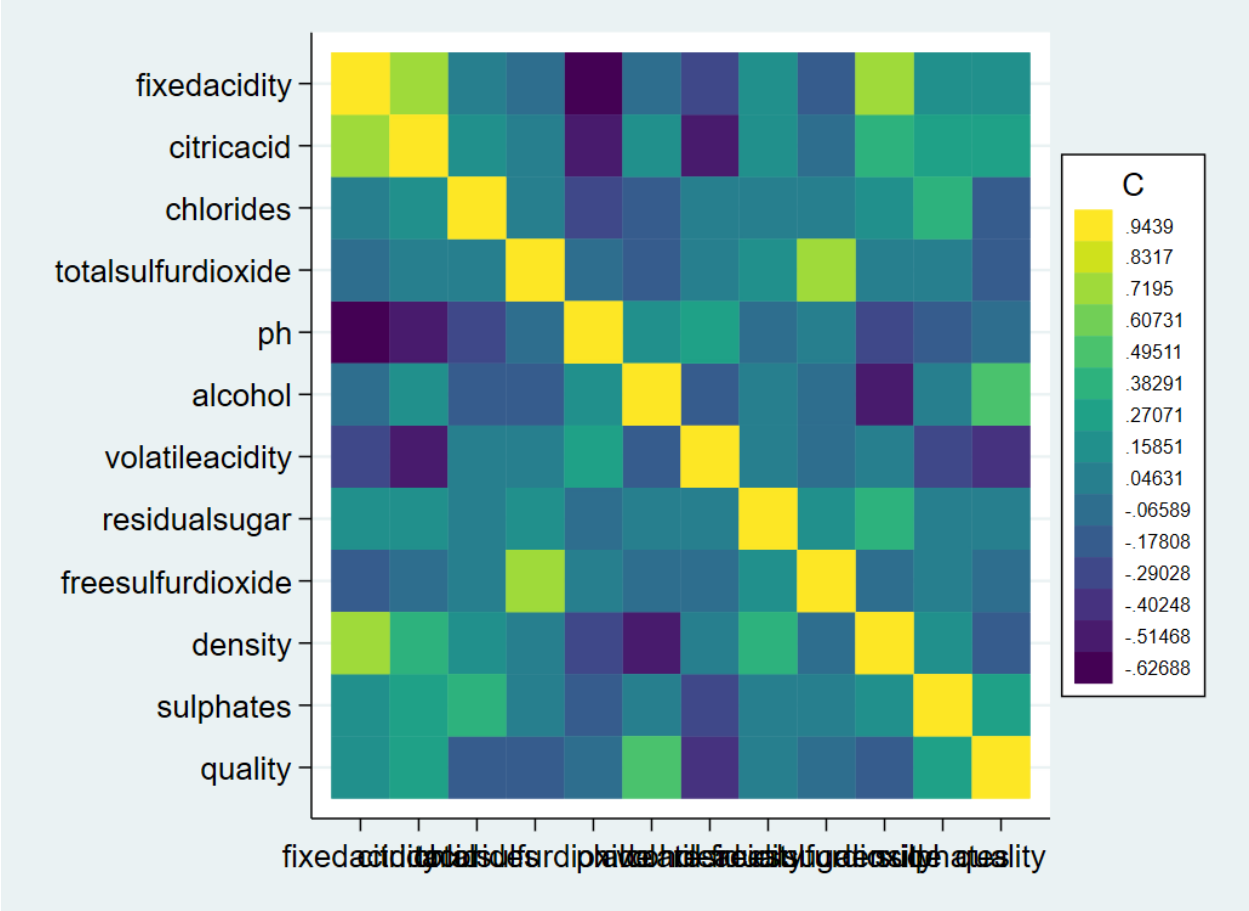
symmetric r(C)[12,12]

	fixedacidity	citricacid	chlorides	totalsulfu~e	ph	alcohol
fixedacidity	1					
citricacid	.67170343	1				
chlorides	.09370519	.20382291	1			
totalsulfu~e	-.11318144	.03553303	.04740047	1		
ph	-.68297819	-.54190414	-.26502614	-.06649456	1	
alcohol	-.06166828	.10990324	-.22114054	-.20565395	.20563252	1
volatileac~y	-.25613089	-.55249568	.06129777	.07647001	.23493727	-.20228803
residualsu~r	.11477672	.14357716	.05560954	.20302788	-.08565241	.04207544
freesulfur~e	-.15379419	-.06097813	.00556215	.66766645	.07037752	-.06940836
density	.66804694	.36494712	.20063248	.07126992	-.34169886	-.49617962
sulphates	.18300567	.31277004	.37126048	.04294683	-.19664759	.09359476
quality	.12405165	.22637251	-.12890656	-.18510029	-.05773139	.47616633

	volatileac~y	residualsu~r	freesulfur~e	density	sulphates	quality
volatileac~y	1					
residualsu~r	.00191788	1				
freesulfur~e	-.01050383	.187049	1			
density	.02202629	.3552834	-.02194575	1		
sulphates	-.26098668	.00552712	.05165757	.14850663	1	
quality	-.39055778	.01373164	-.05065606	-.17491903	.25139708	1

Pada tabel 4, diperlihatkan matriks korelasi pada setiap variabel nya. dapat dilihat pada variabel dependen kualitas (*quality*), variabel alcohol terlihat memiliki korelasi yang kuat (0,3-0,49) dengan quality (0,476) dan volatile acidity (-0,39). sementara itu, hubungan moderat (0,1 - 0,29) terlihat pada variabel fixed acidity (0,12), citric acid (0,226), chlorides (-0,129), total sulfur (-0,185), density (-0,175), dan sulphates (0,25). Secara lebih jelas matriks korelasi ini dapat dilihat pada Gambar 4, sehingga dapat memberikan gambaran bahwa variabel sangat mungkin berpengaruh pada kualitas anggur merah (dilampirkan pula matriks korelasi yang lebih jelas pada bagian lampiran).

Gambar 4. Matriks Korelasi antar Variabel



Sumber : UCI Machine Learning Repository, diolah (dilampirkan matriks korelasi yang lebih jelas pada bagian lampiran)

B. Pra Pemrosesan Data (*Data Preprocessing)*

*Data preprocessing s*angat penting dalam analisis data. Hal ini dikarenakan *dataprep* memiliki dampak langsung pada kualitas dan hasil model atau analisis yang dilakukan. Data preprocessing digambarkan sebagai tahapan yang bertujuan untuk mengatasi berbagai masalah yang dapat mengganggu proses pemrosesan data. Masalah tersebut sering kali disebabkan oleh inkonsistensi format data yang ada. Data preprocessing merupakan langkah pertama yang dilakukan sebelum melakukan data mining. Dalam tahapan ini, terdapat beberapa proses seperti membersihkan data, mengintegrasikan data dari berbagai sumber, mentransformasikan data ke dalam format yang lebih sesuai, dan mereduksi data untuk mengurangi kompleksitasnya (Suripto, 2022).

