UNIVERSIDADE FEDERAL DO MARANHÃO
DEPARTAMENTO DE ENGENHARIA DA COMPUTAÇÃO

Prova ___

Curso: Engenharia da Computação	**Ano / Semestre:**	**Data:**
Disciplina: Mineração de Dados	**Professor:** Thales Levi Azevedo Valente	
Aluno:	**Código:**	

OBS: RESPOSTAS DEVEM SER ENTREGUES À CANETA. CASO O ALUNO USE LÁPIS, O MESMO NÃO TERÁ DIREITO A REVISÃO DAS QUESTÕES.

OBS2. PREENCHA TODOS OS CAMPOS ACIMA. QUALQUER CAMPO SEM PREENCHIMENTO COM OS RESPECTIVOS DADOS CORRETOS PODERÁ ACARRETAR EM 0 NA NOTA.

.

Descrição do Cenário

Você foi contratado(a) como engenheiro(a) de machine learning em um hospital referência em tratamento de doenças neurológicas, especializado em análise de imagens médicas (ressonâncias magnéticas) e sinais biomédicos (EEG) para acompanhamento longitudinal da progressão de doenças como Alzheimer e Parkinson. O hospital pretende utilizar técnicas avançadas de mineração de dados e IA para detectar precocemente sinais de progressão das doenças e personalizar o tratamento dos pacientes. A base de dados que você irá trabalhar possui dados de centenas de pacientes coletados ao longo dos últimos 20 anos, incluindo imagens médicas periódicas, registros de sinais EEG coletados no mesmo dia dos exames de imagens e dados tabulares/textuais, como fatores de exames de sangue, prescrição de medicamentos, dosagens, tempo de uso, entre outros. Com base no contexto, responda:

Com base na descrição do cenário baseado em uma entrevista com o cliente, responda:

<u>Questão 1: Conceitos Fundamentais</u>

- **Pergunta 1.1:** Explique as métricas de avaliação que você utilizaria (inclusive matematicamente) no contexto de acompanhamento de doenças neurológicas através da classificação de imagens e sinais biomédicos em saudável e não saudável. Você priorizaria alguma métrica? Caso a resposta seja positiva, indique qual e justique. Caso a resposta seja negativa, apenas justique. Suas justificativas devem ser contextualizadas no cenário.

Resposta. No contexto de acompanhamento longitudinal da progressão de doenças neurológicas, a principal tarefa de classificação envolve separar pacientes em saudáveis e não saudáveis com base em imagens médicas (ressonâncias magnéticas) e sinais biomédicos (EEG). Para avaliar a qualidade do modelo de aprendizado de máquina utilizado nessa classificação, devemos escolher métricas adequadas, considerando que bases médicas são frequentemente desbalanceadas (há mais pacientes saudáveis do que doentes). Dessa forma, **acurácia não é uma métrica confiável nesse caso**. Para entender as métricas de avaliação utilizadas, primeiro precisamos definir alguns termos:

- TP (True Positives) = Pacientes não saudáveis corretamente identificados.
- TN (True Negatives) = Pacientes saudáveis corretamente identificados
- FN (False Negatives) = Pacientes não saudáveis incorretamente classificados como saudáveis.
- FP (False Positives) = Pacientes saudáveis erroneamente classificados como doentes.

Agora podemos definir as métricas de avaliação. As mais comumente utilizadas seriam:

- Recall (Sensibilidade)
  - Mede a proporção de pacientes não saudáveis corretamente identificados pelo modelo.
  - Fórmula: $\mathit{R}\text{ecall} = \text{TP} \,/\, (\text{TP} + \text{FN})$
- Precisão (Precision)
  - Mede a proporção de pacientes identificados como não saudáveis que realmente estão doentes..
  - Fórmula: $\mathit{R}\text{ecall} = \text{TP} \,/\, (\text{TP} + \text{FP})$
- F1-Score
  - O F1-Score é a média harmônica entre recall e precisão, equilibrando ambas as métricas.
  - Fórmula: $\text{F1-Score} = 2 \text{ x Precision x Recall} \,/\, (\text{Precision} + \text{Recall})$

Quanto a priorização das métricas, a resposta é sim, priorizaria Recall (Sensibilidade) como métrica principal, com ajustes posteriores para melhorar a Precisão, sem comprometer, na medida do possível, a efetividade da detecção de pacientes doentes já alcançada.

- **Pergunta 1.2:** Diferencie dados temporais e não temporais, dando exemplos específicos relacionados às imagens médicas e sinais biomédicos descritos no cenário.

