Brayden Blanchard

EXST 7087 – Digital Agriculture

Final Report

**Introduction**

**Topic and Question of Interest**

Sugarcane (*Saccharum spp.* hybrids) is a clonally propagated, polyploidic crop that displays aneuploidy. This makes genetic studies of sugarcane difficult, but variation in base populations highly achievable (Dumont et al., 2019; Kimbeng et al., 2001; Pontif et al., 2021). A major goal of sugarcane breeding efforts around the world is the application of Marker-Assisted Selection (MAS) or Genomic Selection (GS) approaches to achieve genetic gains in breeding programs(Alwala et al., 2006; Andru et al., 2011). There is recent evidence of the potential benefits of GWAS and other GS techniques used in sugarcane breeding programs (Fickett et al., 2019; Sandhu et al., 2022; Yadav et al., 2020). In Louisiana, previous studies have identified important marker-trait associations (MTA) for 11 sugar yield traits using single nucleotide polymorphisms (SNP) and insertion-deletion (Indel) markers using a modern diverse panel of Saccharum spp. clones. This diversity panel consisted of a significant portion of hybrids from Louisiana sugarcane breeding programs as well as other wild-type, ancestral species, and species originating from different regions. It was planted both in 2015 and 2016 in a randomized complete block design. Common check varieties were randomly planted in each block. Preliminary analysis to account for field variation was performed using these checks, but adjustments for field variation were not seen as necessary. Association between markers and traits was determined using a Q-K analysis as previously described (Yu et al., 2006). Since this study, further effort towards finding adequate prediction models has analyzed ridge regression best linear unbiased prediction (rrBLUP) and different Bayesian approaches (Bayes A, Bayes B, Bayes C, Bayesian Lasso, Bayesian ridge regression) for alternative approaches in predicting genomic estimated breeding values (GEBVs) for GS. This study acknowledged the fact that different prediction models may be adequate for different traits of interest, making it necessary to test and compare multiple approaches for each trait. Including large effect QTLs (identified in GWAS) into the GS models will contribute to their increased accuracy (Satpathy et al., 2022). Further evaluation of prediction models can be attempted in GWAS to adequately account for variation due to blocking factors including plant years, replicates, locations, ratoon crops, and genotypes in order to predict phenotypic performance using SNP and Indel markers. Certain machine learning techniques provide the ability to deal with abundant variation and blocking factors, similar to that experienced in the diversity panel study for the GWAS previously cited (Fickett et al., 2019). This is necessary due to the nature of these genetic studies in sugarcane with potential for confounding effects. The usefulness of machine learning techniques is apparent in multiple applications of sugarcane breeding, including selection (Zhou et al., 2011). The objective of this study is to utilize machine learning techniques such as a random forest model and neural network to examine the accuracy that can be acquired by such techniques in predicting phenotypic performance with SNP and Indel markers while accounting for other variation in a diversity panel of modern sugarcane hybrids. Future evaluation can identify important markers that consistently appear to have a significant effect in multiple different models and approaches to validate the use of those important QTL.

**Dataset Description**

Both phenotypic and genotypic data are provided for 107 clones in this study. Phenotypic traits measured as factors of sucrose content were measured using near-infrared (NIR) technology. Among these variables measured were Brix, Theoretical Recoverable Sucrose (TRS), Fiber, Moisture %, Juice % Sucrose (POL), and Purity. For the purposes of this study, predictions were made for TRS, Brix, and Fiber. Important cane yield traits were also measured. These included Population, Stalk Height, Stalk Weight, and Stalk Diameter. Again, for this study, predictions were made for Bundle Weight (weight of ten stalks) and Stalk Diameter. The experiment was planted in two different years, replicated by location on light and heavy soil, and data was taken for 4 years from plant-cane crops to third ratoon. This provided over 1200 unique observations of phenotypic data. Diameter was only able to be measured in certain settings, causing that subset to be just over 300 observations. The models were adjusted for each individual trait to account for both the different number of observations in the datasets as well as the unique nature of the trait data observed. Blocking factors for potential variation include field row, field column, PlantYear, Crop, and Location. These factors were included as input variables for the models. 5815 SNP and Indel markers were also used as inputs. A separate dataset was created for each trait of interest with the blocking factors and markers as input and the trait of interest as an output. From here, each dataset would be split into training and testing sets for cross validation. Tables for dataset descriptions and summary statistics are provided (Table 1-4).