B.1 Pengecekan *Missing Value*

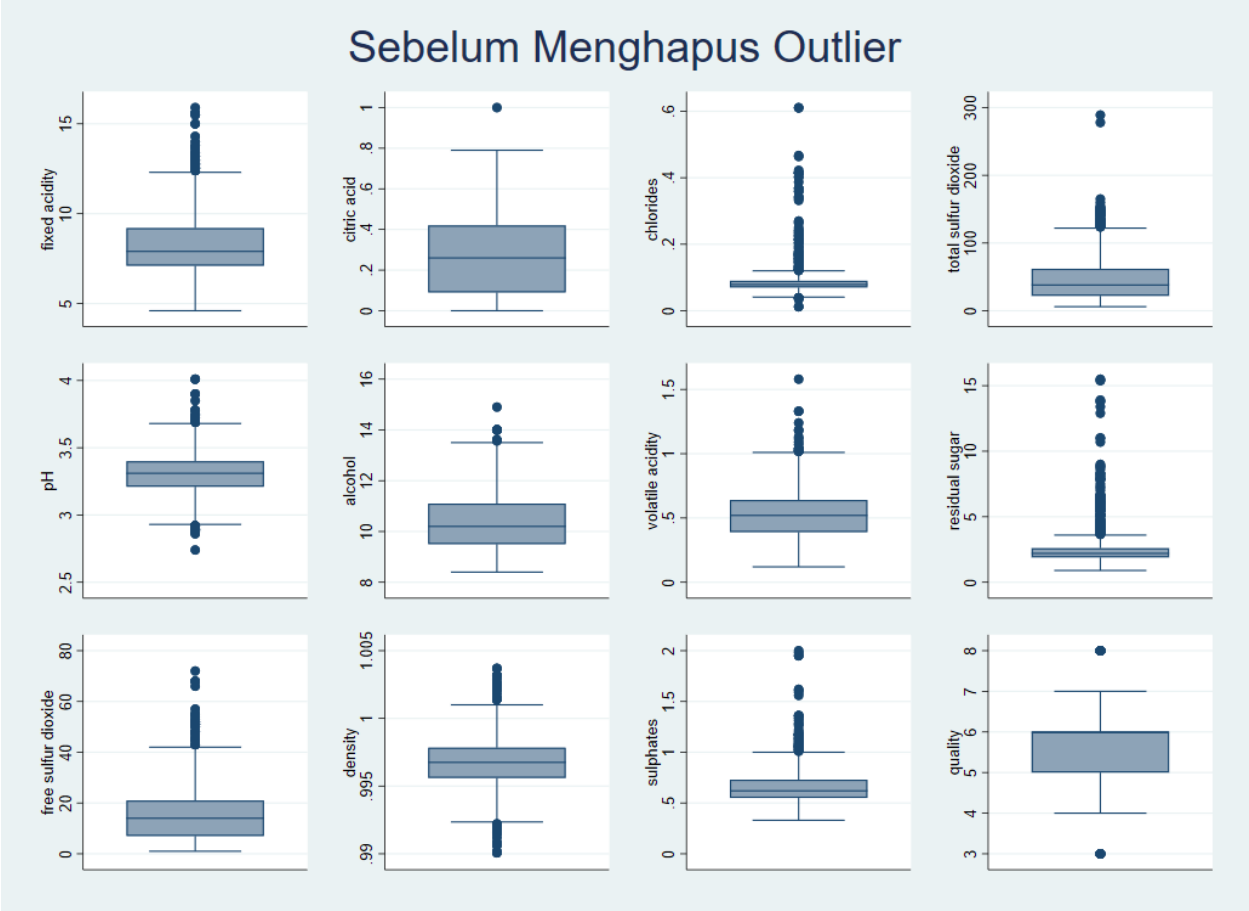
Tabel 5. Tabel Missing Value

Variable	Missing	Total	Percent Missing
fixedacidity	0	1,599	0.00
volatileac~y	0	1,599	0.00
citricacid	0	1,599	0.00
residualsu~r	0	1,599	0.00
chlorides	0	1,599	0.00
freesulfur~e	0	1,599	0.00
totalsulfu~e	0	1,599	0.00
density	0	1,599	0.00
ph	0	1,599	0.00
sulphates	0	1,599	0.00
alcohol	0	1,599	0.00
quality	0	1,599	0.00

B.2 Penghapusan Outlier

Outlier merupakan observasi yang berbeda dari kumpulan observasi lainnya, muncul pada satu ekstrim dari sebagian besar data (Kleinbaum, 2008). Salah satu metode yang dapat dilakukan adalah dengan menggunakan grafik *boxplot* dengan nilai kuartil dari jangkauan. Selain itu, *boxplot* ini juga memudahkan membayangkan besaran pencilan data yang ada. Berdasar Gambar 5, dapat dilihat kedua belas variabel memiliki outlier, dan sebagian besar memiliki outlier batas atas (upper bound). Hal ini didukung oleh distribusi data yang telah diterangkan pada analisis univariate sebelumnya.

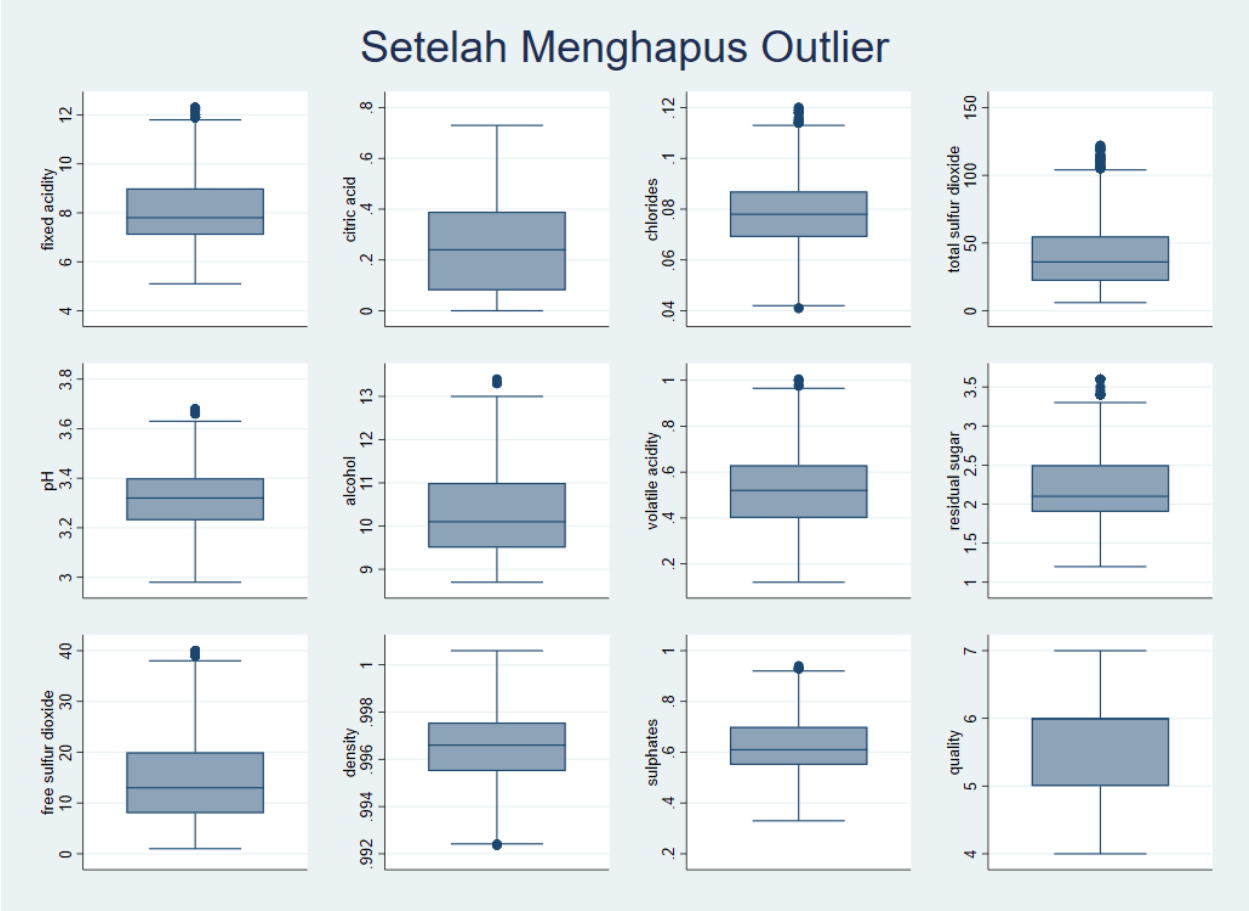
Gambar 5. Boxplot Sebelum Outlier dipotong



