Aula10

DATA SCIENCE IPT

TURMA 02

LINKED EDUCATION

V2019

Participação do Prof. Rodrigo Mello (ICMC-USP)

Ensemble Methods

Conteúdo

Os Ensemble Methods combinam as predições de vários algoritmos visando tratar erros de bias e variância...o objetivo é melhorar a generalização das predições.
Há dois tipos clássicos :

Averaging methods : geram a média de várias predições independentes.
Exemplo : Bagging...(normalmente melhora variância)

Boosting methods : usam várias predições e, sequencialmente, há a tentativa de reduzir o bias da predição combinada. A ideia é produzir uma estimativa melhor combinando muitos estimadores “fracos (?)”.
Exemplo : Adaboost ...(normalmente melhora bias)

Bagging meta-estimadores

Para gerar várias predições (e depois obter a media delas) , um bagging meta-estimador parte de amostras do data set de treinamento. Há algumas variantes de algoritmos :

Pasting : gera as predições coletando amostras do data set de treinamento **sem reposição**.

Bagging (+ usual) : gera as predições coletando amostras do data set de treinamento **com reposição**.

Random Subspaces : seleção randômica de features.

Random Patches : seleção randômica de features e amostras do data set.

Bagging meta-estimadores

Exemplo de amostras selecionadas por bagging e pasting (com e sem reposição) :

Data set de treinamento : 1,3,5,8,9,12,15

Exemplos de amostras randômicas para **bagging** (podem ter o mesmo número de elementos do data set original, pois há reposição) :
1,1,5,9,9,12,15
1,3,4,5,9,12,12
1,2,3,5,5,9,15

Exemplos de amostras randômicas para **pasting** (número de elementos menor que o data set original, pois não há reposição) :
1,3,5,8
1,9,12,15
1,5,8,9

Bagging meta-estimadores

Conteúdo

Partindo de baggin1.ipynb, faça a montagem de amostras com repetição e obtenha a média dos resultados...

Bagging meta-estimadores

Conteúdo

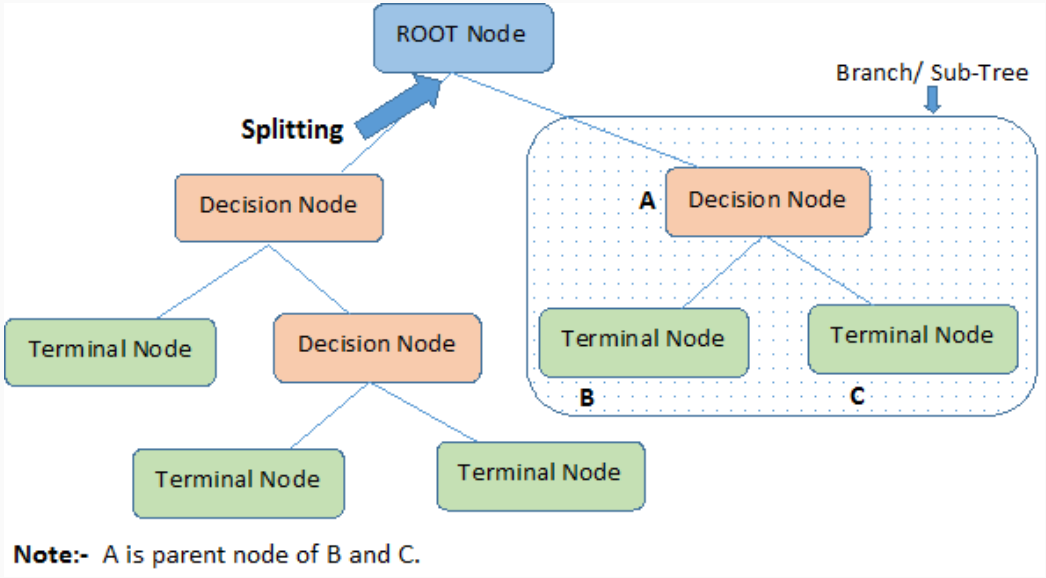
Partindo de bagging2.ipynb, analisar o código, obter o gráfico final do modelo bagging.

Explicar o parâmetro boostrap....

Random Forest

Conteúdo

Inicialmente, falaremos de Decision Trees....e antes ainda, da terminologia delas.