**Model/Technique**

For typical genetic marker studies, multiple models are used to correlate DNA markers with phenotypic traits. Linear regression is the most common method, but typically requires certain assumptions to be met such as homogenous variances, normally distributed residuals, and that the relationship of the dependent and independent variables is linear. These assumptions are often violated in plant breeding studies due to the nature of data collection and the necessary experimental designs. This provides the need for more robust methods for prediction. Machine learning algorithms can provide this robust approach. Neural networks are comprised of layers and nodes through which input data is passed through and adjusted to predict the following layer and, eventually, the output variable. Parameters can be set to adjust these nodes and layers to refine the prediction and increase the accuracy. The assumptions of linear regression do not need to be met for neural networks or random forest models. Radom forest models are another form of machine learning technique that produces a prediction model based on a network of decision trees to predict an output variable. Random forest models are similar to neural networks in that their parameters, mainly decision trees, can be adjusted to increase the accuracy of the predictions. Cross validation strategies can be employed for all three methods that will assess the efficacy of the models to predict the trait of interest. For each trait prediction, a neural network, linear regression, and random forest model were fit to predict the phenotypic performance. Multiple iterations of each were performed to provide reliability of the results. The mean squared error, root mean squared error, and r-squared of the prediction were reported to evaluate the models’ performance in cross validation. A series of tests were performed with the EarlyStopping function of Tensorflow in python to find an adequate architecture of the neural network to achieve stable MSEs in the predictions. Early stopping was set to patience of 50 where the model would stop running after 50 returned MSEs did not change significantly. This gave an estimate of how many epochs should be set that would achieve stable MSEs. The MSE of each epoch would be saved as an array. For linear regression, the train/test split was applied for 50 iterations to create 50 random train/test split datasets to run the linear model. This provided an array of 50 MSEs achieved with linear regression. The random forest model was trained and iterated 5 times to provide 5 MSEs to compare to the MSEs of the other models. The architecture of these NN and RF were unique to each trait and dataset. A table with the given architecture, iterations, and decision trees used for each trait is provided (Table 5). All Train/Test split procedures for cross validation were carried out with an 80/20 split function. All three models’ MSEs for each trait were statistically compared to each other using Welch's t-test due to an unbalanced number of MSEs to be compared between the three.

**Analysis**

For all three models and for every trait of interest, MSE, RMSE, and R2 were recorded for every iteration of the prediction in cross validation. MSE for each of the three models was compared statistically using Welch's t-test. The models tested performed differently relative to each other for every trait except diameter, however the random forest model generally performed the best for all traits. The best model to predict the trait of interest was selected based on the lowest MSE. For TRS, the random forest model performed statistically the best (p < 0.01) with an average MSE of 285, average RMSE of 16.9, and an average R2 of the predicted values of 0.69. For Brix (a component of TRS that is easily measured in the field), the random forest model performed statistically the best (p < 0.0001) with an average MSE of 0.71, average RMSE of 0.84, and an average R2 of 0.65. Fiber is a very important trait for the processing of sugarcane and can impact the efficiency of sugar recovery in a mill. The random forest model once again performed statistically better (p < 0.001) for Fiber than both the neural network and linear regression with an average MSE of 0.64, average RMSE of 0.80, and average R2 of 0.58. Stalk diameter is a component of cane yield, and in Louisiana, smaller stalk diameters generally correlate with higher stalk populations and often higher yield. This causes a negative correlation between stalk diameter and yield. It is also difficult and laborious to measure stalk diameter in the field. Human error and consequently statistical noise and bias interfere with accurate diameter estimates, even though heritability of stalk diameter is known to be relatively high (Milligan et al., 1990a; Milligan et al., 1990b). In the example of this study, data collection for stalk diameter was limited because of these restraints. For stalk diameter, both the neural network and the random forest performed statistically similar (p > 0.05) while the linear regression model seemed to be a poor fit possibly due to the statistical noise and confounding that may have been affecting the analysis. The neural network performed numerically better than the random forest with an average MSE of 4.17, an average RMSE of 2.04, and an average R2 of -0.008. The random forest achieved an average MSE of 4.79, an average RMSE of 2.19, and an average R2 of 0.085. Lastly, bundle weight (the weight of a ten-stalk bundle), is used to calculate average stalk weight which is related to stalk diameter and a component of cane yield. Similar to stalk diameter, it is often negatively correlated with cane yield because smaller stalks usually accompanies higher stalk populations and higher cane yield. Once again, the random forest model predicted bundle weight the best with an average MSE of 5.22, an average RMSE of 2.29, and an average R2 of 0.58. Figure 1 displays each of the three models fit for predicting each of the five traits, their reported MSEs from the cross validation, and the statistical difference between the models’ MSEs. Statistically, the random forest technique was the most effective at predicting the traits of interest. Figure 2 is a similar plot with reported R2 for prediction accuracy using the testing splits of the datasets.

