Exame Preparatório Avançado – CertNexus Certified Artificial Intelligence Practitioner (AIP-210)

Nome do simulado: Preparatório Desafiador – Questões Amplas e Complexas por Tópico
Número de questões: Mais de 100 (máximo por tópico, variando tipos: múltipla escolha, verdadeiro/falso, contextual/cenário, e questões de análise aberta para raciocínio)
Formato: Múltipla escolha (A/B/C/D), Verdadeiro/Falso (V/F), Contextual (com cenários reais ou hipotéticos), e Análise (explicar ou discutir brevemente).
Duração sugerida: 60–90 minutos por lição/tópico (total 10–15 horas para estudo completo).

Instruções: As questões são projetadas para serem amplas, difíceis e integradoras, cobrindo conceitos avançados, cenários reais e interseções entre tópicos. Elas baseiam-se no eBook oficial do curso (CNX0016S_eBook_v10), glossário, soluções e conhecimentos gerais da certificação AIP-210. Use para revisão profunda; responda primeiro, depois verifique respostas e justificativas.

As questões estão organizadas por lição/tópico do curso, com o máximo possível (5–12 por lição, dependendo da profundidade). Incluí V/F para verificação rápida, múltipla escolha para decisão, contextual para aplicação prática, e análise para raciocínio crítico.

Lesson 1: Solving Business Problems Using AI and ML

(Questões amplas sobre identificação de soluções AI/ML, formulação de problemas, abordagens e ética em contextos empresariais complexos. Máximo: 10 questões.)

Questão 1.1 (Contextual/Múltipla Escolha)
Cenário: Uma empresa global de logística enfrenta flutuações imprevisíveis na demanda de frete devido a eventos geopolíticos, mudanças climáticas e variações econômicas. Eles coletaram dados históricos de 10 anos, mas o CEO questiona se AI/ML é viável, considerando custos altos e riscos éticos como viés em alocação de recursos para regiões subdesenvolvidas.
Qual abordagem integrada de formulação de problema ML seria mais apropriada para maximizar o valor de negócio enquanto mitiga riscos éticos?
A) Formular como regressão simples, ignorando ética para focar em precisão inicial.
B) Usar DOE para experimentar variáveis controláveis (ex.: rotas otimizadas), integrando análise de viés e transparência como métricas de performance.
C) Aplicar clustering não supervisionado sem formulação explícita, assumindo que padrões emergirão naturalmente.

D) Priorizar reinforcement learning para simulações reais, sem considerar stakeholders externos.

Resposta correta: B

Justificativa: A formulação de problemas ML (Lesson 1, Topic B) envolve DOE para variáveis independentes/dependentes, enquanto ética (ex.: viés, transparência) deve ser

integrada desde o início (Lesson 1, Topic A). Isso preserva integridade e alinha com stakeholders, evitando soluções isoladas.

Questão 1.2 (Verdadeiro/Falso)
Verdadeiro ou Falso: Em um projeto AI/ML para prever churn de clientes em um banco, a probabilidade de sucesso deve ser avaliada apenas pelo volume de dados disponíveis, ignorando fatores como aleatoriedade estocástica e requisitos de stakeholders.
Resposta correta: Falso

Justificativa: A probabilidade de sucesso (Lesson 1, Topic B) inclui aleatoriedade, requisitos de stakeholders (Lesson 1, Topic A) e recursos, não só dados. Ignorar isso leva a falhas de generalização.

Questão 1.3 (Múltipla Escolha)
Qual é o impacto mais crítico da aleatoriedade e incerteza em um modelo ML estocástico ao resolver problemas governamentais como previsão de crises de saúde pública?
A) Garante previsões perfeitas para eventos individuais.
B) Permite padrões gerais, mas exige mitigação via ensemble methods para reduzir variância.
C) Elimina a necessidade de experimentação DOE.

D) Torna todos os modelos determinísticos com dados suficientes.

Resposta correta: B

Justificativa: Modelos estocásticos (Lesson 1, Topic B) capturam padrões gerais, mas aleatoriedade causa variância; ensemble (Lesson 1, Topic C) mitiga, especialmente em cenários incertos como saúde pública.

