Drew Chaudhari
COGS 118A Final Project

Abstract

This following project discusses the use of machine learning algorithms such as classifiers including Artificial Neural Networks (ANN), Logistic Regression (LOGREG), and K-Nearest Neighbors (KNN). These classifiers are performed and analyzed amongst three datasets which regard the Taiwanese Bankruptcy Prediction, Breast Cancer Coimbra, and Website Phishing. An important factor in this study is the use of multiple trials to ensure the performance of the classifiers is similar amongst numerous evaluations of that dataset. This is completed by using different data partitions, as the three train-test splits that were used include 20% training/80% testing, 50% training/50% testing, and 80% training/20% testing. From this, we are able to perform hyperparameter tuning to understand which classifier performs the best through analyzing the training and testing accuracy.

Introduction

The performance of these machine learning models are influenced by the datasets included in the study as well as the hyperparameters which affect the binary classification tasks. One primary performance metric used in this study to evaluate binary classification tasks was accuracy. The accuracy helps to identify the proportion of correctly classified predictions in relation to the total number of predictions. This performance metric allows for the classes in binary classification to be distinguished and understood which allows us to identify the effectiveness of each classifier in different datasets. Using this performance metric, we aim to conduct multiple trials to compare the performance of each classifier within the datasets under different train-test splits. The three classifiers we use each have methods that have unique models in respect to classifying the data. KNN is an algorithm that calculates the distance between data and is well adaptable to various datasets. Logistic regression maps predicted values to probabilities and is used for linearly separable datasets. ANN use of non-linear activation allows this model to interact with complex datasets and perform robustly.

Data and Problem Description

The datasets chosen and used in this study from the UCI Machine Learning Repository include the Taiwanese Bankruptcy Prediction, Breast Cancer Coimbra, and Website Phishing. In the Taiwanese Bankruptcy Prediction dataset, business regulations were taken and used to define bankruptcy as individuals were classified based on their financial status. Breast Cancer Coimbra depicts the features of sixty four patients with breast cancer and uses quantitative predictors to indicate the presence of breast cancer. Lastly, the Website Phishing dataset uses numerical and categorical data to identify different features regarding legitimate and phishy websites. This study identifies problems regarding the best performing classifier within different datasets under specific hyperparameters and data size as well as the adaptability of each classifier with various data features.

Method Description

Dataset Preparation

Preprocessing the datasets was used in a variety of ways such as standardizing the range within the data, train-test data splits, and binary labeling. The range of the features were standardized so that they had a comparable range before applying the algorithm. Separating the datasets into three different train test splits included 20/80, 50/50, and 80/20, which evaluated the classifiers at different types of training data. The classes of the datasets were joined when needed so that binary classification was implemented before the classifiers performed through each dataset and train test split.

