10.18605/2175-7275/cereus.v16n3p131-150

<< Recebido em: 30/04/2024 Aceito em: 21/08/2024. >>

ARTIGO ORIGINAL

REVISTA CEREUS ISSN: 2175-7275

Comparação de modelos de regressão logística na detecção precoce do risco de diabetes

Comparison of logistics regression models in early detection of diabetes risk

Camilla Ramos da Silva[1], Luiz Fernando Caldeira Ribeiro[2]

[1] MBA Data Science and Analytics. Universidade de São Paulo - USP Email:deborah.queiroz@hotmail.com
ORCID:https://orcid.org/0009-0008-5332-3918

[2] Doutor em Agronomia. Prof. Na Universidade do Estado de Mato Grosso - UNEMAT
ORCID:https://https://orcid.org/0000-0002-8637-6425

RESUMO

Diabetes é um problema de saúde pública mundial que afeta 463 milhões de pessoas e, segundo estudos, até 2045 haverá um incremento de 51% no número de indivíduos acometidos pela doença. A taxa de pessoas pré-diabéticas assintomáticas também é elevada, aproximadamente 84% dos casos, e essa particularidade impede que o doente receba os devidos cuidados e tratamentos antes que a enfermidade se agrave e leve à outras complicações ou até a morte. O diagnóstico precoce da doença se mostra benéfico diante deste cenário e a ciência de dados pode contribuir para que isso seja feito. O objetivo deste trabalho foi propor modelos de predição precoce de diabetes utilizando os métodos supervisionados de Regressão Logística Binária e Regressão Logística Binária Multinível, avaliando qual modelo apresenta resultados acurados. O trabalho foi baseado em estudos anteriores, com metodologias diferentes aplicadas sem a abordagem multinível. Respostas de um questionário aplicado à pacientes – diabéticos e saudáveis – do Sylhet Diabetes Hospital de Bangladesh foram utilizados nas modelagens, o qual continha perguntas relacionadas a sintomas geralmente associados ao diagnóstico do diabetes. Tal estudo possibilitou chegar à modelos com resultados satisfatórios, e evidenciou que a modelagem multinível apresenta melhores resultados se comparado a Regressão Logística convencional.

Palavras-chave: Ciência de dados; Aprendizado supervisionado; Modelagem hierárquica; Estruturas aninhadas.

ABSTRACT

Diabetes is a global public health issue that affects 463 million people, and studies predict a 51% increase in the number of individuals affected by the disease by 2045. The prevalence of asymptomatic prediabetes is also high, accounting for approximately 84% of cases, a factor that hinders timely care and treatment before the disease progresses to more severe complications or even death. Early diagnosis of diabetes is beneficial in this context, and data science can play a significant role in achieving this. The objective of this study was to propose early prediction models for diabetes using supervised methods, namely Binary Logistic Regression and Multilevel Binary Logistic Regression, and to evaluate which model yields more accurate results. The study builds on previous research that applied different methodologies without the multilevel approach. Data for the modeling were obtained from responses to a questionnaire administered to both diabetic and healthy patients at the Sylhet Diabetes Hospital in Bangladesh, containing questions related to symptoms commonly associated with diabetes diagnosis. This study led to the development of models with satisfactory results and demonstrated that the multilevel modeling approach produces better outcomes compared to conventional Logistic Regression.

Keywords: Data Science; Supervised Learning; Hierarchical Modeling; Nested Structures.

1. INTRODUÇÃO

O diabetes mellitus é um problema de saúde pública mundial que afeta a humanidade independente da condição socioeconômica e da localização geográfica, tornando-se uma das maiores síndromes crônicas (OLIVEIRA, 2009). A doença é causada pela produção insuficiente ou má absorção de insulina pelo organismo. A insulina é a responsável pela quebra das moléculas de glicose. Essa deficiência pode levar a aumento da taxa de glicemia no sangue e levar a complicações no coração, artérias, olhos, rins e nos nervos, podendo levar o indivíduo à óbito em casos mais graves (MINISTÉRIO DA SAÚDE, 2020).

