1.1 Exercícios Conceituais

1. Com suas palavras, forneça uma definição para a aprendizagem de máquina:

Aprendizado de Máquina é um método de análise de dados que automatiza o desenvolvimento de modelos analíticos. é uma forma de fazer o computador aprender como se comportar com dados desconhecidos partindo da analise de um conjunto de dados conhecidos.

2. Cite, pelo menos, três problemas reais nos quais técnicas de Machine Learning poderiam ser utilizadas.

- Recomendação de filmes, noticias.
- Tradução automática.
- Detecção de objetos em imagens.

3. Diferencie aprendizagem supervisionada da não supervisionada.

Na aprendizagem supervisionada se tem um conjunto de dados e seus rotulos previamente conhecido, e com esse conjunto se treina como o programa se deve comportar com dados desconhecidos. Na aprendizagem não supervisionada o conjunto de dados utilizado não possui nenhum tipo de rótulo. O objetivo desse tipo de aprendizagem é descobrir similaridades entre os objetos.

4. Qual o significado dos dados de treinamento rotulados (label traning dataset)?

São os dados conhecidos que serão utilizados para treinar o modelo.

5. Defina, com suas palavras, o que é um modelo de machine learning.

Um modelo de machine learning é um arquivo que foi treinado para reconhecer determinados tipos de padrões. Você treina um modelo em um conjunto de dados, fornecendo a ele um algoritmo que pode ser usado para ponderar e aprender com esses dados.

6. Que tipo de algoritmo de machine learning, em termos de categoria, poderia ser usado para segmentar clientes em múltiplos grupos?

Aprendizagem não supervisionada(K-Means, Hierarquical Cluster Analysis - HCA)

7. Explique, com suas palavras, as principais diferenças entre aprendizagem online e offline

Na aprendizagem offline todos os dados são processados para construir o modelo enquanto o aprendizado online é uma abordagem que processa os dados uma observação de cada vez.

8. Qual é a diferença entre os parâmetros e hiperparâmetros em um modelo de ML?

tudo que nós informamos para um modelo ou algoritmo antes dele começar o treino é um hiperparâmetro, e tudo que ele aprende/adapta com o treino é um parâmetro.

9. Explique a diferença entre os modos de aprendizagem que são baseados em modelos ou instâncias.

O aprendizado baseado em instância identifica todas as instâncias de um determinado recurso nos dados de treinamento e usa uma medida de similaridade para generalizar para novos casos. A aprendizagem baseada em modelo usa recursos nos dados de treinamento para prever o resultado / variável de interesse, o modelo usado para especificar a relação entre o preditor e resultado é então generalizado em novos casos.

10. Se um modelo de ML atinge um bom desempenho sobre os dados de treinamento, mas não generaliza bem

para novos dados (teste), o que pode estar acontecendo? O que poderia ser realizado para melhorar a generalização do modelo de ML?

Nesse caso o modelo ficou muito especializado nos dados de treino. Deve-se verificar se o modelo está sendo treinado com dados de testes, verificar as métricas.

1.2 Exercícios de Múltipla Escolha(a opção escolhida está em negrito).

1. Exercício 1 (Fundamentos de Machine Learning) Considere um algoritmo de aprendizagem de máquina que interpreta marcações de e-mail (spam ou não spam) realizadas por um usuário. Baseado nesta observação, o algoritmo aprende a filtrar os e-mails de forma mais eficaz. Neste caso, a tarefa T da definição de aprendizagem de máquina consiste em

- **a) Classificar um e-mail como spam ou não spam.**
- b) Observar as marcações de e-mail como spam ou não spam.
- c) O número ou razão de e-mails corretamente classificados como spam ou não spam.
- d) Não é possível aplicar aprendizagem de máquina neste caso do enunciado

2. Exercício 2 (Fundamentos de Machine Learning) Considere um algoritmo de aprendizagem de máquina que interpreta marcações de e-mail (spam ou não spam) realizadas por um usuário. Baseado nesta observação, o algoritmo aprende a filtrar os e-mails de forma mais eficaz. Neste caso, a métrica P da definição de aprendizagem de máquina consiste em

- a) Classificar um e-mail como spam ou não spam.
- b) Observar as marcações de e-mail como spam ou não spam.
- **c) O número ou razão de e-mails corretamente classificados como spam ou não spam.**
- d) Não é possível aplicar aprendizagem de máquina neste caso do enunciado.

