# **Determining the importance of features in a random forest classification**

## Random Forest feature importance with Gini Impurity Measure

[sci-hub.st/10.1016/j.rse.2020.112105](https://sci-hub.st/10.1016/j.rse.2020.112105)

*GINI can be biased when predicting variables have missing values or are of different types (Strobl et al., 2007). It is only used in CART-RF (in R package randomForest).*

[Random Forest Feature Importance Computed in 3 Ways with Python | MLJAR](https://mljar.com/blog/feature-importance-in-random-forest/)

*This biggest advantage of this method is a speed of computation - all needed values are computed during the Radom Forest training. The drawbacks of the method is to tendency to prefer (select as important) numerical features and categorical features with high cardinality. What is more, in the case of correlated features it can select one of the feature and neglect the importance of the second one (which can lead to wrong conclusions)*

## Permutation Based Feature Importance / Permutated variable importance measure (PVIM)

[Random Forest Feature Importance Computed in 3 Ways with Python | MLJAR](https://mljar.com/blog/feature-importance-in-random-forest/)

*The permutation based importance is computationally expensive. The permutation based method can have problem with highly-correlated features, it can report them as unimportant.*

[sci-hub.st/10.1016/j.rse.2020.112105](https://sci-hub.st/10.1016/j.rse.2020.112105)

*It assesses the average reduction of OOB accuracy after permuting a variable over all trees. The key advantage of the PVIM is that it covers the multivariate interactions within covariates (Nicodemus, 2011). However, both theoretical and empirical studies found PVIM unstable and biased depending on the number of correlated variables (Archer and Kimes, 2008; Gregorutti et al., 2017). PVIM is available for both CART-RF and CIT-RF (in R package randomForest).*

Conditional permutation variable importance measure (CPVIM): *Nicodemus et al., (2010) found that the PVIM generates instable results due to correlations, while the ranking of conditional permutated importance is more reliable and consistent with expectations (in R package party).*

[Feature Selection with the Boruta Package (researchgate.net)](https://www.researchgate.net/profile/Witold-Rudnicki/publication/280138095_Feature_Selection_with_Boruta_Package/links/02bfe51407cd7a64a3000000/Feature-Selection-with-Boruta-Package.pdf)

[Permutation importance: a corrected feature importance measure | Bioinformatics | Oxford Academic (oup.com)](https://academic.oup.com/bioinformatics/article/26/10/1340/193348?login=true)

## Feature Importance computed with SHAP values

[Random Forest Feature Importance Computed in 3 Ways with Python | MLJAR](https://mljar.com/blog/feature-importance-in-random-forest/)

*It is using the Shapley values from game theory to estimate the how does each feature contribute to the prediction.*

[IJGI | Free Full-Text | Application of Random Forest and SHAP Tree Explainer in Exploring Spatial (In)Justice to Aid Urban Planning | HTML (mdpi.com)](https://www.mdpi.com/2220-9964/10/9/629/htm)

## Guided Regularized Random Forest (GRRF)

[An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S030324341930594X)

## Random Forest mean decrease in accuracy (MDA)

[sci-hub.st/10.1080/15481603.2017.1408892](https://sci-hub.st/10.1080/15481603.2017.1408892)

## Correlation-based feature selection (CFS) (Filter without crating RF model)

[sci-hub.st/10.1080/15481603.2017.1408892](https://sci-hub.st/10.1080/15481603.2017.1408892)

## Minimum Depth

[sci-hub.st/10.1016/j.rse.2020.112105](https://sci-hub.st/10.1016/j.rse.2020.112105)

*Ishwaran (2007) suggested that the structure of trees can reveal the variable importance. Since variables that split close to the root node tend to have strong predicting power, if the largest subtree, with the root node split by X, has the shortest distance, or minimum depth, to the root node, then X is considered to have the highest predicting power. It is also only calculated in CARTRF (in R package RandomForestExplainer).*

## Intervention of prediction measure (IPM)

[sci-hub.st/10.1016/j.rse.2020.112105](https://sci-hub.st/10.1016/j.rse.2020.112105)

*The percentage of times a variable has been selected for splitting this case over all trees is called the intervention of this variable for this case.*