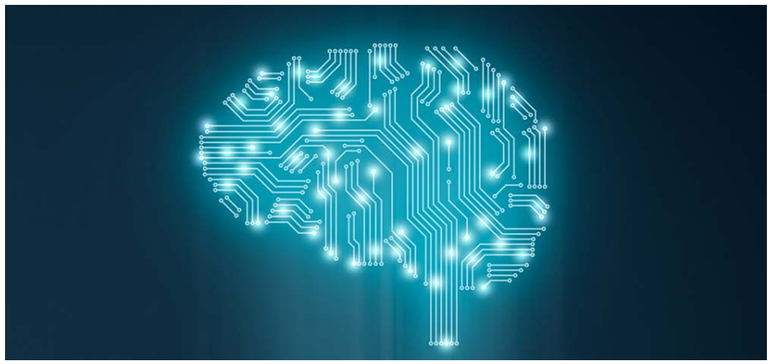


Áp Dụng Các Phương Pháp Học Máy để:

Chọn Các SNP Và Dự Đoán Tính Trạng Của Gạo Từ Dữ Liệu Chiều Cao GWAS

Giảng Viên Phụ Trách Thực Tập

TS. Vũ Tiến Dũng

Giảng Viên Hướng Dẫn

TS. Lê Đức Hậu

Bảo Vệ:

Tạ Văn Nhân

VNU - HUS

NỘI DUNG

- GIỚI THIỆU
- DỮ LIỆU
  - MÔ TẢ DỮ LIỆU
  - TIỀN XỬ LÝ DỮ LIỆU
- CÁC MÔ HÌNH HỒI QUY
  - LASSO VÀ ELASTIC NET
  - SUPPORT VECTOR REGRESSION
  - RANDOM FOREST REGRESSION
- THÍ NGHIỆM
- KẾT QUẢ
- SO SÁNH GIỮA CÁC MÔ HÌNH

GIỚI THIỆU

Single Nucleotide Polymorphism (SNP): là một sự khác nhau giữa các nucleotide của hai chuỗi DNA. Cứ khoảng từ 100 đến 300 cặp Nucleotide của thực vật sẽ xuất hiện một SNP.

Mục đích: thu gọn các SNP ảnh hưởng đến các tính trạng của cây lúa, kết quả có thể được áp dụng trong sơ đồ chọn giống.

DỮ LIỆU VÀ PHƯƠNG PHÁP

Mô Tả Dữ Liệu

Bộ dữ liệu trong nghiên cứu về dữ liệu GWAS của Oryza Sativa từ dự án 3000 gen lúa , có thể được download tại: *https://snp-seek.irri.org/_download.zul*.

Dữ liệu bao gồm 404,388 SNPs của 1869 giống lúa. Chúng ta sẽ xem xét bộ dữ liệu này với đầu ra là "grain weight" (trọng lượng hạt) và "time to flowering" (thời gian ra hoa).

DỮ LIỆU VÀ PHƯƠNG PHÁP

Tiền Xử Lý Dữ Liệu

Vì số SNP rất lớn nên trước tiên ta cần lọc bớt các SNP . Bước này được thực hiện dựa trên P-value của các SNP.

Cụ thể, với biến phụ thuộc là "grain weight" ta chỉ chọn các SNP với P-value từ 5.10^{-5} (14 SNPs) đến 5.10^{-2} (15407 SNPs). Với biến đầu ra là "time to flowering" ta giới hạn P-value từ 5.10^{-6} (5 SNPs) đến 5.10^{-2}(28307 SNPs).

CÁC MÔ HÌNH HỒI QUY

Mô hình hồi quy được áp dụng để ước lượng mối quan hệ giữa các biến phụ thuộc và các biến độc lập.

