



A Comparative Study of Machine Learning Algorithms for Predictive Data Analytics

Ivailo Trivedi-Rao

Department of International Business and Trade, Lagoon Polytechnic of Technology, Maldives

ivailo.trivedi.rao@lpt-mv.net

Peer Review Information

Submission: 10 Jan 2018

Revision: 25 Feb 2018

Acceptance: 12 March 2018

Keywords

*Predictive Analytics;
Machine Learning;
Classification; Model
Comparison; Cross-
Validation*

Abstract

Predictive data analytics uses historical observations to estimate future classes, outcomes, or risks, but the quality of a prediction depends strongly on the learning algorithm, data representation, and evaluation procedure. This study presents a controlled comparison of seven widely used supervised machine learning algorithms: logistic regression, decision tree, random forest, support vector machine (SVM), k-nearest neighbours (k-NN), Gaussian naive Bayes, and gradient boosting. A common experimental pipeline is applied to three established classification datasets distributed with scikit-learn. Scale-sensitive models are standardized, while all models are evaluated using identical stratified five-fold cross-validation splits. Accuracy, macro-averaged F1 score, and model fitting time are used as comparative criteria. The observed results show that SVM, logistic regression, k-NN, and random forest provide consistently strong predictive accuracy in this limited benchmark, while simpler probabilistic and single-tree approaches vary more across datasets. The study also emphasizes that no algorithm is universally superior; algorithm selection should consider data characteristics, computational cost, interpretability, and the cost of prediction errors. The proposed methodology provides a reproducible framework for comparative predictive analytics without relying on a single dataset or a single performance measure.

Introduction

Predictive data analytics combines data preparation, statistical learning, and computational modeling to estimate unknown or future outcomes from observed data. Machine learning has become a central component of predictive systems because it can learn complex decision rules directly from examples instead of requiring every rule to be programmed manually. Domingos (2012) notes that successful machine learning depends not only on the selected algorithm but also on generalization, representation, evaluation, and control of overfitting. Consequently, comparative studies are important because a model that performs well on one dataset may not maintain the same advantage when the data distribution,

dimensionality, noise level, class structure, or sample size changes.

Theoretical and empirical research also cautions against assuming that one learning method will dominate all predictive tasks. Wolpert and Macready (1997) formalized the broader “no free lunch” principle, while Caruana and Niculescu-Mizil (2006) showed through large-scale experiments that algorithm rankings can vary with the selected performance criterion and problem. Hand (2006) further argued that apparent gains from sophisticated classifiers can be overstated when experimental comparisons ignore practical sources of uncertainty. These observations motivate a methodology in which competing algorithms are trained under the

same data partitions and evaluated by more than one indicator.

The objective of this study is therefore to compare representative algorithms from different learning families under a single reproducible predictive-analytics pipeline. The selected models include linear, probabilistic, distance-based, tree-based, kernel-based, and ensemble approaches. The study asks three questions: (i) how consistently do the algorithms perform across datasets with different sizes and class structures; (ii) whether high accuracy is accompanied by similarly high macro-F1 performance; and (iii) what trade-offs arise between predictive performance, computational effort, and interpretability. The contribution is methodological rather than domain-specific: it provides a compact experimental design that can be reused when a practitioner must select an appropriate classifier for a new predictive analytics problem.

Literature Review

Decision trees are among the most interpretable supervised learning methods. Quinlan (1986) described the induction of decision trees from labeled examples and demonstrated how information-based splitting can construct classification rules that are understandable to users. The principal attraction of a tree is its transparent rule structure and ability to model nonlinear boundaries without feature scaling. However, a single tree can be sensitive to small changes in training data and may overfit when allowed to grow without adequate regularization.

Support vector machines introduced a different approach based on maximum-margin decision boundaries. Cortes and Vapnik (1995) showed that nonlinear mappings and kernel functions can enable strong generalization even when the original input space is not linearly separable. SVMs consequently became important for high-dimensional classification, although their performance depends on feature scaling and the choice of kernel and regularization parameters. Their predictive strength must therefore be balanced against reduced transparency compared with a small decision tree or a linear model.

Ensemble learning addresses the instability of individual learners by combining multiple predictive models. Breiman (2001) proposed random forests, which aggregate randomized decision trees and can reduce variance while providing useful estimates of variable importance. Friedman (2001) developed

gradient boosting as stage-wise optimization in function space, showing how sequentially fitted weak learners can form a strong predictive model for classification and regression. These ensemble methods are widely used in predictive analytics because they can capture complex interactions, but they generally increase training cost and reduce direct interpretability.