Splitting : divisão do nó em um ou mais sub-nós

Leaf / terminal node : nó final, não dividido, folha.

Prunning : Remoção de um sub-nó...poda

Imagem fonte : https://www.analyticsvidhya.com

Random Forest

Conteúdo

Decision Trees

São algoritmos de aprendizado supervisionado não paramétricos, usados em classificação e regressão.

Decision trees são modelos que fazem previsão de um output com base em regras simples sobre as features.

Exemplo :

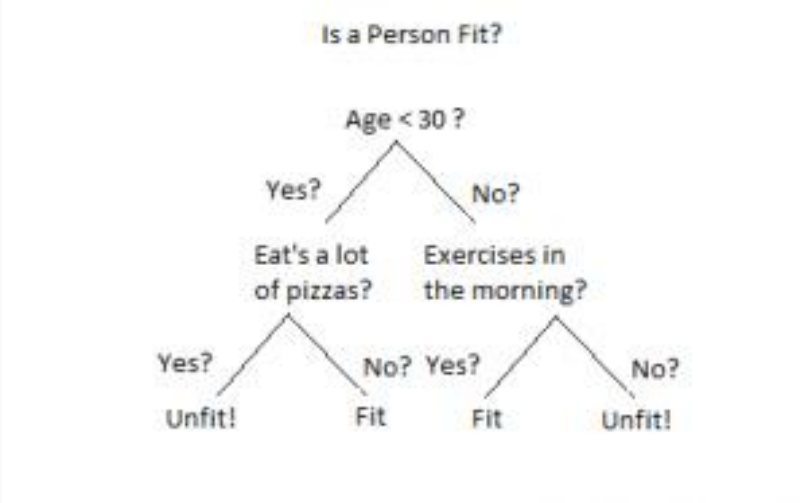


Imagem fonte : https:///www.xoriant.com

Random Forest

Conteúdo

Decision Trees : vantagens e desvantagens

Vantagens :

É simples

Mostra o “processo decisório”

Não requer preparação dos dados (standardization...)

Desvantagens :

Tendência para overfitting (não generaliza bem)

Obter a árvore mais otimizada pode ser um problema computacionalmente complexo

Alguns problemas não são bem modelados por árvores. Exemplo : XOR

Random Forest

Conteúdo

O algoritmo “genérico” para construir a árvore

Escolha um nó raiz

Faça split do nó raiz de acordo com algum critério de “otimização”

Continuar (recursivamente) com splits de nós até toda navegação chegar a um nó terminal (folha)