**Conclusion**

The process of evaluating prediction models for predicting phenotypic traits with DNA marker data is important for the use of selection in breeding programs. The nature of data collection in breeding populations makes modeling difficult and subject to many factors. Machine learning techniques provide a much-needed robustness to improve prediction with phenotypic data collected in selection trials. It is important for a breeder to utilize all the tools he has at his disposal to refine his prediction models and cross validate to increase the reliability of his results and discoveries. In the context of plant breeding, this refers to the certainty of the performance of a new variety, whether that be for early-stage selection to advance through the program or for final release to commercial farmers. As progress is made in the fields of genomics and molecular biology, GWAS and GS techniques will improve and continue to be a useful tool for breeders. Prediction modeling is crucial to the success of these techniques and should continually be tested, modified, and validated as their use and efficacy can vary per the crop, environment, and situation in which they are used. For this study, a random forest model provided much needed simplicity and robustness to predict the performance of varieties for five traits of interest using a diverse panel of Louisiana hybrid clones.

**Tables**

Table 1. Phenotypic Data variables collected from field tests and compiled into a dataframe. This dataset was further subset to contain only one trait of interest, Clone ID, and all blocking factors for the five traits to be predicted.

|  |  |  |
| --- | --- | --- |
| Variables | Type | Description |
| Clone ID | Chr (factor) | Variety name |
| Location | Chr (factor) | Light or Heavy sites (used as replication and block) |
| Crop | Num (factor) | Age of crop (plantcane, first, second, or third ratoon as 0,1,2,3 respectively) |
| PlantYear | Num (factor) | 2015 or 2016 (test was repeated two years in a row) |
| Brix | Num (trait) | % Soluble solids obtained with NIR |
| TRS | Num (trait) | Theoretical Recoverable Sucrose obtained with NIR |
| Fiber | Num (trait) | Fiber obtained with NIR |
| Bu Wt | Num (trait) | Bundle weight that can be used to calculate average stalk weight (Bu Wt/10) |
| Diam | Num (trait) | Average of three diameter measurements of the 5th node |
| Col | Num (factor) | Designation of field row by first number from “Plot” |
| Row | Num (factor) | Designation of field plot by last two numbers from “Plot” |

