**Random Forest Regression:**

Performed on Target Sugar Columns**:** Sucrose, Raffinose

Hyperparameters: n\_estimators : [90, 120, 150, 200]

MSE on Validation data set:

# for estimators 90, 120, 150, 200: 'mean\_test\_score': array([-0.08940205, -0.08714979, -0.08025755, -0.08243478])  
# Best estimator score : -0.08025755   
# n\_estimators=150

The best score does not go better than 0.08 approx.

Considering only Sucrose as Target Variable:

Hyperparameters: n\_estimators : [90, 120, 150, 200]

MSE on Validation data set:

# Considering only Sucrose as a Target Variable, Results are:  
# 'mean\_test\_score': array([-0.10467369, -0.10876152, -0.10717806, -0.11746728])  
# for n\_estimators=90  
# clf.best\_score\_: -0.10467369338788483

Considering only Raffinose as Target Variable:

Hyperparameters: n\_estimators : [90, 120, 150, 200]

MSE on Validation data set:

# Considering Only Raffinose as a Target Variable, Results are:  
# 'mean\_test\_score': array([-0.05895804, -0.05314805, -0.05405331, -0.05012744])  
# clf.best\_params\_: {'n\_estimators': 200}  
# clf.best\_score\_: -0.050127440774539636

MSE on Training data set:

# Considering Only Raffinose as a Target Variable, Results are:  
# 'mean\_test\_score': array([-0.05895804, -0.05314805, -0.05405331, -0.05012744])  
# clf.best\_score\_: -0.050127440774539636  
# clf.best\_params\_: {'n\_estimators': 200}

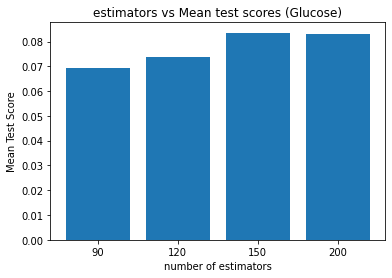
**Observations:**

* MSE value (i.e., 0.1 approx.) is higher when considered a single Target variable as Sucrose. Hence not a good fit as Target variable.
* MSE value on Training & Validated dataset (i.e., 0.05 approx.) is reasonably good when considered a single Target variable as Raffinose. Hence, an acceptable target variable. (Since it suggests that Random forest algorithm for our dataset is neither overfitting nor underfitting the Test instances)
* Target Sugar Columns**:** Sucrose, Raffinose Combined together could also be considered as target variables. Since the MSE value on Validated dataset is 0.08 approx. close to that of considering Raffinose as Single Target Variable.

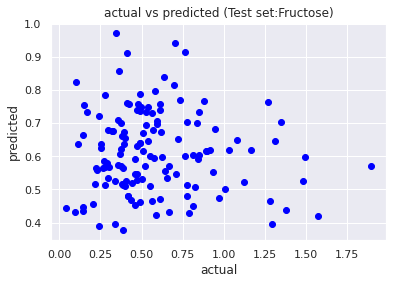
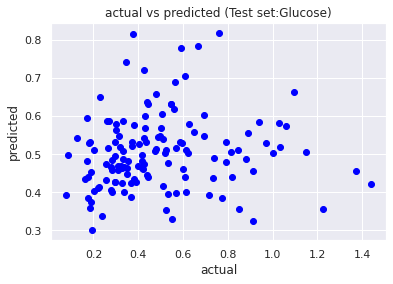
**Choose the best parameters using GridSearchCV:**

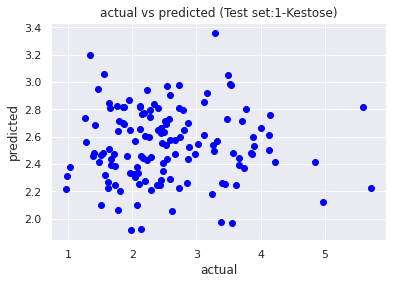
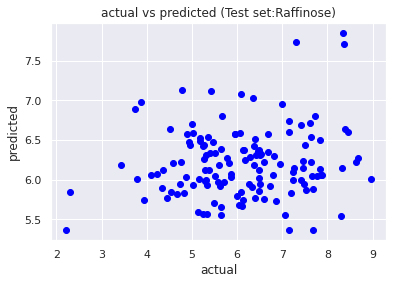
**Considering Single Target Glucose:**

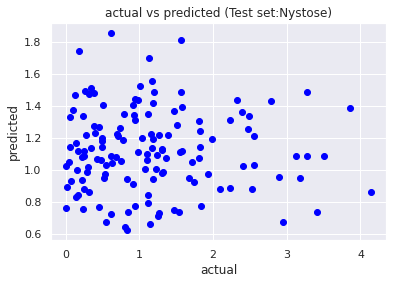
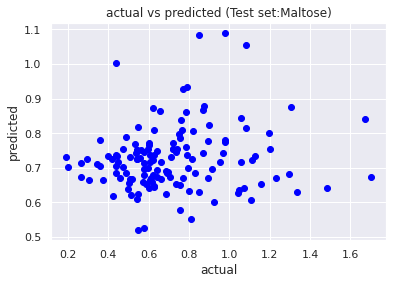
Here the best score 0.0692 is found for number of estimators = 90