Resposta. Os dados temporais são aqueles em que a ordem cronológica é fundamental para a análise. Esses dados representam uma sequência de medições ao longo do tempo, e a relação entre pontos sucessivos influencia a interpretação dos padrões. Caso contrário, os dados são não temporais. No contexto descrito acima, acompanhamento longitudinal da progressão de doenças como Alzheimer e Parkinson, todos os dados são temporais uma vez que se referem a exames periódicos para acompanhamento longitudinal da progressão de doenças. Mais especificamente:

- Sinais de EEG: esses sinais são coletados como séries temporais, onde cada ponto de dados representa uma medição em um instante específico.
- Imagens longitudinais: o hospital coleta imagens periódicas dos pacientes ao longo dos anos para monitorar a progressão da doença. Comparar imagens de um mesmo paciente em diferentes momentos pode revelar mudanças estruturais no cérebro, ajudando a detectar o momento aproximado do surgimento ou a prever a evolução da doença.
- Dados de Tratamento e Prescrição de Medicamentos: o histórico de uso de medicamentos (dose, tempo de uso, mudanças na prescrição) ao longo do tempo é um dado temporal, pois sua evolução influencia a resposta do paciente ao tratamento.

Questão 2: Pré-processamento de Dados

- **Pergunta 2.1:** Imagine que alguns pacientes possuem imagens faltantes, sinais biomédicos faltantes e dados tabulares faltantes, em quaisquer uma das combinações possíveis. Imagine também que você tem conhecimento para criar modelos multimodais (alimentados por todo tipo de dados ao mesmo tempo) e modelos de tipo

de dado único, com recursos computacionais suficientes. Você sabe que um modelo multimodal poderia se beneficiar de um compartilhamento maior de informações, mas também sabe que emsemble de modelos pode ser uma outra alternativa. Explique como você analisaria os dados de todos os pacientes para tomar a decisão se criaria um modelo multimodal ou não. Considere utopicamente que não houve variações de tipo de aparelho para coleta.

Resposta. Para a tomada de decisão em relação a criar um modelo multimodal ou não, primeiro seria realizada uma **análise exploratória** para verificar quantos pacientes possuem todos os tipos de dados disponíveis (imagens, sinais biomédicos e dados tabulares) e quantos possuem dados ausentes em uma ou mais modalidades. Essa análise pode ser feita gerando uma matriz de disponibilidade, onde cada paciente tem um indicador de presença (1) ou ausência (0) para cada tipo de dado.
Exemplo de matriz de disponibilidade dos dados:

Paciente	Imagem (MRI)	EEG	Dados Tabulares
001	1	1	1
002	1	0	1
003	0	1	1
004	1	1	0
005	1	1	1

Com essa matriz, poderia ser calculado o percentual de pacientes com dados completos e o percentual de pacientes com apenas um ou dois tipos de dados disponíveis. Se **a uma quantidade considerável dos pacientes tiver dados completos**, um modelo multimodal pode ser viável. Caso contrário, pode ser avaliado outras possibilidades, como multimodal utilizando as duas modalidades com mais dados e um modelo para o outro tipo de dado, ou um emsemble considerando as 3 modalidades.

- **Pergunta 2.2:** Os dados coletados frequentemente apresentam desafios como ruído nos sinais EEG, variações nas resoluções das imagens médicas, problemas na coleta de dados tabulares por paciente, diferenças de escalas nos dados, entre outros. Em 2 décadas, aparelhos, medicamentos e processos e coleta mudam. Considerando esses fatores, descreva detalhadamente as principais etapas de pré-processamento que você realizaria em todos os tipos de dados antes da modelagem. Caso você considere executar alguma etapa importante antes disso, descreva-a considerando sua base de dados.

Resposta. Antes do pré-processamento, seria **essencial realizar primeiramente uma Análise Exploratória dos Dados (EDA)** para identificar problemas e entender a base. Somente após essa análise, seria possível decidir executar corretamente as etapas de pré-processamento específicas para cada tipo de dado. Essa etapa poderia seguir passos específicos a depender do tipo de dados:

- Sinais EEG: verificação visual através de plotagens, avaliações estatísticas (ex: médias, desvios-padrão, range de valores), análise de outliers, entre outros.
- Imagens Médicas: avaliação visual das imagens para detectar problemas como baixa qualidade, artefatos e inconsistências nas resoluções, verificação visual através de plotagens (histogramas), análise de presença de variação de escalas, entre outros.
- Dados tabulares: análise de valores faltantes, análise de outliers, análise de distribuição, colinearidade de variáveis, entre outros.

A seguir, segue alguns exemplos de pré-processamento que poderiam ser executados a depender da EDA realizada, por tipo de dado e de forma detalhada.

- Sinais EEG: aqui vou citar a escolha do método de normalização. Os sinais poderiam ser normalizados entre 0 e 1 ao se conhecer os ranges de valores (máximo e mínimo) por equipamento. Com uma base grande, isso seria perfeitamente possível.

- Imagens Médicas: aqui vou citar o redimensionamento das imagens caso houvesse mudança de escala nas imagens de um aparelho para outro. Normalmente o dado precisa estar na mesma escala para se trabalhar. Uma forma de aplicar esse redimensionamento seria redimensionar todas as imagens para a menor resolução encontrada na base.
- Dados tabulares: aqui vou citar a normalização das features numéricas somente na base de treino. Os fatores devem ser salvos e aplicados na validação e teste, tratando casos particulares (ex: usou min-max e no teste um valor passou do range). A escolha de qual técnica de normalização a ser aplicada dependeria EDA. Colunas com valores mínimos e máximo bem definidos se beneficiariam do mim-max scaler. Colunas com presença de outliers poderiam se beneficiar do robust-scaler.