De acordo com a 9ª edição do Atlas do Diabetes da International Diabetes Federation (IDF) de 2019, mundialmente 1 a cada 11 adultos entre 20 e 79 anos tem diabetes, totalizando aproximadamente 463 milhões de pessoas com a doença. Além disso, 1 a cada 2 adultos diabéticos ainda não foram diagnosticados, somando um total de 232 milhões de pessoas. Esses números acabam tendo também um impacto econômico, onde tem-se que aproximadamente 10% dos gastos globais com saúde são destinados à diabetes, computando um montante de 760 bilhões de dólares (INTERNACIONAL DIABETES FEDERATION, 2019). Tratando-se da América do Sul e América Central, o Brasil está em primeiro lugar no ranking de números de pessoas entre a fase adulta que possuem diabetes, totalizando 16,8 milhões de pessoas (INTERNACIONAL DIABETES FEDERATION, 2019).

Em 2018, os gastos diretos (hospitalizações, procedimentos ambulatoriais e medicamentos) com diabetes no Sistema Único de Saúde (SUS) alcançaram aproximadamente 1,03 bilhão de reais (NILSON et al.., 2020). O IDF estima que em 2045 teremos um incremento de 51% no número de diabéticos no mundo, sendo 33% na América do Norte e Caribe, 55% na América do Sul e Central, 143% na África, 96% no Oriente Médio e Norte da África, 74% no Sudeste Asiático, 31% no Pacífico Ocidental e 15% na Europa.

As consequências do diabetes no organismo do indivíduo são diversas, e a doença se destaca pela alta taxa de morbimortalidade (PACE, 2006). A partir do registro nacional de diabetes e hipertensão gerido pelo Ministério da Saúde, o SisHiperdia, analisando 1,6 milhões de registros tem-se que: 4,3% dos casos registrados tinham transtorno do pé diabético e 2,2% uma amputação prévia, 7,8% tinham doença renal crônica, 7,8% haviam

tido infarto do miocárdio e 8% haviam tido derrame (SCHMIDT, 2014). Além disso, a mortalidade de indivíduos com diabetes foi 57% maior do que na população em geral (SCHMIDT, 2014).

Há estimativas que indicam que 84% das pessoas com pré-diabetes, estágio onde o nível de açúcar no sangue é elevado e o indivíduo não apresenta sintomas, não sabem que possuem este problema. Atualmente os tratamentos de diabetes iniciam-se apenas quando o paciente já desenvolveu a doença ou até mesmo quando já possui complicações vasculares em decorrência da mesma.

Uma abordagem precoce ajudaria a reduzir o desenvolvimento do diabetes, seus sintomas, consequências e custos gerais de tratamento e cuidados com a saúde (ZANIELLI, 2019). Com o aumento da oferta de informações e com o auxílio da ciência de dados, uma possível detecção precoce de doenças se torna cada vez mais viável e proveitosa. A disponibilidade de tais dados em conjunto com dados e metodologias tradicionais, apresentam-se como uma ferramenta de grande potencial para uma o surgimento de novos sistemas de vigilância epidemiológica, sendo capaz de preencher lacunas existentes na infraestrutura de Saúde pública (SALATHÉ et al.., 2012). Devido à grande quantidade de dados pode tornar inviável a formulação de teorias e hipóteses sem o auxílio de métodos de mineração de dados (SHMUELI, 2010). Sendo assim, torna-se necessário expandir o campo da epidemiologia para novos campos, que possuam um pensamento transdisciplinar (DICLEMENTE et al., 2019).

A ciência de dados apresenta-se como uma ferramenta de grande potencial para a área de saúde pública e privada, através da transformação de massas de dados em recursos para a otimização econômica de sua operação (IBM, 2012). Utilizando-se de métodos de Machine Learning pode-se chegar a um modelo que ajude a fornecer dados e insumos para que a administração pública, instituições de saúde e a população possa lidar de maneira mais eficaz com esse problema que afeta todo o mundo, sendo possível pensar em iniciativas que consigam amenizar os efeitos negativos dessa doença na qualidade de vida dos indivíduos e otimizar gastos em saúde, focando em prevenção e tratamento precoce.

