

Machine Learning com a Linguagem R

Da linguagem R aos algoritmos de aprendizado de máquina que sustentam decisões de negócio. Uma jornada estruturada em **16 módulos** que combina fundamentos de programação, estatística descritiva, redução de dimensionalidade, segmentação de consumidores e algoritmos de regressão, classificação e recomendação aplicados a casos reais de crédito, saúde, operações e varejo — encerrada com um projeto prático de concessão de crédito (Credit Scoring).

16	15+	12	01
Módulos	Packages de R	Algoritmos de Machine Learning	Projeto final prático

SUA JORNADA DE FERRAMENTAS E COMPETÊNCIAS

- 1

R
- 2

Estatística
- 3

EDA com Packages
- 4

Engenharia de Variáveis
- 5

PCA
- 6

Segmentação
- 7

Regressão
- 8

Classificação
- 9

Recomendação
- 10

Credit Scoring

12 algoritmos de ML
Regressão, classificação e recomendação

Do zero ao avançado
Base sólida em R e estatística

Projeto prático de portfólio
Credit Scoring aplicado

100% em R
RStudio e 15+ packages

Transforme dados em modelos preditivos e decisões inteligentes

Saber a sintaxe de um algoritmo de Machine Learning não é o mesmo que saber quando e por que usá-lo. É comum encontrar profissionais que conseguem rodar um Random Forest ou um KMeans em R, mas travam diante da pergunta mais importante: "esse é o algoritmo certo para este problema, e o resultado é confiável o suficiente para embasar uma decisão?". É exatamente essa lacuna que esta formação foi desenhada para fechar.

Ao longo de **16 módulos progressivos**, você vai desenvolver competências práticas em Machine Learning aplicado a negócios utilizando a linguagem R — da revisão de fundamentos estatísticos à engenharia de variáveis, passando por técnicas de redução de dimensionalidade (PCA) e segmentação de consumidores, até doze algoritmos de regressão, classificação e recomendação, cada um aplicado a um estudo de caso diferente — crédito, saúde, manutenção industrial, churn e varejo. A formação é encerrada com um projeto prático integrador de Credit Scoring, no qual você vai aplicar tudo o que aprendeu para decidir, com rigor estatístico, se um cliente deve ou não receber crédito.

A jornada segue uma lógica intencional: primeiro você consolida a base em R e estatística descritiva; depois aprende a preparar e transformar variáveis; em seguida, explora técnicas não supervisionadas (PCA e Segmentação) que organizam e simplificam os dados; e só então avança para os algoritmos supervisionados de regressão, classificação e recomendação — a etapa em que os modelos passam a prever e decidir, não apenas descrever.

Mais do que ensinar sintaxe de algoritmos, esta formação constrói raciocínio de Machine Learning: a capacidade de escolher o algoritmo certo para cada tipo de problema, validar sua performance com métricas adequadas e traduzir a saída de um modelo estatístico em uma decisão de negócio defensável — como aprovar ou negar um empréstimo.

PARA QUEM É ESTA FORMAÇÃO

- Analistas de dados e BI que querem evoluir para Ciência de Dados
- Estudantes de estatística, computação, economia e engenharia
- Profissionais de risco e crédito que querem entender modelos de Credit Scoring
- Quem já concluiu a formação em Estatística para Tomada de Decisão com R e quer avançar

PRÉ-REQUISITOS

Recomenda-se conhecimento básico de R (estruturas de dados) e estatística descritiva — conteúdo equivalente aos dois primeiros módulos desta própria formação. Não é necessária experiência prévia com Machine Learning.

METODOLOGIA

Cada algoritmo é apresentado com o fundamento teórico mínimo necessário e imediatamente aplicado em um estudo de caso prático, culminando no projeto prático de Credit Scoring.

O CAMINHO DA FORMAÇÃO

01 Introdução à Linguagem R	02 Estatística Básica
03 Explorando os Dados com Pacotes — EDA	04 Engenharia de Variáveis com a Linguagem R
05 PCA — Redução de Dimensionalidade	06 Comportamento de Consumidores — Segmentação
07 Redução de Custo — Regressão Linear	08 Estimativa de Score de Crédito — Random Forest Regressão
09 Previsão de Internação Hospitalar — XGBoosting Regressão	10 Manutenção de Equipamentos Inteligente — SVR Regressão
11 Análise de Churn — Árvore de Decisão	12 Classificação de Doenças — KNN
13 Doenças Cardiovasculares — Random Forest	14 Predição de Células Cancerígenas — SVM
15 Sistema de Recomendação — Cesta Inteligente	16 Projeto Final: Credit Scoring ★

CONTEXTO

Toda jornada em Machine Learning começa pela base: entender como o R organiza e manipula dados. Este módulo de abertura ensina a instalar e configurar o ambiente R e RStudio, e apresenta as estruturas de dados — vetores, matrizes, listas e data frames — que vão sustentar a construção de todo algoritmo estudado nesta formação.

