The 37th Canadian Conference on Artificial Intelligence
DOI:

Canadian AI 2024
27—31 May 2024, Guelph, Ontario

A Comparative Analysis of Machine Learning Algorithm Across Diverse Datasets

Xiaoting Huang(B00685239) [†,*], Xuelian Xi(B00977221) [‡], Siqi Wang(B00976996) [¥]

[†] Department of Electrical and Computer Engineering, Dalhousie University, Halifax, Nova Scotia; Xiaoting.Huang@dal.ca

[‡] Department of Computer Science, Dalhousie University, Halifax, Nova Scotia; xl884022@dal.ca

[¥] Department of Computer Science, Dalhousie University, Halifax, Nova Scotia; sq214680@dal.ca

GitHub link: https://github.com/XiaotingH/ML-model-analysis-across-diverse-datasets.git

Abstract

This research investigates the nuanced performance dynamics of machine learning algorithms across datasets of varying sizes. Employing three distinct datasets characterized by different sample sizes, we implement four machine learning algorithms—Logistic Regression, K Nearest Neighbors (KNN), Naïve Bayes, and Random Forest. This paper delineates the application methodologies, hyperparameter tuning procedures, and the resultant performance outcomes. Our findings reveal that Logistic Regression and Naïve Bayes exhibit relatively minor sensitivity to dataset size fluctuations. In stark contrast, KNN is markedly influenced by dataset size, showcasing significant performance variations. Notably, Random Forest demonstrates consistently superior performance across the four machine learning models considered, particularly excelling with the largest datasets. The implications of these observations are discussed, contributing to a comprehensive understanding of the interplay between machine learning algorithms and dataset characteristics. The effectiveness of a machine learning algorithm does not rely solely on expanding the dataset size. Our research indicates that, once a certain threshold is reached, merely increasing the dataset size does not proportionally improve performance.

Keywords: Supervised Learning, Data size, Random Forest, Logistic regression, KNN, Naïve Bayes

1. Introduction

The field of machine learning is at the forefront of modern computing advances, providing transformative solutions for a wide range of fields. The core of its capabilities lies in the ability to adapt different algorithms and apply them to different datasets. This study titled "A Comparative Analysis of Machine Learning Algorithm Across Diverse Datasets" embarks on a journey of discovery to understand the complexity and performance of various machine learning algorithms when applied to different datasets.

In the ever-expanding field of data, choosing the right machine learning algorithm becomes crucial. This choice is often influenced by many factors, such as the size of the dataset, its complexity, and the character of the features involved. Recognizing this, our research delves into a comprehensive comparison of four widely used machine learning algorithms: logistic regression, Naïve Bayes, random forests, and K nearest neighbors (KNN). Each of these algorithms has a unique approach to learning from data for three different datasets: adult income, airline satisfaction, and credit approval.

Each dataset presents a unique set of challenges and characteristics. For example, the Adult Income dataset reflects variables that influence income levels, while the Airline Satisfaction dataset provides insights into customer satisfaction parameters in the airline

[*] corresponding_author@example.ca

This article is © 2024 by author(s) as listed above. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (https://creativecommons.org/licenses/by/4.0/legalcode), except where otherwise indicated with respect to material included in the article. The article should be attributed to the author(s) identified above.

industry. On the other hand, the credit approval dataset contains factors that influence credit card approval decisions. This diversity of datasets allows for a robust analysis of how dataset characteristics affect the performance and applicability of each algorithm.

The purpose of this study is not only to compare the effectiveness of these algorithms on different datasets, but also to provide a framework to help select the most appropriate algorithm based on dataset characteristics. By examining the interactions between algorithms and datasets, this study aims to reveal the decisions that shape the future of machine learning applications.

Through this comparative analysis, we aspire to contribute to a broader understanding of machine learning algorithms, fostering more informed algorithmic choices and application approaches in different real-world scenarios.

