COMPARISON OF REGRESSION ANALYSIS ALGORITHMS

Patrick Rocha

CS4490Z

Department of Computer Science

Roberto Solis-Oba, Dept. of Computer Science

Nazim Madhavji, Dept. of Computer Science

Introduction

- Regression algorithms help us understand and predict variables in data.
- There is a need for a deeper understanding of these algorithms through comparison.
- While direct comparisons have been made, not many thorough analyses of the results exist.
- Six algorithms have been compared using four different datasets.
- New conclusions have been made by analysing the results.
- These conclusions can further our ability to use these algorithms effectively.

Table of Contents

Background & Related Work

- There are many works in regression analysis.
- These works typically involve analysing new methods or algorithms.
- Examples:
  - Research that explores categorical variable encoding techniques (Seger, 2018).
  - Research on Genomic selection (Ogutu et al., 2012).
- Some other papers are more comparative but focus on the attributes of one specific dataset.
- Example:
  - A comparison of regression algorithms on ammonium exchange (Karadag et al., 2007).

Research Methodology

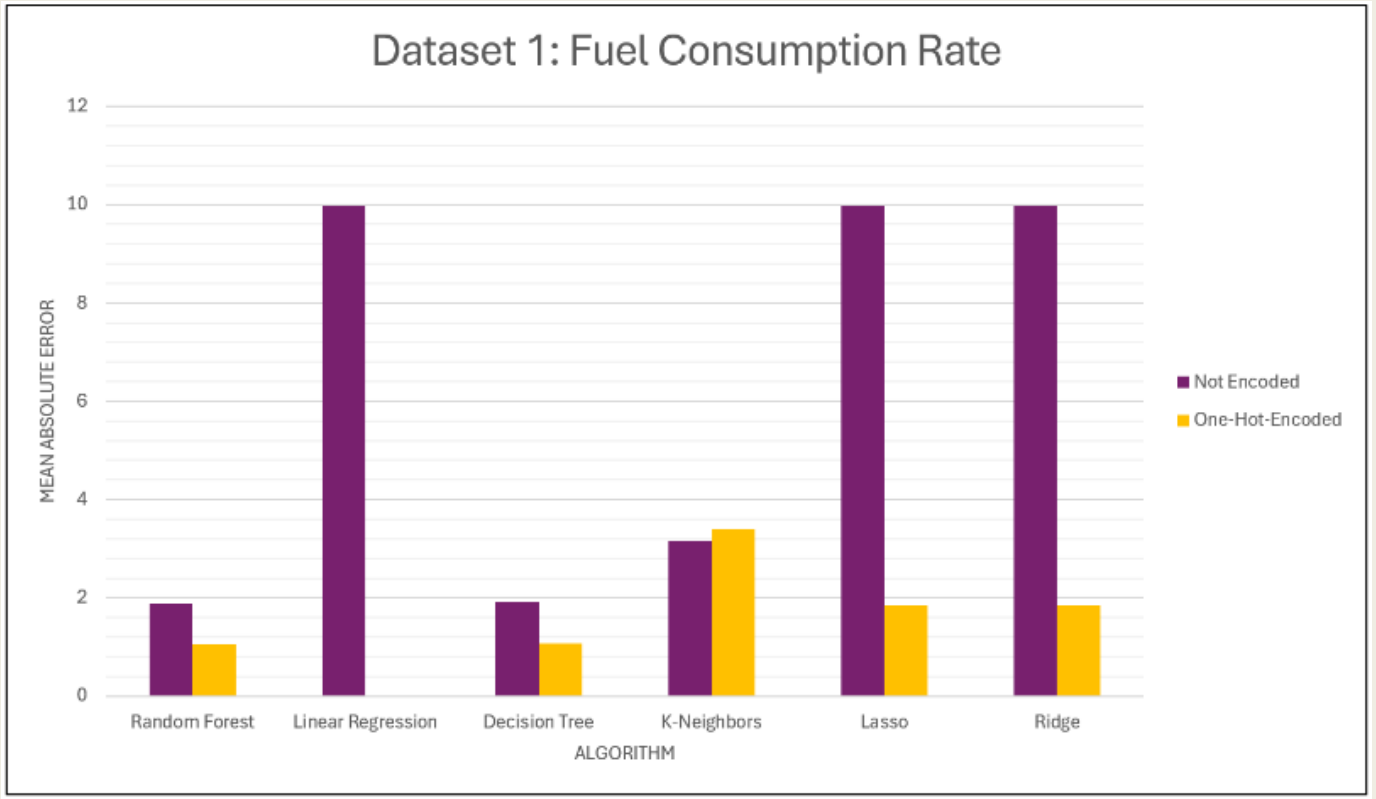
- This comparison will consist of six regression algorithms:
  - Random Forest, Linear Regression, Decision Tree, K-Neighbour, Lasso, and Ridge.
- The programming was done on Python with the help of Scikit-learn and Pandas.
- “Python has been focused by researchers due to its concise, elegant and clear language” (Rong et al., 2018).
- Models will be created for each algorithm on four different datasets.
- MAE (mean absolute error) values will be obtained using cross validation.
- The parameters will be optimized using GridSearchCV.

