MACHINE LEARNING E DATA SCIENCE COM R DE A À Z

INSTRUTOR: JONES GRANATYR [MESTRADO E DOUTORADO EM IA]

MINHAS ANOTAÇÕES DE AULA # SÃO ANOTAÇÕES QUE FAREM AO LONGO DO CURSO

PROF. CÍCERO QUARTO - CICERO@ENGCOMP.UEMA.BR
DEPARTAMENTO DE ENGENHARIA DA COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DO MARANHÃO (UEMA)

IA (INTELIGÊNCIA ARTIFICIAL) -> PARA O INSTRUTOR, IA É O CAMPO DE TI QUE + ESTÁ CRESCENDO ATUALMENTE

POR QUE DE A À Z?
- A: VAI DE CONTEÚDOS + BÁSICOS INTRODUTÓRIOS DE MACHINE LEARNING
- Z: ATÉ CONCEITOS + AVANÇADOS USANDO A LINGUAGEM R

1. CONTEÚDO DO CURSO

~~CONTEÚDO~~

INTRODUÇÃO # CONCEITOS BÁSICOS SOBRE APRENDIZAGEM DE MÁQUINA

PARTE 1 - CLASSIFICAÇÃO

PRÉ-PROCESSAMENTO, NAIVE BAYES, ÁRVORES DE DECISÃO, REGRAS, INSTÂNCIAS, REGRESSÃO LOGÍSTICA, MÁQUINAS DE VETORES DE SUPORTE, REDES NEURAIS ARTIFICIAIS, AVALIAÇÃO DE ALGORITMOS, COMBINAÇÃO E REJEIÇÃO DE CLASSIFICADORES, BASE DE DADOS BOM OU MAL PAGADOR, BASE DE DADOS PARA PREVISÃO DA CLASSE DO SALÁRIO.

PARTE 2 - REGRESSÃO

REGRESSÃO LINEAR, REGRESSÃO POLINOMIAL, ÁRVORES DE DECISÃO, RANDOM FOREST, VETORES DE SUPORTE, REDES NEURAIS ARTIFICIAIS, BASE DE DADOS PLANO DE SAÚDE, BASE DE DADOS PREVISÃO DO PREÇO DE CASAS.

PARTE 3 - REGRAS DE ASSOCIAÇÃO

ALGORITMO APRIORI, ALGORITMO ECLAT, BASE DE DADOS DO MERCADO

PARTE 4 - AGRUPAMENTO

K-MEANS, AGRUPAMENTO HIERÁRQUICO, DBSCAN, BASE DE DADOS DE GASTOS NO CARTÃO DE CRÉDITO.

PARTE 5 - TÓPICOS COMPLEMENTARES

Observação: O instrutor ressalta que é recomendável fazer primeiro a parte 1 e só depois fazer a parte 2, mesmo que o aluno queira pular diretamente para a parte 2. Ou seja, o instrutor destaca que é melhor entender classificação para depois passar para regressão.

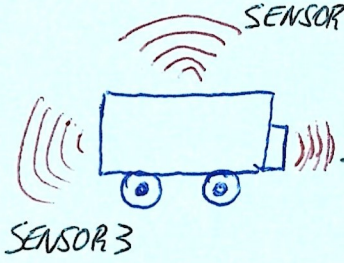
Pré-requisitos

- Conhecimentos sobre lógica de programação, principalmente estruturas condicionais e de repetições;
- Conhecimentos básicos em R são desejáveis, embora seja possível acompanhar o curso sem saber essa linguagem em profundidade;
- Conhecimentos básicos sobre instalação de softwares básicos;
- Nível: todos os níveis

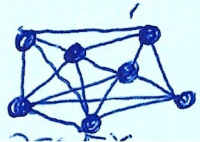
APLICAÇÕES DE MACHINE LEARNING

- RECONHECIMENTO FACIAL

- KINECT (MICROSOFT) # REPRODUÇÃO DE MOVIMENTOS POR CAPTURA DE IMAGENS
  - ↳ O COMPUTADOR VAI (VIA CÂMERA) APRENDER COMO VOCÊ JOGA E AÍ ELE PODE EFETUAR UMA JOGADA SUA, OU SEJA, MARCAR UM GOL DE CABEÇA AO COBRAR UM ESCANTEIO.
- HUMANOIDES DO GOOGLE → DETECTAM PERCEPÇÕES AO SEU REDOR PARA AGIR SOBRE O AMBIENTE
- CARROS AUTÔNOMOS → USAM SENSORES PARA DETECTAR PERCEPÇÕES E AGIR SOBRE O AMBIENTE

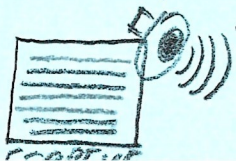


- ROBÔ MÉDICO → NOVOS MEDICAMENTOS, PREVER DOENÇAS NO FUTURO
- COLABORAÇÃO → REDE DE COMUNICAÇÃO COLABORATIVA PARA SISTEMAS DE RECOMENDAÇÃO. EX. NETFLIX RECOMENDA FILMES AFINS PARA ATENDER TENDÊNCIAS DE GOSTOS



RECOMENDAR MÚSICAS BASEADO NO PERFIL DO CLIENTE (USUÁRIO) QUE ELE TEM MAIS TENDÊNCIA DE GOSTAR;

- HARMONY → SITE SÓ DE ENCONTROS. VOCÊ FAZ O CADASTRO DE SUAS CARACTERÍSTICAS PESSOAIS, FÍSICAS, SOCIAIS, GEOGRÁFICAS, DENTRE OUTRAS PARA ENCONTRAR PARCEIROS OU PARCEIRAS AFINS COM O SEU PERFIL E GOSTOS, OBJETIVANDO QUE O RELACIONAMENTO DER CERTO
- AMAZON.COM → QUEM COMPROU O PRODUTO A, TAMBÉM COMPROU O PRODUTO B E/OU C E AÍ SUGERE A RECOMENDAÇÃO PARA VOCÊ
- NASA → PARA EXPLORAR UM PLANETA DESCONHECIDO, ENVIA UM ROBÔ E ESTE VAI REALIZAR VÁRIAS AÇÕES (DETECTAR OBSTÁCULOS, TIRAR FOTOS DO AMBIENTE,)
- FACEBOOK ADS → ANÚNCIOS DO FACEBOOK, PERSONALIZANDO PARA AS PESSOAS QUE QUEREM OU PRETENDEM COMPRAR UM DETERMINADO PRODUTO
- TEXT-SPEAK → ESCREVE → SAI A TRADUÇÃO EM SOM → SOM (REPRODUÇÃO DO TEXTO ESCRITO EM SOM)



OU AO CONTRÁRIO → STEPHEN HAWKING

- CIA → CENTRAL INTELLIGENCE AGENCY
  - ↳ USAM ALGORITMOS DE APRENDIZAGEM DE MÁQUINA PARA DETECTAR QUEM TEM OU NÃO CHANCE DE SER UM TERRORISTA, CRIMINOSO
- REALIDADE VIRTUAL (RV) → USO DE ÓCULOS PARA REPRODUZIR A REALIDADE.
- GOOGLE → USA ALGORITMOS DE APRENDIZAGEM DE MÁQUINA PARA AUTOMATIZAR BUSCAS

ALGUMAS NOTÍCIAS

SOFTWARES PARA INFECTAR O COMPUTADOR ← MALICIOUS SOFTWARE (SOFTWARES MALICIOSOS)

- COMO A IA PODE AJUDAR PREVENIR MALWARE

VÍRUS, WORMS, CAVALOS DE TROIA, SPYWARE, ETC...

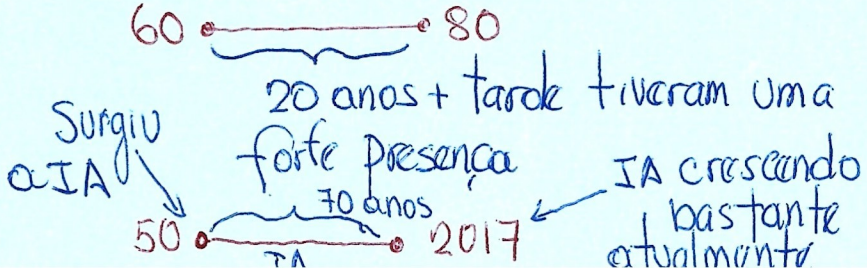
- SITE PORNOGRÁFICO VAI USAR IA PARA IDENTIFICAR ATRIZES E ATORES
- IA PODE AJUDAR A REDUZIR ROUBO DE ELETRICIDADE NO BRASIL
- IA TENTA DESCOBRIR QUEM TRAIU ANNE FRANK
- GOOGLE USA IA PARA ENCONTRAR VÍRUS NA PLAY STORE
- TESLA E AMD ESTÃO CRIANDO CHIP DE IA PARA CARROS AUTÔNOMOS (TENDÊNCIA EM ALTA)
- MICROSOFT PIX EDITA IMAGENS ATRAVÉS DE IA;
- IA DETECTA MAL DE ALZHEIMER UMA DÉCADA ANTES DE SINTOMAS APARECEREM

- IA ESTÁ ESCREVENDO O FIM DE GAME OF THRONES
- IA CONSEGUE RECRIAR FASES DE SUPER MÁRIO;
- ROBÔ FAZ EM SEGUNDOS O QUE DEMORAVA 360 MIL HORAS PARA UM ADVOGADO;
- IA, UMA DAS TENDÊNCIAS QUE REVOLUCIONAM O SERVIÇO DE ATENDIMENTO AO CLIENTE
- CANTORES ESTÃO COMPONDO UM ALBUM INTEIRO USANDO IA

PARA FINALIZAR O TÓPICO DE AULA SOBRE APLICAÇÕES DA IA, O INSTRUTOR RESGATOU UMA MATÉRIA DA REVISTA VEJA INTITULADA "DE MÃOS DADAS COM A INTELIGÊNCIA ARTIFICIAL"

Por que aprender Machine Learning?

1. O instrutor fez um resgaste da linha do tempo dos computadores desde quando surgiram na década de 60, com expansão na década de 80.



Redes Neurais + Deep Learning => Forte tendência de uso

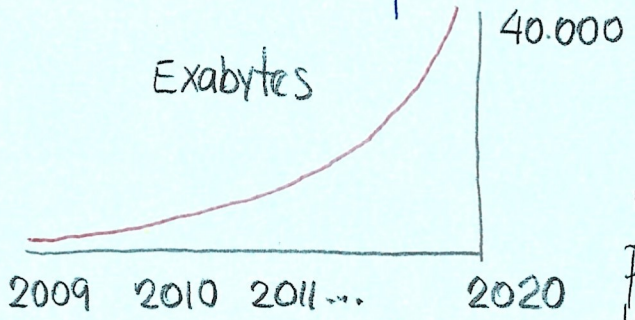
1956 → Um HD de Computador precisava ser transportado de Avião, pois era muito grande e pesado

1980 → HD de 10 Megabyte, entranto <, $3,495

2017 → SeanDisk Ultra de 256GB, entretanto + caro

Storage Limits # Neste item, o instrutor resgatou a linha do tempo da capacidade de armazenamento de dados baseado em HD (Hard Disk), Flash memory e Bacterial DNA.

*Exabytes > Capacidade de armazenamento



Moral da História: temos cada vez + dados e menos capacidade de profissionais para manipular esses dados... e aí entra o Machine Learning para ajudar a manipular esses dados.

Empresas Brasileiras que precisam de Profissionais De Análise de dados e de Machine Learning.

- Itaú e/ou qualquer outro banco estão desenvolvendo aplicações que usam Machine Learning [BRÁCIS;
- IBM #Contém Laboratório de estudos nas áreas:
  - Gestão Inteligente de Recursos Naturais (petróleo, mineração, agricultura);
  - Computação Cognitiva com ênfase em análise de dados sociais;
  - Big-Data;
  - Visual analytics e Comprehension;
  - Ciência e Tecnologia para Aplicações Industriais voltada para [redacted] recursos naturais, ciências da vida e internet das Coisas.

Lab da IBM no Brasil
↓
atua de forma integrada c/ a rede global da IBM Research, formada por 12 Laboratórios no mundo onde trabalham mais de 3 mil cientistas.

Ibm.biz/brazilresearchlab

Deep Learning ↓

- NVIDIA #Trabalha muito c/ RNAs e DL
  - GPU (Unidade de processamento Gráfico)

↓

Disponibilizam vagas no LinkDin nas áreas de IA → Machine Learning

Algumas Notícias

- Empresas recorrem à IA para melhorar negócios;
- A$ 100 startups que Lideram a revolução da IA;
- Salesforce vai investir U$50 milhões em IA;
- Brasil está atrasado em IA;
- A IA no centro da guerra dos smartphones;
- Intel trabalha com Facebook no projeto de chip para IA

- Brainwave é o novo projeto para IA da MS, sendo que a MS (MicroSoft) abriu no Brasil um Laboratório somente de IA;
- Ainda estamos na Internet discada de onde a IA chegará;
- China quer se tornar líder em IA até 2025;
- Mastercard adquire empresa que desenvolve software de IA;
- U$300 a U$500 mil dólares?? # Um profissional de IA, nos EUA, pode ganhar isso/ano

TERMINOLOGIA DA ÁREA DE IA

- **INTELIGÊNCIA ARTIFICIAL (IA)**
  - **SISTEMAS ESPECIALISTAS** → CONSIDERADOS UM DOS PRIMEIROS ESFORÇOS DA ÁREA DE IA QUE CONSISTE DA CONSTRUÇÃO DE SISTEMAS DE BASE DE CONHECIMENTO BASEADO NO CONHECIMENTO DE UM ESPECIALISTA HUMANO → PEGA TODO O CONHECIMENTO QUE O ESPECIALISTA TEM E JOGA PARA O SISTEMA COMPUTACIONAL, PARA ESTE SER CAPAZ DE DAR AS RESPOSTAS A PERGUNTAS FEITAS
  - **VISÃO COMPUTACIONAL** → SIMULAR A VISÃO HUMANA PARA DETECÇÃO DE OBJETOS, RECONHECIMENTO FACIAL. TEM APLICAÇÕES NAS ÁREAS DE SEGURANÇA, ROBÓTICA, ETC...
  - **MACHINE LEARNING** → APRENDIZAGEM DE MÁQUINA, CONSISTINDO DE VÁRIOS ALGORITMOS QUE VÃO FAZER OS COMPUTADORES A APRENDEREM ATRAVÉS DE UMA BASE DE DADOS
  - **PROCESSAMENTO DE LINGUAGEM NATURAL** ENVOLVE BASICAMENTE O COMPUTADOR ENTENDER A SEMÂNTICA DE UMA FRASE, TANTO ESCRITA COMO FALADA (INTERPRETAÇÃO DE TEXTO
  - **ALGORITMOS GENÉTICOS** →
  - **SISTEMAS MULTIAGENTES**
  - **MINERAÇÃO DE DADOS** # O INSTRUTOR RESSALTA QUE ~~EXISTE~~ ↓ EXISTE UMA ≠ SUTIL ENTRE MACHINE LEARNING E MINERAÇÃO DE DADOS.
  - **RESOLUÇÃO DE PROBLEMAS POR MEIO DE BUSCA** EXEMPLOS, COMO ROTAS (GOOGLE MAPS) PARA ENCONTRAR MELHOR CAMINHOS
  - **LÓGICA FUZZY (NEBULOSA)** → TEM MUITA APLICAÇÃO NA ÁREA INDUSTRIAL
  - **RACIOCÍNIO BASEADO EM CASOS**
  - **REDES NEURAIS**

▶▶ ROBÓTICA → PODE SER CONSIDERADA UMA ÁREA INDEPENDENTE, ENTRETANTO PARA FAZERMOS APLICAÇÕES MAIS INTERESSANTE VAI TER QUE INTEGRAR IA NESSES ROBORES

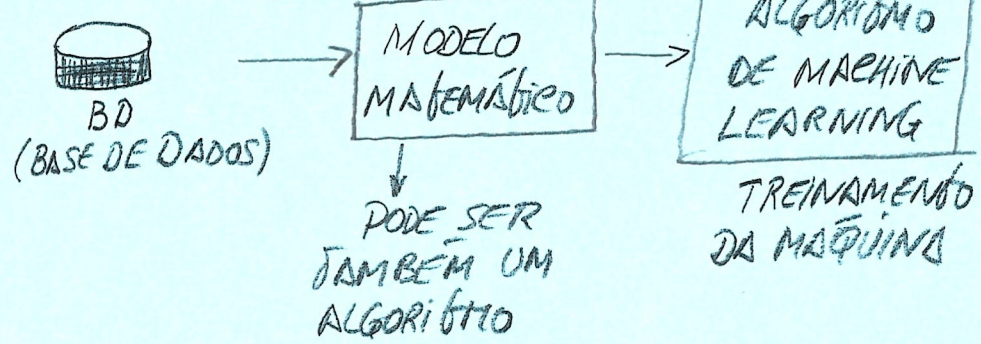
- INTELIGÊNCIA ARTIFICIAL (IA)
  - ÁREA DA CIÊNCIA DA COMPUTAÇÃO RESPONSÁVEL PELO DESENVOLVIMENTO DE SISTEMAS QUE SIMULEM A CAPACIDADE HUMANA DE RESOLVER PROBLEMAS
  - TERMO + GERAL
- INTELIGÊNCIA COMPUTACIONAL
  - REDES NEURAIS, COMPUTAÇÃO EVOLUCIONÁRIA, LÓGICA NEBULOSA (FUZZY)
  - BOM CANDIDATO PARA SUBSTITUIR O TERMO "IA"

↗ TEMOS NO BRASIL O EVENTO ENIAC (ENCONTRO NACIONAL DE # IA E COMPUTACIONAL.

NOTA DO INSTRUBOR } O INSTRUBOR FRISA QUE MUITOS CIENTISTAS DA COMPUTAÇÃO NÃO GOSTAM E/OU EVITAM USAR O TERMO IA, POIS ESSES ACREDITAM QUE ESSE TERMO NÃO REGRESSA OU NÃO TRAZ NADA DE COMPUTAÇÃO EM "INTELIGÊNCIA ARTIFICIAL". O INSTRUBOR PONTUA QUE É UMA CONFERÊNCIA DO IEEE QUE É ESPECÍFICA EM "INTELIGÊNCIA COMPUTACIONAL" E ESTA SÓ ACEITA TRABALHOS QUE ABORDAM SOBRE RNA, COMPUTAÇÃO EVOLUCIONÁRIA, LÓGICA NEBULOSA (FUZZY).

- MACHINE LEARNING (APRENDIZAGEM DE MÁQUINA)
  - MÉTODOS MATEMÁTICOS PARA TREINAR ALGORITMOS.

ILUSTRAÇÃO



- DATA MINING (MINERAÇÃO DE DADOS)
  - EXTRAIR CONHECIMENTO DE BASE DE DADOS, USANDO MÉTODOS DE APRENDIZAGEM DE MÁQUINA
- REDES NEURAIS
  - É UM TIPO DE APRENDIZAGEM DE MÁQUINA É UM ALGORITMO QUE PODE SER UTILIZADO PARA FAZER O COMPUTADOR APRENDER
- DEEP LEARNING (APRENDIZADO PROFUNDO)
  - MUITO MAIS DADOS E PROCESSADORES MAIS POTENTES
  - REDE NEURAL COM MUITAS CAMADAS

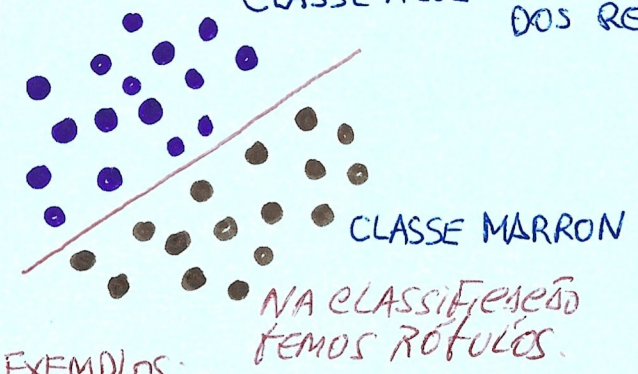
- Big Data # demanda de algoritmos de Machine Learning para extrair conhecimento dos dados
  - Imenso volume de dados
- Ciência de Dados
  - Exploração e análise de dados
  - Ciência da Computação + Estatística
  - Machine Learning

MÉTODOS PREDITIVOS

Aprendizagem de Máquina (MACHINE LEARNING)

Métodos Preditivos	Métodos Descritivos
Classificação	Associação
Regressão	Agrupamento
	Detecção de desvios
	Padrões sequenciais
	Sumarização

CLASSIFICAÇÃO # DEFINE CLASSES PARA CADA UM DOS REGISTROS

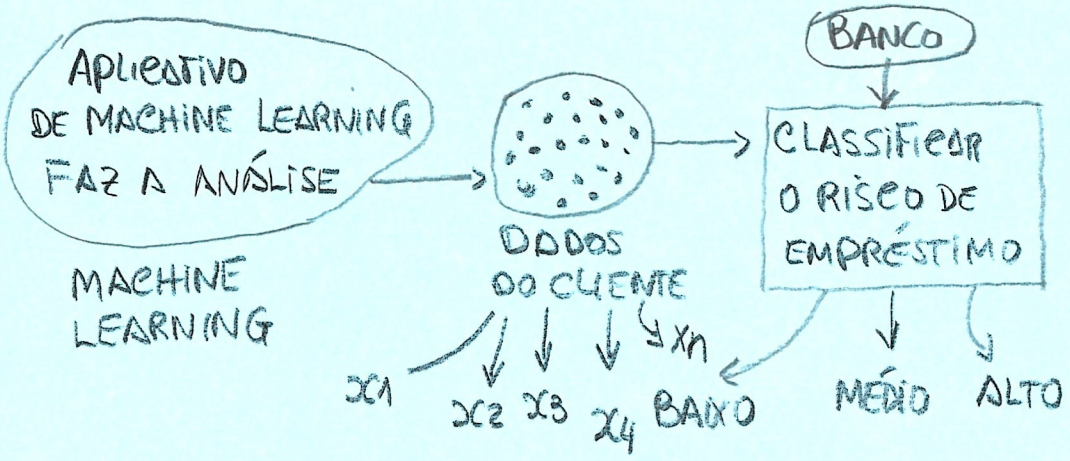


EXEMPLOS:

- MARKETING DIRETO → VOCÊ TEM UM NOVO PRODUTO E NÃO SABE PARA ONDE MANDAR?

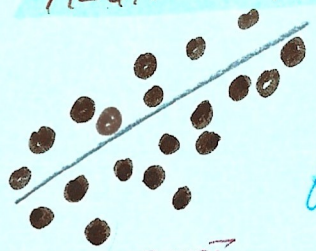
  O PROBLEMA É CLASSIFICAR SE A PESSOA VAI COMPRAR OU NÃO O PRODUTO

- INSATISFAÇÃO DE CLIENTES → PODEMOS CLASSIFICAR SE O CLIENTE ESTÁ SATISFEITO OU INSASTISFEITO COM ALGUM PRODUTO, COM O ATENDIMENTO, PRESTAÇÃO DE SERVIÇOS
- RISCO DE CRÉDITO → BASTANTE UTILIZADO PELOS BANCOS PARA SABER SE O CLIENTE APRESENTA RISCO BAIXO, MÉDIO OU ALTO PARA LIBERAR UM EMPRÉSTIMO.