2. Literature Review

In recent studies, researchers persist in exploring and comparing multiple machine learning models across various scenarios and datasets. Some of the study such as **Error! Reference source not found.** focused on evaluating the performance of various models on a small dataset. Khanam, J.J.et al. **Error! Reference source not found.** applied seven different ML algorithms on diabetes prediction, including Decision Tree(DT), KNN, Random Forest(RF), Naïve Bayes(NB), Adaptive Boosting, Logistic Regression, SVM to predict and evaluate their performance. The dataset used is Pima Indian diabetes datasets (PIDD) with 9 attributes and 768 observations. Most of the models are intuitively provide over 70% accuracy. Whereas the models built with Logistic regression and SVM deliver better performance, the accuracy achieve 78.78% and 77.71% respectively. However, Dris, A.B.et al. [2] focus more on investigating the impact of model performance on small datasets with different number of features and instances. Among the three employed machine learning algorithms, Decision Tree, SVM and Naïve Bayes, SVM delivered overall the best classification accuracy and Naïve Bayes has worse performance.

Concurrently, scholars are directed towards the investigation of large datasets, given their inherent susceptibility to overfitting. Addressing the primary concern associated with large datasets, the exploration of diverse machine learning models to identify those that effectively mitigate the impact of overfitting emerges as a pivotal topic within the research field. Catal, C.et al. [3] used five public NASA datasets from PROMISE repository to construct and evaluate software fault prediction model on the dimensions of dataset size and features selection. According to the study, Random Forest was concluded to deliver the higher prediction accuracy for large dataset, and Naïve Bayes provides best performance on the small dataset.

Similar relate works [4], Althnian, A.et al. aim to investigate the impact of dataset size on the performance of widely used supervised machine learning models, where two large datasets and three size reduced subset of dataset from the two large datasets are employed. According to the experiment result, AB and NB emerged as the most robust models for limited size of dataset, whereas Decision Tree exhibited the poorest performance. Additionally, the results imply the overall model performance is influenced more by the distribution of dataset rather than its size.

In this research work, three datasets from different areas of classification problem are selected, while two large datasets and one small dataset with close number of features are selected. Four commonly used supervised machine learning models, Logistic regression (LR), Random Forest (RF), K-Nearest Neighbor and Naïve Bayes are applied for performance evaluation.

3. Methods and Experiments

3.1. Datasets

In the context of this project, three datasets have been chosen from the UC Irvine repository and Kaggle website. Among them, two are sizable datasets—dataset 2 and dataset 3—each encompassing over a thousand instances with varying numbers of features. Specifically, dataset 3 is focused on predicting passengers' satisfaction based on their inflight habits and personal information, while dataset 2 is geared towards predicting income ranges, distinguishing between those above and below five thousand. On the other hand, dataset 1 is comparatively smaller, consisting of approximately 700 instances and a similar number of features. This dataset pertains to the prediction of credit approval decisions.

	Dataset Name	**Number of features**	**Number of Instances**
Dataset1	Credit Approval	16	690
Dataset2	Adult	14	32561
Dataset3	Airlines	25	103904

Table 1: Datasets information.

3.2. Data preprocessing

3.2.1. Normalization

In the chosen datasets, a significant number of features exhibit skewness and imbalance to some degree. Specifically, for strongly skewed features like "captain loss" and "captain gain," within the income prediction dataset, Min-Max scaling is the preferred standardization method due to its robustness to outliers. Conversely, for features with a more bell-shaped distribution, Z-score scaling is under consideration.

3.2.2. Feature encoding

In the preprocessing phase, two primary encoding techniques are employed, which are label encoding and ordinal encoding. Label encoding is selected in this task work due to its simplicity and memory usage reduction. However, when wording with features exhibiting a meaningful order, such as the in-flight class and airline passengers' satisfaction prediction dataset or the education level in an income prediction dataset, it is preferrable to apply ordinal encoding.

3.2.3. Feature selection

In the phase of data preprocessing, meticulous scrutiny was applied to the features, leading to the removal of those exhibiting high correlations. For instance, in Dataset 3, a heatmap was generated to visually assess correlations among features. Subsequently, a detailed comparison and data analysis were conducted to identify and eliminate redundant features based on the observed correlations.

3.3. Model Construction

3.3.1. Logistic Regression Algorithm

Logistic Regression is a predictive analysis algorithm and a type of classification algorithm. It is used when the response variable is categorical in nature. For binary classification problems, the algorithm predictions are binary or dichotomous, such as Yes/No, Positive/Negative, 1/0, and so on.

In our study, the initial model of Logistic Regression was applied to all three datasets. We further refined the model through hyperparameter tuning using GridSearchCV, which searches through a specified parameter grid to determine the combination that results in the best cross-validation score. The parameters tuned included 'C' for regularization strength, 'penalty' for the norm used in the penalization, 'solver' for the optimization algorithm, and 'tol' for the tolerance for stopping criteria. Because the most optimal hyperparameter contributes most to a LR model.

3.3.2. Naïve Bayes Algorithm

Naive Bayes is a versatile and efficient probabilistic supervised machine learning algorithm widely employed for classification tasks. The fundamental concept behind this algorithm involves calculating the probability of instances belonging to a specific class using Bayes' theorem. It learns the prior probability and likelihood of each feature occurring in each class, subsequently making predictions based on these probability calculations. One of the most significant advantages of Naive Bayes is its lack of hyperparameters, coupled with computational efficiency, making it a favorable choice for handling large datasets. However, it is important to note that the "naive" assumption, which assumes variables are conditionally independent and follow specific distributions, can be limiting. This limitation becomes apparent when the assumption is violated, particularly in real-world scenarios where ideal conditions may not always apply.

In the context of this study, three Naive Bayes models were constructed for different datasets, each assuming features follow Gaussian, Multinomial, and Bernoulli distributions, respectively. The datasets encompass a variety of predictor types, including continuous, discrete, and Boolean features, each corresponding to a well-suited distribution. Consequently, the key to constructing effective Naive Bayes models lies in determining the most suitable distribution type for each dataset scenario.

3.3.3. K-Nearest Neighbors Algorithm

K-Nearest Neighbors (KNN) stands out as a robust and intuitive supervised machine learning algorithm, providing an effective and straightforward approach for both classification and regression tasks. As the name suggests, KNN employs a neighborhood-centric learning strategy, making predictions based on the characteristics of nearby neighbors. The parameter "k", representing the number of neighbors taken into consideration, plays a crucial role in shaping the predictive outcome.

In this task, a KNN model was constructed, utilizing the Euclidean distance as the metric to locate the surrounding k neighbors. To determine the optimal number of neighbors for the model, an exhaustive search ranging from one to fifty with a step of one was conducted. Subsequently, stratified k-fold cross-validation was applied to the dataset, and the cross-validation scores were obtained. The parameter "k" corresponding to the highest cross-validation score was chosen as the optimal number of neighbors for this study.

3.3.4. Random Forest

Random Forest is a widely adopted ensemble learning approach employed for tasks involving both classification and regression in the domain of supervised learning. Classified within the family of decision tree-based methods, it is recognized for its notable accuracy, resilience, and effectiveness in handling substantial datasets characterized by a multitude of features. Ensemble learning means that it combines the predictions from multiple models to make a final prediction. In the case of Random Forest, these models are decision trees. Random Forest uses a technique called bagging, where multiple decision trees are trained on different subsets of the training data. Each subset is created by randomly sampling with replacement (bootstrap sampling) from the original dataset. For classification tasks, Random Forest typically uses a majority voting scheme, where the class predicted by most trees is selected as the final prediction. For regression tasks, it averages the predictions of individual trees.

For this classification task, Random Forest was selected as one of the comparative learning algorithms. The Sklearn library was utilized to construct the Random Forest classifier using the "RandomForestClassifier" command on the training data. Subsequently, we conducted a comparison between Random Search and Grid Search to identify optimal hyperparameters.

These hyperparameters included the number of trees in the forest (n_estimators), the minimum number of samples required to split an internal node (min_sample_split), the minimum number of samples required at a leaf node (min_sample_leaf), and the number of features to consider when searching for the best split (max_depth), also referred to as the maximum depth of trees. Finally, the determined optimal hyperparameters were applied to fit the test dataset, and the performance was evaluated.

4. Results

Based on the results presented in the following tables, LR, KNN, NB, and RF were evaluated using performance metrics such as accuracy, precision, recall, and F1-score. Notably, LR exhibited consistent performance across the three datasets, with its best performance observed in the credit approval prediction dataset, the smallest among the three with multiple multiclass features. This suggests that LR is less influenced by dataset size and is more affected by the types of features and their quantity. In contrast, Random Forest demonstrated the highest overall accuracy among the four models, particularly excelling in predicting airline passengers' satisfaction. This dataset, characterized by multiclass features and a mix of continuous values, showcased RF's superior performance. Naïve Bayes (NB) yielded varying accuracy levels across three distinct types, strongly influenced by feature distribution and types. Similar to RF, NB performed exceptionally well in predicting airline passengers' satisfaction. Conversely, KNN exhibited the most fluctuating accuracy, reaching over 90 percent in the airline dataset but only delivering 73 percent accuracy in the credit dataset. This indicates that KNN is the most sensitive model among the four, responding differently to variations in dataset characteristics.

Simultaneously, due to the presence of imbalances and skewness in the selected datasets, both macro-average and weighted-average evaluation metrics are computed. This approach offers insights into the imbalance within the datasets. Upon comparing the two averaging methods, it is observed that only the adult income prediction dataset exhibits a significant difference between macro and weighted average evaluation metrics. This discrepancy underscores that the adult income prediction dataset is the most imbalanced among the three.

Dataset1: Credit	**Accuracy**	**Precision (MacroAverage)**	**Recall (MacroAverage)**	**F1-Score (MacroAverage)**
Logistic Regression	0.85	0.86	0.86	0.85
KNN	0.73	0.74	0.70	0.70
Naive Bayes - GaussianNB	0.82	0.83	0.80	0.80
Naive Bayes - BernoulliNB	0.83	0.85	0.81	0.82
Naive Bayes - MultinomialNB	0.72	0.72	0.72	0.72
Random Forest	0.90	0.90	0.90	0.90

Table 2: Performance evaluation of LR, KNN, NB, RF, dataset Credit, macro average.

Dataset2: Adult	Accuracy	Precision (MacroAverage)	Recall (MacroAverage)	F1-Score (MacroAverage)
Logistic Regression	0.84	0.80	0.74	0.76
KNN	0.84	0.79	0.74	0.76
Naive Bayes - GaussianNB	0.72	0.70	0.77	0.69
Naive Bayes - BernoulliNB	0.79	0.72	0.76	0.74
Naive Bayes - MultinomialNB	0.79	0.71	0.59	0.60
Random Forest	0.87	0.84	0.77	0.80

Table 3:Performance evaluation of LR, KNN, NB, RF, dataset Adult, macro average

Dataset3: Airline	Accuracy	Precision (MacroAverage)	Recall (MacroAverage)	F1-Score (MacroAverage)
Logistic Regression	0.87	0.87	0.87	0.87
KNN	0.93	0.94	0.93	0.93
Naive Bayes - GaussianNB	0.86	0.86	0.86	0.86
Naive Bayes - BernoulliNB	0.78	0.77	0.78	0.77
Naive Bayes - MultinomialNB	0.82	0.82	0.81	0.81
Random Forest	0.96	0.96	0.96	0.96

Table 4:Performance evaluation of LR, KNN, NB, RF, dataset Airline, macro average

Dataset1: Credit	Accuracy	Precision (WeightedAverage)	Recall (WeightedAverage)	F1-Score (WeightedAverage)
Logistic Regression	0.85	0.86	0.85	0.85
KNN	0.73	0.74	0.73	0.72
Naive Bayes - GaussianNB	0.82	0.82	0.82	0.81
Naive Bayes - BernoulliNB	0.83	0.84	0.83	0.83
Naive Bayes - MultinomialNB	0.72	0.72	0.72	0.72
Random Forest	0.90	0.90	0.90	0.90

Table 5: Performance evaluation of LR, KNN, NB, RF, dataset Credit, weighted average.

Dataset2: Adult	Accuracy	Precision (WeightedAverage)	Recall (WeightedAverage)	F1-Score (WeightedAverage)
Logistic Regression	0.84	0.84	0.84	0.84
KNN	0.84	0.83	0.84	0.83
Naive Bayes - GaussianNB	0.72	0.83	0.72	0.74
Naive Bayes - BernoulliNB	0.79	0.82	0.79	0.80
Naive Bayes - MultinomialNB	0.79	0.76	0.79	0.75
Random Forest	0.86	0.86	0.86	0.85

Table 6: Performance evaluation of LR, KNN, NB, RF, dataset Adult, weighted average.

Dataset3: Airline	Accuracy	Precision (WeightedAverage)	Recall (WeightedAverage)	F1-Score (WeightedAverage)
Logistic Regression	0.87	0.87	0.87	0.87
KNN	0.93	0.93	0.93	0.93
Naive Bayes - GaussianNB	0.86	0.86	0.86	0.86
Naive Bayes - BernoulliNB	0.78	0.78	0.78	0.78
Naive Bayes - MultinomialNB	0.82	0.82	0.82	0.82
Random Forest	0.96	0.96	0.96	0.96

Table 7: Performance evaluation of LR, KNN, NB, RF, dataset Airline, weighted average.

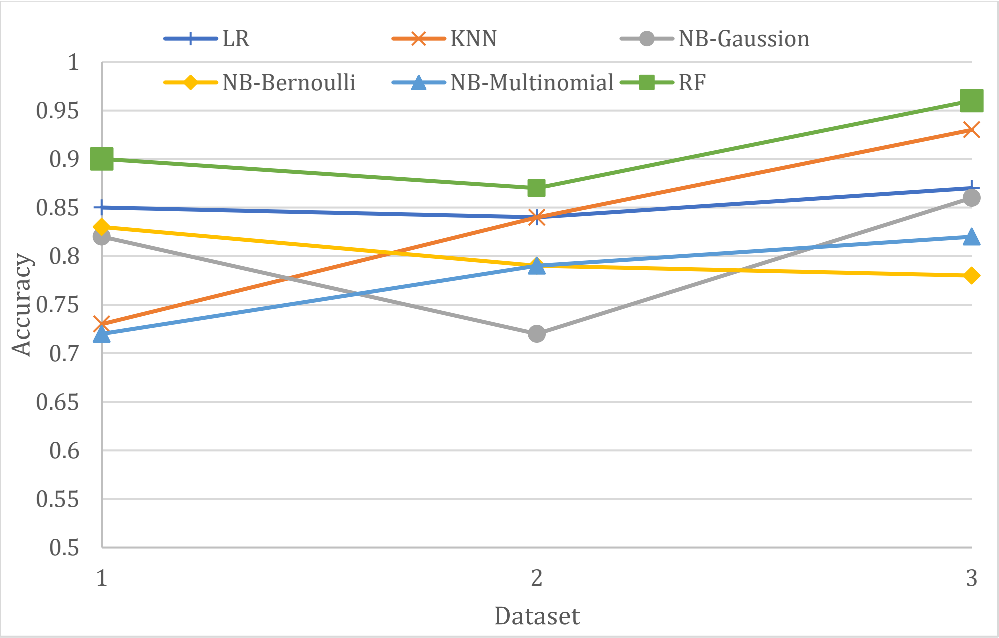

Figure 1: ML models performance comparison over all three datasets.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Macro - Average\ Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \tag{2}$$

$$Weighted - Average\ Precision = \sum_{i=1}^{N} w_i \cdot \frac{TP_i}{TP_i + FP_i} \tag{3}$$

$$Macro - AverageRecall = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i} \tag{4}$$

$$Weighted - AverageRecall = \sum_{i=1}^{N} w_i \cdot \frac{TP_i}{TP_i + FN_i} \quad (5)$$

$$\text{Macro-Average F1 Score} = 2 \cdot \frac{\text{Macro-Average Precision} \times \text{Macro-Average Recall}}{\text{Macro-Average Precision} + \text{Macro-Average Recall}} \quad (6)$$

$$\text{Weighted-Average F1 Score} = \sum_{i=1}^{N} w_i \cdot 2 \cdot \frac{\frac{TP_i}{TP_i + FP_i} \times \frac{TP_i}{TP_i + FN_i}}{\frac{TP_i}{TP_i + FP_i} + \frac{TP_i}{TP_i + FN_i}} \quad (7)$$

5. Discussion

According to the results above, we discussed two main aspects of the relationship between the algorithms of machine learning (LR, KNN, NB and RF) and the sizes of datasets. One is the performance stability analysis of each specific learning algorithm on varied sizes of datasets. And the other is the comparison analysis of different algorithms' performance on large or small datasets.

From the experiment results, we find that LR is the least sensitive algorithm to datasets change, while the performance of KNN fluctuates a lot on varied sizes of datasets. Accuracy values of LR for the three different experiment datasets are all above 0.84, indicating a satisfactory result. According to [5], for observational studies that involve logistic regression in the analysis, [5] recommends a minimum sample size of 500 to derive statistics that can represent the parameters in the targeted dataset. The smallest dataset we have is 690 samples. Therefore, the correct minimum size of dataset lays the foundation for LR effective prediction. Meanwhile, the performance of LR is stable, only differing within 3% on accuracy. This finding is the same with [6] - Researchers tried to find the effect of dataset size and interactions on prediction performance of logistic regression and deep learning models and their result is whatever the interaction order, increasing the dataset size did not significantly affect model performance, especially that of machine learning models – logistic regression.

On the flip side, the performance of KNN exhibits discernible fluctuations with variations in dataset size. The outcomes suggest a relatively positive correlation, indicating that as the dataset size increases, KNN tends to deliver improved performance. Just as J. Scott Olsson's research [7], the accuracy of KNN is contingent on the parameter 'k', especially with little available data, and the optimal 'k' tends to rise with the size of the training data.

Similar observations hold true for Naïve Bayes, demonstrating robutst and consistent performance across the three selected datasets. In our experiments, we developed three Naïve Bayes model: Gaussian NB, Bernoulli NB, and Multinomial NB. The performance of these models is intricately linked to the type of features, effectively capturing the prevalent feature distribution across the entire dataset. The model with the highest accuracy is chosen to represent the ultimate evaluation performance of Naïve Bayes. Examining the performance details presented in the result table, it becomes apparent that the size of the dataset may not be a critical factor influencing Naïve Bayes performance. Instead, the performance is more reliant on the attributes of the features. For instance, in dataset 1, the performance of the constructed Multinomial NB model is slightly inferior to the other two models. This implies that the variable distribution in dataset 1 aligns more closely with a Gaussian or Bernoulli distribution. In such cases, if the NB model is constructed based on the assumption that features follow a multinomial distribution, the performance will be notably degraded.

However, from the perspective of Random Forest, its performance is also satisfactory, consistently achieving accuracy values above 0.87. Notably, Random Forest demonstrates its best performance on the largest dataset, dataset 3, with accuracy surging to 0.96. It further proves the capability of Random Forest to effectively handle large and complex datasets, leveraging its ability to learn intricate details as the dataset size increases. Moreover, Random Forest exhibits commendable robustness in dealing with outliers and displays a high level of generalization across different types of features. Nevertheless, one notable drawback of Random Forest is its high complexity. The procedure of hyperparameters tunning to construct an optimal model that best fits the dataset can be time-consuming, rendering the computational cost relatively hight.

In general, all learning algorithms, except KNN, exhibit satisfactory performance in the smallest dataset (Dataset 1). In the case of the medium-sized dataset (Dataset 2), there is no notable difference in the performance of any algorithm. However, when dealing with the largest dataset (Dataset 3), stability is upheld by Logistic Regression (LR), while the performance of all other algorithms displays improvement. Notably, significant enhancements are observed, particularly for KNN and Random Forest (RF).