Questão 1.4 (Análise/Contextual)
Cenário: Em uma ONG combatendo pobreza, um modelo ML usa dados de surveys para alocar recursos. Discuta como formular o problema considerando DIK hierarchy, ética (privacidade) e abordagens ML (supervised vs. unsupervised), e por que unsupervised pode ser arriscado.
Resposta correta (Análise Esperada): Formule como: Task (prever necessidades); Experience (dados de surveys); Performance (equidade em alocação). DIK: Transforme dados em conhecimento acionável. Ética: Proteja privacidade via anonymization. Supervised é preferível para labels claras (ex.: níveis de pobreza); unsupervised arriscado por clusters enviesados sem ground truth.

Justificativa: Integra DIK (Lesson 1, Topic A), formulação (Topic B) e abordagens (Topic C), destacando riscos éticos.

Questão 1.5 (Verdadeiro/Falso)
Verdadeiro ou Falso: Em problemas de interesse público como mudança climática, AI/ML sempre supera programação tradicional, independentemente da disponibilidade de dados acionáveis.
Resposta correta: Falso

Justificativa: Lesson 1, Topic B enfatiza avaliar se ML é apropriado vs. métodos simples; falta de dados reduz probabilidade de sucesso.

Questão 1.6 (Múltipla Escolha)
Em um cenário de pesquisa com gaps éticos (ex.: viés em estudos médicos), qual estratégia de comunicação com stakeholders maximiza transparência e accountability?
A) Relatar apenas resultados positivos.
B) Desenvolver plano que inclua experts do domínio, feedback público e auditorias éticas.
C) Ignorar governos, focando em parceiros internos.

D) Usar apenas métricas técnicas, sem contexto de negócios.

Resposta correta: B

Justificativa: Comunicação (Lesson 1, Topic A) envolve stakeholders para ética; transparência é chave em AI/ML.

Questão 1.7 (Contextual/Múltipla Escolha)
Cenário: Uma startup de fintech usa ML para detecção de fraude, mas ignora randomness, levando a overfitting em dados históricos enviesados. Como reformular o problema para maior probabilidade de sucesso?
A) Aumentar k em k-NN sem experimentação.
B) Incluir DOE para variáveis aleatórias e métricas como AUC para viés.
C) Mudar para unsupervised sem labels.

D) Reduzir dados para simplicidade.

Resposta correta: B

Justificativa: Reformulação (Lesson 1, Topic B) usa DOE para randomness; métricas como AUC avaliam sucesso.

Questão 1.8 (Verdadeiro/Falso)
Verdadeiro ou Falso: A hierarquia DIK implica que conhecimento é sempre derivado diretamente de dados brutos, sem necessidade de contexto de negócios.
Resposta correta: Falso

Justificativa: DIK (Lesson 1, Topic A) requer agregação e insights para transformar dados em conhecimento acionável.

Questão 1.9 (Análise)
Discuta por que reinforcement learning pode ser inadequado para problemas comerciais como otimização de estoque, comparado a supervised approaches.
Resposta correta (Análise Esperada): Reinforcement exige ambiente simulável com rewards; inadequado se dados históricos forem limitados ou riscos altos (ex.: custos de trial-and-error). Supervised usa labels existentes para previsões diretas, alinhando melhor com negócios.

Justificativa: Abordagens ML (Lesson 1, Topic C) destacam tradeoffs em learning modes.

Questão 1.10 (Múltipla Escolha)
Qual risco ético é mais proeminente em AI/ML para segurança governamental, como reconhecimento facial?
A) Excesso de transparência.
B) Falta de accountability em decisões autônomas.
C) Baixa aleatoriedade.

D) Sobredados disponíveis.

Resposta correta: B

Justificativa: Riscos éticos (Lesson 1, Topic A) incluem accountability em cenários sensíveis.

Lesson 2: Preparing Data

(Questões sobre coleta, transformação, feature engineering e unstructured data. Máximo: 12 questões, focando em desafios avançados como ética em dados e escalabilidade.)

Questão 2.1 (Contextual/Múltipla Escolha)
Cenário: Uma empresa de saúde coleta dados de wearables para prever riscos cardíacos, mas enfrenta 30% de missing values, outliers de sensores defeituosos e dados não estruturados de relatórios médicos.
Qual estratégia integrada de preparação de dados minimiza viés enquanto preserva integridade estatística?
A) Excluir todos dados com missing values para simplicidade.
B) Usar imputação mediana para outliers, feature engineering com Box-Cox para normalização, e embedding para unstructured data, com verificação ética para PII.
C) Aplicar deduplication sem transformação.