Ví dụ mô hình hồi quy chỉ có một biến đầu ra (single-task) là ma trận y cỡ *n x 1*, dữ liệu đầu vào là ma trận X cỡ *n x p* , ma trận tham số w cỡ *p x 1*, hệ số tự do b và sai số ε:

$$y = Xw + b + \varepsilon$$

CÁC MÔ HÌNH HỒI QUY

Lasso và Elastic Net

Để ước lượng tham số, người ta thường tối ưu hóa tổng bình phương sai số (bài toán least-square):

$$\min_w \frac{1}{2n}\sum_{i=1}^{n}(y_i - (x_i w + b))^2$$

Tuy nhiên, đối với dữ liệu chiều cao, khi số lượng các đặc trưng p rất lớn so với số mẫu n bài toán least-square có thể trở nên khó thực hiện. Để giải quyết vấn đề này người ta đưa vào các hàm điều chỉnh chuẩn l1, l2 theo các cách khác nhau vào các mô hình.

CÁC MÔ HÌNH HỒI QUY

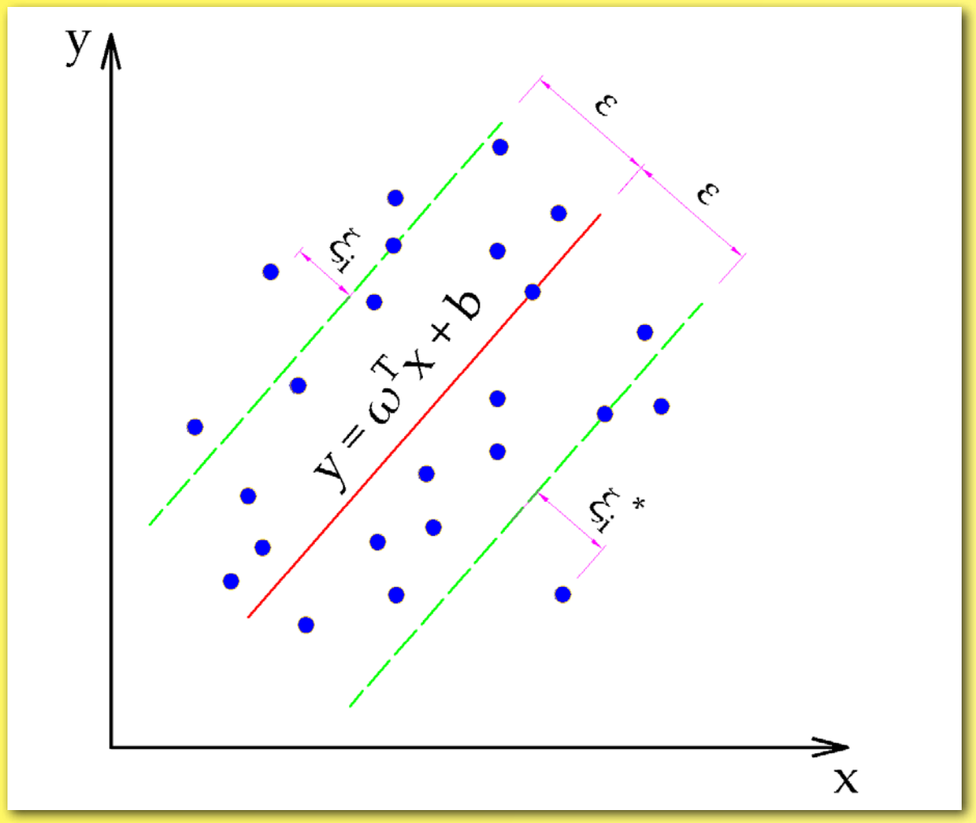
Lasso và Elastic Net

- Lasso $$\min_w \frac{1}{2n}\sum_{i=1}^{n}(y_i-(x_i w+b))^2+\alpha\sum_{j=1}^{p}\left|w_j\right|$$
- Multitak Lasso $$\min_W \frac{1}{2n}\sum_{i=1}^{n}(Y_i-(x_i W+b))^2+\alpha\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{p}w_{ij}^2}$$
- Elastic Net $$\min_w \frac{1}{2n}\sum_{i=1}^{n}(y_i-(x_i w+b))^2+\alpha\rho\sum_{j=1}^{p}\left|w_j\right|+\frac{\alpha(1-\rho)}{2}\sum_{j=1}^{p}w_j^2$$
- Multitask Elastic Net $$\min_W \frac{1}{2n}\sum_{i=1}^{n}(Y_i-(x_i W+b))^2+\alpha\rho\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{p}w_{ij}^2}+\frac{\alpha(1-\rho)}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}w_{ij}^2$$

Trong đó $\alpha\geq 0$, $0\leq \rho \leq 1$. Chúng ta có thể sử dụng các mô hình này để lựa chọn các biến quan trọng có hệ số khác 0.