3. Exercício 3 (Fundamentos de Machine Learning) Considere um algoritmo de aprendizagem de máquina que interpreta marcações de e-mail (spam ou não spam) realizadas por um usuário. Baseado nesta observação, o algoritmo aprende a filtrar os e-mails de forma mais eficaz. Neste caso, a experiência E da definição de aprendizagem de máquina consiste em

- **a) Classificar um e-mail como spam ou não spam.**
- b) Observar as marcações de e-mail como spam ou não spam.
- c) O número ou razão de e-mails corretamente classificados como spam ou não spam.
- d) Não é possível aplicar aprendizagem de máquina neste caso do enunciado.

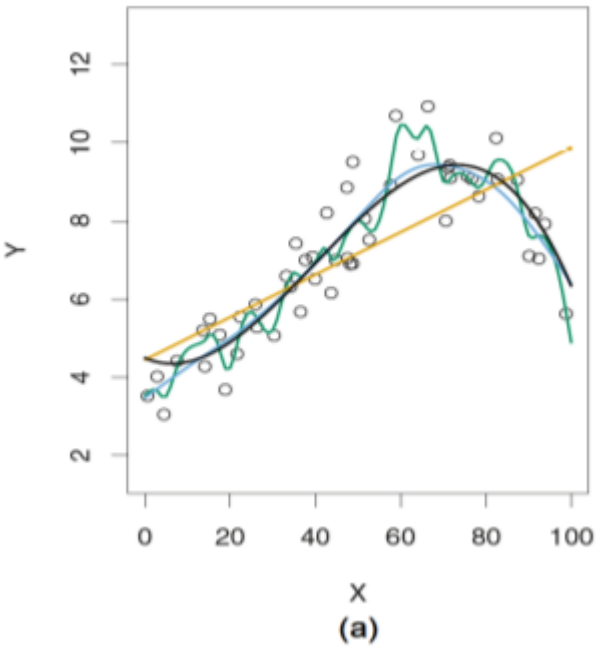
4. Exercício 4 (Métricas de Desempenho) A avaliação de performance ou desempenho de modelos de machine learning é um ponto de relevância e, de fato, temos uma etapa de avaliação que pode fazer parte de um projeto de ciência de dados. Uma métrica amplamente conhecida na literatura é colocada abaixo:

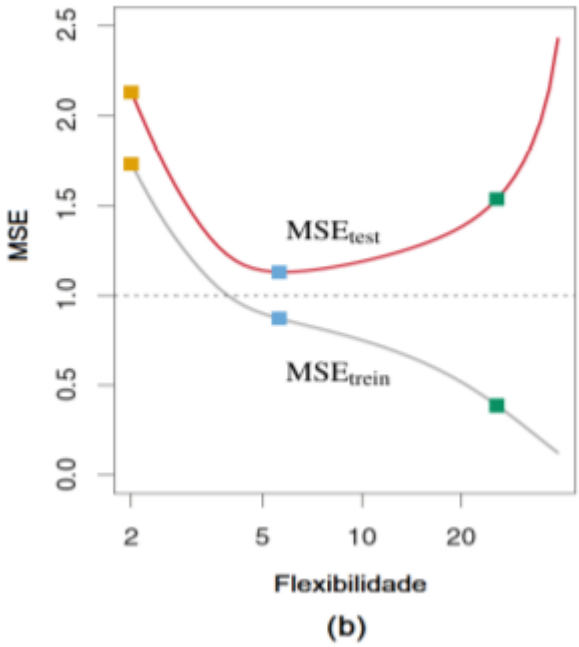
$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{h}(\mathbf{x}_i))^2$$