D) Ignorar unstructured data, focando apenas em numéricos.

Resposta correta: B

Justificativa: Preparação (Lesson 2, Topic B/C) inclui imputação robusta, Box-Cox para distribuições, embedding para unstructured (Topic D), e ética para PII.

Questão 2.2 (Verdadeiro/Falso)
Verdadeiro ou Falso: Em datasets com multicollinearity alta, feature selection via lasso sempre elimina variáveis redundantes sem impacto em performance.
Resposta correta: Falso

Justificativa: Feature engineering (Lesson 2, Topic C) nota que lasso reduz, mas performance depende de contexto; multicollinearity pode persistir.

Questão 2.3 (Múltipla Escolha)
Qual é o desafio mais crítico ao trabalhar com unstructured audio data em um sistema de detecção de fraude vocal?

A) Baixa amplitude ignora sampling rate.
B) Necessidade de Fourier transformation e MFCCs para feature extraction, lidando com noise e periodicity.
C) Conversão direta para texto sem preprocessing.

D) Ignorar spectrograms para velocidade.

Resposta correta: B

Justificativa: Unstructured data (Lesson 2, Topic D) requer Fourier, MFCCs para audio; noise e periodicity afetam qualidade.

Questão 2.4 (Análise/Contextual)
Cenário: Dados de vídeo de vigilância para detecção de anomalias têm resolution variada e augmentation necessária. Explique como preparar, incluindo reshaping, perturbation e ética (privacidade).
Resposta correta (Análise Esperada): Reshape para dimensões uniformes; perturbe com flips/rotations para augmentation; ética: anonymize faces via quasi-identifiers.

Justificativa: Video data (Lesson 2, Topic D) envolve preprocessing; ética em coleta (Topic A).

Questão 2.5 (Verdadeiro/Falso)
Verdadeiro ou Falso: Data binning sempre melhora model performance em continuous variables, independentemente da distribuição.
Resposta correta: Falso

Justificativa: Transformation (Lesson 2, Topic B) nota que binning discretiza, mas pode perder informação em distribuições uniformes.

Questão 2.6 (Múltipla Escolha)
Em um dataset com 50% unstructured text de reviews, qual sequência de preprocessing é mais eficaz para sentiment analysis?
A) Tokenization > Stemming > Bag of words > Stop words removal.
B) Stop words removal > Tokenization > Lemmatization > Embedding.
C) Embedding direto sem tokenization.

D) Deduplication > Normalization > Binning.

Resposta correta: B

Justificativa: Text data (Lesson 2, Topic D) segue stop words, tokenization, lemmatization, embedding para vetores densos.

Questão 2.7 (Contextual/Múltipla Escolha)
Cenário: ETL para big data em healthcare com issues de qualidade (duplicates, irregularities). Como endereçar escalabilidade e ética?
A) Usar ETL manual para privacidade.
B) Automatizar deduplication e imputation, com hashing para PII e verificação de viés.
C) Ignorar duplicates para velocidade.

D) Focar apenas em quantity issues.

Resposta correta: B

Justificativa: ETL (Lesson 2, Topic A), cleaning (Topic B), ética em dados.

Questão 2.8 (Verdadeiro/Falso)
Verdadeiro ou Falso: Feature scaling (normalization vs. standardization) é irrelevante em algorithms como decision trees.
Resposta correta: Verdadeiro

Justificativa: Feature engineering (Lesson 2, Topic C) nota que trees não dependem de scale, ao contrário de distance-based.

Questão 2.9 (Análise)
Discuta riscos éticos em feature engineering para dados sensíveis (ex.: raça em hiring models) e como mitigar com quasi-identifiers.
Resposta correta (Análise Esperada): Riscos: Viés amplificado. Mitigação: Remover quasi-identifiers, usar fairness metrics.

Justificativa: Ética em features (Lesson 2, Topic C).

Questão 2.10 (Múltipla Escolha)
Qual impacto da curse of dimensionality em large datasets?
A) Aumenta performance automaticamente.
B) Reduz capacidade de aprender padrões, exigindo dimensionality reduction.
C) Elimina need de feature selection.

D) Torna todos features relevantes.

Resposta correta: B

Justificativa: Dimensionality (Lesson 2, Topic C).