Na medicina diagnóstica, a ciência de dados pode ser utilizada como um sistema de apoio à decisão, e tem ganhado adeptos e angariado esforços em alguns locais do mundo. São exemplos o sistema Deep Mind e o supercomputador Watson (LOBO, 2017). O sistema inglês Deep Mind, adquirido pela Google, em 2017 já processava 1,6 milhão

de prontuários médicos analisando dados dos pacientes, gerando alertas sobre os mesmos e auxiliando na indicação de medicamentos nos hospitais do Serviço Nacional de Saúde da Inglaterra (NHS). Já a International Business Machines Corporation (IBM), proprietária do supercomputador Watson, assimilou diversos livros de medicina e bibliotecas online de medicina e biomedicina, além de milhares de casos clínicos com a assistência de 14.770 médicos, para buscar melhorar a exatidão de diagnósticos médicos. Este supercomputador continua em um sistema de aprendizado e está mais "inteligente" a cada ano (LOBO, 2017).

A partir da Regressão Logística e da Regressão Logística Multinível, objetiva-se estimar o melhor modelo para a previsão de diabetes em estágio inicial baseado em sintomas geralmente associados à doença, possibilitando a determinação de qual a probabilidade do indivíduo que possui aqueles atributos se tornar diabético no futuro, e espera-se encontrar uma performance superior para a modelagem multinível..

2. MATERIAIS E MÉTODOS

Foram utilizados dados do estudo "Likelihood Prediction of Diabetes at Early Using Data Mining Techniques" (ISLAM et al., 2020). O estudo utilizou a base de dados do Sylhet Diabetes Hospital de Bangladesh.

Foram totalizadas 520 observações de indivíduos com ou sem diagnóstico de diabetes. As variáveis analisadas foram: idade; sexo (feminino ou masculino); poliúria (vontade frequente de urinar – sim ou não); Polidipsia (sede excessiva persistente – sim ou não); perda repentina de peso (sim ou não); fraqueza (sim ou não), polifagia (fome excessiva – sim ou não); infecções genitais (sim ou não); embaçamento visual (sim ou não); coceira (sim ou não); irritabilidade (sim ou não); má cicatrização (sim ou não); paresia muscular (sim ou não); rigidez muscular (sim ou não); alopecia (sim ou não); obesidade (sim ou não); classe (se o indivíduo é ou não diabético – sim ou não) (Islam et al., 2020).

O estudo original de Islan et al. (2020) utilizou as metodologias Naive Bayes, J48 Decision Tree, Regressão Logística e Randon Forest. Neste trabalho, foi utilizado o modelo de Regressão Logística Binária e implementou-se um modelo de Regressão Logística Binária com a abordagem multinível, assumindo o constructo de que a relação indivíduo - sexo e indivíduo - faixa etária apresentam estruturas aninhadas de dados, e verificar o desempenho de cada um dos modelos. Posteriormente tal constructo foi avaliado pela correlação intra-classe.

Na estrutura de dados aninhados tem-se que variáveis apresentam valores que se modificam entre indivíduos em determinado nível (grupo) e que se mantém constantes para outro nível, sendo que esses grupos representam um nível superior (FÁVERO; BELFIORE, 2021).

A Regressão Logística Binária tem o objetivo de estudar a probabilidade de ocorrência de um evento, que é definido de maneira dicotômica, sendo Y = 1 para descrever do evento de interesse e Y =0 para descrever a não ocorrência do evento de interesse (FÁVERO; BELFIORE, 2021). Já a Regressão Logística Multinível considera a existência de estruturas aninhadas de dados, levando em consideração e permitindo que sejam identificadas e analisadas a heterogeneidade individual de cada observação e entre os grupos a quais esses indivíduos pertencem (FÁVERO; BELFIORE, 2021). Assim, tem-se uma regressão com componentes de efeitos fixos e com componentes de efeitos aleatórios. Os componentes fixos indicam a dimensão das associações entre as variáveis, e os componentes aleatórios indicam o impacto dos grupos e as variâncias nos diferentes níveis considerados no modelo (MERLO, 2003).