- FILTROS DE SPAM → UTILIZADOS EM TODOS OS SISTEMAS DE E-MAIL, PEGANDO UMA MENSAGEM E CLASSIFICANDO ESTA COMO SPAM, PROMOÇÕES, SOCIAIS, ... OU SEJA EM PASTAS CORRESPONDENTES.
- SEPARAÇÃO DE NOTÍCIAS → DADO VÁRIOS TEXTOS OU DOCUMENTOS PODEMOS SEPARAR AS NOTÍCIAS EM ESPORTE, POLÍTICA, SAÚDE, TECNOLOGIAS, ECONOMIA, ...
- RECONHECIMENTO DE VOZ: UMA VOZ PERTENCE A UMA PESSOA A, B, C, ...
- RECONHECIMENTO DE FACE: UMA DETERMINADA FACE DE UM SISTEMA DE SEGURANÇA PERTENCE À PESSOA A, B, C, ...
- PREVISÃO DE DOENÇAS → BASEADO NOS DADOS DO PACIENTE FAZER UMA PREVISÃO SE O PACIENTE VAI DESENVOLVER OU NÃO VAI DESENVOLVER UMA DETERMINADA DOENÇA.

REGRESSÃO → FAZER PREVISÃO DE VALORES NUMÉRICOS. POR EXEMPLO, QUANTO UMA EMPRESA GANHARIA DE LUCRO, BASEADO NOS DADOS HISTÓRICOS



NA REGRESSÃO TEMOS NÚMEROS

QUANTITATIVA ← VARIÁVEL NUMÉRICA

→ TUDO DEPOIS DA SETA SÃO NÚMEROS.

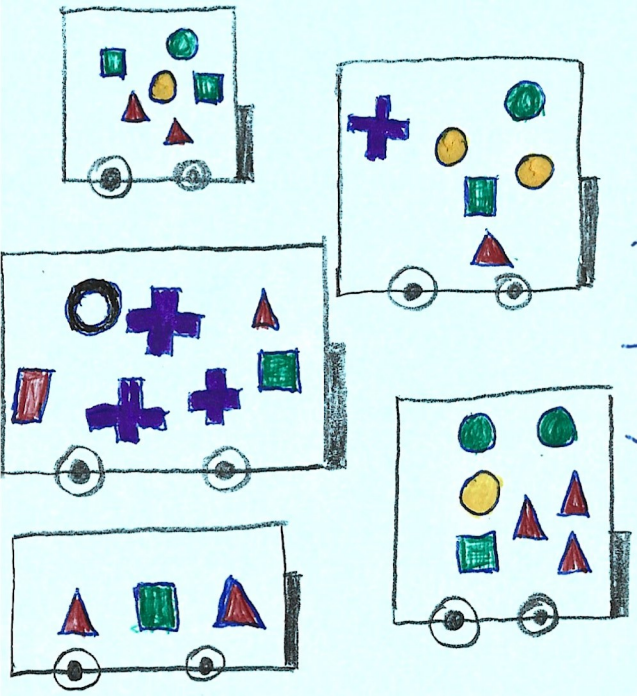
ALGUNS EXEMPLOS:

- GASTOS PROPAGANDA → VALOR DE VENDA
- TEMPERATURA, UMIDADE E PRESSÃO DO AR → VELOCIDADE DO VENTO
- FATORES EXTERNOS → VALOR DO DÓLAR ($)
- RESULTADOS DO EXAME → PROBABILIDADE DE UM PACIENTE SOBREVIVER
- RISCO DE INVESTIMENTO
- GASTOS NO CARTÃO DE CRÉDITO, HISTÓRICO → LIMITE DO CARTÃO. OU SEJA, O CLIENTE NÃO PRECISA PEDIR PARA AUMENTAR SEU LIMITE DE CRÉDITO, POIS BASEADO EM SEU HISTÓRICO DE COMPRAS E PAGAMENTO DAS FATURAS, ISSO É PREVISTO RECONHECIDO DE REGRESSÃO.
- Valores anteriores → Valores de produtos

MÉTODOS DESCRITIVOS

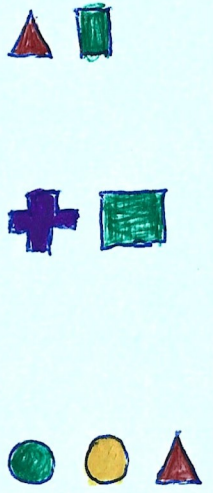
- Associação
- Agrupamento
- Detecção de Desvios
- Padrões Sequenciais
- Sumarização

ASSOCIAÇÃO



CARRINHOS DE COMPRAS DE SUPERMERCADO

OBJETIVO É ENCONTRAR ALGUM TIPO DE ASSOCIAÇÃO



REGRAS DE ASSOCIAÇÃO (RA)

$RA_S = \{RA1, RA2, RA3, RA4, RA5, \ldots RA_N\}$

RA1: QUEM COMPRA UM PRODUTO ▲ → COMPRA TAMBEM ■

RA2: QUEM COMPRA UM PRODUTO ● E ● → COMPRA TAMBEM ▲

APLICAÇÕES DAS REGRAS DE ASSOCIAÇÃO:

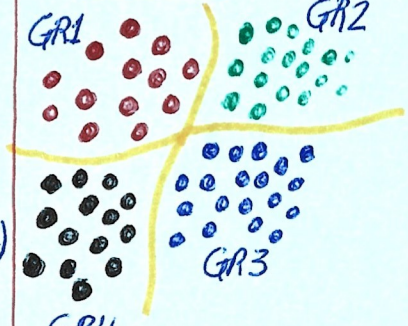
- PRATELEIRAS DE SUPERMERCADO
  - PROMOÇÕES COM ITENS QUE SÃO VENDIDOS EM CONJUNTO
- PLANEJAR CATÁLOGOS DAS LOJAS E FOLHETOS DE PROMOÇÕES
- CONTROLE DE EVASÃO EM UNIVERSIDADES # PODEMOS ANALISAR OS DADOS DAS PESSOAS QUE DESISTEM DE UM DETERMINADO CURSO. POR EXEMPLO, SE A NOTA FOR X, Y: RENDA E Z: CURSO, ENTÃO ELA TEM + CHANCE DE DESISTIR OU UMA DETERMINADA PORCENTAGEM

AGRUPAMENTO

→ OBJETIVO É ANALISAR OS DADOS E ENCONTRAR GRUPOS. EXEMPLOS DESTA ÁREA:

→ SEGMENTAÇÃO DE MERCADO

→→ PEGA UMA BASE DE DADOS DE CLIENTES E SEGMENTAR PARA MANDAR PROPAGANDAS SOMENTE PARA OS CLIENTES ESPECÍFICOS, OU SEJA, MANDAR UMA PROPAGANDA DE UM NOVO VIDEOGAME PARA UM CLIENTE QUE COMPRA VIDEOGAMES.



GR1: GRUPO1
⋮
GR4: GRUPO4

→→ ENCONTRAR GRUPOS DE CLIENTES QUE IRÃO COMPRAR UM PRODUTO (MALA DIRETA)

→→ AGRUPAMENTO DE DOCUMENTOS/NOTÍCIAS # PASSAR COMO PARÂMETROS VÁRIOS DOCUMENTOS E O ALGORITMO VAI AGRUPAR POR CONTEÚDO AQUELES CONTEÚDOS QUE SÃO PARECIDOS E AGRUPAR EM PASTAS NOTÍCIAS ESPECÍFICAS, COMO: NOTÍCIAS DE ESPORTE NA PASTA DE ESPORTE, ECONOMIA NA PASTA ECONOMIA, E ASSIM SUCESSIVAMENTE.

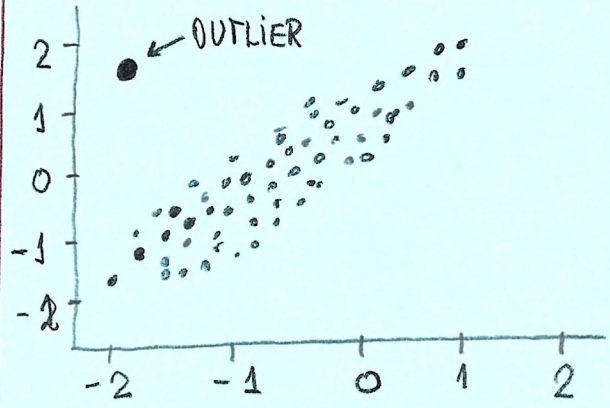
→→ AGRUPAMENTO DE PRODUTOS SIMILARES

A PARTIR DE UMA BASE DE DADOS, RODA-SE ~~UMA~~ UM ALGORITMO DE MACHINE LEARNING PARA FAZER A SEPARAÇÃO DOS PRODUTOS

→→ PERFIS DE CLIENTES (NETFLIX) # MELHORAR AS RECOMENDAÇÕES DE FILMES FUTUROS EM FUNÇÃO DO PERFIL DO CLIENTE BASEADO NO HISTÓRICO DE FILMES ASSISTIDOS PORESTE CLIENTE

→→ ANÁLISE DE REDES SOCIAIS → ENCONTRAR GRUPOS DE USUÁRIOS MAIS INFLUENTES DA REDE E ATRAVÉS DESTES ENVIAR PROMOÇÕES PARA ESTES USUÁRIOS INFLUENTES PODEREM DIVULGAR TAIS PRODUTOS

DETECÇÃO DE DESVIOS (OUTLIERS)



APLICAÇÕES

- FRAUDE EM CARTÃO DE CRÉDITO, OU SEJA, SAIU DO PADRÃO DE COMPRAS QUE VOCÊ TEM;
- INTRUSÃO EM REDES # DETECÇÃO DE INVASÃO EM UMA REDE
- USO DE ENERGIA ELÉTRICA, ÁGUA OU TELEFONE # SE O CONSUMO DESTES SERVIÇOS ESTÁ AUMENTANDO MUITO PODE SER UM PROBLEMA, POIS SAIU DO PADRÃO. NO CASO DO CONSUMO ELEVADO DE ÁGUA, DIFERENTE DOS MESES ANTERIORES, PODE SER UM INDICATIVO DE UM VAZAMENTO. NO CASO DE ENERGIA ELÉTRICA, PODE SER UM INDICATIVO DE ROUBO DE ENERGIA. EM AMBOS OS CASOS, A EMPRESA DE FORNECIMENTO pode acionar um técnico para investigar o desvio de padrão
- Desempenho de atletas (doping) # De mãos de todo o histórico de um atleta e em uma determinada corrida, um atleta que tem um padrão de fazer uma corrida em 15 segundos e ele faz em 10 segundos,

→

ou seja, saiu muito do padrão dele. Nesse caso, seria interessante chamar esse atleta para fazer um exame de doping, de forma a detectar um produto ilegal que o fez aumentar seu desempenho além do padrão

- Monitorar máquinas em um data center

Se num data Center alguma maquina não estiver funcionando corretamente ou fora do padrão, o sistema vai gerar um outlier, provocando um desempenho abaixo do padrão.

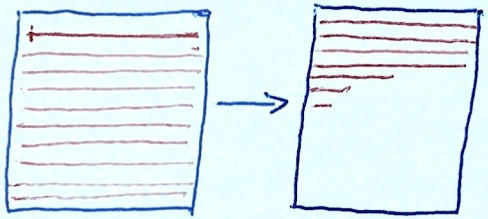
DESCOBERTA DE PADRÕES SEQUÊNCIAIS

TECNICAS INTERESSANTES QUE ALGUMAS LOJAS PODEM UTILIZAR, COMO POR EXEMPLO VOCÊ COMPRA O 1º LIVRO DO HARRY POTTER, E DEPOIS DE DOIS MESES UM E-MAIL É ENVIADO PARA VOCÊ DA EMPRESA DE COMPRA OFERTANDO + LIVROS DO HARRY POTTER

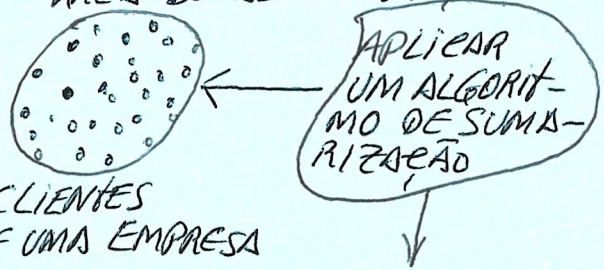
- LIVRARIAS, LOJAS DE EQUIPAMENTOS DE ATLETISMO, COMPUTADORES;
- MARKETING DIRECIONADO PARA PESSOAS QUE TEM MAIORES CHANCHES DE ADQUIRIR UM NOVO PRODUTO;

- PREVENÇÃO DE DOENÇAS → SE VOCÊ SABE QUE DETERMINADOS SINTOMAS VÃO CAUSAR UMA DOENÇA, ENTÃO ISSO É UMA SEQUÊNCIA. POR EXEMPLO, SINTOMAS A, B, C, ... N VÃO CAUSAR DOENÇA X
- NAVEGAÇÃO EM SITES → VOCÊ PODE MAPEAR A SEQUÊNCIA DE MENUS ACESSADOS EM SEU SITE

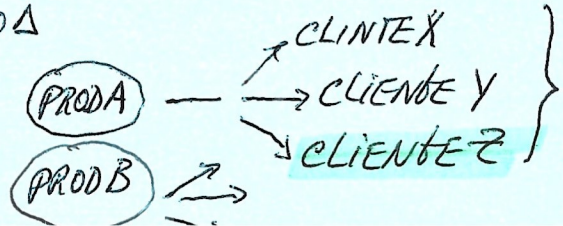
SUMARIZAÇÃO



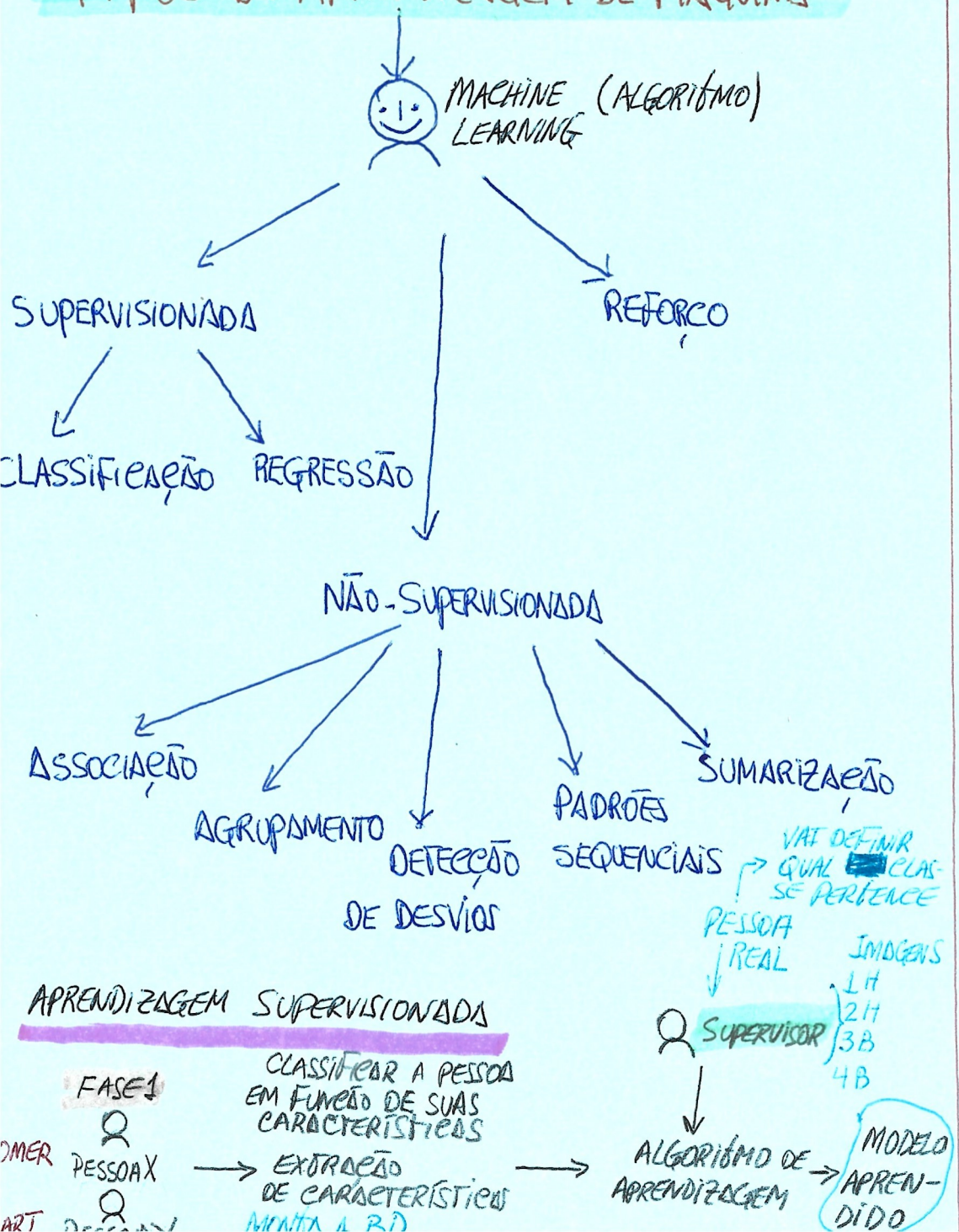
- SÃO OUVINTES DO PROGRAMA HOMENS NA FAIXA DE 25 A 35 ANOS, COM NÍVEL SUPERIOR E QUE TRABALHAM NA ÁREA DE ADMINISTRAÇÃO



- SEGMENTAÇÃO DE MERCADO → VOCÊ PODE FAZER UMA DETERMINADA VENDA DE UM PRODUTO EM FUNÇÃO DE DETERMINADAS CARACTERÍSTICAS DA PESSOA



TIPOS DE APRENDIZAGEM DE MÁQUINA



APRENDIZAGEM SUPERVISIONADA

FASE 1

HOMER — PESSOA X — BART — PESSOA Y

CLASSIFICAR A PESSOA EM FUNÇÃO DE SUAS CARACTERÍSTICAS

EXTRAÇÃO DE CARACTERÍSTICAS

MONTA A BD

ALGORITMO DE APRENDIZAGEM → MODELO APRENDIDO

PESSOA REAL → SUPERVISOR (VAI DEFINIR QUAL CLASSE PERTENCE)

IMAGENS: 1H, 2H, 3B, 4B

FASE 2

PESSOA Y → EXTRAÇÃO DE CARACTERÍSTICAS → MODELO APRENDIDO → BART SIMPSON

PERSONAGEM DA SÉRIE AMERICANA OS SIMPSONS.

ENTÃO, PRECISAMOS FAZER A EXTRAÇÃO DAS CARACTERÍSTICAS DAS IMAGENS E DEPOIS CLASSIFICAR O QUE TEM EM UMA IMAGEM E NÃO TEM NA OUTRA (CABEÇA REDONDA, CABEÇA ESPETADA, TEM BIGODE, CALÇA, SHORT, COR DA ROUPA, ...)

OU SEJA, EXTRAIR TODAS AS CARACTERÍSTICAS DOS PERSONAGENS

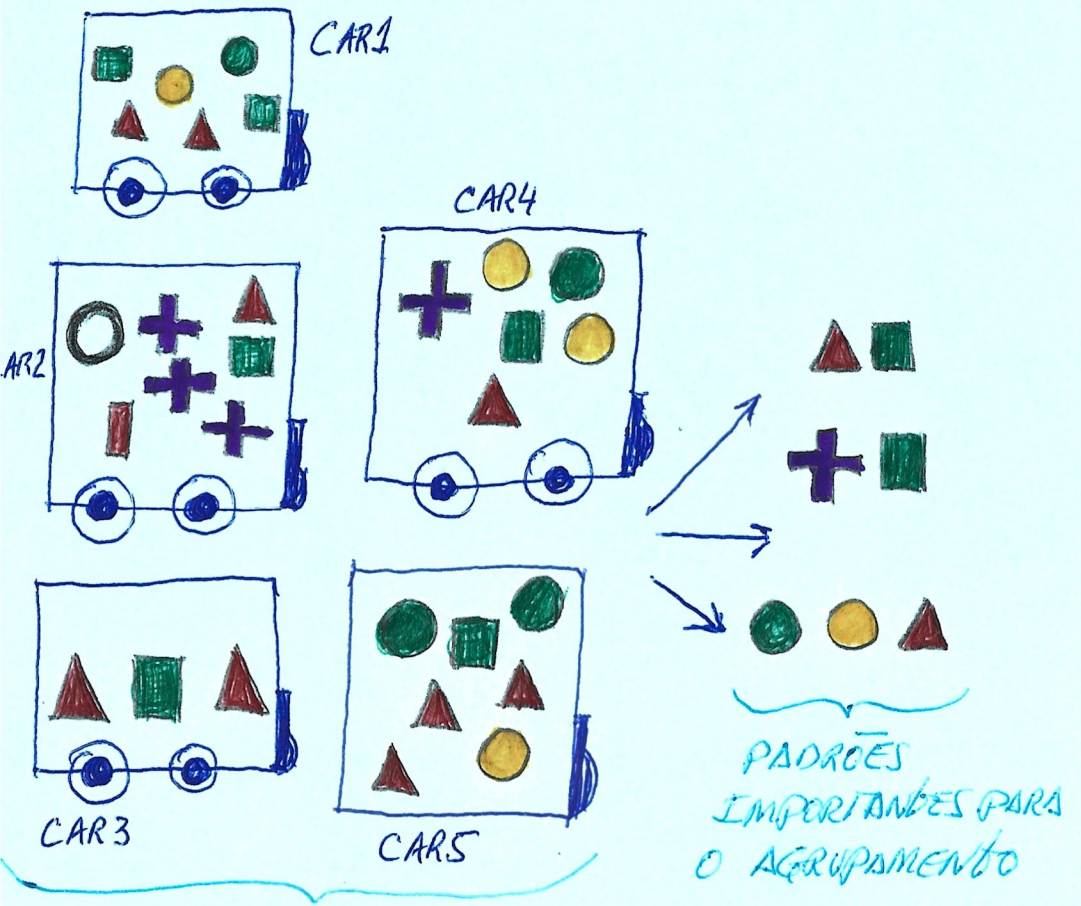
MODELO APRENDIDO → CONHECIMENTO QUE O ALGORITMO GEROU

FASE 2 → NÃO SEI A QUAL CLASSE A IMAGEM É, OU SEJA, NÃO SEI SE É UMA IMAGEM DO HOMER OU DO BART. VAMOS EXTRAIR AS CARACTERÍSTICAS E DAR PRA O MODELO APRENDIDO ESSAS CARACTERÍSTICAS PARA QUE O ALGORITMO (MODELO APRENDIDO) DER A RESPOSTA DE QUEM É A IMAGEM LIDA.