Sobre a interpretação dessa métrica de desempenho, marque a alternativa correta:

- a) A aplicação da métrica MSE só ocorre na aprendizagem não supervisionada.
- b) Na equação, o termo yi consiste na i-ésima estimativa do modelo de ML.
- c) Só é possível estimar, empiricamente, o MSE para o conjunto de treinamento.
- **d) Quanto maior é o valor do MSE, pior será o desempenho do modelo de ML avaliado.**

5. Exercício 5 (Métricas de Desempenho) As Figuras (a) e (b) abaixo, extraídas do livro An Introduction to Statistical Learning, discutem a relação entre o MSEtrein e MSEtest, ou seja, o desempenho dos modelos de ML nas fases de treinamento e teste. Sobre tal relação, marque a alternativa correta:





- a) A razão pela qual o MSEtest não segue o decaimento do MSEtrein reside na falha de generalização do modelo smoothing splines utilizado.
- b) O comportamento em "U" para curva do MSEtest ocorre porque a função hipótese verdadeira é do tipo não linear.
- c) Nota-se que existe uma garantia de performance de teste (i.e., baixo MSEtest) se nós ajustarmos o modelo de ML com os dados de treinamento.
- **d) A diferença entre MSEtrein e MSEtest é explicada pelo fato de que o processo de aprendizagem das técnicas de ML se baseia na minimização do MSEtrein e, por conta**

disso, não pode garantir ótima generalização para os dados de teste (i.e., baixo MSEtest).

6. Exercício 6 (Técnicas de ML) No estudo de machine learning, realizar a associação de técnicas de aprendizagem de máquina de acordo com o supervisionamento aplicado no processo de treinamento do modelo é um aspecto importante. Sobre esse tópico, marquea alternativa correta:

- a) O algoritmo K-Nearest Neighbors ou K-Vizinhos mais próximos pode ser aplicado em problemas somente de forma não supervisionada.
- b) Não existe uma relação entre os algoritmos - Árvores de Decisão (Decision Threes) e Florestas Aleatórias (Random Forests).
- **c) Redes neurais artificiais podem ser usadas em problemas considerando a aprendizagem supervisionada e não supervisionada.**
- d) A análise de componentes principais (PCA) é uma técnica supervisionada de aprendizagem de máquina.

7. Exercício 7 (Desempenho de Classificadores) Um hospital conta com uma equipe de pesquisadores em ciência de dados e inteligência artificial, avaliando diversos classificadores construídos para análise de problemas pulmonares dos pacientes. O objetivo consiste em levantar a performance dos classificadores para compreender, adotar e posteriormente testar os melhores modelos treinados. Cada classificador atua para apontar riscos de doenças pulmonares em futuros pacientes, isto é, o resultado da aplicação de modelo de ML aponta a presença ou ausência de risco de doença pulmonar de um determinado paciente que dá entrada no hospital com sintomas relacionados à parte respiratória-pulmonar*. O exemplo abaixo consiste nos resultados de desempenho sobre o treinamento de um classificador construído a partir de centenas de imagens médicas pulmonares armazenadas no banco de dados de um hospital.

Matriz de Resultados de Treinamento do Classificador

		Predição do Classificador	
		Risco presente	Risco ausente
Realidade	Risco presente	VP = 100	FN = 70
	Risco ausente	FP = 150	VN = 1200

Considerando as informações apresentadas, marque a alternativa correta:

- a) A acurácia do classificador é 85%, caracterizando o desempenho de forma completa.
- **b) A precisão calculada permite dizer que o classificador tem alto desempenho.**

- c) O desbalanceamento dos dados, especialmente com a quantidade de pacientes sem risco, aumenta a acurácia do modelo.
- d) Falsos positivos e negativos têm impacto equivalente sobre o suporte às decisões dos resultados do classificador.

