Study on M/L Models: from comparing with Dataset

↓

Wine Quality.

total models trained → 7.

---

1. Logistic Regression. Usecase & Best case → linearly seperable Data.

Linear model → classification. ↓ rel. b/w Data vari.

estimates Probability using Sigmoid func. ←

Adv.

↳ Simple, fast, interpretable.

Dis Adv.

X — Non-linear Pattern Complex

Struggle.

intution → find a Straight line that Seperates classes.

---

2. K-Nearest Neighbour. → Recom. Sys, Patter. Recog. } use case.

Classifi. — regression. → Small DS, non-linear feat.

intution

↳ A Data pt. – Similar to neigh. (major. vote nrst pt)

No training phase – Lazy Algorithm (Distance metrics)

↳ Simple, No Assumption

↳ Slow, noise & Scaling → Sensitive

] – Adv. & Dis.

3. Decision Tree —usecase:→ Healthcare Decisions, final Scoring.

/ \

classi. regressi.

intution:
Split DS - based on condi.
(if → else)

Diff:
Rule-based model.

Adv:
=> Easy understd., Non-linear Data - ✓

x - control = ↑ overfitting

Dis&Adv:
↳ interpretability ✓
complex ptr. DS ✓

---

4. Random Forest —usecase→ frd. Detect. } feature. imp. tsk
Stck. predic

/ \

classi. regressi.

intution:
=> Build many. D. Tree
↳ Maj. votes taken] — Ensemble mthd

Adv:
↳ Stable, Accurate, Noise ✓
Slower, interpretability ↓

Dis&Adv:
x Large. DS - complex ptrn.

5. Support vector Machine —usecase→ Bio inform, img recog, text class.
↳ classification

intution:
↳
find hyperplane w/t max margin b/w classes.

kernel ——→ non-linear Seperation

Adv:
↳ High-Dimensional Data

DisAdv: large DS — x , tuning ↓

red → DS-complex. ptrn
↓
Best. preferred.

---

6. Gaussian naive Bayes. —usecase→ email Spam filtering, Document classification
↳ classification (Probabilistic Appr.ch)

intution:
Bayes theorem: Assumes features are independent.

Probabilistic classifier. → Difference ↳ often unrealistic.

Adv:
↳ Very. fast, small Data Handling.

Best case: Baseline model.

---

7. Gradient Boosting classifier: ——→ financial modelling. tabular Datasets.
↳ class. & regression.

intution:
Build tree sequentially — each new tree fix Previous error

Boosting ensemble.

Adv:
↳ ↑ Accuracy, Non-linear Pttrn ✓

DisAdv. → Slower training, overfitting ↑ Prone

Best case:
Accur = imp.
compt. time = Accpt ✓

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 -

Recall → How many Actual +ve Detected. f1-Score → Harmonic balan

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, feture Distri.

Result →

1. Ensemble methods → Best Performance.
   ↓
   1. Random forest
   2. Gradient Boosting

   Reason:
   1. ↓ variance
   2. Capture complex pattern

2. 1. Logistic Regression
   2. Naive Bayes
   → Moderate Performance
   Reason: the Dataset isn't Perfectly linear.

3. KNN, SVM } Accuracy varied because they Depend Strongly On feature Scaling and Parameter tuning.