Sumber : UCI Machine Learning Repository, diolah

Dalam data preprocessing penghapusan outlier adalah langkah penting dalam mengatasi data ekstrim atau yang tidak representatif. Hal ini dikarenakan untuk mencegah bias analisis, memastikan kualitas data, memperbaiki asumsi statistik (supaya lebih simetris), meningkatkan keakuratan model. Setelah melakukan *treatment* untuk membuang observasi yang masuk pada outlier atas dan bawah dihasilkan observasi yang berdistribusi sebagaimana boxplot pada Gambar 6. Dapat dilihat, boxplot tersebut menunjukkan distribusi yang lebih baik (lebih simetris) dibandingkan sebelum *treatment* dilakukan. Pada Gambar 6 juga terlihat masih terdapat sedikit pencilan hal ini dikarenakan ketika melakukan pemotongan pada outlier, memungkin statistik deskriptif seperti median, kuartil 1, kuartil 3 dapat berubah sehingga outlier baru yang mempunyai definisi yang baru pula muncul namun sangat sedikit dibandingkan sebelumnya.

Gambar 6. Boxplot Setelah Outlier dipotong



Sumber : UCI Machine Learning Repository, diolah

Berikut disajikan (Tabel 6) deskripsi statistik dataset setelah dilakukan pemotongan outlier pada data. Terlihat masing-masing observasi bersisa 1161 dari sebelumnya 1599. Hal itu berarti terdapat 438 baris yang masuk pada outlier. Penghapusan outlier ini memberikan dampak yang baik pada penyebaran data yang dapat dilihat dari rata-rata setiap variabel dengan deviasi median.

Tabel 6. Deskriptif Statistik *Treatment* Outlier

Variable	Obs	Mean	Std. Dev.	Min	Max
fixedacidity	1,161	8.145478	1.447691	5.1	12.3
citricacid	1,161	.2457623	.1793901	0	.73
chlorides	1,161	.0785814	.0144042	.041	.12
totalsulfu~e	1,161	42.09044	26.08394	6	122
ph	1,161	3.324746	.1300177	2.98	3.68
alcohol	1,161	10.35373	.9671793	8.7	13.4
volatileac~y	1,161	.5237511	.1638676	.12	1.005
residualsu~r	1,161	2.183979	.4386323	1.2	3.6
freesulfur~e	1,161	14.87511	8.588529	1	40
density	1,161	.9965605	.0015864	.99236	1.0006
sulphates	1,161	.6280276	.1124446	.33	.94
quality	1,161	5.624462	.7241316	4	7

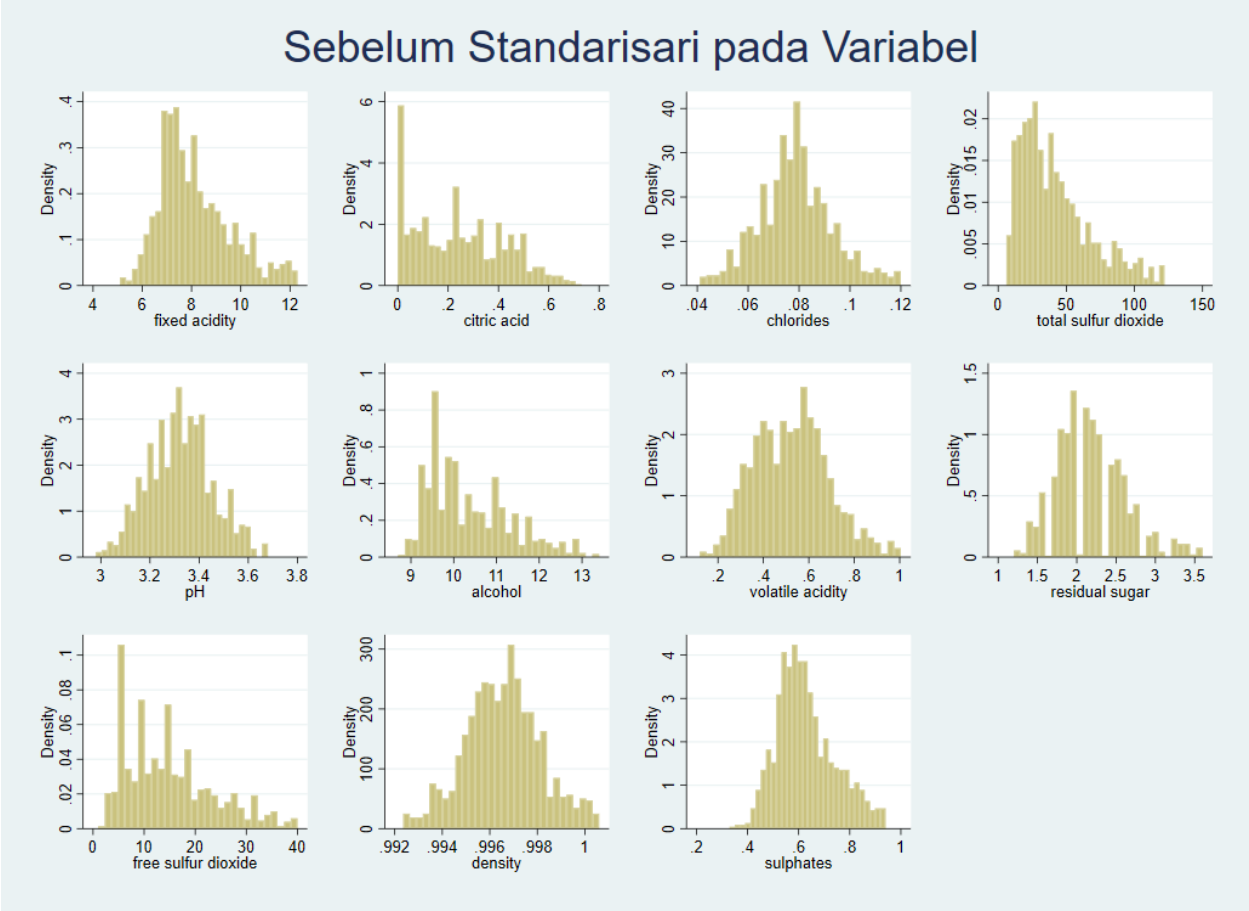
Sumber : UCI Machine Learning Repository, diolah

B. 3 Standarisasi Data

Standarisasi merupakan teknik pengolahan data untuk melakukan perubahan skala, dimana data yang dimiliki akan diubah dengan distribusi yang simetris, sehingga rata-rata = 0 (terpusat) dan standar deviasi nya = 1 (Anzihory, 2021).

