Machine Learning Analysis on MNIST and California Housing Datasets

Hussein Shaa'lan [1], Omar Bakir Habib [1], Alaa Hussein [1], Saif Qahtan [1]
[1]Informatics Institute for Postgraduate Studies, University of Information Technology and Communications, Baghdad, Iraq
*Corresponding Author: omar.baker.student@uoitc.edu.iq

Abstract

This researcj explores the performance of machine learning models on two basic types of problems: classification and regression. The MNIST data was used to test a Logistic Regression model on handwritten digit classification with a 91.76 accuracy which indicates good baseline performance by a linear model. In the regression task, both Linear Regression and Random Forest Regressor were used to analyze the California Housing data.The results of the experiments show that the Random Forest model is much more effective than Linear Regression, and its R 2 score is 0.8088 in contrast to 0.6195, and its error rate is smaller. This has been enhanced by the capacity of the ensemble techniques in describing non-linear and complicated relationships in real life data. The results emphasize the need to choose the right models depending on the nature of the data. Although simple linear models may be effective in structured data sets, complex regression tasks need more sophisticated models. This work offers a comparative approach via which the complexity of the models affects the predictive performance of the various machine learning problems.

Keywords

Machine Learning, Classification, Regression, MNIST, California Housing, Logistic Regression, Random Forest, Model Evaluation, Predictive Modeling

1. Introduction

Machine Learning (ML) has become an inseparable part of modern data-driven systems because it enables computers to find the patterns in data and make smart decisions without being explicitly programmed [1]. It finds numerous applications in other areas like image recognition, natural language processing, healthcare, and financial analysis [2]. One of the most basic machine learning tasks is classification and regression, which are applied to solve numerous real-world problems [3]. Classification involves categorizing input data into fixed groups, and regression deals with the issue of estimating numerical values [3].

The model selection is a vital element of machine learning, as the selection strongly depends on the characteristics and the complexity of data. In structured and linearly separable data, simpler models can be used but more complicated models are needed to learn non-linear correlations in

real-world data [4]. The study incorporates the utilization of two benchmark datasets in order to assess the performance of machine learning models.

The MNIST is a handwritten digit classification dataset that is used as a benchmark to measure classification algorithms [5]. The California Housing data, on the other hand, is a regression problem, with various interacting variables affecting the prices of houses [6]. This paper compares the performance of the various machine learning models on these datasets. Logistic Regression is utilised as the baseline classification model of MNIST and Linear Regression and Random Forest Regressor to model the California Housing dataset.

The key issue of this research is establishing the performance of various machine learning models when used on data with different degrees of complexity. In particular, the research questions are whether simple linear models are adequate in terms of classification and regression, or more complex models are needed to learn non-linear, complex relationships in real world data. Whereas there are many studies that have implemented machine learning models to either classification or regression tasks separately, few studies have made a direct comparison of model performance across the types of problems using a single experimental system [7]. This paper will help close this gap by comparing classification and regression models to each other under uniform conditions, giving a better insight into the impact of model complexity on model performance.

2. Work Datasets Description

2.1 MNIST Dataset

MNIST dataset is a well-known benchmark dataset in machine learning literature to recognize hand written digits [5].The modified national standards of measurement and technology (MNIST) database is a popular benchmark among the machine learning community to test handwritten digit recognition systems [5]. It was originally a collection of tests that LeCun et al. used as a benchmark to test classification algorithms because of its standard format and clearly defined evaluation procedure. It has grey scale images of hand written figures (0-9).It contains grayscale images of handwritten decimal numbers (0-9), with all of them centered and resized to the same size (28x28 pixels). All pictures are coded in 784 pixel values.The images are flattened into a 784-pixel intensity array (28x 28) with each value between 0 (black) and 255 (white) being the grayscale intensity of the respective pixel. The dataset is especially well-suited to the analysis of linear classifiers since the digit classes are quite distinct in pixel space.

2.2 California Housing Dataset

The California Housing dataset can be employed in a regression task in the context of machine learning and data analysis [6].California Housing data is a standard to use in regression in machine learning studies [6]. It was initially based on the U.S. Census of 1990 and enumerated by Pace and Barry and has since been extensively used to assess regression algorithms because it is complex in real-life applications and has rich feature interactions. It contains articles related to housing in Californian districts.The dataset is composed of aggregated demographic and geographic data at the block-group level in which each record characterizes a cluster of households in a given geographic location in California. The median house value, the variable of interest, displays non-linear relationships with several predictors, and thus it

is especially difficult to use purely linear models and an excellent testbed to ensemble techniques.