Support Vector Regression

Support Vector Regression được xây dựng dựa trên ý tưởng của thuật toán soft margin. Các biến slack ξ_i và ξ_i^* được thêm vào mô hình để giải bài toán tối ưu có ràng buộc.



$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

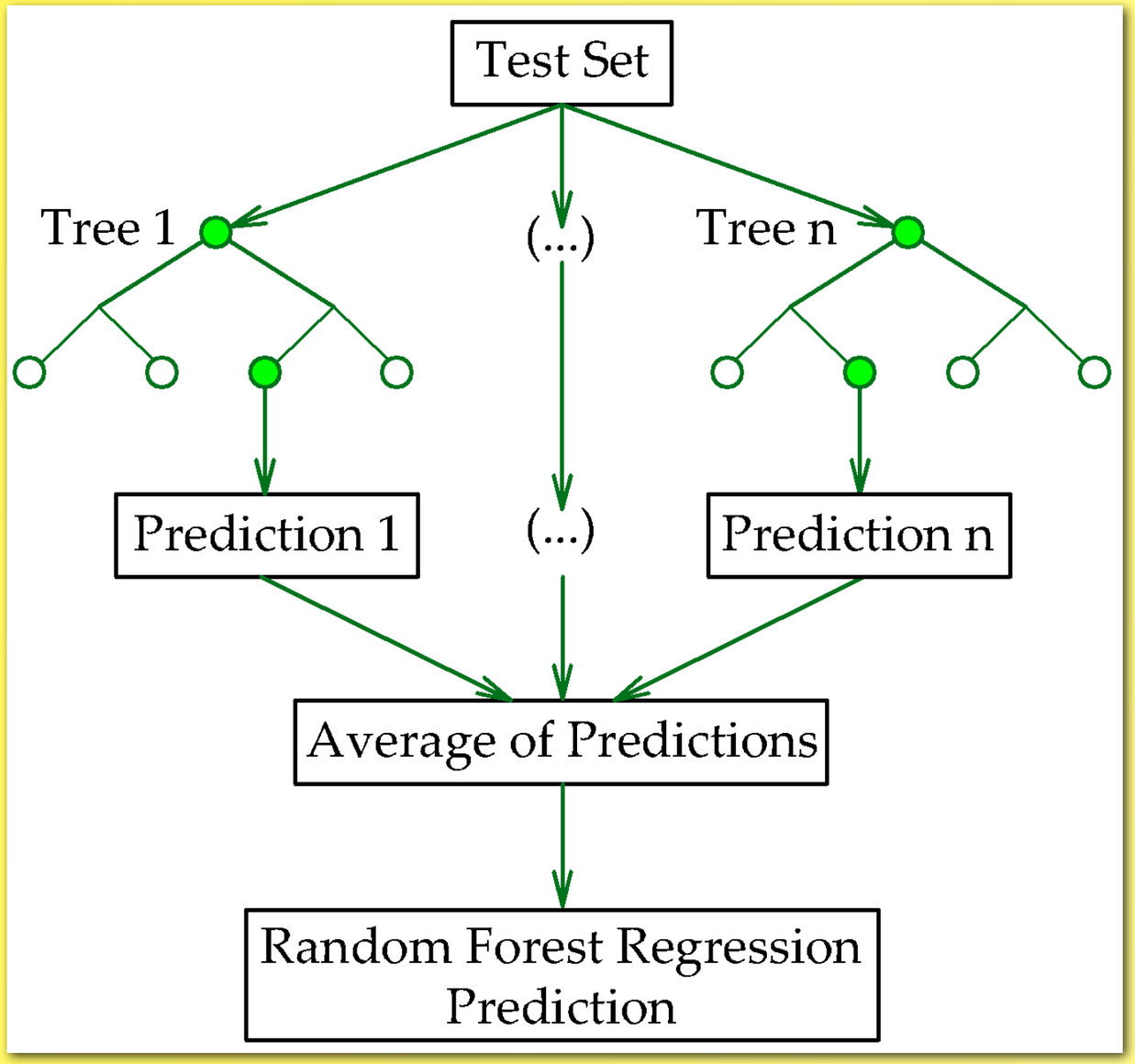
s.t. $y_i - x_i w - b \le \varepsilon + \xi_i, \quad i = 1, \ldots n,$