$X^{'} = \frac{X-\mu}{\sigma}$, dimana μ adalah rata-rata dan σ adalah nilai standar deviasinya. Setelah melakukan treatment sebelumnya akan didapatkan distribusi sebagaimana Gambar 7.

Gambar 7. Histogram Variabel Sebelum Standarisasi



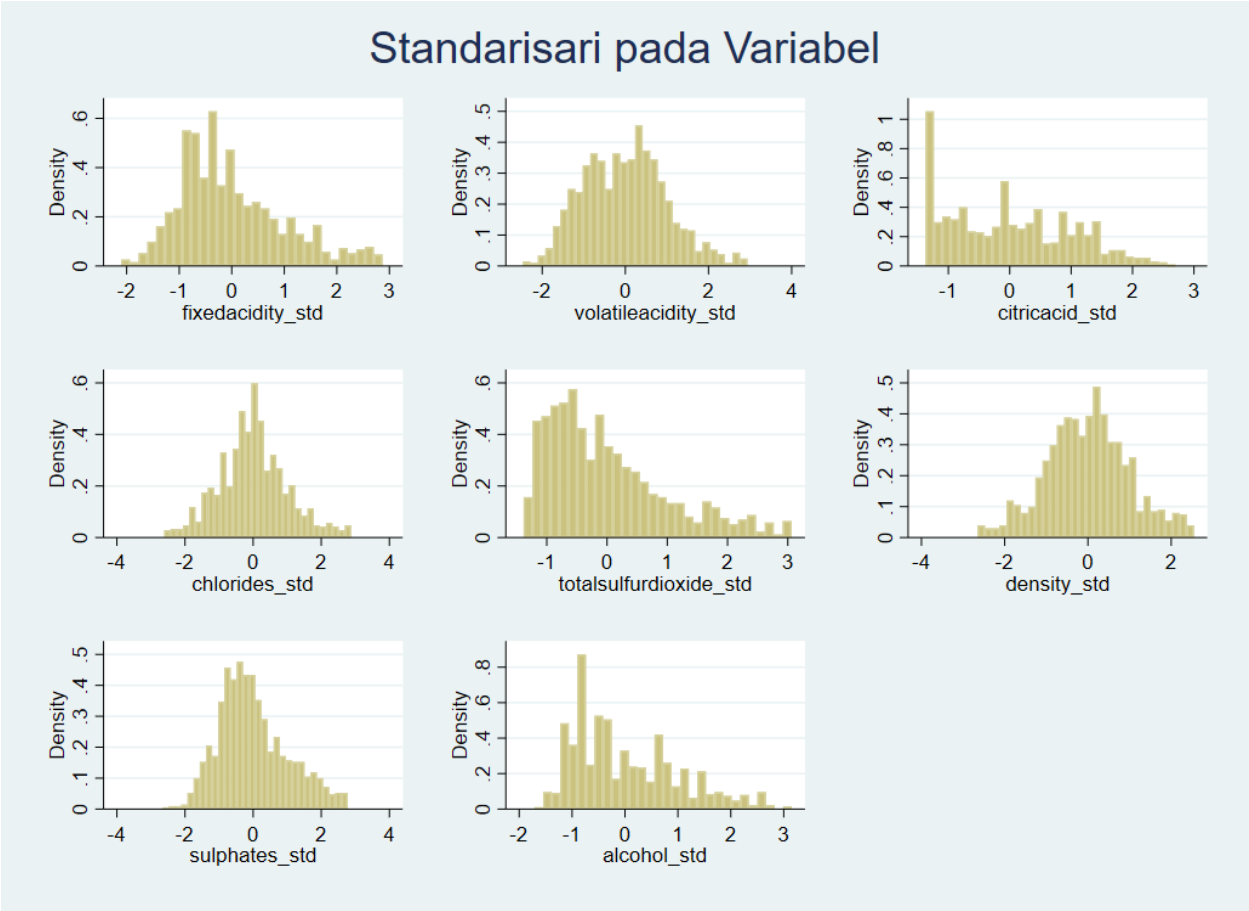
Sumber : UCI Machine Learning Repository, diolah

Berdasar Gambar 7, terlihat masih terdapat distribusi data yang tidak simetris pada beberapa variabel. Penulis memutuskan untuk melakukan standarisasi pada sembilan variabel, yaitu variabel fixed acidity, volatile acidity, citric acid, chlorides, total sulfur dioxide, density, sulphates, dan alcohol. Meskipun beberapa diantaranya terlihat memiliki distribusi yang simetris. Standardisasi data sangat penting karena memungkinkan sistem yang berbeda untuk bertukar data secara konsisten. Tanpa standardisasi, akan sulit bagi komputer untuk berkomunikasi satu sama lain dan bertukar data. Standardisasi juga memudahkan untuk mengolah dan menganalisis data serta menyimpannya dalam database (Simplilearn, 2023). Selain itu standarisasi data juga membuat kesalahan dalam mengambil kesimpulan dari analisis semakin kecil.

Dengan treatment standarisasi yang telah dilakukan, di dapatkanlah gambar. terlihat pada Gambar 8, bahwa variabel tersebut disimpan dengan nama `var'+'std' dimana `var' berarti nama

asli dari variabel tersebut sedangkan "std" berarti telah mengalami standarisasi. Variabel tersebut terlihat memiliki skala yang lebih seragam dengan mean berada pada kisaran 0.

Gambar 8. Histogram Setelah Variabel Terstandarisasi



Sumber : UCI Machine Learning Repository, diolah

Berikut disajikan (Tabel 7) deskripsi statistik dataset setelah dilakukan treatment standarisasi pada fitur tertentu. Terlihat masing-masing observasi tetap pada 1161. Standarisasi ini memberikan dampak yang baik penyeragaman skala yang sangat baik bagi algoritma *machine learning*, utamanya regresi logistik, SVM (Support Vector Machine) yang mengandalkan proses iterasi untuk memaksimalkan solusi optimalnya.

Tabel 7. Deskriptif Statistik Setelah Treatment Standarisasi dan Biner pada Quality

Variable	Obs	Mean	Std. Dev.	Min	Max
quality_good	1,161	.1171404	.3217261	0	1
residualsu~r	1,161	2.183979	.4386323	1.2	3.6
ph	1,161	3.324746	.1300177	2.98	3.68
fixedacidi~d	1,161	-1.75e-09	1	-2.10368	2.869757
citricacid~d	1,161	-5.23e-09	1	-1.369988	2.699356
totalsulfu~d	1,161	-1.84e-09	1	-1.383627	3.063554
sulphates_~d	1,161	1.11e-09	1	-2.65044	2.774456
freesulfur~e	1,161	14.87511	8.588529	1	40
volatileac~d	1,161	-6.67e-10	1	-2.463886	2.936816
chlorides_~d	1,161	-2.74e-09	1	-2.609066	2.875461
density_std	1,161	6.54e-11	1	-2.647865	2.546414
alcohol_std	1,161	-4.30e-10	1	-1.709851	3.149641

B.4. Pengubahan Data *Dependent Variable* (quality) ke Data Biner

Pada tahapan *preprocessing data* ini juga terdapat pemrosesan data yang melibatkan pengubahan variabel dependen "kualitas anggur merah" menjadi variabel biner dengan nilai 1 dan 0. Pengubahan ini dilakukan atas pertimbangan berikut.

