[[1]](#footnote-1)

Reproducing “An Empirical Comparison of Supervised Learning Algorithms

Albert Leon Wu  
A12377791

*Abstract*— To further learn and reinforce the several supervised learning algorithms presented in COGS118A, I proceeded to confirm the large scale empirical results of “An Empirical Comparison of Supervised Learning Algorithms” by Caruana and Niculescu-Mizil on three different supervised learning algorithms: SVMs, K-Nearest Neighbors, and Decision Trees. These methods were evaluated based on the train and test accuracies. Moreover, an additional analysis was added based on the performance of these methods based on train-test split percentages and how they affected performance.

# INTRODUCTION

M

ANY research papers focus on the exploration and discovery of forefront and groundbreaking advancements of their respective fields. However, many times reproducing the results of others is just as important as the research itself. If the results of another experiment cannot be reproduced, then should we even accept the findings of the original claims?

With machine learning becoming such a hot-topic, and its applications are beginning to be used in a wide variety of industries, it is important to understand when to use one algorithm over another. The industry dictates what a “good” algorithm is. which performance metrics, such as precision, recall, and lift. For example, when analyzing earthquakes, we want high recall over precision. Missing just one earthquake could be devastating. On the other hand, the removal of brain cells in brain surgery would require high precision – we cannot just remove all the brain cells to ensure that we get rid of all the bad ones! A good balance between the two, is the F-Score, which accounts for both at once.

As we proceed, the purpose of this paper is to reproduce the results of “An Empirical Comparison of Supervised Learning Algorithms”, and add another dimension to the analysis: how does splitting percentage between training and testing sets affect the overall accuracy of the examined supervised learning methods?

# Data and Problem Description

The Adult Data Set, which can be found from the online UCI Data Repository, was used as the backbone of the findings. There are several attributes, such as ‘age’, ‘education’, ‘maritial status’, ‘sex’, etc.

Given these attributes, for each supervised learning algorithm, we want to classify whether a person makes less than or equal to $50,000 USD.

# Methodology

## Data Cleansing

The Adult dataset was put into a data-frame. The python method “get\_dummies” applies a one-hot encoding of the categorical variables. Since it classifies income as “<= 50k” and “>50k”, one of these columns was removed entirely, and acts as the “y” set for supervised training. We use this to better represent categorical data.

## Learning Algorithms

SVMs: We used a Linear SVM and varied the regularization parameter by factors of ten from 10^-7 to 10^3.

KNN: We use 26 values of K ranging from K = 1 to K = 26

Decision Trees: We use values 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 for depth parameter

## Performance Metrics

The above methods were evaluated based on the accuracy, recall, and F-Score. Then, the methods were ranked based on order of accuracy.