APRENDIZAGEM NÃO-SUPERVISIONADA

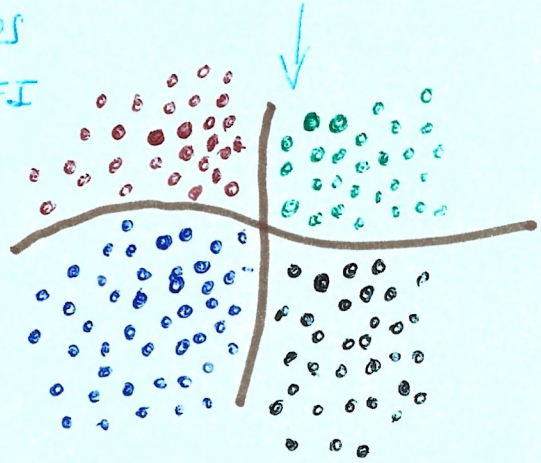
- ANALISAR AUTOMATICAMENTE OS DADOS (ASSOCIAÇÃO, AGRUPAMENTO);
- NECESSITA ANÁLISE PARA DETERMINAR O SIGNIFICA-

- DO OOS PADRÕES ENCONTRADOS



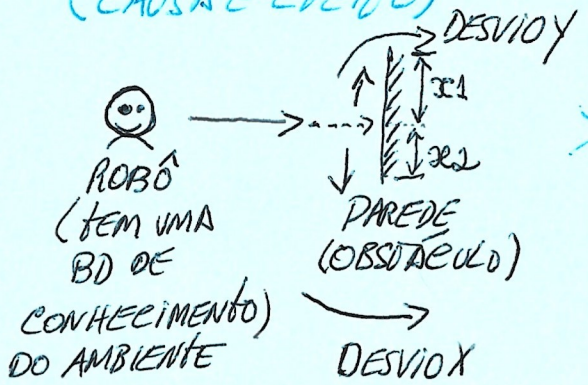
PADRÕES IMPORTANTES PARA O AGRUPAMENTO

DESCREVENDO OS DADOS DA BD QUE EU NÃO SEI O QUE TEM NELA



APRENDIZAGEM POR REFORÇO

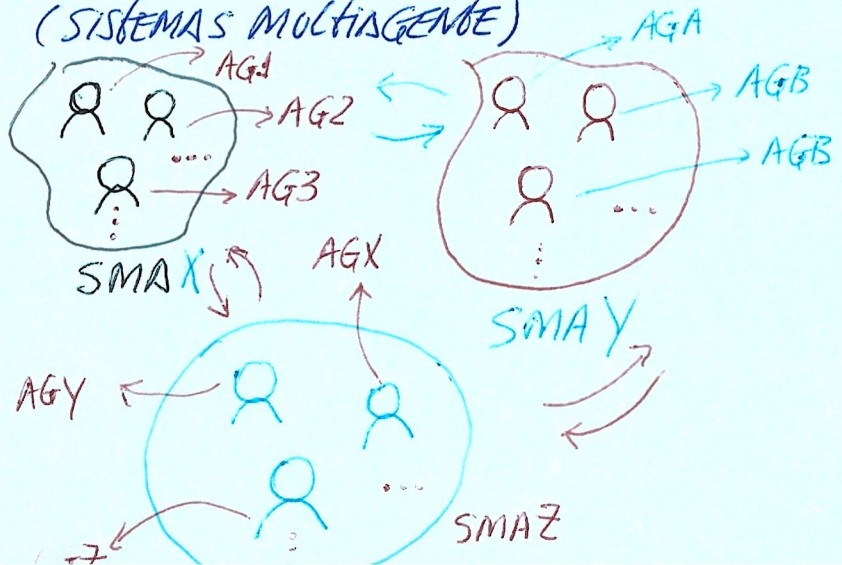
- APRENDER COM AS INTERAÇÕES COM O AMBIENTE (CAUSA E EFEITO)



⇒ APRENDER COM A SUA PRÓPRIA EXPERIÊNCIA

- APRENDER COM SUA PRÓPRIA EXPERIÊNCIA
- ROBÔ COLETANDO LIXO APRENDENDO A ANDAR EM UM AMBIENTE.

NOTA DO INSTRUTOR } A APRENDIZAGEM POR REFORÇO TEM MUITA APLICAÇÃO NA ÁREA DE SMA (SISTEMAS MULTIAGENTE)



O MÉTODO PREDITIVO CLASSIFICAÇÃO

PARA ESSE TÓPICO DE APRENDIZAGEM DE MÁQUINA, UTILIZAREMOS UMA BASE DE DADOS BANCÁRIO COM OS ATRIBUTOS ABAIXO.

BASE DE DADOS: RISCO DE CRÉDITO

HISTÓRIA DO CRÉDITO	DÍVIDA	GARANTIAS	RENDA ANUAL	RISCO
RUIM	ALTA	NENHUMA	< 15.000	ALTO
DESCONHECIDA	ALTA	ADEQUADA	>= 15.000 A <= 35.000	ALTO
DESCONHECIDA	BAIXA	NENHUMA	> 35.000	MODERADO
BOA	ALTA	ADEQUADA	>= 15.000 A <= 35.000	BAIXO
...	...	...	...	...

(CLIENTES)

RISCO → META/CLASSE

ATRIBUTOS PREVISORES

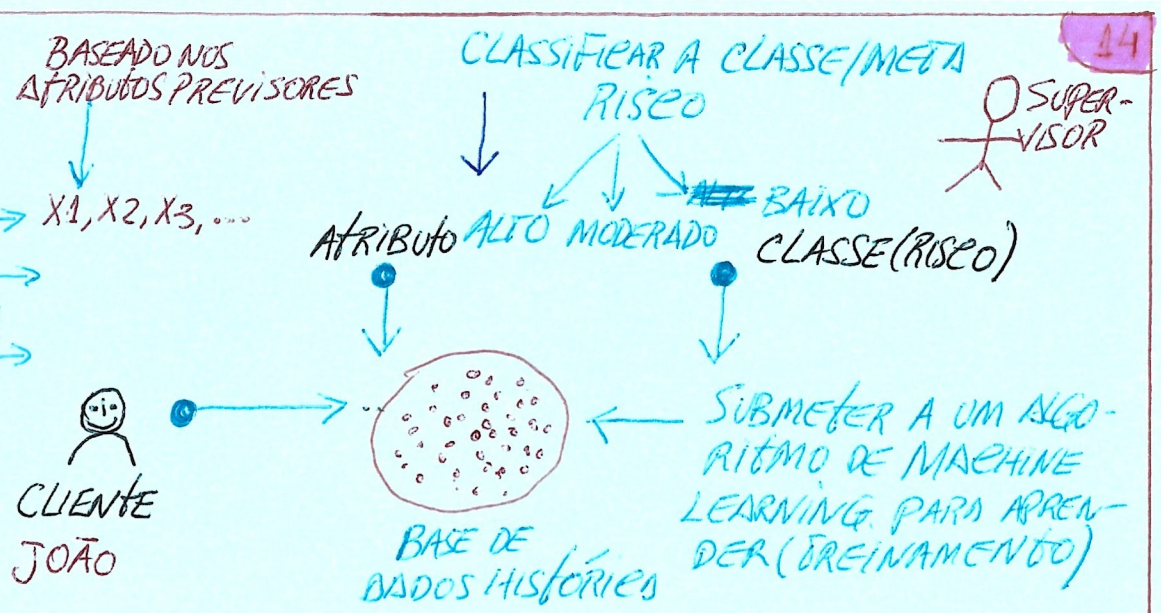
TREINAMENTO

BENS, TAIS COMO: CARROS, CASAS, APTS, TERRENOS, ...

RISCO ALTO
- HISTÓRICO DO CRÉDITO = RUIM
- DÍVIDA = ALTA
- GARANTIAS = NENHUMA
- RENDA ANUAL = < 15.000

RISCO MODERADO
- HISTÓRIA DO CRÉDITO: DESCONHECIDA
- DÍVIDA: BAIXA
- GARANTIAS: NENHUMA
- RENDA ANUAL: >= 15.000 A <= 35.000

RISCO BAIXO
- HISTÓRIA DO CRÉDITO: BOA
- DÍVIDA: ALTA
- GARANTIAS: NENHUMA



- ALGORITMO DE NAVE BAYES → GERA UMA TABELA DE PROBABILIDADES
- ÁRVORE DE DECISÃO → GERA UM ÁRVORE
- REDES NEURAIS → GERAM PESOS

...

N

AGORA, A PARTIR DE UMA NOVA BASE DE DADOS, FAREMOS AS CLASSIFICAÇÕES PARA OS NOVOS CLIENTES.

TESTE

HISTÓRIA DO CRÉDITO	DÍVIDA	GARANTIAS	RENDA ANUAL	RISCO
RUIM	ALTA	ADEQUADA	< 15.000	?
DESCONHECIDA	ALTA	ADEQUADA	< 15.000	?
DESCONHECIDA	BAIXA	NENHUMA	> 35.000	?
BOA	ALTA	ADEQUADA	>= 15.000 A <= 35.000	?

CLASSIFICAÇÃO (VENDA DE LIVROS)

x_1 SEXO	x_2 PAIS	x_3 IDADE	y COMPRAR
M	FRANÇA	25	SIM
M	INGLATERRA	21	SIM
F	FRANCA	23	SIM
F	INGLATERRA	34	SIM
F	FRANÇA	30	NÃO
M	ALEMANHA	21	NÃO
M	ALEMANHA	20	NÃO
F	ALEMANHA	18	NÃO
F	FRANÇA	34	NÃO
⋮	⋮	⋮	⋮

ATRIBUTOS PREVISORES | META

BASE DE DADOS HISTÓRICA

TREINAMENTO

x_1, x_2 e x_3 → VARIÁVEIS INDEPENDENTES

y → VARIÁVEL DEPENDENTE

X1 SEXO	X2 PAIS	X3 IDADE	META COMPRAR
M	FRANÇA	38	?
F	INGLATERRA	25	?
M	ALEMANHA	55	?
F	FRANÇA	20	?
⋮	⋮	⋮	⋮

TESTE

CLASSIFICAÇÃO (PREVER O ESPORTE)

TREINAMENTO

COR DOS OLHOS	CASADO	SEXO	CABELO	ESPORTE
CASTANHO	SIM	M	LONGO	FUTEBOL
AZUL	SIM	M	CURTO	FUTEBOL
CASTANHO	SIM	M	LONGO	FUTEBOL
CASTANHO	NÃO	F	LONGO	AERÓBICA
CASTANHO	NÃO	F	LONGO	AERÓBICA
AZUL	NÃO	M	CURTO	FUTEBOL
⋮	⋮	⋮	⋮	⋮

TESTE

COR DOS OLHOS	CASADO	SEXO	CABELO	ESPORTE
CASTANHO	SIM	M	CURTO	?
CASTANHO	NÃO	M	LONGO	?
AZUL	NÃO	F	LONGO	?
AZUL	SIM	M	LONGO	?
⋮	⋮	⋮	⋮	⋮

NOTA DO INSTRUTOR: ESSA TÉCNICA É MUITA USADA EM TESTE VOCACIONAL. POR EXEMPLO, UMA UNIVERSIDADE QUANDO VAI FAZER DIVULGAÇÃO DE VESTIBULAR APLICA O SUPRACITADO TESTE E A PARTIR DAÍ, CRIAR UMA BASE DE DADOS COM TODAS AS PESSOAS QUE VÃO FAZER O TESTE E COM BASE NAS CARACTERÍSTICAS DAS PESSOAS PREVER QUAL CURSO A PESSOA VAI FAZER.

BASE DE DADOS HISTÓRICOS [CARACTERÍSTICAS CLIMÁTICAS]

CLASSIFICAÇÃO (JOGAR TÊNIS)

ATRIBUTOS PREVISORES →
- TEMPO
- TEMPERATURA
- UMIDADE
- VENTO

META/CLASSE → JOGAR TÊNIS

TEMP	TEMPERATURA	UMIDADE	VENTO	JOGAR
ENSOLARADO	QUENTE	ALTA	FRACO	NÃO
ENSOLARADO	QUENTE	ALTA	FORTE	NÃO
NUBLADO	QUENTE	ALTA	FRACO	SIM
CHUVOSO	MODERADA	ALTA	FRACO	SIM
CHUVOSO	AGRADÁVEL	NORMAL	FORTE	NÃO
NUBLADO	AGRADÁVEL	NORMAL	FORTE	SIM
ENSOLARADO	MODERADA	ALTA	FRACO	NÃO
CHUVOSO	MODERADA	NORMAL	FRACO	SIM
⋮	⋮	⋮	⋮	⋮

TREINAMENTO

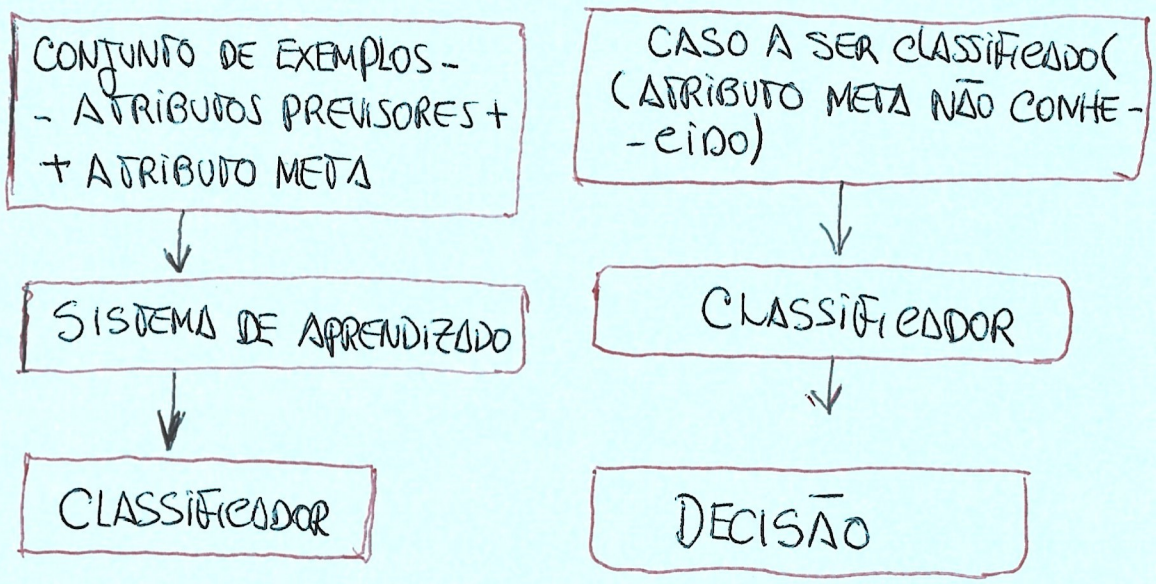
TEMPO	TEMPERATURA	UMIDADE	VENTO	JOGAR TÊNIS
ENSOLARADO	MODERADA	NORMAL	FORTE	?
CHUVOSO	AGRADÁVEL	NORMAL	FRACO	?
NUBLADO	QUENTE	NORMAL	FRACO	?
NUBLADO	AGRADÁVEL	ALTA	FORTE	?

TESTE

ATENÇÃO → CLASSIFICAÇÃO

- CADA REGISTRO PERTENCE A UMA CLASSE E POSSUI UM CONJUNTO DE ATRIBUTOS PREVISORES;
- OBJETIVA-SE DESCOBRIR UM RELACIONAMENTO ENTRE OS ATRIBUTOS PREVISORES E O ATRIBUTO META;
  - OS ALGORITMOS DE MACHINE LEARNING VÃO TENTAR ENCONTRAR ESSES RELACIONAMENTOS
- O VALOR DO ATRIBUTO META É CONHECIDO (APRENDIZAGEM SUPERVISIONADA)

Representação da classificação (MÉTODO INDUTIVO)



NOTA: IR PARA A PÁGINA 12 E VER ILUSTRAÇÃO DA REPRESENTAÇÃO DA CLASSIFICAÇÃO APLICADA NA RESOLUÇÃO DE UMA TAREFA DO MUNDO REAL.

INSTALAÇÃO DO R E DO RSTUDIO

- NA BUSCA DO BROWSER, DIGITAR DOWNLOAD R E CLICAR EM THE R PROJECT FOR STATISTICAL COMPUTING (O Projeto R para Computação Estatística) → HTTPS://WWW.R-PROJECT.ORG. DEPOIS, CLICAR EM ~~DOWNLOAD~~ CRAN, DEPOIS ESCOLHER O ESPELHO HTTPS://CLOUD.R-PROJECT.ORG/ E DEPOIS, ESCOLHER O SISTEMA OPERACIONAL USADO. NO MEU CASO, SO WINDOWS, DEPOIS CLICAR EM BASE (EM SUBDIRETÓRIOS.

  Rx64 → CARREGAR ESSE ÍCONE, VIA MENU DE PROGRAMAS (JANELA DO WINDOWS)

```
> 1+1
[1] 2
```

Nota do Instrutor } NÃO VAMOS TRABALHAR COM O R, E SIM COM UMA IDE RSTUDIO → FERRAMENTA + POPULAR PARA TRABALHAR COM O R

DOWNLOAD RSTUDIO # BAIXAR O ARQUIVO DE INSTALAÇÃO E EXECUTÁ-LO ATÉ À SUA CONCLUSÃO.

A	B
D	C

ONDE:
A, B, C E D SÃO AS INTERFACES DO IDE RSTUDIO

PRÉ-PROCESSAMENTO NAS BASES DE DADOS

↳ SÃO PROCESSOS QUE DEVEM SER EXECUTADOS ANTES DE FAZERMOS APLICAÇÃO DOS ALGORITMOS DE APRENDIZAGEM DE ~~DADOS~~ MÁQUINA.

CONTEÚDO

- TIPOS DE VARIÁVEIS
- BASE DE DADOS
  - DADOS DE CRÉDITO
  - CENSO
- CARREGAMENTO DAS BASES DE DADOS
- VALORES INCONSISTENTES (POR EXEMPLO, IDADE < 0 (NEGATIVA) → NÃO PODE ACONTECER ISSO. PORTANTO, PRECISAMOS CORRIGIR ESSE PROBLEMA.
- VALORES FALTANTES (DADOS AUSENTES, NÃO DISPONÍVEIS) E ISSO PREJUDICA O APRENDIZADO DOS ALGORITMOS
- ESCALONAMENTO DE ATRIBUTOS → PODE TER ATRIBUTOS NUMÉRICOS FORA DE ESCALA EM RELAÇÃO AOS DEMAIS, PARA CIMA OU PARA BAIXO
- TRANSFORMAÇÃO DE VARIÁVEIS CATEGÓRICAS → VARIÁVEIS DE TEXTO E NESSE CASO VAI PRECISAR TRANSFORMAR PARA NÚMEROS, POIS ALGUNS OU A MAIORIA DOS ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NÃO LEEM TEXTOS
- INTRODUÇÃO A AVALIAÇÃO DE ALGORITMOS
- BASE DE TREINAMENTO E BASE DE TESTE

TIPOS DE VARIÁVEIS

existem algoritmos de maquinas que não reconhecem variáveis numéricas, como é o caso de regras de associação.

- Numéricas
  - Contínua
    - Números reais
    - temperatura, altura, Peso, Salário
  - Discreta
    - Conjunto de valores finito (inteiros)
    - CONTAGEM DE ALGUMA COISA
      1 BAIXO
      2 MEDIO BAIXO
      3 MÉDIO
      4 MEDIO ALTO
      5 - ALTO
- CATEGÓRICAS {STRING}
  - NOMINAL
    - DADOS NÃO MENSURÁVEIS
    - SEM ORDENAÇÃO: COR DOS OLHOS, GÊNERO, ID, NOME
  - ORDINAL
    - CATEGORIZADO SOB UMA ORDENAÇÃO.
    - TAMANHO, P, M E G

BASE DE DADOS DE CRÉDITO

↓

1ª BASE DE DADOS A SER USADA COMO ESTUDO DE CASO PARA O DESENVOLVIMENTO DO CURSO.

VARIÁVEIS

CLIENTID: ID DO CLIENTE
INCOME: RENDA
LOAN: EMPRÉSTIMO (DÍVIDA)
DEFAULT: CLASSE

} ATRIBUTOS PREVISORES

} ATRIBUTO CLASSE

0: NÃO PAGOU A DÍVIDA
1: PAGOU A DÍVIDA

DEFAULT:
- 0 → NÃO PAGOU O EMPRESTIMO
- 1 → PAGOU O EMPRESTIMO

NO IDE DO RSTUDIO

- FILE
  - NEW FILE
    - R SCRIPT
      - SALVAR O ARQUIVO COM O NOME: PRE_PROCESSAMENTO_CREDIT_DATA
        ↓ ESTE ARQUIVO PASSA TER EXTENSÃO .R

NO MENU "MORE", ESCOLHER A OPÇÃO "SET AS WORKING DIRECTORY"

```
>SETWD("D:/MACHINE LEARNING E DATA SCIENCE COM R")
```

CARREGANDO A BASE DE DADOS .CSV QUE ESTÁ NA PASTA SUPRACITADA PARA DENTRO DO R

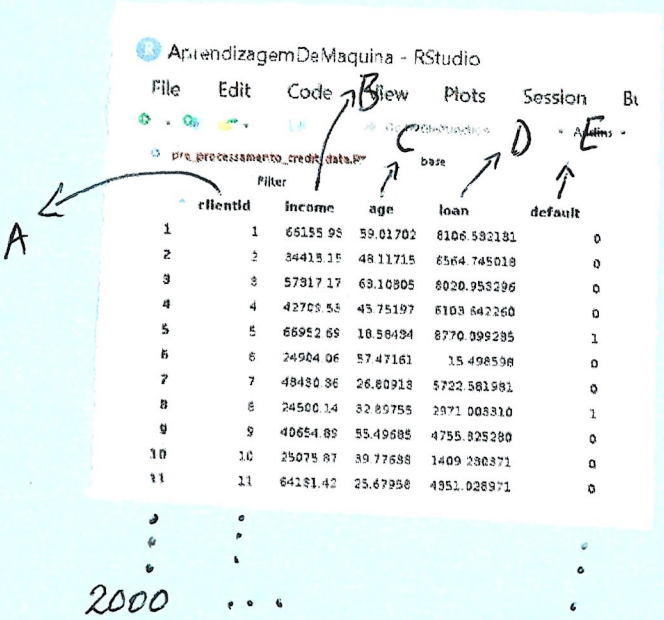
NA ÁREA DE SCRIPT DO ARQUIVO "PRE_PROCESSAMENTO_CREDIT_DATA.R", DIGITAR OS COMANDOS QUE SEGUEM:

→

BASE <- READ.CSV('CREDIT_DATA.CSV') # SELECIONAR TUDO E EXECUTAR

↓

2000 OBS. OF 5 VARIABLES



FAREMOS AGORA A VINCULAÇÃO DOS ATRIBUTOS DA BASE COM AQUELES TIPOS DE VARIÁVEIS DESCRITOS NA PÁGINA 18

A: ID DO CLIENTE

↓

CHAVE PRIMÁRIA

↓

~~VARIÁVEL NOMINAL~~ VARIÁVEL CATEGÓRICA NOMINAL

Nota do instrutor: PARA OS ALGORITMOS DE MACHINE LEARNING ESSE TIPO DE CAMPO NÃO É IMPORTANTE PARA FAZER UM TIPO DE PREVISÃO. PORTANTO, VAMOS APAGAR ESSE CAMPO DA BASE DE DADOS BASE.