1. Setelah melakukan percobaan estimasi dengan model ordered, hasil estimasi menggunakan model biner menghasilkan hasil yang lebih baik dengan tingkat klasifikasi yang lebih baik. Dalam konteks ini,
    - klasifikasi yang baik (=1), dianggap ketika kualitas anggur merah memiliki nilai 7 atau lebih tinggi,
    - klasifikasi yang tidak baik (=0) berlaku untuk nilai kualitas data sensori di bawah 7.
2. Alasan lain untuk mengubah variabel kualitas anggur merah menjadi biner adalah rekomendasi dari UCI Machine Learning Repository. Repositori ini menyarankan untuk menyederhanakan kasus ini menjadi klasifikasi biner, mengingat penelitian ini difokuskan pada kualitas anggur yang baik atau tidak, tanpa mempertimbangkan tingkatan kualitas yang lebih spesifik.

Pengambilan keputusan ini dilakukan dengan tujuan meningkatkan kualitas estimasi model dan fokus pada aspek kualitas anggur merah yang paling relevan untuk penelitian ini. Dengan mengubah variabel kualitas anggur merah menjadi biner, penelitian dapat lebih difokuskan pada identifikasi faktor-faktor yang berpengaruh terhadap kualitas anggur yang baik dan tidak baik. Klasifikasi kualitas anggur merah ini juga sudah dijelaskan pada tabel 7.

C. Pemodelan Logit dan Probit

Untuk menganalisis bagaimana variabel prediktor seperti fixed acidity (x_1), volatile acidity (x_2), citric acid (x_3), residual sugar (x_4), chlorides (x_5), free sulfur dioxide (x_6), total sulfur dioxide (x_7), density (x_8), ph (x_9), sulphates (x_{10}), dan alcohol (x_{11}) mempengaruhi variabel kualitas anggur merah, dilakukan pemodelan menggunakan metode logit dan probit dengan menggunakan data akhir setelah dilakukan proses treatment seperti dijelaskan pada bagian sebelumnya.

Tabel 8. Hasil Pemodelan Logit

```
Iteration 0:   log likelihood = -419.33978
Iteration 1:   log likelihood = -315.53544
Iteration 2:   log likelihood =  -276.7314
Iteration 3:   log likelihood = -274.18907
Iteration 4:   log likelihood = -274.16998
Iteration 5:   log likelihood = -274.16997

Logistic regression                             Number of obs     =      1,161
                                                LR chi2(11)       =     290.34
                                                Prob > chi2       =     0.0000
Log likelihood = -274.16997                     Pseudo R2         =     0.3462
```

quality_good	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
residualsugar	.0365944	.320017	0.11	0.909	-.5906274	.6638161
ph	-1.149284	1.532147	-0.75	0.453	-4.152238	1.853669
fixedacidity_std	.167654	.2625738	0.64	0.523	-.3469812	.6822893
citricacid_std	-.1004891	.2047398	-0.49	0.624	-.5017717	.3007936
totalsulfurdioxide_std	-.5775302	.2283057	-2.53	0.011	-1.025001	-.1300592
sulphates_std	.8449647	.1197432	7.06	0.000	.6102723	1.079657
freesulfurdioxide	.0081484	.0190754	0.43	0.669	-.0292388	.0455355
volatileacidity_std	-.4266219	.1795796	-2.38	0.018	-.7785915	-.0746523
chlorides_std	.0730769	.1173057	0.62	0.533	-.1568381	.3029919
density_std	-.1523435	.2680727	-0.57	0.570	-.6777564	.3730694
alcohol_std	1.052897	.2069556	5.09	0.000	.6472712	1.458522
_cons	.3989562	5.216218	0.08	0.939	-9.824643	10.62256

Tabel 9. Hasil Pemodelan Logit dengan variabel signifikan

```
Iteration 0:   log likelihood = -419.33978
Iteration 1:   log likelihood = -314.37563
Iteration 2:   log likelihood = -278.36467
Iteration 3:   log likelihood = -276.34088
Iteration 4:   log likelihood = -276.33206
Iteration 5:   log likelihood = -276.33206

Logistic regression                             Number of obs     =      1,161
                                                LR chi2(4)        =     286.02
                                                Prob > chi2       =     0.0000
Log likelihood = -276.33206                     Pseudo R2         =     0.3410
```

quality_good	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
totalsulfurdioxide_std	-.5748949	.1550481	-3.71	0.000	-.8787836	-.2710062
sulphates_std	.8357143	.1136444	7.35	0.000	.6129753	1.058453
volatileacidity_std	-.4504559	.1260981	-3.57	0.000	-.6976036	-.2033082
alcohol_std	1.073812	.1164734	9.22	0.000	.8455281	1.302095
_cons	-3.211395	.1921956	-16.71	0.000	-3.588091	-2.834698

Tabel 10. Hasil Pemodelan Probit

```
Iteration 0:   log likelihood = -419.33978
Iteration 1:   log likelihood = -286.99168
Iteration 2:   log likelihood = -274.83642
Iteration 3:   log likelihood = -274.60678
Iteration 4:   log likelihood = -274.60659
Iteration 5:   log likelihood = -274.60659

Probit regression                               Number of obs     =      1,161
                                                LR chi2(11)       =     289.47
                                                Prob > chi2       =     0.0000
Log likelihood = -274.60659                     Pseudo R2         =     0.3451
```

quality_good	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
residualsugar	.0061234	.1731838	0.04	0.972	-.3333106	.3455575
ph	-.5559636	.8411096	-0.66	0.509	-2.204508	1.092581
fixedacidity_std	.11082	.1445591	0.77	0.443	-.1725107	.3941507
citricacid_std	-.077449	.1114911	-0.69	0.487	-.2959675	.1410695
totalsulfurdioxide_std	-.307092	.119922	-2.56	0.010	-.5421348	-.0720491
sulphates_std	.443631	.0655884	6.76	0.000	.3150802	.5721818
freesulfurdioxide	.0058366	.0102506	0.57	0.569	-.0142543	.0259274
volatileacidity_std	-.2421433	.0976134	-2.48	0.013	-.4334621	-.0508244
chlorides_std	.0242609	.0640097	0.38	0.705	-.1011958	.1497176
density_std	-.0806644	.1473815	-0.55	0.584	-.3695269	.2081981
alcohol_std	.5634647	.112708	5.00	0.000	.3425611	.7843684
_cons	-.0064743	2.873499	-0.00	0.998	-5.638429	5.62548

Tabel 11. Hasil Pemodelan Probit dengan Variabel signifikan

```
Iteration 0:   log likelihood = -419.33978
Iteration 1:   log likelihood = -289.57452
Iteration 2:   log likelihood = -276.90828
Iteration 3:   log likelihood = -276.80429
Iteration 4:   log likelihood = -276.80428

Probit regression                                 Number of obs     =      1,161
                                                  LR chi2(4)        =     285.07
                                                  Prob > chi2       =     0.0000
Log likelihood = -276.80428                       Pseudo R2         =     0.3399
```

quality_good	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
totalsulfurdioxide_std	-.2962571	.0818535	-3.62	0.000	-.456687	-.1358272
sulphates_std	.4408483	.0623434	7.07	0.000	.3186575	.5630391
volatileacidity_std	-.2482206	.0676219	-3.67	0.000	-.3807572	-.1156841
alcohol_std	.574203	.0617384	9.30	0.000	.4531978	.6952081
_cons	-1.747126	.0898113	-19.45	0.000	-1.923153	-1.571099

C.1. Estimasi Parameter

Tabel 12. Hasil Estimasi Parameter

Pemodelan	Estimasi Parameter
Logit	g(x) = -3,21 - 0,45 volatileacidity_std (x2) - 0,575 totalsulfur_std (x7) + 0,836 sulphates_std (x10) + 1,074 alcohol_std (x11)
Probit	Z = -1,747 - 0,248 volatileacidity_std (x2) - 0,296 totalsulfur_std (x7) + 0,441 sulphates_std (x10) + 0,5742 alcohol_std (x11)

Dapat dilihat pada kedua model memberikan arah yang sama pada masing-masing variabel. Tabel tersebut menjelaskan estimasi B0 pada model probit lebih tinggi daripada model probit begitupun dengan total sulfur, volatile acidity. Sebaliknya pada variabel sulphates dan alcohol model logit memberikan nilai yang lebih tinggi dimana kedua variabel itu berarah positif. Sehingga dapat disimpulkan, model logit disini cenderung memberikan nilai yang lebih tinggi pada arah positif dan lebih rendah pada arah negatif sebaliknya dengan model probit. Atau dengan kata lain, model logit lebih cenderung sensitif.