Results

- The MAE values will be obtained on the following datasets:
  - Fuel consumption rate, wine quality, individual income, housing prices.
- There will be two values for each model: one that includes one-hot encoding and one without it.
  - This will help further the depth at which the analysis can reach.
- Afterwards, an analysis will be made to make conclusions from the results.

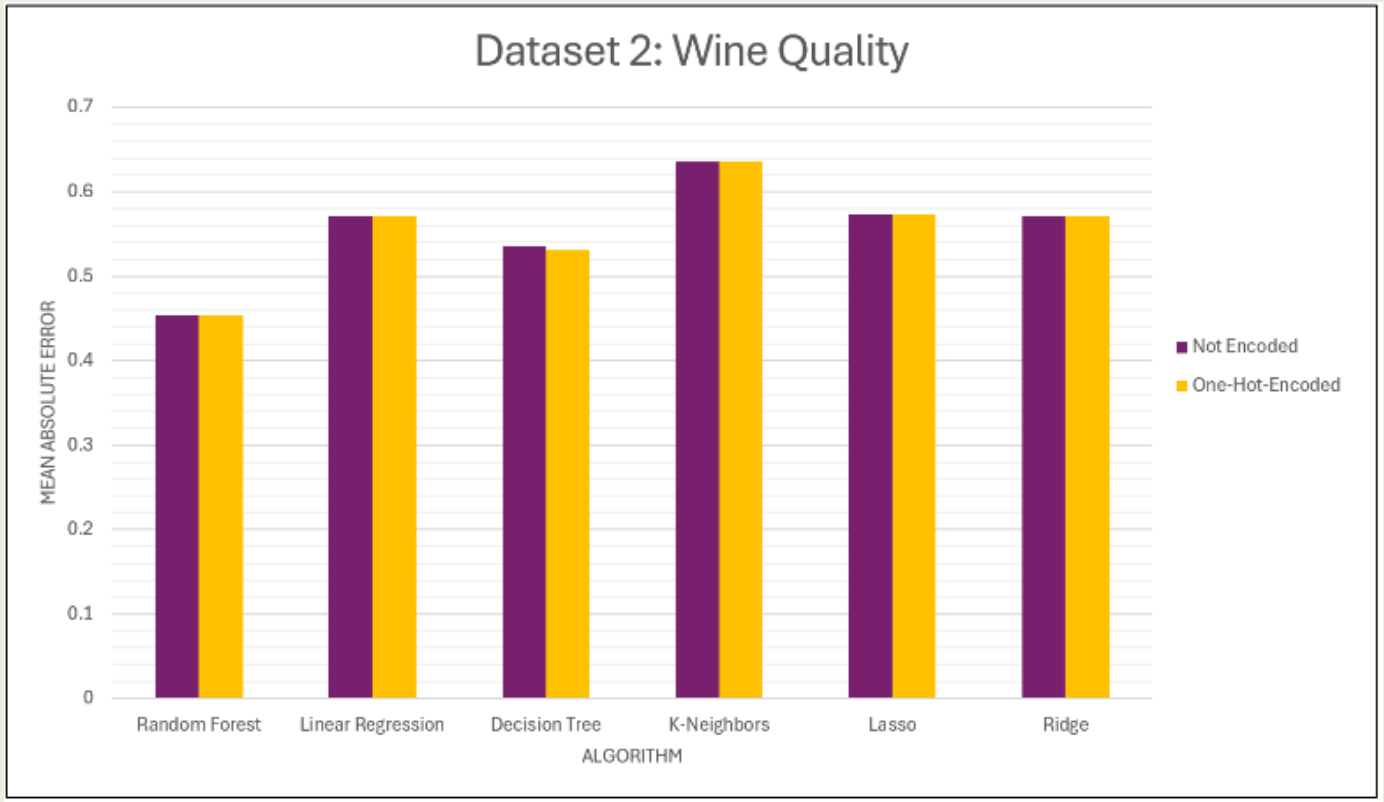
Dataset 1: Fuel Consumption Rate

- There are 5431 rows and 15 columns in this dataset.
- The one-hot encoded value for linear regression was abnormally high.
- The categorical columns have a lot of unique values.
- Adding many columns can have negative effects on regression algorithms.