PRINCIPAIS TÓPICOS

- Instalação do R e RStudio
- Vetores numéricos e caracteres
- Matrizes
- Listas
- Data Frame
- Estrutura de repetição
- Funções
- Família Apply

NA PRÁTICA

Antes de treinar qualquer modelo, você precisa saber manipular a matriz de variáveis que alimenta o algoritmo. Você vai praticar a criação, indexação e transformação de estruturas de dados, além de automatizar tarefas repetitivas com a família apply — uma habilidade que você vai reutilizar em quase todos os módulos seguintes.

DESAFIO DO MÓDULO

Recrie um pequeno conjunto de dados de clientes (idade, renda, score de crédito) como data frame, e escreva uma função que classifique cada cliente em faixas de renda usando a família apply.

ERRO COMUM A EVITAR

Confundir vetor com lista é o deslize mais frequente de quem começa em R — vetores exigem que todos os elementos sejam do mesmo tipo; listas, não.

DICA DO AUTOR

Domine bem as estruturas de dados agora — todos os algoritmos de Machine Learning desta formação partem de uma matriz ou data frame bem organizado.

O QUE VOCÊ VAI CONQUISTAR

Você vai dominar a sintaxe do R e as estruturas de dados que sustentam toda a formação: vetores, matrizes, listas, data frames e funções.

- Sintaxe fluente em R
- Scripts reprodutíveis
- Base para todos os algoritmos de ML da formação

Início da jornada → **01 R** → Estatística Básica

02 Estatística Básica

Os fundamentos estatísticos que antecedem qualquer algoritmo de Machine Learning.

CONTEXTO

Antes de qualquer algoritmo de Machine Learning, é preciso entender o comportamento estatístico das variáveis que você vai alimentar nos modelos. Este módulo revisa os fundamentos da estatística descritiva — variabilidade, assimetria e curtose — tanto para variáveis categóricas quanto numéricas, base conceitual que vai orientar decisões de pré-processamento nos módulos seguintes.

PRINCIPAIS TÓPICOS

- Estatística e a Ciência de Dados
- Variabilidade de variáveis numéricas
- Variabilidade de variáveis categóricas
- Assimetria e curtose
- Análise descritiva de variáveis categóricas
- Análise descritiva de variáveis numéricas

NA PRÁTICA

Um algoritmo como o KNN ou o KMeans é extremamente sensível à escala e à distribuição das variáveis. Reconhecer a variabilidade e a assimetria dos seus dados antes de modelar evita erros silenciosos que só apareceriam depois, com o modelo já treinado — e já errado.

DESAFIO DO MÓDULO

Calcule a assimetria de uma variável de renda ou faturamento e avalie se essa distribuição, sem tratamento, comprometeria um algoritmo sensível a outliers como o KMeans.

ERRO COMUM A EVITAR

Alimentar um modelo com variáveis em escalas muito diferentes (idade em anos e renda em milhares) sem normalizar é uma causa comum de modelos distorcidos, especialmente em algoritmos baseados em distância.

DICA DO AUTOR

Sempre visualize a distribuição de cada variável antes de decidir se ela precisa de transformação. Um histograma evita muitos retrabalhos de modelagem.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de descrever e interpretar a variabilidade de variáveis categóricas e numéricas, antecipando problemas que afetariam algoritmos de Machine Learning.

- Leitura crítica de medidas estatísticas
- Diagnóstico de assimetria e curtose
- Base para pré-processamento em ML

Introdução à Linguagem R → **02 Estatística Básica** → EDA com Pacotes

03 Explorando os Dados com Pacotes — EDA

Análise exploratória de dados (EDA) com pacotes especializados do R.

CONTEXTO

Assim como na análise estatística tradicional, a análise exploratória (EDA) é a etapa que antecede qualquer modelo de Machine Learning. Este módulo apresenta três pacotes de R que aceleram drasticamente essa exploração — SummaryTools, SmartEDA e DataExplorer — permitindo identificar rapidamente outliers, dados faltantes e relações entre variáveis antes de treinar qualquer algoritmo.

PRINCIPAIS TÓPICOS

- Introdução ao package SummaryTools
- Introdução ao package SmartEDA
- Introdução ao package DataExplorer

NA PRÁTICA

Antes de treinar um modelo de PCA, Segmentação ou classificação, um cientista de dados precisa responder: existem outliers que vão distorcer a distância entre pontos? há variáveis com dados faltantes concentrados? Você vai responder a essas perguntas em minutos, usando relatórios automatizados de EDA.

DESAFIO DO MÓDULO

Rode um relatório de EDA com DataExplorer em uma base de crédito e identifique, antes de qualquer modelagem, quais variáveis têm maior proporção de valores ausentes.

