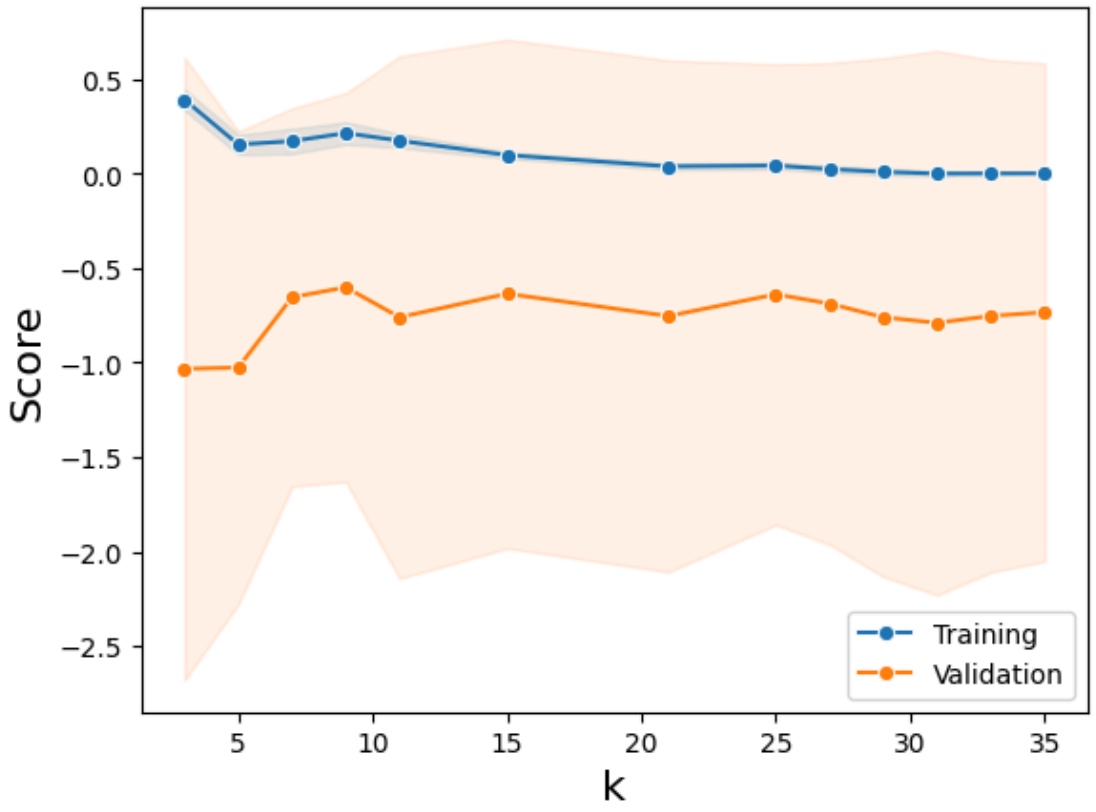
Project: Applied Machine Learning with scikit-learn

