COGS 118A Classification Final Project

Alvin Xiao

***Abstract*—A comprehensive study was recently conducted on supervised learning methods. To support those findings, this paper conducts additional trials on binary classification using specific classifiers (Logistic Regression, K-Nearest Neighbors, Random Forest). Different partitions of data and hyper parameter tuning through cross validation are experimented with during the training process.**

I. Introduction

Over the last few decades, machine learning has become an increasingly studied tool for researchers and businesses alike. The computational and predictive power that machine learning offers at the expense of less human labor makes it a popular field to try and optimize. In particular, some of the common problems that machine learning is suited to solve are classification problems. Much is devoted to studying models that take a set of observations with different features, and trying to predict some target class.

This paper builds on the comprehensive empirical study of supervised learning algorithms done by Caruana and Niculescu-Mizil by examining three specific classifiers: Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN). These classifiers were chosen because of their different strengths and applications in different problems. Random Forest is an ensemble method that considers multiple decision trees to get an output. Logistic Regression is a statistical method that measures the probability of an output based on the linear relationship between variables and models. KNN measures distance between a data point to labeled classes and determines what to label the new data point based on that distance.

II. Methodology

A. Learning Algorithms

For all algorithms, the Python scikit-learn library is used to train and fit the models.

Logistic Regression: the regularization parameter varies by factors of 10 from 10^{-3} to 10^{2}. Less regularization is tested with larger datasets due to lack of computational power. Max iterations of 10,000 are also established to allow more complexity and ensure convergence.

KNN: the number of neighbors varies from the range of 1 to the minimum of 20 and the length of the dataset, inclusive. Distance weighted KNN is specifically chosen to reproduce results from Caruana and Niculescu-Mizil. Euclidean distance is measured.

Random Forest: the size of the feature set considered at each split varies from 2 to 10 inclusive, with a step size of 2. This is reflected as the max features parameter.

B. Evaluation Metric

To get a general sense of how all the classifiers perform against one another, accuracy as the evaluation metric. Accuracy can be calculated as the following:

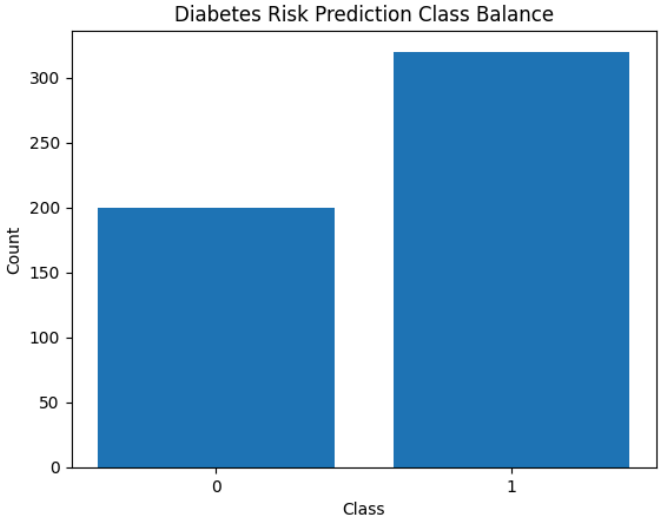
$$accuracy = \frac{\#\ of\ correct\ predictions}{\#\ of\ total\ predictions}$$

C. Datasets

All datasets were pulled from the UCI Machine Learning Repository.

Early Stages Diabetes Risk Prediction: the original dataset has 520 instances and 16 features. The positive class being predicted is diabetic likelihood or condition and the negative class is no diabetic condition being found or predicted. Most features are binary and simple dictionaries were used to map the binary strings to numerical inputs.

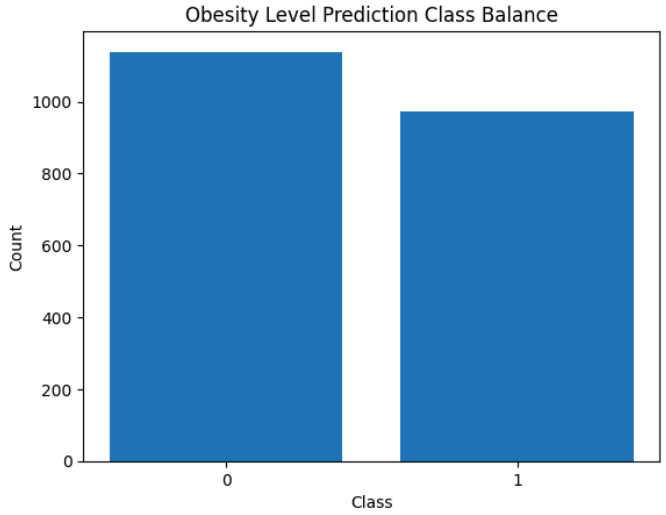
FIGURE 1. Target class balance for diabetes risk dataset



Estimation of Obesity Levels Based On Eating Habits and Physical Condition: the original dataset has 2111 instances and 16 features. The original dataset was prepared for multiclass classification, so to convert it into binary classification, the new positive class is any level of obesity and the new negative class is no obesity being detected.

