📚 Comaprision between Machine Learning models

Feature/Algorithm	Logistic Regression	Decision Tree	Random Forest	K-Nearest Neighbors (KNN)
Use Cases	Binary classification (e.g., spam detection).	Classification with clear decision boundaries (e.g., loan approval).	Complex classification tasks (e.g., image recognition).	Classification where proximity of data points is relevant (e.g., recommendation systems).
Pros	Simple, efficient, interpretable. Good for small datasets and binary classification.	Easy to understand and interpret. Non-parametric.	Handles large datasets well. Robust to outliers and noisy data.	Simple, effective, and non-parametric. Good for multi-class problems.
Cons	Assumes linear relationship. Not suitable for complex relationships in data.	Prone to overfitting. Not ideal for very large datasets.	Can be computationally intensive. Less interpretable.	Sensitive to noisy data. Computationally intensive with large datasets.
When to Use	When the relationship between variables is approximately linear.	When decisions can be made based on a set of rules.	For ensemble learning to improve accuracy. When dealing with unbalanced datasets.	When the dataset is small, and the distance between data points is a meaningful indicator.
Metrics	Accuracy, Precision, Recall, F1-Score, ROC-AUC.	Accuracy, Precision, Recall, F1-Score, ROC-AUC.	Accuracy, Precision, Recall, F1-Score, ROC-AUC.	Accuracy, Precision, Recall, F1-Score, ROC-AUC.

📚 Binary- and Multi classifications

Feature	Binary Classification	Multi-Class Classification
Definition	Categorizing data into one of two distinct groups.	Categorizing data into more than two groups.
Use Cases	- Spam Detection (spam or not spam).	- Image Classification (animals, cars, landscapes).

Feature	Binary Classification	Multi-Class Classification
	- Medical Diagnosis (disease positive or negative).	- Text Categorization (sports, politics, entertainment).
	- Credit Approval (approve or deny).	- Handwriting Recognition (different characters).
Characteristics	- Only two categories.	- More than two categories.
	- Estimate probability of a single class.	- Estimate probability of multiple classes.
	- Class prediction based on higher probability.	- Class prediction based on highest probability.
Common Algorithms	- Logistic Regression.	- Multinomial Logistic Regression.
	- Support Vector Machine (SVM).	- Decision Trees.
	- Neural Networks (specific configurations).	- Random Forests.
		- Neural Networks.
Metrics	- Accuracy, Precision, Recall, F1-Score, ROC-AUC.	- Accuracy, Precision, Recall, F1-Score (averaged).
	- Focus on distinguishing between two classes.	- Focus on correctly identifying multiple classes.

📚 Metrics : Classifications

Metric	Definition	Pros	Cons	Use Cases	Examples
Accuracy	The ratio of correctly predicted observations to the total observations.	Simple and intuitive.	Can be misleading with imbalanced datasets.	General classification tasks when class distribution is balanced.	Correctly identifying 90 out of 100 emails as spam or not spam.
Precision	The ratio of correctly predicted positive observations to the total predicted positives.	Focuses on the positive class. Avoids false positives.	Does not consider false negatives.	When the cost of false positives is high.	Predicting spam emails where false positives (marking important emails as spam) are critical.
Recall	The ratio of correctly predicted positive observations to the all	Focuses on covering the actual positive cases.	Does not consider false positives.	When missing a positive is costly.	Medical diagnostics where missing a disease (false negative) is critical.

Metric	Definition	Pros	Cons	Use Cases	Examples
	observations in the actual class.				
F1-Score	The weighted average of Precision and Recall.	Balances Precision and Recall.	Not as easy to interpret as accuracy.	When there's an uneven class distribution and both false positives and false negatives are important.	Classifying fraudulent transactions where both false negatives and false positives are important.
ROC-AUC	The performance measurement for the classification problems at various thresholds settings.	Provides an aggregate measure of performance across all possible classification thresholds.	Can be overly optimistic with imbalanced datasets.	Comparing different classifiers.	Evaluating a model's ability to distinguish between patients with and without a disease.

📚 Metrics: Regressions

Metric	Definition	Pros	Cons	Use Cases	Examples
MAE (Mean Absolute Error)	The average of the absolute differences between the predicted values and observed values.	Simple to understand and interpret.	Can be less sensitive to outliers.	General regression tasks.	Predicting house prices, where each error term is equally important.
MSE (Mean Squared Error)	The average of the squared differences between the predicted values and observed values.	Punishes larger errors more.	Can be more sensitive to outliers.	When large errors are particularly undesirable.	Financial risk modeling, where large errors can be very costly.
RMSE (Root Mean Squared Error)	The square root of the MSE.	Easily interpretable in the units of the response variable.	Sensitive to outliers.	When the magnitude of the error is important.	Forecasting sales figures, where large errors are especially undesirable.

