Tabel Perbandingan 8 Algoritma Machine Learning

Parameter	Logistic Regression	GLM (Generalized Linear Model)	Naive Bayes	Decision Tree	Neural Networks	Linear Regression	K-Means	Random Forest
Tipe Algoritma	Supervised (Klasifikasi)	Supervised (Regresi/Klasifikasi)	Supervised (Klasifikasi)	Supervised (Klasifikasi/Regresi)	Supervised (Klasifikasi/Regresi)	Supervised (Regresi)	Unsupervised (Clustering)	Supervised (Klasifikasi/Regresi)
Asumsi Data	Hubungan linear antara log-odds dan fitur	Distribusi spesifik (Normal, Binomial, dll)	Independensi fitur	Tidak ada asumsi distribusi	Tidak ada asumsi eksplisit	Linearitas, homoskedastisitas	Cluster berbentuk spherical	Tidak ada asumsi kuat
Kompleksitas Model	Rendah	Sedang	Sangat Rendah	Sedang (tergantung kedalaman)	Sangat Tinggi	Rendah	Sedang	Tinggi
Interpretabilitas	**Tinggi** (Koefisien jelas)	**Tinggi**	**Tinggi**	**Tinggi** (Visualisasi pohon)	**Sangat Rendah** (Black-box)	**Tinggi**	**Sedang** (Analisis cluster)	**Sedang** (Feature importance)
Preprocessing	Normalisasi (opsional)	Tergantung distribusi	Kategorikal → Frekuensi	Tidak perlu	Wajib (Normalisasi, Encoding)	Normalisasi (jika skala beda)	Wajib (Scaling)	Tidak perlu
Keunggulan	- Cepat dan sederhana	- Fleksibel untuk berbagai distribusi data	- Sangat cepat untuk dataset besar	- Mudah diinterpretasi	- Akurasi sangat tinggi untuk data kompleks	- Interpretasi mudah	- Skalabilitas baik	- Akurasi tinggi tanpa overfitting
Kelemahan	- Hanya untuk hubungan linear	- Pemilihan distribusi krusial	- Gagal jika fitur dependen	- Rentan overfitting	- Butuh data dan komputasi besar	- Sensitif terhadap outlier	- Harus tentukan K manual	- Komputasi intensif
Toleransi Outlier	Rendah	Tergantung distribusi	Tinggi	Sedang	Sedang	Sangat Rendah	Rendah	Tinggi
Hyperparameter Kunci	- Regularisasi (L1/L2)	- Family distribution (Binomial, Poisson)	- Smoothing parameter (Laplace)	- Max depth, Min samples split	- Layers, Learning rate, Epochs	- Regularisasi (Ridge/Lasso)	- Jumlah cluster (K)	- Jumlah pohon, Max depth

Parameter	Logistic Regression	GLM (Generalized Linear Model)	Naive Bayes	Decision Tree	Neural Networks	Linear Regression	K-Means	Random Forest
Aplikasi Ideal	- Klasifikasi biner (e.g., spam detection)	- Data count (Poisson), Binary (Logistik)	- Text classification	- Rules-based systems (e.g., risiko kredit)	- Image/NLP/Time series	- Prediksi harga, sales	- Segmentasi pelanggan	- Kaggle competitions, High-accuracy tasks
Tools RapidMiner	`Logistic Regression`	`Generalized Linear Model`	`Naive Bayes`	`Decision Tree`	`Neural Net` / `Deep Learning`	`Linear Regression`	`K-Means`	`Random Forest`
Waktu Training	Sangat Cepat	Cepat	Sangat Cepat	Cepat	Sangat Lambat	Cepat	Sedang	Lambat (tergantung jumlah pohon)
Handling Missing Data	Median/Mean Imputation	Median/Mean Imputation	Ignoransi atau imputasi sederhana	Split ke cabang terpisah	Wajib imputasi	Median/Mean Imputation	Wajib imputasi	Handle otomatis via bagging
Dimensionality	Cocok untuk fitur sedikit (~10-100)	Cocok untuk fitur sedikit	Cocok untuk fitur tinggi (e.g., text)	Fitur menengah (hingga ribuan)	High-dimensional (e.g., gambar)	Fitur sedikit	Reduksi dimensi disarankan (PCA)	Fitur tinggi (ribuan)
Overfitting Risk	Rendah (dengan regularisasi)	Sedang	Sangat Rendah	Sangat Tinggi	Tinggi (perlu dropout/regularisasi)	Sedang (dengan regularisasi)	N/A (Unsupervised)	Rendah (karena bagging)
Metrik Evaluasi	- Accuracy, AUC-ROC	- Deviance, AIC	- Precision, Recall	- Accuracy, Gini Importance	- Loss function (e.g., Cross-Entropy)	- RMSE, R^2	- Silhouette Score, WCSS	- OOB Error, Feature Importance