$x_i w + b - y_i \le \varepsilon + \xi_i^*, \; i = 1, \ldots, n,$

$\xi_i \ge 0, \; i = 1, \ldots, n,$

$\xi_i^* \ge 0, \; i = 1, \ldots, n.$

Random Forest Regression



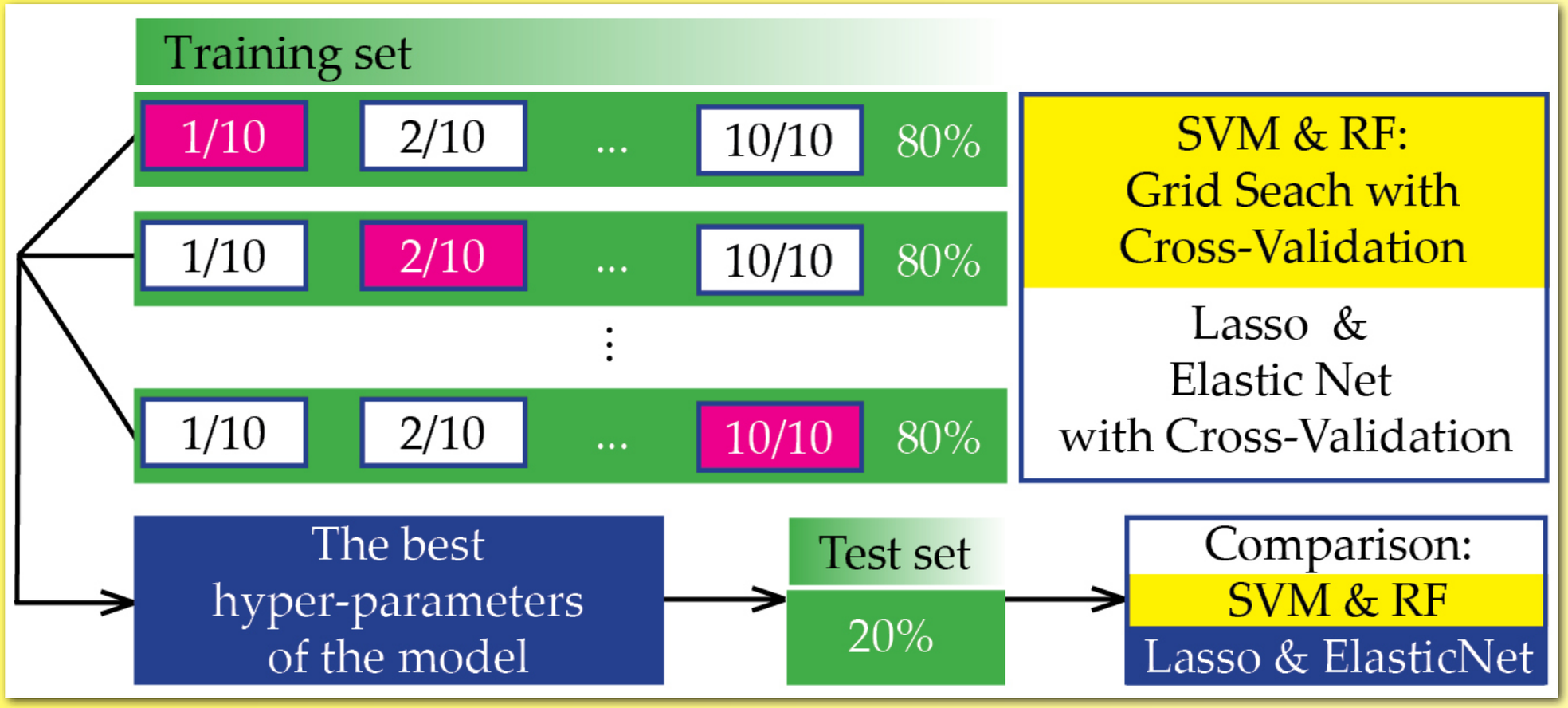
Random Forest Regression là kết hợp của nhiều cây hồi quy để tránh overfitting cho các cây riêng lẻ.