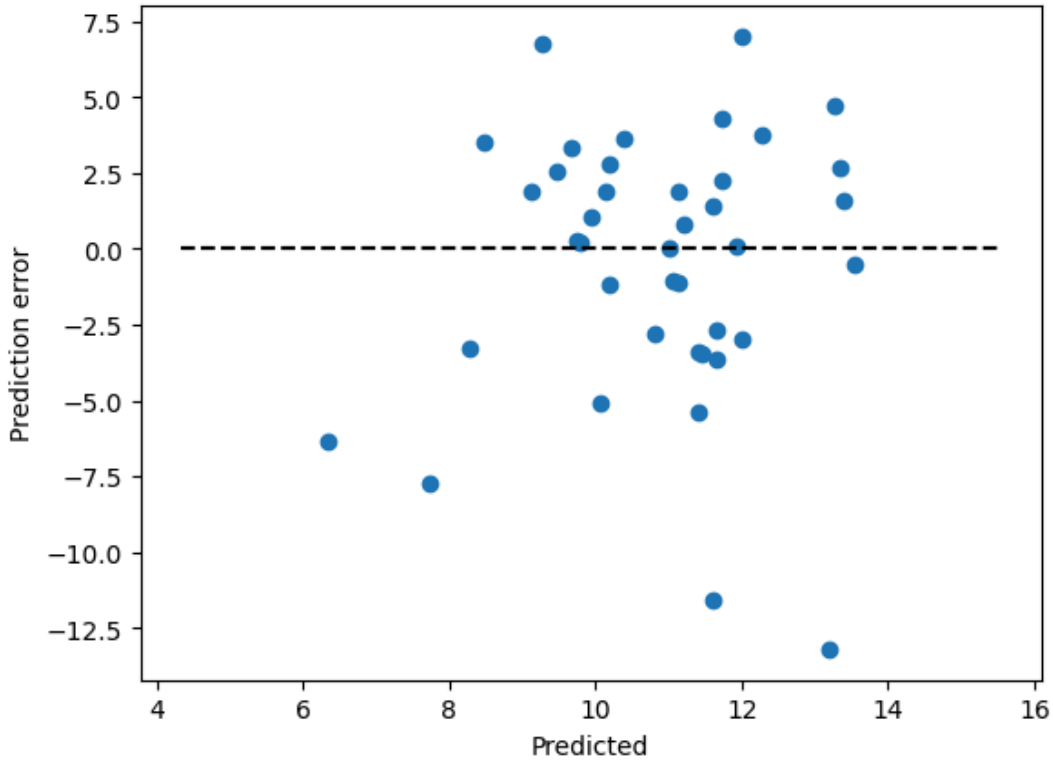
Group Members: Nathaniel Madrigal and Alexander Madrigal

Student Performance Dataset:

The student performance dataset contained many numerical and categorical features to gauge the performance of a student. The predicted feature was the student's final score in their grade. Regression was chosen for this dataset to use the features to predict the student's performance. K-nearest neighbor and Elastic Net models were chosen. Linear regression was not chosen as non-linear relationships may exist with the large number of features.

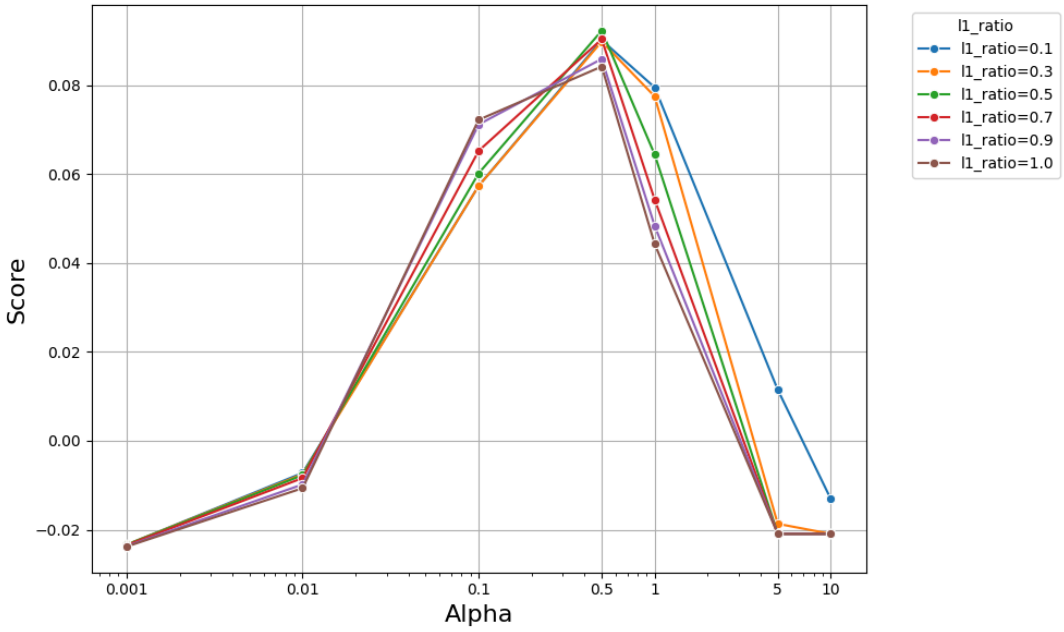
K-nearest neighbors validation and metrics:

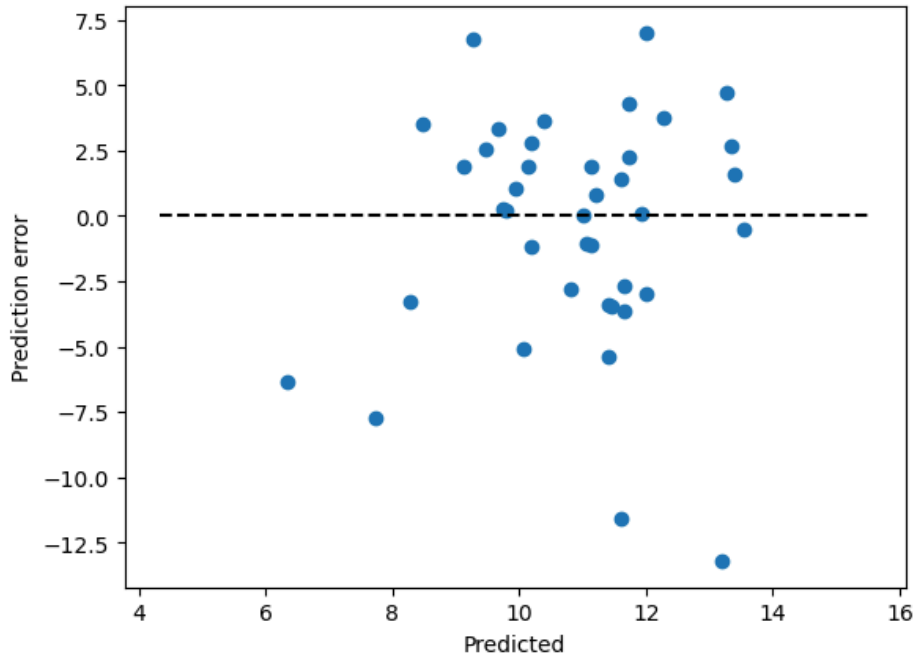




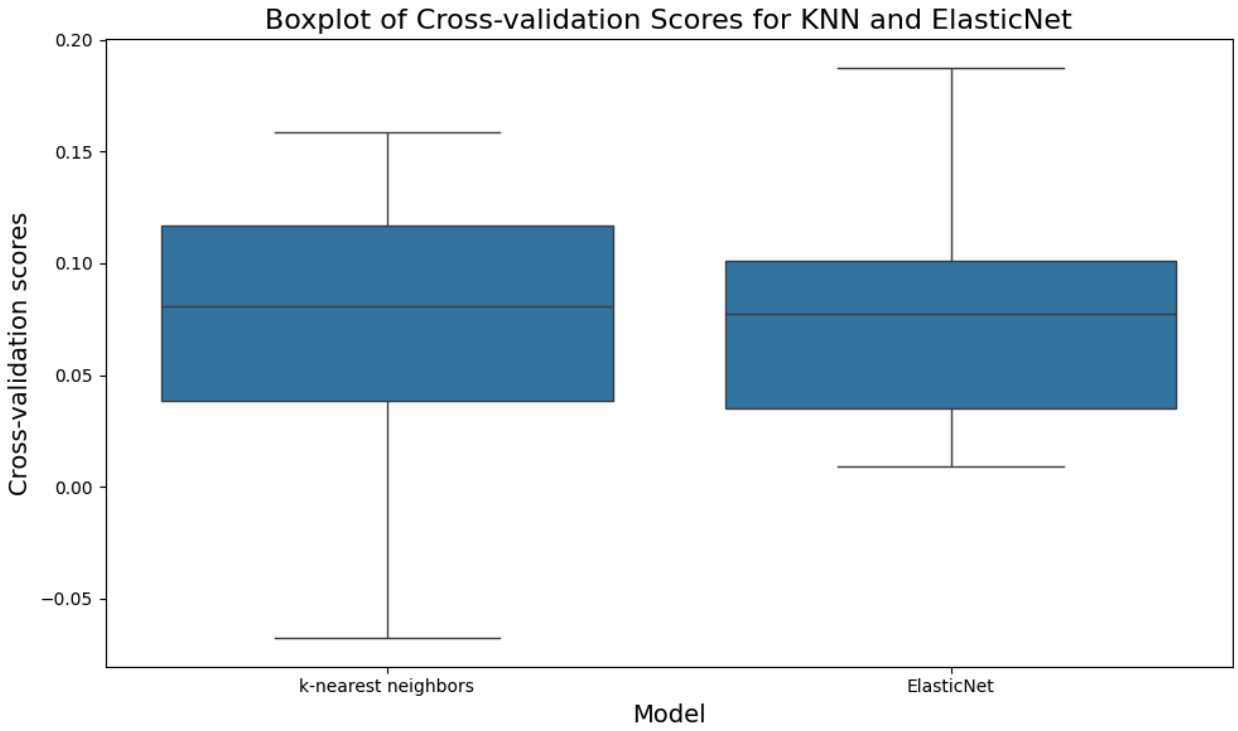
MAE: 3.343
MSE: 19.204
RMSE: 4.382
R-squared: 0.139