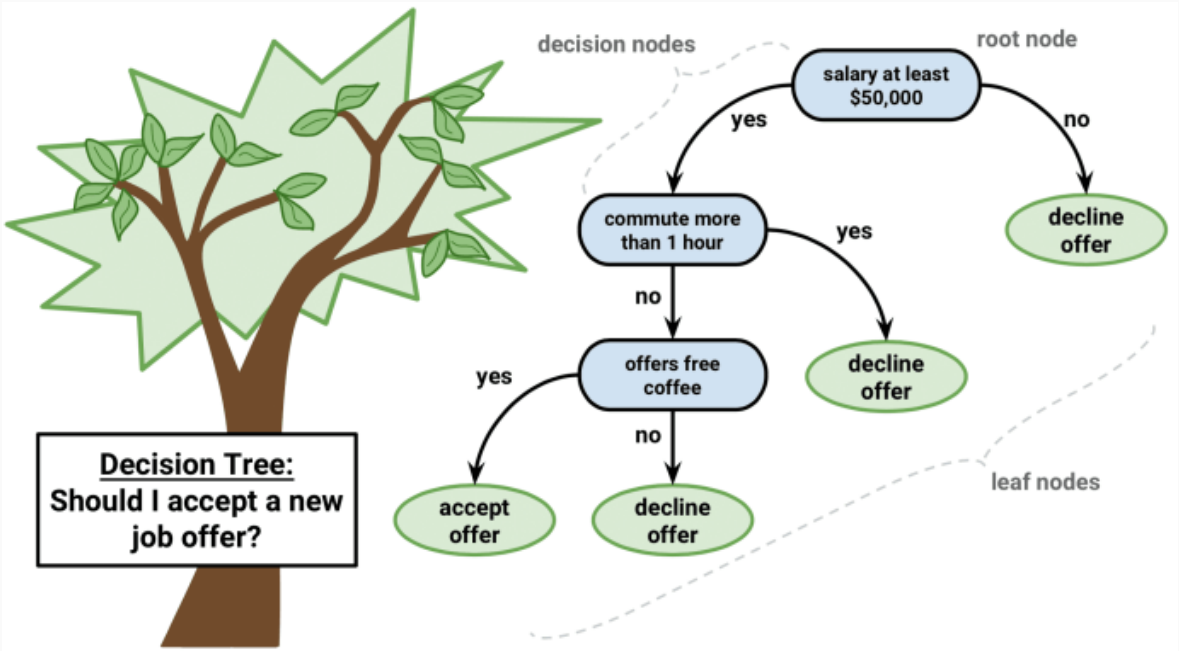


Imagem fonte : https:///dataaspirant.co

Como escolher o split (critério)?

Alternativa : buscar aumentar a “pureza” após o split.

A maior “pureza” ocorre quando todos os elementos de um nó pertencem a uma mesma classe...ou seja, não é necessário mais split..temos o nó folha.

Assim, escolher o split que aumentar a pureza é um bom critério...mas como medir a pureza?

Alternativa : index **Gini (há outras...)**

Índice Gini

Se um dataset **T** contém exemplos de **n** classes, o índice gini é dado por :

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

Onde p_j é a frequência relativa da classe j em T

Quanto menor o índice, maior a pureza...se todos os elementos (10, por exemplo) forem de uma única classe, o índice é zero $(1 – (10/10)^2)$...se há duas classes balanceadas (exemplo 5 elementos em cada):

$Gini(T)=1- ((5/10)^2 +(5/10)^2)= 0.5$

Random Forest

Após um split, o Índice Gini pode ser calculado como :

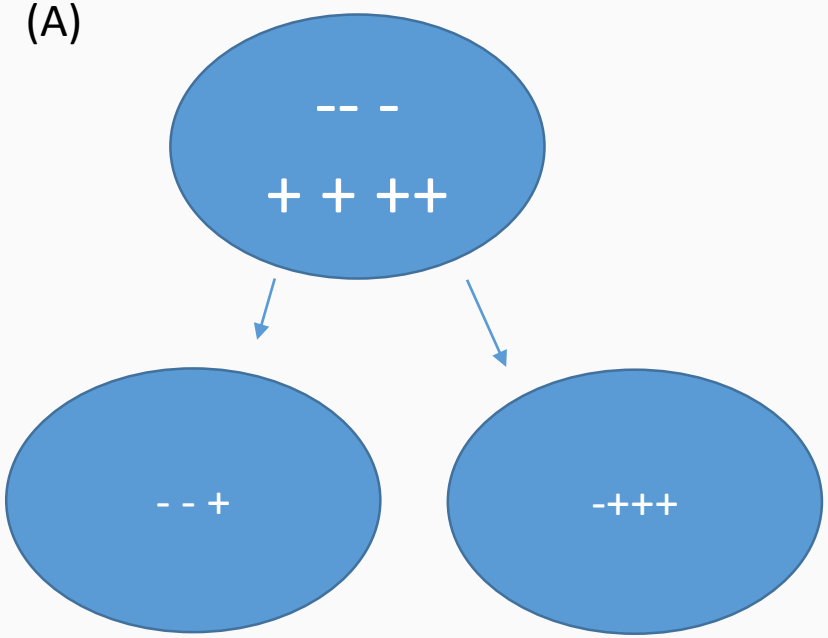
$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

