COSC 4557: ML Algorithm Selection

Beckham Carver

July 30, 2024

1 Introduction

The purpose of this assignment is to learn how to select a best performing ML algorithm for a given dataset. With algorithm selection at the forefront the data chosen is somewhat arbitrary, our goal being to investigate the presence of predictive elements in the data. The target prediction is the pride index of various universities, that rating being from survey data taken by the Campus Pride Index ranking the LGBTQ+ inclusivity of universities. This prediction will be made from features relating to population/financial statistics in the campus city. Our hypothesis is that there will be little correlation between the population/financial features and the pride rating, as the presence of a university alone will normalize the financial and population characteristics between cities.

2 Data Description

The data used for this assignment was sourced from the Campus Pride Index, unitedstateszipcodes.org, and US-Census 5-year housing data. These datasets source data between 2019-2024, the time discrepancy between them when crossing is considered to be negligible for our purposes. The pride index data contains classification of size for the campus city, student count, rating, as well as various tags for the programs offered by the university. The Zipcode data contains information regarding population estimate, and the census contains a plethora of statistics regarding housing data by Zipcode. For combining the data, the Pride Index only contains City/State information for the campus, the Zipcode data contains City/State/Zipcode information, and the Census only Zipcode. Combining these datasets involved aggregating housing and population statistics and merging by Zipcode, then City/State. After this merging the dataset, we ended with 119 rows, which was smaller than ideal.

3 Experimental Setup

While aggregating the data we created two features for measuring financial characteristics of the campus locations. These fields being the 'Zipcode Income Variance Score' and 'Housing Income Score' which rank the variance in median income between zipcodes in the city, and the median household income for the city respectively. The distribution of family/household types was also used, those being households with families, non-families, married families, and unmarried families. These features were included as they may serve as an indicators for the cultural/social makeup of the campus city. We then chose a bottom-up approach for testing models, with a selection of models from the categories listed on the scikit-learn superviseed learning documentation. The models and types chosen were: Linear (RidgeClassifier), Support Vectors (SVC), Stochastic Gradient (SGDClassifier), Neighbors (KneighborsClassifier), Naive Bayes (GaussianNB), Decision tree (DecisionTreeClassifier), Percepitron (MLP Classifier), and Ensembles (Bagging Classifier). These models were tested using model-selection.cross-validate using 'balanced-accuracy' scoring and default hyper-parameters.

4 Results

Our results varied with 5/8 models scoring at $0.17 \pm .4$, two models with greater than 0.975 scores, and one outlier with an average score 0.40. The large gap between the 5 poor models and the two seemingly perfect models suggests that there was over-fitting, in this case we would then accept the hypothesis that there is no significant correlation between the financial/population statistics and the pride ranking as all other models have scores less than 0.9. In the case the models are not over-fit then both the Decision Tree and Bagging Classifier have efficacy greater than 0.95 and fit the model justly. Additionally, assuming the models are not over-fit the convusion matrices support their efficacy, as even in cases of low accuracy predictions trend towards the center with few outliers.

Best	Avg	Std Dev	Model
0.60	0.40222	0.13051	Linear, RidgeClassifier
0.30	0.17889	0.04732	Support Vector, SVC
0.33	0.16556	0.11933	Stochastic Gradient, SGDClassifier
0.60	0.22889	0.19734	Neighbors, KneighborsClassifier
0.50	0.21667	0.16997	Naive Bayes, GaussianNB
1.00	0.98667	0.04989	Decision Tree, DecisionTreeClassifier
0.33	0.13333	0.09981	Percepitron, MLPClassifier
1.00	0.97556	0.06262	Ensembles, BaggingClassifier