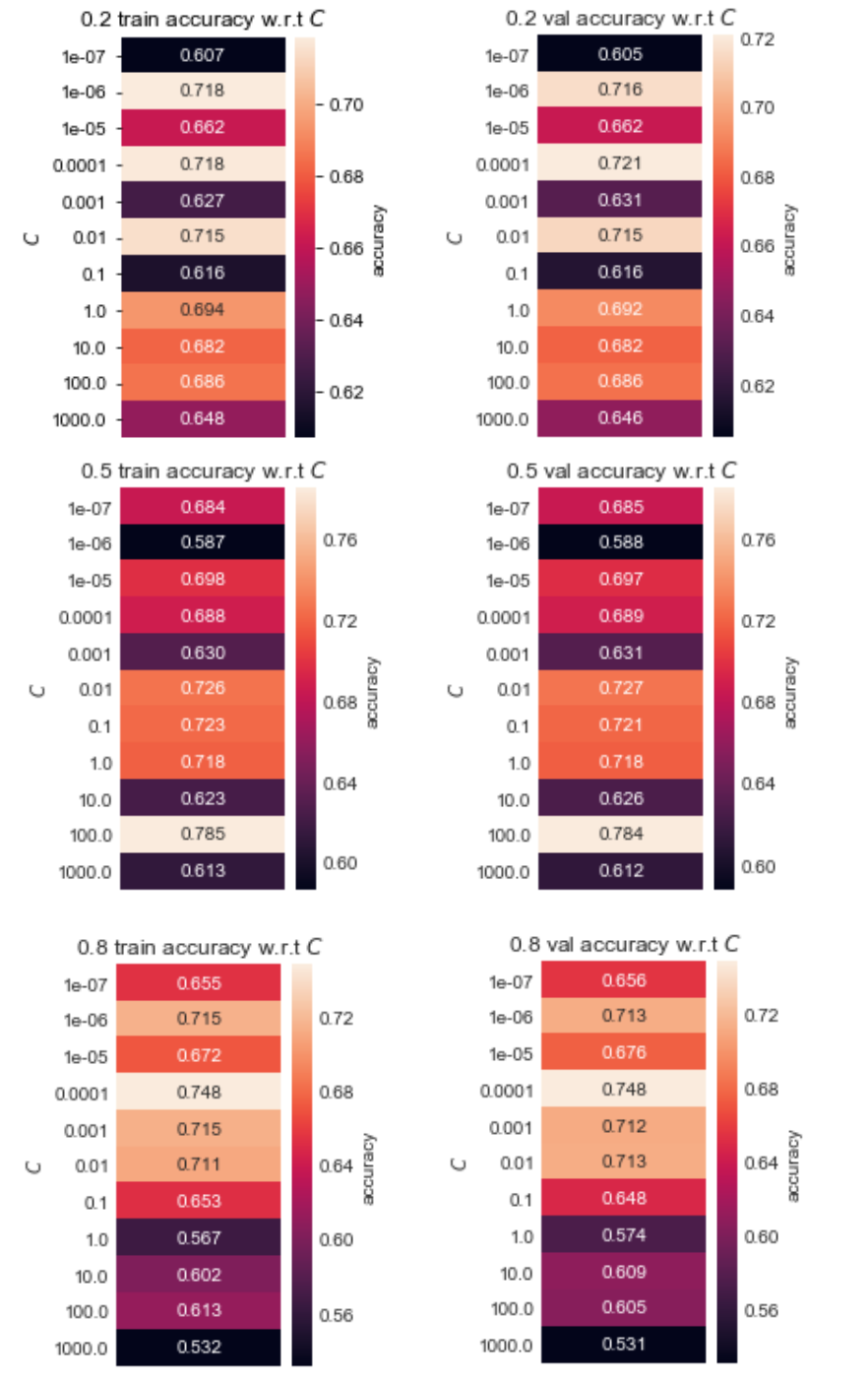
## Procedure

We gather the results by hyper tuning the parameters through Grid Search to see which parameters lead to the best accuracy for each methodology.

To get a better accuracy and attribute the findings less to randomness, we run three trials for each split-train-test percentage 0.2, 0.5, 0.8. We then return the grid-search train and test scores and average the findings for each parameter ( Regularization Parameter C for Linear SVM, Parameter k for k Nearest Neighbors, and S size of feature set for Random Forests.