ERRO COMUM A EVITAR

Pular a etapa de EDA e ir direto para o treinamento de um algoritmo é a causa mais comum de modelos que "funcionam nos testes mas quebram na produção".

DICA DO AUTOR

Trate a EDA como uma investigação, não como uma formalidade — é nela que você encontra os problemas antes que o modelo os encontre por você (e falhe).

O QUE VOCÊ VAI CONQUISTAR

Você vai aprender a acelerar a análise exploratória de dados utilizando pacotes como SummaryTools, SmartEDA e DataExplorer, ganhando velocidade sem perder rigor antes da modelagem.

- Relatórios exploratórios automatizados
- Detecção rápida de outliers e missing
- Preparação acelerada para modelagem

Estatística Básica → **03 EDA com Pacotes** → Engenharia de Variáveis

FERRAMENTAS DO MÓDULO

[summarytools](#) [SmartEDA](#) [DataExplorer](#)

04 Engenharia de Variáveis com a Linguagem R

Transforme variáveis para que os algoritmos de Machine Learning funcionem corretamente.

CONTEXTO

Muitos algoritmos de Machine Learning — de PCA a KMeans, de SVM a Regressão Logística — são sensíveis à escala e à distribuição das variáveis. Este módulo ensina as principais técnicas de transformação: logaritmo, raiz quadrada, Z-score e normalização, preparando as variáveis para os algoritmos das próximas etapas.

PRINCIPAIS TÓPICOS

- Introdução à Engenharia de Variáveis
- Transformação de variáveis
- Testando a normalidade
- Logaritmo
- Logaritmo + 1
- Transformação pela SQRT
- Transformação pelo Z-Score
- Transformação pela normalização
- bestNormalize

NA PRÁTICA

Um algoritmo de PCA, por exemplo, é extremamente sensível a variáveis em escalas diferentes — sem padronização (Z-score), uma variável de faturamento em milhões vai dominar completamente uma variável de idade em anos. Você vai aprender a decidir qual transformação aplicar a cada situação, inclusive de forma automatizada com o bestNormalize.

DESAFIO DO MÓDULO

Aplique Z-score e normalização Min-Max na mesma variável e compare visualmente como cada transformação altera sua distribuição — decida qual seria mais adequada para um algoritmo de segmentação.

ERRO COMUM A EVITAR

Aplicar PCA ou KMeans em variáveis não padronizadas é um dos erros mais comuns e mais custosos em Machine Learning — variáveis com escalas maiores dominam artificialmente o resultado.

DICA DO AUTOR

bestNormalize existe para tirar de você a decisão manual de qual transformação usar — mas entenda a lógica de cada uma antes de automatizar.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de aplicar as principais técnicas de transformação de variáveis para preparar dados adequados aos algoritmos das próximas etapas.

- Testes formais de normalidade
- Padronização e normalização de variáveis
- Pré-requisito direto para PCA e Segmentação

EDA com Pacotes → **04 Engenharia de Variáveis** → PCA — Redução de Dimensionalidade

05 PCA — Redução de Dimensionalidade

Reduza a dimensionalidade dos dados sem perder informação relevante.

CONTEXTO

No dia a dia, bases de dados costumam ter dezenas ou centenas de variáveis correlacionadas entre si. O PCA (Análise de Componentes Principais) resolve esse problema, reduzindo a dimensionalidade dos dados sem perder informação relevante — uma técnica fundamental tanto para visualização quanto para alimentar outros algoritmos com menos ruído e menos multicolinearidade.

PRINCIPAIS TÓPICOS

- Redução de dimensionalidade
- KMO — correlação conjunta
- Gráfico Scree Plot
- Rotação Varimax
- Scores

NA PRÁTICA

Você vai aplicar o teste KMO para verificar se suas variáveis têm correlação suficiente para justificar um PCA, interpretar o Scree Plot para decidir quantos componentes reter, e usar a Rotação Varimax para tornar esses componentes mais interpretáveis — gerando scores que podem alimentar modelos de segmentação ou classificação.

DESAFIO DO MÓDULO

Aplique o teste KMO em uma base de indicadores financeiros de clientes e decida, com base no Scree Plot, quantos componentes principais reter para representar a maior parte da variância dos dados.

ERRO COMUM A EVITAR

Aplicar PCA em variáveis com KMO baixo (abaixo de 0,6) é um erro comum — nesses casos, a correlação entre as variáveis não é suficiente para justificar a redução de dimensionalidade.

DICA DO AUTOR

Os componentes principais não têm significado de negócio automático — é seu trabalho, depois da Rotação Varimax, interpretar o que cada componente representa na prática.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de reduzir a dimensionalidade de uma base complexa, preservando a informação relevante e preparando os dados para segmentação e classificação.

