Ecole d'hiver 6-EGC

11h30-13h00

Machine Learning and interpretability : examples in precision medicine

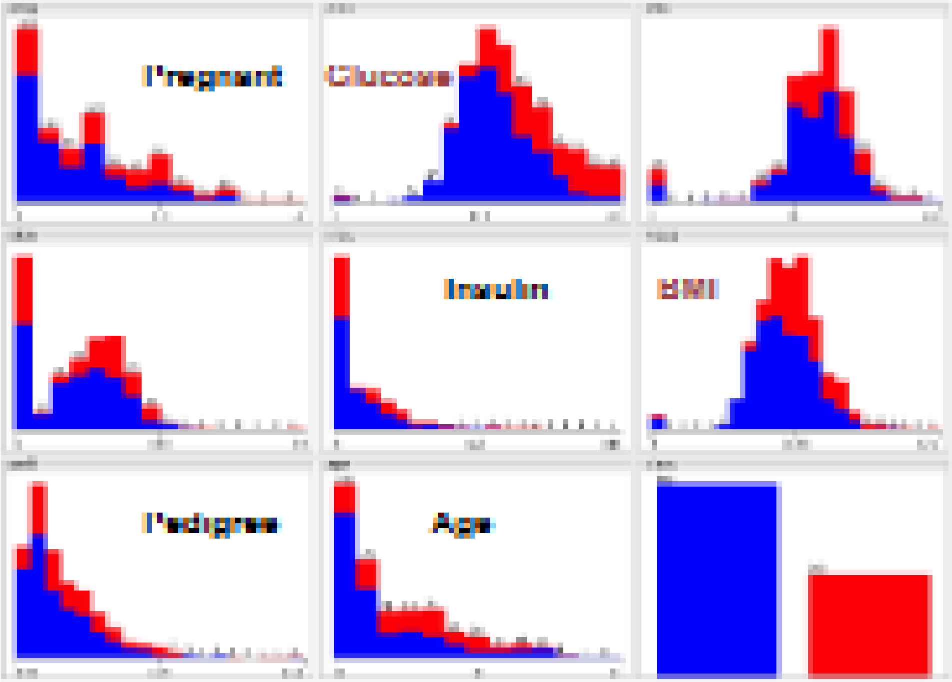
Jean-Daniel Zucker

DR IRD

Travail en collaboration E. Prifti (IRD), Y. Chevaleyre (Dauphine), B. Hanczar (Evry), K. Clément (INSERM) & N. Sokolovska (SU)

Partial Dependancy Plots : they show the marginal effect of values of one or two variables

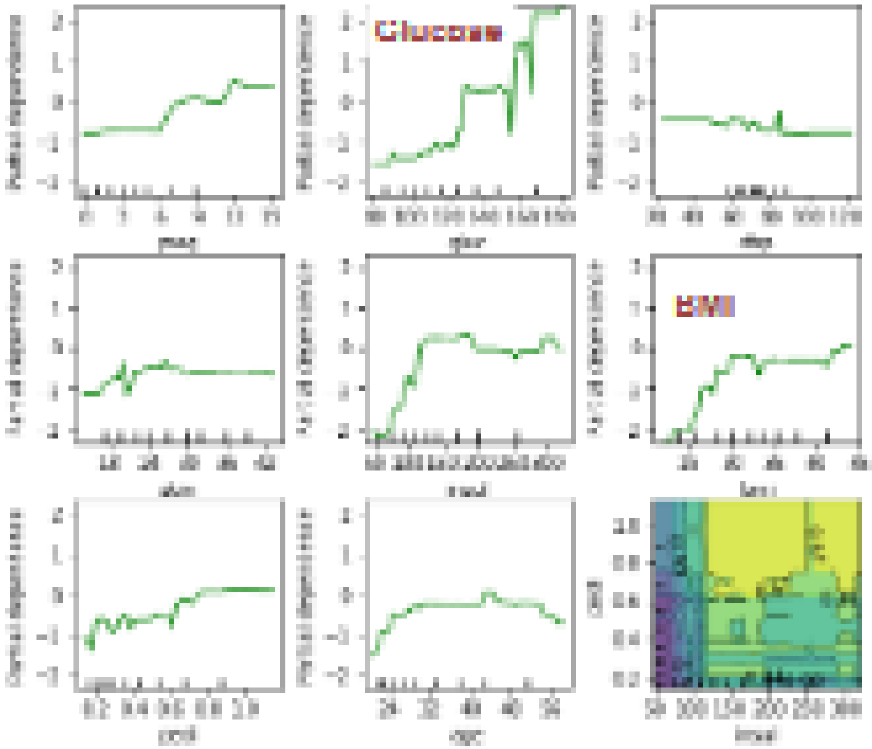
Feature label	Variable type	Range
Number of times pregnant	Integer	0–17
Plasma glucose concentration in a 2 h oral glucose tolerance test	Real	0–199
Diastolic blood pressure	Real	0–122
Triceps skin fold thickness	Real	0–99
2 h serum insulin	Real	0–846
Body mass index	Real	0–67.1
Diabetes pedigree function	Real	0.078–2.42
Age	Integer	21–81
Class	Binary	Tested positive for diabetes = 1



Partial Dependancy Plots : they show the marginal effect of values of one or more variables

- If you are familiar with linear or logistic regression models, partial dependence plots can be interpreted similarly to the coefficients in those models.
- But partial dependence plots can capture more complex patterns from your data, and they can be used with any model.

Y axis: « diabetes partial dependance»



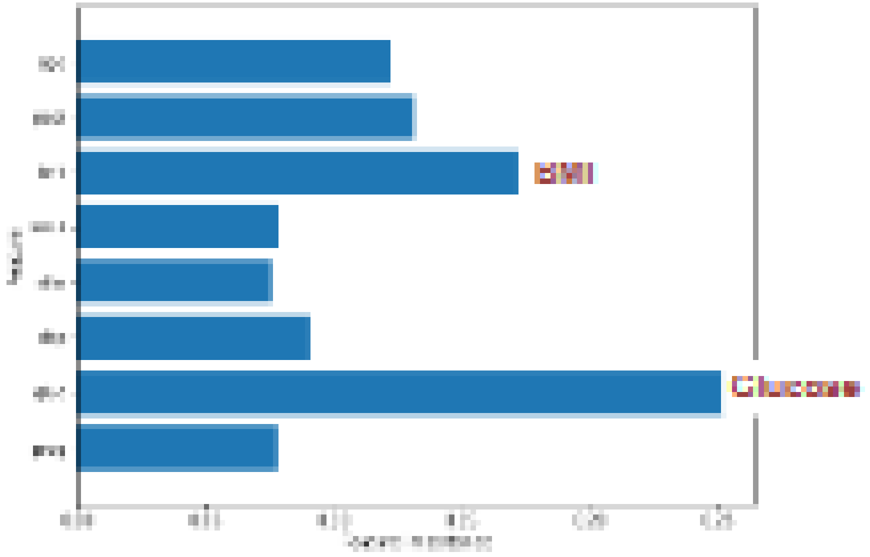
Variable Importance: Global, Model-Agnostic or not

Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way.

- To measure the importance of the ith feature after training, the values of the i-th feature are permuted among the training data and the out-of-bag error is again computed on this perturbed data set.

- The importance score for the i-th feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees.

- The score is normalized by the standard deviation of these differences.



Code Python [illegible]