Random Forest Regression

Pre-processing Steps:

* Normalized the image data by addition and division of the differences in the day light intensity.
* Eliminated a few columns such as Print Info, Sample #, Extraction Date, Measurement Date from the sugar values dataset.
* 1,1,1- Kestopentoase had 4 negative sugar readings which were dropped from our dataset as outliers.
* Training Set is taken as 80% of our entire dataset.
* For Stratified sampling, we took 60% of the available total number of samples for each genotype into the training set and the remaining into testing set.
* For some of the genotypes(Sanaa and Calgary), there were not enough samples to perform a stratified sampling( samples where less than 2 which did not allow split) so the stratified sampling was not performed on them.
* As suggested by the workflow and Uwe, the sugar values were normalized by dividing them with the TR\_1 values for the corresponding batch.
* Hyperparameters for GridSearchCV : n\_estimators:{100,300,500,600}

Results Phase 1: Normalized by addition

|  |  |  |  |
| --- | --- | --- | --- |
| **Training on entire dataset** | | | |
| **Sugar** | **R2 Score** | **RMSE** |  |
| Glucose | 0.851498 | 0.113 |  |
| Fructose | 0.8497 | 0.1458 |  |
| Sucrose | 0.8554 | 0.8763 |  |
| Raffinose | 0.8552 | 0.5236 |  |
| 1-Kestose | 0.8509 | 0.33 |  |
| Nystose | 0.845 | 0.338 |  |
| Maltose | 0.8672 | 0.103 |  |
| 1,1,1-Kestopentaose | 0.845 | 0.104 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Training on 80% dataset with 20% test set** | | | |
| **Sugar** | **Training R2 Score** | **Training RMSE** |  |
| Glucose | 0.8436 | 0.112 |  |
| Fructose | 0.8368 | 0.144 |  |
| Sucrose | 0.846 | 0.897 |  |
| Raffinose | 0.849 | 0.547 |  |
| 1-Kestose | 0.848 | 0.338 |  |
| Nystose | 0.846 | 0.329 |  |
| Maltose | 0.865 | 0.102 |  |
| 1,1,1-Kestopentaose | 0.841 | 0.100 |  |

|  |  |
| --- | --- |
| **Testing Score** | |
| **Sugar** | **R2 Score** |
| Glucose | -0.009 |
| Fructose | -0.04 |
| Sucrose | -0.11 |
| Raffinose | -0.09 |
| 1-Kestose | -0.072 |
| Nystose | 0.073 |
| Maltose | -0.101 |
| 1,1,1-Kestopentaose | -0.077 |

|  |  |
| --- | --- |
| **Cross Validation Results with 5 folds** | |
| **Sugar** | **R2 Score** |
| Glucose | -0.204 |
| Fructose | -0.212 |
| Sucrose | -0.105 |
| Raffinose | -0.079 |
| 1-Kestose | -0.163 |
| Nystose | -0.181 |
| Maltose | -0.139 |
| 1,1,1-Kestopentaose | -0.077 |

A close up of a map

Description automatically generated A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated A close up of text on a white background

Description automatically generated

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA close up of a map

Description automatically generated

The results show that the model is overfitting highly as it is not able to generalize properly as when verified by the testing data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Stratified Sampling - Training Set** | | | |
| **Sugar** | **R2 Score** | **RMSE** |  |
| Glucose | 0.851 | 0.111 |  |
| Fructose | 0.854 | 0.135 |  |
| Sucrose | 0.843 | 11.58 |  |
| Raffinose | 0.848 | 0.524 |  |
| 1-Kestose | 0.843 | 0.327 |  |
| Nystose | 0.848 | 0.338 |  |
| Maltose | 0.853 | 0.107 |  |
| 1,1,1-Kestopentaose | 0.85 | 0.099 |  |

Results Phase 2:

Normalization using division and the sugar values normalized as suggested in the workflow.

|  |  |  |  |
| --- | --- | --- | --- |
| **Training on entire dataset** | | | |
| **Sugar** | **R2 Score** | **RMSE** |  |
| Glucose | 0.8385 | 0.180 |  |
| Fructose | 0.837 | 0.197 |  |
| Sucrose | 0.844 | 0.064 |  |
| Raffinose | 0.841 | 0.07 |  |
| 1-Kestose | 0.847 | 0.116 |  |
| Nystose | 0.841 | 0.269 |  |
| Maltose | 0.847 | 0.115 |  |
| 1,1,1-Kestopentaose | 0.844 | 0.321 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Training on 80% training dataset** | | | |
| **Sugar** | **R2 Score** | **RMSE** |  |
| Glucose | 0.831 | 0.181 |  |
| Fructose | 0.837 | 0.195 |  |
| Sucrose | 0.849 | 0.063 |  |
| Raffinose | 0.846 | 0.078 |  |
| 1-Kestose | 0.855 | 0.113 |  |
| Nystose | 0.844 | 0.27 |  |
| Maltose | 0.841 | 0.116 |  |
| 1,1,1-Kestopentaose | 0.839 | 0.332 |  |

|  |  |
| --- | --- |
| **Testing Set** | |
| **Sugar** | **R2 Score** |
| Glucose | -0.103 |
| Fructose | -0.07 |
| Sucrose | -0.009 |
| Raffinose | -0.24 |
| 1-Kestose | -0.13 |
| Nystose | -0.161 |
| Maltose | -0.100 |
| 1,1,1-Kestopentaose | -0.211 |

|  |  |
| --- | --- |
| **Cross Validation Results with 5 folds** | |
| **Sugar** | **R2 Score** |
| Glucose | -0.142 |
| Fructose | -0.183 |
| Sucrose | -0.101 |
| Raffinose | -0.159 |
| 1-Kestose | -0.116 |
| Nystose | -0.110 |
| Maltose | -0.172 |
| 1,1,1-Kestopentaose | -0.151 |

A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA close up of text on a white background

Description automatically generated

These results seem to be better as compared to the previous ones although even these results are not satisfactory.