Classifiers

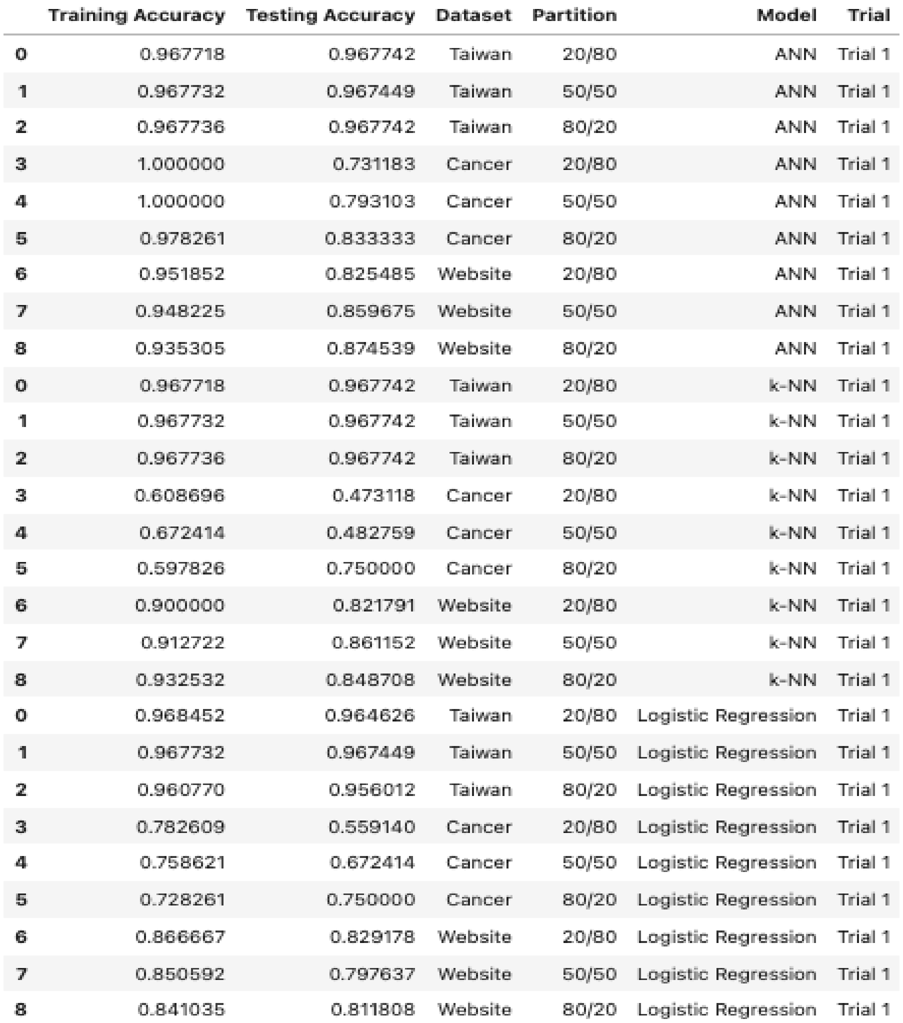
KNN used a 5 fold cross validation for the grid search to find the best combination of neighbors and distance metrics including both Euclidean and Manhattan, as well as using the values 3, 5, and 10 for the number of neighbors applied. For Logistic Regression, the solvers used included liblinear and lbfgs, and the regularization strength (C) was set from 0.01 to 100 for the grid. ANN used configurations of hidden layers such as one layer (100 nodes) and two-layer networks (100, 50) nodes, as functions ReLU and Tanh were activated with SGD and Adam optimizers.

Algorithm Setup

Performance metrics such as testing and training accuracy were implemented for each classifier while using optimized hyperparameters. These hyperparameters were obtained from a completed grid search using a 5-fold cross validation. The algorithm was completed three times so that these multiple trails can depict the accuracy and reliability of the data from the classifier which is also shown in the bar plot visualization.

