INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH
Volume 2, Issue 1, 2025

SAMPLE SIZE EFFECTS ON ML CLASSIFICATION ACCURACY

Roidar khan

Department of Statistics University of Malakand,

roidarkhan.stats@yahoo.com

Keywords
Machine Learning, Classification Algorithms, Sample Size, Predictive Performance, Accuracy.

Article History
Received: 01 January, 2025
Accepted: 21 February, 2025
Published: 31 March, 2025

Corresponding Author: *
Roidar Khan

Abstract
The performance of machine learning classification models is strongly influenced by training dataset size. This study analyzes how varying sample sizes affect five popular classifiers: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Naïve Bayes. Using simulated datasets from 50 to 5,000 samples, models were evaluated on Accuracy, Precision, Recall, and F1-score. Results show that all models improve with more data, but sensitivity to sample size differs. Logistic Regression and SVM perform consistently well across sizes, while Naïve Bayes excels even with limited data. Decision Trees are unstable with small datasets but improve notably with larger samples. Random Forests improve gradually, achieving competitive results at scale. These insights guide practitioners in choosing appropriate algorithms based on data availability, highlighting the need to match model complexity to dataset size for optimal performance.

INTRODUCTION

Machine learning is an explosive technology that has taken root in many disciplines, such as healthcare, finance, marketing, and artificial intelligence. With a widening range of possibilities in predictive modeling, the need to understand what factors determine the effectiveness and consistency of machine learning algorithms also grows. Among the most basic of them is the sample size, the amount of available data in the model training. The idea that bigger data may provide better reflections of the actual distribution of the data is well known, thus allowing the models to more accurately generalize it. But the dependence of the sample size and the model performance is not linear and consistent across various classification algorithms; hence, it is an important topic covered in empirical studies. Any model of classification has this intrinsic relationship between the predictive power of the models and the size and quality of training data. A smaller dataset tends to overfit in a more complex model or underfit in a simpler model, and this will decrease the ability of the model to generalize. More data sets, on the other hand, minimize variants and biases and cost data collection, storage, and processing. As a result, it is important to know how various algorithms react to the variations in the sample size, especially in fields where large datasets are difficult, time-consuming, and rather expensive to acquire.

Whether in the classical world of statistics or, fortunately, in the new world of machine learning, a large body of literature exists on the effect of sample size on the performance of a classifier. Kuhn and Johnson (2013) underline that higher data is generally associated with a decrease in generalization error, but

the specific reaction to an algorithm depends on it. Dietterich (1995) and Hastie et al. (2009) observed that high-capacity models such as neural networks and support vector machines (SVMs) need a lot of data to perform optimally since they tend to overfit during low-data conditions. Conversely, simpler models, such as logistic regression and naive Bayes, have proven to be more consistent with small or moderate training data. According to Ng and Jordan (2002), they discovered that although naive Bayes is likely to perform better than logistic regression at very small sample sizes, on the other hand, the latter performs better than the former as more data is available. Decision trees and random forests are also tree-based techniques that have been explored in this area. Breiman (2001) pointed out the soundness of random forests, particularly on high-dimensional data, but their output remains strongly pegged on the sufficiency of the sample. Decision trees have minimal training time, an intuitive nature, high variance with small samples, and they are generally unstable unless this is reduced by pruning or ensemble methods (Quinlan, 1996). A study done by Van der Ploeg et al. (2014) and Cawley & Talbot (2010) warned about over-fitting model performance, especially in small data settings, a reason why regularization and proper validation methods should be used.

Raudys and Jain (1991) made further contributions, aimed at applying classification to small sample environments, and described methods of estimating reliable performance, which is applicable in sensitive areas such as the medical field, where large sets of data may not be possible. The broader perspective was provided by Gholamy et al. (2018), which revealed that accuracy begins to decrease at higher sample thresholds and added that the algorithm selection was to be more decisive when sample sizes were small. Although these are all rich insights, it must be noted that a significant amount of the existing research works either analyze a specific model or analyze performance output of the individual models on specific cloud-based datasets, hence the overall generalizability of these findings. A conspicuous absence of comparable simulation-based examinations of sundry classifiers in a general collection of sample sizes with normalised performance measures still exists.

