IRTA
RECERCA I TECNOLOGIA AGROALIMENTÀRIES

Machine learning transcriptome analysis to identify genes associated with feed efficiency in pig

Miriam Piles, Carlos Fernandez-Lozano, María Velasco, Olga González, Juan Pablo Sánchez, Raquel Quintanilla, María Ballester

Molecular mechanisms underlying **feed efficiency** are still unknown

Machine-learning applied into a resampling strategy can provide a good assessment of the generalizability of the results

Objective

To identify genes associated with feed efficiency using transcriptomic (RNA-Seq) data from pigs phenotypically extreme for residual feed intake

$$\boldsymbol{RFI}_{ijk} = \boldsymbol{FI}_{ijk} - \left[S_j + \beta_{\boldsymbol{MBW}_j} \times \boldsymbol{MBW}_i + \beta_{\boldsymbol{ADG}_j} \times \boldsymbol{ADG}_i + \beta_{\boldsymbol{BFG}_j} \times \boldsymbol{BFG}_i\right]$$

$\beta_{\boldsymbol{MBW}_j}$, $\beta_{\boldsymbol{ADG}_j}$ and $\beta_{\boldsymbol{BFG}_j}$ partial regressions coefficients (within sex *j*).

FI: feed intake   ***MBW***: metabolic body weight   ***ADG***: average daily gain   ***BFG***: backfat gain

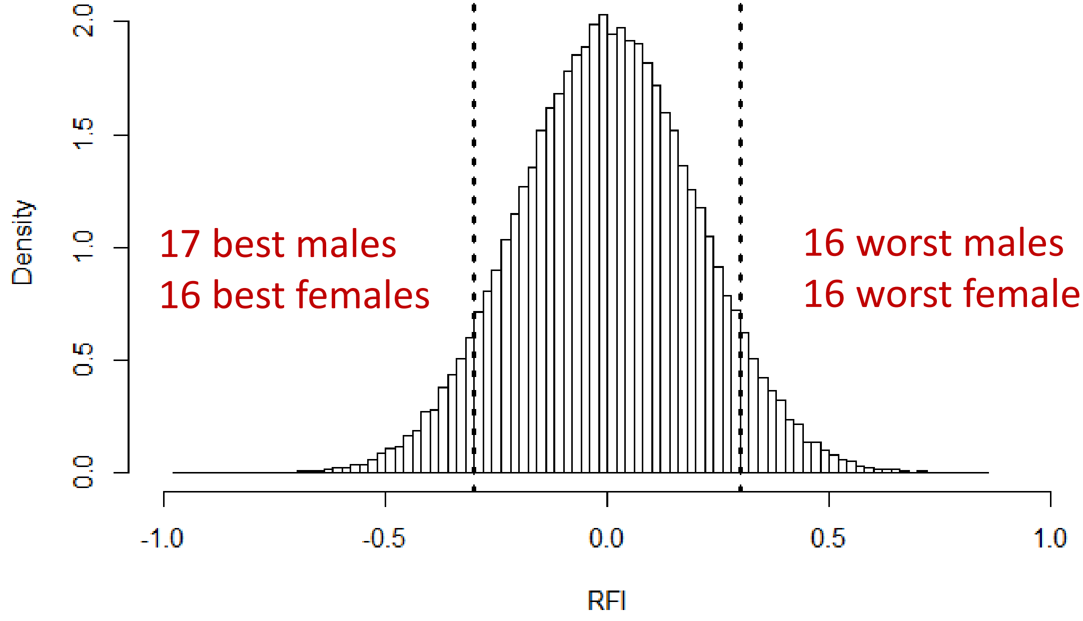
25 Kg of body weight

IVO-automatic feeders
Body weight and backfat measured at
1, 3, 6, 9, 12 and 15 weeks

