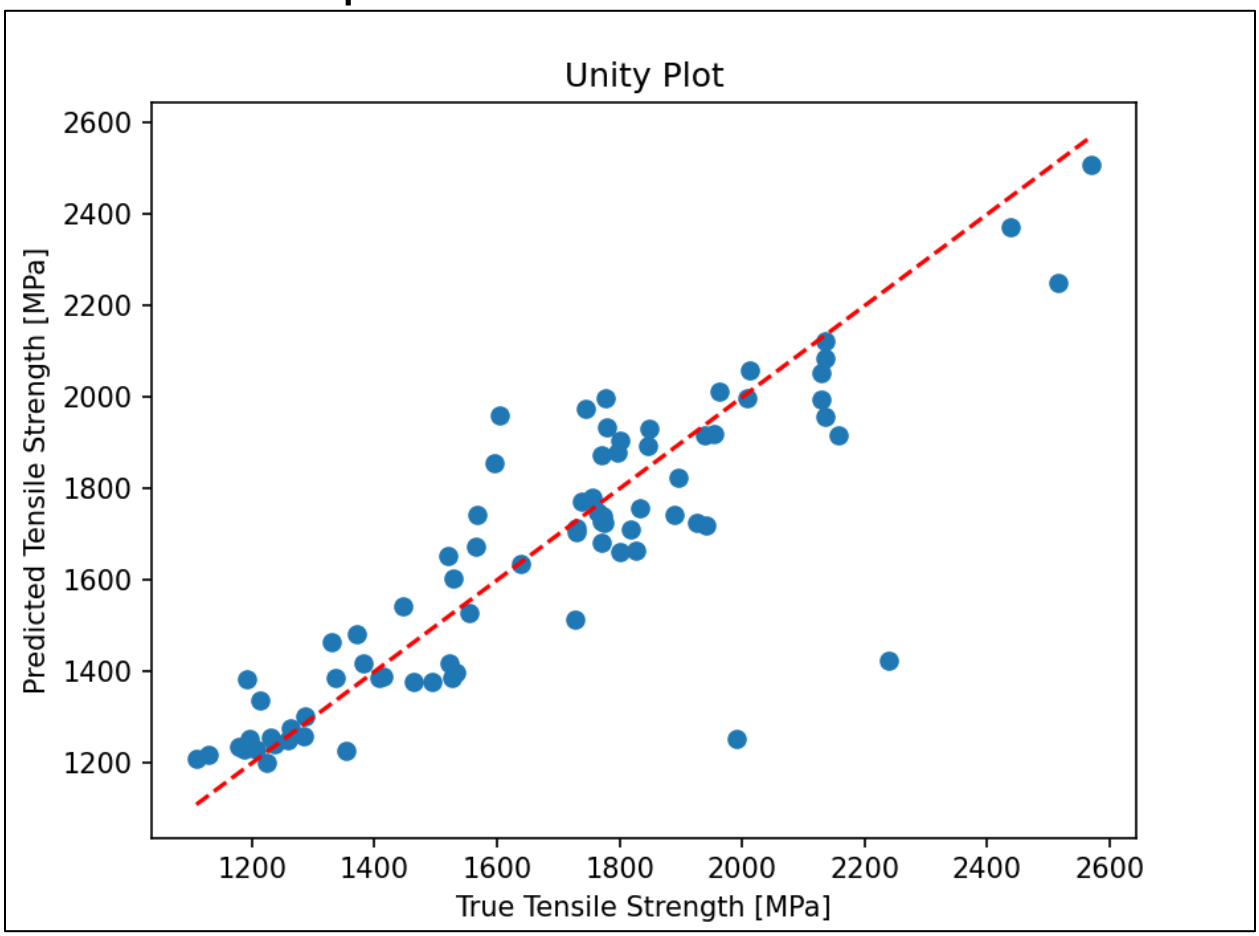
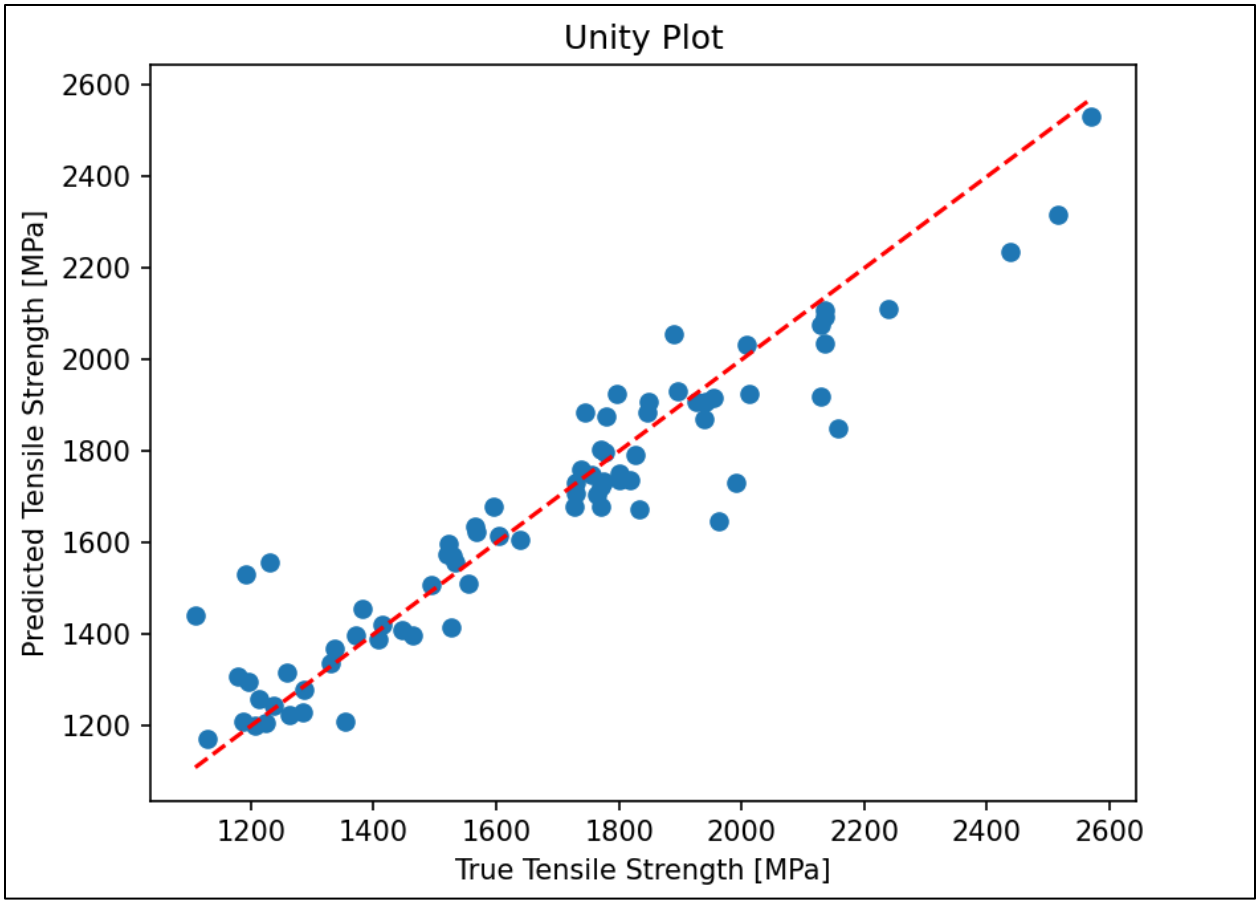
A5: SVR Interpretation