BASE$CLIENTID <- NULL

↓

BASE 2000 OBS. OF 4 VARIABLES

B: INCOME (RENDA) -> VARIÁVEL NUMÉRICA CONTÍNUA

C: AGE (IDADE) -> VARIÁVEL NUMÉRICA CONTÍNUA

D: LOAN (DÍVIDA/EMPRÉSTIMO) -> VARIÁVEL NUMÉRICA CONTÍNUA

E: DEFAULT (CLASSE/METAS) -> VARIÁVEL NUMÉRICA DISCRETA.

SUMMARY(BASE)

	INCOME	AGE	LOAN	DEFAULT
MIN.	: 2004	: -52.42	: 1.378	: 0.0000
1st QU.	: 32796	: 28.99	: 1939.709	: 0.0000
MEDIAN	: 45789	: 41.32	: 3974.719	: 0.0000
MEAN	: 45332	: 40.81	: 4444.370	: 0.1415
3RD QU.	: 57791	: 52.59	: 6432.411	: 0.0000
MAX	: 69996	: 63.97	: 13766.051	: 1.0000
		NA's : 3		

Tabela de resumo de estatística da BD base

1st Qu: 1ª fatia da base = 25% da base

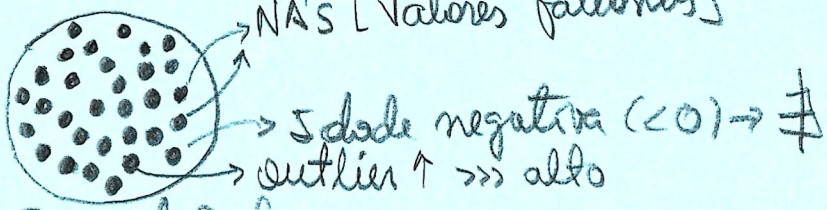
Median (mediana) = 50% da base

3td Qu: 75% da base

PRÓXIMA AULA SERÁ:

↳ Tratamento de Valores inconsistentes - base crédito.

Tratamento de Valores Inconsistentes - Base Crédito



A partir do Comando summary(base) podemos perceber que tem-se no Campo "age" um valor negativo para a idade, que é inconsistente

age = -52.42 → Uma inconsistência e isso pode influenciar negativamente no desempenho e aprendizagem do algoritmo.

Nos registros 16, 22, 27 apresentam age < 0. Desta forma, precisamos achar uma forma de corrigir essa inconsistência.

Nota do Instrutor } Para ver se tem + registros com "age" < 0, siga os comandos.

No script criado, digite:

base[base$age<0,]

	income	age	loan	default
16	50501.73	-28.21	3977.	0
22	32197.62	-52.42	4244.	0
27	63287.04	-36.49	9595.	0
NA	NA	NA	NA	NA
NA.1	NA	NA	NA	NA
NA.2	NA	NA	NA	NA

base[base$age<0, 1:2] #NA (Not Available)

NA → 29, 31 e 32

base[base$age<0 & !is.NA(base$age),]

	income	age	loan	default
16	50501.73	-28.21	3977.	0
22	32197.62	-52.42	4244.	0
27	63287.04	-36.49	9595	0

IDADE_INVÁLIDA <- BASE[BASE$AGE<0 & !is.NA(BASE$$AGE),]

IDADE_INVALIDA

NOTA DO INSTRUTOR } A BASE DE DADOS NÃO PODE FICAR COM ESSES TRÊS VALORES COM ~~NAS~~ IDADES INVÁLIDAS

MANEIRAS DE TRATAR DADOS INVÁLIDOS (AGE<0)

1ª OPÇÃO: APAGAR A COLUNA INTEIRA # PEGAR A COLUNA "AGE" E APAGAR A COLUNA INTEIRA, OU SEJA EXCLUIR A VARIÁVEL "AGE". ENTRETANTO, ESSA NÃO É UMA OPÇÃO OU SOLUÇÃO RECOMENDÁVEL, POIS NÃO É POR CAUSA DE 3 REGISTROS, VAMOS PERDER 1997 REGISTROS BONS. SE FOSSE O INVERSO, AÍ SIM, SERIA VIÁVEL.

2ª Opção: APAGAR SOMENTE OS REGISTROS COM PROBLEMAS

```
BASE <- BASE[BASE$AGE > 0, ] #A BASE VAI FICAR COM 1997 REGISTROS
```

ESSA opção É MENOS PIOR QUE A 1ª opção, ENTRANTO, VAMOS PERDER DADOS.

3ª Opção: PREENCHER OS DADOS MANUALMENTE

↳ PEGAR OS 3 REGISTROS, ENTRAR EM CONTATO COM ESSAS TRÊS PESSOAS E PERGUNTAR AS IDADES DOS MESMOS # É UMA OPÇÃO + CORRETA, ENTRETANTO É + BRAÇAL E MANUAL.

4ª Opção: CALCULAR A MÉDIA DA IDADE E PREENCER OS VALORES INVÁLIDOS COM O VALOR DA MÉDIA

```
MEAN(BASE$AGE)
[1] NA # O R retorna NA porque existem NA's na base. Portanto, devemos informar ao R que não queremos NA's.
mean(base$age, na.rm=TRUE)
[1] 40.80756 # Entretanto esse valor não é verdadeiro, pois ele considera o Calculo da média usando os valores negativos ou seja, age<0. Temos que dizer ao R para não incluir os valores de age<0.
MEAN(BASE$AGE[BASE$AGE>0], NA.RM=TRUE)
[1] 40.9277 → ESSA É A MÉDIA REAL QUE VAMOS CONSIDERAR NOS CALCULOS PARA RESOLVER O PROBLEMA DE DADOS INVÁLIDOS.
BASE$AGE <- IFELSE(BASE$AGE<0, 40.92, BASE$AGE)
```

	INCOME	AGE	LOAN	DEFAULT
16	50501.73	40.9200	3977....	0
22	32197.62	40.9200	4244....	0
27	63287.04	40.9200	9595....	0

```
MEAN(BASE$AGE, NA.RM=TRUE)
[1] 40.92769 # MÉDIA ATUALIZADA
```

NOTA DO INSTRUTOR } COLOCANDO O CURSOR SOBRE UM COMANDO E APERTAR "F1", MOSTRA O HELP SOBRE ESSE COMANDO.

TRATAMENTO DE VALORES FALTANTES - BASE CRÉDITO

↳ NO ATRIBUTO PREVISOR "AGE", TEMOS TRÊS VALORES NA'S, OS REGISTROS 29, 31 E 32.

```
BASE[is.NA(BASE$AGE),] # VAI MOSTRAR OS REGISTROS COM NA'S, CONFORME TRAZIDO ABAIXO.
```

	INCOME	AGE	LOAN	DEFAULT
29	59417.81	NA	2082.626	0
31	48528.85	NA	6155.785	0
32	23526.30	NA	2862.010	0

```
BASE$AGE <- IFELSE(IS.NA(BASE$AGE), MEAN(
(BASE$AGE, NA.RM=TRUE),
, BASE$AGE)
```

ESCALONAMENTO DE ATRIBUTOS - BASE CRÉDITO

CAMPO 15 → 63061
14 → 27268
} REGISTRO INCOME → 63061 − 27268 → 35793

61
39
} REGISTRO AGE → 61 − 39 = 22

PRECISAMOS COLOCAR OS CAMPOS "INCOME" E "AGE" NUMA MESMA ESCALA DIMENSIONAL

↳ ALGORITMOS BASEADOS EM DISTÂNCIA DEVE SER LEVADO EM CONSIDERAÇÃO ESSA QUESTÃO DE DISTÂNCIA (ESCALONAMENTO)

↳ DEIXAR OS ATRIBUTOS NUM MESMO PADRÃO.

TÉCNICAS DE ESCALONAMENTO

- NORMALIZAÇÃO (NORMALIZATION)

$$X = \frac{X - \text{MÍNIMO}(X)}{\text{MÁXIMO}(X) - \text{MÍNIMO}(X)}$$

PRINCIPALMENTE PARA TRATAR OUTLIER

- PADRONIZAÇÃO (STANDARDISATION)

$$X = \frac{X - \text{MÉDIA}(X)}{\text{DESVIO PADRÃO}(X)}$$

← ESSA TÉCNICA É A MAIS RECOMENDADA PARA ESCALONAMENTO.

ONDE:

X É O VALOR DO RESPECTIVO ATRIBUTO QUE QUEREMOS NORMALIZAR. SUPONHAMOS QUE QUEIRAMOS NORMALIZAR UMA DETERMINADA IDADE. POR EXEMPLO:

PACIENTE → AGE = 40

$$\frac{40 - 20}{60 - 20} = 0{,}5$$

40 → VALOR QUE QUEREMOS NORMALIZAR
20 → VALOR MÍNIMO (AGE)

FAZENDO O ESCALONAMENTO SOMENTE DOS ATRIBUTOS PREVISORES

```
BASE[, 1:3] <- SCALE(BASE[, 1:3])
```

BASE DE DADOS DO CENSO

↳ ∃ na WEB a BASE UCI MACHINE LEARNING REPOSITÓRIO (REPOSITORY)

ARCHIVE.ICS.UCI.EDU/ML/INDEX.PHP → Repositório de machine learning. para diversas tarefas de mineração de dados.

A BASE QUE USAREMOS É A "ADULT"

ADULT DATA SET → Em DOWNLOAD, clicar em Data Folder, depois em adult.data

CRIANDO NOVO SCRIPT PARA ESSA NOVA BD

↳ SALVANDO O script com o NOME DE "PRE_PROCESSAMENTO_CENSO

1. IMPORTANDO A BASE DE DADOS PARA O R;

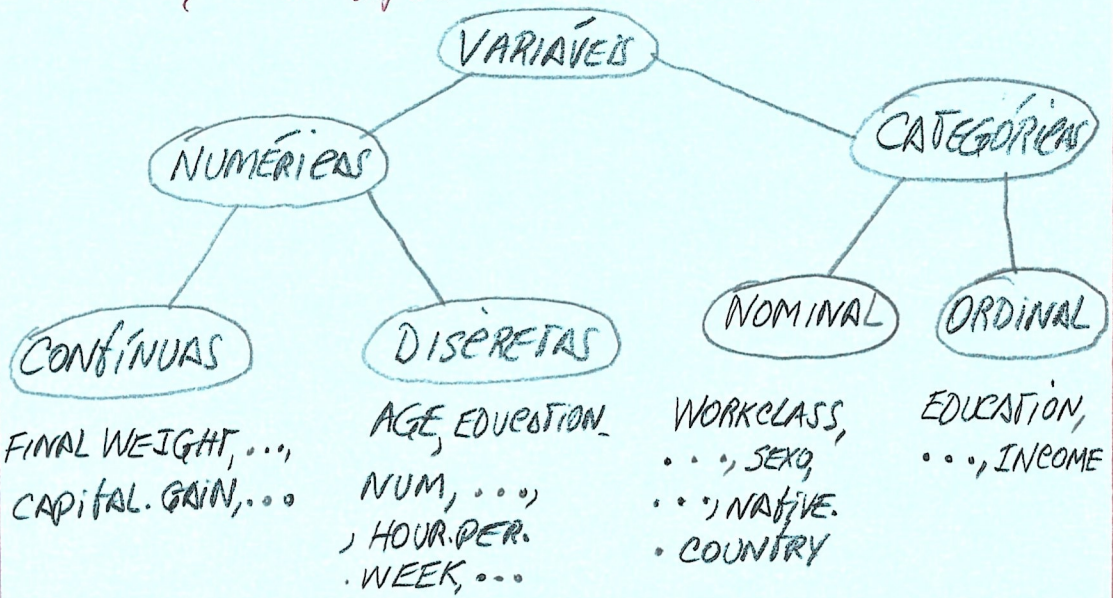
```
BASE_CENSO <- READ.CSV('CENSUS.CSV')
```

↳ 30162 obs. OF 16 VARIABLES

2. VAMOS APAGAR A COLUNA "X" (ID), POIS ESTE NÃO SERVIRÁ PARA AS PREVISÕES.

```
BASE_CENSO$X <- NULL
```

VINCULAÇÃO DE TIPOS DE VARIÁVEIS # PASSO 3



NOTA DO INSTRUTOR } NESSA BASE, NÃO TEMOS VALORES NA'S E VALORES INCONSISTENTES. OU SEJA, A BASE ESTÁ MAIS ORGANIZADA QUE A BASE "CREDIT_DATA".

TRANSFORMAÇÃO DE ATRIBUTOS CATEGÓRICOS - BASE CENSO

↳ PRECISAMOS TRANSFORMAR ATRIBUTOS CATEGÓRICOS PARA NUMÉRICOS, POIS ALGUNS ALGORITMOS DE MACHINE LEARNING NÃO RECONHECEM ESSES TIPOS DE VARIÁVEIS NO MODELO.

→

PORTANTO, NA BASE DE DADOS CENSUS TEMOS VÁRIOS ATRIBUTOS NOMINAIS QUE PRECISAM SER TRANSFORMADOS PARA ATRIBUTOS NUMÉRICOS DO TIPO DISCRETO. POR EXEMPLO, ATRIBUTO GÊNERO, M ∧ F, PRECISAMOS COLOCAR UM VALOR NUMÉRICO PARA A MÁQUINA ENTENDER. M → 0 E F → 1 OU VICE-VERSA. PARA TAL, VAMOS FAZER UM COMANDO DO R, CONFORME TRAZIDO ABAIXO:

```
TABLE(BASE_CENSO$SEX) ↵
```

FEMALE	MALE
9782	20380

```
UNIQUE(BASE.BENSO$SEX) ↵
"MALE"   "FEMALE"
```

VAMOS FAZER A MODIFICAÇÃO DO TIPO DE ATRIBUTO

```
base_censo$sex <- factor(base_censo$sex, levels =
= c(' Female', ' Male'), labels = c(0,1))
```

ESPAÇO ESPAÇO } OLHAR COMO ESTÁ A SAÍDA NO COMANDO UNIQUE(BASE_CENSO$SEX)

```
BASE_CENSO[is.na(BASE_CENSO$SEX)]
Data Frame with 0 columns and 30162 rows
```

ESCALONAMENTO DE ATRIBUTOS - CENSO

FINAL.WEIGHT	EDUCATION
77516	10
83311	10

} # MUITO GRANDE DE ESCALAS. ISSO É PROBLEMA, POIS HÁ ALGORITMOS DE MACHINE LEARNING (KNN) QUE LEVA EM CONSIDERAÇÃO AS DISTÂNCIAS.

↓ TEMOS QUE COLOCAR ESSES ATRIBUTOS NA MESMA ESCALA.

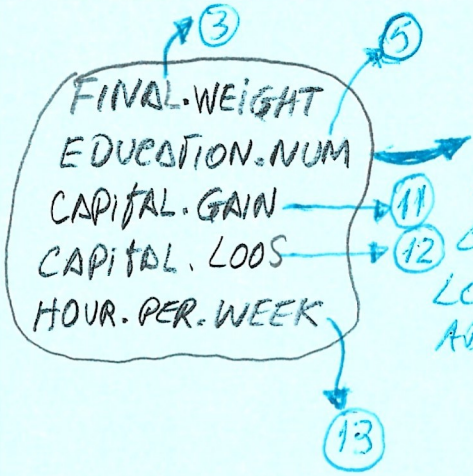
```
BASE_CENSO <- SCALE(BASE_CENSO)
> BASE_CENSO <- SCALE(BASE_CENSO)
  Error in colMeans(x, na.rm = TRUE) : 'x'
  must be numeric
BASE_CENSO[, 1] <- SCALE(BASE_CENSO[, 1])
```

COLUNAS/BASE

- WORKCLASS → 2
- EDUCATION → 4
- MARITAL.STATUS → 6
- OCCUPATION → 7
- RELATIONSHIP → 8
- RACE → 9
- SEX → 10
- NATIVE.COUNTRY → 14
- INCOME → 15

TODOS ESSES SÃO DO TIPO ORIGINAL FACTOR, E PRECISAM, DESTA FORMA, SEREM CONVERTIDOS PARA NUMERIC ANTES DO ESCALONAMENTO.

→



ATRIBUTOS NUMÉRICOS QUE NÃO PRECISAM SER TRANSFORMADOS E PODEM SER ESCALONADOS LOGO DIRETAMENTE, CONFORME ABAIXO

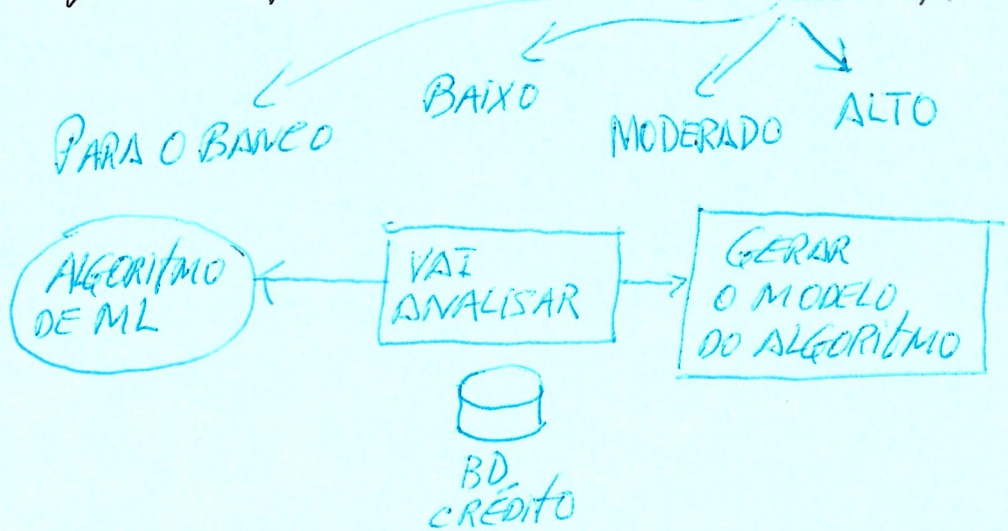
```
• BASE_CENSO[,3] <- SCALE(BASE_CENSO[,3]) ↵
• BASE_CENSO[,5] <- SCALE(BASE_CENSO[,5]) ↵
• BASE_CENSO[,11:13] <- SCALE(BASE_CENSO[,11:13]) ↵
```

COM ISSO, CONCLUIMOS O ESCALONAMENTO DOS ATRIBUTOS NUMÉRICOS DA BASE DE DADOS SENSUS

INTRODUÇÃO A AVALIAÇÃO DE ALGORITMOS

→ VAMOS APRENDER SE O DESEMPENHO DO ALGORITMO É BOM OU RUIM. VAMOS DIVIDIR AS BASES DE DADOS EM UMA PORÇÃO SOMENTE PARA TREINAMENTO E UMA PORÇÃO SOMENTE PARA TESTE

Para esse tópico, vamos usar a Base de crédito, já tratada e discutida na página 14, cujos atributos previsores são: História do crédito, dívida, garantias e renda anual, de forma a prever a meta (classe = Risco).



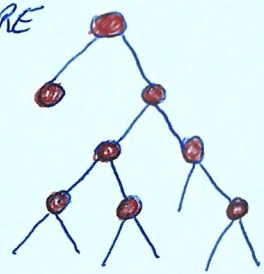
POR EXEMPLO, O ALGORITMO NAIVE BAYES VAI GERAR UMA TABELA DE PROBABILIDADES, CONFORME ILUSTRAÇÃO ABAIXO:

%	%	%
%	%	%
%	%	%

} PROBABILIDADES DAS CLASSE P(A), P(M) E P(B)

↑ RISCO ALTO ↓ RISCO MODERADO ↓ RISCO BAIXO.

JÁ O ALGORITMO DE DECISÃO, VAI CONSTRUIR UMA ÁRVORE

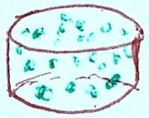


DEPOIS QUE EU QUISER CLASSIFICAR UM NOVO REGISTRO (CLIENTE), EU NÃO SEI A QUE CLASSE ESSES NOVOS REGISTROS PERTENCEM

TABELA:

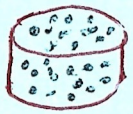
HISTÓRIA DO CRÉDITO	DÍVIDA	GARANTIAS	RENDA ANUAL	RISCO
X1	X2	X3	X4	
RUIM	ALTA	ADEQUADA	<15.000	?
DESCONHECIDA	ALTA	ADEQUADAS	>35.000	?
BOA	BAIXA	NENHUMA	>=15.000 A <=35.000	?

SUBMETEREMOS ESSES NOVOS REGISTROS AO MODELO APRENDIDO DO ALGORITMO E AÍ ESTE FARÁ A NOVA CLASSIFICAÇÃO.



