FIOCRUZ

Concurso Público Fiocruz 2023

Pesquisador em Saúde Pública

Prova Discursiva

PE62 - Cientista de Dados

Questão 01

Prever o peso de recém-nascidos, ainda no período gestacional, é uma prática fundamental para a equipe médica, uma vez que o peso ao nascer é um indicador vital da saúde do recém-nascido e ainda é um preditor significativo de riscos para doenças crônicas e necessidade de cuidados especiais. Motivados por este cenário, um grupo de pesquisadores utilizou o modelo de árvores XGBoost para prever o peso ao nascer de bebês em função de variáveis como a idade gestacional; o sexo do bebê; o ganho de peso da mãe durante a gravidez; o estado emocional e estresse materno; entre outras.

Neste estudo foi utilizada uma amostra com 524 gestantes. A base de dados foi dividida de forma aleatória em treino e teste, sendo 75% da base para treino e os outros 25% para teste. Os pesquisadores ajustaram seis modelos XGBoost distintos, 3 deles com 100 árvores e 3 deles com 500 árvores. Além da diferença no número de árvores, os modelos também se diferem pelas variáveis independentes utilizadas. A Tabela 1 apresenta a descrição de cada um dos 6 modelos.

Tabela 1: Descrição dos modelos XGBoost de previsão do peso ao nascer.

Modelo	Nº de árvores	Variáveis independentes
1	100	todas (modelo completo)
2	100	todas, menos ganho de peso gestacional
3	100	todas, menos sexo do bebê
4	500	todas (modelo completo)
5	500	todas, menos ganho de peso gestacional
6	500	todas, menos sexo do bebê

O valor do R^2 foi utilizado como medida de comparação dos resultados. Tais valores, tanto para a base de treino quanto para a base de teste, estão descritos na Tabela 2. A Figura 1 apresenta gráficos de barras para os valores da Tabela 2.

Tabela 2: *Valores do R^2 para cada modelo ajustado nas bases de treino e teste.*

Modelo	base de treino	base de teste
1	0,899	0,852
2	0,758	0,705
3	0,895	0,841
4	0,955	0,830
5	0,850	0,699
6	0,949	0,802

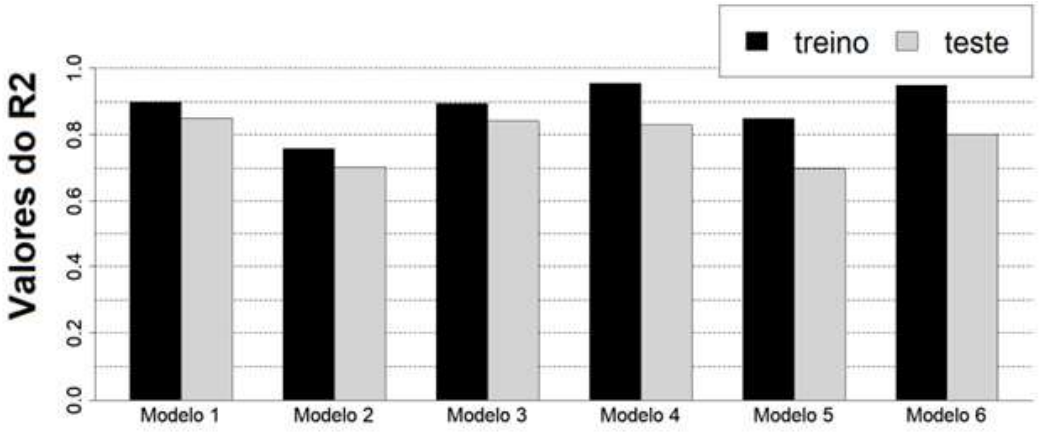

Figura 1: *Representação gráfica para os valores de R^2 para cada modelo, nas bases de treino e teste.*

De acordo com as informações acima responda aos itens a seguir. Desenvolva a resposta em um texto com o mínimo de 50 e o máximo de 150 linhas.

1) Discorra sobre a importância em dividir a base de dados em treino e teste.
2) Quais as conclusões da comparação entre os resultados dos modelos com mesma quantidade de árvores, porém variáveis diferentes? Isto é, comparar os Modelos 1, 2 e 3 entre si e também os Modelos 4, 5 e 6 entre si.
3) Quais as conclusões da comparação entre os resultados dos modelos com as mesmas variáveis, porém quantidade de árvores diferentes? Isto é, comparar o Modelo 1 com o Modelo 4, comparar o Modelo 2 com o Modelo 5 e comparar o Modelo 3 com o Modelo 6.
4) Se você fizesse parte da equipe de pesquisadores, qual dos modelos você recomendaria para realizar as previsões? Justifique.

Questão 02

O rastreamento do câncer de mama representa uma estratégia crucial para viabilizar a detecção precoce e assegurar uma resposta mais eficaz no tratamento. Um conjunto de pesquisadores dedica-se à elaboração de um modelo preditivo para o diagnóstico do câncer de mama, baseado em dados de fácil obtenção durante consultas de rotina e análises de exames de sangue. A amostra compreendeu 116 pacientes do sexo feminino, entre os quais 64 foram diagnosticadas com a patologia. As variáveis utilizadas no estudo estão detalhadas na Tabela 01.