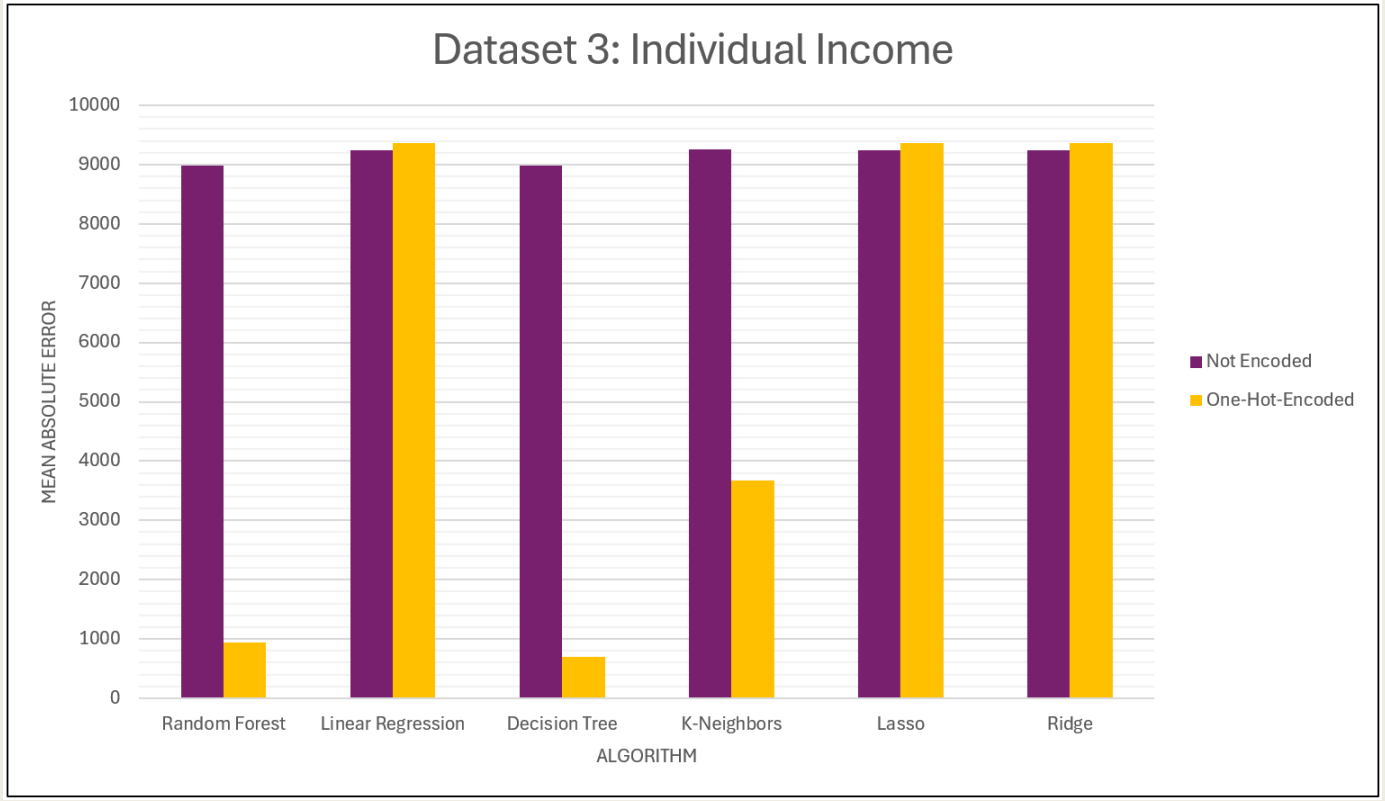
Dataset 2: Wine Quality

- There are 6497 rows and 13 columns in this dataset.
- Only one column is categorical and has two unique values.
- The difference in MAE values between one-hot encoding and no encoding is minimal.
- The feature variables contain relevant information for predicting the target variable.