- **Pergunta 2.3**: Hipoteticamente, o hospital sugeriu que você focasse primeiramente nos sinais EEG dos pacientes. Após preparar o dado, você verificou que possui dados de 50.000 pacientes ao longo de 5 anos, com dados coletados dos pacientes coletados em 1 dia, de 3 em 3 meses, regularmente. Os pacientes possuem rótulos de saudável ou não saudável a cada dia da realização do exame. Explique como você realizaria o processo de divisão dos dados considerando 2 cenários: (1) classificação em saudável ou não saudável e (2) previsão de saudável ou não saudável.

  Resposta.
  - Cenário 1. Classificação pontual (saudável/não saudável). Neste cenário, o objetivo é classificar cada exame individualmente sem levar em conta o histórico temporal explicitamente. Neste caso, os dados poderiam ser divididos por paciente, pois pacientes diferentes devem ser exclusivos de cada conjunto (evita vazamento de informação para validação e teste) garantindo que o modelo generalize corretamente para novos pacientes. Após a divisão por paciente, poderia ser aplicada a técnica hold-out de divisão de base usando a divisão estratificada, no qual poderiam ser utilizados os valores 80-10-10 para treino, validação e teste, respectivamente. A divisão estratificada garantiria um melhor equilíbrio entre as classes.
  - Cenário 2. Previsão ou acompanhamento longitudinal (saudável/não saudável). Se o objetivo for prever o estado futuro de saúde com base no histórico anterior (cenário realístico de acompanhamento da doença ao longo do tempo), realizaria a mesma divisão anterior somando 2 detalhes adicionais. Primeiro a divisão também deveria considerar o fator tempo, ou seja, exames anteriores (anos iniciais) no conjunto de treino e exames posteriores (anos mais recentes) no conjunto de teste. Isso evitaria vazamentos temporais, permitindo avaliar a capacidade real do modelo em prever o estado futuro dos pacientes. O outro detalhe seria o “janelamento” do dado, uma vez que em um problema de previsão é necessário analisar uma sequência de dados para inferir n passos a frente do dado analisado.

<u>Questão 3: Análise Crítica</u>

- **Pergunta 3.1:** Comente criticamente sobre a afirmação "temos dados suficientes" feita pelo hospital especialmente no contexto de uma análise longitudinal em contexto hospitalar, onde os dados são: sensíveis, sujeitos conhecimento de domínio específico, restrições éticas e de privacidade, múltiplas fontes de dados ao longo de anos e dados não-rotulados com informações clínicas dispersas em relatórios médicos. Após essa crítica, proponha um plano estruturado e detalhado para preparar esses dados para análise preditiva desde o entendendimento do dado, considerando os problemas relacionados a esse contexto.

  Resposta.
  Apesar da existência de centenas de pacientes a quantidade isolada não garante que esses dados sejam suficientes para uma análise longitudinal robusta, especialmente devido aos seguintes pontos críticos:
  - Dados Faltantes: Em contextos hospitalares, é comum haver períodos sem coleta, exames incompletos e perda de seguimento de pacientes.
  - Qualidade e Padronização: ao longo dos anos, equipamentos e protocolos clínicos evoluem, podendo levar à inconsistência nos dados.

- Dependência Temporal: Para uma análise longitudinal, é necessário garantir uma quantidade significativa de dados por paciente ao longo do tempo. Poucas medições ou intervalos muito espaçados podem ser insuficientes.
- Questões éticas e legais relacionadas ao **sigilo e privacidade** podem limitar o acesso integral às informações.

Um plano de ação estruturado poderia envolver:

- Tratamento Ético dos dados, entando atento a remoção de informações pessoais sensíveis para cumprir exigências legais, especialmente a LGPD. Contratos de sigilo e protocolos éticos também devem ser estabelecidos nessa etapa inicial.
- Consulta a Especialistas (Médicos e Técnicos): discutir com médicos e equipes clínicas sobre variáveis relevantes dos dados, como interpretar corretamente os dados e entender claramente os critérios de diagnóstico e progressão das doenças neurológicas.
- Análise exploratória detalhada (**EDA**) dos dados disponíveis, já discutido a sua importância e passos nas questões anteriores
- Padronização e Harmonização dos Dados através de integração e pré-processamento inicial dos dados a depender da EDA realizada. Sugestões de pré-processamento também já foram discutidos em questões anteriores.
- Divisão da base, onde a escolha da divisão se daria de acordo com os cenários descritos na questão 2.3.
- Realização de experimentos iniciais com modelos simplificados para avaliar se a qualidade e quantidade dos dados são suficientes para análises mais complexas

Por fim, faria uma documentação rigorosa de todas as etapas, decisões metodológicas e resultados obtidos, facilitando auditorias, reprodução e futuras validações clínicas e éticas.