105 Kg of body weight

123 males and 121 females

Low RFI



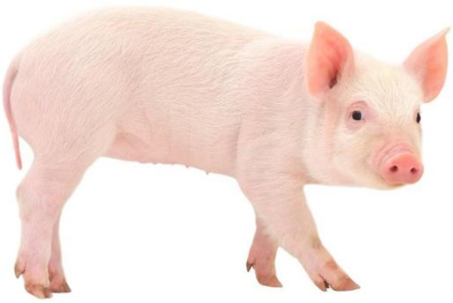
17 best males
16 best females

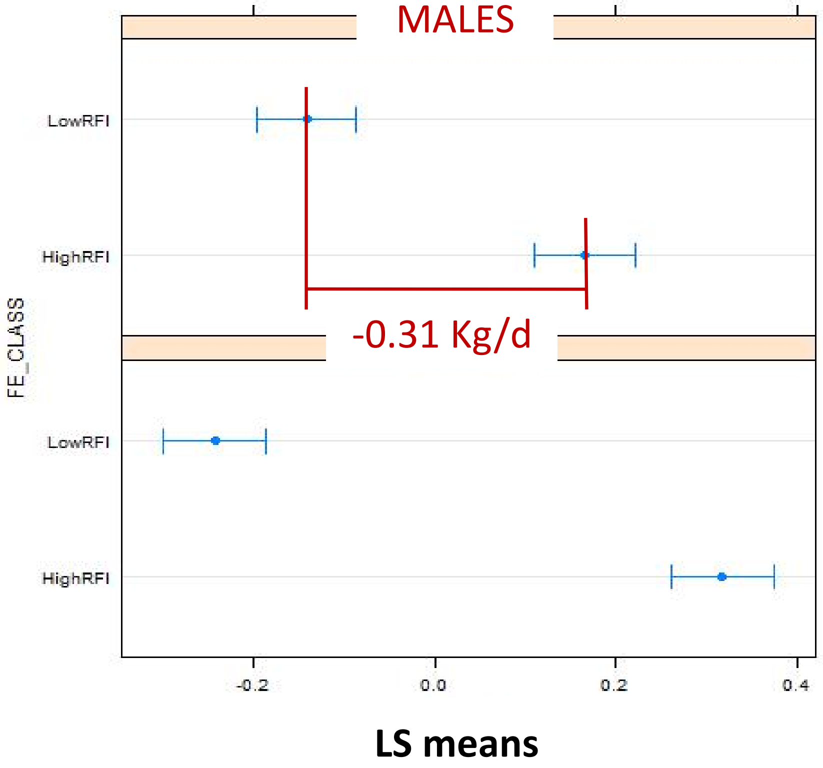
16 worst males
16 worst females

High RFI

Differences at the phenotypic level

Low RFI

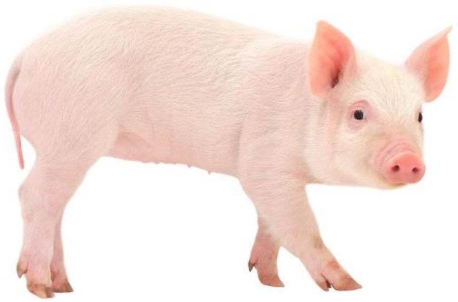


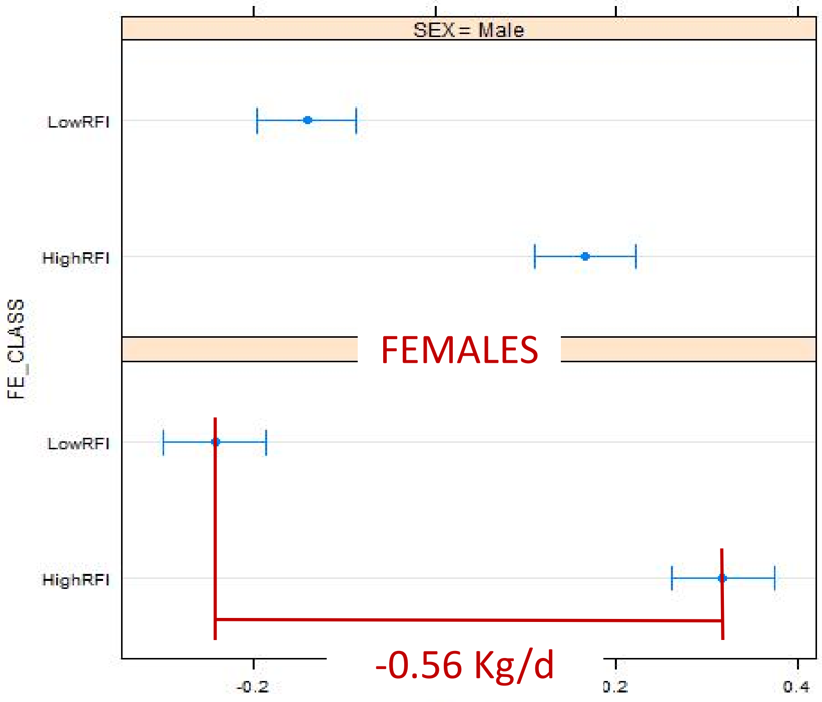


High RFI

Differences at the phenotypic level

Low RFI





FEMALES

-0.56 Kg/d

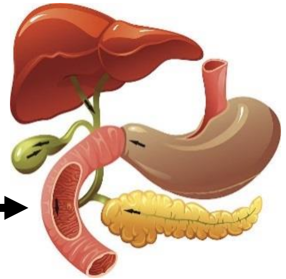
LS means

High RFI

65 animals

65 samples from liver

65 samples from duodenum



RNA-Seq with an Illumina Hiseq2000

Differential expression analyses: Machine learning algorithms

Functional categorization of DE genes: IPA

Differential expression analysis

1st step: Gene ranking based on permutation accuracy importance score

Classification using Random Forest algorithm based on conditional inference

Data (**X**, **y**)

X is the **65 x 13990 predictor matrix of RNA-Seq data**

y is the **class vector: Low or High RFI**

⇩

2nd step: Classification using Machine learning algorithms

Data (**X**, **y**)