1. Uji signifikansi parameter serentak (Likelihood Ratio Test)

Tabel 13. Pengujian Parameter Serentak

Pemodelan	LR chi2(4)
Logit	286.02l
Probit	285.07

Tabel tersebut memperlihatkan model logit memberikan nilai LR chi2 yang lebih besar dibandingkan model probit, sehingga nilai estimasi koefisien yang paling signifikan adalah model logit. Namun demikian dengan tingkat signifikansi (α)= 5%, kedua model tersebut disimpulkan signifikan, artinya variabel prediktor secara bersama-sama signifikan mempengaruhi variabel respons.

2. Uji Signifikansi Parameter Secara Parsial

Tabel 14. Pengujian Parameter Parsial

β	Logit	Probit
β_0	-3,21	-1,747
β_2 volatileacidity_std (x2)	- 0,45	- 0,248
	Wald = 12.76 P-Value = 0.0004	Wald = 13.47 P-Value =0.0002
β_7 totalsulfur_std (x7)	- 0,575	- 0,296
	Wald = 13.75 P-Value = 0.0002	Wald = 13.10 P-Value = 0.0003
β_{10} sulphates_std (x10)	0,836	0,441
	Wald = 54.08 P-Value = 0.0000	Wald = 50.00 P-Value = 0.0000
β_{11} alcohol_std (x11)	1,074	0,5742
	Wald = 85.00	Wald = 86.50

	P-Value = 0.0000	P-Value = 0.0000

Tabel 14 tersebut menjelaskan setelah pengujian pertama dengan uji Wald, dihasilkan variabel-variabel tersebut yang signifikan. Dapat dilihat bahwa variabel yang signifikan sama antara logit dan probit. Dari tabel tersebut dapat dilihat model logit memiliki nilai Wald yang paling besar pada β_4 , yaitu alkohol yang artinya bahwa variabel alkohol paling signifikan pada model logit. Begitu pula dengan model probit β_4 menjadi variabel dengan nilai Wald terbesar. Dengan demikian, kedua model menyatakan variabel alkohol paling signifikan berpengaruh pada bagus tidaknya kualitas anggur merah.

3. Ketepatan Klasifikasi dengan APER (*Apparent Error Rate)*

Tabel 15. Ketepatan Klasifikasi Model Logit

Logistic model for quality_good

Classified	True D	True ~D	Total
+	51	23	74
-	85	1002	1087
Total	136	1025	1161

Classified + if predicted Pr(D) >= .5
True D defined as quality_good != 0

Sensitivity	Pr(+| D)	37.50%
Specificity	Pr(-|~D)	97.76%
Positive predictive value	Pr(D| +)	68.92%
Negative predictive value	Pr(~D| -)	92.18%
False + rate for true ~D	Pr(+|~D)	2.24%
False - rate for true D	Pr(-| D)	62.50%
False + rate for classified +	Pr(~D| +)	31.08%
False - rate for classified -	Pr(D| -)	7.82%
Correctly classified		90.70%

Tabel 16. Ketepatan Klasifikasi Model Probit

Probit model for quality_good

Classified	True D	True ~D	Total
+	49	20	69
-	87	1005	1092
Total	136	1025	1161

Classified + if predicted Pr(D) >= .5
True D defined as quality_good != 0

Sensitivity	Pr(+| D)	36.03%
Specificity	Pr(-|~D)	98.05%
Positive predictive value	Pr(D| +)	71.01%
Negative predictive value	Pr(~D| -)	92.03%
False + rate for true ~D	Pr(+|~D)	1.95%
False - rate for true D	Pr(-| D)	63.97%
False + rate for classified +	Pr(~D| +)	28.99%
False - rate for classified -	Pr(D| -)	7.97%
Correctly classified		90.78%

Tabel 17. Matrik APER

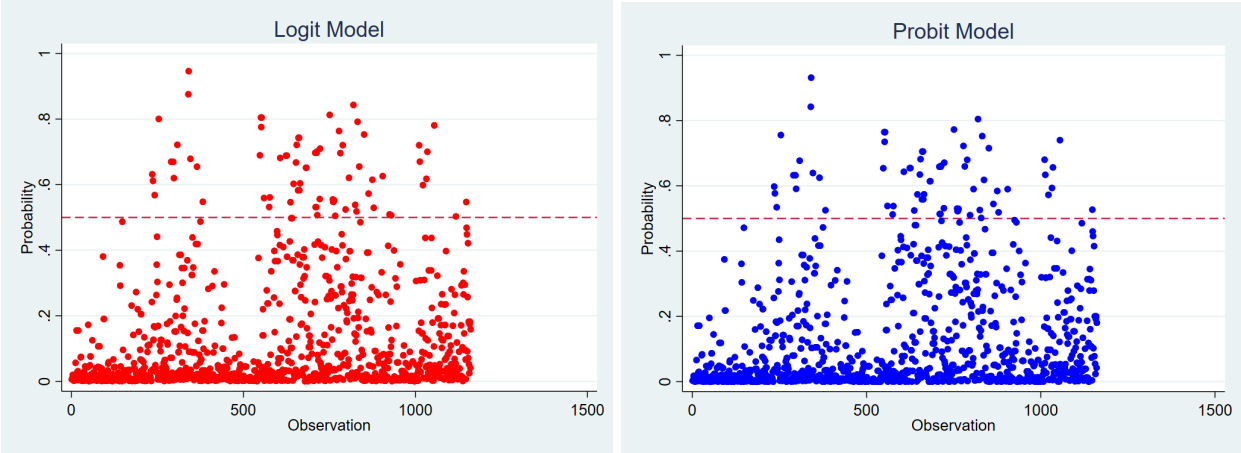
Pemodelan	APER	Ketepatan Klasifikasi

Logit	9,3%	90,7%
Probit	9,21%	90,79%

Berdasar tabel tersebut, evaluasi model probit memberikan hasil yang lebih baik dibanding model logit dengan selisih performa ketepatan 0,09% (atau hampir 1%).

4. Plot Probabilitas

Gambar 9. Plot Probabilitas Logit Probit



Sumber : UCI Machine Learning Repository, diolah

Dengan menggunakan nilai parameter η yang telah ditentukan, kita dapat menghitung probabilitas kualitas anggur merah pada model logit dan probit. Probabilitas ini menggambarkan kemungkinan terjadinya kualitas anggur merah yang baik (di atas atau sama dengan 7) berdasarkan nilai variabel prediktor yang telah diobservasi. Gambar tersebut menjelaskan bahwa probabilitas plot pada kedua model (logit dan probit) cenderung sama.