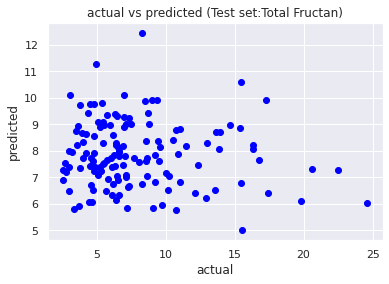
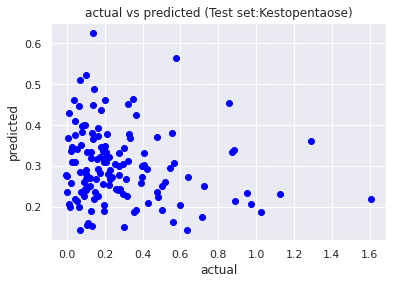


**Using Estimator = 90 :**









For scaled attributes & Take estimator = 90 as parameter for Random Forest Algorithm:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target Sugars | R2 Train set | R2 Test set | MSE Train set | MSE Test set |
| Glucose | 0.8426 | -0.0662 | 0.0138 | 0.0764 |
| Fructose | 0.8392 | -0.1896 | 0.0222 | 0.1344 |
| Sucrose | 0.8415 | -0.1511 | 0.8602 | 6.0531 |
| Raffinose | 0.8509 | -0.0164 | 0.2979 | 1.7000 |
| 1-Kestose | 0.8411 | -0.0793 | 0.1170 | 0.8850 |
| Maltose | 0.8603 | -0.0229 | 0.0112 | 0.0766 |
| Nystose | 0.8446 | -0.0929 | 0.1125 | 0.9120 |
| Kestopentaose | 0.8437 | -0.2078 | 0.0108 | 0.0927 |
| Total Fructan | 0.8351 | -0.1436 | 2.8616 | 21.0892 |

**Ridge Regression vs Linear Regression:**

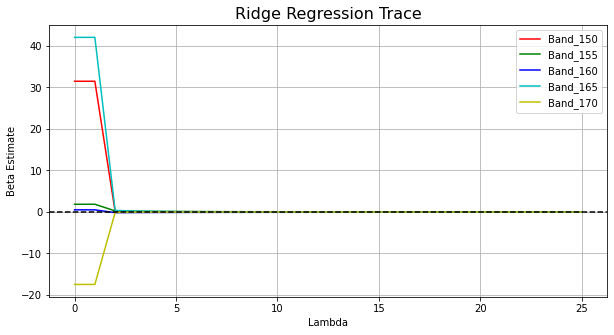
1. For an ordinary least square method (OLS), If the number of parameters is high (let us say for e.g. 10), we would need equal number of observations/data points (i.e., 10 respectively for 10 parameters). OLS has high variance while trying to fit the test data points to the curve/function.
2. It tries to only minimize the “sum of the least square residuals”.
3. Hence, we go for Ridge regression in such scenarios where there is not enough data.
4. Introduces a slight bias in order to reduce the variance.
5. It tries to minimize the “sum of the least square residuals as well as the penalty”. (lambda\*slope^2)

**Applying Ridge Regression to our data:**

**Taking Raffinose as Target Sugar:**

As the model complexity changes, the beta estimates decrease and converge to almost 0. However, none of the attributes i.e., 5 bands (Band\_150, Band\_155, Band\_160, Band\_165, Band\_170) are not playing a major role in finding the target sugar (Raffinose) value.

A similar resultant observation is made for another set of attributes (i.e., bands namely Band\_200, Band\_205, Band\_208, Band\_210, Band\_215)



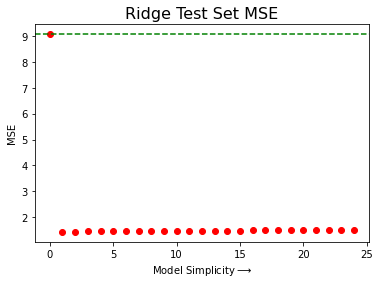
**A picture containing map, sitting, white

Description automatically generated**

The MSE (Mean Square Error) was calculated for different set of lambdas (range: 0-25) and plotted below. It also compares the OLS Regression (Ordinary Least Squares) , Ridge Regression in terms of MSE metric for different lambdas.