X is the **65 x p predictor matrix of RNA-Seq data**

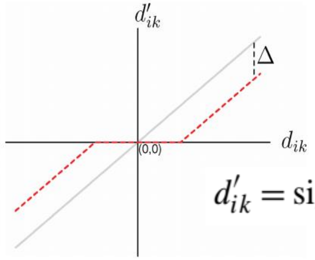
Subsets of *p* predictors ***p* = 50, 75, 100, 125, 150, 200, 250, 300, 350, 400**

y is the **class vector: Low or High RFI**

Machine learning algorithms

mlr R package

Nearest Shrunken Centroids (PAMR)



$$\bar{x}'_{ik} = \bar{x}_i + m_k s_i d'_{ik}$$

$$d'_{ik} = \operatorname{sign}(d_{ik})(|d_{ik}| - \Delta)_+$$

For a test sample $x^* = (x_1^*, x_2^*, \ldots x_p^*)$

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \bar{x}'_{ik})^2}{s_i^2} - 2\log \pi_k$$

$$C(x^*) = \ell \text{ if } \delta_\ell(x^*) = \min_k \delta_k(x^*)$$

Elastic net (CVGLMNET)

$$\hat{\beta} \quad = \quad \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

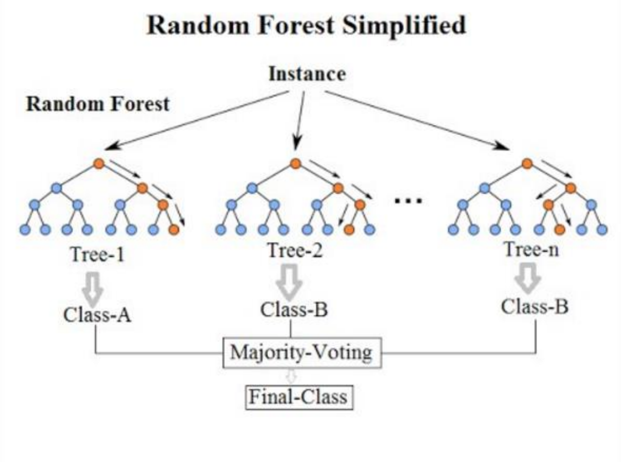
The elastic net penalty

$$J(\beta) = \alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1$$

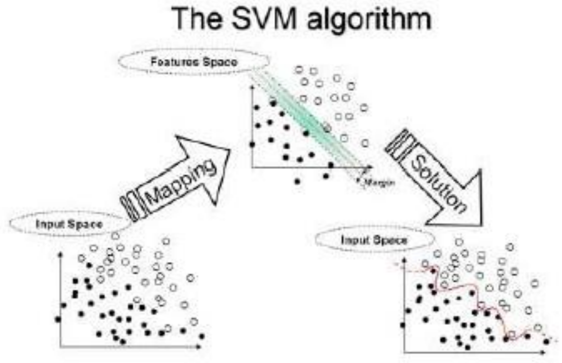
(with $\alpha = \frac{\lambda_2}{\lambda_2+\lambda_1}$)

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 \text{ s.t. } J(\beta) \le t.$$

Random Forest (randomForest)



Support Vector Machine (e1071)