- Redução de dimensionalidade com PCA
- Interpretação de Scree Plot e KMO
- Geração de scores para outros modelos

Engenharia de Variáveis → **05 PCA** → Segmentação de Consumidores

FERRAMENTAS DO MÓDULO

psych

FactoMineR

prcomp

06 Comportamento de Consumidores — Segmentação

Agrupe consumidores por comportamento de compra e descubra perfis de cliente.

CONTEXTO

Depois de reduzir a dimensionalidade com PCA, o próximo passo natural é agrupar consumidores com comportamentos semelhantes entre si. Este módulo ensina as duas grandes famílias de algoritmos de clusterização — o método hierárquico e o KMeans/KModes — aplicadas à segmentação de comportamento de consumo, um dos casos de uso mais valiosos de Machine Learning em marketing e negócios.

PRINCIPAIS TÓPICOS

- Análise de Cluster
- Método hierárquico
- Dendrograma e ponto de corte
- KMeans & KModes
- Método do cotovelo (elbow method)
- Perfil de comportamento de consumo

NA PRÁTICA

Você vai segmentar uma base de consumidores em grupos com perfis de comportamento semelhantes — por exemplo, por frequência de compra, ticket médio e categorias preferidas — usando tanto o método hierárquico, para explorar visualmente a estrutura dos dados, quanto o KMeans, para segmentações mais robustas e escaláveis.

DESAFIO DO MÓDULO

Segmente uma base de consumidores em 3 a 5 clusters usando KMeans e escreva, para cada cluster, uma frase que resuma o comportamento de compra predominante daquele grupo.

ERRO COMUM A EVITAR

Escolher o número de clusters (K) de forma arbitrária, sem usar critérios como o método do cotovelo, é um erro frequente que gera segmentações pouco interpretáveis ou instáveis.

DICA DO AUTOR

Um cluster de comportamento de consumo só tem valor se puder orientar uma ação de marketing concreta — uma campanha, uma oferta, um canal. Se você não consegue imaginar essa ação, o cluster ainda não está bem definido.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de segmentar consumidores em grupos homogêneos por comportamento de compra e de interpretar o perfil de cada cluster formado.

- Segmentação de comportamento de consumo
- Clusterização hierárquica e KMeans
- Interpretação de perfis de cluster

PCA — Redução de Dimensionalidade → **06 Segmentação de Consumidores** → Redução de Custo

FERRAMENTAS DO MÓDULO

stats (kmeans) klaR hclust

07 Redução de Custo — Regressão Linear

Preveja custos operacionais e identifique onde reduzi-los com Regressão Linear.

CONTEXTO

Este módulo marca a entrada da formação nos algoritmos preditivos supervisionados voltados a variáveis numéricas, aplicados a um problema clássico de operações: prever e reduzir custos. Você vai construir um modelo de Regressão Linear para estimar o custo de um processo a partir de variáveis explicativas, identificando quais delas mais pesam no resultado final.

PRINCIPAIS TÓPICOS

- Revisão de correlação
- Regressão Linear Múltipla
- Regressão Linear Simples
- Interpretação de coeficientes de custo

NA PRÁTICA

Você vai prever uma variável de custo operacional — por exemplo, custo logístico ou custo de atendimento — a partir de variáveis como volume, distância ou tempo de processo, e vai interpretar os coeficientes do modelo para identificar oportunidades concretas de redução de custo.

DESAFIO DO MÓDULO

Construa um modelo de Regressão Linear para prever um custo operacional e aponte, com base nos coeficientes, duas variáveis que representam as maiores oportunidades de redução.

ERRO COMUM A EVITAR

Interpretar um R^2 alto como garantia de um bom modelo, sem validar em dados de teste, é um erro comum — um modelo pode se ajustar bem aos dados de treino e generalizar mal.

DICA DO AUTOR

Regressão Linear é o modelo mais fácil de explicar a uma diretoria de operações — comece sempre por ela antes de justificar um modelo mais complexo.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de construir, avaliar e interpretar modelos de Regressão Linear aplicados a problemas de redução de custo.

- Modelagem com Regressão Linear
- Interpretação de coeficientes de custo
- Benchmark para modelos mais complexos

Segmentação de Consumidores → **07 Redução de Custo** → Score de Crédito (Random Forest)

CONTEXTO

Nem toda relação entre variáveis é linear. Este módulo apresenta o Random Forest, um algoritmo de ensemble baseado em árvores de decisão, aplicado à estimativa de um score de crédito contínuo — capaz de capturar relações não lineares e interações complexas entre variáveis financeiras e comportamentais do cliente.

PRINCIPAIS TÓPICOS

- Random Forest — Regressão
- Importância de variáveis
- Estimativa de score contínuo
- Comparação com a Regressão Linear

NA PRÁTICA

Você vai treinar um Random Forest para estimar o score de crédito de um cliente a partir de variáveis como renda, histórico de pagamento e tempo de relacionamento, comparando o erro (RMSE) desse modelo com o de uma Regressão Linear simples.