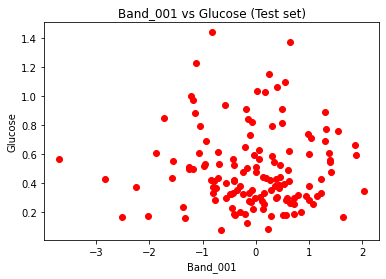
ridge\_reg.intercept\_ : 6.177895592419269

ols\_reg.intercept\_ : 6.206258214766167

****

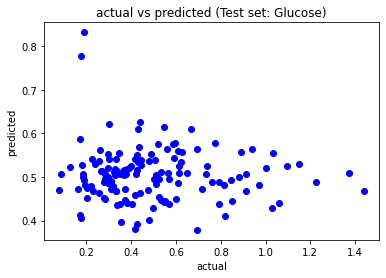
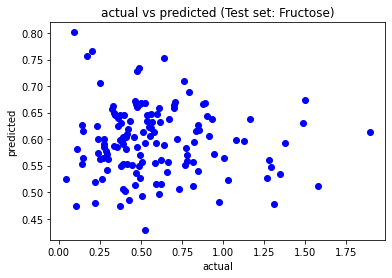
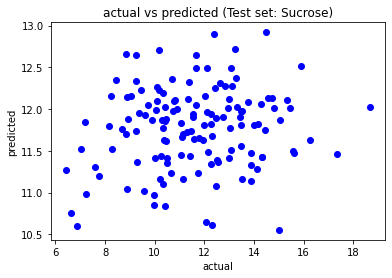
**Correlation b/w Band\_001 & Raffinose Target:**

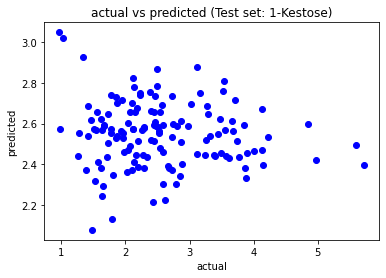
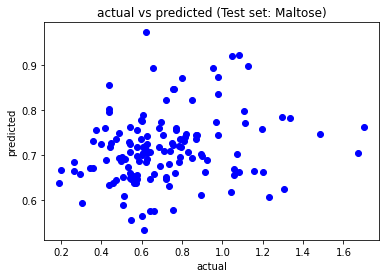
**A screenshot of a cell phone

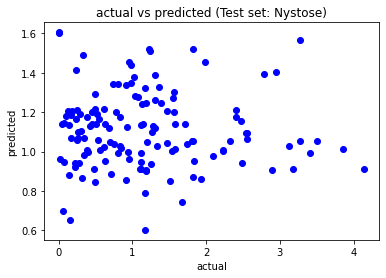
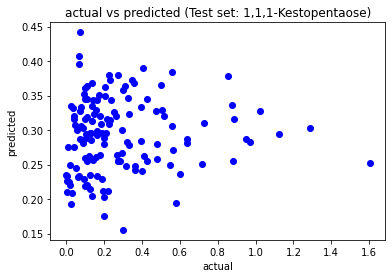
Description automatically generated **

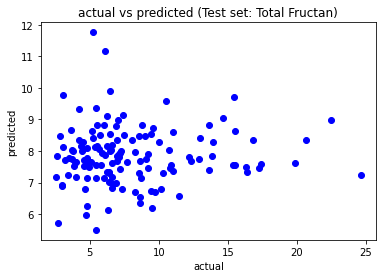
**Using Ridge Regression - Actual vs Predicted values for Target Sugars:**

**A screenshot of a cell phone

Description automatically generated   **

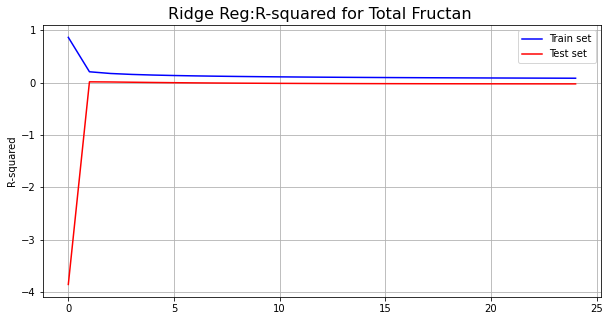
** **

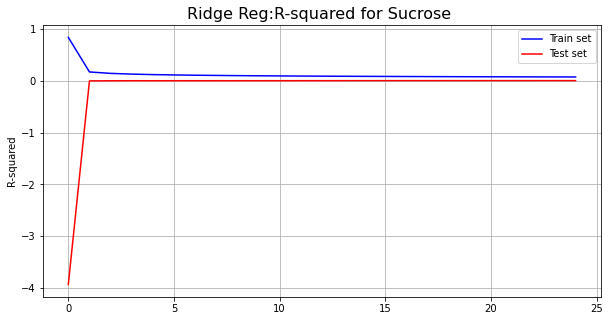
** **

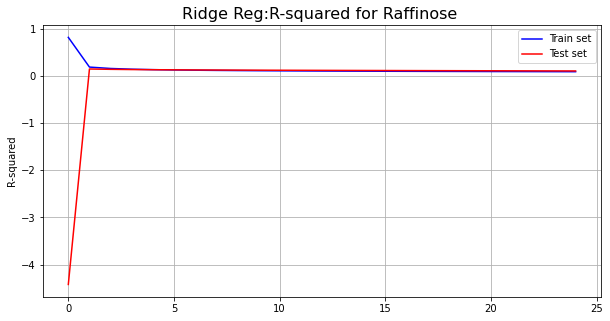
****

**-----------------------------------------------------EOD---------------------------------------------------------------------------**

**Please ignore the below: It needs some correction since all plots are same**

****

****

****