Elastic net metrics:





MAE: 3.209
MSE: 17.498
RMSE: 4.183
R-squared: 0.215



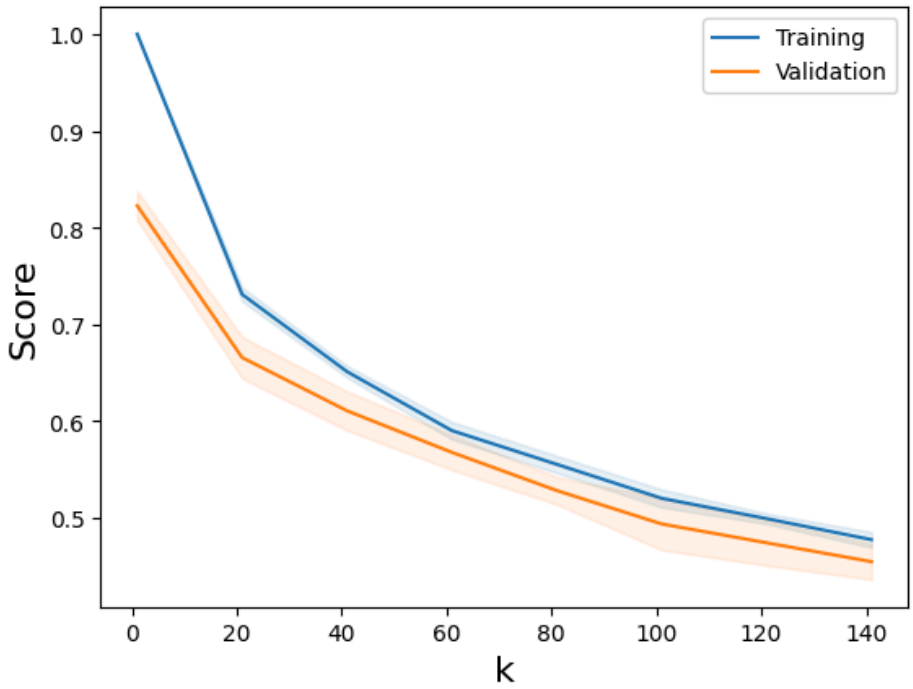
Student Performance Regression Models Discussion:

A struggle with the student performance features is that the large number of numerical and categorical features introduce a lot of noise between input features and output features. Also, the features themselves may not have strong correlation with the output feature. A limitation of the k-nearest neighbor model and the elastic net model is that they do not consider the complex relationships between the input features that may exist, as well as not ignoring the noisy or outlier data. The real world relevance of this model is extremely important. The dataset says that while predicting the performance of students is difficult, it is the most useful application of this data. In trying to predict the performance of students using demographic, socioeconomic, and other important student features, this can provide insight into what features lead to an increase in student performance and likewise which features decrease student performance. This type of model can be used to target groups of students having certain features getting the assistance they need to perform better in school.