5. Model Terbaik

Tabel 18. AIC, BIC Model Logit

Akaike's information criterion and Bayesian information criterion

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	1,161	-419.3398	-276.3321	5	562.6641	587.9493

Note: N=Obs used in calculating BIC; see [R] BIC note.

Tabel 19. AIC, BIC Model Probit

Akaike's information criterion and Bayesian information criterion

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	1,161	-419.3398	-419.3398	1	840.6796	845.7366

Note: N=Obs used in calculating BIC; see [R] BIC note.

Pemilihan model terbaik ini akan didasarkan pada nilai McFadden's R2 terbesar dan AIC dan BIC terkecil yang ditunjukkan pada tabel berikut.

Tabel 20. Pengujian Model Terbaik

Parameter Model Terbaik	Logit	Probit
McFadden's R2	0.329	0.340
AIC (estat)	562,6641	840,6796
AIC (fitstat)	0.485	0.485
AIC*n (fitstat)	562.664	563.609
BIC (estat)	587,95	845,74

Berdasarkan nilai McFadden's R2 model probit unggul dibandingkan dengan model logit. Hal itu berarti model probit memberikan penjelasan lebih baik terhadap variasi dalam data dibandingkan model logit. sebaliknya berdasar AIC (*Akaike Information Criterion*) estat, model logit lebih baik dibandingkan probit dengan ditunjukkannya nilai yang lebih rendah. Hal itu berarti model logit memberikan penjelasan lebih baik dengan mempertimbangkan kompleksitas model. Namun dengan menggunakan perintah *fitstat,* AIC keduanya sama tetapi BIC dan AIC*n untuk logit lebih rendah. Kesimpulan penulis, penulis menilai model probit yang lebih baik digunakan dalam kasus ini. Hal ini didukung dari tingkat APER yang lebih rendah atau dengan kata lain ketepatan klasifikasi yang lebih tinggi, akibat dari model yang lebih kompleks sebagaimana dijelaskan pada AIC dan BIC.

V. KESIMPULAN DAN SARAN

Penelitian ini telah menemukan setidaknya sembilan hasil yang patut dicatat. Pertama, berdasar dengan 1.599 dataset original yg didapat dari UCI Machine Learning Repository yang kemudian dilakukan treatment sehingga menjadi 1.161 dataset per masing-masing variabel didapatkan hasil terdapat beberapa variabel independen yang mempengaruhi variabel dependen kualitas anggur merah. Kedua, dengan melakukan *correlation matrix*, terdapat tujuh variabel independen yang diduga mempengaruhi kualitas anggur merah, yaitu alcohol, volatile acidity, fixed acidity, citric acid, chlorides, total sulfur, density, dan sulphates yang tersebar dengan kategori korelasi moderat hingga kuat. Ketiga, setelah mengetahui dugaan kuat pada variabel-variabel tersebut penulis lebih memperhatikan variabel tersebut dengan melakukan treatment-treatment khusus seperti memasukkannya pada standarisasi data/observasi untuk mendapatkan distribusi yang simetris sehingga menghasilkan hasil estimasi yang optimal. Keempat, setelah melakukan pemodelan dengan logit dan probit didapatkan empat variabel yang sama yang signifikan berdasarkan uji Wald, yaitu volatileacidity_std (x2), totalsulfure_std (x7), sulphates_std (x10), dan alcohol_std (x11), sehingga disimpulkan (-) jumlah asam, (-) total sulfur (S02), (+) aditif anggur (sulphates), dan (+) persen kandungan alkohol berpengaruh pada kualitas anggur. Kelima, berdasarkan evaluasi model yang dilakukan menggunakan uji Likelihood Ratio Test disimpulkan bahwa kedua model dapat menjelaskan hubungan antara variabel independen dengan variabel dependen. Keenam, dengan Wald test disimpulkan bahwa kedua model menghasilkan arah yang searah dalam konstanta dan masing-masing variabel yang signifikan dan disimpulkan bahwa model estimasi logit lebih sensitif dibanding probit dengan ekstrimnya nilai koefisien variabel. Ketujuh, berdasarkan APER (*Apparent Error Rate)* didapatkan model probit mempunyai ketepatan klasifikasi yang lebih tinggi (90,79%) hampir 1% dibandingkan logit (90,7%). Kedelapan, berdasarkan nilai McFadden R^2 model probit mendapat nilai yang lebih tinggi dalam memberikan penjelasan dibanding logit namun dengan mempertimbangkan kompleksitas model, model logit lebih baik dalam memberi penjelasan dibanding probit ditinjau dari lebih rendahnya nilai BIC dan AIC. Kesembilan, dengan berbagai pertimbangan penulis memberi kesimpulan bahwa pada kasus penentuan faktor determinasi kualitas anggur merah model probit merupakan model terbaik.

Namun, semua hasil ini harus diperlakukan dengan hati-hati, dan tidak dapat dengan mudah diekstrapolasi ke kasus lain atau ekonomi lain. Setidaknya ada dua alasan untuk ini: pertama, karena koefisien determinasi (Pseudo R2) yang tidak terlalu tinggi, yaitu 34% pada model terbaik. Sehingga masih banyak variabel yang mempengaruhi kualitas anggur merah, seperti *treatment* penanaman dan pengolahan anggur merah, kondisi anggur merah, negara asal anggur merah dan sebagainya. Kedua, dalam penelitian ini penulis menggunakan dataset UCI Machine Learning Repository (Cortez et al., 2009) dengan sampling Vinho Verde, dari Portugal Utara, sehingga tidak dapat dipastikan untuk terhindar dari *selection bias.* Tentu hal tersebut, membutuhkan kajian mendalam dan spesifik pada regional secara random untuk mendapat kesimpulan yang bersifat umum.

DAFTAR PUSTAKA

Agresti, A. (2019). An introduction to categorical data analysis. John Wiley & Sons.

Anzihory, E. (2021, July 19). Normalisasi vs Standarisasi. Medium. https://anzihory.medium.com/normalisasi-vs-standarisasi-101093633e18

Basalamah, S. (2020, October 23). Cara Mengidentifikasi Dan Penanganan data outlier. Medium. https://salsabilabasalamah.medium.com/cara-mengidentifikasi-dan-penanganan-data-outlier-d2fe16c6d62c

Enderlein, G. (1987). McCullagh, P., J. A. Nelder: Generalized Linear Models. Chapman and Hall London – New York 1983, 261 S., £ 16,–. Biometrical Journal, 29(2), 206–206. https://doi.org/10.1002/bimj.4710290217

Fact.MR – red wine market by product type (Shiraz, merlot, cabernet sauvignon, Pinot Noir & others), by body type (light bodied red wine, medium bodied red wine), by sweetness level (dry red wine, sweet red wine), by Sales Channel & by region - global market insights 2022 - 2032. Fact.MR, Market Research Company. (n.d.). https://www.factmr.com/report/160/red-wine-market

Gomes, G. S., & Ludermir, T. B. (2008). Complementary log-log and probit: Activation functions implemented in Artificial Neural Networks. 2008 Eighth International Conference on Hybrid Intelligent Systems. https://doi.org/10.1109/his.2008.40

Greene, W. (2005). Reconsidering heterogeneity in panel data estimators of the Stochastic Frontier Model. Journal of Econometrics, 126(2), 269–303. https://doi.org/10.1016/j.jeconom.2004.05.003

Hardin, J. W., & Hilbe, J. M. (2012). Generalized Estimating Equations. https://doi.org/10.1201/b13880

Indonesia, D. (n.d.). Peta Produsen Anggur di indonesia pada 2021, Bali terbesar. Dataindonesia.id. https://dataindonesia.id/agribisnis-kehutanan/detail/peta-produsen-anggur-di-indonesia-pada-2021-bali-terbesar

Introduction to Glms: Stat 504. PennState: Statistics Online Courses. (n.d.-a). https://online.stat.psu.edu/stat504/lesson/6/6.1

Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2014). Applied regression analysis and other multivariable methods. Cengage Learning.