3. Methodology

To analyze the performance of machine learning models in a classification as well as a regression task, a systematic approach was chosen.In order to critically assess the behavior of machine learning models in a classification and regression task, we developed a structured and reproducible experimental pipeline. This method uses model selection, performance analysis, preprocessing and training of data.The four key steps of the pipeline are: (1) data preprocessing and normalization, (2) model selection and configuration, (3) model training on specified training splits and (4) performance evaluation through task-specific metrics. This single workflow guarantees similar and comparable conditions of the two datasets.as depicted in figure (1), as represented in Figure 1.

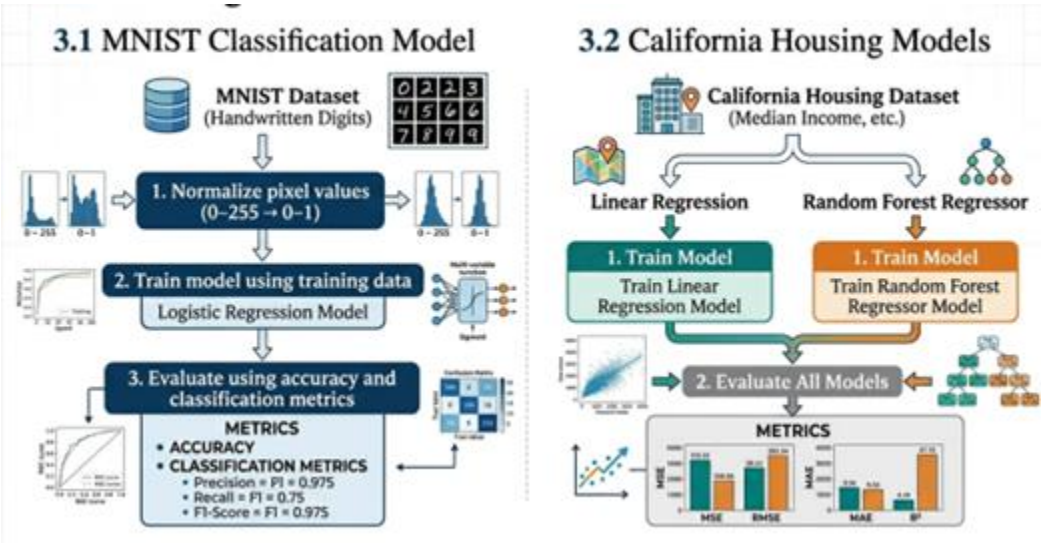

Figure 1: research method process

3.1 MNIST Classification Model

Multi-class classification was performed using a Logistic Regression model. model was used as a baseline classifier on the multi-class classification task (digits 0-9). Logistic Regression generalizes binary classification to multicast classification using One-vs-Rest (OvR) approach, where a binary classifier is trained per class of digit. Although it is linear, it has been found to be a good baseline to determine the separability of the MNIST feature space and can be used to give interpretable class probabilities through the softmax function.

Steps:

1. Normalize pixel values (0–255 → 0–1)
2. Train model using training data

Evaluate using accuracy and classification metrics

3.2 California Housing Models

Two regression models have been used: To assess the impact of model complexity on the quality of regression, two regressions of different levels of expressiveness were used and compared:

1. Linear Regression: A parametric model which assumes a linear relationship between input features and the target variable. It approximates feature coefficients by Ordinary Least Squares (OLS) minimization, and it is the default regression model in this paper.
2. Random Forest Regressor :A type of ensemble learning that builds a number of decision trees using bootstrapped subsets of the data and averages their predictions. Random Forest can reduce the overfitting problem and can model non-linear interactions among features that are not available in linear models by aggregating a large number of weak learners.

Measurements of evaluation: The two regression models were assessed based on four complementary measures of different facets of prediction quality:

- Mean Squared Error (MSE): penalizes large errors more than average squared errors; averages the difference between predicted and actual values.
- Root Mean Squared Error (RMSE) : square root of MSE, which is in the same units as the target variable, making it easier to interpret.

- Mean Absolute Error (MAE) : averages the plain difference between predictions and actuals, and is resistant to outliers in comparison to MSE.
- R^2 Score: the coefficient of determination, which is the percentage of the variance in the target that is explained by the model; a value of 0 to 1 where higher values imply a better fit.