Questão 2.11 (Verdadeiro/Falso)
Verdadeiro ou Falso: Imputação com mean é robusta a outliers em distribuições skewed.
Resposta correta: Falso

Justificativa: Imputation (Lesson 2, Topic B) prefere median para skewed data.

Questão 2.12 (Múltipla Escolha)
Em unstructured images, por que augmentation com perturbation é essencial para robustness?
A) Aumenta noise.
B) Simula variações reais (ex.: iluminação), reduzindo overfitting.
C) Reduz resolution.

D) Ignora aspect ratio.

Resposta correta: B

Justificativa: Image data (Lesson 2, Topic D).

Lesson 3: Training, Evaluating, and Tuning a Machine Learning Model

(Máximo: 10 questões, focando em overfitting, metrics avançadas e tuning iterativo.)

Questão 3.1 (Contextual/Múltipla Escolha)
Cenário: Um modelo de detecção de câncer tem alta accuracy no treino (98%), mas cai para 70% em validação devido a imbalance e noise.
Qual estratégia de tuning e avaliação integrada resolve overfitting e melhora generalização?
A) Aumentar epochs sem cross-validation.
B) Usar k-fold stratified cross-validation, learning curves para bias-variance, e regularization.
C) Ignorar metrics como F1, focando accuracy.

D) Reduzir dados para simplicidade.

Resposta correta: B

Justificativa: Evaluation (Lesson 3, Topic B), tuning (Topic B), cross-validation para imbalance.

Questão 3.2 (Verdadeiro/Falso)
Verdadeiro ou Falso: Goodhart's Law implica que depender de uma métrica única (ex.: accuracy) pode distorcer avaliação em imbalanced datasets.
Resposta correta: Verdadeiro

Justificativa: Metrics (Lesson 3, Topic B).

Questão 3.3 (Múltipla Escolha)
Qual é o papel das learning curves em diagnosticar irreducible error?
A) Mostram quando mais dados não reduzem error.
B) Ignoram bias.
C) Sempre indicam underfitting.

D) Substituem cross-validation.

Resposta correta: A

Justificativa: Tuning (Lesson 3, Topic B).

Questão 3.4 (Análise/Contextual)
Cenário: Modelo com high variance em GPU-parallelized training. Explique como usar ensemble e regularization para tuning.
Resposta correta (Análise Esperada): Ensemble reduz variance combinando modelos; regularization (L1/L2) penaliza complexidade. Parallelize para eficiência.

Justificativa: Training (Lesson 3, Topic A), tuning.

Questão 3.5 (Verdadeiro/Falso)
Verdadeiro ou Falso: LOOCV é eficiente para large datasets.
Resposta correta: Falso

Justificativa: Cross-validation (Lesson 3, Topic B) nota LOO é computacionalmente caro.

Questão 3.6 (Múltipla Escolha)
Em um modelo black box, como avaliar performance além de accuracy?
A) Apenas com AUC.
B) Usar PRC para imbalance, F1 para tradeoff, e explainability tools.
C) Ignorar variance.

D) Focar em training time.

Resposta correta: B

Justificativa: Metrics (Lesson 3, Topic B).

Questão 3.7 (Contextual/Múltipla Escolha)
Cenário: Treino iterativo com overfitting. Como otimizar hyperparameters?
A) Grid search exaustivo.
B) Bayesian optimization para efficiency em large spaces.
C) Randomized search sem distribuição.

D) Manual tuning.

Resposta correta: B

Justificativa: Tuning (Lesson 3, Topic B).

Questão 3.8 (Verdadeiro/Falso)
Verdadeiro ou Falso: Irreducible error é sempre devido a overfitting.
Resposta correta: Falso

Justificativa: Error types (Lesson 3, Topic B).

Questão 3.9 (Análise)
Discuta bias-variance tradeoff em deep learning models.
Resposta correta (Análise Esperada): High bias (underfit) vs. high variance (overfit); balance via data aug, regularization.

Justificativa: Generalization (Lesson 3, Topic B).

Questão 3.10 (Múltipla Escolha)
Qual métrica é melhor para imbalanced classification?
A) Accuracy.
B) F1 score.
C) MSE.

D) R2.

Resposta correta: B

Justificativa: Metrics (Lesson 3, Topic B).