Experiments

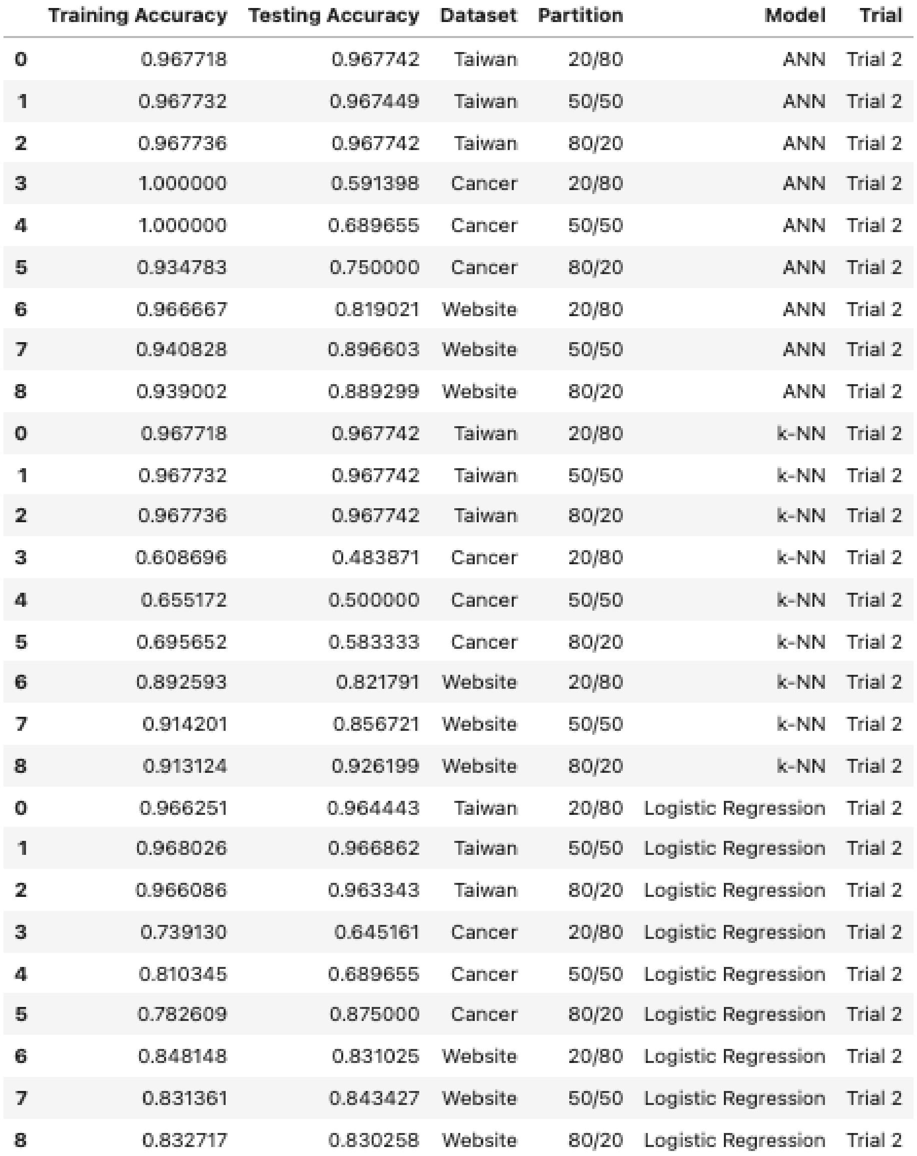
Table 1



	Training Accuracy	Testing Accuracy	Dataset	Partition	Model	Trial
0	0.967718	0.967742	Taiwan	20/80	ANN	Trial 1
1	0.967732	0.967449	Taiwan	50/50	ANN	Trial 1
2	0.967736	0.967742	Taiwan	80/20	ANN	Trial 1
3	1.000000	0.731183	Cancer	20/80	ANN	Trial 1
4	1.000000	0.793103	Cancer	50/50	ANN	Trial 1
5	0.978261	0.833333	Cancer	80/20	ANN	Trial 1
6	0.951852	0.825485	Website	20/80	ANN	Trial 1
7	0.948225	0.859675	Website	50/50	ANN	Trial 1
8	0.935305	0.874539	Website	80/20	ANN	Trial 1
0	0.967718	0.967742	Taiwan	20/80	k-NN	Trial 1
1	0.967732	0.967742	Taiwan	50/50	k-NN	Trial 1
2	0.967736	0.967742	Taiwan	80/20	k-NN	Trial 1
3	0.608696	0.473118	Cancer	20/80	k-NN	Trial 1
4	0.672414	0.482759	Cancer	50/50	k-NN	Trial 1
5	0.597826	0.750000	Cancer	80/20	k-NN	Trial 1
6	0.900000	0.821791	Website	20/80	k-NN	Trial 1
7	0.912722	0.861152	Website	50/50	k-NN	Trial 1
8	0.932532	0.848708	Website	80/20	k-NN	Trial 1
0	0.968452	0.964626	Taiwan	20/80	Logistic Regression	Trial 1
1	0.967732	0.967449	Taiwan	50/50	Logistic Regression	Trial 1
2	0.960770	0.956012	Taiwan	80/20	Logistic Regression	Trial 1
3	0.782609	0.559140	Cancer	20/80	Logistic Regression	Trial 1
4	0.758621	0.672414	Cancer	50/50	Logistic Regression	Trial 1
5	0.728261	0.750000	Cancer	80/20	Logistic Regression	Trial 1
6	0.866667	0.829178	Website	20/80	Logistic Regression	Trial 1
7	0.850592	0.797637	Website	50/50	Logistic Regression	Trial 1
8	0.841035	0.811808	Website	80/20	Logistic Regression	Trial 1

The first trial that was completed demonstrates each classifier using three different train-test splits for all three datasets. From this, we can observe that the best performing classifier was Logistic Regression with a testing accuracy of 82.67%, which particularly performed extremely well within the Taiwanese Bankruptcy Prediction dataset due to the numerical separability it contained. ANN had a testing accuracy of 80.12% and excelled in datasets with complex components containing non-linear decision boundaries such as the Website Phishing dataset. KNN seemed to struggle with specific scaling data as depicted in the Website Phishing which resulted in the lowest testing accuracy score of 76.33%. The 20/80 train-test split demonstrated that KNN struggled while Logistic Regression successfully adapted to the datasets. ANN improved with more data through the 50/50 split and Logistic Regression still maintained the highest testing accuracy, as all the models improved for the 80/20 train-test split.

