

Study on ML Models: from comparing with Dataset



Wine Quality

Total Models Trained  $\rightarrow$  7.

1. Logistic Regression.  $\xrightarrow{\text{Use code \& Best code}}$  linearly separable Data.

Linear model  $\rightarrow$  classification. well b/w Data w.r.t.

Adv. Estimates Probability using Sigmoid func.

↳ Simple, fast, interpretable.

Dig Adv.

$\times$  — Non-linear Pattern Complex

Struggle.

Intuition  $\rightarrow$  find a straight line that separates classes.

2. K-Nearest Neighbour.

$\rightarrow$  Recom. Sys

Pattern. Recog.

{ Use case.

Classification

Regression.

$\rightarrow$  Small DS, non-linear

Intuition  
A data pt. — Similar to neigh. (Major. vte next pt)

No training phase - Lazy algorithm (Distance metrics)

↳ Simple, No Assumption

]- Adv. & D.G.

↳ Slow, noise  $\infty$  Staling  $\rightarrow$  sensitive

3. Decision Tree  $\xrightarrow{\text{use case:}}$  Healthcare Decisions, final Scoring.

classi.      regressi.

Diff:

Rule-based model.

Adv:

$\Rightarrow$  Easy underst., Non-linear Data ✓

$x - \text{Control} = \uparrow \text{overfitting}$

DisAdv:

$\hookrightarrow$  Interpretability ✓

Complex pr. DS ✓

intuition:

Split DS - based on condi.  
(if  $\rightarrow$  else)

4. Random Forest  $\xrightarrow{\text{use case}}$  fnd. Detect. } feature. imp. fsk  
Stack. predic

classi.      regressi.

intuition:

$\Rightarrow$  Build many. D. Tree

Adv:

$\hookrightarrow$  Stable, Accurate, Noise ✓

Slower, interpretability ↓

$\hookrightarrow$  Maj. voted taken

DisAdv:

$\times$  Large-DS - complex alg. ftn.

5. Support vector Machine → use case Bio-inform. img recog, text clothe.

↳ classification

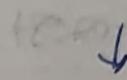
intuition:

find hyperplane w/ max margin b/w classes.

Adv: kernel → non-linear separation.

↳ High-Dimensional Data

red → DS-complex. ptn



Best. preferred.

DisAdv: Large DS - x, tuning ↓

6. Gaussian naive Bayes. use case → Email Spams filtering,

↳ classification (Probabilistic Appr.ch) Document classification

intuition:

Bayes theorem: Assumes features are independent.

Probabilistic classif. → difference

↳ often

unrealistic.

Adv: ↳ very. fast, small Data handling.

Best case: Baseline model.

7. Gradient Boosting classifier: → financial modelling.

↳ classif. & regression.

tabular Datasets.

intuition:

Build tree Sequentially — each new tree fix Prev. error

Boosting ensemble.

Adv:

↳ ↑ Accuracy, Non-linear ptn ✓

DisAdv:

→ slower training, overfitting → prone

Best case:

Accur = imp.

compt. time = accept ✓

# Models & Evaluated metrics on the Dataset

ML Model	Metrics	Accuracy	Precision	Recall	f1-Score
Logistic Regression		0.9722	0.9741	0.9722	0.9720
K-Nearest Neighbour		0.9721	0.9747	0.9722	0.9724
Support vector machine		0.9721	0.9741	0.9722	0.9720
Decision Tree		0.9722	0.9741	0.9722	0.9721
Random Forest		1.0000	1.0000	1.0000	1.0000
Naive Bayes		0.9722	0.9744	0.9722	0.9723
Neural networks		0.9722	0.9741	0.9722	0.9720

Accuracy → Overall correctness

Precision → How many Predicted +ve were

Recall → How many Actual +ve Detected.

f1-Score → Harmonic mean

Why there is a variance →

1. Learn Pattern in Different ways.
2. Noise Handling
3. Models Assume feature relationship
4. Complexity of Decision boundary
5. DS size, complex, scaling, feature Distn.

## Result →

### 1. Ensemble Methods



1. Random forest
2. Gradient Boosting

→ Best Performance.

Reason:

1. ↓ Variance
2. Capture complex pattern

### 2. 1. Logistic Regression

2. Naïve Bayes

→ Moderate Performance

Reason: the Data Set isn't perfectly linear.

### 3. KNN

SVM

{ Accuracy varied because they Depend Strongly  
On feature Scaling and Parameter tuning.