COGS 118A: Final Project

Jiafeng Wu
UCSD

Abstract

This paper builds upon Caruana and Niculescu-Mizil's 2006 study, *An Empirical Comparison of Supervised Learning Algorithms*, which evaluated the performance of various machine learning algorithms across multiple datasets and metrics. This study analyzes the performance of three algorithms—K-Nearest Neighbors (KNN), Decision Trees (DT), and Random Forests (RF)—on the Wine Quality, Abalone, and Student Performance datasets. Each model was optimized through hyperparameter tuning using grid search. The results demonstrate that Random Forest has best performance and test accuracy is higher when having more training data. These findings align with those of the original study.

1. Introduction

Classification is an important application of machine learning. This study experiments on the performance of three classifiers—Decision Trees (DT), K-Nearest Neighbors (KNN), and Random Forests (RF)—using datasets with varying characteristics. The selected datasets—Wine Quality, Abalone Age Prediction, and Student Performance—represent variation in features, sample sizes, and distributions. All datasets were sourced from the UCI Machine Learning Repository. Inspired by Caruana and Niculescu-Mizil's work, this study aims to compare model performance based on accuracy, with metrics such as precision, recall, and confusion matrices to evaluate. The study also explores the effects of hyperparameter tuning and train-test splits on the performance of each model.

2. Methodology

This experiment evaluates the performance of three supervised learning algorithms using accuracy, precision, recall, and confusion matrix metrics. The datasets were preprocessed to handle missing values and ensure compatibility with the classifiers.

Datasets:

1. Wine Quality Dataset has 4,898 samples and 12 features. Target variable (wine quality) is converted into a binary classification: scores ≥7 (high quality) and

scores <7 (low quality). It is the largest dataset used in the study with imbalanced data.
2. Abalone Dataset has 4,177 samples and 9 features. Target variable ("rings") is converted into a binary classification task: rings ≥10 (older) and rings <10 (younger).
3. Student Performance Dataset has 395 samples and 33 features. Target variable (G3, final grade) is transformed into a binary classification: scores ≥10 (pass) and scores <10 (fail). It is the small dataset used with imbalanced distribution.

Data preprocessing performs replacing for missing values and standardization using StandardScaler. Categorical variables in the Abalone dataset ("Sex") are encoded using LabelEncoder, and the Student Performance dataset uses one-hot encoding.

Learning Algorithms and Hyperparameters:

1. Decision Trees (DT) uses max_depth (5, 10, 15) and min_samples_split (2, 5, 10).
2. K-Nearest Neighbors (KNN) uses n_neighbors (3, 5, 7, 9), weights ("uniform" and "distance"), and p (1 for Manhattan, 2 for Euclidean).
3. Random Forest (RF) uses n_estimators (50, 100, 150), max_depth (10, 20, 30), min_samples_split (2, 5, 10), and max_features ("sqrt" and "log2").

Hyperparameter tuning was conducted using GridSearchCV with five fold cross validation. The best parameters were chosen, ensuring optimal performance.

3. Experiment and Results

The results of the experiments are summarized in tables that show training, validation, and test accuracy, along with precision, recall, and error rates for all classifiers across the three datasets. Results demonstrate that Random Forest consistently performs better than Decision Trees and KNN across all datasets and partitions. An increase in test accuracy with additional training data is also observed.

Classifier	Train/Test Split	Train Accuracy	Training Error	Validation Accuracy	Validation Error	Test Accuracy	Test Error	Precision	Recall
Decision Tree	80/20	0.9748	0.0252	0.8127	0.1873	0.8248	0.1752	0.5651	0.5937
	50/50	0.9138	0.0862	0.8100	0.1900	0.8006	0.1994	0.5507	0.4600
	20/80	0.8965	0.1035	0.7988	0.2012	0.7923	0.2077	0.5662	0.3704

KNN	80/20	1.0000	0.0000	0.8636	0.1364	0.8728	0.1272	0.7205	0.6326
	50/50	1.0000	0.0000	0.8474	0.1526	0.8516	0.1484	0.6733	0.5818
	20/80	1.0000	0.0000	0.8202	0.1798	0.8126	0.1874	0.5937	0.4510
Random Forest	80/20	0.9986	0.0014	0.8732	0.1268	0.8718	0.1282	0.8054	0.5760
	50/50	0.9977	0.0023	0.8511	0.1489	0.8632	0.1368	0.7698	0.5174
	20/80	0.9942	0.0058	0.8318	0.1682	0.8290	0.1710	0.6708	0.4160

Table 1. Training, validation, and test accuracy, error, recall, and precision for DT, KNN, and RF across all partitions of the Wine Quality dataset.