DESAFIO DO MÓDULO

Treine um Random Forest para estimar o score de crédito de uma base de clientes e compare o RMSE com o de uma Regressão Linear — decida qual modelo você recomendaria e por quê.

ERRO COMUM A EVITAR

Escolher um algoritmo mais complexo, como o Random Forest, só porque é mais "moderno", sem comparar sua performance com modelos mais simples, é um erro comum — às vezes uma Regressão Linear bem construída supera modelos sofisticados.

DICA DO AUTOR

A importância de variáveis do Random Forest é uma ótima forma de validar, com dados, se as regras de crédito atuais da empresa fazem sentido estatístico.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de aplicar o Random Forest para estimar variáveis numéricas de risco, como o score de crédito de um cliente.

- Modelagem com Random Forest
- Leitura de importância de variáveis
- Estimativa de score contínuo

Redução de Custo → **08 Score de Crédito (Random Forest)** → Internação Hospitalar

FERRAMENTAS DO MÓDULO

randomForest

CONTEXTO

Este módulo aplica o XGBoosting, um dos algoritmos de ensemble mais usados no mercado, a um problema de saúde: prever o tempo de internação hospitalar de um paciente a partir de variáveis clínicas e administrativas — uma informação crítica para o planejamento de leitos e recursos de um hospital.

PRINCIPAIS TÓPICOS

- XGBoosting — Regressão
- Hiperparâmetros principais
- Importância de variáveis clínicas
- Comparação com outros algoritmos de regressão

NA PRÁTICA

Você vai treinar um modelo de XGBoosting para prever o número de dias de internação de um paciente a partir de variáveis como idade, comorbidades e tipo de procedimento, e vai comparar sua performance com a de um Random Forest.

DESAFIO DO MÓDULO

Treine um XGBoosting para prever o tempo de internação de uma base de pacientes e compare o RMSE com o do Random Forest do módulo anterior.

ERRO COMUM A EVITAR

Usar o XGBoosting com os hiperparâmetros padrão, sem ajustar taxa de aprendizado e profundidade das árvores, é um erro comum que pode limitar bastante a performance do modelo.

DICA DO AUTOR

Em problemas de saúde, um erro de previsão tem custo direto em capacidade hospitalar — sempre valide o modelo com dados de teste antes de usá-lo em um planejamento real de leitos.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de aplicar o XGBoosting a problemas de previsão numérica em contextos de saúde e operações hospitalares.

- Modelagem com XGBoosting
- Ajuste de hiperparâmetros
- Aplicação em planejamento hospitalar

Score de Crédito (Random Forest) → **09 Internação Hospitalar (XGBoosting)** → Manutenção de Equipamentos

FERRAMENTAS DO MÓDULO

xgboost

CONTEXTO

Este módulo fecha o bloco de algoritmos de regressão com o SVR (Support Vector Regression), aplicado a um problema clássico de manutenção preditiva: estimar o tempo restante até a falha de um equipamento a partir de dados de sensores, uso e histórico de manutenção.

PRINCIPAIS TÓPICOS

- SVR — Support Vector Regression
- Kernels e hiperparâmetros
- Manutenção preditiva
- Comparação final entre os algoritmos de regressão

NA PRÁTICA

Você vai treinar um modelo de SVR para estimar o tempo restante até a falha de um equipamento a partir de dados de sensores, fechando a comparação entre os algoritmos de regressão estudados na formação — Regressão Linear, Random Forest, XGBoosting e SVR.

DESAFIO DO MÓDULO

Treine um SVR para estimar o tempo até a falha de um equipamento e monte uma tabela comparativa do RMSE com os demais algoritmos de regressão estudados.

ERRO COMUM A EVITAR

Usar o SVR com os hiperparâmetros padrão, sem testar diferentes kernels, é um erro comum que pode limitar bastante a performance do modelo em dados de sensores.

DICA DO AUTOR

SVR tende a funcionar melhor em bases com variáveis bem padronizadas — revise a engenharia de variáveis do módulo 04 antes de treinar o modelo.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de aplicar o SVR a problemas de manutenção preditiva e de comparar criticamente os quatro principais algoritmos de regressão da formação.

- Modelagem com SVR
- Manutenção preditiva de equipamentos
- Comparação entre quatro algoritmos de regressão

Internação Hospitalar (XGBoosting) → **10 Manutenção de Equipamentos (SVR)** → Análise de Churn

FERRAMENTAS DO MÓDULO

e1071

CONTEXTO

Se o bloco anterior tratou de prever números, este foca em prever categorias — a tarefa mais comum em Machine Learning aplicado a negócios. Você vai abrir o bloco de classificação com a Árvore de Decisão, aplicada a um problema clássico de negócio: prever se um cliente vai ou não cancelar (churn) o serviço.