# Experiment

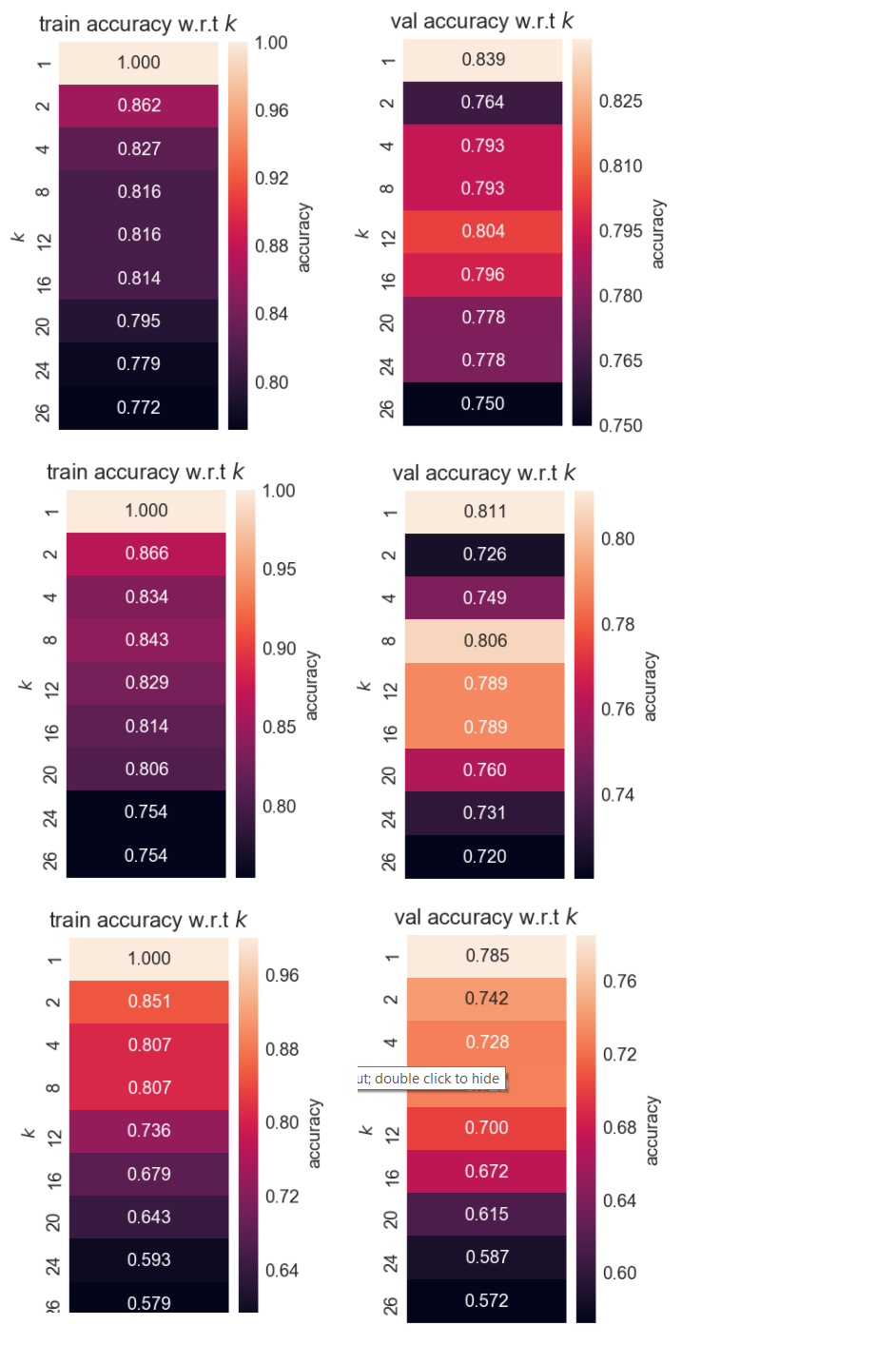
Table 1 refers to the heatmaps for Linear SVM. We can see that