Lesson 4: Building Linear Regression Models

Questão 4.1 (Contextual / Múltipla Escolha)
Cenário: Você está construindo um modelo de precificação dinâmica de imóveis em uma grande cidade europeia. Os dados apresentam alta multicolinearidade entre área construída, número de quartos e valor do terreno, além de outliers extremos devido a mansões de luxo.
Qual abordagem integrada é a mais robusta para obter coeficientes interpretáveis e minimizar impacto de multicolinearidade e outliers?
A) Usar regressão linear simples com normal equation e sem regularização
B) Aplicar Ridge regression (L2) combinada com análise VIF e remoção seletiva de variáveis altamente colineares
C) Aplicar Lasso regression (L1) sem análise prévia de multicolinearidade

D) Usar Batch Gradient Descent sem regularização

Resposta correta: B

Justificativa: Ridge (L2) reduz impacto da multicolinearidade sem zerar coeficientes (mantém interpretabilidade), enquanto VIF ajuda a identificar e tratar colinearidade explicitamente (Lesson 4, Topic B). Lasso zera variáveis, o que pode ser indesejado quando todas são teoricamente relevantes.

Questão 4.2 (Verdadeiro/Falso)
Verdadeiro ou Falso: A normal equation sempre é preferível ao gradient descent em problemas de regressão linear com mais de 10.000 observações e 100 features.
Resposta correta: Falso

Justificativa: Normal equation tem complexidade $O(n^3)$ na inversão da matriz, tornando-a inviável para datasets grandes ou alta dimensionalidade (Lesson 4, Topic A vs. Topic C).

Questão 4.3 (Análise / Contextual)
Explique por que, em um cenário de previsão de consumo energético residencial com dados mensais de 15 anos, a aplicação de elastic net regression pode ser superior tanto ao Ridge quanto ao Lasso isoladamente.

Resposta correta (esperada): Elastic Net combina L1 (seleção de features) e L2 (estabilidade em multicolinearidade), sendo particularmente útil quando há grupos de variáveis correlacionadas (ex.: temperatura média, mínima, máxima, umidade). Mantém interpretabilidade parcial e lida melhor com multicolinearidade do que Lasso puro.

Questão 4.4 (Múltipla Escolha)
Qual das seguintes afirmações é **falsa** sobre a regularização em regressão linear?
A) Ridge penaliza a soma dos quadrados dos coeficientes
B) Lasso pode zerar coeficientes e realizar seleção automática de variáveis
C) Elastic Net nunca zera coeficientes quando λ é pequeno

D) Ridge mantém todos os coeficientes diferentes de zero (exceto em casos extremos)

Resposta correta: C

Justificativa: Elastic Net pode zerar coeficientes (herda propriedade do Lasso), mas de forma mais controlada que Lasso puro (Lesson 4, Topic B).

Questão 4.5 (Verdadeiro/Falso)
Verdadeiro ou Falso: Em um problema de regressão linear com 500 features e apenas 800 observações, a aplicação direta da normal equation sem regularização provavelmente resultará em overfitting severo e instabilidade numérica.
Resposta correta: Verdadeiro

Justificativa: Alta dimensionalidade + poucas observações → curse of dimensionality + multicolinearidade → instabilidade (Lesson 4, Topic A).

Lesson 5: Building Forecasting Models

Questão 5.1 (Contextual / Múltipla Escolha)
Cenário: Uma rede de supermercados quer prever vendas diárias de produtos perecíveis considerando sazonalidade semanal, feriados nacionais e promoções regionais. Os dados apresentam tendência crescente e heterocedasticidade.
Qual pipeline de modelagem é o mais apropriado?
A) ARIMA simples sem diferenciação
B) SARIMA com componentes sazonais + regressores exógenos (SARIMAX)
C) VAR sem verificação de estacionariedade

D) ARIMA com ordem (0,0,0)

Resposta correta: B

Justificativa: SARIMAX lida com sazonalidade, tendência (via diferenciação) e variáveis exógenas (promoções, feriados) — ideal para forecasting multivariado com sazonalidade (Lesson 5, Topic A/B).

Questão 5.2 (Verdadeiro/Falso)
Verdadeiro ou Falso: Em séries temporais multivariadas, o modelo VAR assume que todas as variáveis são endógenas e não exige diferenciação se as séries forem cointegradas.
Resposta correta: Verdadeiro

Justificativa: VAR modela interdependências endógenas; se houver cointegration, usa-se VECM em vez de VAR em níveis (Lesson 5, Topic B).