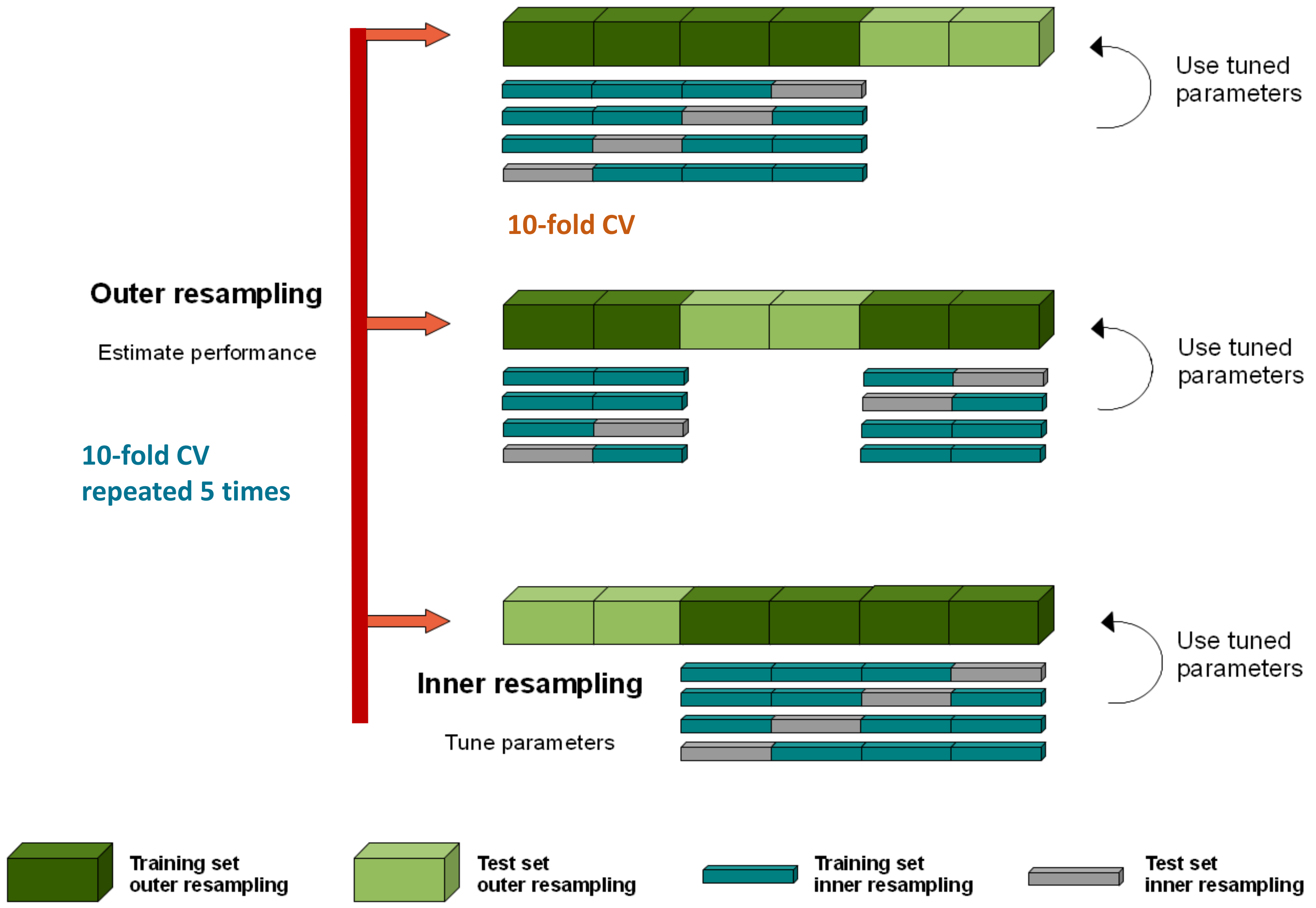


$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M \quad \text{subject to} \quad \sum_{j=1}^{p}\beta_j^2 = 1,$$

$$y_i\left(\beta_0 + \sum_{i\in S}\hat{\alpha}_i\langle x, x_i\rangle\right) \ge M(1-\varepsilon_i)$$