Classifier	**Split**	**Predicted Negative (Actual Negative)**	**Predicted Positive (Actual Negative)**	**Predicted Negative (Actual Positive)**	**Predicted Positive (Actual Positive)**
DT	80/19	2071	273	242	354
	50/50	5144	599	866	738
	20/80	8357	821	1621	958
KNN	80/19	2180	150	224	386
	50/50	5354	441	649	903
	20/80	8403	801	1402	1151
RF	80/19	2177	93	284	386
	50/50	5526	244	761	816
	20/80	8688	524	1487	1058

Table 2. Confusion matrix across all conditions for the Wine Quality dataset.

Wine Quality Dataset

Random Forest outperformed DT and KNN in two of the three partitions (Table 1). In the 80/20 split, KNN has the highest test accuracy of 87.28% with a test error of 12.72%, followed by Random Forest with 87.18% test accuracy and 12.82% test error. Decision Tree has a lower score, with 82.48% test accuracy and 17.52% test error. In the 50/50 split, Random Forest has the highest score with 86.32% test accuracy and 13.68% test error. KNN and Decision Tree followed, with test accuracies of 85.16% and 80.06%, respectively. In the 20/80 split, Random Forest again demonstrated better performance with 82.90% test accuracy and 17.10% test error. KNN and Decision Tree showed declines in this partition, having 81.26% and 79.23% test accuracies, respectively, indicating challenges dealing with reduced training data. Random Forest

consistently achieved the highest validation accuracy across all partitions compared to the others. It also has the best balance between precision and recall in the 80/20 split, with 80.54% precision and 57.60% recall, minimizing false negatives. In contrast, Decision Tree has lower performance, with precision of 55.07% and recall dropping to 37.04% in the 20/80 split, reflecting its difficulty with generalization. Random Forest had the lowest number of false negatives and false positives across all partitions (Table 2). For instance, in the 80/20 split, it misclassified 284 actual positives as negatives, compared to Decision Tree's 242 false negatives and KNN's 224. In the 20/80 split, Random Forest misclassified 1,487 false negatives, fewer than Decision Tree's 1,621 but slightly more than KNN's 1,402.

Classifier	Train/Test Split	Train Accuracy	Training Error	Validation Accuracy	Validation Error	Test Accuracy	Test Error	Precision	Recall
Decision Tree	80/20	0.8075	0.1925	0.7815	0.2185	0.7675	0.2325	0.7428	0.7900
	50/50	0.8078	0.1922	0.7696	0.2304	0.7626	0.2374	0.7732	0.7481
	20/80	0.8647	0.1353	0.7481	0.2519	0.7474	0.2526	0.7685	0.7127
KNN	80/20	0.9381	0.0619	0.7763	0.2237	0.7899	0.2101	0.7792	0.8107
	50/50	0.8883	0.1117	0.7786	0.2214	0.7699	0.2301	0.7605	0.7867
	20/80	0.8958	0.1042	0.7912	0.2088	0.7685	0.2315	0.7469	0.8070
Random Forest	80/20	0.9364	0.0636	0.7983	0.2017	0.8026	0.1974	0.7895	0.8381
	50/50	0.9393	0.0607	0.7971	0.2029	0.7937	0.2063	0.7790	0.8196
	20/80	0.9533	0.0467	0.7920	0.2080	0.7808	0.2192	0.7671	0.8041

Table 3. Training, validation, and test accuracy, error, recall, and precision for DT, KNN, and RF across all partitions of the Abalone dataset.

Classifier	Split	Predicted Negative (Actual Negative)	Predicted Positive (Actual Negative)	Predicted Negative (Actual Positive)	Predicted Positive (Actual Positive)
DT	80/19	974	330	253	951

	50/50	2436	697	791	2343
	20/80	3903	1086	1447	3590
KNN	80/19	960	289	238	1021
	50/50	2365	775	667	2460
	20/80	3692	1361	960	4013
RF	80/19	936	287	208	1077
	50/50	2407	728	565	2567
	20/80	3824	1221	977	4004

Table 4. Confusion matrix across all conditions for the Abalone dataset

Abalone Dataset

Random Forest showed the best results, having the highest test accuracy across all partitions: 80.26% (80/20), 78.96% (50/50), and 80.00% (20/80). Similarly, it has the highest validation accuracy among all. Decision Tree and KNN have lower validation and test accuracy, reflecting difficulties in dealing with the imbalance. Random Forest consistently achieved better precision and recall compared to the others, particularly in the 20/80 split, with precision of 76.0% and recall of 80.8%. Decision Tree and KNN demonstrated lower scores. However, KNN slightly outperformed Decision Tree in precision (76.5% vs. 74.3%), showing a better predictivity of the positive class. Random Forest had the lowest number of false negatives and false positives across all splits. In the 80/20 split, Random Forest misclassified only 208 false negatives, compared to Decision Tree of 253 and KNN of 238.