PRINCIPAIS TÓPICOS

- Árvore de Decisão
- Leitura da árvore e regras de decisão
- Poda da árvore (pruning)
- Aplicação em análise de churn

NA PRÁTICA

Você vai treinar uma Árvore de Decisão para prever o cancelamento de clientes a partir de variáveis como tempo de contrato, uso do serviço e histórico de atendimento, e vai ler a árvore resultante como um conjunto de regras de negócio explicáveis para a equipe de retenção.

DESAFIO DO MÓDULO

Treine uma Árvore de Decisão em uma base de churn e escreva, em linguagem de negócio, as duas regras de decisão mais relevantes encontradas pela árvore.

ERRO COMUM A EVITAR

Avaliar um classificador apenas pela acurácia, especialmente em bases desbalanceadas (poucos casos de cancelamento), é um erro grave — sensibilidade e especificidade contam uma história muito mais honesta.

DICA DO AUTOR

Uma árvore de decisão só é útil para a equipe de retenção se puder ser lida como um fluxograma de regras — evite árvores tão profundas que percam essa clareza.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de aplicar a Árvore de Decisão para prever cancelamento de clientes e de traduzir o modelo em regras de negócio explicáveis.

- Classificação com Árvore de Decisão
- Leitura de regras de decisão
- Avaliação por matriz de confusão

Manutenção de Equipamentos (SVR) → **11 Análise de Churn** → Classificação de Doenças (KNN)

FERRAMENTAS DO MÓDULO

rpart

CONTEXTO

Este módulo apresenta o KNN (K-Nearest Neighbors) aplicado a um problema de saúde: classificar o diagnóstico de um paciente com base em pacientes semelhantes já diagnosticados. Por depender diretamente da distância entre pontos, o KNN é um dos algoritmos mais sensíveis à escala das variáveis clínicas.

PRINCIPAIS TÓPICOS

- KNN — K-Nearest Neighbors
- Escolha do valor de K
- Sensibilidade à escala das variáveis clínicas
- Comparação com a Árvore de Decisão

NA PRÁTICA

Você vai treinar um classificador KNN para prever o diagnóstico de um paciente a partir de variáveis clínicas, testar diferentes valores de K, e comparar sua performance com a Árvore de Decisão treinada no módulo anterior.

DESAFIO DO MÓDULO

Treine um classificador KNN em uma base de diagnóstico, testando pelo menos três valores diferentes de K, e compare a acurácia de cada configuração.

ERRO COMUM A EVITAR

Treinar um KNN sem padronizar as variáveis clínicas é um erro comum e especialmente grave — como o algoritmo se baseia em distância, uma variável em escala maior domina artificialmente o resultado.

DICA DO AUTOR

Não existe um K "certo" universal — teste alguns valores e escolha aquele que equilibra performance nos dados de teste e capacidade de generalização.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de aplicar o KNN a problemas de classificação de diagnóstico e de reconhecer quando a escala das variáveis compromete o resultado do modelo.

- Classificação com KNN
- Escolha do valor de K
- Reforço da importância da padronização

Análise de Churn → **12 Classificação de Doenças (KNN)** → Doenças Cardiovasculares

FERRAMENTAS DO MÓDULO

class (KNN)

CONTEXTO

Random Forest já apareceu no bloco de regressão — e este módulo mostra como o mesmo algoritmo se adapta a problemas de classificação, aplicado à predição de risco de doença cardiovascular a partir de variáveis clínicas como pressão arterial, colesterol e histórico familiar.

PRINCIPAIS TÓPICOS

- Random Forest — Classificação
- Importância de variáveis clínicas
- Avaliação de risco cardiovascular
- Comparação com KNN e Árvore de Decisão

NA PRÁTICA

Você vai treinar um Random Forest para classificar o risco cardiovascular de um paciente a partir de variáveis clínicas, e vai comparar sua performance — acurácia, sensibilidade e especificidade — com a do KNN e da Árvore de Decisão já treinados.

DESAFIO DO MÓDULO

Treine um Random Forest de classificação em uma base de risco cardiovascular e monte uma tabela comparativa de sensibilidade e especificidade com os algoritmos dos módulos anteriores.

ERRO COMUM A EVITAR

Escolher o algoritmo de classificação apenas pelo maior valor de acurácia geral, ignorando sensibilidade e especificidade, é um erro recorrente — especialmente em saúde, onde deixar de identificar um paciente de risco tem um custo muito maior que um falso alarme.

DICA DO AUTOR

Em problemas de saúde, priorize sempre a sensibilidade do modelo — o custo de não identificar um paciente de risco costuma ser muito maior do que o de um falso positivo.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de aplicar o Random Forest a problemas de classificação de risco em saúde e de comparar sua performance com outros classificadores.