Giá trị dự đoán của mô hình bằng trung bình giá trị dự đoán của các cây hồi quy.

THÍ NGHIỆM

Workflow

Trong khi Gridsearch ước lượng tham số tốt nhất thì k-fold cross-validation giúp giảm overfitting cho tập huấn luyện



Các Độ Đo

Mean Square Error (MSE): trung bình phương sai của giá trị dự đoán và giá trị quan sát.

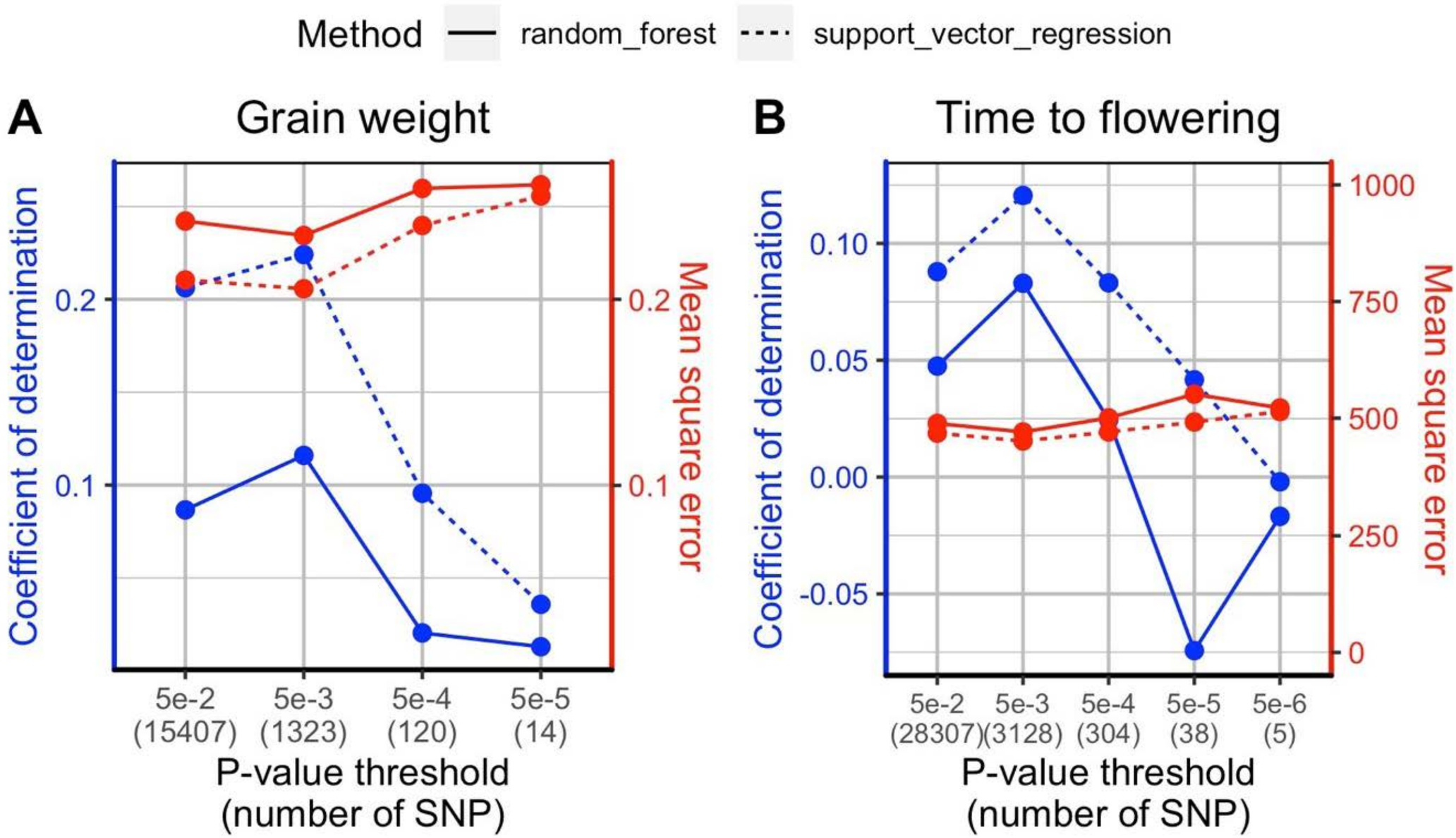
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{obs}^{(i)} - y_{pred}^{(i)})^2.$$