4. Experimental Results

4.1 MNIST Classification Results

Metric	Value
Accuracy	91.76%
Precision	0.92
Recall	0.92
F1-score	0.92

Observations**:**

- High accuracy indicates strong classification performance
- Some confusion exists between similar digits (e.g., 5, 8, 9)

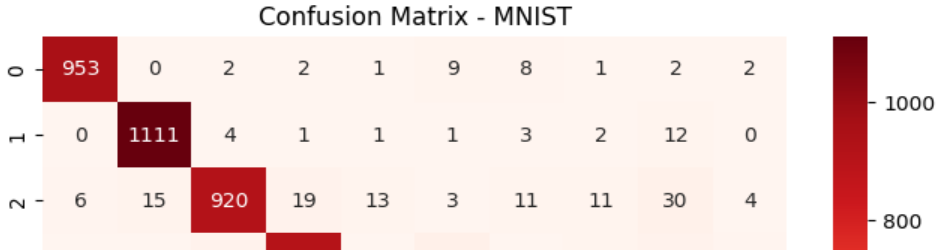

Figure 2: Confusion Matrix for MNIST

The performance of the Logistic Regression model in MNIST is shown in the confusion matrix. Most of the values are on the diagonal implying that there are correct classifications.The prevalent values lie down the main diagonal, which means that the model is able to identify the enormous majority of the digits correctly. This general trend validates that the learned decision boundaries are generally applicable to all ten classes, and no one class has a serious systematic error. There are minor errors between visually similar digits such as 5, 8, and 9.The strongest off-diagonal concentrations are between visually similar pairs of digits - especially 4/9, 3/5, and 5/8 - which have non-discriminable overlap in their stroke patterns, and cannot be reliably disambiguated by a linear classifier without spatial features present.

4.2 California Housing Results

Model	R^2 Score	RMSE	MAE
Linear Regression	0.6195	69,765	50,352
Random Forest	0.8088	49,457	32,190

Figure 3: Actual vs Predicted Values (Random Forest)

The scatter plot identifies the correlation between the predicted and actual houses values with the help of the Random Forest model. The data points coincide well on the diagonal line and this implies that it is correct.The points are bunched about the ideal 45° diagonal (predicted = actual), which means there is a high predictive power over the entire range of housing values. This is a testimony that the model can capture non-linear and relatively complex relationships in the data.A little dispersion can be observed at the upper end of the value range (high-value properties), which indicates that extreme values are more difficult to forecast - a typical drawback when the training data is sparse at its high-price end. On balance, the plot confirms the fact that Random Forest is effective to capture the non-linear, multi-layered format of the California Housing data.

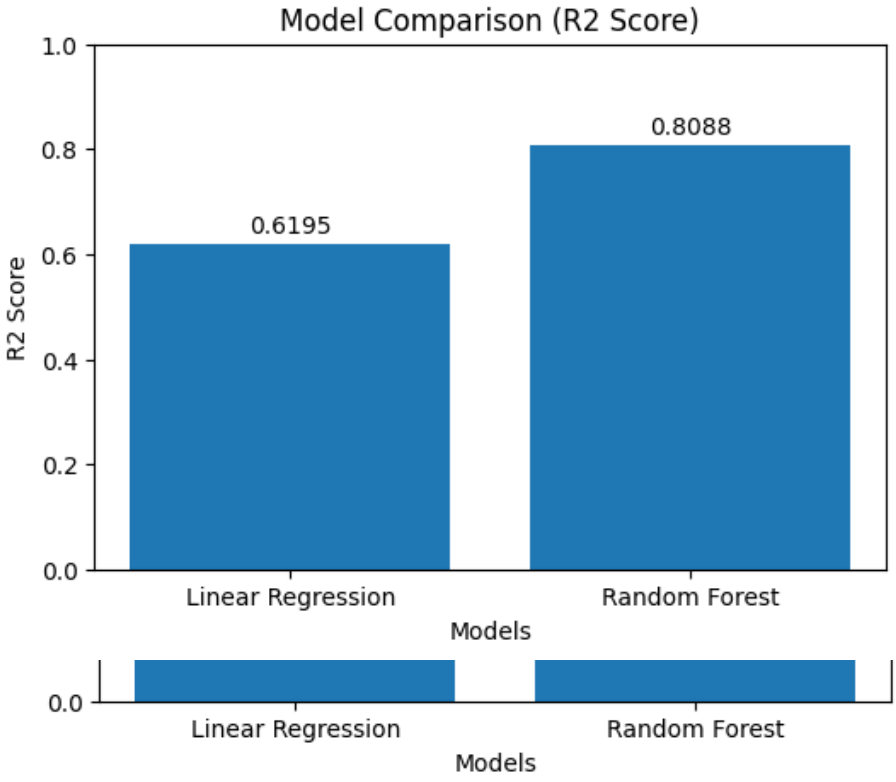

Figure 4: Model Comparison Based on R^2 Score

The bar chart shows the performance of Linear Regression and Random Forest models in terms of R 2 score. The Random Forest model has a higher R 2 (0.8088) than Linear Regression (0.6195), which means it has a higher predictive power and it can also explain non-linear relationships that are not linear in the California Housing data.