$$\sum_{i=1}^{n}\varepsilon_i \le C \qquad \varepsilon_i \ge 0$$

Nested Resampling



Outer resampling

Estimate performance

10-fold CV
repeated 5 times

10-fold CV

Use tuned parameters

Use tuned parameters

Inner resampling

Tune parameters

Use tuned parameters

Training set outer resampling

Test set outer resampling

Training set inner resampling

Test set inner resampling

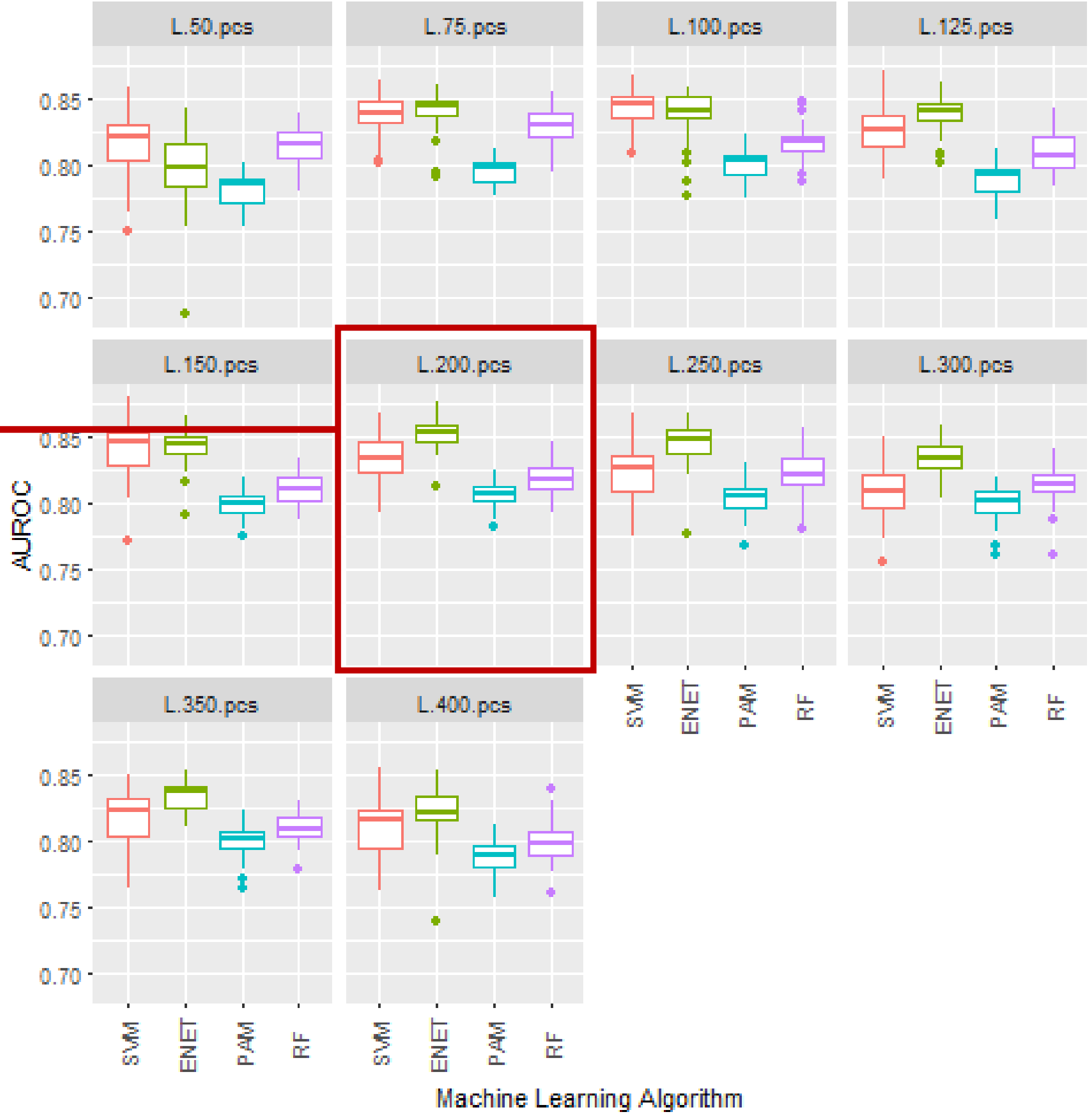
Classification Performance

LIVER

AUROC

AUROC = 0.85

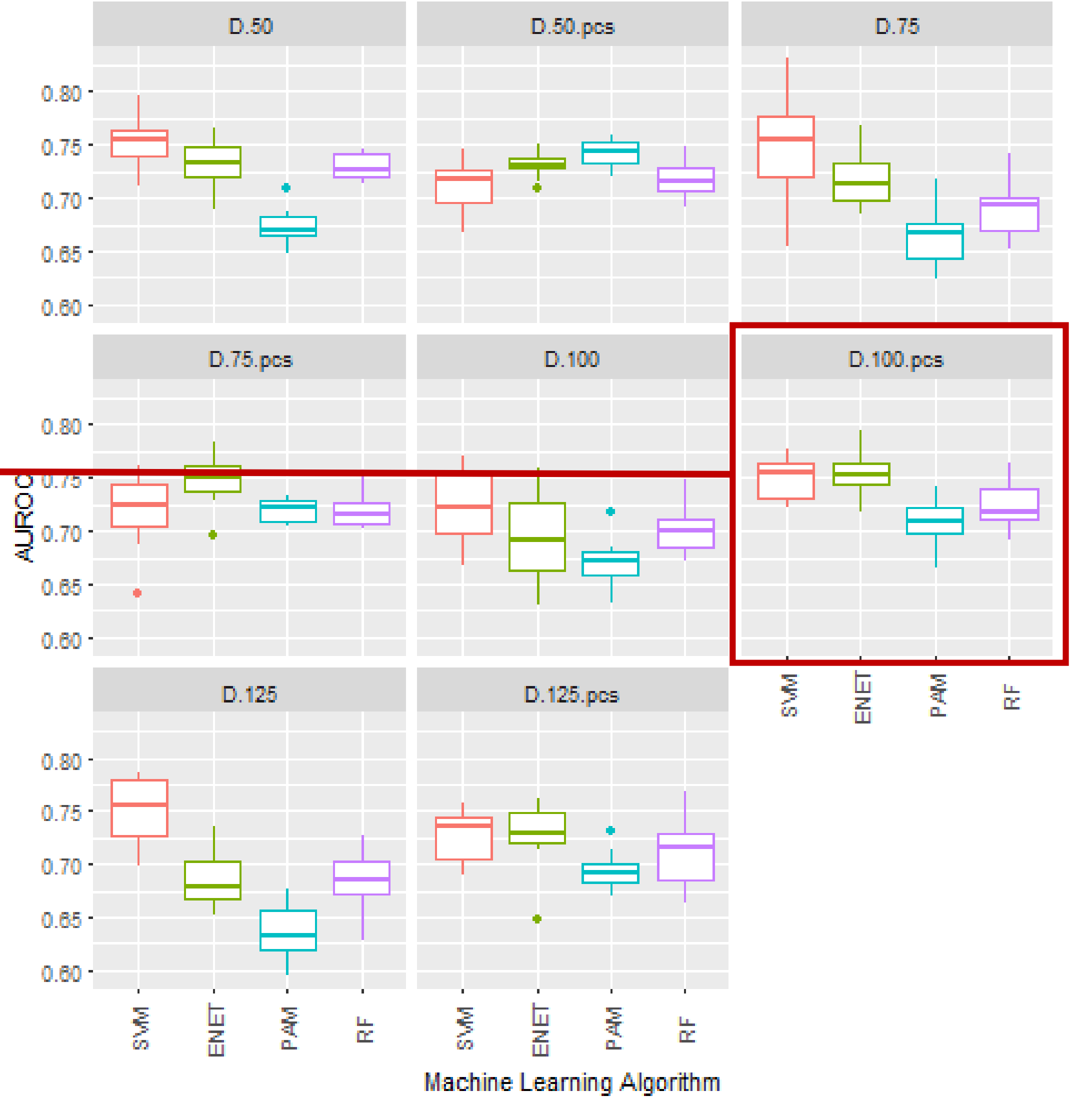
Accuracy = 0.78



Classification Performance

DUODENUM
AUROC

AUROC = 0.76
Accuracy = 0.69



Functional analysis: most significant overrepresented canonical pathways in liver

Canonical Pathways	-log(p-value)	Molecules
Melatonin Degradation II	3,41E00	SMOX,IL4I1
Sirtuin Signaling Pathway	2,01E00	WRN,TUBA1A,GLS,TOMM20,CLOCK,TUBA3E,PFKM
PPARα/RXRα Activation	1,8E00	PRKAR2B,CLOCK,GNA14,CYP2C8,AIP
Tryptophan Degradation X (Mammalian, via Tryptamine)	1,76E00	SMOX,IL4I1
GPCR-Mediated Nutrient Sensing in Enteroendocrine Cells	1,2E00	PRKAR2B,GNA14,CCK
Xenobiotic Metabolism Signaling	6,81E-01	SMOX,IL4I1,CYP2C8,AIP

Functional analysis: most significant overrepresented canonical pathways in duodenum

Canonical Pathways	-log(p-value)	Molecules
NRF2-mediated Oxidative Stress Response	3,04E00	DNAJC6,MAPK8,DNAJC1,PRKD3