BD1 TREINAMENTO

BD TREINAMENTO	BD TESTE	ALGORITMO	
ALTO	ALTO	ACERTO	50% DE ACERTO E 50% ERRO
ALTO	MODERADO	ERRO	
BAIXO	BAIXO	ACERTO	
MODERADO	ALTO	ERRO	



BD2 TESTE

PROVA DE MATEMÁTICA DE REVISÃO 10 QUESTÕES

≠

DIA DA PROVA 10 QUESTÕES DIFERENTES

TREINAMENTO

TESTE

ESTA PROVA VOCÊ VAI REALMENTE MEDIR SE O ALUNO APRENDEU DE FATO.

$$\text{EFICIÊNCIA DO ALGORITMO} = \frac{\text{Nº DE ACERTOS}}{\text{Nº TOTAL DE REGISTROS}}$$

BASE DE TREINAMENTO E TESTE

↳ NESTE TÓPICO FAREMOS A DIVISÃO DAS BASES DE DADOS DE TREINAMENTO E TESTE

PRÉ_PROCESSAMENTO_CREDIT_DATA.R

```
IMPLEMENTACAO1 <- READ.CSV('CREDIT_DATA.CSV')
IMPLEMENTACAO1$CLIENTID = NULL
SUMMARY(IMPLEMENTACAO1)
IDADE_INVALIDA <- IMPLEMENTACAO1[IMPLEMENTACAO1$
                  $AGE < 0 & !IS.NA(IMPLEMENTACAO1$
                  $AGE), ]

IMPLEMENTACAO1$AGE <- NULL
IMPLEMENTACAO1 <- IMPLEMENTACAO1$AGE > 0, ]
MEAN(IMPLEMENTACAO1$AGE, NA.RM = TRUE)
MEAN(IMPLEMENTACAO1$AGE[IMPLEMENTACAO1$AGE > 0],
     NA.RM = TRUE)

IMPLEMENTACAO1$AGE <- IFELSE(IMPLEMENTACAO1$AGE < 0,
                      40.92, IMPLEMENTACAO1$AGE
IMPLEMENTACAO1[IS.NA(IMPLEMENTACAO1$AGE), ]
IMPLEMENTACAO1$AGE <- IFELSE(IS.NA(IMPLEMENTACAO1$
                      $AGE), MEAN(IMPLEMENTACAO1$AGE,
                      NA.RM = TRUE), IMPLEMENTACAO1$AGE)
IMPLEMENTACAO1[, 1:3] = SCALE(BASE[, 1:3])
```

ONDE: _____ EXECUTAR SOMENTE OS COMANDOS →

VAMOS AGORA FAZER UMA INSTALAÇÃO ADICIONAL DE UM PACOTE DO R PARA QUE CONSIGAMOS FAZER A DIVISÃO DA BASE DE DADOS EM BASE DE TREINAMENTO E BASE DE TESTE.

```
INSTALL.PACKAGES("caTools") #INSTALANDO O PACOTE
LIBRARY(caTools)            #USANDO O PACOTE

SET.SEED(1)
DIVISAO <- SAMPLE.SPLIT(BASE$DEFAULT, SPLITRATIO =
                        = 0.75
```

NOTA DO INSTRUTOR } NESSA BASE "DIVISÃO" TEM VÁRIOS "TRUE" E VÁRIOS "FALSE", QUE SIGNIFICAM OS REGISTROS QUE VÃO ENTRAR NA BASE (TRUE) E OS QUE NÃO VÃO ENTRAR NA BASE (FALSE).

```
BASE_TREINAMENTO <- SUBSET(BASE, DIVISAO == TRUE)
BASE_TESTE <- SUBSET(BASE, DIVISAO == FALSE)
```

OBSERVAÇÃO QUANDO FORMOS TRABALHAR COM A BASE CENSO, CONSIDERAR 0.85 PARA A VARIÁVEL DIVISÃO E O RESTANTE PARA TESTE

MUDAR PARA INCOME

```
SET.SEED(1)
DIVISAO <- SAMPLE.SPLIT(BASE$DEFAULT, SPLITRATIO =
                        = 0.85
BASE_TREINAMENTO <- SUBSET(BASE, DIVISAO == TRUE)
BASE_TESTE <- SUBSET(BASE, DIVISÃO == FALSE)
```

RESUMINDO

- BASE → REFERE-SE À BASE DE DADOS CREDIT_DATA (A)
- BASE_CENSO → REFERE-SE À BASE DE DADOS CENSUS_DATA (B)

BASE_TESTE → (A)
BASE_TREINAMENTO → (A)
DIVISÃO → (A)

BASE_TESTE2 → (B)
BASE_TREINAMENTO2 → (B)
DIVISÃO2 → (B)

USADO EM TAREFAS COMO CLASSIFICAÇÃO E MINERAÇÃO DE TEXTOS.

MÓDULO APRENDIZAGEM BAYESIANA

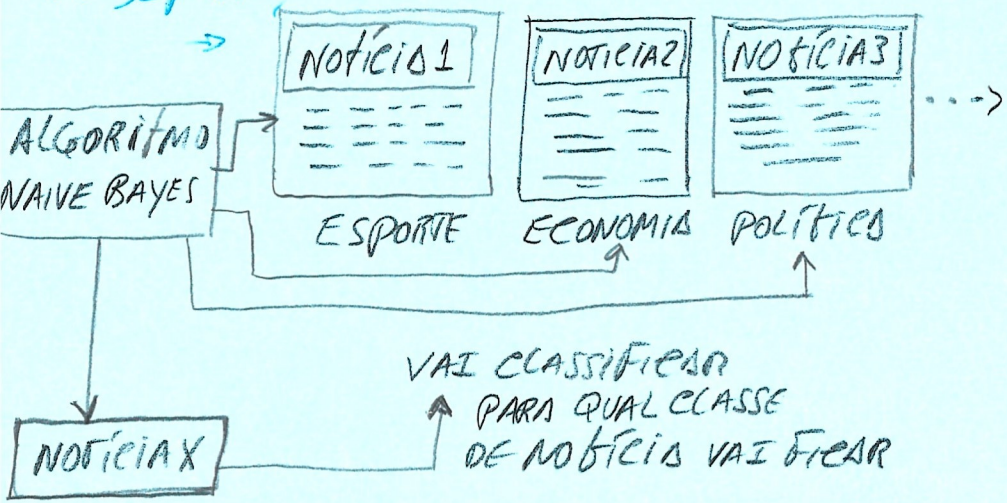
CONTEÚDO

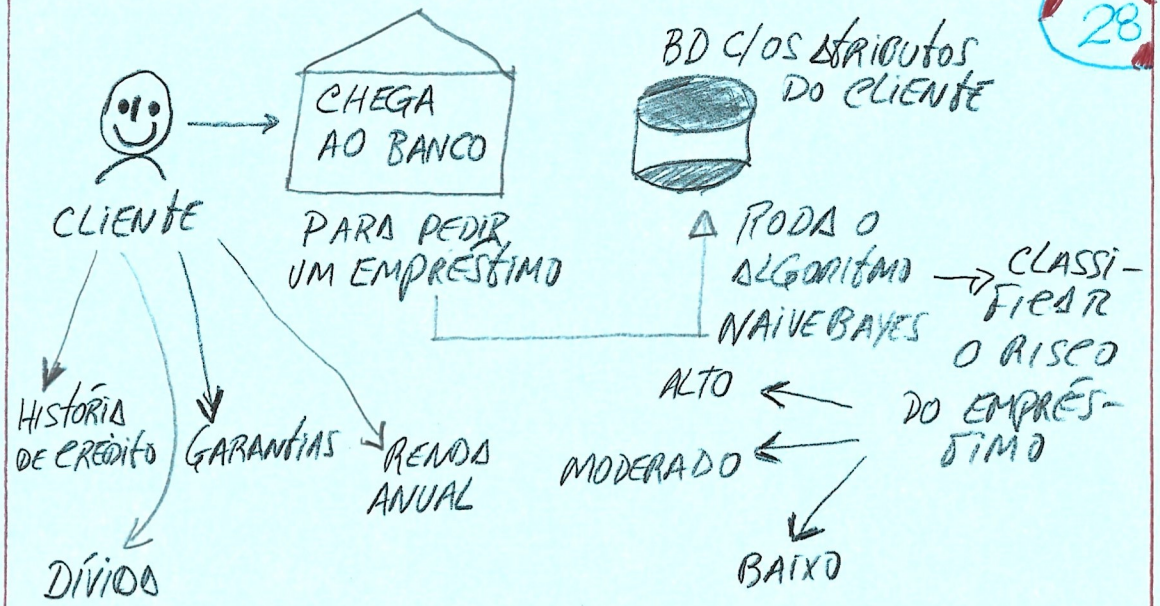
- TEORIA SOBRE O NAIVE BAYES
- NAIVE BAYES EM R - BASE RISCO DE CRÉDITO (BIBLIOTECA e1071)
- NAIVE BAYES EM R - BASE CRÉDITO
- NAIVE BAYES EM R - BASE CENSO

NAIVE BAYES - INTRODUÇÃO # BASEADO EM PROBABILIDADES

- ABORDAGEM PROBABILÍSTICA (TEOREMA DE BAYES)
- EXEMPLOS
  - FILTROS DE SPAM;

- MINERAÇÃO DE EMOÇÕES (😊 ; ☹ ; 😢 , ...)
  DADA UMA FRASE, MINERAR AS EMOÇÕES QUE SE ENCONTRAM NO TEXTO ESCRITO
- SEPARAÇÃO DE DOCUMENTOS →





— BASE ORIGINAL ATRIBUTOS PREVISORES → | CLASSES/META

HISTÓRIA DO CRÉDITO	DÍVIDA	GARANTIAS	RENDA ANUAL	RISCO
RUIM	ALTA	NENHUMA	<15.000	ALTO
DESCONHECIDA	ALTA	NENHUMA	>=15.000 A <=35.000	ALTO
DESCONHECIDA	BAIXA	NENHUMA	>=15.000 A<=35.000	MODERADO
DESCONHECIDA	BAIXA	NENHUMA	>35.000	BAIXO
DESCONHECIDA	BAIXA	ADEQUADA	>35.000	BAIXO
RUIM	BAIXA	NENHUMA	<15.000	ALTO

NAIVE BAYES

	HISTÓRIA DO CRÉDITO			DÍVIDA		GARANTIAS		RENDA ANUAL		
RISCO DE CRÉDITO	BOA 5	DESCONH. 5	RUIM 4	ALTA 7	BAIXA 7	NENHUMA 11	ADEQUADA 3	<15000 3	>=15000 <=35000 4	>35.000 7
ALTO 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0	3/6	2/6	1/6
MODERADO 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3	0	2/3	1/3
BAIXO 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5	0	0	5/5

$\frac{6}{14}$: SIGNIFICA 6 REGISTROS DA CLASSE ALTA DE UM TOTAL DE 14 REGISTROS E ASSIM SUCESSIVAMENTE PARA OS DEMAIS VALORES DA TABELA

NAIVE BAYES - APRENDIZAGEM

	HIST. DE CRED.			DÍVIDA		GARANT.		RENDA ANUAL		
RISCO DE CRED.	BOA 5	DESCONH. 5	RUIM 4	ALTA 7	BAIXA 7	NENH. 11	ADEQ. 3	A 3	B 4	C 7
ALTO 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0	3/6	2/6	1/6
MODER. 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3	0	2/3	1/3
BAIXO 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5	0	0	5/5

HIST/CRÉDITO	RISCO
RUIM	ALTO
DESCONH.	ALTO
DESCONH.	MOD.
DESCONH.	ALTO
DESCONH.	BAIXO
DESCONH.	BAIXO
RUIM	ALTO
RUIM	MOD.
BOA	BAIXO
BOA	BAIXO
BOA	ALTO
BOA	MOD.
BOA	BAIXO
RUIM	ALTO

A: <15.000

B: >=15.000
<=35.000

C: >35.000

TABELA DE PROBABILIDADES

↓

MACHINE LEARNING -> DE NAIVE BAYES

ALGORITMO

NAIVE BAYES - CLASSIFICAÇÃO

↳ ESSE ALGORITMO TRABALHA ANALISANDO UMA BASE HISTÓRICA DE CRÉDITO PARA A CONSTRUÇÃO DE UMA TABELA DE PROBABILIDADES, COM O OBJETIVO DE CLASSIFICAR NOVOS CLIENTES SOLICITANDO UMA LINHA DE CRÉDITO.

POR EXEMPLO, SUPONHAMOS QUE TENHAMOS UM NOVO REGISTRO (CLIENTE) COM OS ATRIBUTOS ABAIXO, E VAMOS SABER QUAL SERIA A CLASSE DELE.

CLIENTE

- HISTÓRIA = BOA
- DÍVIDA = ALTA
- GARANTIAS = NENHUMA
- RENDA = > 35.000

↓

RISCO DE CRED.	BOA 5	ALTA 7	NENHUMA 11	> 35.000 7	RISCO
ALTO 6/14	1/6	4/6	6/6	1/6	BAIXO
MODERADO 3/14	1/3	1/3	2/3	1/3	
BAIXO 5/14	3/5	2/5	3/5	5/5	

$P(ALTO) = 6/14 * 1/6 * 4/6 * 6/6 * 1/6 = 0,0079 = 0,79\%$

$P(MODERADO) = 3/14 * 1/3 * 1/3 * 2/3 * 1/3 = 0,0052$

$P(BAIXO) = 5/14 * 3/5 * 2/5 * 3/5 * 5/5 = 0,0514$

$\sum P(A,M,B) = 0,0079 + 0,0052 + 0,0514 = 0,0645$ 100%

$P(ALTO) = 0,0079/0,0645 * 100 = 12,24\%$

$P(MOD.) = 0,0052/0,0645 * 100 = 8,07\%$

$P(BAIXO) = 0,0514/0,0645 * 100 = 79,69\%$

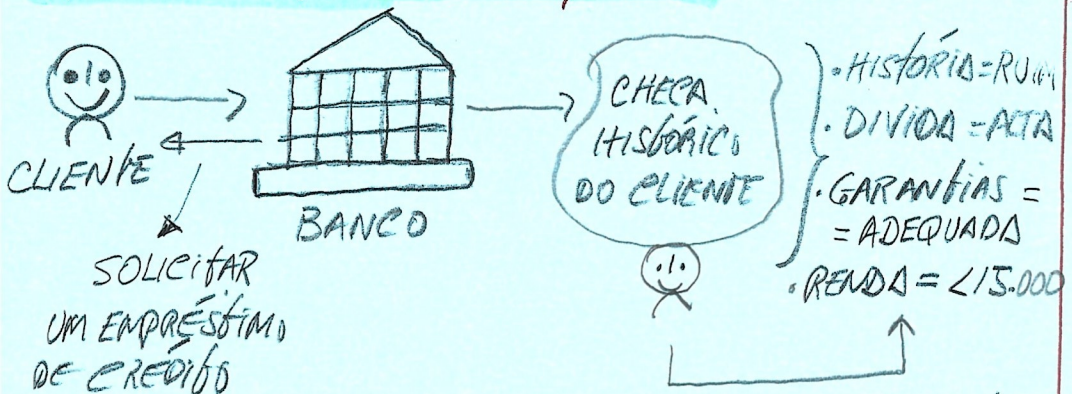
P(B) = 79,69%

↓

PODE SER USADO EM CIMA DO VALOR DO EMPRÉST.

RESULTADO FINAL -> CRÉDITO CONCEDIDO, POIS O RISCO É BAIXO, PODENDO SER CONCEDIDO

NAIVE BAYES - CORREÇÃO LAPLACIANA



FAZENDO O MESMO PROCEDIMENTO FEITO NA PÁGINA 29, ACHAREMOS O SEGUINTE RESULTADO FINAL, ENTRETANTO, ESSE EXEMPLO TEM UMA PARTICULARIEDADE, CONFORME DESCRITO ABAIXO.

$P(ALTO) = 6/14 * 3/6 * 4/6 * 0 * 3/6 = 0$

$P(MODERADO) = 3/14 * 1/3 * 1/3 * 1/3 * 0 = 0$

$P(BAIXO) = 5/14 * 0 * 2/5 * 2/5 * 0 = 0$

OU SEJA, AS TRÊS P(N) = 0 ↓ NESSE CASO, NÃO VAMOS FAZER A PREVISÃO DE RISCO DO FINANCIAMENTO

PARA RESOLVER ESSE PROBLEMA, RECORREMOS A UM ARTIFÍCIO DO ALGORITMO CHAMADO DE CORREÇÃO LAPLACIANA.

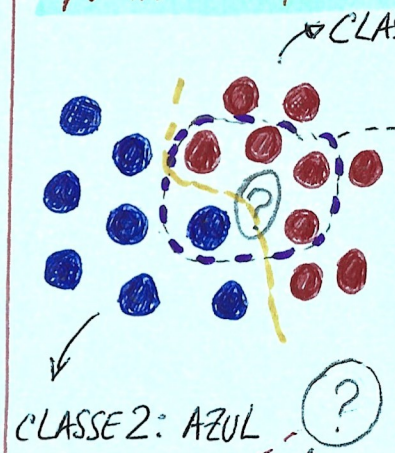
CORREÇÃO LAPLACIANA → ADICIONARÁ UM REGISTRO

NO REGISTRO "RUIM", VALOR = 0, VAI FICAR 1/6. EM BOA E DESCONHECIDA, VAI FICAR 3/6 E 2/6, RESPECTIVAMENTE. NA COLUNA "RISCO DE CRÉDITO", VALOR "BAIXO" VAI FICAR 6/15. A QUANTIDADE DE "RUIM = 4, VAI MUDAR PARA 5.

$P(BAIXO) = 6/15 * 1/6 * 3/6 * 2/6 * ?$

Nota do instrutor ESSA AULA É SOMENTE ILUSTRATIVA DE COMO FUNCIONA ESSA CORREÇÃO. PARA TER OS VALORES EXATOS PARA SEREM USADOS NA FÓRMULA, O IDEAL É ADICIONAR OS NOVOS REGISTROS NA BASE DE DADOS E DEPOIS CONSTRUIR A TABELA DE PROBABILIDADES.

NAIVE BAYES - + CONCEITOS



$P(VERMELHA) = 9/17$

$P(AZUL) = 8/17$

OS ALGORITMOS NAIVE BAYES TÊM UM PARÂMETRO CHAMADO DE RADIUS = RAIO QUE SÃO ALGUNS PONTOS QUE ELE VAI PEGAR QUE ESTÃO DENTRO DESSA ÁREA CHAMADA "RADIUS" PEGANDO ALGUNS VALORES, OU SEJA 4 VERMELHOS E 1 AZUL

$P'(VERMELHO) = 4/9$

$P'(AZUL) = 1/8$

P E P' SÃO CHAMADAS DE PROBABILIDADES APRIORI

PORTANTO, A BOLINHA? VAI SER VERMELHA

VOCÊ TEM UM PONTO QUE NÃO SABEMOS A QUE CLASSE PERTENCE

$P''(VERMELHO) = 9/17 * 4/9 = 0,24$

$P''(AZUL) = 8/17 * 1/8 = 0,06$

... NAIVE BAYES - + CONCEITOS ... CONT...

P'' SÃO CHAMADAS DE PROBABILIDADES POSTEORI

VANTAGENS X DESVANTAGENS

- **VANTAGENS**
  - RÁPIDO;
  - SIMPLICIDADE DE INTERPRETAÇÃO;
  - TRABALHA COM ALTAS DIMENSÕES;
  - BOAS PREVISÕES EM BASES PEQUENAS
- **DESVANTAGENS**
  - COMBINAÇÃO DE CARACTERÍSTICAS (ATRIBUTOS INDEPENDENTES) - CADA PAR DE CARACTERÍSTICAS SÃO INDEPENDENTES - NEM SEMPRE É VERDADE.

BOA * ALTA * < 15 -> INDEPENDENTES

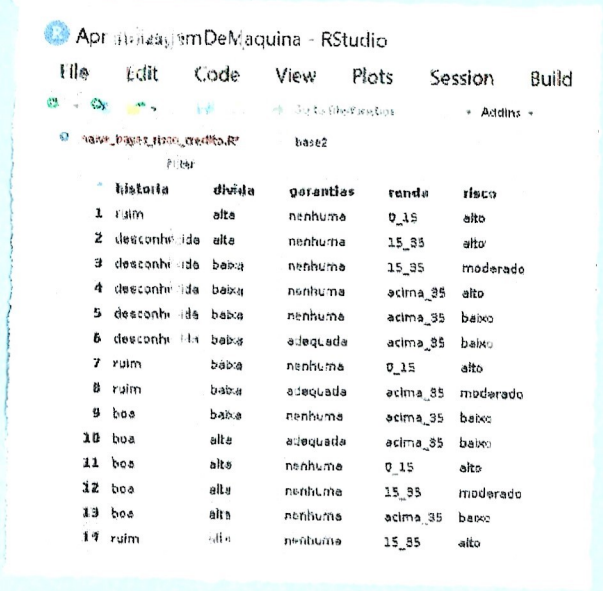
RENDA } PAR
DIVIDA } INDEPENDENTE

IMPLEMENTAÇÃO NAIVE BAYES EM R } BASE RISCO DE CRÉDITO

```
setwd("D:/Machine Learning e Data Science com R") # Set As Working Directory
```

NO CONSOLE DO RSTUDIO, VAMOS CRIAR NOVO SCRIPT DE COMANDOS DE IMPLEMENTAÇÃO → SALVANDO-O COM O NOME DE "NAIVE_BAYES_RISCO_CREDITO.R" # A EXTENSÃO .R, O R SALVA AUTOMATICAMENTE.

```
BASE2 <- read.csv('risco_credito.csv')
BASE2
```



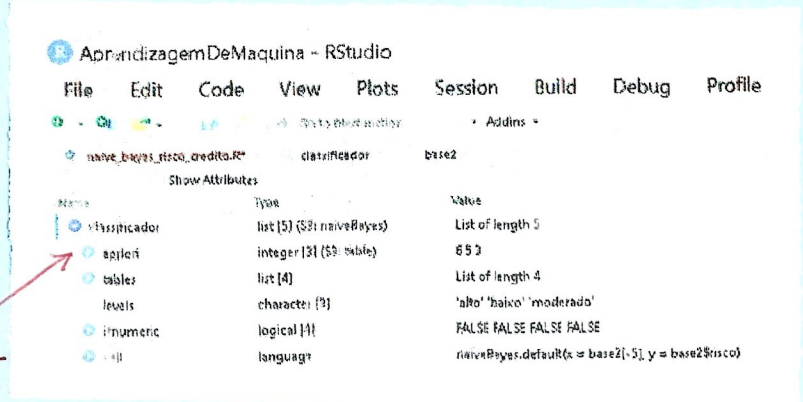
	historia	divida	garantias	renda	risco
1	ruim	alta	nenhuma	0_15	alto
2	desconhecida	alta	nenhuma	15_35	alto
3	desconhecida	baixa	nenhuma	15_35	moderado
4	desconhecida	baixa	nenhuma	acima_35	alto
5	desconhecida	baixa	nenhuma	acima_35	baixo
6	desconhecida	baixa	adequada	acima_35	baixo
7	ruim	baixa	nenhuma	0_15	alto
8	ruim	baixa	adequada	acima_35	moderado
9	boa	baixa	nenhuma	acima_35	baixo
10	boa	alta	adequada	acima_35	baixo
11	boa	alta	nenhuma	0_15	alto
12	boa	alta	nenhuma	15_35	moderado
13	boa	alta	nenhuma	acima_35	baixo
14	ruim	alta	nenhuma	15_35	alto

```
install.packages("e1071") # Instalando Pacote.
library(e1071)            # Carregando o Pacote
```

```
# Criando o Classificador
classificador <- naiveBayes(x = base2[-5], y = base2$risco)
```

↓

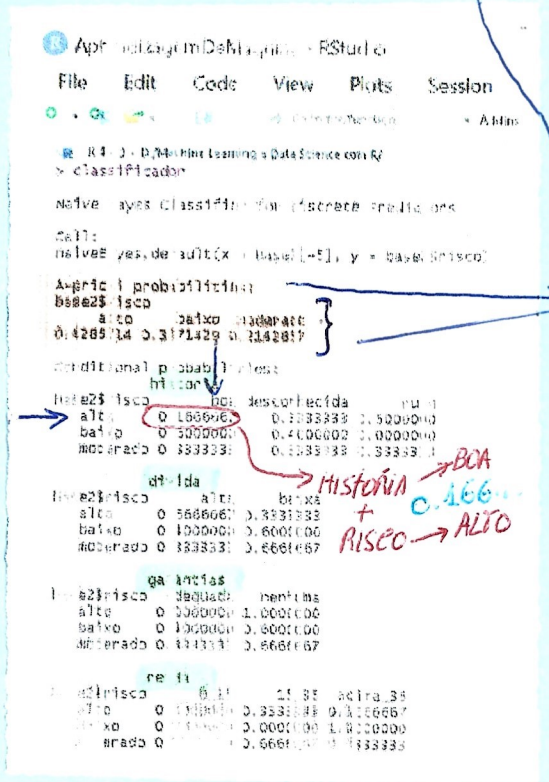
DEPOIS QUE CRIAR O CLASSIFICADOR NAIVE BAYES NO SCRIPT, RODA-SE O COMANDO ACIMA E GERA-SE A TABELA (QUADRO) A SEGUIR



EXPANDINDO-SE, TEREMOS:

RISCOS → ALTO = 6
→ BAIXO = 5
→ MODERADO = 3

O TREINAMENTO DO ALGORITMO É JUSTAMENTE A GERAÇÃO DA TABELA DE P(n).