Empirical benchmarking has repeatedly shown why comparisons must be broad and controlled. Baesens et al. (2003) compared conventional and kernel-based classifiers for credit scoring and found that advanced approaches could perform very well, while simpler models remained competitive. Caruana and Niculescu-Mizil (2006) evaluated multiple supervised learning algorithms with several performance criteria and demonstrated that conclusions depend partly on the chosen metric. Lessmann et al. (2015) later benchmarked a large collection of classifiers across multiple real-world credit datasets and emphasized the value of multi-dataset, multi-metric evaluation for predictive decision support.

Evaluation methodology is equally important. Fawcett (2006) explained the use of receiver operating characteristic analysis and highlighted common pitfalls in classifier assessment. Hand (2006) emphasized that small numerical differences between methods can be less important than uncertainty arising from sampling, population changes, or modeling assumptions. Together with Domingos (2012), these studies indicate that the central methodological requirement is fair evaluation: the same training-test protocol, appropriate preprocessing, clearly stated metrics, and conclusions that remain limited to the evidence produced by the experiment.

Methodology

1. Research Design

The study follows a comparative experimental design. Three multiclass or binary classification datasets available in the scikit-learn distribution were used: Breast Cancer Wisconsin (569 instances, 30 predictors, 2 classes), Wine (178 instances, 13 predictors, 3 classes), and Digits (1,797 instances, 64 predictors, 10 classes). These datasets were selected because they differ in sample size, number of predictors, and class structure, allowing the algorithms to be observed under heterogeneous but manageable conditions. The experiments were executed with Python and scikit-learn 1.8.0 using a fixed random seed of 42 wherever an estimator required randomness.

2. Algorithms and Preprocessing

Seven classifiers were compared: logistic regression, decision tree, random forest, SVM with a radial basis function kernel, k-NN with $k = 5$, Gaussian naive Bayes, and gradient boosting. The set covers distinct inductive biases rather than many minor variants of one technique. Standardization was applied inside the validation pipeline for logistic regression, SVM, and k-NN because these methods depend on feature magnitudes or distances. Tree-based models and Gaussian naive Bayes were evaluated on the original feature scale. Default or conservative standard settings were retained to keep the comparison reproducible and to avoid giving one algorithm an advantage through extensive tuning.

3. Validation and Evaluation Procedure

For each dataset, stratified five-fold cross-validation with shuffling was used. In every fold, four partitions were used for model fitting and the remaining partition for evaluation; the process was repeated until each partition served once as the validation fold. Crucially, the same fold-generation strategy was applied to every

algorithm. The primary measure was classification accuracy. Macro-averaged F1 was added because it assigns equal importance to each class and therefore provides a useful complement when class frequencies differ. Mean fitting time was recorded as a simple indicator of computational demand. Although ROC/AUC is an established evaluation tool (Fawcett, 2006), macro-F1 was preferred here because the benchmark includes both binary and multiclass datasets and the goal was to retain a single directly comparable secondary measure.

The interpretation strategy follows prior comparative research (Caruana and Niculescu-Mizil, 2006; Lessmann et al., 2015): performance is examined across datasets rather than from one isolated score. Differences are treated descriptively because only three datasets are used, which is too small a sample for a strong cross-dataset significance claim. This restriction is important because the no-free-lunch perspective implies that any apparent advantage must be tied to the tested problem distribution rather than presented as a universal ranking (Wolpert and Macready, 1997).

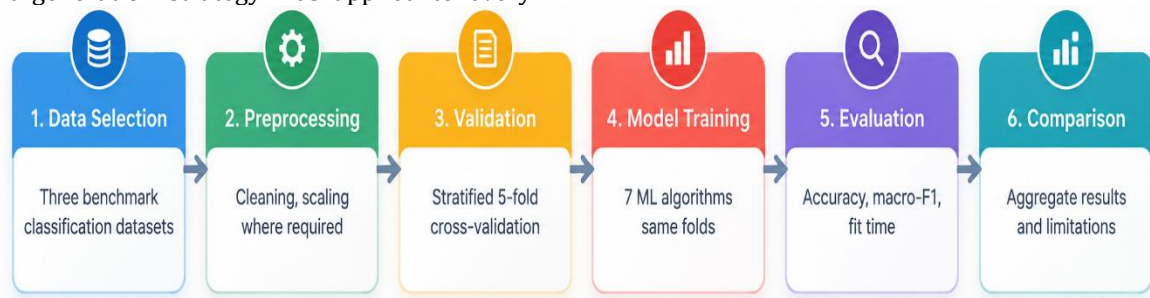


Figure 1. Comparative machine learning methodology used in the study.

Results

Table 1 reports mean five-fold cross-validation results. Under this specific experimental setting, SVM produced the highest overall mean accuracy (98.15%) and macro-F1 (98.09%). Logistic regression followed closely with a mean accuracy of 97.60%, while k-NN and random forest also

exceeded 97% mean accuracy. The gap between the leading methods is small compared with the larger decline observed for a single decision tree and Gaussian naive Bayes on the Digits dataset. This pattern supports the general lesson that algorithm behavior is data-dependent rather than fixed across tasks.

Table 1. Five-fold cross-validation performance of the compared algorithms

Algorithm	Breast Cancer Accuracy	Wine Accuracy	Digits Accuracy	Mean Accuracy	Mean Macro-F1	Mean Fit Time (s)
SVM (RBF)	97.72%	98.33%	98.39%	98.15%	98.09%	0.0149
Logistic Regression	97.37%	98.33%	97.11%	97.60%	97.51%	0.0191
k-NN	96.31%	97.17%	97.89%	97.12%	97.03%	0.0025
Random Forest	95.43%	97.75%	97.89%	97.02%	96.94%	0.6014
Gradient Boosting	94.91%	92.14%	96.38%	94.48%	94.41%	3.1461
Gaussian Naive Bayes	93.85%	97.19%	84.53%	91.86%	91.75%	0.0015
Decision Tree	91.04%	89.32%	85.47%	88.61%	88.45%	0.0126

The SVM results are consistent with the strong generalization properties associated with maximum-margin learning (Cortes and Vapnik, 1995), while the random forest results illustrate the stability benefits expected from tree aggregation (Breiman, 2001). Logistic regression remained highly competitive despite its simpler functional form. This observation is important in applications in which explanation, coefficient inspection, implementation cost, or probability estimation matters as much as a small change in classification accuracy. It also aligns with Hand's (2006) warning that greater model complexity does not automatically translate into practically meaningful superiority.

Gradient boosting achieved strong accuracy but required the highest average fitting time in this benchmark. Its performance was also less consistent on the small Wine dataset than SVM or logistic regression. The result does not contradict the effectiveness of boosting described by Friedman (2001); rather, it shows that untuned models may react differently to dataset size and structure. Likewise, Gaussian naive Bayes performed very well on Wine but substantially worse on Digits, reflecting the consequences of its distributional and conditional-independence assumptions. The decision tree produced the lowest aggregated accuracy, but it remains attractive where transparent rules and low prediction complexity are operational priorities.

The mean fitting times reinforce the same trade-off. k-NN and naive Bayes required little fitting work, whereas ensemble methods were more computationally demanding. Random forest nevertheless delivered a favorable balance between predictive quality and computational cost, while gradient boosting was the most expensive among the tested settings. The practical implication is that model selection should be based on a joint assessment of predictive quality, resource requirements, stability, and interpretability, as also suggested by empirical benchmarking literature (Baesens et al., 2003; Lessmann et al., 2015).

Conclusion and Discussion

This methodology paper compared seven supervised machine learning algorithms for predictive data analytics using a common preprocessing and stratified five-fold cross-validation framework. The experiment showed that SVM, logistic regression, k-NN, and random forest were the most consistent methods across the three tested datasets. SVM achieved the highest aggregated accuracy and macro-F1 in this benchmark, but the differences among the

strongest models were relatively small. The results therefore support selecting algorithms on evidence from the target data rather than choosing a method solely because it is newer or more complex. From a methodological perspective, three practices are central. First, comparisons should use identical data partitions and preprocessing rules so that performance differences are attributable to the learners rather than to experimental inconsistency. Second, more than one metric should be examined because accuracy alone can hide class-specific behavior; macro-F1 provided a useful secondary view in the present study, while ROC/AUC may be preferable in applications where ranking performance and decision thresholds are central (Fawcett, 2006). Third, computational cost and interpretability should be reported alongside predictive quality, especially when the difference in predictive performance is small. The study has limitations. Only three benchmark datasets and standard model settings were used, no exhaustive hyperparameter search was performed, and fitting times depend on hardware and software configuration. Therefore, the numerical results should not be generalized as a universal ranking. The no-free-lunch argument (Wolpert and Macready, 1997) and empirical comparison literature (Caruana and Niculescu-Mizil, 2006; Lessmann et al., 2015) both support this cautious interpretation. Future work can extend the same framework to larger real-world datasets, regression tasks, class imbalance, nested cross-validation, calibration, feature selection, and cost-sensitive evaluation. The central conclusion is that reliable predictive analytics depends as much on experimental design and evaluation discipline as on the learning algorithm itself.

References

- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6), 627–635. <https://doi.org/10.1057/palgrave.jors.2601545>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, 161–168. <https://doi.org/10.1145/1143844.1143865>

- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. <https://doi.org/10.1007/BF00994018>
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87. <https://doi.org/10.1145/2347736.2347755>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1), 1–14. <https://doi.org/10.1214/088342306000000060>
- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1, 81–106. <https://doi.org/10.1007/BF00116251>
- Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82. <https://doi.org/10.1109/4235.585893>