8. Exercício 8 (Técnicas de ML) Alguns livros de ML trazem a informação de que a aprendizagem não supervisionada é desafiadora, no comparativo com o treinamento de modelo de forma supervisionada. Claro, o treinamento de um modelo não supervisionado ocorre de modo diferente pelo simples fato de não possuirmos dados rotulados com as saídas conhecidas. Nesse sentido, a aprendizagem não supervisionada é conduzida como parte da análise exploratória de dados e uma das técnicas mais conhecidas da aprendizagem estatística é a análise de componentes principais, PCA - Principal Component Analysis. Marque a alternativa que descreve corretamente qual é o conceito fundamental da PCA:

- **a) A PCA é uma técnica de aprendizagem de máquina não supervisionada e se refere a um processo de cômputo dos componentes principais de um grupo maior de variáveis (características), permitindo descrever a variabilidade estatística contida nos dados com um grupo menor de variáveis.**
- b) PCA é uma técnica de aprendizagem semisupervisionada e se refere a um processo de aglomeração de variáveis a partir de um conjunto menor de características, chamadas de componentes principais.
- c) A PCA é uma técnica de aprendizagem não supervisionada e se refere ao processo de predição de variáveis ou características de interesse, a partir de componentes principais não ortogonais entre si.
- d) A PCA é uma técnica de aprendizagem não supervisionada e se refere a um processo de visualização de dados para que todas as dimensões e características possam ser analisadas por meio de gráficos de correlação entre todas as variáveis.

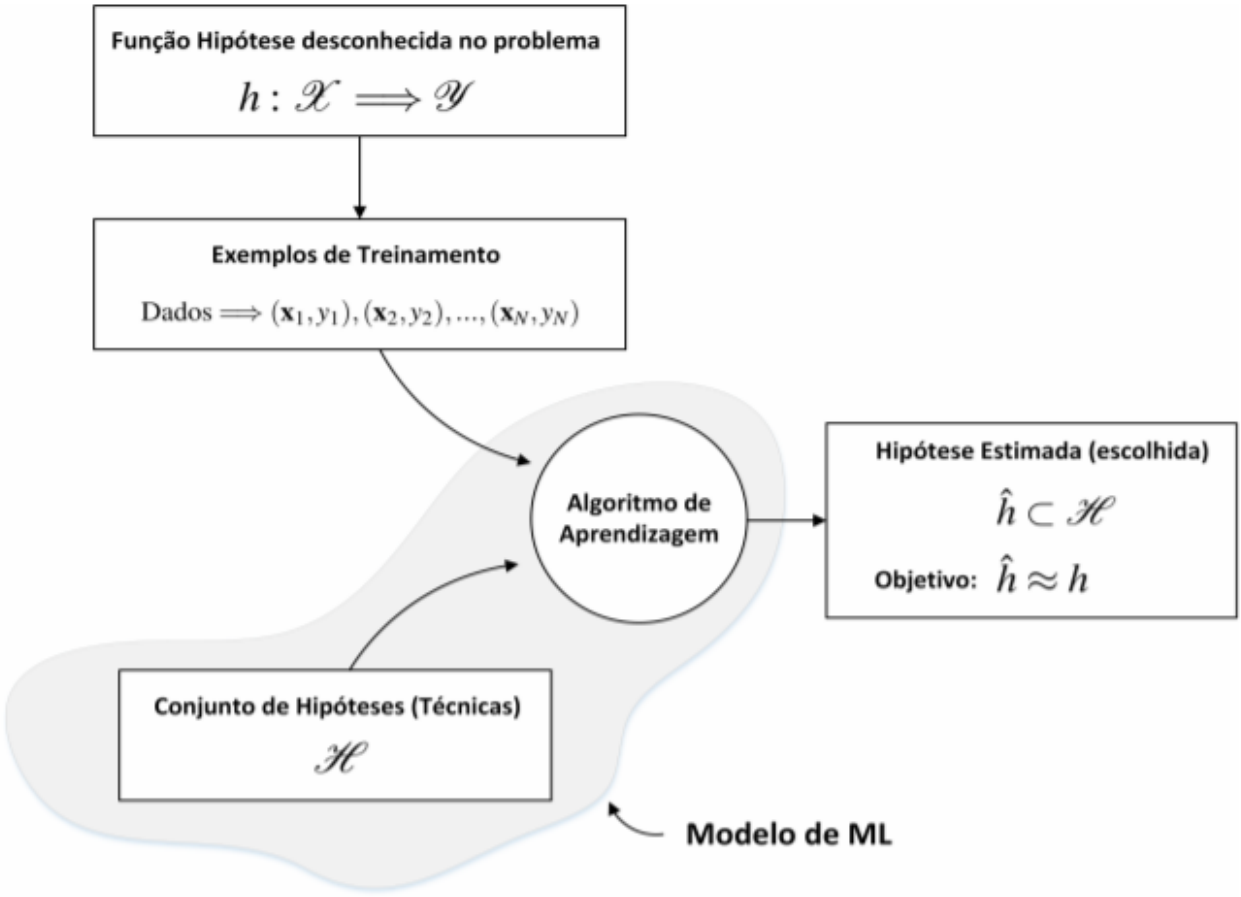
9. Exercício 9 (Modos de Aprendizagem) Na categoria modos de aprendizagem, marque a alternativa que descreve corretamente qual é a diferença de generalização entre modo de aprendizagem baseado em instâncias e modelos:

- **a) Ambos os modos de aprendizagem, baseados em instâncias e modelos, se fundamentam apenas em métricas de similaridade. – Na classificação de um potencial motorista para os serviços de corrida (aplicativos de transporte privado como Uber) as distâncias entre as coordenadas de localização de uma pessoa (por meio de seu smarphone) e os potenciais motoristas são um exemplo de informação que pode ser incluída no cálculo de métricas de similaridade.**
- b) Na aprendizagem baseada em instâncias a generalização realizada pelo modelo de ML é baseada em métricas de similaridade, enquanto a aprendizagem baseada em modelos formula equações matemáticas que são usadas para fazer a generalização. – Na classificação de um potencial motorista para os serviços de corrida (aplicativos de transporte privado como Uber) é possível aplicar modelos baseados em instâncias para

classificação de potenciais motoristas como também a construção de classificadores a partir de modelos matemáticos usados para tal tarefa (classificação).

- c) A aprendizagem baseada em modelos realiza sua generalização por meio de equações matemáticas, tal como o algoritmo K-NN, enquanto a generalização por instâncias se baseia em métricas de desempenho.
- d) Não existem diferenças entre os modos de aprendizagem de máquina baseados em instâncias e modelos.

10. Exercício 10 (Fundamentos de ML) Em vários livros e artigos científicos, além dos diversos materiais que encontramos na internet, vemos o uso intercambiável ou equivalente entre as expressões algoritmos de aprendizagem supervisionados, técnicas e/ou modelos de machine learning supervisionadas. A Figura abaixo apresenta uma terminologia adequada e em sintonia com diversas literaturas de alto nível da área e nos permite esclarecer os conceitos e diferenças entre essas expressões para o caso supervisionado.