SE JOGAR ESSES VALORES NA TABELA DE PROBABILIDADES CONSTRUÍDA PELO ALGORITMO (VER PÁGINA 29) VAMOS ACHAR ESSES MESMOS VALORES.

HISTÓRIA → BOA → 0.166... + RISCO → ALTO

VAMOS SUPOR QUE QUEIRAMOS CLASSIFICAR UM NOVO REGISTRO, CONFORME TRAZIDO ABAIXO:

HISTÓRIA: BOA, DÍVIDA: ALTA, GARANTIAS: NENHUMA, RENDA: >35 (1)

HISTÓRIA: RUIM, DÍVIDA: ALTA, GARANTIAS: ADEQUADA, RENDA: <15 (2)

QUAL SERÁ A CLASSE PARA ESSES DOIS NOVOS REGISTROS { ALTO, BAIXO, MODERADO } RISCO?

PARA ISSO, VAMOS PRECISAR CRIAR UM NOVO DATAFRAME COM ESSES NOVOS REGISTROS, CONFORME TRAZIDO ABAIXO.

```
historia <- c("boa")
divida <- c("alta")
garantias <- c("nenhuma")
renda <- c("acima_35")
df <- data.frame(historia, divida, garantias, renda)
previsao <- predict(classificador, newdata = df)
```

PARA VER O RESULTADO, BASTA DIGITAR NO PROMPT DO RSTUDIO O COMANDO print(previsao) E SAI O RESULTADO ABAIXO:

[1] BAIXO. #PORTANTO, O CLIENTE 1 TEM RISCO BAIXO PARA O FINANCIAMENTO DE CRÉDITO.

... QUERENDO VER AS PROBABILIDADES, TEM-SE:

```
PREVISAO <- PREDICT(CLASSIFICADOR, NEWDATA =
                    = DF, 'RAW')
PRINT(PREVISAO) ↵ # SAÍDA MOSTRADA ABAIXO
```

	ALTO	BAIXO	MODERADO
[1,]	0.1227496	0.7941734	0.08183306
[2,]	0.8993525	0.0007194	0.09992805

NAIVE BAYES EM R - BASE CRÉDITO

CRIA-SE UM NOVO ARQUIVO DE SCRIPT COM O NOME: NAIVE_BAYES_CREDIT_DATA COM AS LINHAS DE CÓDIGO ABAIXO:

```
#LEITURA DA BASE DE DADOS
BASE <- READ.CSV('CREDIT_DATA.CSV')
#APAGA A COLUNA CLIENTID
BASE$CLIENTID = NULL

#TRATA VALORES INCONSISTENTES
BASE$AGE <- IFELSE(BASE$AGE < 0, 40.92,
                   , BASE$AGE)

#FAZENDO O ESCALONAMENTO
BASE[, 1:3] <- SCALE(BASE[, 1:3])
```

```
#DIVISÃO ENTRE TREINAMENTO E TESTE
LIBRARY(CATOOLS)
SET.SEED(1)
DIVISAO <- SAMPLE.SPLIT(BASE$INCOME
        , SPLITRATIO = 0.75)

BASE_TREINAMENTO <- SUBSET(BASE, DIVISAO ==
                 == TRUE)

BASE_TESTE <- SUBSET(BASE, DIVISAO == FALSE)
```

~~VAMOS CRIAR UM NOVO SCRIPT~~

```
CLASSIFICADOR2 <- NAIVEBAYES(X = BASE_TREINAMENTO[
               -4], Y = BASE_TREINAMENTO$DEFAULT

PRINT(CLASSIFICADOR2) # MOSTRA QUE 86% É À CLASSE
    0 E 14% É À CLASSE 1. SENDO
    QUE "0" NÃO PAGA O FINANCIAMENTO E "1" PA-
    GA O FINANCIAMENTO BANCÁRIO
```

NOTA DO INSTRUTOR } ESSA BASE DE TREINAMENTO TEM SOMENTE DADOS (VARIÁVEIS NUMÉRICAS)

↓

OU SEJA, A TABELA DE PROBABILIDADES TAMBÉM É GERADA

```
#VAMOS AGORA FAZER AS NOVAS PREVISÕES USANDO
 A BASE DE TESTE JÁ CRIADA (VER PÁGINA 27)
```

... CONTINUAÇÃO

```
PREVISOES <- predict(classificador, newdata =
                     base_teste[-4])

VIEW(PREVISOES)
```

NAME	TYPE	VALUE
PREVISOES	FACTOR	FACTOR WITH 2 LEVELS: "0", "1"

DEFAULT → OS VALORES DAS BASES DE TREINAMENTO E DE TESTE ESTÃO COMO VARIÁVEIS (VALORES NUMÉRICOS) E NA REALIDADE ELES SÃO NA REALIDADE VARIÁVEIS CATEGÓRICAS.

0
0
1
⋮

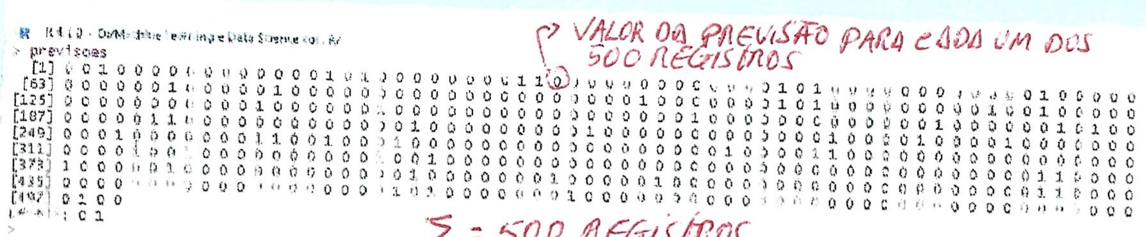
```
# ENCODE DA CLASSE
BASE$DEFAULT <- FACTOR(BASE$DEFAULT, LEVELS =
                       = c(0, 1))
```

NOTA DO INSTRUTOR } VAMOS REFAZER TUDO NOVAMENTE PARA ADEQUAR A BASE ORIGINAL JÁ TRANSFORMANDO AS VARIÁVEIS, INCLUSIVE A CLASSE META QUE É A DEFAULT.

```
# LEITURA DA BASE DE DADOS
BASE <- READ.CSV('credit_data.csv')
# APAGANDO A COLUNA CLIENTID
BASE$CLIENTID <- NULL
# TRATANDO VALORES INCONSISTENTES
BASE <- IFELSE(BASE$AGE < 0, 40.92, BASE$AGE)
# TRATANDO VALORES FALTANTES (NAS)
BASE$AGE <- IFELSE(IS.NA(BASE$AGE), MEAN(
            (BASE$AGE, NA.RM = TRUE), BASE$
            $AGE)
# FAZENDO O ESCALONAMENTO
BASE[, 1:3] <- SCALE(BASE[, 1:3])
# ENCODE DA CLASSE
BASE$DEFAULT <- FACTOR(BASE$DEFAULT, LEVELS=
             = c(0, 1))
# DIVISÃO DA BASE EM TREINAMENTO E TESTE
LIBRARY(caTools)
SET.SEED(1)
DIVISAO <- SAMPLE.SPLIT(BASE$INCOME, SPLITRATIO =
        =0.75
#CRIANDO AS BASES TREINAMENTO E TESTE
BASE_TREINAMENTO <- SUBSET(BASE, DIVISAO ==
                 == TRUE)
BASE_TESTE <- SUBSET(BASE, DIVISAO == FALSE)
LIBRARY(e1071)
CLASSIFICADOR1 <- NAIVEBAYES(X= BASE_TREINAMEN-
```

... TO[-4], Y = BASE_TREINAMENTO $DEFAULT)
PRINT(CLASSIFICADOR1)

PREVISOES <— PREDICT(CLASSIFICADOR1, NEWDATA =
= BASE_TESTE[-4])



↱ VALOR DA PREVISÃO PARA CADA UM DOS 500 REGISTROS

Σ = 500 REGISTROS

	INCOME	AGE	LOAN	DEFAULT
4	- 0.1830...	0.3640...	0.5448...	0

ENTRETANTO, NO CONTEXTO DESTA TABELA, "0" É ACERTO.

↘ PEGA ESSE CAMPO E SUBMETE À TABELA DE PROBABILIDADES GERADA PELO ALGORITMO NAIVE BAYES E CLASSIFICOU O REGISTRO PARA A CLASSE 0 (NÃO VAI PAGAR)
E ASSIM, É FEITO PARA TODOS OS 500 REGISTROS.

MATRIZ_CONFUSAO <— TABLE(BASE_TESTE[, 4], PREVISOES)
PRINT(MATRIZ_CONFUSAO)

	PREVISOES	
	0	1
0	421	6
1	26	47

MATRIZ DE CONFUSÃO AJUDA A AVALIAR O DESEMPENHO DO ALGORITMO EM TERMOS DE ACERTOS E ERROS NAS PREVISÕES

NOTA DO INSTRUTOR } PERCENTUAL DE ACERTOS =
= 421 + 47 = 468/500 = 0,936 X
X 100 = 93,6 %

↓

PERCENTUAL DE ERROS

26 + 6 = 32/500 = 0,064 X 100 = 6,4 %

VISUALIZANDO O PERCENTUAL DE ACERTO

INSTALL.PACKAGES("CARET")
LIBRARY(CARET) # CLASSIFICATION AND REGRESSION TRAINING.

CONFUSIONMATRIX(MATRIZ_CONFUSAO)

	PREVISOES	
	0	1
0	421	6
1	26	47

PRECISÃO DO ALGORITMO ↗

ACCURACY : 0.936
95% CI : (0.9108, 0.9558)

NO INFORMATION RATE : 0.894
P-VALUE [ACC > NIR] : 0.0007810

KAPPA : 0.7105

MCNEMAR'S TEST P-VALUE : 0.0007829

→

...CONT

```
SENSITIVITY : 0.9418
SPECIFICITY : 0.8868
POS PRED VALUE: 0.9859
NEG PRED VALUE: 0.6438
PREVALENCE : 0.8940
DETECTION RATE: 0.8420
DETECTION PREVALENCE : 0.8540
BALANCED ACCURACY: 0.9143
'POSITIVE' CLASS : 0
```

VAMOS SALVAR OS RESULTADOS EM UM ARQUIVO

↳ RESULTADOS_CREDIT_DATA.TXT

```
1 RESULTADOS CREDIT DATA
2 ----------
3
4 0.936 - NAIVE BAYES - INCONSISTENCIAS + FALTANTES +
  + ESCALONAMENTO

- NAIVE BAYES - INCONSISTENCIAS + FALTANTES

- NAIVE BAYES - SEM PRÉ-PROCESSAMENTO
```

```
# LEITURA DA BASE DE DADOS
BASE <- READ.CSV('CREDIT_DATA.CSV')
# APAGA-SE A COLUNA CLIENTID
BASE$CLIENTID <- NULL
# ENCODE DA CLASSE
BASE$DEFAULT <- FACTOR(BASE$DEFAULT, LEVELS =
                       = C(0, 1))
# DIVISÃO ENTRE TREINAMENTO E TESTE
LIBRARY(CATOOLS)
SET.SEED(1)
DIVISAO <- SAMPLE.SPLIT(BASE$INCOME, SPLITRA-
                        TIO = 0.75)
BASE_TREINAMENTO <- SUBSET(BASE, DIVISAO == TRUE)
BASE_TESTE <- SUBSET(BASE, DIVISAO == FALSE)

LIBRARY(E1071)
CLASSIFICADOR <- NAIVEBAYES(X = BASE_TREINAMENTO[
                 [-4], Y = BASE_TREINAMENTO$DEFAULT)
#PRINT(CLASSIFICADOR)
PREVISOES <- PREDICT(CLASSIFICADOR, NEWDATA =
             = BASE_TESTE[-4])
MATRIZ_CONFUSAO <- TABLE(BASE_TESTE[, 4], PRE-
                   VISOES)
# PRINT(MATRIZ_CONFUSAO)
# INSTALL.PACKAGES("CARET")
LIBRARY(CARET)
CONFUSIONMATRIX(MATRIZ_CONFUSAO)
```

...cont

SEM PRÉ-PROCESSAMENTO ⇒ ACCURACY = 0.932

NAIVE BAYES EM R - BASE CENSO

↳ VAMOS CRIAR NOVO SCRIPT E SALVAR O MESMO COM O NOME "NAIVE_BAYES_CENSUS.R"...

DEPOIS DE RODARMOS TODO O TEMPLAYTE, VAMOS AGORA IMPLEMENTAR O ALGORITMO NAIVE BAYES PARA A BASE CENSO.

```
# 1º Vamos carregar a biblioteca necessária
library(e1071)
# Criando o novo classificador Naive Bayes
classificador <- naiveBayes(x = base_treinamento[-15], y = base_treinamento$income)
print(classificador) # mostra a tabela de probabilidades

#Criando as previsões
previsoes <- predict(classificador, newdata = base_teste[-15])
Previsoes # digita no prompt de comando.

#Vamos agora comparar os valores da base de teste que já sabemos com os valores previstos de forma a ver o quão o algoritmo acertou e errou.
#Vamos construir a matriz de confusão
matriz_confusao <- table(base_teste[, 15], previsoes)
print(matriz_confusao)
```

	Previsões 0	Previsões 1
0	3166	232
1	576	550

Erros do algoritmo: (1,0) (0,1)

Acertos do algoritmo: (0,0) (1,1)

- Acertou 3166 "0", ou seja, classificados como "0";
- Teve 232 erros para a classe "0";
- Teve 576 erros para a classe "1", ou seja errou + que acertou (550) para esta classe

```
# Carregando a biblioteca "caret"
library(caret) # permite ver o percentual de acerto
confusionMatrix(matriz_confusao)
```

Confusion Matrix and Statistics

	Previsões 0	Previsões 1
0	3166	232
1	576	550

... Cont

Accuracy : 0.8214
95% CI : (0.8099, 0.8325)
No Information Rate : 0.8271
P-Value [ACC > NIR] : 0.8512

Kappa : 0.468

Mcnemar's Test P-Value : < 2.e-16

Sensitivity : 0.9461
Specificity : 0.7033
Pos Pred Value : 0.9317
Neg Pred Value : 0.4985
Prevalence : 0.8271
Detection Rate : 0.6978
Detection Prevalence : 0.7511
Balanced Accuracy : 0.7747

'Positive' Class : 0

Podemos, também, fazer os mesmos procedimentos para avaliar os resultados

Resultados Censos

- Naive Bayes - Categóricos + escalonamento
- Naive Bayes - Categóricos
- Naive Bayes - Escalonamento
- Naive Bayes - sem pré-processamento

Nota do instrutor } O algoritmo Naive Bayes não utiliza cálculos de distâncias, por isso que não é necessário fazer o escalonamento como não queime.

Referências Complementares

- O instrutor recomenda seu curso Mineração de Emoção em Textos com Python e NLTK para quem deseja usar o algoritmo Naive Bayes para tarefas de classificação;
- O livro Thoughtful Machine Learning de Matthew Kirk: o Capítulo 4 apresenta mais sobre a teoria do Naive Bayes e mostra como contribuir um classificador de spam;
- Livro Data Algorithms de Mahmoud Parsian: o Capítulo 14 apresenta explicações sobre a teoria do Naive Bayes.

APRENDIZAGEM POR ÁRVORES DE DECISÃO

Outro paradigma de aprendizado de máquina, o qual terá o conteúdo programático abaixo:

- Teoria sobre árvores de decisão;
- Árvores de decisão em R - base risco de crédito;
- Árvores de decisão em R - base censo;
- Poda em árvores de decisão (base Censo);
- Random Forests (Florestas randômicas) # Essa técnica é usada no Knect da MS para detecção de movimentos

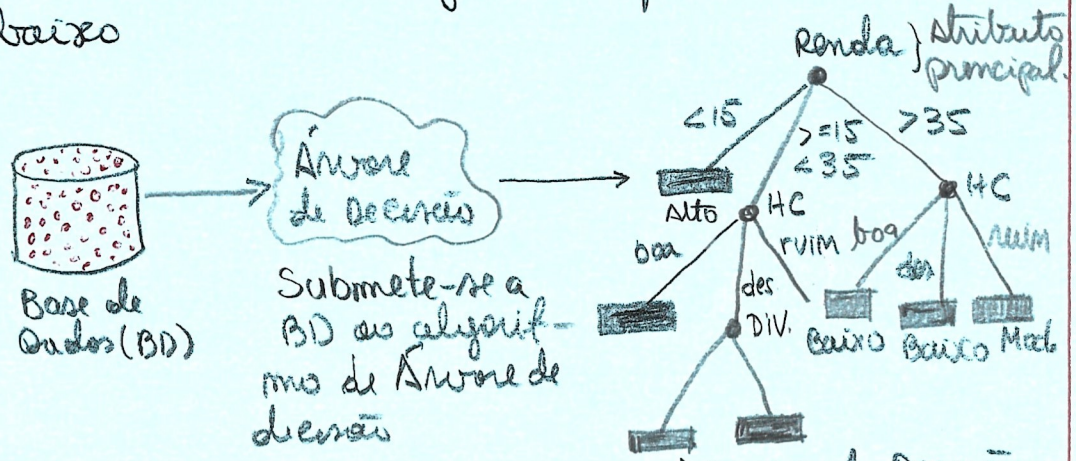
1) INTRODUÇÃO

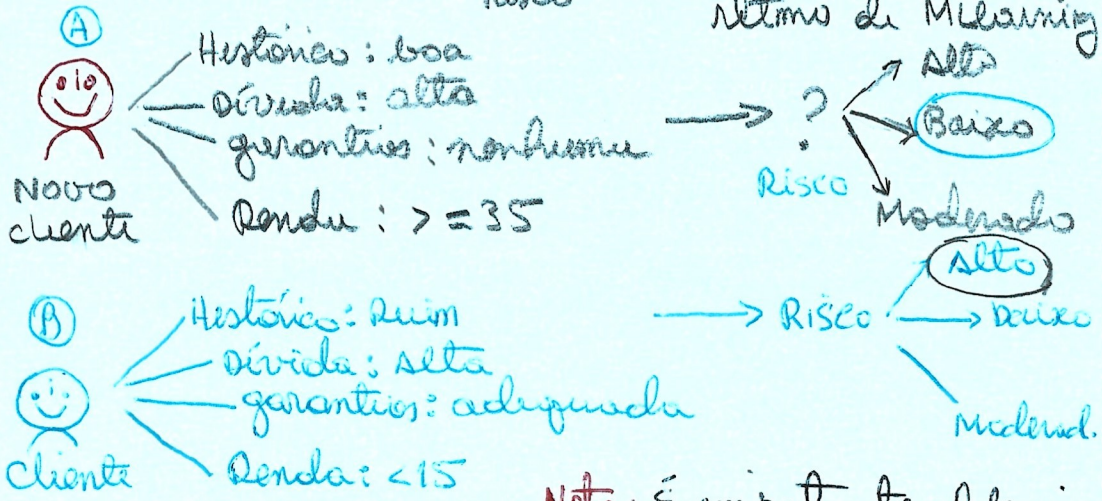
* Moderado
Clone meta

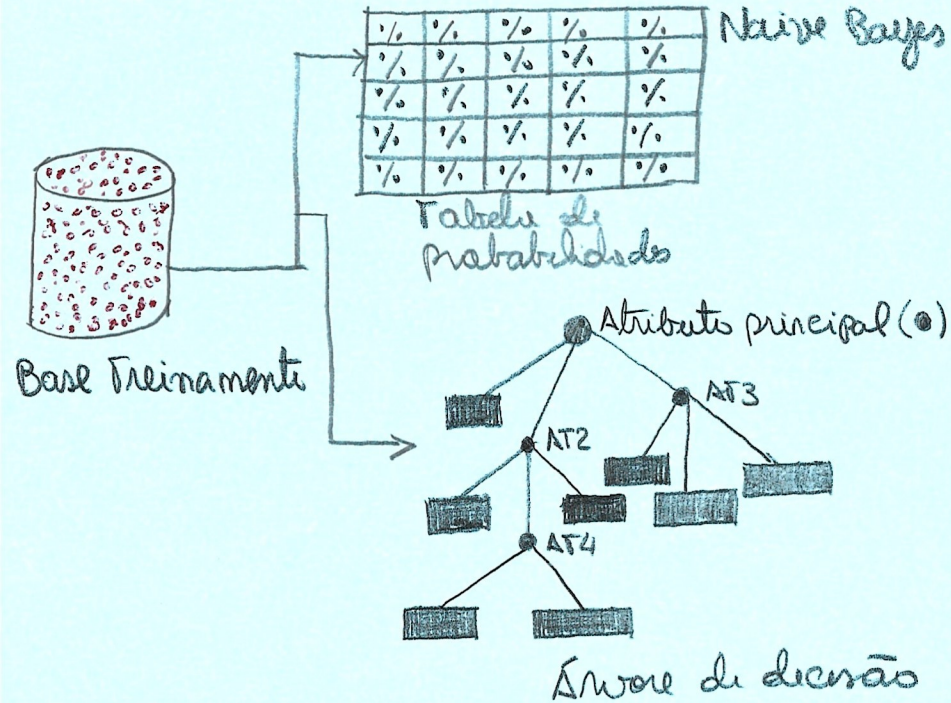
Base Original

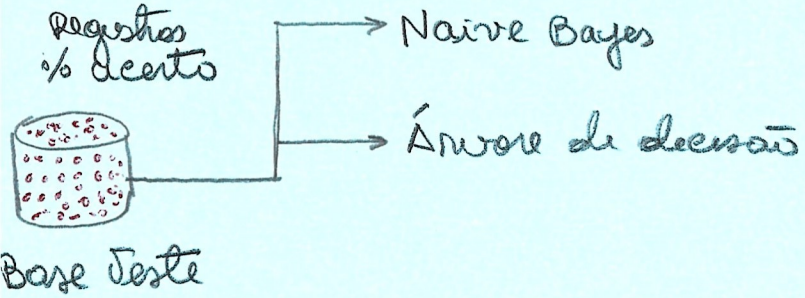
Histórico de crédito	Dívida	Garantias	Renda Anual	Risco
Ruim	alta	nenhuma	< 15 000	alto
Desconhecida	alta	nenhuma	>= 15000 a <= 35.000	alto
Desconhecida	baixa	nenhuma	>= 15000 a <= 35000	M*
Desconhecida	baixa	nenhuma	> 35000	alto
⋮	⋮	⋮	⋮	⋮

A ideia é fazer a previsão do risco em "alto", "moderado" ou "baixo" baseado nos atributos preditores (histórico de crédito, dívida, garantias e renda anual) usando algoritmo de árvore de decisão, veja o esquema ilustrativo abaixo