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	2,45E00	MAPK8,CASP1,PRKD3
Aldosterone Signaling in Epithelial Cells	2,2E00	DNAJC6,DNAJC1,PRKD3
Cholecystokinin/Gastrin-mediated Signaling	1,67E00	MAPK8,PRKD3
IL-8 Signaling	1,14E00	MAPK8,PRKD3
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	1,15E00	MAPK8,PRKD3
Unfolded protein response	9,35E-01	MAPK8
Protein Ubiquitination Pathway	9,25E-01	DNAJC6,DNAJC1

Conclusions

- Good performance of ML algorithms and RNA-Seq expression data for classifying pigs into high or low RFI groups.

- Expression difference of genes involved in feed efficiency are clearer in liver than in duodenum tissues.

- Among biological pathways identified in liver and duodenum tissues, those related to response to oxidative stress, aldosterone signaling and melatonin degradation seemed consistently involved in FE related traits.

Funding

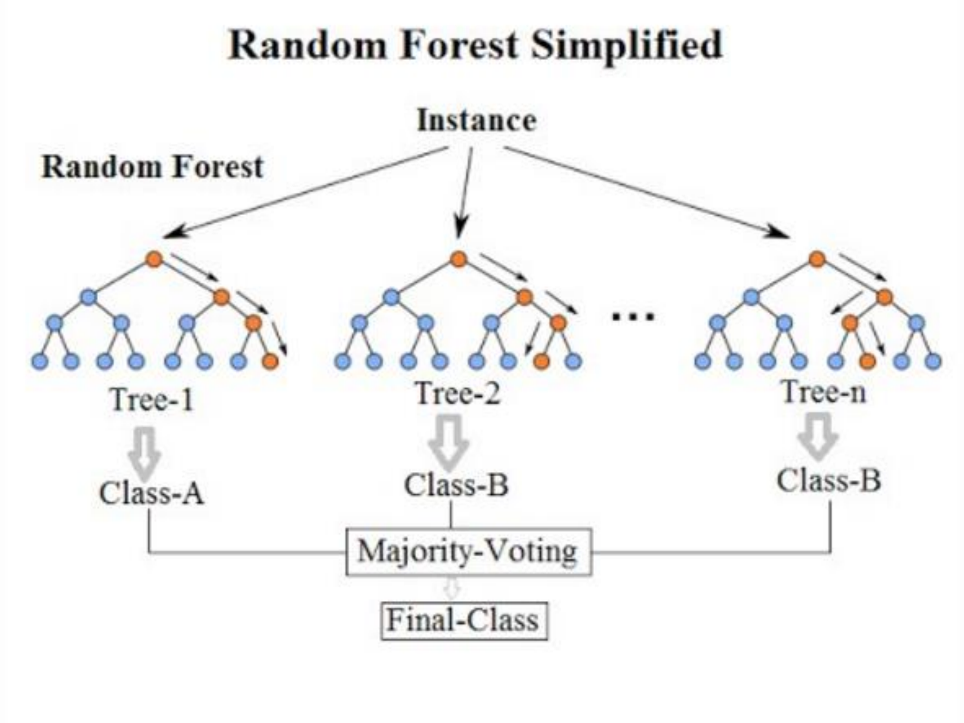
This work was funded by the **European Union Seventh Framework Programme** (FP7/2007–2013) as part of the **ECO-FCE project** under grant agreement No 311794 and the **Spanish Ministry of Economy, Industry and Competitiveness (MINECO)** Carlos Fernandez-Lozano received a grant from **Juan de la Cierva Fellowship Program** (Ref. FJCI-2015-26071) and M. Ballester received a grant from **Ramón y Cajal Post-doctoral Fellowship Program** (RYC-2013-12573).