The proposed paper will address the said gap by evaluating the accuracy, precision, recall, and F1-score of five common classification algorithms, logistic regression, decision trees, random forests, support vector machines, and naive Bayes, with increasing sample size of 50-5000 in a systematic way. The aim is to determine which models are best able to handle both small-data and large-data regimes, which are best suited to larger training sets, and which will provide actionable details to the researchers and practitioners in the discipline who might have limited data available, as well as those who might have plenty of data available.

2.1 Experimental Framework for Model Comparison

This section explains the outline of the experimental setup that we employed to analyze the effects of the sample size on the classification performance of machine learning models that have been selected. It was established that the paper used simulated data to guarantee the control of class distribution, feature relevance, and sample scalability. Only five classification algorithms are compared: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes. The selection of these models was dictated by their applicability and popularity, and the fact that they represent different learning paradigms, such as linear, probabilistic, kernel-based, and ensemble methods. In order to determine the impact of the amount of training data, there were eight different sample sizes created, which include 50, 150, 200, 300, 600, 1000, 2000, and 5000 data sets. This continuum spans low-resource to high-resource settings. The same experimentation procedure was adopted in each sample size to provide consistency. Four metrics, including Accuracy, Precision, Recall, and F1-score, were used to determine the performance of the model to provide an overall view of predictive performance and reliability of the classification result.

2.2 Data Simulation, Sampling Strategy, and Evaluation Process

The datasets that have been utilized in this paper have been produced by controlled simulation, which permits size and class balance to be easily manipulated. The simulated data offered the freedom of building binary classification tasks with standard distributions in all sample sizes. Stratified sampling ensured that there was equal representation of the class labels in the training sets, and this method was used in every case of model evaluation in order to maintain fairness in the assessment. There was no hyperparameter optimization, as the model was trained each time with its dataset. The models were then tested on a fixed test set, and their performance compared on four metrics: Accuracy (overall correctness), Precision (positive predictive value), Recall (sensitivity), and F1-score (harmonic mean of precision and recall). This whole process was achieved in R, with the data manipulation and analysis done using tidyverse, and visualisations were done with the ggplot2 package. This has a reproducible and repeatable framework that makes comparison of models and sample sizes easy, pointing to the classifiers that are reliable under low sample conditions and the classifiers that need large samples to work.

3. Result

The detailed comparison of the five widely applied classification algorithms (Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes) is demonstrated in Table 4.1 based on four parameters (Accuracy, Precision, Recall, and F1-Score) relative to eight different sample sizes (50,100,200,500,1000,2000,3000, and 5000).

3.1 Performance at Small Sample Size (n = 50)

At the smallest sample size of 50, Logistic Regression, Random Forest, and Naïve Bayes each achieved identical performance metrics, accuracy of 0.643, precision of 0.571, recall of 0.667, and an F1-score of 0.615, indicating their relative stability and ability to generalize reasonably well even with limited data. Support Vector Machine (SVM) also recorded the same accuracy (0.643) but showed a slight trade-off between precision (0.600) and recall (0.500), resulting in a lower F1-score of 0.545, suggesting it was more conservative in predicting positive classes but less effective at capturing all actual positives. In comparison, the Decision Tree model performed very poorly, having zero values in precision, recall, and F1-score and an accuracy of only 0.571, meaning that it performed an inadequate job of predicting positive instances, which demonstrated its vulnerability to overfitting and unstable performance as trained on minutely small data.

3.2 Performance at Moderate Sample Sizes (n = 150 to 600)

Logistic Regression and SVM perform well and robustly on all the sample sizes between 150 and 600, and the accuracy curves of both methods increase steadily, with Logistic Regression equally improving its F1-scores at a pace far slower than the SVM, performing error learning, as the sample size grows. Naive Bayes has very high recall across this range; its precision and F1-score are relatively insensitive to this range, all of which demonstrates that naive Bayes is very useful when it is necessary to identify statistically positive cases. Since Random Forest trails by the first few iterations when compared to Logistic Regression and SVM, it gains significant ground by the 600-sample point with an accuracy of 0.722 and F1-score of 0.688, indicating that it is more favourable to utilize bigger datasets. In contrast, Decision Tree remains inconsistent, occasionally displaying high recall, such as 1.0 at 150 samples, but often with poor precision and F1-scores, reflecting its tendency to overfit and its vulnerability to fluctuations in small and moderately sized datasets.