Table 2. Summary Statistics sample of sugar yield traits of interest for every clone. Mean values presented here for 107 total clones.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Clone ID | N | Brix | TRS | Fiber | Bundle Weight | Diam |
|  |  |  |  |  |  |  |
| 00-930 | 12 | 19.79 | 242 | 11.8 | 11.95 | 20.69 |
| 00-950 | 12 | 20.84 | 257 | 11.9 | 12.89 | 20.79 |
| 01-040 | 12 | 17.1 | 186 | 14.1 | 13.03 | 23.34 |
| 01-147 | 3 | 18.57 | 217 | 10.8 | 14.23 | 20.81 |
| 01-249 | 4 | 19.45 | 239 | 12.7 | 11.5 | 18.47 |
| 01-283 | 12 | 19.85 | 239 | 10.6 | 11.76 | 21.43 |
| 01-299 | 73 | 19.57 | 242 | 12.7 | 12.73 | 19.31 |
| 01-315 | 12 | 18.88 | 224 | 11.8 | 8.79 | 19 |
| 01-517 | 9 | 19.81 | 246 | 10.8 | 15.43 | 25.48 |
| 01-523 | 12 | 19.68 | 235 | 13 | 10.71 | 19.44 |
| 02-610 | 12 | 19.54 | 242 | 13.1 | 13.79 | 20.63 |
| 02-618 | 12 | 19.22 | 235 | 12 | 9.08 | 17.59 |
| 02-623 | 12 | 19.85 | 251 | 12.2 | 12.27 | 20.53 |
| 04-838 | 12 | 19.67 | 242 | 13.3 | 11.5 | 18.85 |
| 04-847 | 12 | 19.55 | 231 | 11.1 | 16.09 | 21.56 |
| 05-448 | 12 | 19.49 | 236 | 11.8 | 12.48 | 23.3 |
| 05-457 | 12 | 19.95 | 246 | 13 | 10.91 | 19.24 |
| 05-466 | 9 | 19.78 | 241 | 14.4 | 10.54 | 21.01 |
| 05-499 | 3 | 19.76 | 233 | 13.9 | 11.95 | 19.33 |
| 05-961 | 12 | 19.45 | 235 | 12.2 | 13.14 | 21.21 |
| 06-001 | 12 | 20.32 | 254 | 11.6 | 15.9 | 23.14 |
| 06-038 | 12 | 19.83 | 240 | 11.5 | 14.78 | 24.24 |
| 06-040 | 12 | 19.85 | 243 | 12.9 | 13.63 | 21.88 |
| 06-530 | 12 | 19 | 213 | 11.6 | 13.48 | 21.77 |
| 06-563 | 12 | 19.79 | 243 | 12.7 | 14.62 | 20.52 |
| 07-057 | 12 | 19.19 | 233 | 13.2 | 10.72 | 18.56 |
| 07-613 | 12 | 19.76 | 245 | 11.1 | 13.19 | 22.09 |
| 07-617 | 12 | 20.17 | 249 | 11.2 | 10.73 | 20.95 |
| 08-088 | 19 | 19.46 | 234 | 12.2 | 11.71 | 19.3 |
| 08-090 | 19 | 20.05 | 248 | 12.4 | 12.64 | 22.05 |
| 08-092 | 12 | 20.26 | 249 | 12.5 | 13.07 | 22.11 |
| 08-709 | 12 | 19.57 | 233 | 12.8 | 10.77 | 19.04 |
| 08-711 | 12 | 18.65 | 221 | 10.3 | 13.22 | 22.75 |
| 08-717 | 12 | 18.67 | 228 | 12.8 | 10.91 | 19.95 |
| 08-723 | 8 | 18.79 | 216 | 11.9 | 11.13 | 19.34 |
| 08-726 | 12 | 20.84 | 264 | 11.1 | 13.88 | 20.98 |
| 09-099 | 12 | 18.83 | 227 | 12.9 | 11.91 | 21.8 |
| 09-105 | 12 | 18.87 | 223 | 13 | 12.54 | 19.63 |
| 09-107 | 12 | 19.03 | 228 | 13.5 | 13.5 | 21.51 |
| 09-108 | 12 | 20.15 | 243 | 12.6 | 9.28 | 19.18 |
| 09-112 | 19 | 19.83 | 240 | 12.9 | 16 | 22.33 |
| 09-114 | 12 | 19.91 | 237 | 12.5 | 9.47 | 19.31 |
| 09-117 | 9 | 20.8 | 271 | 10.5 | 8.69 | 19.8 |
| 09-118 | 11 | 18.72 | 206 | 11.7 | 13.23 | 22.02 |
| 09-121 | 12 | 19.25 | 215 | 13.6 | 10.53 | 17.05 |
| 09-123 | 12 | 19.81 | 233 | 13.4 | 13.35 | 19.07 |
| 09-129 | 12 | 19.47 | 237 | 13.5 | 8.81 | 17.38 |
| 09-803 | 9 | 21.18 | 260 | 11.4 | 8.64 | 18.43 |
| 09-810 | 12 | 18.87 | 224 | 11.5 | 11.13 | 19.49 |
| 09-822 | 12 | 19.67 | 242 | 13 | 12.71 | 20.2 |
| 09-824 | 12 | 19.78 | 248 | 14.6 | 12.69 | 21.08 |
