# SCC0276 - Aprendizado de Máquina

## Aula Random Forest

Profa. Dra. Roseli Aparecida Francelin Romero
SCC - ICMC - USP

2019

# Sumário

# O que é Random Forest

- suponha que é hora do jantar e sua mãe pergunta a você e ao seu irmão o que ela deve cozinhar para o jantar, para obter uma resposta de ambos, ela pede que vocês dois joguem sim e não.
- Ela começa fazendo perguntas como deveria ser o jantar: algo assado ou frito? Ela investiga mais perguntando se deve conter legumes?
- Ao fazer perguntas desta maneira decisiva, ela está tentando combinações diferentes para chegar a uma conclusão sobre o tipo de prato que ela deve fazer e como resultado dessas perguntas, ela é capaz de chegar a uma conclusão do prato final para preparar.

# O que é Random Forest

- A técnica Random Forest (RF) funciona da mesma maneira, onde você e seu irmão são árvores em uma floresta, suas preferências alimentares são amostras diferentes de dados e, finalmente, o prato é a decisão que a RF faz.
- Definição Formal: RF é um poderoso método de *ensemble* usada para construir modelos preditivos através da construção de um grande número de árvores não correlacionadas para resolver o problema do overfitting, que envolve a média da saída do preditor de diferentes modelos de árvore que são distribuídos ajudando a reduzir a variancia.

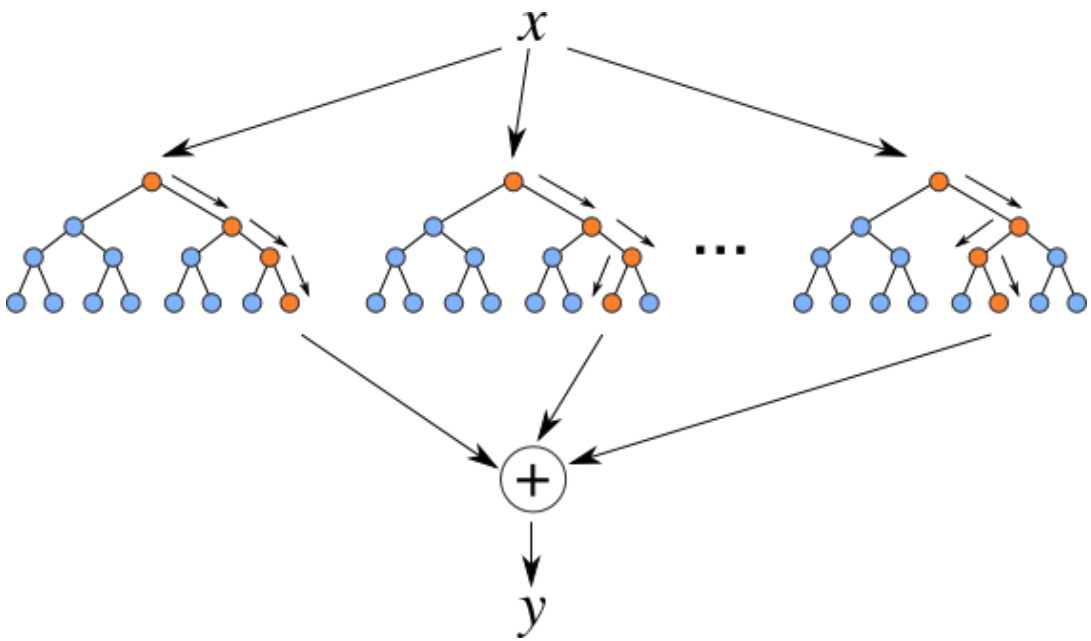
# O que é Random Forest

- O Algoritmo de RF é composto de diferentes árvores de decisão, cada uma com os mesmos nós, mas usando dados diferentes que levam a folhas diferentes. Ele mescla as decisões de várias árvores de decisão para encontrar uma resposta, que representa a média de todas essas árvores de decisão.

# O que é Random Forest

- A Random Forest cria uma coleção de árvores não-correlacionadas usando uma seleção aleatória de características, e dados e, em seguida, calcula a média deles, o que reduz a variancia.
- Aprendizado Supervisionado

# Exemplo




- O algoritmo usa as folhas, ou decisões finais, de cada nó para chegar a uma conclusão própria.
- Isso aumenta a precisão do modelo, pois ele observa os resultados de muitas árvores de decisão diferentes e encontra uma média.