Tabela 01 – Variáveis utilizadas para a construção do modelo preditivo.

Nome da variável	Descrição	Código/Categorias
Câncer	Possui câncer de mama?	0: Não 1: Sim
Idade	Faixa etária	0: Inferior a 40 anos 1: De 40 a 69 anos 2: 70 anos ou mais
IMC	Índice de Massa Corpórea	0: Peso normal 1: Acima do peso 2: Obeso
Insulina	Nível de insulina no sangue	-
Resistina	Nível de resistina no sangue	-

Os pesquisadores conduziram o ajuste de dois modelos de regressão logística múltipla, considerando diferentes conjuntos de variáveis independentes, conforme ilustrado na Tabela 02. Adicionalmente, foram fornecidos os valores do Critério de Informação de Akaike (AIC) e a Área Sob a Curva ROC (AUC) para cada um dos modelos.

Tabela 02 – Resultados dos dois modelos de regressão logística múltipla ajustados.

Variável		Modelo 1			Modelo 2		
		Coef	SE	p-valor	Coef	SE	p-valor
Intercepto		-9,123	2,120	<0,0001	-8,754	2,100	<0,0001
Idade	Inferior a 40 anos	-	-	-	-	-	-
	De 40 a 69 anos	1,675	0,765	0,029	1,975	0,665	0,003
	70 anos ou mais	0,163	0,993	0,869	0,184	0,893	0,837
IMC	Peso normal	-	-	-	-	-	-
	Acima do peso	-0,122	0,618	0,844	-	-	-
	Obeso	-0,456	0,453	0,314	-	-	-
Insulina		0,084	0,031	0,007	0,094	0,041	0,022
Resistina		0,049	0,021	0,020	0,069	0,023	0,002
AIC		119,26			106,15		
AUC		0,78			0,85		

De acordo com as informações fornecidas acima, responda aos itens abaixo. Desenvolva a resposta em um texto com o mínimo de 50 e o máximo de 150 linhas.

1) Discorra sobre o porquê de o modelo de regressão logística ser adequado para o problema. Apresente os motivos de não se utilizar o modelo de regressão normal e discuta sobre a interpretação dos parâmetros do modelo logístico e suas vantagens para o contexto apresentado.
2) Utilizando o Modelo 2, calcule e interprete a estimativa pontual para a razão de chances (também conhecida como odds ratio) das variáveis idade, considerando a categoria de 40 a 69 anos, e insulina no sangue.
3) Com base em um nível de significância de 5%, indique quais as variáveis independentes que são importantes para a predição de câncer de mama no Modelo 1. Justifique sua resposta.
4) Dentre os dois modelos apresentados, qual modelo você escolheria. Justifique sua resposta.

Rascunho da Questão 01

RASCUNHO

Rascunho da Questão 01

RASCUNHO

Rascunho da Questão 01

RASCUNHO

Rascunho da Questão 01

RASCUNHO

Rascunho da Questão 01

RASCUNHO

Rascunho da Questão 02

RASCUNHO

Rascunho da Questão 02

RASCUNHO

Rascunho da Questão 02

RASCUNHO

Rascunho da Questão 02

RASCUNHO

Rascunho da Questão 02

RASCUNHO

Instruções - Questões Discursivas

1. Cada questão discursiva deverá ter um Limite mínimo de 50 linhas e máximo de 150 linhas.

2. Transcreva sua resposta para a parte pautada no Caderno de Respostas. Não assine, rubrique ou coloque qualquer marca que o identifique, sob pena de ter sua prova anulada. Assim, a detecção de qualquer marca identificadora no espaço destinado à transcrição do texto definitivo acarretará nota ZERO na respectiva prova discursiva.

3. O tempo total de duração da prova será de 4 (quatro) horas, incluindo o tempo para o preenchimento da Resposta Definitiva da Questão Discursiva. Nenhum rascunho SERÁ LEVADO EM CONTA.

4. Verifique se a prova é para o **PERFIL** para o qual concorre.

5. Somente após autorizado o início da prova, verifique se este Caderno de Questões está completo e em ordem. **Folhear o Caderno de Questões antes do início da prova implica na eliminação do candidato.**

6. Verifique, no **Caderno de Respostas**, se seu nome, número de inscrição, identidade e data de nascimento estão corretos. Caso contrário, comunique ao fiscal de sala.

7. O rascunho do **Caderno de Questões** poderá ser utilizado para anotações, mas somente as respostas assinaladas no **Caderno de Respostas** serão objeto de correção.

8. Observe as seguintes recomendações relativas ao **Caderno de Respostas**:

- **.** não haverá substituição por erro do candidato;
- **.** não pode ser dobrado, amassado, rasurado, manchado ou conter qualquer registro fora dos locais destinados às respostas;

9. O fiscal não está autorizado a alterar quaisquer dessas instruções.

10. Você só poderá retirar-se da sala após 60 minutos do início da prova.

11. Quaisquer anotações só serão permitidas se feitas no **Caderno de Respostas**.

12. Os três últimos candidatos deverão permanecer na sala até que o último candidato entregue o **Caderno de Respostas**.

13. Ao terminar a prova, entregue ao fiscal de sala, **obrigatoriamente**, o **Caderno de Questões** e o **Caderno de Respostas**.