ECOFCE

GOBIERNO DE ESPAÑA
MINISTERIO
DE ECONOMÍA, INDUSTRIA
Y COMPETITIVIDAD

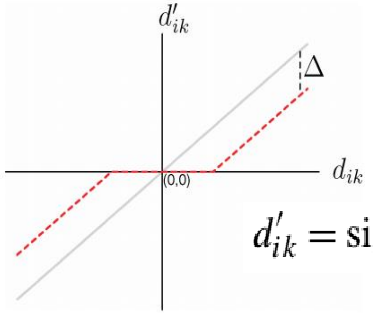
Machine learning algorithms

Random Forest



Machine learning algorithms

Nearest Shrunken Centroids (PAMR)



$$\bar{x}'_{ik} = \bar{x}_i + m_k s_i d'_{ik}$$

$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+$$

For a test sample $x^* = (x_1^*, x_2^*, \ldots x_p^*)$

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \bar{x}'_{ik})^2}{s_i^2} - 2 \log \pi_k$$

$$C(x^*) = \ell \text{ if } \delta_\ell(x^*) = \min_k \delta_k(x^*)$$

Machine learning algorithms

Elastic net (CVGLMNET)

$$\hat{\beta} \quad = \quad \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

The elastic net penalty

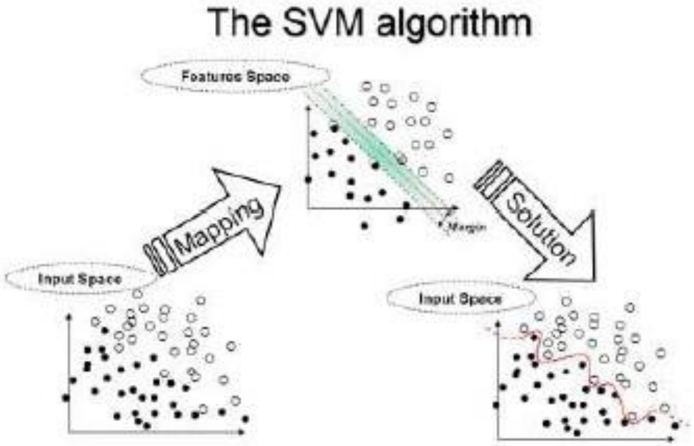
$$J(\beta) = \alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1$$

(with $\alpha = \frac{\lambda_2}{\lambda_2+\lambda_1}$)

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 \text{ s.t. } J(\beta) \leq t.$$

Machine learning algorithms

Support Vector Machine (SVM)



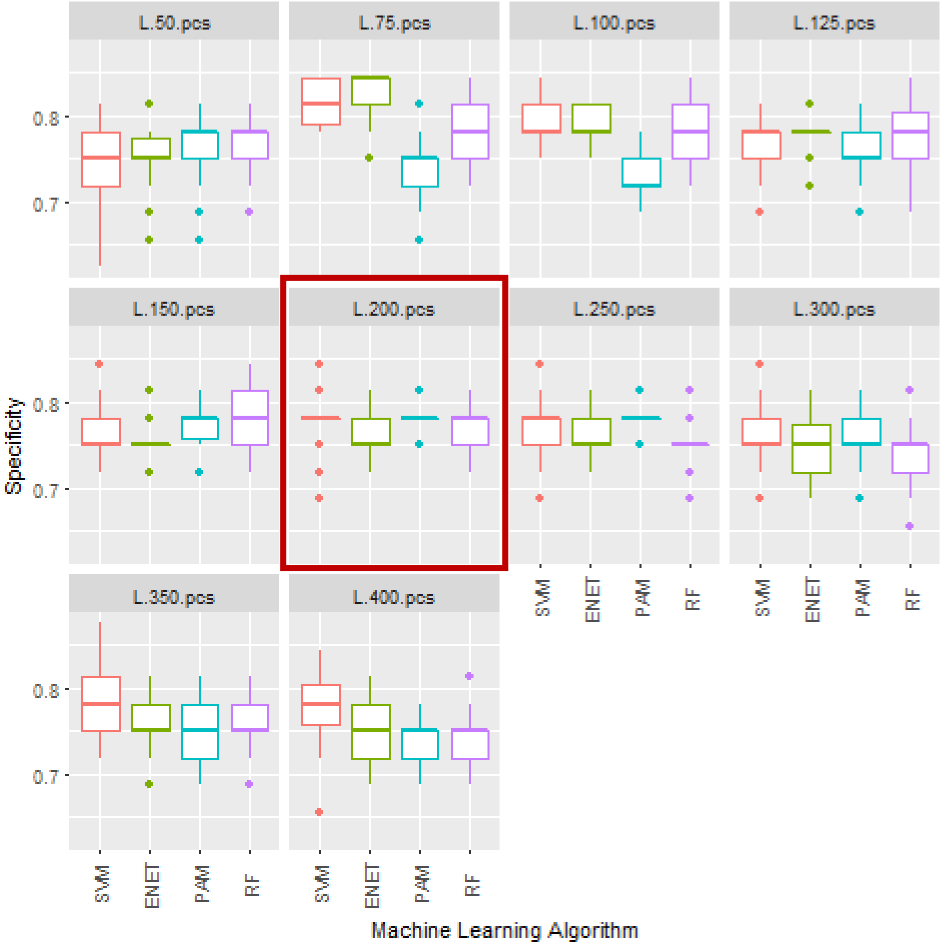
$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M \quad \text{subject to} \quad \sum_{j=1}^{p}\beta_j^2 = 1,$$