| 09-825 | 12 | 19.91 | 248 | 12.3 | 10.82 | 19.3 |
| 09-827 | 12 | 18.68 | 221 | 10.7 | 9.68 | 19.3 |
| 09-831 | 12 | 19.87 | 241 | 11.8 | 14.23 | 21.95 |
| 09-840 | 12 | 19.31 | 231 | 12 | 8.65 | 19.18 |
| 09-841 | 12 | 19.18 | 231 | 11.3 | 11.62 | 21.66 |
| 10-147 | 4 | 20.26 | 253 | 10.9 | 13.53 | 22.11 |
| 11-168 | 5 | 19.18 | 226 | 12.3 | 12.01 | 19.24 |
| 11-172 | 7 | 19.7 | 235 | 12.4 | 13.8 | 20.25 |
| 11-183 | 7 | 20.35 | 248 | 12.4 | 12.69 | 21.22 |
| 11-187 | 7 | 20.85 | 257 | 13.2 | 12 | 18.57 |
| 12-201 | 7 | 19.88 | 254 | 11.2 | 15.31 | 21.46 |
| 12-202 | 7 | 21.01 | 260 | 12.8 | 15.24 | 21.69 |
| 12-218 | 7 | 19.36 | 227 | 14.1 | 15.18 | 20.49 |
| 12-227 | 7 | 20.89 | 251 | 13.3 | 16 | 21.58 |
| 52-068 | 7 | 18.89 | 222 | 12.4 | 12.21 | 21.37 |
| 72-370 | 12 | 19.18 | 233 | 12 | 11.68 | 21.29 |
| 74-383 | 12 | 18.36 | 218 | 11.2 | 13.6 | 22.64 |
| 77-407 | 12 | 20.1 | 238 | 10.9 | 11.74 | 19.66 |
| 79-010 | 12 | 19.15 | 221 | 13.3 | 11.21 | 19.25 |
| 79-318 | 12 | 19.71 | 241 | 13.2 | 13.98 | 21.03 |
| 79-348 | 12 | 18.26 | 209 | 12.3 | 17.49 | 22.61 |
| 81-010 | 12 | 19.07 | 224 | 13 | 17.73 | 23.42 |
| 81-030 | 12 | 20.28 | 250 | 12 | 14.46 | 21.89 |
| 82-089 | 12 | 20.21 | 248 | 11.7 | 14.03 | 21.96 |
| 83-644 | 12 | 20.29 | 248 | 10.9 | 15.12 | 22.07 |
| 85-376 | 12 | 19.8 | 240 | 12.1 | 13.37 | 21.61 |
| 85-380 | 3 | 18.84 | 214 | 12.6 | 13.75 | 18.85 |
| 85-384 | 12 | 19.44 | 236 | 11.4 | 10.43 | 18.84 |
| 85-830 | 9 | 17.55 | 193 | 13 | 14.17 | 22.37 |
| 85-845 | 12 | 19.15 | 229 | 12.9 | 11.77 | 21.98 |
| 86-454 | 12 | 20.05 | 240 | 11 | 14.27 | 25.54 |
| 89-846 | 12 | 19.5 | 243 | 12.2 | 11.1 | 19.85 |
| 89-889 | 12 | 19.91 | 238 | 12.5 | 14.28 | 22.59 |
| 91-552 | 12 | 19.58 | 234 | 14 | 12.29 | 22.66 |
| 92-618 | 12 | 21.04 | 263 | 12.9 | 12.6 | 21.49 |
| 92-624 | 12 | 20.37 | 250 | 12.9 | 11.67 | 21.87 |
| 92-648 | 12 | 19.64 | 238 | 10.9 | 13.77 | 24.4 |
| 94-426 | 12 | 19.6 | 240 | 12.6 | 14.19 | 21.52 |
| 94-428 | 12 | 19.53 | 243 | 11.1 | 14.29 | 23.36 |
| 94-432 | 12 | 20.61 | 254 | 12.3 | 11.99 | 20.96 |
| 95-951 | 12 | 20.26 | 245 | 14.7 | 9.58 | 17.24 |
| 95-988 | 12 | 19.98 | 250 | 11.6 | 14.03 | 21.88 |
| 96-540 | 67 | 19.28 | 233 | 11.6 | 13.37 | 20.77 |
| 96-561 | 12 | 19.69 | 242 | 11.9 | 11.6 | 17.3 |
| 97-128 | 12 | 20.14 | 246 | 12.8 | 14.22 | 21.92 |
| 97-218 | 1 | 18.4 | 223 | 11.9 | 19.45 | NA |
| 97-609 | 12 | 19.96 | 243 | 12 | 14.02 | 20.75 |
| 98-207 | 12 | 20.15 | 247 | 12.9 | 9.23 | 18.94 |
| 98-209 | 12 | 19.35 | 238 | 11.4 | 10.18 | 19.6 |
| 99-226 | 12 | 20.35 | 250 | 12.2 | 17.04 | 22.78 |
| 99-233 | 4 | 18.75 | 218 | 13.6 | 10.59 | 20.84 |
| CP52-68 | 4 | 17.98 | 206 | 12.1 | 12.48 | 24.04 |
| N27 | 12 | 15.945 | 160.5 | 11.15 | 17.64 | 21.16 |
| NCO310 | 12 | 19.61 | 229.5 | 13.1 | 8.45 | 19.895 |
| POJ234 | 12 | 19.37 | 236 | 11.9 | 15.74 | 23.95 |
| NA | 4 | 17.67 | 199 | 9.1 | 14.6 | 17.56 |