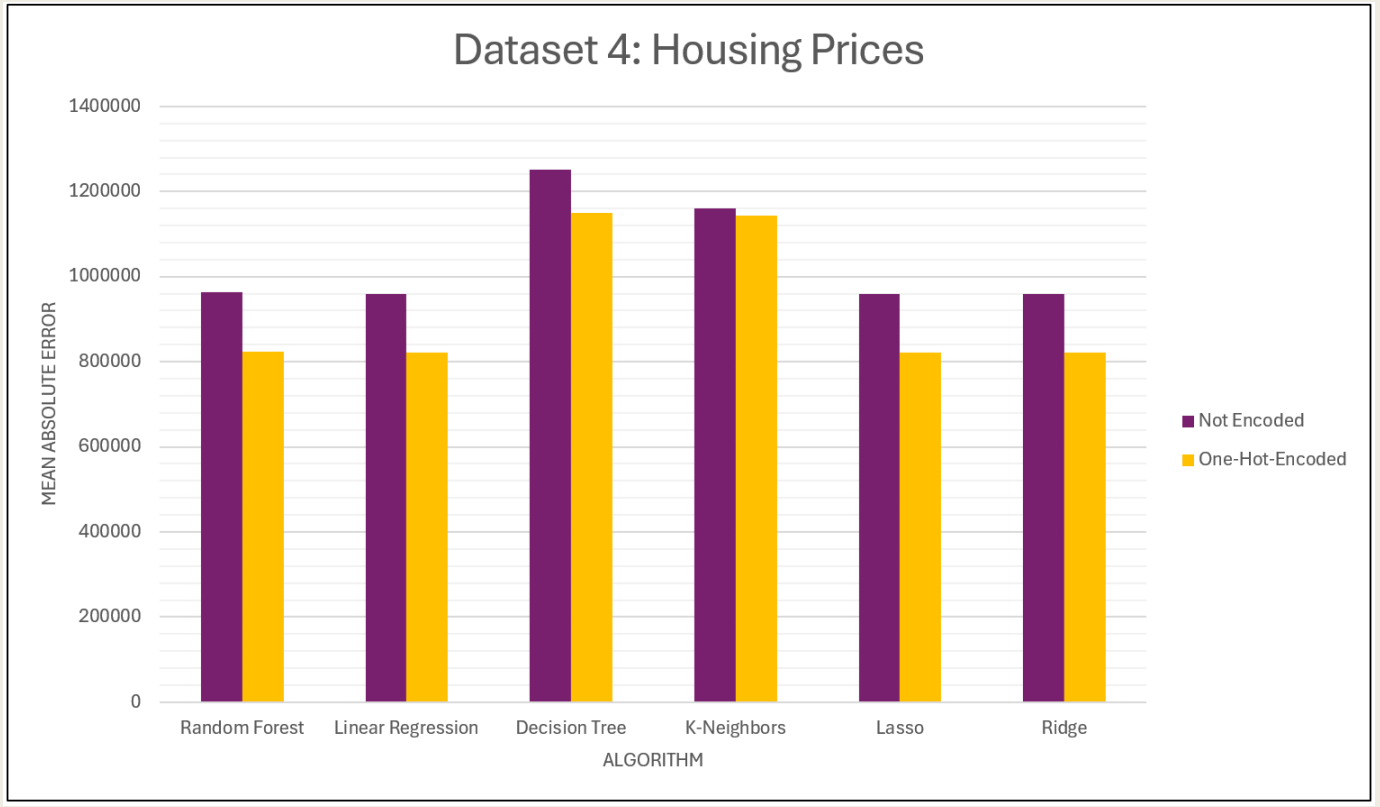
Dataset 3: Individual Income

- There are 1338839 rows and 9 columns in this dataset.
- Without one-hot encoding, every algorithm had similar results.
- When one-hot encoding was used, major differences occurred in the random forest, decision tree, and k-neighbour algorithms.
- The other three algorithms did not see much of a difference.



Dataset 4: Housing Prices

- There are 545 rows and 13 columns in this dataset.
- Each categorical column has very few unique values.
- Each algorithm sees small and consistent improvements when one-hot encoding is used.
- The differences between the MAE values of each algorithm are small.



Critical Analysis

- When many columns were created from one-hot encoding, the linear regression, lasso, and ridge algorithms performed much worse.
- These are all linear algorithms, which suggests that there is a correlation between the performance of linear algorithms and the reliance on categorical values.
- Random forest and decision tree were consistently effective. One-hot encoding always provided benefits.
- The k-nearest neighbours algorithm lies midway between the previous two groups.
- Overall, random forest and decision tree saw the best results.

Novelty

- The results showcase strengths and weaknesses of the included algorithms that are not acknowledged in the field.
- These conclusions are not only useful within the four datasets, but within any dataset that shares similar attributes.
  - These attributes include things such as number of rows, number of columns, and the ratio between numerical columns and categorical columns.
- While there are many different regression algorithms, they are all grouped together into categories (linear algorithms, tree algorithms, etc). The results of each algorithm used in the experiment are meaningful to other algorithms.

Limitations of Results

- There are many ways to create models and fit them onto data. Each method can have differences on the results.
- Scikit-learn may limit the extent at which the results can be utilised.
- This does not mean the results are meaningless if using other methods.
  - Each method must abide by the definition of an algorithm.
  - The patterns discovered in the experiment can be tested on other methods if one is sceptical.

Impact on Theory & Practice