Questão 5.3 (Análise)
Explique a diferença prática entre usar ARIMA vs. SARIMA vs. SARIMAX em um problema de previsão de demanda elétrica horária com forte sazonalidade diária e semanal, e influência de temperatura.

Resposta correta (esperada): ARIMA → apenas univariado sem sazonalidade explícita. SARIMA → captura sazonalidade (ex.: P,D,Q,s). SARIMAX → adiciona regressores

exógenos (temperatura), essencial para melhorar precisão em variáveis externas conhecidas.

Questão 5.4 (Múltipla Escolha)
Qual teste estatístico é mais apropriado para verificar estacionariedade em uma série temporal antes de aplicar ARIMA?
A) Teste de Dickey-Fuller aumentado (ADF)
B) Teste de Shapiro-Wilk
C) Teste de Levene

D) Teste qui-quadrado

Resposta correta: A

Justificativa: ADF testa presença de raiz unitária (não-estacionariedade) — padrão em forecasting (Lesson 5, Topic A).

Lesson 6: Building Classification Models Using Logistic Regression and k-Nearest Neighbor

Questão 6.1 (Contextual / Múltipla Escolha)
Cenário: Classificação de risco de crédito em uma fintech com forte desbalanceamento (apenas 3% de inadimplentes) e alta dimensionalidade (200 features).
Qual combinação de técnicas oferece o melhor trade-off entre interpretabilidade, performance e robustez?
A) Logistic Regression + SMOTE + L1 regularization
B) k-NN com k=1 + sem balanceamento
C) Logistic Regression sem regularização + undersampling aleatório

D) k-NN com k=50 + sem feature selection

Resposta correta: A

Justificativa: Logistic Regression oferece interpretabilidade + coeficientes; L1 faz feature selection; SMOTE lida com imbalance (Lesson 6, Topic A/B/E).

Questão 6.2 (Verdadeiro/Falso)
Verdadeiro ou Falso: A função de custo da regressão logística (cross-entropy) é convexa, garantindo convergência global com gradient descent adequado.
Resposta correta: Verdadeiro

Justificativa: Convexidade é propriedade fundamental da logistic regression (Lesson 6, Topic A).

Questão 6.3 (Múltipla Escolha)
Em um problema de classificação binária com threshold padrão 0.5, mover o threshold para 0.7 geralmente causa:
A) Aumento de recall e diminuição de precision
B) Aumento de precision e diminuição de recall
C) Aumento simultâneo de ambos

D) Diminuição simultânea de ambos

Resposta correta: B

Justificativa: Threshold mais alto → mais conservador → menos falsos positivos → maior precision, menor recall (Lesson 6, Topic D).

Lesson 7: Building Clustering Models

Questão 7.1 (Análise / Contextual)
Cenário: Segmentação de clientes de e-commerce com comportamento altamente não esférico (clusters em espiral ou alongados).
Por que k-means provavelmente falhará e qual alternativa hierárquica seria mais apropriada?

Resposta correta (esperada): k-means assume clusters esféricos e baseados em centróides → falha em formas complexas. Hierarchical agglomerative clustering (com linkage ward ou complete) + dendrograma permite visualizar e cortar em clusters não esféricos.

Questão 7.2 (Verdadeiro/Falso)
Verdadeiro ou Falso: O índice de Davies-Bouldin é minimizado quando os clusters são compactos e bem separados.
Resposta correta: Verdadeiro

Justificativa: DB index mede razão within-cluster / between-cluster (Lesson 7, Topic A/B).

Questão 7.3 (Múltipla Escolha)
Qual linkage é mais robusto a outliers em hierarchical clustering?
A) Single linkage
B) Complete linkage
C) Average linkage

D) Ward linkage

Resposta correta: B

Justificativa: Complete linkage considera a distância máxima → menos sensível a outliers que single linkage.

Lesson 8: Building Decision Trees and Random Forests

Questão 8.1 (Contextual / Múltipla Escolha)
Cenário: Modelo de churn em telecom com 1.2 milhões de linhas e 300 features. Alta variância observada em árvores individuais.
Qual ensemble oferece o melhor equilíbrio entre performance, interpretabilidade parcial e tempo de treinamento?
A) Uma única árvore CART profunda
B) Random Forest com 500 árvores + max_features = sqrt(n)
C) Gradient Boosting com learning rate 0.01 e 2000 estimadores

D) Bagging com árvores rasas sem feature subsampling

Resposta correta: B

Justificativa: Random Forest reduz variância via bagging + random feature selection, escalável e com feature importance (Lesson 8, Topic B).