# Sumário

# Vantagens: Random Forest

- é um algoritmo rápido para treinar
- RF pode ser usada tanto para classificação e regressão.
- overfitting (controversia)

# Desvantagens: Random Forest

- Lento na criação de previsões, uma vez que o modelo é feito.
- Deve ter cuidado com outliers e buracos nos dados.

# Sumário

# Exemplo de Regressão

- Supor que desejamos estimar a renda familiar média em sua cidade. Pode-se facilmente encontrar uma estimativa usando o Algoritmo RF.
- Deve-se começar distribuindo questionários pedindo que as pessoas respondam a várias perguntas diferentes. Dependendo de como eles responderam a essas perguntas, uma renda familiar estimada seria gerada para cada pessoa.

# Exemplo de Regressão

- Depois de encontrar as árvores de decisão de várias pessoas, pode-se aplicar o Algoritmo RF a esses dados.
- Deve-se examinar os resultados de cada árvore de decisão e usar a técnica RF para encontrar uma renda média entre todas as árvores de decisão.
- A aplicação desse algoritmo fornecerá uma estimativa precisa da renda familiar média das pessoas pesquisadas.

# Exemplo de Classificação

- Supor que uma nova empresa deseja saber que tipo de pessoa pode comprar seus produtos.
- Começar apresentando um questionário a um grupo de pessoas no mesmo mercado-alvo, com uma série de perguntas sobre seus comportamentos de compra e o tipo de produto que preferir.
- Com base em suas respostas, pode-se classificá-las como um cliente em potencial ou não como um cliente em potencial.
- Aplicar o RF. Se o algoritmo concluir que a maioria das pessoas nesse mercado-alvo não são clientes em potencial, pode ser uma boa ideia para a empresa repensar seus produtos com esses tipos de pessoas em mente.

# Sumário

# Matemática da Classificação com RF

- Ao executar RF com base em dados de classificação, você deve saber que está usando frequentemente o índice de Gini ou a fórmula usada para decidir como os nós em uma ramificação da árvore de decisão.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

- Essa fórmula usa a classe e a probabilidade para determinar o Gini de cada ramificação em um nó, determinando qual ramificação é mais provável de ocorrer. $p_i$ representa a frequência relativa da classe que se está observando no conjunto de dados e c representa o número de classes.

# Matemática da Classificação com RF

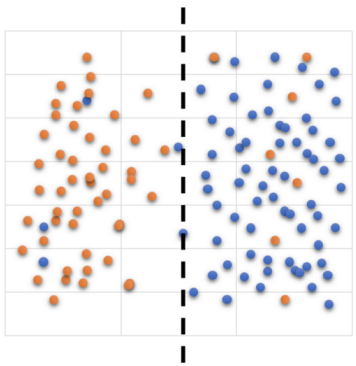
- Pode-se também usar a entropia para determinar como os nós se ramificam em uma árvore de decisão.

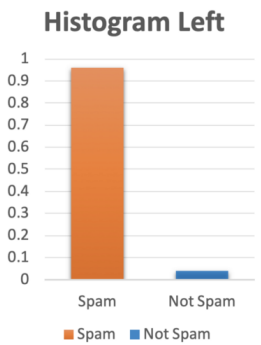
$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

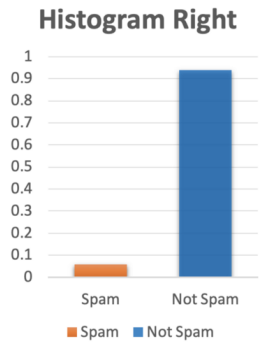
# Criterio de corte para separar os dados

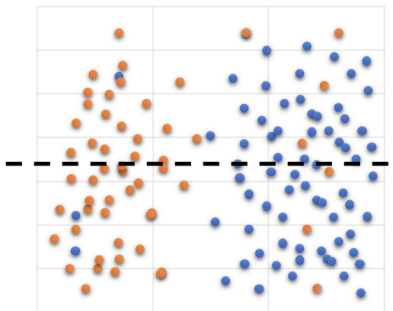
- Suponha que temos dados de SMS que podem ser ou nao spam, no plano.
- Quando cada árvore é construída, nós são criados de forma a separar os pontos de uma maneira que divida esses pontos em duas classes, produzindo uma distribuição homogênea.
- Isso é feito tomando alguns pontos de dados aleatórios e selecionando o limiar de divisão (coordenada x-y) com base no ponto de dados que maximiza o Ganho de Informações. Essas divisões podem ser verticais, horizontais ou lineares dependendo da distribuição de dados.