Kothawade, R. D., Raj Gupta, H., & Mondal, M. (2023). Building a classification model based on feature engineering for the prediction of wine quality by employing supervised machine learning and Ensemble Learning Techniques. 2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE). https://doi.org/10.1109/iccece51049.2023.10085272

Learning, U. M. (2017, November 27). Red Wine Quality. Kaggle. https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

Red Wine Market Size, trends and Global Forecast to 2032. The Business Research Company. (n.d.). https://www.thebusinessresearchcompany.com/report/red-wine-global-market-report

Reyvan. (n.d.). Jigsaw is now part of UNext. Jigsaw is now Part of UNext | UNext. https://www.jigsawacademy.com/blogs/business-analytics/univariate-analysis/

Setyaningrum, D. A., & Sirait, T. (2021). Pemodelan logit, probit, Dan Complementary Log-log. Seminar Nasional Official Statistics, 2020(1), 429–438. https://doi.org/10.34123/semnasoffstat.v2020i1.383

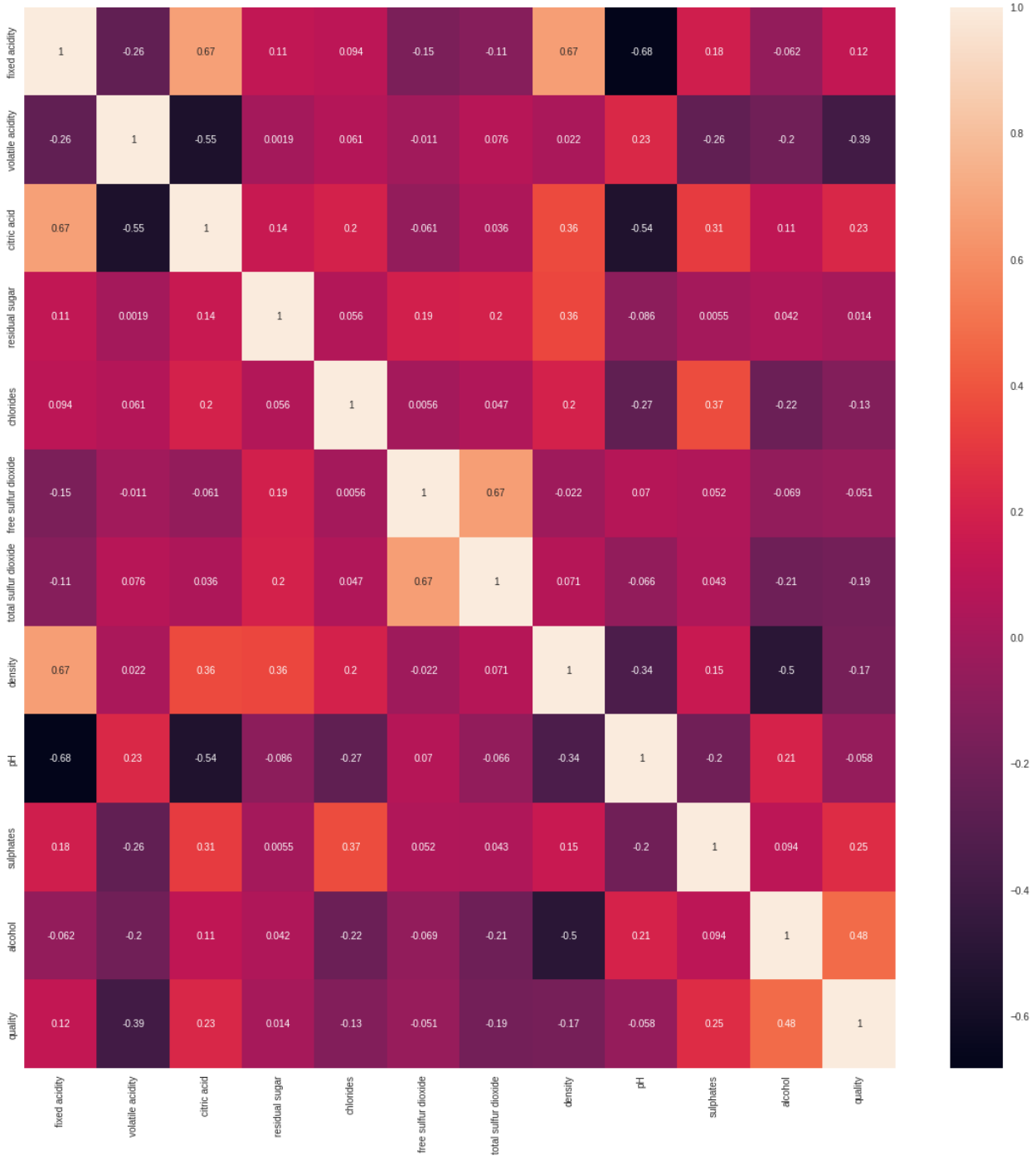
Single categorical predictor: Stat 504. PennState: Statistics Online Courses. (n.d.-b). https://online.stat.psu.edu/stat504/lesson/6/6.2

Teknik pre-processing Dan Classification Dalam Data Science. Master of Industrial Enginering. (2022, August 26). https://mie.binus.ac.id/2022/08/26/teknik-pre-processing-dan-classification-dalam-data-science/

Wine quality. UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/186/wine+quality

Lampiran

Lampiran 1. Korelasi Matriks antar Variabel



Lampiran 2. Fitstat Pemodelan Logit

Measures of Fit for **logit** of **quality_good**

Log-Lik Intercept Only:	-419.340	Log-Lik Full Model:	-276.332
D(1156):	552.664	LR(4):	286.015
		Prob > LR:	0.000
McFadden's R2:	0.341	McFadden's Adj R2:	0.329
Maximum Likelihood R2:	0.218	Cragg & Uhler's R2:	0.424
McKelvey and Zavoina's R2:	0.533	Efron's R2:	0.309
Variance of y*:	7.041	Variance of error:	3.290
Count R2:	0.907	Adj Count R2:	0.206
AIC:	0.485	AIC*n:	562.664
BIC:	-7605.271	BIC':	-257.787

Lampiran 3. Fitstat Pemodelan Probit

Measures of Fit for **probit** of **quality_good**

Log-Lik Intercept Only:	-419.340	Log-Lik Full Model:	-276.804
D(1156):	553.609	LR(4):	285.071
		Prob > LR:	0.000
McFadden's R2:	0.340	McFadden's Adj R2:	0.328
Maximum Likelihood R2:	0.218	Cragg & Uhler's R2:	0.423
McKelvey and Zavoina's R2:	0.515	Efron's R2:	0.306
Variance of y*:	2.061	Variance of error:	1.000
Count R2:	0.908	Adj Count R2:	0.213
AIC:	0.485	AIC*n:	563.609
BIC:	-7604.326	BIC':	-256.843