Questão 8.2 (Verdadeiro/Falso)
Verdadeiro ou Falso: O out-of-bag error em Random Forest é uma estimativa não enviesada do erro de generalização, similar ao leave-one-out cross-validation.
Resposta correta: Verdadeiro

Justificativa: OOB error usa amostras não usadas em cada bootstrap (Lesson 8, Topic B).

Lesson 9: Building Support-Vector Machines

Questão 9.1 (Análise)
Explique por que o kernel RBF (Gaussian) permite separar dados circularmente concêntricos, mas aumenta o risco de overfitting em datasets pequenos.

Resposta correta (esperada): Kernel trick mapeia para espaço de alta dimensão onde hiperplano separa círculos; γ alto → fronteira muito irregular → overfitting (Lesson 9, Topic A).

Questão 9.2 (Múltipla Escolha)
Em SVM para regressão (SVR), o parâmetro ε controla:
A) A largura da margem de tolerância
B) A penalidade por violações da margem
C) O grau do kernel polinomial

D) O raio do kernel RBF

Resposta correta: A

Justificativa: ε define a faixa de erro aceitável sem penalidade (Lesson 9, Topic B).

Lesson 10: Building Artificial Neural Networks

Questão 10.1 (Contextual / Múltipla Escolha)
Cenário: Detecção de defeitos em peças metálicas via imagens industriais de alta resolução.
Qual arquitetura e configuração é a mais indicada considerando eficiência computacional e capacidade de capturar texturas finas?
A) MLP com 5 camadas densas
B) CNN com várias camadas convolucionais + batch normalization + global average pooling
C) RNN com LSTM para pixels sequenciais

D) Transformer sem convolução

Resposta correta: B

Justificativa: CNN + pooling + BN é padrão ouro para visão computacional industrial (Lesson 10, Topic B).

Questão 10.2 (Verdadeiro/Falso)
Verdadeiro ou Falso: Em RNNs vanilla, o vanishing gradient ocorre principalmente devido à multiplicação repetida de gradientes < 1 ao longo do tempo.
Resposta correta: Verdadeiro

Justificativa: Problema clássico resolvido por LSTM/GRU (Lesson 10, Topic C).

Lesson 11: Operationalizing Machine Learning Models

Questão 11.1 (Análise / Contextual)
Cenário: Modelo de recomendação de produtos em produção há 9 meses. Latência aumentou 400% e acurácia caiu 12%.
Quais etapas de MLOps você priorizaria para diagnosticar e corrigir?

Resposta correta (esperada): 1. Monitoramento de drift (concept/data) 2. Re-treino automatizado via CI/CD 3. Endpoint com auto-scaling 4. Versionamento de modelos + rollback 5. Logging de predições vs. real.

Questão 11.2 (Múltipla Escolha)
Qual é a principal vantagem de usar Docker + Kubernetes para deployment de modelos ML em comparação a servidores bare-metal?
A) Menor latência em predição
B) Portabilidade, escalabilidade horizontal e isolamento de dependências
C) Menor consumo de memória

D) Maior interpretabilidade do modelo

Resposta correta: B

Justificativa: MLOps e deployment (Lesson 11, Topic A/B).

Lesson 12: Maintaining Machine Learning Operations

Questão 12.1 (Contextual / Múltipla Escolha)
Cenário: Após ataque adversarial em modelo de reconhecimento facial corporativo, a empresa quer proteger o pipeline inteiro.
Qual conjunto de medidas de segurança é o mais abrangente?
A) Apenas hashing do modelo treinado
B) RBAC + model signing + adversarial robustness training + monitoramento contínuo de inputs + pen-test
C) Somente logging de predições

D) Atualização semanal do modelo sem verificação

Resposta correta: B

Justificativa: Segurança em ML (Lesson 12, Topic A) inclui múltiplas camadas: acesso, integridade, robustness, monitoramento.

Questão 12.2 (Verdadeiro/Falso)
Verdadeiro ou Falso: Model drift sempre implica que o retraining deve ser feito imediatamente, independentemente da severidade da queda de performance.
Resposta correta: Falso

Justificativa: Monitoramento (Lesson 12, Topic B) deve avaliar impacto antes de retrain; pode haver drift benigno.