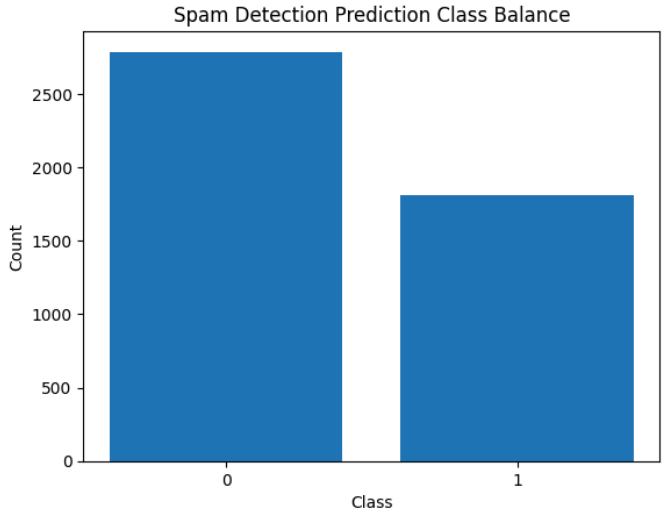
Features are of various data types, including continuous and categorical values, where there were both ordinal and nominal data among the categorical features. Ordinal encoding to preserve order was applied on the ordinal data and one-hot encoding was applied on the nominal data.

FIGURE 2. TARGET CLASS BALANCE FOR OBESITY LEVEL DATASET



Spambase: the original dataset has 4601 instances and 57 features. The positive class signified that spam, and the negative class was not spam. All features in this dataset were continuous and of different scales.

FIGURE 3. TARGET CLASS BALANCE FOR SPAM DETECTION DATASET



III. EXPERIMENT

Three classifiers have been selected from those tested in Caruana and Niculescu-Mizil, and three datasets have been chosen from the UCI repository. Each classifier will be tested on each dataset for a total of nine models. Each model on each dataset employs cross-validation in order to find the best respective parameters. The specific parameters being optimized were the same parameters being tested in Caruana and Niculescu-Mizil, such as regularization values for Logistic Regression or feature set sizes for Random Forest. However, not all parameters tested in the empirical study are tested in this paper due to lack of computational power. Each classifier on each dataset will consider three different data partitions, specifically the following: 20% training and 80% testing split, 50% training and 50% testing split, and 80% training and 20% testing split. Three trials will be conducted for each data partition, and the training, validation, and testing errors will be averaged across the three trials per partition. Cross validation to optimize parameters is practiced during each split, so each partition's classifier may have different parameters. Accuracy is used as the evaluation metric for each trial.

Data on the validation accuracies during parameter tuning were stored in two dimensional arrays. The following heatmaps show the parameters and their attained accuracies during different partitions of data. The validation accuracies are averaged across trials.