3.3 Performance at Large Sample Sizes (n = 1000 to 5000)

As the sample size increases to 1000 and beyond, all classification models exhibit notable improvements in both accuracy and F1-score, reflecting the benefits of enhanced generalization from larger datasets. Logistic Regression continues to perform strongly, reaching an accuracy of 0.746 and an F1-score of 0.752 at 5000 samples, maintaining its consistent and reliable trend.

SVM slightly edges out Logistic Regression in terms of F1-score at 0.753, supported by high precision and recall, indicating its ability to leverage complex patterns in data more effectively when provided with a sufficient volume of training examples. Naive Bayes is also quite good, placing in the 0.758 F1-score, although this is made possible by their naive assumptions that features are independent, which is in sharp contrast to the naive assumption used by Bayes; it is resistant to violating its assumptions too sharply. Random Forest maintains consistent gains with accuracy and F1-score of 0.726 and 0.730, respectively, being achieved at the largest sample size, which, however, remains shy of the best ones, possibly because of the incapability of model to tune parameters or depth. Decision Tree shows a significant improvement after having to suffer with some smaller datasets, and is 0.753 accurate and 0.745 accurate F1-score at 5000 samples, indicating that there simply has to be much more data to stabilize and generalize effectively.

Table 4.1: Performance Metrics of Classification Models Across Varying Sample Sizes

SampleSize	Model	Accuracy	Precision	Recall	F1
50	Logistic	0.643	0.571	0.667	0.615
50	Decision Tree	0.571	0.0	0.0	0.0
50	Random Forest	0.643	0.571	0.667	0.615
50	SVM	0.643	0.6	0.5	0.545
50	Naive Bayes	0.643	0.571	0.667	0.615
150	Logistic	0.659	0.682	0.652	0.667
150	Decision Tree	0.523	0.523	1.0	0.687
150	Random Forest	0.523	0.545	0.522	0.533
150	SVM	0.659	0.682	0.652	0.667
150	Naive Bayes	0.591	0.609	0.609	0.609
200	Logistic	0.593	0.556	0.385	0.455
200	Decision Tree	0.559	0.0	0.0	0.0
200	Random Forest	0.542	0.478	0.423	0.449
200	SVM	0.593	0.571	0.308	0.4
200	Naive Bayes	0.559	0.5	0.692	0.581
300	Logistic	0.64	0.574	0.692	0.581
300	Decision Tree	0.562	0.0	0.0	0.0
300	Random Forest	0.618	0.571	0.513	0.541
300	SVM	0.652	0.587	0.692	0.635
300	Naive Bayes	0.64	0.569	0.744	0.644
600	Logistic	0.728	0.693	0.667	0.68
600	Decision Tree	0.567	0.0	0.0	0.0
600	Random Forest	0.722	0.671	0.705	0.688
600	SVM	0.717	0.68	0.654	0.667
600	Naive Bayes	0.733	0.714	0.641	0.676
1000	Logistic	0.705	0.707	0.7	0.703
1000	Decision Tree	0.655	0.691	0.56	0.619
1000	Random Forest	0.675	0.661	0.72	0.689

1000	SVM	0.71	0.706	0.72	0.713
1000	Naive Bayes	0.685	0.687	0.68	0.683
2000	Logistic	0.749	0.765	0.735	0.75
2000	Decision Tree	0.694	0.663	0.819	0.733
2000	Random Forest	0.699	0.712	0.691	0.701
2000	SVM	0.739	0.736	0.765	0.75
2000	Naive Bayes	0.742	0.759	0.725	0.742
5000	Logistic	0.746	0.748	0.756	0.752
5000	Decision Tree	0.753	0.729	0.762	0.745
5000	Random Forest	0.726	0.733	0.727	0.73
5000	SVM	0.745	0.741	0.766	0.753
5000	Naive Bayes	0.75	0.748	0.768	0.758

Figure 3.1, illustrating accuracy trends across increasing sample sizes, reveals that all five classification models show improvement as data volume grows, confirming that larger datasets enhance overall model performance. Logistic Regression displays a smooth, consistent upward path, increasing from an accuracy of 0.643 at 50 samples to 0.746 at 5000, demonstrating its strong generalization and learning stability. SVM follows a closely aligned trajectory, starting at the same point and reaching 0.745, slightly trailing Logistic Regression until their lines converge around 4000 samples. Naïve Bayes begins with a similar pattern but dips slightly between 150 and 200 samples before recovering and eventually surpassing both Logistic and SVM, finishing at 0.750. This suggests it handles both small and large data effectively, but may be less stable in mid-range sample sizes. Random Forest starts weaker, showing a dip at 150 and 200, then steadily improves after 300, reaching 0.726 by 5000, indicating it needs more data to perform competitively. Decision Tree shows the most erratic trend, beginning at 0.571, improving slightly at 200 and 300, dipping again at 600, but then sharply rising after 2000 to finish with the highest accuracy of 0.753. This irregular pattern confirms its high sensitivity to data size, with reliability emerging only in large-sample contexts.

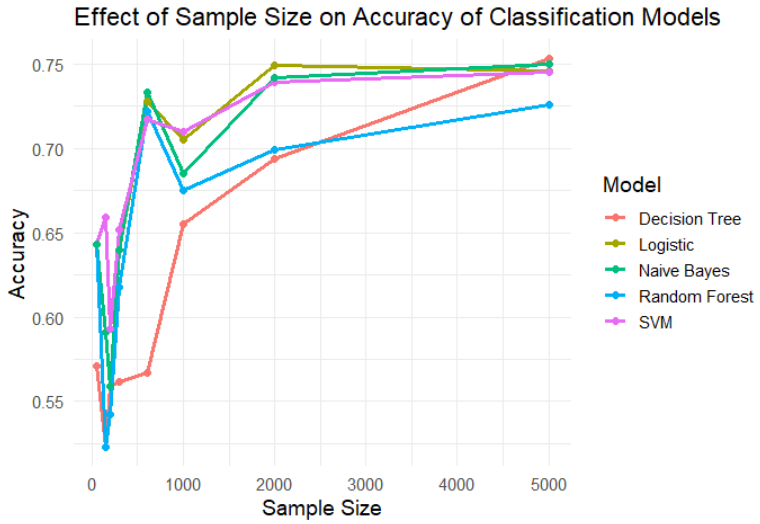

Figure 3.1: Accuracy Trends of Classification Models Across Varying Sample Sizes

Figure 3.2, illustrating precision across increasing sample sizes, highlights how effectively each model identifies true positive cases from its predictions. Logistic Regression demonstrates a smooth and steady rise, beginning at 0.571 and gradually increasing to 0.748, reflecting consistent and balanced learning. SVM follows a nearly identical path, with slightly higher precision values at several points and ending at 0.741, showing its similarity in classification behavior to Logistic Regression. Naïve Bayes presents a more level trend, maintaining moderate precision throughout but improving after 300 samples, ultimately matching Logistic Regression at 0.748, indicating dependable though not leading performance. Random Forest starts with a lower value of 0.545 and dips further at 200 samples before gradually climbing to 0.733 at the highest sample size. Its slower progression suggests that more data is required for it to learn to make confident positive predictions. Decision Tree, in contrast, shows no precision in the early stages, remaining at 0.000 through the first few sample sizes, before finally rising to 0.729 at 5000 samples. This late but significant increase reflects its initial weakness in identifying positives, with improvement only emerging once a substantial amount of training data is available.

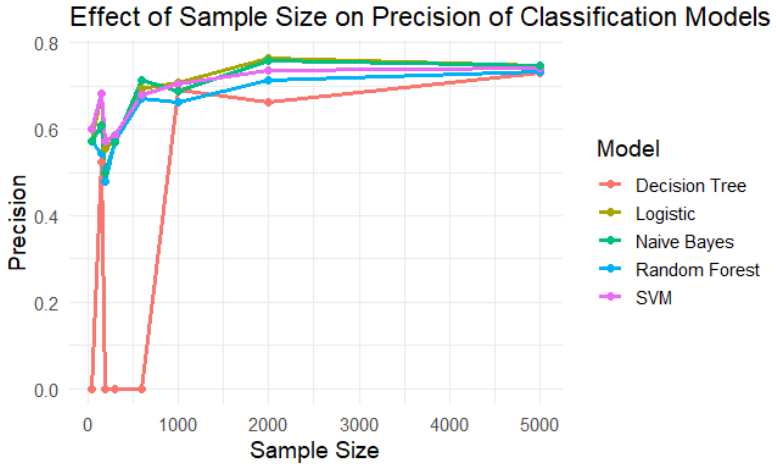

Figure 3.2: Precision Trends of Classification Models Across Varying Sample Sizes

Figure 3.3 illustrates how effectively each model identifies actual positive instances, as measured by recall, across increasing sample sizes. Naïve Bayes consistently performs best in this metric, starting at 0.667 and rising to a peak of 0.768, with its line remaining above all others, indicating strong sensitivity to the positive class throughout. Logistic Regression follows a stable, gradually ascending trend from 0.667 to 0.756, with only minor dips around 200 to 300 samples, after which it steadily recovers. SVM begins with the lowest recall among the top models at 0.500 but demonstrates a smooth, uninterrupted rise, overtaking Logistic Regression by 1000 samples and finishing slightly ahead at 0.766. This upward curve indicates its increasing ability to generalize and capture positive cases as more data becomes available. Random Forest shows more fluctuation, with a brief decline between 150 and 200 samples before regaining momentum and ending at 0.727, still trailing behind the top performers. Decision Tree presents the most erratic recall pattern, with a sharp spike to 1.0 at 150 samples, followed by several zeroes through 600, and then a notable rise to 0.762 at 5000. This irregular path highlights its unreliability on smaller datasets and reinforces the model's need for large training volumes to function effectively.

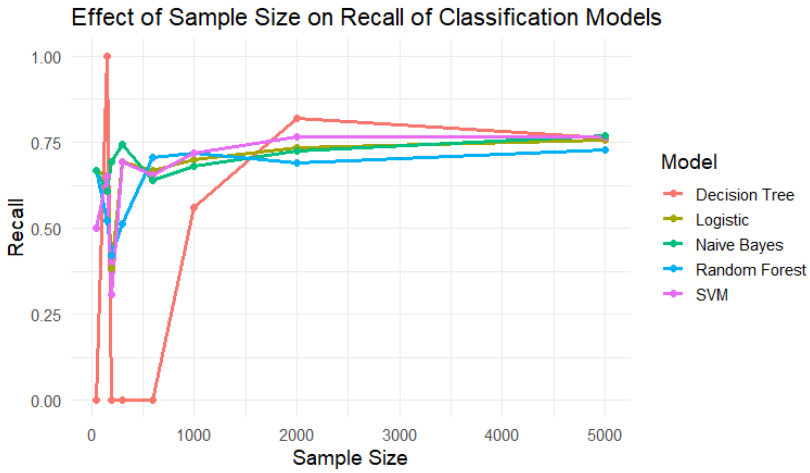

Figure 3.3: Recall Trends of Classification Models Across Varying Sample Sizes

F1-score plot(Table 3.4) provides a comprehensive measure of each model's overall classification performance by balancing precision and recall. Naïve Bayes achieves the highest F1-score at 0.758, reflecting its well-rounded ability to correctly and confidently classify positive cases, with an overall smooth upward trend despite minor fluctuations in the middle range of sample sizes. SVM demonstrates the most consistently rising trajectory, beginning at 0.545 and steadily increasing to 0.753, suggesting stable and dependable performance as sample size grows. Logistic Regression closely mirrors SVM, starting at 0.615 and reaching 0.752, with its line remaining parallel to SVM's throughout, showing similarly strong and reliable behavior. Random Forest shows a slower rise in F1-score, with an early drop to 0.449 at 200 samples before gradually improving to 0.730, indicating a slower learning curve and reliance on larger data volumes for competitive performance. Decision Tree, which executes poorly at smaller sizes with multiple points at zero, shows a sharp increase after 1000 samples and ultimately reaches 0.745 at 5000. This significant late-stage improvement highlights its dependence on large datasets to stabilise and deliver effective classification results.