Unity Plot (not normalized)



Unity Plot (normalized)

Data is NOT normalized:

RMSE on the training set: 93.53167104564383

RMSE on the test set: 170.97988255842617

R-squared on the test set: 0.7445366176469796

Date is normalized using StandardScaler:

RMSE on the training set: 94.67656450510049

RMSE on the test set: 114.63020882418492

R-squared on the test set: 0.885174909539825

1. Best Hyperparameter Values: (Best C: 1000 & Best Gamma: 0.1)

Significance of the parameters:

Parameter C: The value of C in an SVR model controls the trade-off between achieving a low training error and a low testing error. A smaller C makes the model more tolerant of errors on the training set (allowing some slack), which might be useful for handling outliers. Conversely, a larger C makes the model less tolerant of training errors and might result in a narrower margin.

Parameter Gamma: The gamma parameter determines how much influence a single training example has. Smaller values of gamma make the RBF kernel relatively wide (each example has a more global influence), while larger values of gamma make the kernel relatively narrow (each example has a more local influence). Narrow kernels can lead to more complex models that are sensitive to individual data points.

Based on the best hyperparameter values obtained (Best C: 1000 & Best Gamma: 0.1), the model's characteristics can be inferred. With a value of C = 1000, it suggests a relatively low tolerance for outliers. High C values indicate that the model is less tolerant of training errors and aims to fit the training data more closely, which may result in a narrow margin hyperplane. In the context of Support Vector Regression (SVR), this means the model would be less tolerant of outliers, potentially leading to sensitivity to individual data points. Additionally, with a value of Gamma = 0.1, the RBF kernel is relatively wide compared to other values in the search grid. A smaller gamma value implies that each training example has a more global influence on the model, making the kernel relatively wide. In SVR with an RBF kernel, a wider kernel can result in smoother and more generalized predictions.

2. Trusting the Learned Model for Novel Predictions:

The trustworthiness of the learned model in predicting the tensile strength of novel material compositions hinges on various factors. The model's performance is directly influenced by the chosen hyperparameters, such as Best C: 1000 and Best Gamma: 0.1, which suggest that it is relatively intolerant of outliers and performs best within the range of training data. This is the reason that I decided to normalize the data as well and test both outcomes. However, the model's reliability also depends on the quality and representativeness of the training dataset; if it encompasses a broad spectrum of steel alloy compositions, the model is more likely to make accurate predictions. Conversely, if the dataset is limited or biased, its reliability diminishes. Additionally, the model's trustworthiness is contingent on the extent to which the chemical composition alone determines tensile strength. Factors not considered in the dataset, like microstructure or manufacturing processes, could impact the model's effectiveness. In conclusion, trust in the model's predictions for novel material compositions varies with hyperparameters, data quality, and the complexity of factors influencing tensile strength, making it advisable to exercise caution and potentially seek additional validation when dealing with materials substantially different from the training data or when dealing with outliers.

3. Number of Support Vectors:

The initial model with an epsilon of 50 had 135 support vectors. Now, if we were to re-run the code with an epsilon of 100, the number of support vectors is 90. The epsilon parameter in the SVR model defines the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value.

When you increase the epsilon from 50 to 100, you're increasing the width of this tube. This means that the model becomes more tolerant of errors up to 100 units away from the true value, as opposed to 50 units in the previous model. As a result, fewer support vectors are needed to define the decision boundary, hence the decrease from 135 to 90 support vectors.

In other words, a larger epsilon creates a larger margin of tolerance for errors, which can result in fewer support vectors as more points fall within this margin. Conversely, a smaller epsilon results in a smaller margin and potentially more support vectors.

This is consistent with the principle of SVMs, where the decision function is entirely determined by a subset of the training samples - the support vectors. By increasing epsilon, you're effectively simplifying the model, which is reflected in the reduced number of support vectors.