- There is always contention on what algorithms to use given a certain kind of dataset.
  - “Least squares method is used for finding the parameters of the models, but linear regression is criticized since it results in different linearized forms” (Karadag et al., 2007).
- To help come to a consensus, a deep understanding of when to use certain algorithms and why they work when they do is crucial.
- This research project builds upon the knowledge we have regarding these ideas and will help ensure other researchers create optimal models.

Conclusions

- The main objective of this research project was to discover new information regarding the relationship between certain dataset qualities and the performance of regression algorithms.
- The patterns found in the analysis can help improve the understanding of when it is optimal to use each pattern.
- As stated in the introduction, previous comparative experiments have been made. Rather than doing something new, this research project adds onto the current knowledge we have on machine learning.
- It is meant to push the world’s understanding of regression algorithms.

Future Work & Lessons Learnt

- There are many ways to expand upon the results. These include:
  - Adding more algorithms.
  - Adding more datasets.
  - Testing different libraries and programming languages.
  - Using different imputation and encoding methods.
- It is important to remember that the purpose of a comparative project like this is to discover patterns that can be applied to other scenarios.
- There is a limit to the amount you should add, but there would also be a benefit to adding more than what was done here.
- In addition to the patterns found, this thesis project showed how much room there is for improvement among comparative analyses.

References

- Karadag, D., Koc, Y., Turan, M., & Ozturk, M. (2007). *A comparative study of linear and non-linear regression analysis for ammonium exchange by Clinoptilolite Zeolite*.
- Ogutu, J. O., Schulz-Streeck, T., & Piepho, H.-P. (2012). *Genomic Selection Using Regularized Linear Regression Models: Ridge Regression, Lasso, Elastic Net and Their Extensions*.
- Rong, S., & Bao-wen, Z. (2018). *The research of regression model in Machine Learning Field*. MATEC Web of Conferences, 176, 01033. https://doi.org/10.1051/matecconf/201817601033
- Seger, C. (2018). *An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing*