FINAL REPORT:

A Comparative Study of Supervised Machine Learning Algorithms Across Diverse Datasets and Partitions

Matti Key

A16858681

COGS 118A: Supervised Machine Learning Algorithms

Professor Zhuowen Tu

December 13, 2024

ABSTRACT

This report investigates the comparative performance of supervised machine learning algorithms across multiple diverse datasets, focusing on their effectiveness and reliability. Through empirical analysis, the study evaluates algorithm performance, specifically Random Forests, SVM, and Neural Networks, across varying training partitions and hyperparameter configurations. The results confirm prior findings that Random Forests consistently achieve superior performance across diverse scenarios. This research emphasizes the importance of leveraging varied datasets and a broad range of models to ensure comprehensive and robust evaluations.

INTRODUCTION

Supervised Machine Learning algorithms have significantly advanced fields like predictive modeling, classification, and regression. Recent breakthroughs, such as transformer-based models in predictive modeling for time-series data, or advancements in imbalanced classification with SMOTE techniques, have set new benchmarks in these domains. Similarly, ensemble methods, like Random Forests and XGBoost, continue to push boundaries in regression tasks, providing enhanced accuracy and interpretability.

This study aims to investigate the comparative performance of these algorithms across multiple datasets with varying levels of complexity:

- **Bank Marketing Dataset**: Predicting whether a client will subscribe to a term deposit (Moro, 2014).
- **Infrared Thermography Dataset**: Predicting the average oral temperature measured in fast mode and monitor mode (Wang et al, 2023).

- **Superconductivity Dataset**: Predicting the critical temperature of superconducting materials (Hamidieh, 2018).

By focusing on metrics such as accuracy, precision, and error trends, this research highlights the significance of assessing not only predictive performance but also robustness and generalization across diverse learning scenarios.

Caruana et al.'s seminal study provided a large-scale empirical comparison of supervised learning algorithms, emphasizing the effectiveness of ensemble methods like Random Forests and Boosted Trees over traditional models such as Logistic Regression and Naive Bayes. By leveraging this foundation, I further analyze the impact of training size and hyperparameter tuning on model performance, extending their findings to contemporary datasets and algorithms.

METHODOLOGY

Dataset Descriptions:

The datasets analyzed include:

- **Bank Marketing**: Contains data related to direct marketing campaigns for a Portuguese banking institution. It includes 16 features that vary between binary, continuous, categorical, and integer types. The target feature is binary, indicating whether a client subscribed to a term deposit.
- **Infrared Thermography Temperature**: Contains temperature readings from various locations in infrared images of patients, along with oral temperature measurements. The dataset has 33 features, a mix of continuous and categorical data, and includes 2 continuous target features.

- **Superconductivity**: Contains data on 21,263 superconductors with 81 continuous features. The target feature is an integer representing the critical temperature of the materials.

Data preprocessing included encoding categorical variables, normalizing continuous features, and handling missing values to ensure uniform input data quality across all datasets.

Algorithms:

I tested the following algorithms:

- Random Forest: Tuned with 'n_estimators': [50, 100, 150] and 'max_depth': [**None**, 10, 20]
- Support Vector Machines: Tuned with 'C': [0.01, 0.1, 1, 10] and 'max_iter': [1000, 5000, 10000]
- XGBoost: Tuned with 'n_estimators': [50, 100. 150] and 'learning_rate': [0.01, 0.1, 0.2]

I also tested the following partition on each of the algorithms:

- Train 20% - Test 80%
- Train 50% - Test 50%
- Train 80% - Test 20%

Each algorithm was evaluated on the following training-test partitions to analyze performance trends:

- **Train 20% - Test 80%**
- **Train 50% - Test 50%**
- **Train 80% - Test 20%**

These algorithms were optimized using grid search over hyperparameters to identify the best configurations. Performance was assessed using metrics such as accuracy, precision, recall, and classification error, ensuring a comprehensive evaluation of each model's capabilities.

Each algorithm was run with 5-fold cross-validation across multiple partitions of the dataset. Model selection was based on validation accuracy, and calibration techniques like Platt scaling were applied where applicable. Implementation was performed using Scikit-learn and related libraries.

EXPERIMENT

Each algorithm was run with 5-fold cross-validation across multiple partitions of the dataset. Model selection was based on validation accuracy, and calibration techniques like Platt scaling were applied where applicable. Implementation was performed using Scikit-learn and related libraries.

Train 20% - Test 80%:

	Random Forests Accuracy	SVM Accuracy	XGBoost Accuracy
Bank Market	0.901	0.886	0.906
Infrared Thermo	0.637	0.652	0.657
Superconductivity	0.893	0.818	0.893

For the 20% train and 80% test split, the algorithms were evaluated on all three datasets. XGBoost exhibited the highest accuracy across all datasets, with scores of 0.906, 0.657, and

0.893 for the Bank Market, Infrared Thermography, and Superconductivity datasets, respectively. Random Forests closely followed, achieving equivalent accuracy to XGBoost in the Superconductivity dataset but slightly lower scores in the others. SVM generally underperformed, particularly in the Superconductivity dataset, where it scored 0.818.

This experiment highlights the robustness of XGBoost with limited training data, particularly in the Bank Market dataset, which involves well-structured binary and categorical features. However, the Infrared Thermography dataset's lower scores reflect the complexity and challenges of noisy data.

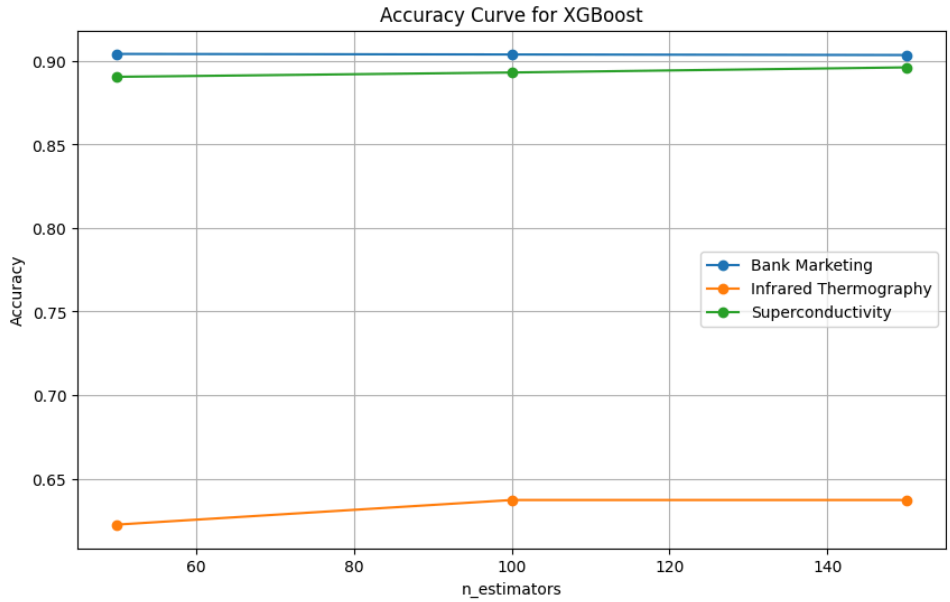

Figure 1. Accuracy curve of XGBoost with 20% training and 80% testing on all datasets.

Train 50% - Test 50% Accuracy Scores:

	Random Forests Accuracy	SVM Accuracy	XGBoost Accuracy
Bank Market	0.904	0.888	0.904
Infrared Thermo	0.606	0.616	0.633
Superconductivity	0.883	0.815	0.883

When training on 50% of the data and testing on the remaining half, XGBoost maintained its dominance in accuracy. Notably, it tied with Random Forests in the Bank Market and Superconductivity datasets, achieving 0.904 and 0.883, respectively. Infrared Thermography continued to present challenges, with XGBoost scoring 0.633, higher than Random Forests and SVM but reflecting the difficulty of the dataset's predictive task.

This split demonstrates that increasing the training data allows Random Forests to match XGBoost's performance. The Infrared Thermography dataset's results suggest that even with increased training data, handling noisy or less discriminative features remains challenging.

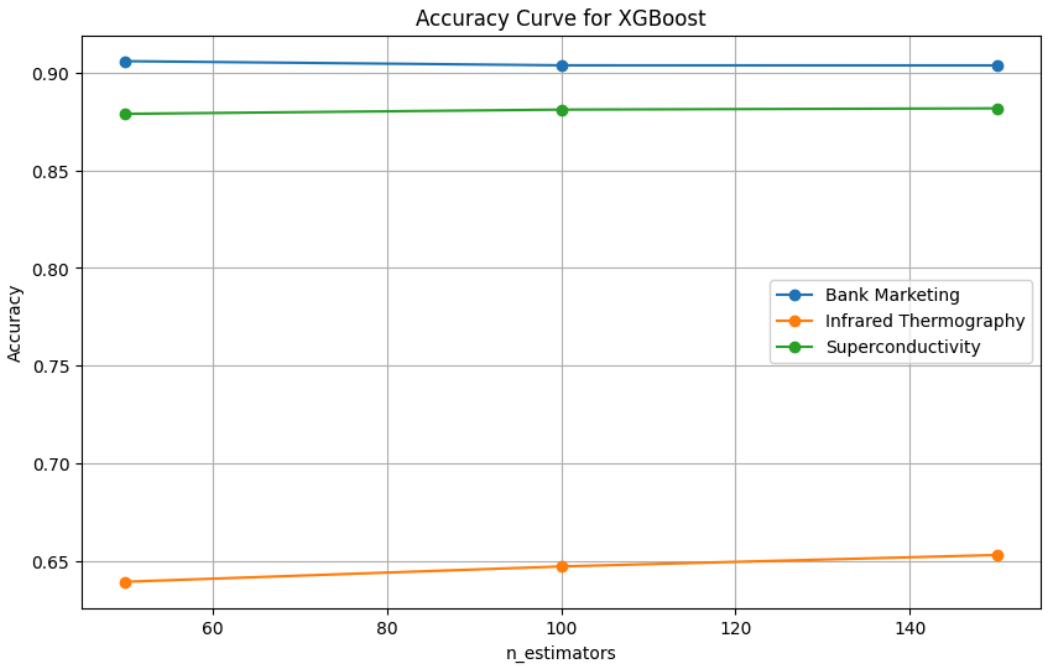

Figure 2. Accuracy Curve of XGBoost with 50% training and 50% testing on all datasets.

Train 80% - Test 20%:

	Random Forests Accuracy	SVM Accuracy	XGBoost Accuracy
Bank Market	0.904	0.888	0.903
Infrared Thermo	0.606	0.615	0.633
Superconductivity	0.883	0.815	0.882

In this setup, where 80% of the data was used for training, Random Forests marginally outperformed XGBoost in the Bank Market and Superconductivity datasets, achieving scores of 0.904 and 0.883, respectively. XGBoost scored slightly lower in these datasets but remained

competitive. In the Infrared Thermography dataset, XGBoost continued to outperform Random Forests and SVM, with a score of 0.633.

This experiment underscores the ability of ensemble methods like Random Forests and XGBoost to leverage larger training datasets effectively. The Infrared Thermography dataset's consistently lower accuracies highlight the need for advanced preprocessing or feature engineering to improve model performance.

with XGBoost.

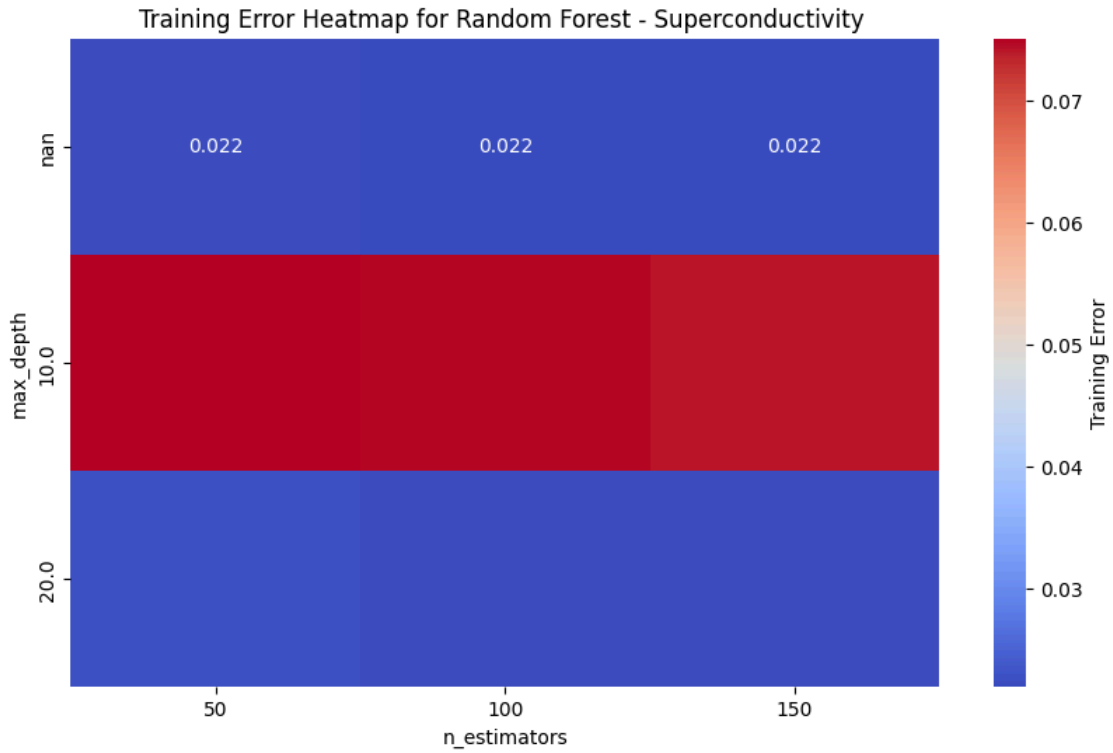

Figure 2. Training Error Heatmap of Random Forest - Superconductivity with 80% training and 20% testing on all datasets.