Metric	Definition	Pros	Cons	Use Cases	Examples
R-squared (R^2)	The proportion of the variance in the dependent variable that is predictable from the independent variables.	Easy interpretation as a percentage.	Does not indicate whether the model is adequate.	Comparing models of the same dependent variable.	Comparing different models predicting employee performance.
Adjusted R-squared	Modified version of R^2 that adjusts for the number of predictors in the model.	Penalizes for unnecessary variables.	Still does not indicate whether a model is adequate.	When multiple models with different numbers of predictors are being compared.	Model selection in multiple regression when deciding how many predictors to include.

📚 Major Hyper Parameters

Model	Hyperparameter	Meaning	Common Values	Example Use Case
Logistic Regression	C	Inverse of regularization strength; smaller values specify stronger regularization.	0.01, 0.1, 1, 10, 100	Credit scoring: balancing bias and variance.
	penalty	Specifies the norm used in the penalization.	'l1', 'l2', 'elasticnet'	Spam detection: feature selection and regularization.
Decision Tree	max_depth	Maximum depth of the tree.	None, 3, 5, 10	Loan approval: preventing overfitting.
	min_samples_split	Minimum number of samples required to split an internal node.	2, 5, 10	Marketing targeting: control tree growth.
Random Forest	n_estimators	Number of trees in the forest.	10, 100, 200	Image classification: accuracy and performance.
	max_features	Number of features to consider when looking for the best split.	'auto', 'sqrt', 'log2'	Biometric authentication: feature selection.
KNN	n_neighbors	Number of neighbors to use.	3, 5, 7, 11	Recommendation systems: tuning model complexity.

Model	Hyperparameter	Meaning	Common Values	Example Use Case
	weights	Weight function used in prediction.	'uniform', 'distance'	Healthcare: weighting closer neighbors more heavily.
SVM	C	Regularization parameter.	0.01, 0.1, 1, 10, 100	Text classification: balancing margin and misclassification.
	kernel	Specifies the kernel type to be used in the algorithm.	'linear', 'poly', 'rbf', 'sigmoid'	Face recognition: choosing appropriate kernel.
K-Means	n_clusters	The number of clusters to form as well as the number of centroids to generate.	3, 5, 10, 20	Market segmentation: identifying distinct customer groups.
	init	Method for initialization.	'k-means++', 'random'	Document clustering: optimizing centroid initialization.
DBSCAN	eps	The maximum distance between two samples for one to be considered as in the neighborhood of the other.	0.3, 0.5, 1	Environmental studies: identifying clusters of geographic locations.
	min_samples	The number of samples (or total weight) in a neighborhood for a point to be considered as a core point.	5, 10, 20	Anomaly detection: determining density thresholds.

Cross Validations

Cross-Validation Method	Definition	Pros	Cons	Use Cases	Exa Scen
KFold	Splits the dataset into K consecutive folds without shuffling.	Simple and easy to understand.	Not suitable for imbalanced datasets.	When dataset is large and relatively balanced.	Valida regres mode balanc datase
	Each fold is then used once as a validation while the K-1 remaining	Good for large datasets.	Each fold might not represent the overall distribution.		

Cross-Validation Method	Definition	Pros	Cons	Use Cases	Exa Scen
	folds form the training set.				
StratifiedKFold	Splits the dataset into K folds, making sure each fold is an appropriate representative of the class proportions.	Reduces variance and bias in imbalanced datasets.	Slightly more complex than KFold.	When dealing with classification problems, especially with imbalanced datasets.	Evalua classif mode where target variab classe imbala
	Like KFold, but each set contains approximately the same percentage of samples of each target class.	Ensures each fold is representative of the class proportions.			
RandomizedStratifiedKFold	Similar to StratifiedKFold, but adds a layer of randomness in splitting.	Combines the benefits of random splitting and stratification.	More complex to implement and understand.	When randomness in split is important to avoid bias.	Testin mode perfor with a selecti the da
	Each fold is a random representative of class proportions, ensuring both randomness and stratification.	Ensures robustness against overfitting.	Might introduce additional randomness to the model evaluation.		