- Classificação com Random Forest
- Avaliação de risco cardiovascular
- Comparação entre classificadores

Classificação de Doenças (KNN) → **13 Doenças Cardiovasculares** → Células Cancerígenas (SVM)

FERRAMENTAS DO MÓDULO

randomForest

14 Predição de Células Cancerígenas — SVM

Classifique células como benignas ou malignas com Support Vector Machines.

CONTEXTO

Este módulo fecha o bloco de classificação com o SVM (Support Vector Machine), aplicado a um dos problemas mais estudados em Machine Learning aplicado à saúde: classificar células como benignas ou malignas a partir de características extraídas de exames — uma tarefa em que a margem de separação entre classes é crítica.

PRINCIPAIS TÓPICOS

- SVM — Classificação
- Classificação de células cancerígenas
- Kernels e margem de separação
- Comparação final entre os classificadores

NA PRÁTICA

Você vai treinar um SVM para classificar células como benignas ou malignas a partir de variáveis extraídas de exames, fechando a comparação entre os algoritmos de classificação estudados na formação — Árvore de Decisão, KNN, Random Forest e SVM.

DESAFIO DO MÓDULO

Treine um SVM em uma base de classificação de células e monte uma tabela comparativa final de acurácia, sensibilidade e especificidade dos algoritmos de classificação estudados neste bloco.

ERRO COMUM A EVITAR

Escolher o kernel do SVM sem testar alternativas é um erro comum — o kernel linear pode não capturar a separação real entre células benignas e malignas em bases mais complexas.

DICA DO AUTOR

Em diagnóstico oncológico, um falso negativo (classificar uma célula maligna como benigna) é o erro mais caro — sempre priorize a sensibilidade do modelo nessa decisão.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de aplicar o SVM a problemas de classificação médica de alta sensibilidade e de comparar criticamente os principais algoritmos de classificação da formação.

- Classificação com SVM
- Leitura de margem de separação
- Comparação final entre classificadores

Doenças Cardiovasculares → **14 Células Cancerígenas (SVM)** → Sistema de Recomendação

FERRAMENTAS DO MÓDULO

e1071

15 Sistema de Recomendação — Cesta Inteligente

Descubra regras de associação entre produtos e apoie estratégias de cross-sell.

CONTEXTO

Este módulo introduz uma família diferente de algoritmos de Machine Learning: sistemas de recomendação, usados por praticamente todo grande varejista e plataforma digital para sugerir produtos com base em padrões de compra. Você vai aprender o Algoritmo Apriori, técnica clássica por trás da "cesta inteligente" de recomendação, que identifica regras de associação entre produtos.

PRINCIPAIS TÓPICOS

- Algoritmo Apriori
- Suporte, confiança e lift

NA PRÁTICA

Você vai aplicar o Apriori a uma base de transações de vendas para descobrir regras do tipo "clientes que compram o produto A também compram o produto B com frequência X" — a lógica por trás de recomendações de cross-sell e estratégias de layout de loja.

DESAFIO DO MÓDULO

Aplique o Apriori a uma base de transações e identifique a regra de associação com maior lift — explique, em uma frase, o que essa regra significaria para uma estratégia comercial.

ERRO COMUM A EVITAR

Interpretar uma regra com alto suporte, mas baixo lift, como uma descoberta relevante é um erro comum — o lift, não o suporte isolado, indica se a associação é realmente mais forte que o esperado pelo acaso.

DICA DO AUTOR

Regras de associação geram muitas combinações — nem todas são úteis. Filtre por lift e valide se a regra faz sentido de negócio antes de agir sobre ela.

O QUE VOCÊ VAI CONQUISTAR

Você será capaz de aplicar o Algoritmo Apriori para identificar regras de associação entre produtos e apoiar estratégias de recomendação e cross-sell.

- Regras de associação (Apriori)
- Estratégias de cross-sell e up-sell
- Leitura de suporte, confiança e lift

Células Cancerígenas (SVM) → **15 Cesta Inteligente** → Projeto: Credit Scoring

FERRAMENTAS DO MÓDULO

arules arulesViz

POR QUE ESTE MÓDULO É MAIS ENXUTO

O Apriori é um algoritmo conceitualmente autocontido — diferente dos módulos de regressão e classificação, aqui o foco é dominar profundamente uma única técnica e sua lógica de regras de associação (suporte, confiança e lift).

CONTEXTO

Este é o projeto final da formação — o momento em que você integra estatística descritiva, engenharia de variáveis e algoritmos de classificação para resolver um dos problemas mais valiosos de Machine Learning aplicado a negócios: decidir se um cliente deve ou não receber crédito. O Credit Scoring é usado por bancos, fintechs e financeiras para equilibrar crescimento de carteira com controle de risco. É aqui que você conhece a **Regressão Logística**, o modelo clássico de mercado para Credit Scoring, e a compara com os demais classificadores aprendidos na formação.