6. Conclusion:

Our comprehensive study offers pivotal insights into the performance dynamics of various machine learning algorithms across datasets of differing sizes. We meticulously evaluated four widely-adopted algorithms—Logistic Regression (LR), K Nearest Neighbors (KNN), Naïve Bayes (NB), and Random Forest (RF)—across three distinct datasets.

Both Logistic Regression and Naïve Bayes algorithms demonstrated a remarkable stability in performance, irrespective of dataset size fluctuations. This robustness can be attributed to their inherent algorithmic structures, making them reliable choices for varied datasets, particularly when computational efficiency and generalizability are pivotal.

The KNN algorithm exhibited significant variability in performance contingent on dataset size. Our findings underscore that KNN's efficacy is largely dependent on the dataset's volume, with larger datasets generally enhancing its predictive accuracy. This suggests that KNN may be more suited for scenarios where sufficient data is available, and its parameter 'k' can be optimally tuned.

Random Forest emerged as a notably powerful algorithm, especially with larger datasets. Its ensemble approach, combining multiple decision trees, allows it to excel in complex scenarios involving substantial datasets. This characteristic makes RF an ideal candidate for applications where large-scale data is prevalent and where model robustness is critical.

The efficiency of a machine learning algorithm is not solely dependent on increasing dataset size. Our study suggests that, beyond a certain threshold, simple enlarging the dataset does not proportionally enhance performance. This finding emphasizes the importance of considering other factors such as feature selection, algorithm tuning, and computational efficiency in the data-algorithm interplay. According to research paper from Mo, H. et al. [8], depending on the size of the data set, adjusting the hyperparameters in the algorithm can significantly affect the performance of the algorithm.

Our research contributes significantly to the understanding of machine learning algorithms' performance across various datasets. It underscores the importance of choosing the right algorithm based on the specific characteristics and size of the dataset, thereby guiding future applications and developments in the field of machine learning.

7. Contribution

Group member	Contribution
Xiaoting Huang	Coding: - KNN and NB model construction - Dataset preprocessing Report: - Literature review - Methods and experiments - Results - Discussion
Xuelian Xi	Coding: - RF model construction - Dataset preprocessing Report: - Abstract - Methods and experiments - Results - Discussion
Siqi Wang	Coding: - LR model construction - Dataset preprocessing Report: - Introduction - Methods and experiments - Results - Conclusion

References

[1] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *Ict Express*, *7*(4), 432-439.

[2] Dris, A. B., Alzakari, N., & Kurdi, H. (2019, June). A Systematic Approach to Identify an Appropriate Classifier for Limited-Sized Data Sets. In *2019 International Symposium on Networks, Computers and Communications (ISNCC)* (pp. 1-6). IEEE.

[3] Catal, C., & Diri, B. (2009). Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem. *Information Sciences*, *179*(8), 1040-1 058.

[4] Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A. B., Alzakari, N., ... & Kurdi, H. (2021). Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Applied Sciences*, *11*(2), 796.

[5] Bujang MA, Sa'at N, Sidik TMITAB, Joo LC. (2018, July) *Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. Malays J Med Sci. 2018 Jul;25(4):122-130. doi: 10.21315/mjms2018.25.4.12. Epub 2018 Aug 30. PMID: 30914854; PMCID: PMC6422534.*

[6] Bailly A, Blanc C, Francis É, Guillotin T, Jamal F, Wakim B, Roy P. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. Comput Methods Programs Biomed. 2022 Jan;213:106504. doi: 10.1016/j.cmpb.2021.106504. Epub 2021 Oct 28. PMID: 34798408.

[7] J. Scott Olsson. 2006. An analysis of the coupling between training set and neighborhood sizes for the kNN classifier. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06). Association for Computing Machinery, New York, NY, USA, 685–686. https://doi.org/10.1145/1148170.1148317

[8] Mo, H., Zhang, Y., & Zhang, Y. (2017). A Performance Evaluation of Classification Algorithms for Big Data. Procedia Computer Science, 122, 1100-1107. https://doi.org/10.1016/j.procs.2017.11.479