5. Analysis of Results

5.1 MNIST Analysis

This means that the Logistic Regression model was effective as it was able to accurately classify 91.76 per cent of the time despite its simplicity. This finding demonstrates that the MNIST data is well structured and can be partitioned relatively easily by a straight line thus indicating that linear models can be effective. The model is also more stable and predictable across varying classes since the accuracy, recall, and F1-score are all quite high (approximately 0.92). It is a significant finding with a linear model that uses raw pixel features, and no convolutional or spatial pre-processing. The uniform accuracy and recall of all ten digit classes (=0.92) further supports the idea that the model is not excessively biased towards a specific digit, which strengthens the stance of the classifier as a baseline. However, it had a few minor errors, particularly in numbers that were resembling, such as 5, 8, and 9. These mistakes are due to the fact that linear decision boundaries can't always pick up on small changes in image patterns. This demonstrates that the Logistic Regression is a good place to start, but more sophisticated models such as Convolutional Neural Networks (CNNs) might even improve this by further representing the spatial characteristics. Convolving filters teach CNNs hierarchical spatial representations, edges, curves, and higher-level shapes, so they can differentiate such ambiguous pairs of digits much more accurately. The current state of the art CNNs regularly reach 99+ percent accuracy on MNIST, demonstrating how much of a gap still exists above linear models.

5.2 Regression Analysis

The results show that the Random Forest model significantly outperformed Linear Regression, as illustrated in Table X.

Model	R^2 Score
Linear Regression	0.6195
Random Forest	0.8088

This improvement has been facilitated by the California Housing dataset that has complicated and non-linear relationships between such features as income, location, and population. Linear Regression requires that there should be a linearity between the input variables and the target. This renders it difficult to model more complex relationships. Random Forest on the other hand an ensemble learning approach is more adept at discovering feature interaction and patterns that do not necessarily have a straight line. As an example, the correlation between median income and house value is very non-linear: the prices increase rapidly at lower income levels and then level off, which is naturally represented by the decision-tree splits of a Random Forest, but which Linear Regression must represent with only a single slope.

Such results indicate that complexity of a model matters greatly in making correct predictions, particularly when dealing with real world data whose relationships among variables are not necessarily linear.

6. Discussion

The results of the present research highlight the importance of selecting appropriate machine learning models based on the characteristics of the data. Logistic Regression performed quite well on the MNIST data, demonstrating that even simple models can be highly accurate when the data is properly structured and there are clear patterns to it. A less complex model in these contexts would also have practical benefits: it can be trained more quickly, has smaller memory requirements, and its coefficients can be more easily interpreted to easily show which parts of the pixels most contribute to the classification of each digit.

The California Housing dataset, however, is more complex since some variables are not related in a linear manner. When this happens, more advanced models like Random Forest are needed to accurately predict these relationships and make accurate predictions. The fact that Rand Forest performs better indicates that it is robust and can be generalized more on complex data. The bagging used by the ensemble to train each tree on a bootstrapped subsample of the data also helps to reduce variance and ensures the final model is more robust to noise and outliers that are typical of real-world tabular data.

Furthermore, the fact that classification and regression tasks are compared shows that there cannot be a single model that will necessarily be the best. Rather, the nature of the problem, complexity of the data and the balance between prediction accuracy and computational efficiency should determine the model selection..Moreover, the direct analogy between classification and regression problems supports one of the key principles of practical machine learning that there is no universal dominating model that performs better on all types of problems. The nature of the task, structural properties of the data and practical trade-off between predictive accuracy and computational cost must guide the selection of the effective model. Resource-constrained deployments, such as those in resource-constrained settings, might not be worth the expensive training time and infrastructure overhead of a complex ensemble.