R-squared (R^2): đại diện cho mức độ giải thích của các biến đầu vào cho các biến đầu ra.

$$R^2 = 1 - \frac{explained\ variance}{total\ variance} = 1 - \frac{\sum_{i=1}^{m}\left(y^{(i)} - y_{pred}^{(i)}\right)^2}{\sum_{i=1}^{m}\left(y^{(i)} - y_{mean}^{(i)}\right)^2}.$$

KẾT QUẢ

Random Forest và Support Vector Regression



KẾT QUẢ

Lasso và Elastic Net

	Grain weight			Time to flowering		
Advanced linear regressors	#SNP	MSE	R^2	#SNP	MSE	R^2
Lasso	1	0.26524	-0.00052	4	514.4029	-0.00064
Multi-task Lasso	4	**0.26484**	**0.00100**	4	**514.4028**	**-0.00064**
Elastic Net	1	0.26524	-0.00052	26	514.7855	-0.00138
Multi-task Elastic Net	26	0.26501	0.00034	26	514.7853	-0.00138

#SNP	Single task lasso on grain weight	Single task lasso on time to flowering	Multi-task lasso
1	172235150	120600601	120600601
2		180393908	180393908
3		208081820	208081820
4		373102125	373102125

KẾT QUẢ

Lasso và Elastic Net

#SNP	Single task elastic net on grain weight	Single task elastic net on time to flowering	Multi-task elastic net
1	172235150	2389069	2389069
2		11760405	11760405
3		34489925	34489925
4		41750977	41750977
5		50628584	50628584
6		61375990	61375990
7		84928121	84928121
8		120600601	120600601
9		125832526	125832526
10		137230497	137230497
11		147879747	147879747
12		180393908	180393908
13		180644828	180644828
14		208081820	208081820
15		227565875	227565875
16		230068383	230068383
17		235911589	235911589
18		239306881	239306881
19		239531623	239531623
20		244281569	244281569
21		274806614	274806614
22		285157810	285157810
23		341280196	341280196
24		345796823	345796823
25		363361795	363361795
26		373102125	373102125

RFR và SVR: chọn được 1323 SNPs cho "Grain weight", và 3128 SNPs cho "time to flowering".

Với đầu ra kết hợp cả 2 trait: Multi - task Lasso chọn được 4 SNPs;

Multi - task Elastic Net chọn được 26 SNPs.

SO SÁNH GIỮA CÁC MÔ HÌNH

- Các mô hình RFR và SVR đạt được độ chính xác cao hơn Lasso và Elastic Net.

- Mô hình SVR có độ chính xác cao hơn RFR.

- Mô hình Elastic Net có độ chính xác cao hơn Lasso.

- Các mô hình Lasso và Elastic Net rút gọn được nhiều biến hơn so với RFR và SVR.

HỎI ĐÁP

CẢM ƠN THẦY, CÔ GIÁO VÀ CÁC BẠN ĐÃ THEO DÕI!