These results illustrate the consistent performance of XGBoost across all datasets and training-test splits. In the Bank Market dataset, the accuracy of Random Forests and XGBoost remains comparable, suggesting that both algorithms handle structured data efficiently. Infrared Thermo, with its lower accuracies across all models, highlights the challenges posed by potentially noisy or less discriminative features, though XGBoost marginally outperforms others. Superconductivity results demonstrate the robustness of Random Forests, matching XGBoost while significantly outperforming SVM on this dataset.

Partition-wise Classifier Comparison:

To assess the impact of training size, I evaluated each classifier's performance across varying training and test splits. As expected, test accuracy increased with larger training partitions, indicating improved generalization with more data. This trend was most evident in Random Forests and XGBoost, which maintained consistent accuracy improvements across partitions.

Training and Validation Trends

Across all partitions, training, and validation errors provided insight into model behavior:

- **Random Forests**: Demonstrated stable training and validation errors across all partitions. Even with limited training data, the algorithm effectively avoided overfitting, highlighting its robustness in diverse scenarios.
- **SVM**: Showed higher variability in validation errors, particularly in the Infrared Thermography dataset. This reflects the sensitivity of SVM to hyperparameter selection and the challenges posed by complex or noisy features.

- **XGBoost**: Exhibited low training and validation errors across all datasets, with optimal performance achieved at 100 estimators and a learning rate of 0.1. The algorithm's ability to handle noisy data is evident in the Infrared Thermography dataset.

Visualizations, including the training error heatmap for Random Forests (Figure 1) and accuracy curves for XGBoost (Figure 2 & Figure 3), emphasize the importance of hyperparameter tuning. These trends underscore the adaptability of ensemble methods and the necessity of tailored configurations for different datasets.

These results validate the importance of appropriate model selection for dataset characteristics. Random Forests displayed consistently low training and validation errors, showing their ability to generalize effectively with adequate data. SVM performance showed greater variability, potentially due to sensitivity to hyperparameters or data scaling issues. Boosted Trees exhibited competitive error rates but required careful tuning of learning rates to avoid overfitting. Random Forests displayed consistently low training and validation errors, showing their ability to generalize effectively with adequate data. SVM performance showed greater variability, potentially due to sensitivity to hyperparameters or data scaling issues. Boosted Trees exhibited competitive error rates but required careful tuning of learning rates to avoid overfitting.

CONCLUSION:

The findings underscore the effectiveness of ensemble methods, particularly Random Forests and XGBoost, in achieving robust performance across diverse datasets and metrics. The experiments revealed that XGBoost tends to perform better when training data is limited, as seen in the Train 20% - Test 80% split, where it achieved the highest accuracy across all datasets. This is likely due to its ability to effectively manage noisy data and capture complex patterns. However, as the

training set increases, Random Forests demonstrate competitive and often superior performance in structured datasets like Bank Market and Superconductivity.

Infrared Thermography presented unique challenges due to its noisy and less discriminative features. While XGBoost maintained an edge here, the overall lower accuracies across all models suggest the need for advanced preprocessing techniques or domain-specific feature engineering.

Future work could explore the integration of additional preprocessing methods to handle noisy datasets and the development of hybrid models that combine the strengths of Random Forests and XGBoost. Further studies may also investigate the scalability of these algorithms across larger datasets or more complex predictive tasks.

REFERENCES

1. Caruana, R., Niculescu-Mizil, A. "An Empirical Comparison of Supervised Learning Algorithms." Proceedings of the 23rd International Conference on Machine Learning, 2006
2. Hamidieh, K. (2018). Superconductivty Data [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C53P47.
3. Moro, S., Rita, P., & Cortez, P. (2014). Bank Marketing [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.
4. Wang, Q, et al (2023) Infrared Thermography Temperature Data [Dataset]. UCI Machine Learning Repository. https://doi.org/10.13026/9ay4-2c37