Table I

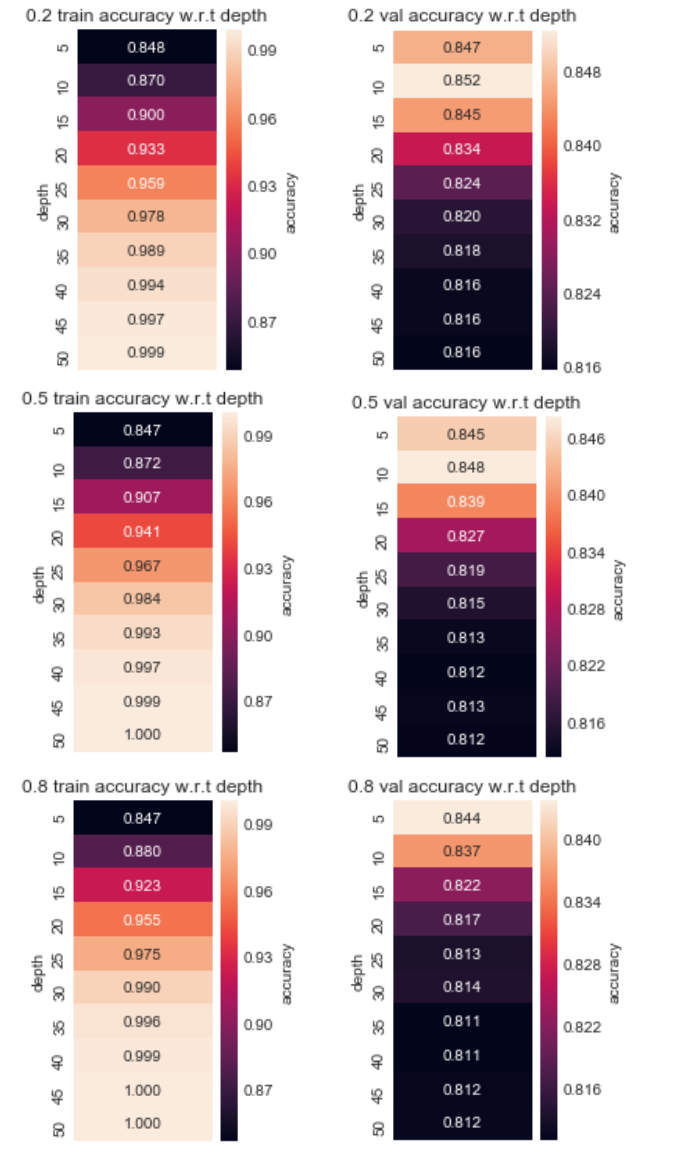
The highest averaged test accuracy given is 79% with the best parameter C being 0.0001. These results are relatively consistent with Caruana and Niculescu-Mizil. As we can see, generally there is a higher train and test accuracy when there is more data fed into the training set. Moreover, the highest train and validation accuracies can be found by looking at the maximum in the heatmaps.

Similarly, the same procedure was applied to k Nearest Neighbors, with the Parameter k being hyper-tuned by Grid Search. Table II refers to the heatmaps for k-Nearest Neighbors.

Table II



We can see that for all split percentages, the train accuracy is 1 when k = 1. To give a better percentage, k=2 was then examined. The highest averaged test accuracy given is 85.1% in the 0.8 test split percentage. These results are consistent with Caruana and Niculsescu-Mizil, whom received around a 75.6%. The difference could be attributed to the differences in cleaning the data. The best parameter is k = 2.

Table III

Lastly, we examine the Decision Tree at different depth levels. The highest averaged test accuracy given is 84.24% with the best parameter depth being 5. These results are relatively consistent with Caruana and Niculescu-Mizil whom got an accuracy of 0.843. From the graphs, we can see that for depth values over 30 we may have overfitted the data. The training accuracy for high values is 1.0, while the validation accuracies of 0.812 are the lowest for split percentages 0.5 and 0.8. However, the highest averaged accuracy that exists is 0.923 in the 0.8 split percentage with parameter depth = 15. Again, we find that the higher split percentage led to the highest averaged accuracy.

# Conclusion

In conclusion, we can see that ranking the methods in order of accuracy, we achieve the same results as Caruana and Niculescu-Mizil. The Linear SVM achieved an accuracy of 79%, KNN achieved an accuracy of 85.1%, and the decision tree received an accuracy 0f 0.923. We have found that the larger split percentage 0.8 proved to give better accuracies in all three different supervised learning algorithms. Thus, the findings are consistent.

# References

<https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf>

1. [↑](#footnote-ref-1)