GENOMIC SELECTION

A ML based tool to predict plant phenotype based on genetic differences

Dinuka Jayalath[1], David Nugroho[1], Lakruvini Marasinhe[1], Ashima Khanna[1]

[1]*Technical University of Munich Campus Straubing for Biotechnology and Sustainability*

Technische Universität München

TUMCS

1 INTRODUCTION

- **Genomic selection** (GS) has increased the speed of plant breeding, leading to growing crop yields over the last decade.
- **Machine Learning** (ML) models can predict complex phenotypic traits such as biomass, yield, biotic, and abiotic stress tolerance by how they process linear and non-linear relationships between the genetic disparity and environmental features extracted from the dataset. This can help ease the process of GS.[1]

OBJECTIVE: Develop an ML model to predict the phenotypic responses of individual plants based on the genetic differences of the available plant samples.

2 APPROACH

DATA PREPROCESSING

- Data imputation
- Checking equal samples for genotype & phenotype data
- Additive Encoding for genotype data
- Initial Data visualization

MODEL SELECTION

- LASSO
- Support Vector Regression (SVR)
  1. 'polynomial' kernel
  2. 'rbf' kernel
- Nu SVR
- K neighbors regression (KNR)
- Kernel ridge regression (KRR)
- Tweedie regression (TR)
- Bayesian Ridge (BR)

MODEL EVALUATION AND OPTIMIZATION

- C and alpha regularization
- Feature Selection methods

- Tested the model with unseen data
- Performance measures comparison with leader board test

Fig 1. Workflow adopted for ML optimization and selection

REFERENCES

1. Danilevicz, F., et al., (2022). Plant Genotype to Phenotype Prediction Using Machine Learning. Front. genet., 13.
2. Pedregosa et al., (2011). Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830.
3. Jones et al., (2001). SciPy: Open source scientific tools for Python. Int. J. Comput. Eng. pp. 296-305.
4. Mittag, F., et al., (2015). Influence of feature encoding and choice of classifier on disease risk prediction in genome-wide association studies. PloS one, 10(8), p.e0135832.

3 TRAINING & TESTING

a) DATA VISUALIZATION & DIMENSIONALITY REDUCTION

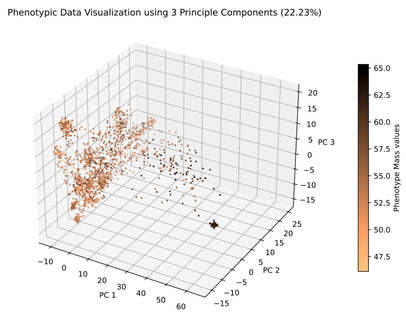

Fig 2. Phenotypical data visualization using three PCs

- PCA was adopted for intuitively visualizing the phenotypical data distribution.
- With the first 3 PCs, the cumulative variance explained ratio was 22% and the phenotypic data seemed to have a sparse relationship (Fig 3).

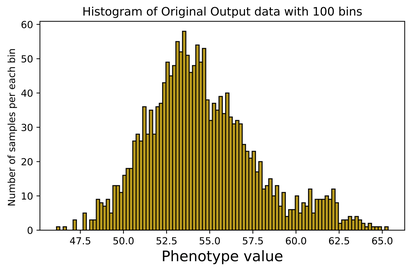

Fig 3. Histogram of Phenotype data with 100 bins

- Mean squared error (MSE) and explained variance (Ex Var) were used as performance measures.
- The distribution was determined to be not normal using Shapiro-Wilk test (P = 4.709e-18)

B) MODEL SELECTION AND PERFORMANCE MEASUREMENT

- Based on the initial 5-fold cross-validation, a few regression models were tested and evaluated using explained variance scores.
- LASSO was chosen for further optimization based on performance metrics.

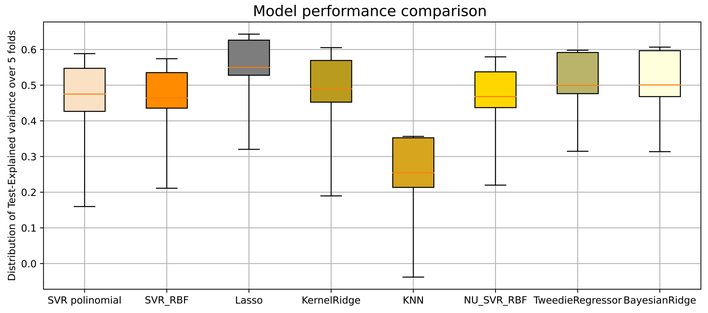

Fig 5: Box plot graph comparing explained variance values of the eight trained models for 5 fold cross validation (whiskers show std over 5 folds)

Table 1: Performance measures for all tested models

	SVR Polynomial	SVR RBF	LASSO	Nu SVR	KRR	KNR	BR	TR
MSE	5.16	5.16	4.33	5.13	4.66	7.75	4.69	4.72
Ex Var	0.44	0.44	0.53	0.45	0.50	0.23	0.50	0.50

4 MODEL OPTIMIZATION

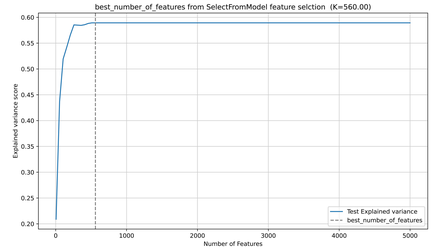

Fig 7. Feature selection using Lasso model

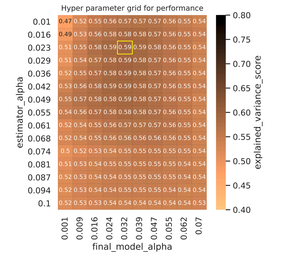

Fig 6. Hyperparameter optimization using nested loops

- Several **feature selection** methods were tested and the **SelectFromModel** method was employed with **LASSO** using the built-in feature selection characteristics of **L1 regularization**.
- **The minimum number of features** required for maximizing performance was systematically analyzed, and **560** features were selected out of **5000** without compromising the model performance (Fig.7).
- **Hyperparameters** were optimized with a **randomized cross-validation grid search** for best-performing models and further optimized with **nested loops** for the selected models and parameters (Fig.6).

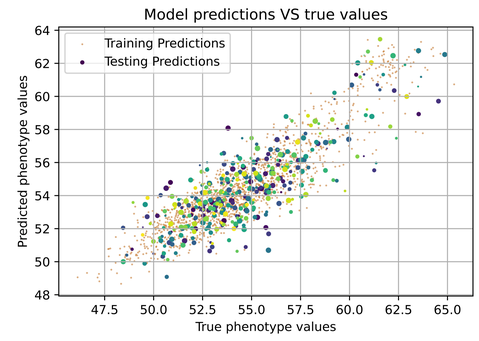

Fig 8: Final model predictions VS true Phenotype values corelation

	MSE	Ex Var
Unoptimized model	4.33	0.53
Leaderboard predictions	3.94	0.65
Optimized model	4.14	0.59
Leaderboard predictions	3.46	0.70

Table 2: Performance measures comparison with the leader board values

- The optimized model generalized for data points within the middle range well, however, could not predict well for those in a higher range.
- Based on the performance metrics of leaderboard predictions, we could conclude that **our optimized model generalized well on unseen data.**

5 FURTHER WORK

- Grid search CV can be used for hyperparameter tuning for complex models like BayesianRidge when computational power and time are not constraints.
- Synthetic data generators can be used to create more samples similar to those of the original dataset.
- Further evaluations for data imbalance might be needed.