Letter Recognition Dataset:

The letter recognition dataset had a total of 20,000 instances and 16 attributes. The output feature was a categorical attribute being one of the 26 capital English letters. The 16 input features were all numerical features containing various information such as height and width in pixels from black and white displays. The goal of this dataset is to create classification models which can predict from the display attributes, the correct capital letter. The three classification models

chosen for this project were: k nearest neighbor classification, gaussian naive bayes, and quadratic discriminant analysis. Both KNN and QDA perform best on well separated and clustered datasets while GaussianNB would work well under the assumption that attributes are normally distributed.



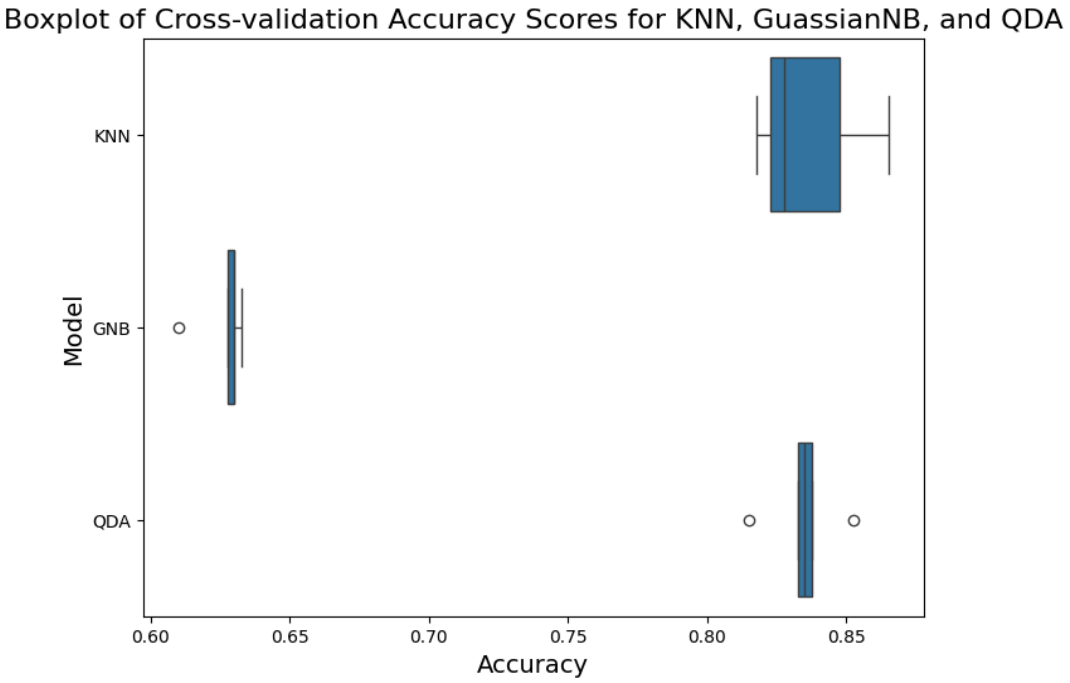
Letter Recognition Model Training and Validation:

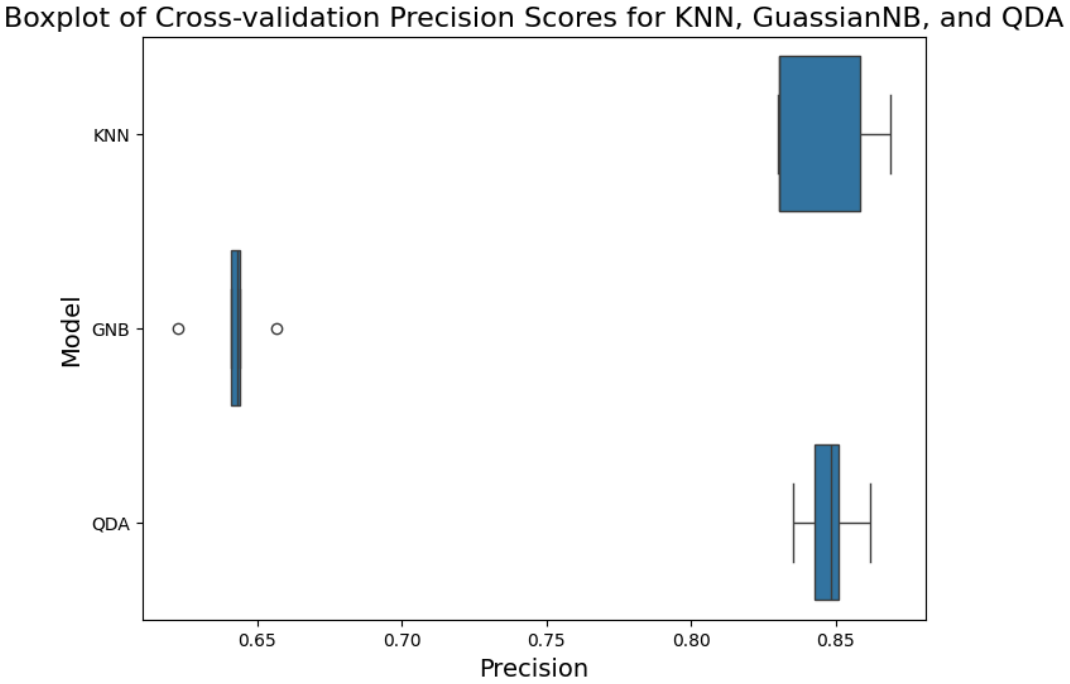
To train the three classification models, a 80-10-10 training/validation/test split was made on the dataset. The dataset was standardized on the training dataset. For KNN, a validation curve was used on the validation dataset to determine that the hyperparameter k=1 is the best solution for the model. After fitting all models, cross validation scores were calculated from the combination of the training and validation sets. Using 5 cross folds, all three models showed consistent results regardless of the fold. No model appears to be largely overfit from the training data.

Letter Recognition Model Metrics:

The accuracy and precision scores for all three models were determined from cross validation on the testing dataset. KNN and QDA performed similarly with higher scores in both accuracy and precision while GNB had both the worst scores by a large margin on both metrics. QDA had the highest mean accuracy and precision of 0.834 and 0.848 respectively, while also having the smallest IQR. Because of the accuracy and consistency of this model, QDA was the best model at classifying letters from the input features.

	K Nearest Neighbors	Gaussian Naive Bayes	Quadratic Discriminant Analysis
Mean Accuracy	0.836	0.626	0.834
Mean Precision	0.844	0.642	0.848





Letter Recognition Discussion:

Through training and evaluating the three classification models, the quadratic discriminant analysis model provided promising results in determining letters from input features on rectangular pixel displays. Our model outperformed the baseline logistic regression model which had an accuracy and precision of 75.760 and 75.555 respectively (Slate). However, our QDA model could not compete with neural network classification which had accuracy and precision scores of 91.980 and 92.171 respectively. This demonstrates that while our dataset had strong clustering and segmentation to facilitate QDA, there are complex connections between input features that can help determine our output feature. A k nearest neighbors approach through binary neural networks could achieve as high an accuracy of 99.7 (Hodge). Further research into neural networks and their connections to KNN models can result in better performing models at the cost of memory and time needed to train and build the models.

Works Cited:

Cortez, Paulo. "Student Performance." UCI Machine Learning Repository, 2008, https://doi.org/10.24432/C5TG7T.

Hodge, Victoria J. et al. “A high performance k-NN approach using binary neural networks.” *Neural networks : the official journal of the International Neural Network Society* 17 3 (2004): 441-58 .

Helwig, Nathaniel E.. “Adding bias to reduce variance in psychological results: A tutorial on penalized regression.” (2017).

Slate, David. "Letter Recognition." UCI Machine Learning Repository, 1991, https://doi.org/10.24432/C5ZP40.