$$y_i\left(\beta_0 + \sum_{i\in S}\hat{\alpha}_i\langle x, x_i\rangle\right) \geq M(1-\varepsilon_i)$$

$$\sum_{i=1}^{n}\varepsilon_i \leq C \qquad \varepsilon_i \geq 0$$

Classification Performance

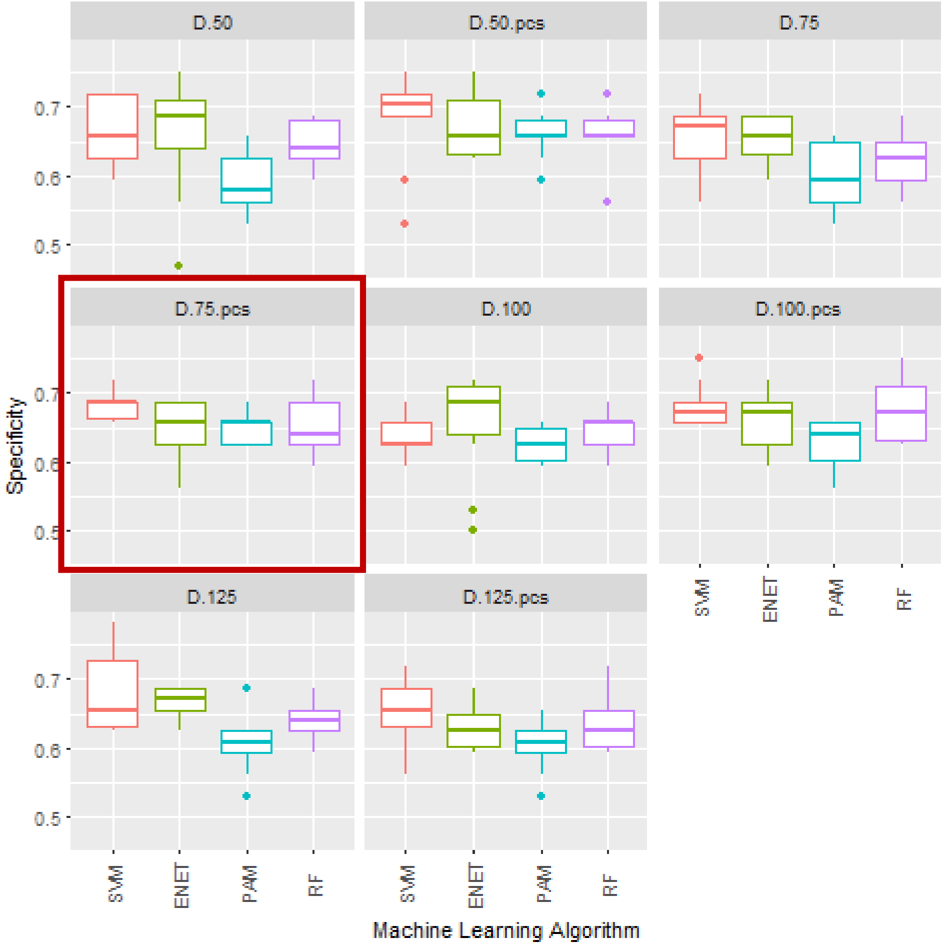
LIVER
Specificity



Classification Performance

DUODENUM

Specificity


D.50
D.50.pcs
D.75
D.75.pcs
D.100
D.100.pcs
D.125
D.125.pcs
Specificity
SVM
ENET
PAM
RF
Machine Learning Algorithm