Diferente dos módulos anteriores, aqui não há uma única técnica isolada: o desafio é integrador. Você vai reaproveitar a engenharia de variáveis do módulo 04 e os algoritmos de classificação dos módulos 11 a 14 (Árvore de Decisão, KNN, Random Forest e SVM) para construir, do zero, um modelo de concessão de crédito defensável perante uma diretoria de risco.

PRINCIPAIS TÓPICOS

- Definição do problema de concessão de crédito
- Regressão Logística — o modelo clássico de Credit Scoring
- Preparação e engenharia de variáveis da base
- Comparação com Árvore, KNN, Random Forest e SVM
- Avaliação por matriz de confusão
- Definição do ponto de corte (cutoff) do score
- Simulação de impacto financeiro por política de crédito
- Apresentação executiva do modelo

NA PRÁTICA

Usando uma base prática de solicitantes de crédito, você vai construir um modelo de Regressão Logística para estimar a probabilidade de inadimplência de cada cliente, comparar seu resultado com os classificadores já treinados na formação, definir um ponto de corte (cutoff) que equilibre a aprovação de bons pagadores e a rejeição de maus pagadores, e simular o impacto financeiro de diferentes políticas de concessão — encerrando com uma apresentação executiva do modelo.

DESAFIO DO MÓDULO

Compare a Regressão Logística com pelo menos dois outros algoritmos de classificação aprendidos na formação (por exemplo, Random Forest e SVM) no mesmo problema de Credit Scoring e defina, com base na matriz de confusão, qual ponto de corte você recomendaria para o negócio.

ERRO COMUM A EVITAR

Definir o ponto de corte do score apenas pela acurácia geral do modelo, sem considerar o custo de negócio de aprovar um mau pagador versus rejeitar um bom pagador, é um erro estratégico grave em Credit Scoring.

DICA DO AUTOR

Um modelo de crédito não é bom ou ruim isoladamente — ele é calibrado para uma política de risco específica. Sempre traduza a saída do modelo em impacto financeiro antes de recomendá-la à diretoria.

O QUE VOCÊ VAI CONQUISTAR

Você vai entregar um projeto prático completo de Credit Scoring, do dado bruto à recomendação executiva — aplicando de forma integrada engenharia de variáveis, Regressão Logística e os demais algoritmos de classificação da formação.

- Projeto de portfólio pronto para currículo
- Modelo de Credit Scoring completo
- Simulação de políticas de crédito

Cesta Inteligente → **16 Credit Scoring** → Fim da formação

POR QUE ESTE É O ENCERRAMENTO IDEAL

Credit Scoring integra praticamente todos os blocos da formação — estatística descritiva, engenharia de variáveis e classificação — em um problema de negócio prático e mensurável financeiramente. O encerramento perfeito para comprovar domínio de ponta a ponta em Machine Learning.

Ao concluir a formação, você terá saído do zero à construção de modelos de Machine Learning

Ao concluir a formação em **Machine Learning com a Linguagem R**, você terá percorrido um caminho completo — da estrutura de dados mais básica do R ao seu primeiro projeto prático de concessão de crédito orientado por algoritmos. Mais do que dominar doze algoritmos isoladamente, você terá desenvolvido a capacidade de escolher a técnica certa para cada tipo de problema, validar sua performance com rigor e traduzir a saída de um modelo em uma decisão de negócio defensável.

SUA JORNADA COMPLETA

1

R

2

Estatística

3

EDA com Packages

4

Engenharia de Variáveis

5

PCA

6

Segmentação

7

Regressão

8

Classificação

9

Recomendação

10

Credit Scoring

- Programação em R: vetores, listas, data frames e funções
- Estatística descritiva e diagnóstico de variabilidade
- Análise exploratória de dados com packages especializados
- Engenharia de variáveis e testes de normalidade
- Redução de dimensionalidade com PCA
- Segmentação de consumidores (Hierárquico, KMeans, KModes)

- Regressão aplicada a custo, crédito, saúde e manutenção (Linear, Random Forest, XGBoosting, SVR)
- Classificação aplicada a churn e saúde (Árvore, KNN, Random Forest, SVM)
- Sistemas de recomendação com o Algoritmo Apriori
- Regressão Logística e comparação de classificadores
- Projeto prático completo de Credit Scoring, do dado bruto à recomendação executiva

PARA ONDE ESSA FORMAÇÃO PODE TE LEVAR

O conjunto de competências construído nesta formação é a base analítica exigida em posições como:

- Cientista de Dados
- Analista de Machine Learning
- Analista de Risco / Crédito
- Analista de CRM / Marketing Analytics
- Consultor(a) de Data Science
- Engenheiro(a) de Modelos Preditivos

Pronto para construir modelos que decidem com você?

Comece agora sua formação em Machine Learning com a Linguagem R.

QUERO COMEÇAR AGORA