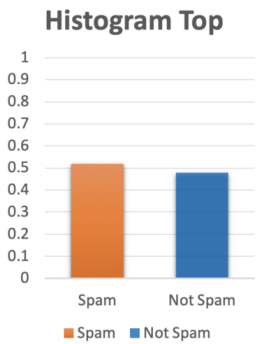
# Criterio de corte para separar os dados

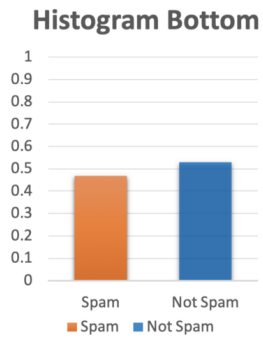




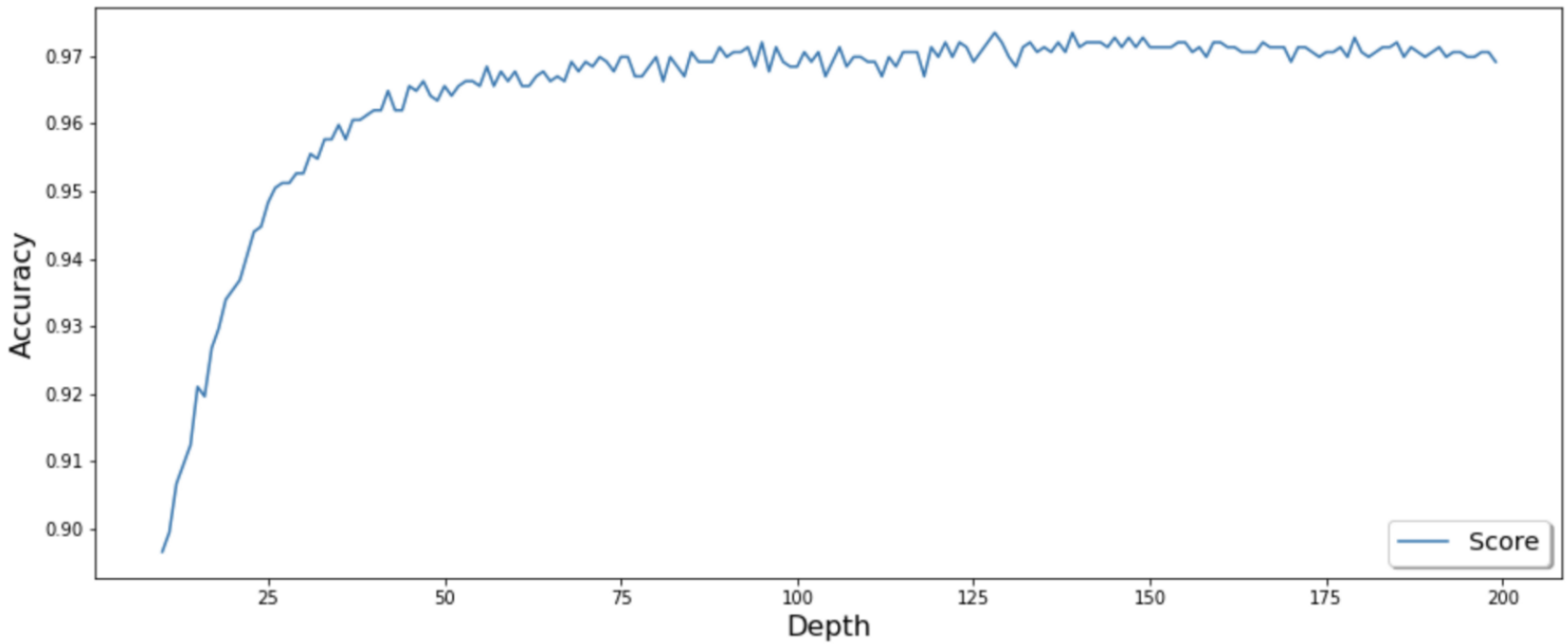



# Dados Importantes: Profundidade D (maxdepth=D)

- random_forest_classifier = RandomForestClassifier(n_estimators=31, random_state=100, max_depth=D)
- Se o valor do parâmetro de profundidade máxima D for grande, o modelo tende a sofrer de ajuste excessivo, enquanto que, quando o valor é pequeno, produz baixa confiança do modelo, resultando em ajuste insuficiente. Portanto, é preciso selecionar cuidadosamente um valor ideal de D.