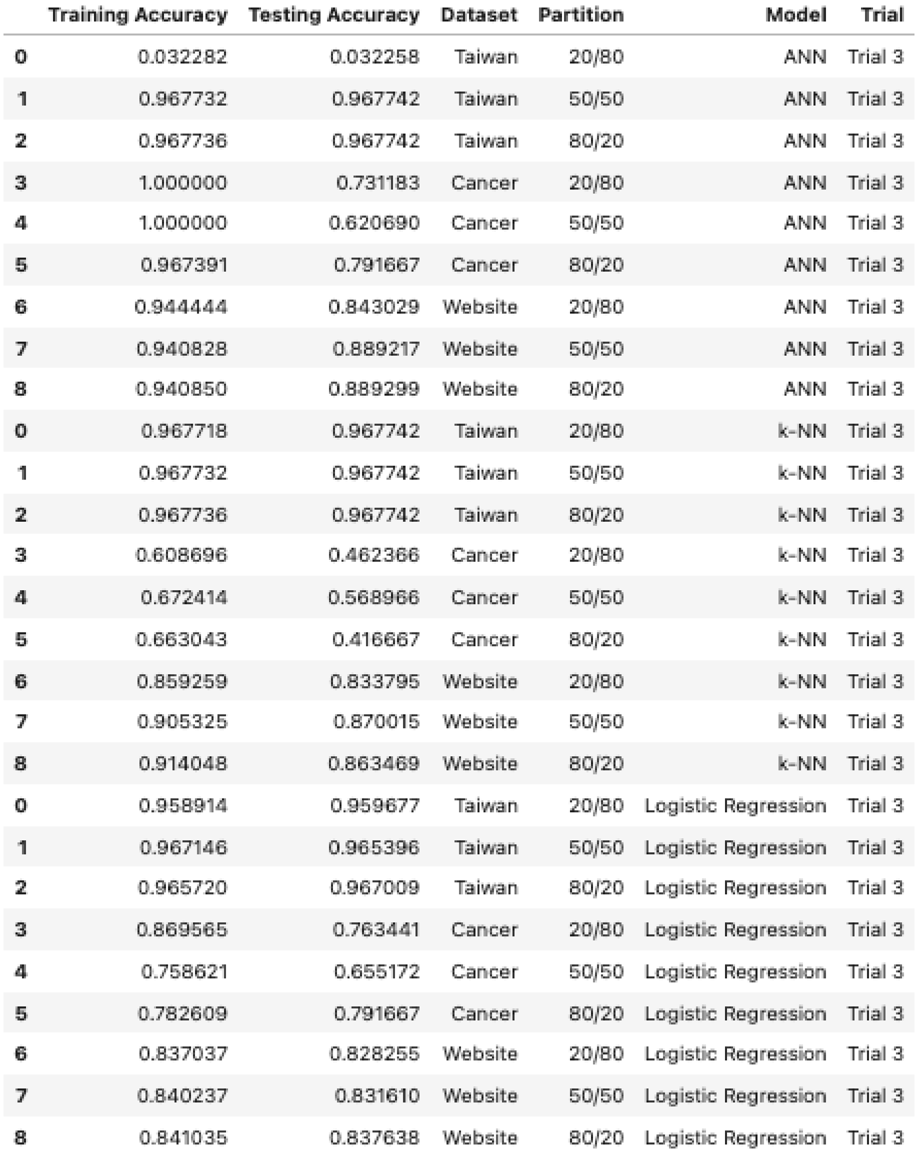
Table 2



	Training Accuracy	Testing Accuracy	Dataset	Partition	Model	Trial
0	0.967718	0.967742	Taiwan	20/80	ANN	Trial 2
1	0.967732	0.967449	Taiwan	50/50	ANN	Trial 2
2	0.967736	0.967742	Taiwan	80/20	ANN	Trial 2
3	1.000000	0.591398	Cancer	20/80	ANN	Trial 2
4	1.000000	0.689655	Cancer	50/50	ANN	Trial 2
5	0.934783	0.750000	Cancer	80/20	ANN	Trial 2
6	0.966667	0.819021	Website	20/80	ANN	Trial 2
7	0.940828	0.896603	Website	50/50	ANN	Trial 2
8	0.939002	0.889299	Website	80/20	ANN	Trial 2
0	0.967718	0.967742	Taiwan	20/80	k-NN	Trial 2
1	0.967732	0.967742	Taiwan	50/50	k-NN	Trial 2
2	0.967736	0.967742	Taiwan	80/20	k-NN	Trial 2
3	0.608696	0.483871	Cancer	20/80	k-NN	Trial 2
4	0.655172	0.500000	Cancer	50/50	k-NN	Trial 2
5	0.695652	0.583333	Cancer	80/20	k-NN	Trial 2
6	0.892593	0.821791	Website	20/80	k-NN	Trial 2
7	0.914201	0.856721	Website	50/50	k-NN	Trial 2
8	0.913124	0.926199	Website	80/20	k-NN	Trial 2
0	0.966251	0.964443	Taiwan	20/80	Logistic Regression	Trial 2
1	0.968026	0.966862	Taiwan	50/50	Logistic Regression	Trial 2
2	0.966086	0.963343	Taiwan	80/20	Logistic Regression	Trial 2
3	0.739130	0.645161	Cancer	20/80	Logistic Regression	Trial 2
4	0.810345	0.689655	Cancer	50/50	Logistic Regression	Trial 2
5	0.782609	0.875000	Cancer	80/20	Logistic Regression	Trial 2
6	0.848148	0.831025	Website	20/80	Logistic Regression	Trial 2
7	0.831361	0.843427	Website	50/50	Logistic Regression	Trial 2
8	0.832717	0.830258	Website	80/20	Logistic Regression	Trial 2

In the second trial, the ranking of how well the classifiers performed remained the same, however, there were some noticeable differences and changes with the testing accuracy from the first trial. ANN improved to a testing accuracy of 82.45%, and this was specifically illustrated in the 20/80 train-test split as this classifier struggled less due to improved regularization. Logistic Regression performed the best in the Taiwanese Bankruptcy Prediction and Breast Cancer Coimbra datasets as it increased to a testing accuracy score of 83.12%. We can infer from the KNN classifier that the reduction in its sensitivity to noise resulted in an increase in the testing accuracy to 77.89%. As the results for the 50/50 split remained the same, the 80/20 split showed a higher testing accuracy for all the classifiers.

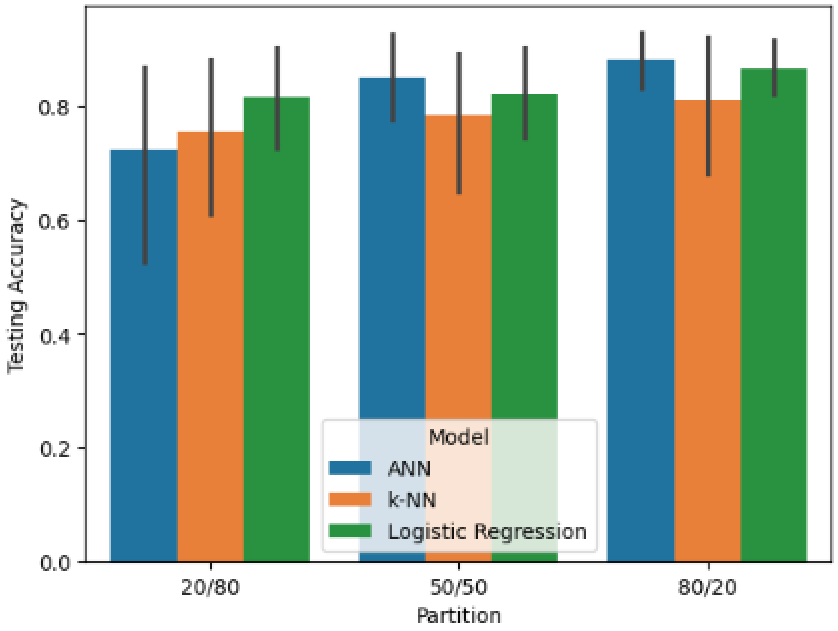
Table 3



	Training Accuracy	Testing Accuracy	Dataset	Partition	Model	Trial
0	0.032282	0.032258	Taiwan	20/80	ANN	Trial 3
1	0.967732	0.967742	Taiwan	50/50	ANN	Trial 3
2	0.967736	0.967742	Taiwan	80/20	ANN	Trial 3
3	1.000000	0.731183	Cancer	20/80	ANN	Trial 3
4	1.000000	0.620690	Cancer	50/50	ANN	Trial 3
5	0.967391	0.791667	Cancer	80/20	ANN	Trial 3
6	0.944444	0.843029	Website	20/80	ANN	Trial 3
7	0.940828	0.889217	Website	50/50	ANN	Trial 3
8	0.940850	0.889299	Website	80/20	ANN	Trial 3
0	0.967718	0.967742	Taiwan	20/80	k-NN	Trial 3
1	0.967732	0.967742	Taiwan	50/50	k-NN	Trial 3
2	0.967736	0.967742	Taiwan	80/20	k-NN	Trial 3
3	0.608696	0.462366	Cancer	20/80	k-NN	Trial 3
4	0.672414	0.568966	Cancer	50/50	k-NN	Trial 3
5	0.663043	0.416667	Cancer	80/20	k-NN	Trial 3
6	0.859259	0.833795	Website	20/80	k-NN	Trial 3
7	0.905325	0.870015	Website	50/50	k-NN	Trial 3
8	0.914048	0.863469	Website	80/20	k-NN	Trial 3
0	0.958914	0.959677	Taiwan	20/80	Logistic Regression	Trial 3
1	0.967146	0.965396	Taiwan	50/50	Logistic Regression	Trial 3
2	0.965720	0.967009	Taiwan	80/20	Logistic Regression	Trial 3
3	0.869565	0.763441	Cancer	20/80	Logistic Regression	Trial 3
4	0.758621	0.655172	Cancer	50/50	Logistic Regression	Trial 3
5	0.782609	0.791667	Cancer	80/20	Logistic Regression	Trial 3
6	0.837037	0.828255	Website	20/80	Logistic Regression	Trial 3
7	0.840237	0.831610	Website	50/50	Logistic Regression	Trial 3
8	0.841035	0.837638	Website	80/20	Logistic Regression	Trial 3

The last trial run ultimately showed the best testing accuracy for all the classifiers. Logistic Regression had a testing accuracy of 83.90%, which maintained similar to its previous trials but slightly increased and remained performing well throughout the trials. On the other hand, ANN reached 84.10% accuracy, as this was the highest accuracy throughout all the datasets and the only classifier to have noticeable change in results for the 20/80 train-test split. Hyperparameter optimization improved the accuracy of KNN to raise its testing accuracy to 79.15%. The 50/50 split proved that Logistic Regression and ANN performed better, however, in the 80/20 split, all models had a higher testing accuracy throughout the datasets.

Visualization



The graph above shows the performance of the KNN, ANN, and Logistic Regression classifiers for all three datasets while performing through each partition including the train-test splits of 20/80, 50/50, and 80/20. As illustrated in the graph, the 20/80 partition has the lowest average testing accuracy amongst all three classifiers. Although there was a reduced amount of data, Logistic Regression held the highest accuracy as ANN followed, but KNN was unable to perform well due to the limited training samples. The 50/50 split demonstrated more of a consistent performance amongst the classifiers as all three slightly improved. The highest average accuracy scores however were represented in the 80/20 partition as classifiers improved their performance as ANN showed the biggest drastic improvement. Logistic Regression outperformed the other classifiers as it maintained a strong and consistent average regardless of the partition size. ANN excelled in the 80/20 train-test split as this depicts that it needs more training data to perform well in complex datasets such as Website Phishing. Although KNN improved throughout the partitions, it remained the most underperforming classifier out of the three because of its sensitivity towards these datasets which included feature scaling.

Conclusion

From this study, we were able to investigate how classifiers such as KNN, Logistic Regression, and ANN vary in their performance with different data train-test splits as well as various datasets. The best performing classifier was Logistic Regression, which was the most reliable and consistent model over the three different partitions for all the datasets. It was able to use its robustness to perform well even when the training data was limited as it became clear that it is successful in linearly separable datasets. ANN was the second best performing classifier as it improved over time in the partitions containing larger training datasets such as the 80/20 train-test split due to excelling in non-linear relationships within datasets. Lastly, KNN improved after hyperparameter tuning and showed an improvement in the 80/20 partition but ultimately was the worst performing classifier and requires more preprocessing in order to succeed. Specific performance metrics in this study such as accuracy provided information regarding the difference in the classifiers performance. Accuracy exemplified that the classifiers performance for Logistic Regression and ANN was strengthened as it was not able to do the same for KNN's sensitivity to imbalance data. These findings in this study allow us to understand the usage of specific classifiers based on certain datasets and partitions that help us acknowledge which machine learning models perform well in given environments.

References

Taiwanese Bankruptcy Prediction [Dataset]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5004D.

Abdelhamid, N. (2014). Website Phishing [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5B301.

Patrcio, M., Pereira, J., Crisstomo, J., Matafome, P., Seia, R., & Caramelo, F. (2018). Breast Cancer Coimbra [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P59.