Table 3. Marker Data sample for 4 clones out of 109 total.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Clone | SNP1 | SNP2 | SNP3 | SNP4 | SNP5 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | SNP5815 |
| 85-376 | 2 | 2 | 2 | 1 | 0 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 1 |
| 79-318 | 1 | 1 | 1 | 1 | 1 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 1 |
| 85-830 | 1 | 1 | 1 | 1 | 1 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 1 |
| 09-099 | 2 | 2 | 2 | 1 | 0 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 1 |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Table 4. Summary Statistics for all traits of interest. Datasets were created with each of these traits as the output variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TRS | Brix | Fiber | Diameter | Bundle Weight |
|  |  | % | % | mm | lbs |
| count | 1215 | 1215 | 1215 | 313 | 1215 |
| mean | 237.2409 | 19.59386 | 12.25529 | 20.88938 | 12.7397 |
| std | 30.73799 | 1.442216 | 1.271772 | 2.338591 | 3.642599 |
| min | 110.443 | 14.08606 | 8.705757 | 15.08333 | 3.98 |
| 25% | 216.1543 | 18.57154 | 11.35969 | 19.11667 | 10.05 |
| 50% | 236.2906 | 19.60957 | 12.20741 | 21 | 12.55 |
| 75% | 262.2505 | 20.7134 | 13.083 | 22.51 | 15.05 |
| max | 302.112 | 22.7637 | 16.74535 | 28.45333 | 26.15 |