# Dados Importantes



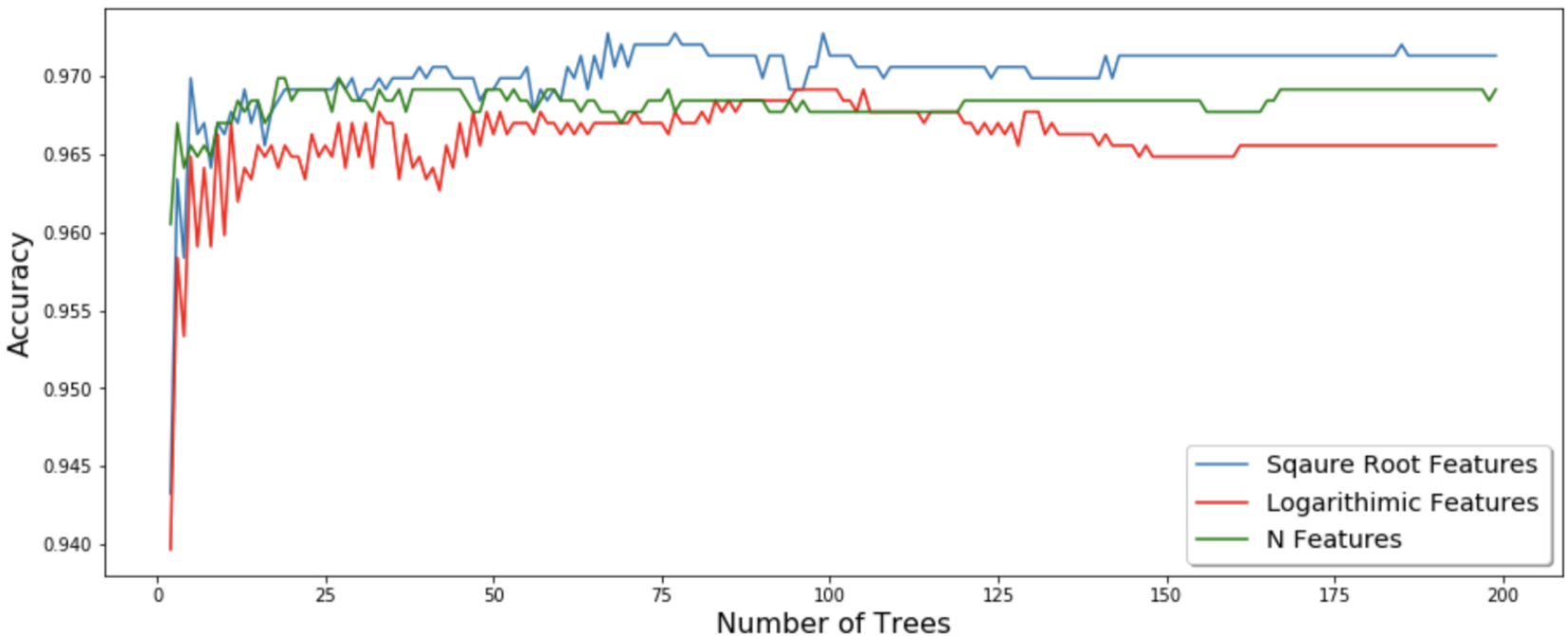
# Dados Importantes: Número de árvores

- Geralmente,quanto maior o número de árvores, melhor a precisão.
- random_forest_classifier = RandomForestClassifier(n_estimators=T, random_state=100)
- Não há penalidade por usar mais árvores, mas é computacionalmente caro.
- Uma maneira de escolher o número de árvores (T) é olhar para a estimativa de erro, quando as árvores estão sendo adicionadas à floresta e pará-la quando a taxa de erro começar a se estabilizar.

# Dados Importantes: Maximo de Features

- random_forest_classifier = RandomForestClassifier(n_estimators=31, random_state=100, max_features=F)
- Esse parâmetro especifica o tamanho de subconjuntos aleatórios de features ao procurar a melhor divisão em cada nó.
- Aumentar muito o no. de features pode melhorar o desempenho no conjunto de treinamento, mas resultará em uma correlação entre as árvores que estamos tentando evitar.
- O valor ideal (F), que é geralmente escolhido como a raiz quadrada do número total de features, pode produzir bons resultados no desempenho, anti-correlação entre árvores e redução de variância.

# Dados Importantes: Maximo de Features



# References


- https://medium.com/sfu-big-data/demystifying-random-forest-1ed89e335fb
- T. Hastie, R. Tibshirani, and J. Friedman. Springer Series in The Elements of Statistical Learning, USA, (2008)