FIGURE 4. DIABETES RISK LOGISTIC REGRESSION PARAMETER TUNING

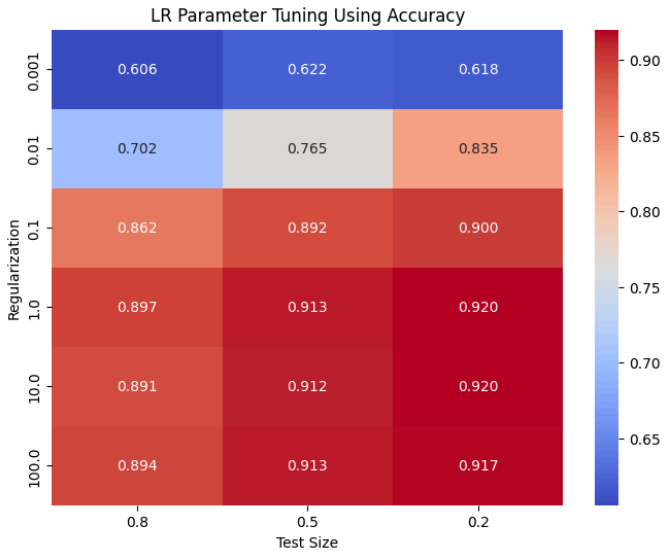


FIGURE 5. Diabetes risk k-nearest neighbors parameter tuning

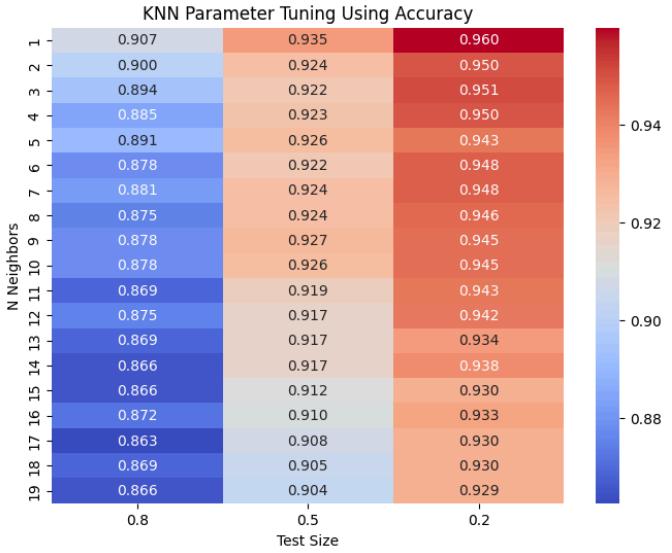
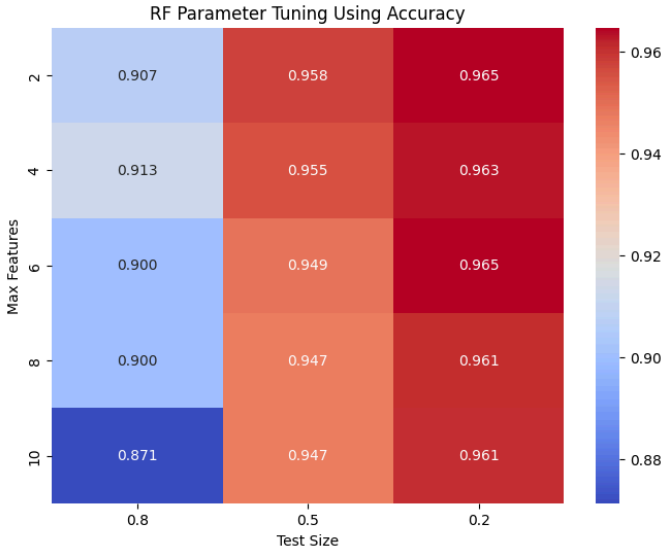


FIGURE 6. Diabetes risk random forest parameter tuning



The following tables show the accuracies of the classifiers for each partition averaged across trials. Each table represents a different dataset.

TABLE I. Results for diabetes risk prediction

MODEL	Mean Accuracy			
	Partition	*Training*	*Validation*	*Testing*
RF	80/20	0.995	**0.969**	**0.968**
RF	50/50	0.994	0.954	0.967
RF	20/80	0.990	0.926	0.922
KNN	80/20	0.994	0.960	0.952
KNN	50/50	0.994	0.940	0.951
KNN	20/80	**1.000**	0.911	0.909
LR	80/20	0.934	0.921	0.929
LR	50/50	0.944	0.917	0.922
LR	20/80	0.949	0.904	0.878

TABLE II. Results for obesity level prediction

MODEL	Mean Accuracy			
	Partition	*Training*	*Validation*	*Testing*
RF	80/20	**1.000**	**0.994**	**0.991**
RF	50/50	**1.000**	0.989	0.990
RF	20/80	**1.000**	0.979	0.985
KNN	80/20	**1.000**	0.980	**0.991**
KNN	50/50	**1.000**	0.973	0.975
KNN	20/80	**1.000**	0.955	0.957
LR	80/20	0.989	0.983	0.985
LR	50/50	0.987	0.979	0.978
LR	20/80	0.983	0.957	0.960

TABLE III. Results for spam detection prediction

MODEL	Mean Accuracy			
	Partition	*Training*	*Validation*	*Testing*
RF	80/20	0.999	**0.951**	**0.955**
RF	50/50	0.999	0.948	0.947
RF	20/80	**1.000**	0.938	0.937
KNN	80/20	0.999	0.820	0.815
KNN	50/50	0.999	0.793	0.806
KNN	20/80	**1.000**	0.771	0.766
LR	80/20	0.933	0.929	0.930
LR	50/50	0.929	0.921	0.924
LR	20/80	0.939	0.922	0.918

IV. Conclusion

Most of the findings are consistent with those from Caruana and Niculescu-Mizil. The Random Forest classifier performed best on all three dataset when testing on unknown data. This suggests that the Random Forest classifier generalized best. The other models performed satisfactory as well, with KNN usually slightly outperforming Logistic Regression. However, as seen in the spam detection dataset where there are many features, KNN performs noticeably worse than Logistic Regression. It is likely that the large amount of features cause the problem to be very high in dimensions, thus causing KNN to struggle more.

Another observation is that data tends to perform better when more training data is used to fit the model. For example, the 80/20 training/testing split usually outperforms the 20/80 split, and suggests that the model is being underfit in the latter. There is some variability in the evaluation due to the randomness of the data partition, but some of the randomness is mitigated by taking the average accuracy across trials.

One of the biggest challenges included the amount of time it took to run these trials. Due to parameter tuning, cross validation, and the heavy computational time required for some models, smaller datasets were preferred for this paper. In the future, it would be interesting to evaluate these classifiers datasets with hundreds of thousands of observations to see if these trends still hold true.

References

[1] Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168).