ÁRVORE DE DECISÃO - APRENDIZAGEM 1

$$Entropy(S) = \sum_{i=1}^{c} -P_i \log_2 P_i$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

CLASSE META RISCO

A	A	M	A	B	B	A	M	B	B	A	M	B.	A

ONDE:

A: ALTO; M: MODERADO; B: BAIXO

ALTO = 6/14

MODERADO = 3/14

BAIXO = 5/14

$$ENTROPY(S) = \sum_{i=1}^{c} -P_i \log_2 P_i$$

$$E(S) = -6/14 * LOG(6/14; 2) - 3/14 * LOG(3/14; 2) - 5/14 * LOG(5/14; 2) = \boxed{1,53}$$

NO EXCEL, FEZ-SE:

ENTROPIA			GANHO DE INFORMAÇÃO
VALOR1	VALOR 2	VALOR 3	0,266429
0,43	0,21	0,36	

RESULTADO
1,53

G(HISTORIA)
G(DIVIDA)
G(GARANTIAS)
G(RENDA)

HISTORIA DE CREDITO (14)

- BOA (5/14)
  - ALTO (1/5)
  - MODERADO (1/5)
  - BAIXO (3/5)
- DESCONHECIDA (5/14)
  - ALTO (2/5)
  - MOD (1/5)
  - BAIXO (2/5)
- RUIM (4/14)
  - ALTO (3/4)
  - MOD (1/4)
  - BAIXO (0)

$E(S) = -1/5 * LOG(1/5; 2) - 1/5 * LOG(1/5; 2) - 3/5 * LOG(3/5; 2) = \boxed{1,37}$ # HISTORIA DE CREDITO BOA

$E(S) = -2/5 * LOG(2/5; 2) - 1/5 * LOG(1/5; 2) - 2/5 * LOG(2/5; 2) = \boxed{1,52}$ # HISTORIA DE CREDITO DESCONHECIDA

$E(S) = -3/4 * LOG(3/4; 2) - 1/4 * LOG(1/4; 2) - 0 * LOG(0; 2) = \boxed{0,81}$

... CONT

$$GAIN(S,A) = ENTROPY(S) - \sum_{v \in VALUES(A)} \frac{|S_v|}{|S|} ENTROPY(S_v)$$

GANHO(HISTÓRIA) = 1,53 − (5/14 * 1,37) − (5/14 * 1,52) − (4/14 * 0,81) = **0,26**

DÍVIDA/RISCO

E(S) = − 4/7 * LOG(4/7; 2) − 1/7 * LOG(1/7; 2) − 2/7 * LOG(2/7; 2) = 1,38

⋮

GANHO(DÍVIDA) = 1,53 − (7/14 * 1,38) − (7/14 * 1,56) = **0,06**

GARANTIAS/RISCO

E(S) = − 6/11 * LOG(6/11; 2) − 2/11 * LOG(2/11; 2) − 3/11 * LOG(3/11; 2) = 1,44

⋮

GANHO(GARANTIAS) = 1,53 − (11/14 * 1,44) − (3/14 * * 0,92) = **0,20**

RENDA/RISCO

E(S) = − 3/3 * LOG(3/3; 2) − 0 * LOG(0; 2) − 0 * LOG(0; 2) = 0,0

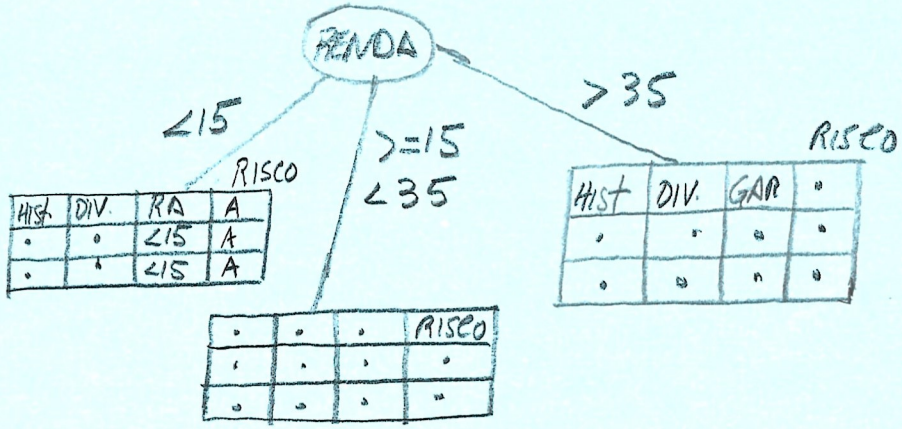
⋮

GANHO(RENDA) = 1,53 − (3/14 * 0,00) − (4/14 * 1,0) − (4/14 * 1,00) − (7/14 * 1,15) = **0,66**

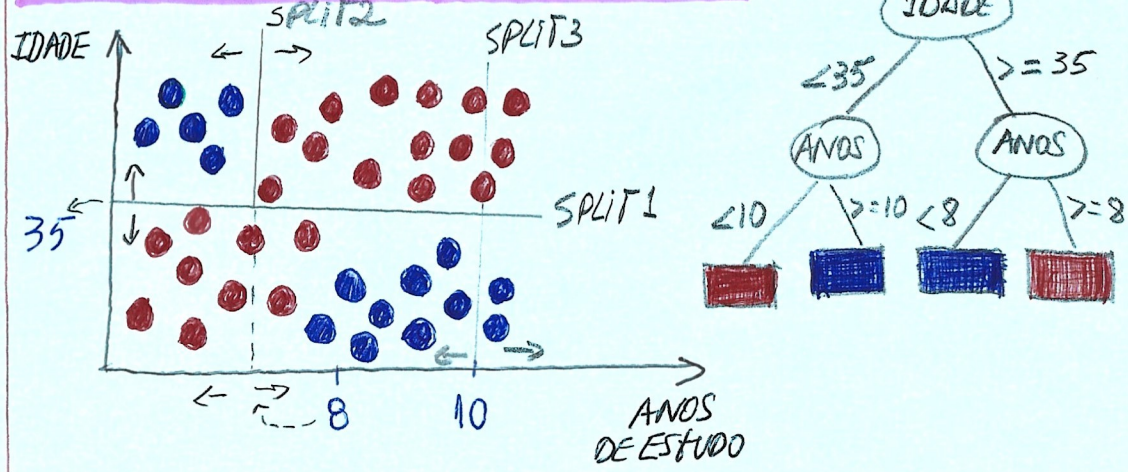
RESUMO | GANHO(S)

- HISTÓRIA DE CRÉDITO = 0,26
- DÍVIDA = 0,06
- GARANTIAS = 0,20
- RENDA = 0,66

Portanto, o NÓ RAIZ da ÁRVORE será "RENDA"



ÁRVORES DE DECISÃO - + CONCEITOS



A SEGUIR, VEREMOS O CONCEITO DE PODA EM ÁRVORES DE DECISÃO

PODA EM ÁRVORES DE DECISÃO

- BIAS (VIÉS)
  - ERROS POR CLASSIFICAÇÃO ERRADA
- VARIÂNCIA
  - ERROS POR SENSIBILIDADE PEQUENAS A MUDANÇAS NA BASE DE TREINAMENTO;
- PODE LEVAR A OVERFITTING; ACONTECE QUANDO O ALGORITMO SE ADAPTA DEMAIS AOS DADOS DE TREINAMENTO, OU SEJA, QUANDO DAMOS UMA NOVA BASE DE DADOS, O ALGORITMO COMETE MUITOS ERROS NO DESEMPENHO.
  - O ALGORITMO TEM UM ÓTIMO DESEMPENHO DE TREINAMENTO.
  - É SEMELHANTE A UM ALUNO QUE DECORA UM CONTEÚDO EM VEZ DE APRENDER.

- VANTAGENS
  - FÁCIL INTERPRETAÇÃO;
  - NÃO PRECISA NORMALIZAÇÃO OU PADRONIZAÇÃO;
  - RÁPIDO PARA CLASSIFICAR NOVOS REGISTROS
- DESVANTAGENS
  - GERAÇÃO DE ÁRVORES MUITO COMPLEXAS
  - PEQUENAS MUDANÇAS NOS DADOS PODE MUDAR A ÁRVORE (PODA PODE AJUDAR).
  - PROBLEMA NP- COMPLETO PARA CONSTRUIR A ÁRVORE

- ERAM MUITO POPULARES EM MEADOS DOS ANOS 90
- Upgrades como random forest (florestas randômicas) melhoram o desempenho (usado no Kinect da Microsoft) #Kinect é um aplicativo para captura dos movimentos do usuário e na sua reprodução.

Imagem do Kinect da MSI soft (google Imagens)

- CART - Classification And Regression Trees

Árvores de decisão em R - Base Risco de Crédito

Vamos nesse tópico, implementar uma árvore de decisão no R usando como Base de dados a base Risco de Crédito

1º passo: Criar novo arquivo de Script, e depois salvar como "arvore_decisao_risco_-credito.R"

```
> base <- read.csv('risco_credito.csv')  # Na área de trabalho do RStudio, podemos observar que base tem 14 obs. de 5 variáveis.
```

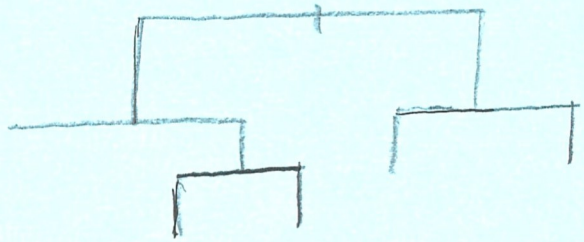
Nota do instrutor } Para implementarmos árvores de decisão no R, precisamos instalar o pacote RPART

```
> install.packages("rpart") ↵
> library(rpart) ↵
# Criando o classificador
> classificador <- rpart(formula = risco ~.,
                         data = base)
> print(classificador)
n= 14
node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 14 8 alto (0.4285714 0.3571429
                   0.2142857)*
```

> plot(classificador) # após a execução deste comando, tem-se um erro:

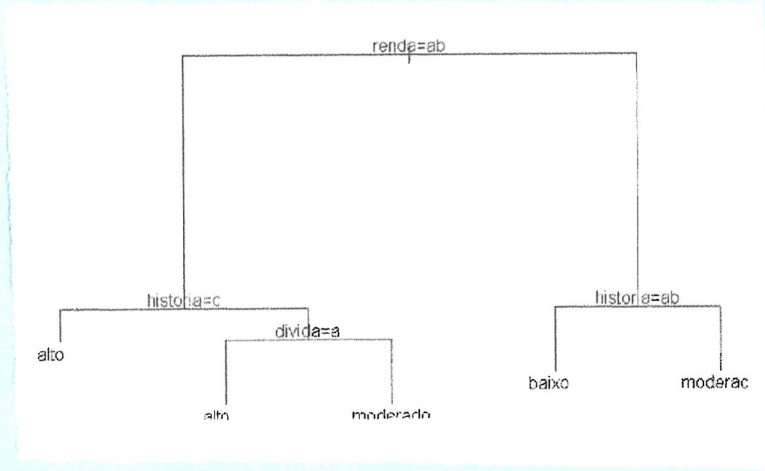
Error in plot.rpart(classificador): fit is not a tree, just a root # Isso ocorre porque a nossa base de dados é muito pequena. Para resolver esse problema, vamos passar um parâmetro adicional dentro da modelagem do classificador, conforme é mostrado a seguir

```
> classificador <- rpart(formula = risco ~.,
    , data = base, control = rpart.control(
    (minbucket = 1)) ↵ # Todo esse trecho é o aprendizado de máquina.
> print(classificador) ↵
> plot(classificador) ↵
```



Para colocar os rótulos dos nós, procede da forma abaixo

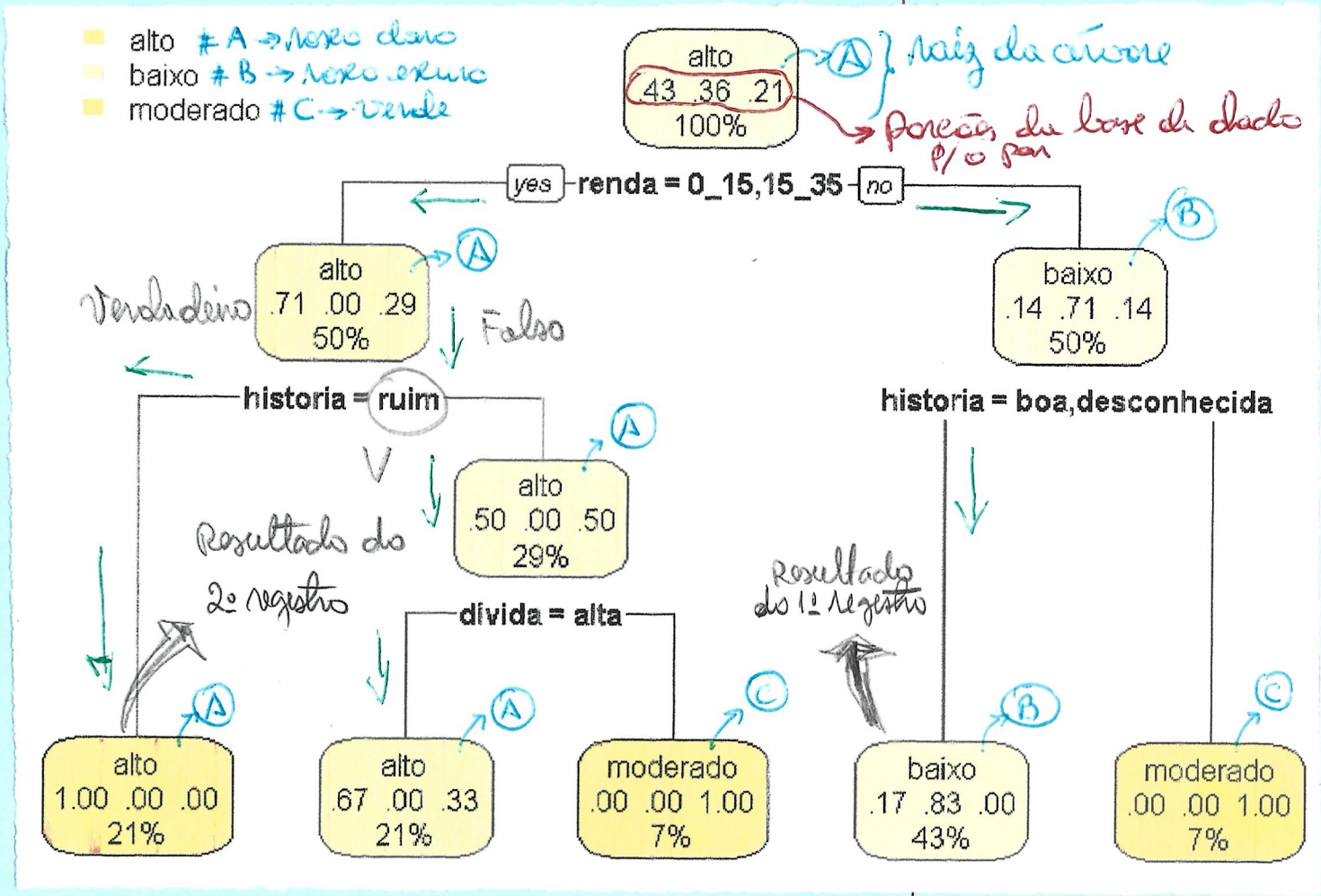
```
> text(classificador)
```



Nota do instrutor } Para visualizarmos a árvore de uma forma + interessante, vamos instalar outro pacote chamado Rpart.plot

```
> install.packages("rpart.plot")
> library(rpart.plot)
```

> rpart.plot(classificador)



> renda2 = c("acima_35", "0_15")

> df <- data.frame(historia, divida, garantias, renda2)

#Vamos agora classificar esses dois registros:

> previsao <- predict(classificador, newdata = df)

> print(previsao)

	alto	baixo	moderado
1	0.1666667	0.8333333	0
2	1.0000000	0.0000000	0

#Criando um data frame novo para testar o classificador

> historia <- c("boa", "ruim")

> divida <- c("alta", "alta")

> garantias <- c("nenhuma", "adequada")

ÁRVORES DE DECISÃO EM R - BASE CRÉDITO

Criar novo arquivo de Script com o nome "arvore_decisao_credit_data.R"

Leitura da base de dados

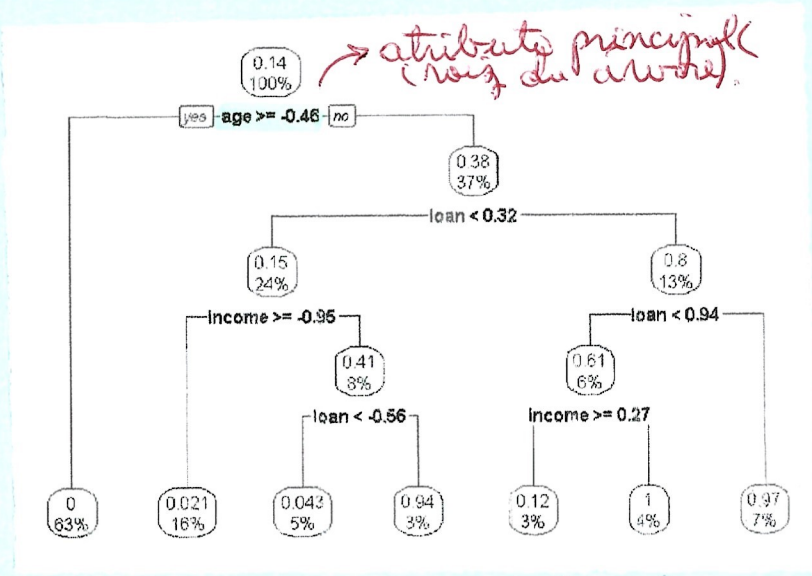
base <- read.csv('credit_data.csv')

apaga a coluna clientid

base$clientid <- NULL

```
# Trata os valores inconsistentes
>base$age <- ifelse(base$age < 0, 40.92, base$age)

# Trata os valores faltantes
>base$age <- ifelse(is.na(base$age), mean(
                (base$age, na.rm = TRUE),
                , base$age)

# Faz o escalonamento
> base[, 1:3] <- scale(base[, 1:3])

# Divide base em base treinamento e teste
>library(caTools)
> set.seed(1)
> divisao <- sample.split(base$income,
                , splitRatio = 0.75)

> base_treinamento <- subset(base, divisao ==
                == TRUE)

> base_teste <- subset(base, divisao == FALSE)
> library(rpart)
> classificador <- rpart(formula = default ~.,
                , data = base_treinamento)
> print(classificador)
```

```
> library(rpart.plot)
> rpart.plot(classificador)
```



→ atributo principal (raiz da árvore)

Árvore de decisão: O algoritmo concluiu que o atributo "age" é o atributo principal

Nota do instrutor: Para que a idade não apareça < 0, basta refazer o pré-processamento sem todo o escalonamento.

```
#Vamos agora submeter a base teste para a
árvore de decisão construída. Esse teste vai ser-
vir para avaliar o desempenho da árvore de
decisão construída na classificação dos registros.
> previsoes <- predict(classificador, newdata =
                = base_teste[-4])

# Encode da classe
> base$default <- factor(base$default,
                , levels = c(0,1))
```

Base teste

	income	age	loan	default
4	-0.16302429	0.364045494	0.54484878	0
6	-1.42587990	1.248421432	-1.45427744	0
7	0.21629826	-1.065401068	0.41971763	0
15	1.23760682	-0.130256093	-0.65177375	0
18	-0.18685210	1.503231576	-0.38807884	0
20	0.94521304	-1.120717011	0.17112519	0
21	-1.55872582	-0.691701288	-1.08236711	0
29	0.98323913	3.000000000	-0.77550929	0
35	0.85530452	-0.321137171	-0.89181618	0
37	-1.35269971	1.902256115	-0.76576309	0
41	1.46084848	-0.606027978	0.85386060	0
43	1.20157166	0.612186204	-0.02574979	0
46	-0.25170858	-1.720334685	-1.07018594	0
52	0.85509898	-0.981508232	-0.36119446	0
61	-0.58543586	0.957212918	-0.76006567	0
68	-1.01960904	-1.410034918	0.49304986	1
70	-0.84318658	0.651854899	0.05832648	0

Previsões

	0	1
4	1.00000000	0.00000000
6	1.00000000	0.00000000
7	0.00000000	1.00000000
15	1.00000000	0.00000000
18	1.00000000	0.00000000
20	0.97916667	0.02083333
21	0.95652174	0.04347826
29	1.00000000	0.00000000
35	1.00000000	0.00000000
37	1.00000000	0.00000000
41	0.88095238	0.11904762
43	1.00000000	0.00000000
46	0.97916667	0.02083333
52	0.97916667	0.02083333
61	1.00000000	0.00000000
68	0.00000000	1.00000000
70	1.00000000	0.00000000
72	0.00000000	1.00000000
76	1.00000000	0.00000000

Valores das probabilidades das classes.

Para retornar se é da classe 0 ou 1, veja o código abaixo:

```
> previsoes <- predict(classificador, newdata = base_teste[-4], type = 'class')
```

Vamos agora gerar a matriz de confusão

```
> matriz_confusao <- table(base_teste[, 4], previsoes)
> print(matriz_confusao)
```

	Previsões	
	0	1
0	420	7
1	11	62

Considerações:

- acertou 420 "0" e 62 "1";
- Errou 11 "1" e 7 "0";
- Se Σ os erros, a máquina errou somente 18 registros

```
> library(caret)
> confusionMatrix(matriz_confusao)
```

```
Confusion Matrix and Statistics

          previsoes
             0  1
  0        420  7
  1         11 62

               Accuracy : 0.964
                 95% CI : (0.9437, 0.9785)
    No Information Rate : 0.862
    P-Value [Acc > NIR] : 1.534e-14

                  Kappa : 0.8523

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 0.9745
            Specificity : 0.8986
         Pos Pred Value : 0.9836
         Neg Pred Value : 0.8493
             Prevalence : 0.8620
         Detection Rate : 0.8400
   Detection Prevalence : 0.8540
      Balanced Accuracy : 0.9365

       'Positive' Class : 0
```

Percentual de acerto da árvore de decisão.

Desempenho da máquina (classificador) em prever novos registros (base Teste).

Comparando o desempenho do Classificador baseado em Naive Bayes e Árvore de decisão

Resultados Credit Data

0.936 - Naive Bayes - Inconsistentes + faltan + Escal
0.936 - Naive Bayes - Inconsistentes + faltantes
0.932 - Naive Bayes - sem pré-processamento

0.964 - Árvore - Inconsistentes + faltantes + escaloram.
0.964 - Árvore - Inconsistentes + faltantes
0.964 - Árvore - sem pré-procemamentos

Próximo tópico, estudaremos Árvores de decisão em R - base Censo.