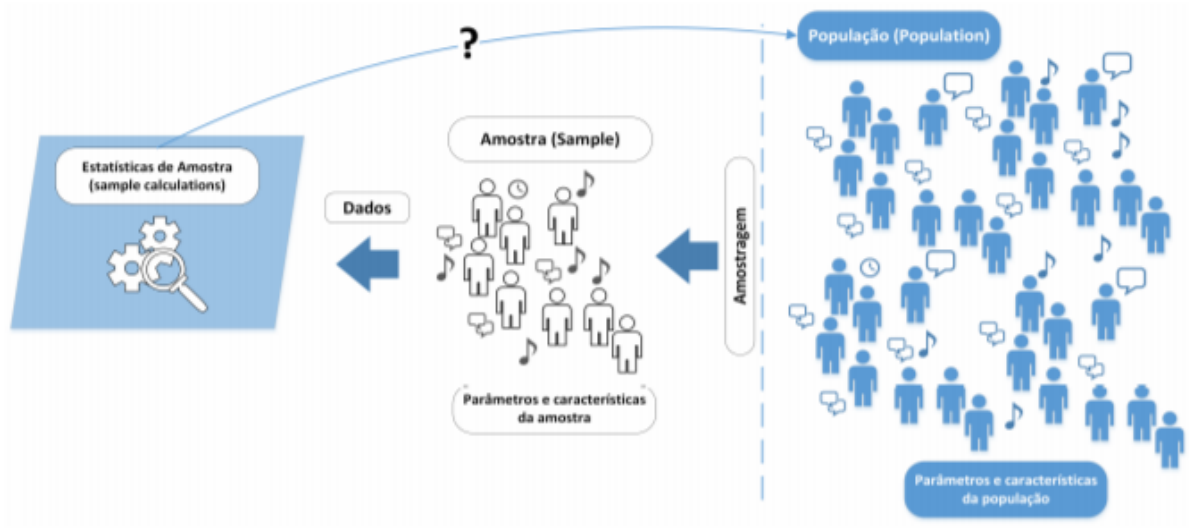
Baseado na figura, marque a alternativa que descreve o conceito de modelo supervisionado de machine learning:

- a) O modelo de aprendizagem de máquina é o algoritmo de aprendizagem usado para treinamento.
- **b) O modelo de aprendizagem de máquina supervisionado consiste na combinação entre uma função hipótese candidata e um algoritmo de aprendizagem.**

- c) O modelo de aprendizagem de máquina consiste em um teste feito entre a função hipótese candidata e a verdadeira.
- d) O modelo de aprendizagem de máquina consiste no conjunto de funções hipóteses candidatas que são combinadas com um único algoritmo de aprendizagem de máquina.

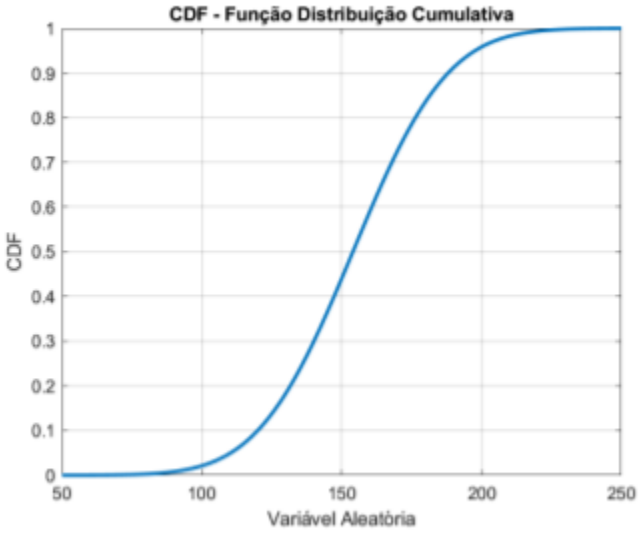
1.3 Exercícios de Revisão - Probabilidade/Estatística

1. Baseado na figura abaixo, apresente suas explicações sobre a diferença entre população e amostra. De que parte da estatística nós estamos falando quando queremos estimar parâmetros populacionais a partir dos dados de uma amostra?



Uma população é um conjunto de pessoas, itens ou eventos sobre os quais você quer fazer inferências. Uma amostra é um subconjunto de pessoas, itens ou eventos de uma população maior que você coleta e analisa para fazer inferências. Para representar a população bem, uma amostra deve ser coletada aleatoriamente e ser adequadamente grande. Esses tópicos fazem parte de **estatística inferencial**

2. Em probabilidade, nós utilizamos as funções PDF (probability density function) e CDF (cumulative distribution function), mostradas abaixo, para a caracterização estatística de variáveis aleatórias. Explique a diferença entre essas funções e como podemos calcular probabilidades a partir de cada uma delas (PDF e CDF).



As funções **PDF** desecrevem valores de densidade probalistica que uma função matemática atribui a cada valor x já a função **CDF** por sua vez, retorna a probalidade acumulada até o valor x.