Onde N1 e N2 são a quantidade de elementos em cada “branch” e N=N1+N2

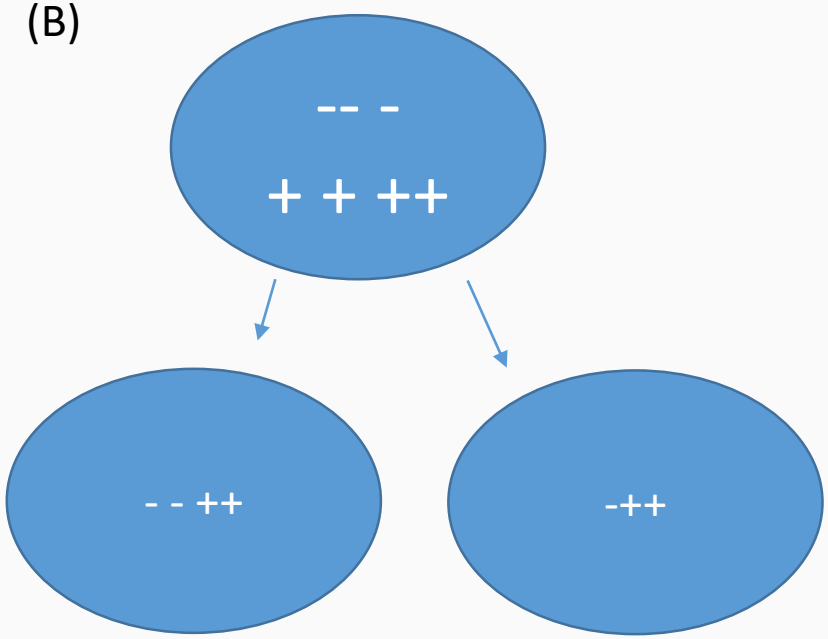
Random Forest

Exemplo do cálculo do índice Gini após dois splits (A) e (B)

(A)



(B)



Gini A : 3/7*(1-(2/3**2+1/3**2)) + 4/7*(1-(1/4**2 + ¾**2))=0.404

Gini B : 4/7*(1-(1/2**2+1/2**2)) + 3/7*(1-(1/3**2 + 2/3**2))=0.476

Assim, split A reduz mais a impureza que o B...A é melhor.

Random Forest

Conteúdo

Prevendo inadimplência em função de salário, contrato e tempo de serviço (**árvore de decisão**)

Partindo de decision_tree.ipynb

Atividade 1 : Analisar o código..o algoritmo é CART com índice gini

Atividade 2 : Alterar a entrada para que a feature salário tenha importância...salário maior, menos inadimplência

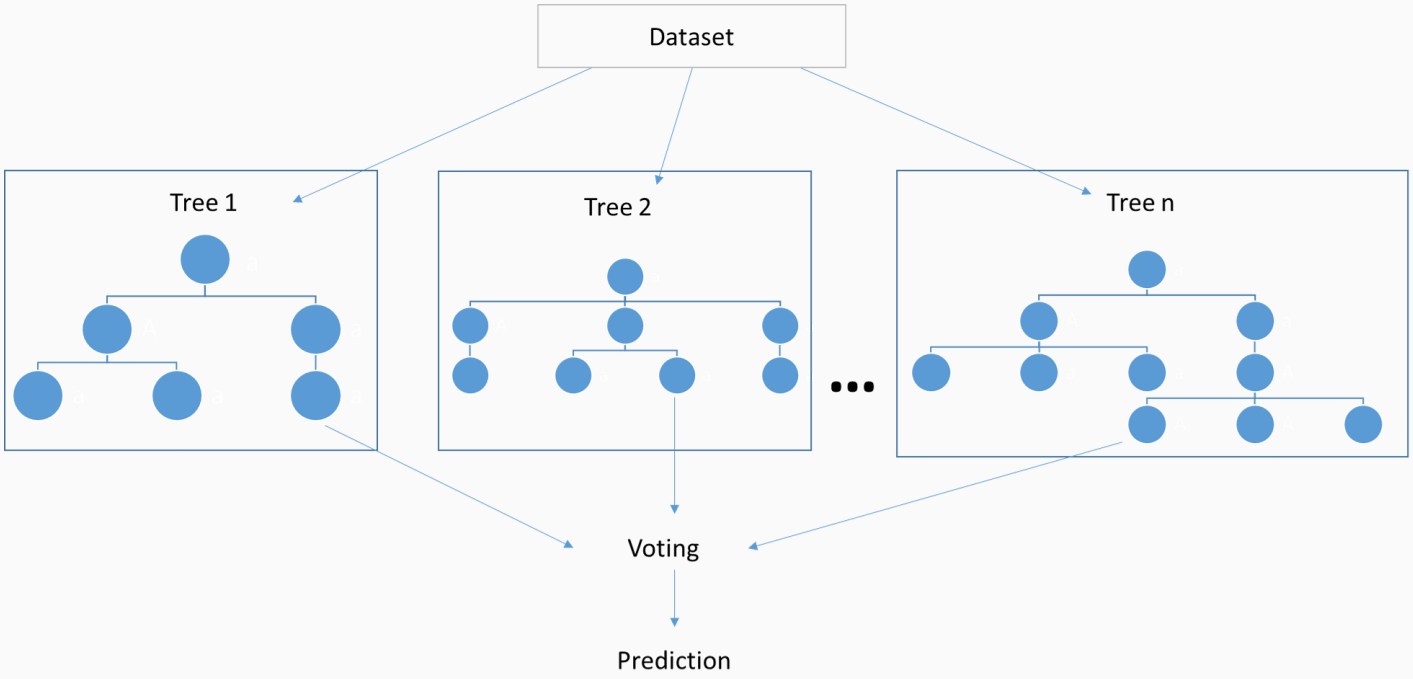
Atividade 3 : imponha um só nível e comprove a importância das features