ÁRVORES DE DECISÃO EM R - BASE CENSO

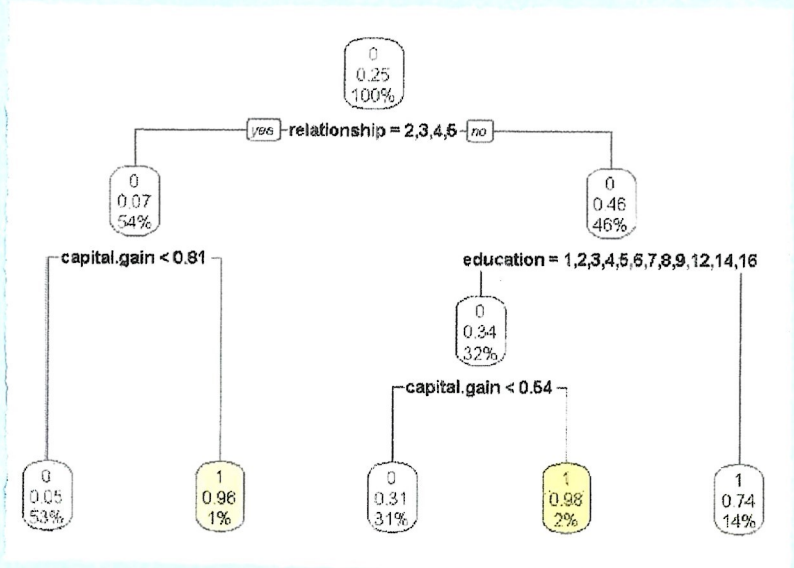
1º PASSO } PRÉ-PROCESSAMENTO DO TEMPLATE BASE SENSO, CONFORME DESCRITO ABAIXO:

- CRIAR NOVO SCRIPT E SALVÁ-LO COM O NOME "ARVORE_DECISAO_CENSUS.RDATA" ↑ A EXTENSÃO O R COLOCA AUTOMATICAMENTE.

NOTA DO INSTRUTOR } A BASE TESTE É PARA FAZER A AVALIAÇÃO DO ALGORITMO.

- LIBRARY(RPART)

```
> CLASSIFICADOR <- RPART(FORMULA = INCOME N., , DATA = BASE_TREINAMENTO) ↵
#VISUALIZANDO A ÁRVORE DE DECISÃO
>RPART.PLOT(CLASSIFICADOR)↵ #MOSTRA A ÁRVORE.
```




```
#FAZENDO AS PREVISÕES
> PREVISOES <- PREDICT(CLASSIFICADOR, NEWDATA = BASE_TESTE[-15], TYPE = 'CLASS')
```

NOTA DO INSTRUTOR } TYPE = 'CLASS' É PARA MOSTRAR A CLASSE META EM '0' E '1' E NÃO AS PROBABILIDADES DAS CLASSES EM CADA REGISTRO.

```
> PREVISOES #MOSTRA TODAS AS PREVISÕES PARA OS REGISTROS DA BASE_TESTE.
#GERANDO A MATRIZ DE CONFUSÃO
>MATRIZ_CONFUSAO <- TABLE(BASE_TESTE[,15], , PREVISOES) ↵
> PRINT(MATRIZ_CONFUSAO)
```

	PREVISÕES		
	0	1	...
0	3216	182	...
1	564	562	...

```
>LIBRARY(CARET) #VER O PERCENTUAL DE ACERTO
> CONFUSIONMATRIX(MATRIZ_CONFUSAO)↵
```

ACCURACY : 0.8351
95% CI : (0.824, 0.8458)

NO INFORMATION RATE : 0.8355

KAPPA : 0.5025

⋮

COMPARATIVO ENTRE O DESEMPENHO DOS ALGORITMOS

RESULTADOS CENSUS

0.8214 - NAIVE BAYES - CATEGÓRICOS + ESCALONAMENTO
0.8216 - NAIVE BAYES - CATEGÓRICOS
0.8214 - NAIVE BAYES - SEM PRÉ-PROCESSAMENTO

---#--- ≠ ---

0.8351 - ÁRVORE DE DECISÃO - CATEGÓRICOS + ESCALONAM.
0.8351 - ÁRVORE DE DECISÃO - CATEGÓRICOS
0.8351 - ÁRVORE DE DECISÃO - ESCALONAMENTO
0.8351 - ÁRVORE DE DECISÃO - SEM PRÉ-PROCESSAMENTO.

ÁRVORES DE DECISÃO EM R - PODA BASE CENSO

RODA-SE TODO O SCRIPT "ARVORE_DECISAO_-CENSUS.RDATA"

LOGO ABAIXO DO CLASSIFICADOR, VAMOS COLOCAR UM CÓDIGO ADICIONAL, CONFORME MOSTRADO ABAIXO:

```
> PODA <- CLASSIFICADOR$CPTABLE[WHICH.MIN(
        (CLASSIFICADOR$CPTABLE[, "XERROR"]),
        , "CP"] # CRIA PODA = 0.01 COMO MENOR ERRO
```

```
> PRINT(CLASSIFICADOR$CPTABLE) ↵
```

	CP	NSPLIT	REL ERROR	XERROR	XSTD
1	0.13232529	0	1.0000000	1.0002000	0.0108...
2	0.06424318	2	0.7353494	0.7353494	0.00970...
⋮	0.01000000	4	0.6342839	0.6342839	0.0091...

> PRUNE(CLASSIFICADOR, PODA)

```
> prune(classificador,poda)
n= 25638

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 25638 6382 0 (0.75107263 0.24892737)
   2) relationship=2,3,4,5 13852  955 0 (0.93105689 0.06894311)
     4) capital.gain< 0.8076171 13597  710 0 (0.94778260 0.05221740) *
     5) capital.gain>=0.8076171 255   10 1 (0.03921569 0.96078431) *
   3) relationship=1,6 11786 5427 0 (0.53953844 0.46046156)
     6) education=1,2,3,4,5,6,7,8,9,12,14,16 8261 2820 0 (0.65863697 0.34136303) *
     7) education=10,11,13,15 3525  918 1 (0.26042553 0.73957447) *
```

ÁRVORES DE DECISÃO - RANDOM FOREST

CLASSIFICADOR 3 → DIZ QUE O RISCO É MODER.
CLASSIFICADOR 1 → DIZ QUE O RISCO É ALTO
CLASSIFICADOR 2 → DIZ QUE O RISCO É BAIXO
CLASSIFICADOR N
... COMBINAR AS DECISÕES

+ DE 01 ÁRVORE, DE ZENAS DE ÁRVORES

ESSE ALGORITMO UTILIZA A DECISÃO DE VÁRIAS ÁRVORES.

RANDOM FOREST

- ENSEMBLE LEARNING (APRENDIZAGEM EM CONJUNTO)
  - "A IDÉIA É CONSULTAR DIVERSOS PROFISSIONAIS PARA TOMAR UMA DECISÃO"
  - VÁRIOS ALGORITMOS JUNTOS PARA CONSTRUIR UM ALGORITMO MAIS "FORTE";
  - USA A MÉDIA (REGRESSÃO) OU VOTOS DA MAIORIA (CLASSIFICAÇÃO) PARA DAR A RESPOSTA FINAL

Nota do instrutor: A QUANTIDADE DE ÁRVORES DE DECISÃO É UMA PARAMETRIZAÇÃO DO ALGORITMO, OU SEJA, VOCÊ PASSA PARA O ALGORITMO A QUANTIDADE DE ÁRVORES QUE VOCÊ QUER.

SUPOSIÇÃO:

RISCO = ALTO

RISCO = BAIXO

RISCO = BAIXO

→ PORTANTO, A RESPOSTA FINAL DO ALGORITMO VAI SER BAIXO.

2B
1A → BAIXO

RANDOM FOREST

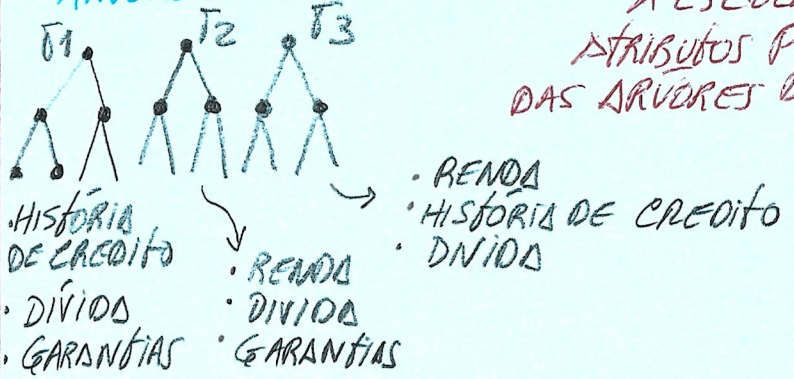
HISTÓRICO DE CRÉDITO	DÍVIDA	GARANTIAS	RENDA ANUAL	RISCO
RUIM	ALTA	NENHUMA	< 15.000	ALTO
DESCONHECIDO	ALTA	NENHUMA	>= 15.000 A <= 35.000	ALTO
Desconhecida	Baixa	Nenhuma	>=15000 a Mod. <=35000	
Desconhecida	Baixa	Nenhuma	>35.000	Baixo
⋮	⋮	⋮	⋮	⋮

Nota do instrutor: A floresta é dita "Random" porque escolhe de forma aleatória K atributos

... para comparação da métrica de pureza/impureza (impureza de gini/entropia)

Suposição:
K = 3 # NÚMERO DE ATRIBUTOS
ÁRVORES = 3

→ A BASE DE DADOS SÓ TEM 4 ATRIBUTOS, ENTRETANTO, K = 3. NESSE CASO, HAVERÁ A ESCOLHA ALEATÓRIA DOS ATRIBUTOS PARA A CONSTRUÇÃO DAS ÁRVORES DE DECISÃO.



ARTIGO: REAL-TIME HUMAN POSE RECOGNITION IN PARTS FROM SINGLE DEPTH IMAGES

RANDOM FOREST EM R - BASE CRÉDITO

- PASSO 1: ABRIR O ARQUIVO "TEMPLATE_CREDIT_DATA";
- PASSO 2: COPIAR TODO O CÓDIGO DESSE SCRIPT;
- PASSO 3: CRIAR NOVO SCRIPT;
- PASSO 4: SALVAR O SCRIPT COM O NOME DE "RANDOM_FOREST_CREDIT_DATA";
- PASSO 5: COLAR O CONTEÚDO, PASSO 2, NO SCRIPT NOVO;
- PASSO 6: EXECUTAR TODO O CÓDIGO;
- PASSO 7: INSTALAR O PACOTE "RANDOMFOREST";

```
#INSTALANDO O PACOTE "RANDOMFOREST"
> INSTALL.PACKAGES("RANDOMFOREST")
> #CARREGANDO PACOTE NECESSÁRIO
> LIBRARY(RANDOMFOREST)
> CRIANDO O CLASSIFICADOR
> CLASSIFICADOR <- RANDOMFOREST(X = BASE_
  - TREINAMENTO[-4], Y = BASE_TREINAMEN-
  TO$DEFAULT, NTREE = 10)
```

NOTA DO INSTRUTOR } ANTES DE EXECUTAR O CLASSIFICADOR, VOLTAR NO SCRIPT E INCLUIR A LINHA DE CÓDIGO, [redacted] DEPOIS DA LINHA DE COMANDO ABAIXO:

```
#ESCALONAMENTO
> BASE[, 1:13] <- SCALE(BASE[, 1:13])
```

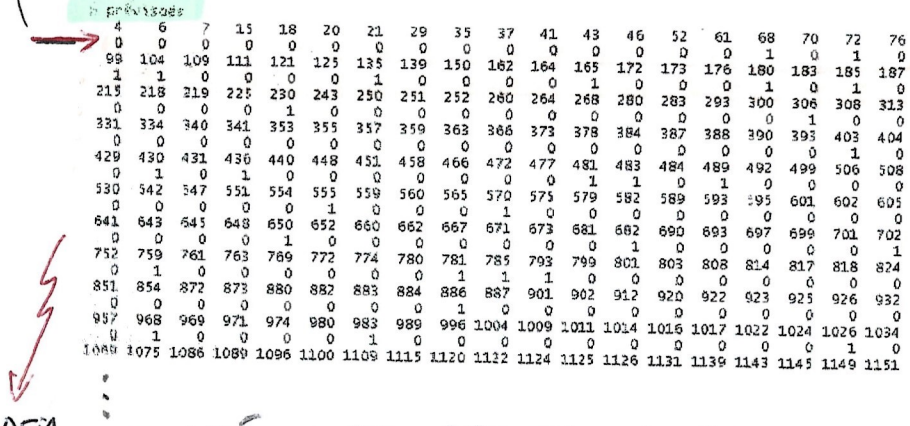
```
#ENCODE PARA A CLASSE
BASE$DEFAULT <- FACTOR(BASE$DEFAULT,
  > LEVELS = C(0, 1))
```

NOTA DO INSTRUTOR } O TREINAMENTO JÁ ESTÁ PRONTO A PARTIR DA CRIAÇÃO DO CLASSIFICADOR, COM UMA FLORESTA DE 10 ÁRVORES.

```
#CRIANDO AS PREVISÕES
> PREVISOES <- PREDICT(CLASSIFICADOR, NEWDATA =
  = BASE_TESTE[-4])
> PREVISOES
```

NA BASE TESTE ESSE VALOR DEU 0, OU SEJA, O ALGORITMO ACERTOU. É ISSO QUE VAMOS COMPARAR

ESSAS PREVISÕES FORAM FEITAS PELO CLASSIFICADOR



4	6	7	15	18	20	21	29	35	37	41	43	46	52	61	68	70	72	76
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
98	104	109	111	121	125	135	139	150	162	164	165	172	173	176	180	183	185	187
1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0
215	218	219	225	230	243	250	251	252	260	264	268	280	283	293	300	306	308	313
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
331	334	340	341	353	355	357	359	363	366	373	378	384	387	388	390	393	403	404
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
429	430	431	436	440	448	451	458	466	472	477	481	483	484	489	492	499	506	508
0	1	0	1	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0
530	542	547	551	554	555	559	560	565	570	575	579	582	589	593	595	601	602	605
0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
641	643	645	648	650	652	660	662	667	671	673	681	682	690	693	697	699	701	702
0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1
752	759	761	763	769	772	774	780	781	785	793	799	801	803	808	814	817	818	824
0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
851	854	872	873	880	882	883	884	886	887	901	902	912	920	922	923	925	926	932
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
957	968	969	971	974	980	983	989	996	1004	1009	1011	1014	1016	1017	1022	1024	1026	1034
0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
1069	1075	1086	1089	1096	1100	1109	1115	1120	1122	1124	1125	1126	1131	1139	1143	1145	1149	1151

⋮

A IDEIA AGORA É COMPARAR COM OS VALORES DA BASE_TESTE QUE JÁ SABEMOS OS VALORES DA CLASSE MESMA PARA VER QUAL FOI O DESEMPENHO DO ALGORITMO DE MACHINE LEARNING, OU SEJA, QUAL FOI SEU PERCENTUAL DE ACERTO E DE ERRO.

VAI ATÉ OS 500 REGISTROS DA BASE_TESTE

#CRIANDO A MATRIZ DE CONFUSÃO

> MATRIZ_CONFUSAO <- TABLE(BASE_TESTE[,4], PREVISOES)

> PRINT(MATRIZ_CONFUSAO)

	PREVISOES	
	0	1
0	423	4
1	7	66

#O ALGORITMO SÓ ERROU 11 REGISTROS DOS 500 REGISTROS TOTAIS DA BASE_TESTE.

-> ERROS

ACERTOS

#MOSTRANDO AS ESTATÍSTICAS DE DESEMPENHO

LIBRARY(CARET) #CARREGA O PACOTE NECESSÁRIO

CONFUSIONMATRIX(MATRIZ_CONFUSAO) ↵

CONFUSION MATRIX AND STATISTICS

PREVISOES

	0	1
0	423	4
1	7	66

ACCURACY: 0,978

95% CI: (0.963, 0.989)

...

NO INFORMATION RATE : 0.86

P-VALUE [ACC > NIR] : < 2e-16

KAPPA : 0.9102

MCNEMAR'S TEST P-VALUE: 0.5465

SENSITIVITY : 0.9837

SPECIFICITY : 0.9429

POS PRED VALUE : 0.9906

NEG PRED VALUE : 0.9041

PREVALENCE : 0.8600

DETECTION RATE : 0.8460

DETECTION PREVALENCE : 0.8540

BALANCED ACCURACY : 0.9633

'POSITIVE' CLASS : 0

Nota do instrutor: REFAZER O TESTE PARA n=15 ÁRVORES # A ACCURACY: 0.982, OU SEJA AUMENTOU.

DEPOIS DISTO, FAZ-SE O COMPARATIVO DE DESEMPENHO DE ALGORITMO PARA AS TECNICAS NAIVE BAYES, ARVORE, RANDOM FOREST 30, BASEADO NOS TIPOS DE PREPARAÇÃO DE DADOS EM:

- INCONSISTENTES + FALTANTES + ESCALONAMENTO;
- INCONSISTENTES + FALTANTES;
- SEM PROCESSAMENTO DE DADOS PRÉVIOS

RESULTADOS CREDIT DATA

- 0.936 - NAIVE BAYES - INCONSISTENTES + FALTANTES + + ESCALONAMENTO
- 0.936 - NAIVE BAYES - INCONSISTENTES + FALTANTES
- 0.932 - NAIVE BAYES - SEM PRÉ-PROCESSAMENTO
- 0.964 - ÁRVORE - INCONSISTENTES + FALTANTES + + ESCALONAMENTO
- 0.964 - ÁRVORE - INCONSISTENTES + FALTANTES
- 0.964 - ÁRVORE - SEM PROCESSAMENTO
- RANDOM FOREST 30 - INCONSISTENTES + FALTANTES + + ESCALONAMENTO (0.984)
- RANDOM FOREST 30 - INCONSISTENTES + FALTANTES (0.984)
- RANDOM FOREST 30 - SEM PRÉ-PROCESSAMENTO (0.988)

NOTA DO INSTRUTOR } PARA O ALGORITMO "RANDOM FOREST", NECESSARIAMENTE, PRECISA-SE CORRIGIR VALORES FALTANTES (NAs), POIS SENÃO DAR ERRO NA HORA DA EXECUÇÃO DOS CÓDIGOS.

IMPLEMENTAÇÃO DO RANDOM FOREST EM R - BASE CENSO

1º PASSO: PEGAR O ARQUIVO TEMPLATE_CENSUS.R;

2º PASSO: COPIAR TODO O CÓDIGO E COLAR NUM NOVO SCRIPT, SALVANDO ESTE COM O NOME "RANDOM_FOREST_CENSUS". APÓS ISSO, RODAR O CÓDIGO, CUJO A SAÍDA É MOSTRADA ABAIXO:

```
BASE            30162 OBS. OF 15 VARIABLES
BASE_TESTE      4524 OBS. OF 15 VARIABLES
BASE_TREINAMENTO 25638 OBS. OF 15 VARIABLES
VALUES
  DIVISAO    LOGI [1:30162] TRUE TRUE TRUE ...
```

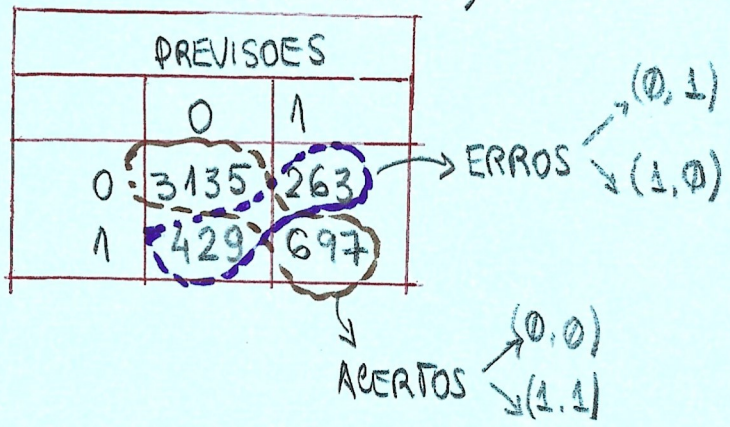
```
> LIBRARY(RANDOMFOREST)
> SET.SEED(1)
> CLASSIFICADOR <- RANDOMFOREST(X = BASE_TREINAMENTO[-15], Y = BASE_TREINAMENTO$ INCOME, NTREE = 10)
#CRIANDO A VARIÁVEL PREVISÕES
PREVISOES <- PREDICT(CLASSIFICADOR, NEWDATA = BASE_-TESTE[-15])
#MOSTRANDO PREVISOES
> PREVISOES ↵
```

...	↓ BASE ORIGINAL INCOME	↓ PREVISÕES (MACHINE LEARNING) INCOME	
4	0	0	# ACERTO
6	0	0	# ACERTO
7	0	0	# ACERTO
10	1	1	# ACERTO
11	1	0	# ERRO

CRIANDO A MATRIZ DE CONFUSÃO

```
> MATRIZ_CONFUSAO <- TABLE(BASE_TESTE[, 15],
                           , PREVISOES)
```

```
> PRINT(MATRIZ_CONFUSAO)
```



VENDO ESTATÍSTICAS DE DESEMPENHO DO MACHINE LEARNING

```
> LIBRARY(CARET) ↵
> CONFUSIONMATRIX(MATRIZ_CONFUSAO) ↵ # SAÍDA ABAIXO
```

CONFUSIONM MATRIX AND STATISTICS

...

PREVISOES

	0	1
0	3135	263
1	429	697

```
                 ACCURACY : 0.847
                 95% CI   : (0.8362, 0.8574)
    NO INFORMATION RATE   : 0.7878
    P-VALUE [ACC > NIR]   : < 2.2e-16
                  KAPPA   : 0.5697
 MCNEMAR'S TEST P-VALUE   : 3.55e-10

            SENSITIVITY : 0.8796
            SPECIFICITY : 0.7260
         POS PRED VALUE : 0.9226
         NEG PRED VALUE : 0.6190
             PREVALENCE : 0.7878
         DETECTION RATE : 0.6930
   DETECTION PREVALENCE : 0.7511
      BALANCED ACCURACY : 0.8028
       'POSITIVE' CLASS : 0
```

REFERÊNCIAS COMPLEMENTARES

- DATA MINING WITH DECISION TREES: THEORY AND APPLICATIONS (MACHINE PERCEPTION AND ARTIFICIAL INTELLIGENCE) DE ODED Z. MAIMON: LIVRO DE FÁCIL COMPREENSÃO E FOCADO EM ÁRVORES DE DECISÃO.
- PROGRAMS FOR MACHINE LEARNING (MORGAN KAUFMANN SERIES IN MACHINE LEARNING) DE J. ROSS QUINLAN: UM DOS LIVROS MAIS CLÁSSICOS SOBRE O ASSUNTO
- DECISION TREES AND RANDOM FOREST: A VISUAL INTRODUCTION FOR BEGGINERS: A SIMPLE GUIDE TO MACHINE LEARNING WITH DECISION TREES DE CHRIS SMITH E MARK KONING: UM DOS LIVROS MAIS FÁCEIS E DIDÁTICOS SOBRE O ASSUNTO.

CONTINUA NA PÁGINA 55 COM REGRAS DE ASSOCIAÇÃO