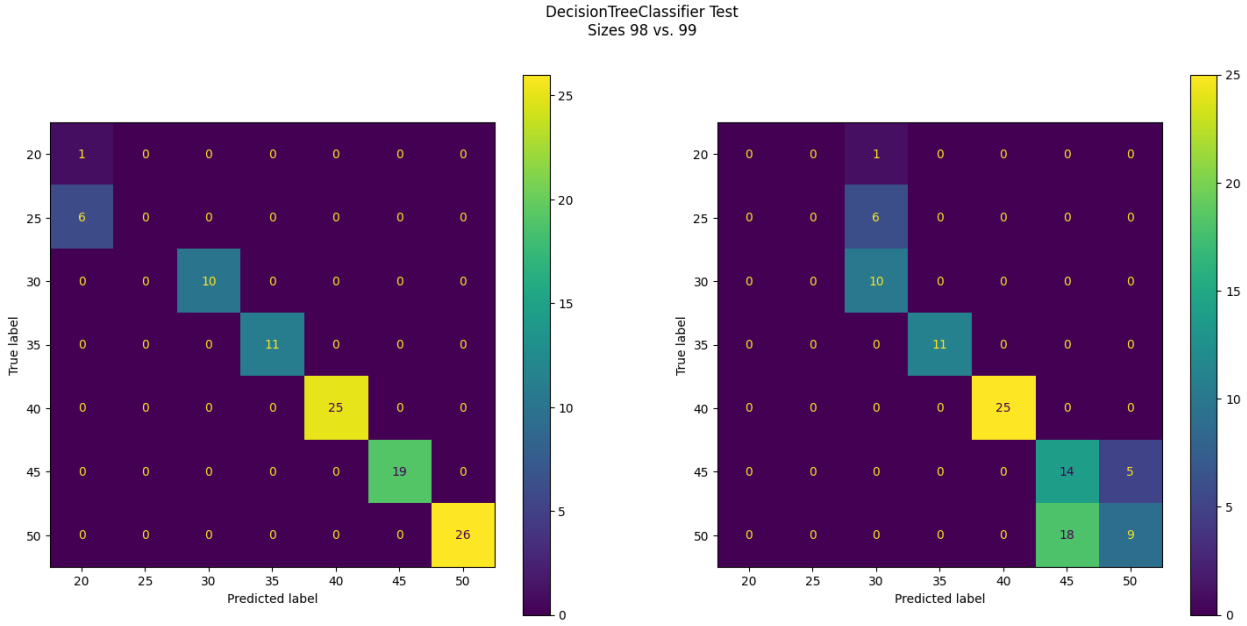

Figure 1: Despite the low standard deviation, when testing the DecisionTreeClassifier results could vary greatly depending on the test size; with a test size of 98 giving 91.84% accuracy, and a size of 99 giving 72.73% accuracy. Nonetheless, the incorrect results trend along the diagonal.

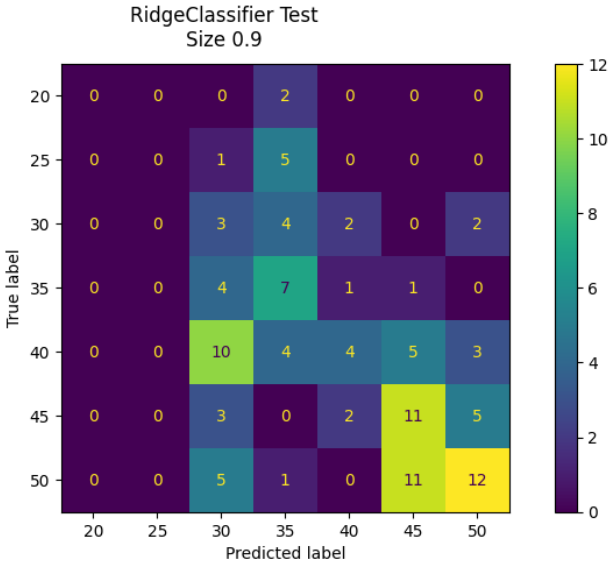

Figure 2: The Ridge Classifier had an average score of 0.40, and an accuracy of 34.26% with a test size of 90%. This model struggles especially with ratings of 30 and 35, having a wide spread of predictions. The overall grouping supports the above average but insignifant score.

References

[Arv] Joakim Arvidsson. *Campus Pride Index - Kaggle*. URL: https://www.kaggle.com/datasets/joebeachcapital/campus-pride-index. (accessed: 06.11.2024).

[Bur] U.S. Census Bureau. *US Census Housing: ACSST5Y2020.S1901*. URL: https://data.census.gov/. (accessed: 07.22.2024).

[] *scikit-learn docs*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html. (accessed: 07.22.2024).

[] *scikit-learn docs*. URL: https://scikit-learn.org/stable/supervised_learning.html. (accessed: 07.22.2024).

[] *seaborn docs*. URL: https://seaborn.pydata.org/generated/seaborn.pairplot.html. (accessed: 07.22.2024).

[Sto] Michael Stoll. *ml-algorithm-selection GitHub*. URL: https://github.com/COSC5557/ml-algorithm-selection-arrangedcupid0. (accessed: 07.22.2024).

[] *UnitedStatesZipcodes.org*. URL: https://www.unitedstateszipcodes.org/zip-code-database/population/. (accessed: 06.11.2024).

[Wol] Milana M. Wolff. *ml-algorithm-selection GitHub*. URL: https://github.com/COSC5557/ml-algorithm-selection-mwolff2021-1. (accessed: 07.22.2024).