Table 5. Model fitting parameters for each of the models and all five traits of interest.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models |  | Parameters | | | | |
|  |  | TRS | Brix | Fiber | Diameter | Bundle Weight |
| Neural Network | Layers | 6 | 6 | 6 | 3 | 6 |
| Loss Function | MSE | MSE | MSE | MSE | MSE |
| Optimizer | Adam | Adam | Adam | Adam | Adam |
| Epochs | 1000 | 1000 | 1000 | 300 | 300 |
| Batch Size | 60 | 60 | 60 | 130 | 50 |
| Linear Regression | Iterations of Train/Test split | 50 | 50 | 50 | 50 | 50 |
| Random Forest | Iterations of Train/Test split | 5 | 5 | 5 | 5 | 5 |
| Decision Trees | 250 | 250 | 250 | 250 | 250 |
| Random Seed | 42 | 42 | 42 | 42 | 42 |

**Figures**

Chart, bar chart

Description automatically generated

Figure 1. Model average MSE plotted for all three models and by each trait of interest. Colors indicate significant differences between MSE (p < 0.05).

Chart, bar chart

Description automatically generated

Figure 2. Model average R2 for predictions plotted for all three models and by each trait of interest.

**Appendix (Codes)**

Github:

<https://github.com/bablanchard/Digital-Ag-Final-Project>

* Code: BB\_FinalProject\_1.2.py
* Data: BB\_pheno\_1.1.rds ImputedSNPs\_original.csv

References:

Alwala, S., Kimbeng, C., Gravois, K., & Bischoff, K. (2006). TRAP, a new tool for sugarcane breeding: comparison with AFLP and coefficient of parentage. *Journal American Society Sugar Cane Technologists*, *26*, 62-86.

Andru, S., Pan, Y. B., Thongthawee, S., Burner, D. M., & Kimbeng, C. A. (2011). Genetic analysis of the sugarcane (Saccharum spp.) cultivar 'LCP 85-384'. I. Linkage mapping using AFLP, SSR, and TRAP markers. *Theor Appl Genet*, *123*(1), 77-93. <https://doi.org/10.1007/s00122-011-1568-x>

Dumont, T., Thong-Chane, A., Barau, L., Siegmund, B., & Hoarau, J. (2019). Genetic Variabilities and Genetic Gains for Yield Components in Regional Sugarcane Breeding Programmes on Réunion Island. *Sugar Tech*, 1-11.

Fickett, N., Gutierrez, A., Verma, M., Pontif, M., Hale, A., Kimbeng, C., & Baisakh, N. (2019). Genome-wide association mapping identifies markers associated with cane yield components and sucrose traits in the Louisiana sugarcane core collection. *Genomics*, *111*(6), 1794-1801. <https://doi.org/https://doi.org/10.1016/j.ygeno.2018.12.002>

Kimbeng, C. A., Froyland, D., Appo, D., Corcoran, A., & Hetherington, M. (2001). *An appraisal of early generation selection in the Central Queensland sugarcane improvement program* Brisbane.

Milligan, S., Gravois, K., Bischoff, K., & Martin, F. (1990a). Crop effects on genetic relationships among sugarcane traits. *Crop science*, *30*(4), 927-931.

Milligan, S., Gravois, K. A., Bischoff, K. P., & Martin, F. (1990b). Crop effects on broad‐sense heritabilities and genetic variances of sugarcane yield components. *Crop science*, *30*(2), 344-349.

Pontif, M., Kimbeng, C., Bischoff, K., Gravois, K., LaBorde, C., Hawkins, G., Sexton, D., Hoy, J., Baisakh, N., & Wilson, B. (2021). Registration of ‘L 11‐183’sugarcane. *Journal of Plant Registrations*, *15*(3), 447-462.

Sandhu, K., Shiv, A., Kaur, G., Meena, M., Raja, A. K., Krishnapriya, V., Mall, A., Skumar, Singh, P., Singh, J., Hemaprabha, G., Pathak, A., Krishnappa, G., & Kumar, S. (2022). Integrated Approach in Genomic Selection to Accelerate Genetic Gain in Sugarcane. *Plants*, *11*, 2139. <https://doi.org/10.3390/plants11162139>

Satpathy, S., Shahi, D., Blanchard, B., Pontif, M., Gravois, K., Kimbeng, C., Hale, A., Todd, J., Rao, A., & Baisakh, N. (2022). Evaluation of Models for Utilization in Genomic Prediction of Agronomic Traits in the Louisiana Sugarcane Breeding Program. *Agriculture*, *12*(9), 1330. <https://www.mdpi.com/2077-0472/12/9/1330>

Yadav, S., Jackson, P., Wei, X., Ross, E. M., Aitken, K., Deomano, E., Atkin, F., Hayes, B. J., & Voss-Fels, K. P. (2020). Accelerating Genetic Gain in Sugarcane Breeding Using Genomic Selection. *Agronomy*, *10*(4). <https://doi.org/10.3390/agronomy10040585>

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203-208. <https://doi.org/10.1038/ng1702>

Zhou, M. M., Kimbeng, C. A., Tew, T. L., Gravois, K. A., & Pontif, M. J. (2011). Artificial Neural Network Models as a Decision Support Tool for Selection in Sugarcane: A Case Study Using Seedling Populations. *Crop science*, *51*(1), 21-31. <https://doi.org/https://handle.nal.usda.gov/10113/48712> https://dx.doi.org/10.2135/cropsci2010.02.0057