A feature importance na decision tree pode ajudar na seleção de faetures do dataset

Random Forest

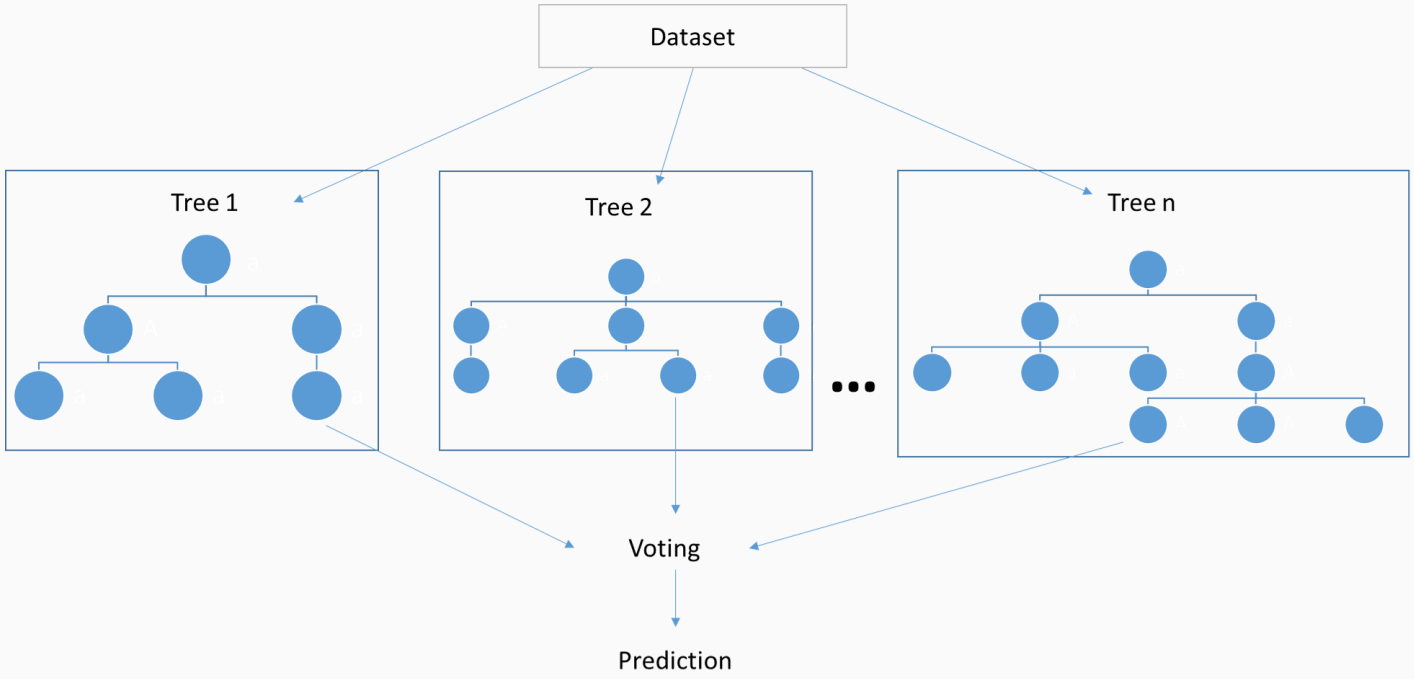
Conteúdo

Random Forest é um ensemble method que utiliza bagging e tem como preditores “Árvores de decisão”...