Classifier	**Train/Test Split**	**Train Accuracy**	**Training Error**	**Validation Accuracy**	**Validation Error**	**Test Accuracy**	**Test Error**	**Precision**	**Recall**
Decision Tree	80/20	0.9694	0.0306	0.8828	0.1172	0.8945	0.1055	0.9358	0.9195
	50/50	0.9712	0.0288	0.8882	0.1118	0.8771	0.1229	0.9032	0.9187
	20/80	0.9536	0.0464	0.8778	0.1222	0.8935	0.1065	0.9321	0.9140
KNN	80/20	0.8196	0.1804	0.7804	0.2196	0.7426	0.2574	0.7341	0.9634
	50/50	0.8477	0.1523	0.7445	0.2555	0.7525	0.2475	0.7965	0.8705
	20/80	0.8439	0.1561	0.7806	0.2194	0.6930	0.3070	0.7044	0.9258
Random Forest	80/20	0.9884	0.0116	0.9261	0.0739	0.8861	0.1139	0.9461	0.8822

	50/50	0.9729	0.0271	0.9086	0.0914	0.9007	0.0993	0.9331	0.9217
	20/80	0.9958	0.0042	0.8894	0.1106	0.8660	0.1340	0.8790	0.9398

Table 5. Training, validation, and test accuracy, error, recall, and precision for DT, KNN, and RF across all partitions of the Student dataset.

Classifier	**Split**	**Predicted Negative (Actual Negative)**	**Predicted Positive (Actual Negative)**	**Predicted Negative (Actual Positive)**	**Predicted Positive (Actual Positive)**
DT	80/19	52	11	14	160
	50/50	149	40	33	372
	20/80	252	45	56	595
KNN	80/19	26	55	6	150
	50/50	85	93	54	362
	20/80	77	244	47	580
RF	80/19	71	8	19	139
	50/50	158	27	32	377
	20/80	220	88	39	601

Table 6. Confusion matrix across all conditions for the Student dataset

Student Performance Dataset

Random Forest still has the best performance across all splits, achieving test accuracies of 88.61% (80/20), 90.07% (50/50), and 86.60% (20/80). KNN has lower performance especially in the 20/80 split with only 69.30% accuracy. Random Forest has high precision and recall across all splits. In the 80/20 split, it has precision of 94.61% and recall of 88.22%, reflecting its ability to correctly classify both passing and failing students. Decision Tree showed high precision of 93.58% and recall 91.95%.in 80/20 split. KNN exhibited the highest recall in the 80/20 split of 96.34%, but its precision dropped to 73.41%, indicating a higher number of false positives.

Algorithm/Dataset	**Wine**	**Abalone**	**Student**

Decision Tree	0.825/0.813/0.975	0.768/0.782/0.865	0.895/0.888/0.971
KNN	0.873/0.864/1.000	0.790/0.791/0.938	0.753/0.781/0.848
Random Forest	0.872/0.873/0.999	0.803/0.798/0.953	0.901/0.926/0.996

Table 7. Test/Validation/Training Accuracy Table

From the summarization in Table 7, it is shown that Random Forest is the most robust classifier, demonstrating better generalization, higher accuracy, and better handling of imbalanced data compared to other methods. KNN achieves perfect training accuracy on the Wine dataset, which suggests it may completely memorize the training data. With a test accuracy of 87.3% and a validation accuracy of 86.4%, this indicates potential overfitting.

4. Conclusion

The experiment conducted across the three datasets shows the performance of the three classifiers. The results consistently demonstrated that Random Forest outperformed the other classifiers in terms of accuracy, precision, recall, and error rates across all train-test partitions.

Random Forest proves to be the most robust model, achieving the lowest error rates and maintaining a balanced precision and recall across all datasets. It minimizes both false positives and false negatives. Decision Tree shows low training error but higher test error, particularly in smaller training data such as the 20/80 split, which leads to reduced generalization and an increase in false negatives. KNN struggles with imbalanced datasets and performs less effectively with smaller training data. Overall, the findings align with prior studies, which have shown that Random Forest consistently demonstrates superior generalization and better performance across datasets. In addition, better test accuracy is seen in larger training data, 80/20 splits.

There are several enhancements to gain deeper insights into classification performance. Future studies could include more classifiers, such as Gradient Boosting Machines and Support Vector Machines, as this study evaluated only three classifiers. Addressing imbalanced datasets through techniques like the Synthetic Minority Oversampling Technique could also improve performance. Furthermore, using larger datasets could help better evaluate the classifiers' ability to discover more complex relationships. Finally, future studies could include additional evaluation metrics. While this study focused on accuracy, precision, recall, error rates, and confusion matrices, metrics such as F1 score and ROC curves would provide more assessment of classifier performance.

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Wine Quality [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.

Nash, W., Sellers, T., Talbot, S., Cawthorn, A., & Ford, W. (1994). Abalone [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C55C7W.

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168).