Altogether, this work points to the idea that although linear models are valuable in the context of the baseline analysis, the ensemble approach has numerous benefits in the framework of real-life data analysis and, thus, is more applicable in practice.Summarily, linear models give an invaluable

baseline, which is fast, interpretable and competitive in well structured problems. But in the case of irregularities and interactions between features typical of real-world data, ensemble techniques like Random Forest can provide significantly better predictive accuracy and that ought to be used where accuracy is the main goal and the computational resources allow.

Key Insights

1. Model Complexity Matters
   Simple models work well for structured data (MNIST)
   Complex models are needed for real-world regression tasks
2. Feature Relationships
   Housing prices depend on multiple interacting variables
   Non-linear models handle this better
3. Performance Trade-offs
   Logistic Regression: fast and efficient
   Random Forest: higher accuracy but more computational cost

7. Limitations

- MNIST model is not optimized (no hyper parameter tuning) regularization strength (C) was set to default values, solver and maximum iterations were default values, systematic tuning through cross-validation may provide quantifiable accuracy gains.
- No feature engineering applied to housing dataset :derived features (like rooms-per-household ratio or population-density ratio) would not have benefited both models, especially Linear Regression, by generating more linearly separable features.
- Missing values (in case they existed) were not dealt with in-depth: a more comprehensive imputation algorithm (say: median imputation or KNN-based filling) and analysis of outliers could enhance the quality of data and model stability.
- Deep learning models were not explored :CNNs to classify images and gradient-boosted trees or neural networks to regress will be natural extensions that would likely bridge most of the remaining performance gap.

8. Conclusion

This research shows that machine learning models can greatly vary in their performance based on the type and complexity of the data. The Logistic Regression model performed very well on the MNIST dataset, which means that linear models can be used when dealing with structured classification issues that have clear patterns.

Conversely, the Random Forest model was much more effective than the Linear Regression in the California Housing data, which underscores the need to employ non-linear and ensemble algorithms in addressing complex data in the real world. The fact that the R squared score and error values were higher with Random Forest attests to its better prediction and model ability to capture complex features relationships.

On the whole, the findings highlight that no one model can be optimal in every task. Rather, the choice of the model must be determined by the data properties and the particular issue under consideration. It demonstrates that data complexity is an important factor to consider

when selecting algorithms to use in machine learning since it directly affects prediction accuracy and reliability.The findings demonstrate that non-linear ensemble algorithms are more effective than linear models in real regression problems, and simpler models may still be effective in structured classification problems.

9. Future Work

Future improvements may include:

- Using Convolutional Neural Networks (CNNs) for MNIST which exploit local receptive fields and weight sharing to encode spatial hierarchies, which are likely to drive the accuracy over 99 percent higher.
- Applying feature engineering to housing data
- Hyperparameter tuning (Grid Search / Random Search)
- Testing advanced models such as XGBoost

References

[1] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 1-21.

[2] Jhaveri, R. H., Revathi, A., Ramana, K., Raut, R., & Dhanaraj, R. K. (2022). A review on machine learning strategies for real-world engineering applications. *Mobile Information Systems*, *2022*(1), 1833507.

[3] Chaudhary, P. S., Khurana, M. R., & Ayalasomayajula, M. (2024). Real-world applications of data analytics, big data, and machine learning. In *Data Analytics and Machine Learning: Navigating the Big Data Landscape* (pp. 237-263). Singapore: Springer Nature Singapore. [4] D. Harrison and D. L. Rubinfeld,

[4] Kapoor, A. (2024). Ml approach: Algorithms, real-world applications and research directions. *Real-World Applications and Research Directions (November 01, 2024)*.

[5] Lwakatare, L. E., Raj, A., Crnkovic, I., Bosch, J., & Olsson, H. H. (2020). Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions. *Information and software technology*, *127*, 106368.

[6] Nozari, H., Ghahremani-Nahr, J., & Szmelter-Jarosz, A. (2024). AI and machine learning for real-world problems. In *Advances in computers* (Vol. 134, pp. 1-12). Elsevier.

[7] Khan, M. A. (2023). Real World Applications And Research Directions For Machine Learning: Challenges And Defies. *Cloud Computing and Data Science*, 2949-2954.

[8] Sarkar, D., Bali, R., & Sharma, T. (2017). Machine learning basics. In *Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems* (pp. 3-65). Berkeley, CA: Apress.

[9] Huang, Y., Li, J., Li, M., & Aparasu, R. R. (2023). Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC medical research methodology*, *23*(1), 268.

[10] Paleyes, A., Urma, R. G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: a survey of case studies. *ACM computing surveys*, *55*(6), 1-29.