Random Forest

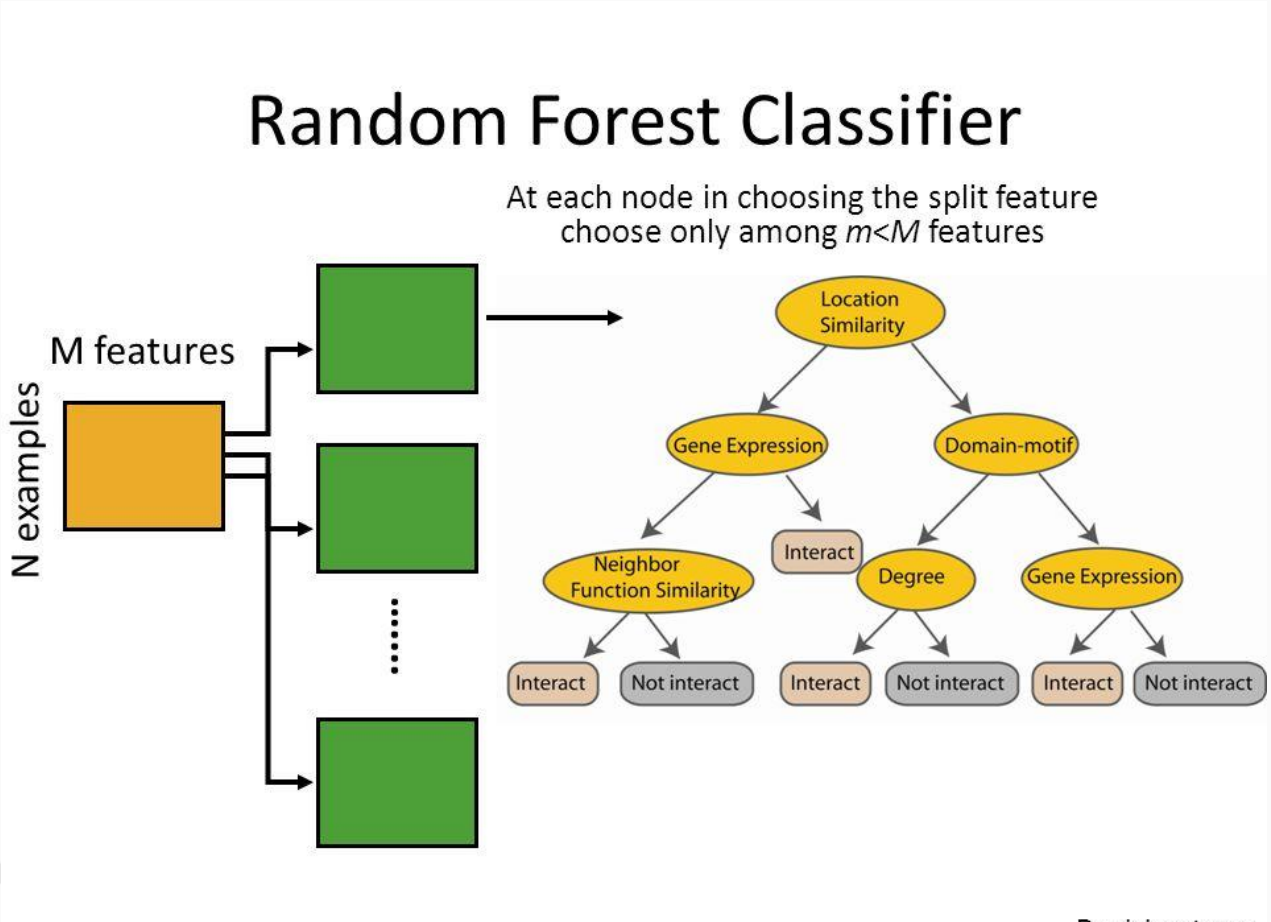
Cada árvore utiliza uma seleção aleatória das amostras (com reposição)



Random Forest

Conteúdo

Em cada split, apenas uma parte das features é considerada..há várias estratégias para isso... A ideia é aumentar o número de variações entre os preditores (várias "opiniões" diferentes para o cálculo consolidado (média)...isso melhora a variância com provável leve aumento de bias



Random Forest

Conteúdo

Com base em random_forest.ipynb

Classificação binária : Virgínica ou não

1)Analisar o código

2)Split 70 30

3)Calcular acurácia no treino e no teste

4)Veja acurácias de treino e teste para classificador binário Versicolor ou não

Cursos com Alta Performance de Aprendizado