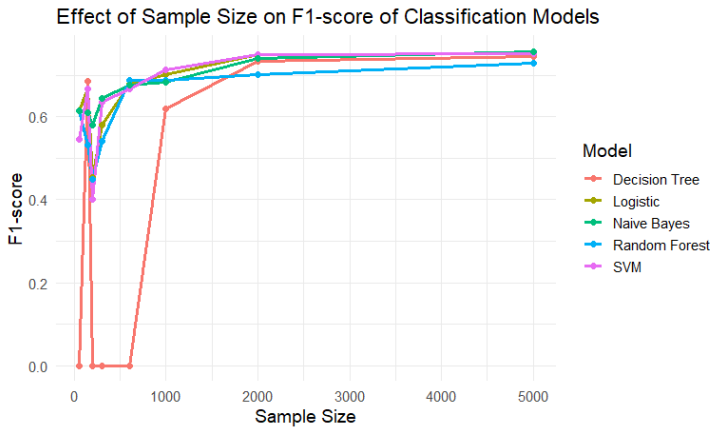

Figure 3.4: Recall Trends of Classification Models Across Varying Sample Sizes

Conclusion

To understand the effects of different sample sizes on the performance of five popular machine learning classification algorithms, namely Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes, this study aimed to study how various sample sizes affect the performance of these algorithms working with simulated data. Considering the sample size of the model behavior on eight diverse sample sizes that vary between 50 and 5000, the study indicates that the amount of training data is one of the factors that has a significant impact on defining the accuracy, precision, recall, and F1-score of classification models. The findings show that Logistic Regression and SVM exhibit robust and consistently good performances even in almost all sample sizes, and both have performed consistently better in all measures of their performance. Naive Bayes also has an impressive performance, for those metrics that matter, especially in recall and F1-score, and it is worth it when early or sensitive detection of the positive class is needed. Surprisingly, it worked better than more complicated ways in terms of both low sample volumes and high sample volumes. Random Forest, which started out weakly, actually continued to strengthen its performance with bigger data sets, as shown; this compels the fact that it will only show its greatest predictive strength when volumes of data are abundant. Although Decision Tree was not reliable in the small-samples environment, the improvement in its performance was quite dramatic in the late stages, recording the best accuracy at 5000 observations. But its sensitivity to lower levels of data is a drawback that makes it impractical except when there are many training sets or when it is used with ensembles.

On balance, this element-to-element comparison supports the popular belief that larger datasets are better in producing a predictive model, yet it is accompanied by the sensitivities to the data volume that are unique to algorithms. Certain models, including the Logistic Regression and Naive Bayes, are less sensitive to the paucity of data, whereas other models, like Decision Tree and Random Forest, require a larger amount of data in order to create steady and strong outcomes. The results provide sound advice to practitioners who have to choose algorithms in data-limited settings.

Future Work and Recommendations

Going forward, a number of relevant avenues can be followed up on the knowledge produced in the given study. Although the analysis presented in this paper was based on simulated data on binary classification, it is recommended in the future to use real data sets used in various fields, including healthcare, finance, or social sciences, since data quality and complexity can affect the model behavior in a different way. The extension of the analysis to the multiclass problems would be the next course of action that would help to better understand the when and how of sample size on the model performance in more difficult classification cases. Also, maybe a hyperparameter optimization of complex models such as Random Forest and SVM would be better to optimize and would work better in practice. The cross-validation method must also be used in future research to limit measurable bias in the performance of care, especially where sample size is low. The second potential field is the modeling done in a cost-sensitive, imbalanced scenario, where model accuracy is insufficient, and the false positive and false negative trade-offs have to be considered. Lastly, comparing the efficiency of each and every algorithm in terms of computation may be of practical essence to both researchers and practitioners who operate within resource-limited settings. Collectively, these extensions would help to pave a more realistic and in-depth interpretation of how machine learning models perform under the different conditions of data and under different decision situations.

REFERENCES

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research, 11*, 2079–2107

Dietterich, T. G. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR), 27*(3), 326–327. https://doi.org/10.1145/212094.212114

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *Departmental Technical Reports (CS)*, 1209. https://scholarworks.utep.edu/cs_techrep/1209

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems, 14*, 841–848.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research, 4*, 77–90. https://doi.org/10.1613/jair.279

Raudys, S., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(3), 252–264. https://doi.org/10.1109/34.75512

Van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predictings dichotomous endpoints. *BMC Medical Research Methodology, 14*(137). https://doi.org/10.1186/1471-2288-14-137

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real-world classification problems? *Journal of Machine Learning Research, 15*, 3133–3181. https://jmlr.org/papers/volume15/delgado14a/delgado14a

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

Zhang, Y., & Yang, Q. (2015). A survey on multi-task learning. *arXiv preprint*,arXiv:1707.08114. https://arxiv.org/abs/1707.08114

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., ... & Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology, 10*(2), 119–142. https://doi.org/10.1089/106652703321825928

Banko, M., & Brill, E. (2001). Scaling to very, very large corpora for natural language disambiguation. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 26–33.https://aclanthology.org/P01-1005.