| ML model | Assumptions | Advantages | Disadvantages | Feature Scaling | Missing Data | Outliers | Suitable for | Learning | Example Use |
|---|---|---|---|---|---|---|---|---|---|
| Naïve Bayes Classifier | Features are independent | • Performs well with categorical variables<br>• Converges faster: less training time<br>• Good with moderate to large training data sets<br>• Good when dataset contains several features | • Correlated features affect performance | No | Can handle missing data (it ignores missing data) | Robust to outliers | • Classification<br>• Multiclass classification | Supervised | • Sentiment Analysis<br>• Document categorisation<br>• Email Spam Filtering |
| Support Vector Machine (SVM) | None | • Good for datasets with more variables than observations<br>• Good performance<br>• Good of-the-shelf model in general for several scenarios<br>• Can approximate complex non-linear functions | • Long training time required<br>• Tuning is required to determine which kernel is optimal for non-linear SVMs | Yes | Sensitive | Robust to outliers | • Classification<br>• Regression | Supervised | • Stock market forecasting<br>• Value at risk determination |
| Linear Regression | • Linear relation between features and target<br>• Residuals are normally distributed<br>• Homoscedasticity | • Interpretability<br>• Little tuning | • Correlated features may affect performance<br>• Extensive feature engineering required | Yes | Sensitive | Sensitive | Regression | Supervised | • Sales forecasting<br>• House pricing |
| Logistic Regression | • Linear relation between features and the log odds<br>• Residuals are normally distributed<br>• Homoscedasticity | • Interpretability<br>• Little tuning | • Correlated features may affect performance<br>• Extensive feature engineering required | Yes | Sensitive | Potentially sensitive | Classification | Supervised | • Risk Assessment<br>• Fraud Prevention |
| Classification and Regression Trees | None | • Interpretability<br>• Render feature importance<br>• Less data pre-processing required | • Do not predict a continuous output (for regression)<br>• It does not predict beyond the range of the response values in the training data.<br>• Overfits | No | Some implementations do not need missing data imputation. The one in Scikit-learn does | Robust to outliers | • Classification<br>• Regression | Supervised | • Risk Assessment<br>• Fraud Prevention |
| Random Forests | None | • Interpretability<br>• Render feature importance<br>• Less data pre-processing required<br>• Do not overfit (in theory)<br>• Good performance /accuracy<br>• Robust to noise<br>• Little if any parameter tuning required<br>• Apt for almost any machine learning problem | • Do not predict a continuous output (for regression)<br>• It does not predict beyond the range of the response values in the training data<br>• Biased towards categorical variables with several categories<br>• Biased in multiclass problems toward more frequent classes | No | Some implementations do not need missing data imputation. The one in Scikit-learn does. | Robust to outliers | • Classification<br>• Regression | Supervised | • Credit Risk Assessment<br>• Predict breakdown of mechanical parts (automobile industry).<br>• Assess probability of developing a chronic disease (healthcare)<br>• Predicting the average number of social media shares |
| Gradient Boosted Trees | None | • Great performance<br>• Apt for almost any machine learning problem<br>• It can approximate most non-linear functions | • Prone to overfit<br>• Needs some parameter tuning | No | Some implementations do not need missing data imputation (e.g. xgboost). The one in Scikit-learn does. | Robust to outliers | • Classification<br>• Regression | Supervised | • Same as Random Forests |
| K-nearest neighbours | None | • Good performance | • Slow when predicting<br>• Susceptible to high dimension (lots of features) | Yes | Sensitive | Robust to outliers | • Classification<br>• Regression | Supervised | • Gene expression<br>• Protein-protein interaction<br>• Content retrieval (of webpages for example) |
| AdaBoost | None | • It doesn't overfit easily<br>• Few parameters to tune | • Can be sensitive to noise and outliers | No | Can handle | Sensitive | • Classification<br>• Regression | Supervised | • Same as Random Forests, less used however, as xgboost and lightGBMs are more popular implementations of gradient boosted machines |
| Neural Networks | None | • Can approximate any function<br>• Great Performance | • Long training time<br>• Several parameters to tune, including neuronal architecture<br>• Prone to overfit<br>• Little interpretability | Yes | Sensitive | Can handle outliers, and it affects performance if they are too many | • Classification<br>• Regression | Supervised | • Image analysis<br>• Forecasting<br>• Text analysis |
| K-Means Clustering | • clusters are spherical<br>• clusters are of similar size | • Fast training | • Need to determine k, the number of clusters<br>• Sensitive to initial points and local optima | Yes | In the Scikit-learn implementation, missing data needs to be imputed | Sensitive | • Segmentation | Unsupervised | • Customer segmentation<br>• Outlier detection |
| Hierarchical clustering | | • No a priori information about the number of clusters required | • Final number of clusters to be decided by the scientist<br>• Slow training | Yes | Sensitive | Sensitive | • Segmentation | Unsupervised | • Customer segmentation<br>• Gene analyses |
| PCA | • Correlation among features | • Captures most of the variance in a smaller number of features | • Number of principal components that explain most of the variance to be determined by the user | Yes | Sensitive | Sensitive | Reducing feature space to train machine learning models | Unsupervised | • Creating few, informative, variables from tons of data |