Os modelos preditivos foram desenvolvidos a partir da linguagem R. O software utilizado na execução dos comandos foi o R Studio. Durante a modelagem os atributos presentes na base de dados foram testados estatisticamente a 95% de significância. Para a avaliação da qualidade do ajuste dos modelos e possível comparação entre eles, foram utilizados atributos como: logaritmo da função de máxima verossimilhança (LogLik), eficiência global do modelo, a especificidade e a sensitividade, além da área abaixo da Curva ROC (Receiver Operating Characteristic).

O critério da máxima verossimilhança é utilizado para estimação dos valores dos coeficientes de cada parâmetro dos modelos (GONZALEZ, 2018). Para o cálculo da eficiência global do modelo, a especificidade e a sensitividade, faz-se necessária a determinação de um ponto de corte, ou também chamado de cutoff, para o qual o valor escolhido foi de 0,5. Para probabilidades abaixo deste valor, o evento estudado é considerado como se não houvesse ocorrido, e caso a probabilidade seja maior ou igual ao ponto de corte o evento é classificado como se houvesse ocorrido.

A partir de tal classificação em evento e não evento foi possível estruturar a tabela de classificação de eventos, conhecida como matriz de confusão. Nela foram plotados os volumes de incidência real de evento e não evento versus as classificações obtidas pelo modelo como evento ou não evento.

A curva ROC, também conhecida como Curva de Sensibilidade, é um gráfico que apresenta a variação da sensitividade do modelo em função do complemento de sua especificidade, ou seja, (1-especificidade). Assim, conforme maior for a área abaixo da curva ROC, maior a eficiência global de previsão do modelo (FÁVERO; BELFIORE, 2021).

Antes da modelagem a base de dados passou por uma preparação, dado que os atributos presentes no dataset não se apresentavam de forma binária, então esse processo de “dummização” das variáveis se fez necessário, além da criação das classes das faixas etárias.

Houve também a divisão da base em treino (70% do total das observações) e teste (80% do total das observações) com o auxílio da função createDataPartition() do pacote caret, fixando-se a semente. Esse procedimento foi adotado para a que a presença de um efeito conhecido como overfitting seja avaliado posteriormente. Tal fenômeno faz com que o modelo se adeque perfeitamente ao conjunto de dados de treinamento, mas que não funcione para as observações que não estavam presentes no seu treinamento, isto é, nos dados de teste, fazendo com que a generalização do modelo não seja possível de maneira correta (YING, 2019).

A Regressão Logística Binária pôde ser aplicada através da função glm do pacote stats, considerando a distribuição de probabilidade da variável resposta do modelo como binomial, e com o auxílio da função step do mesmo pacote para remoção das variáveis que não se mostraram estatisticamente significantes a 95% de significância (ou seja, p-value < 0,05).

Para a Regressão Logística Multinível, dois tipos de aninhamentos de dados foram testados: 1) Nível 1: indivíduo; nível 2: faixa etária E nível 2: sexo. Assim como citado anteriormente, a fonte de dados foi preparada com a criação dos atributos que serão utilizados como níveis, transformando-os no tipo factor, e com a “dummização” dos demais atributos.

Após esses procedimentos iniciou-se a modelagem através do pacote glmmTMB, a partir da função glmmTMB(), definindo sua família como binomial. Este pacote não possui a capacidade de remover as variáveis que não se mostram estatisticamente significantes à 95%, nem a função step() é compatível com este pacote. Sendo assim, a seleção das variáveis relevantes ao modelo foi realizada manualmente através da análise de seus respectivos p-values (p-value < 0,05).

3. RESULTADOS

Regressão Logística Binária

Para a Regressão Logística Binária, os seguintes atributos foram obtidos como resposta:

Tabela 1. Atributos obtidos como resultado da Regressão Logística

Atributo	Estimativa	Erro padrão	Valor z	p
(Intercepto)	1,2473	0,5480	2,2760	0,0228
Sexo Masculino	-5,0362	0,7878	-6,3924	0,0000
Poliúria	4,1032	0,6953	5,9010	0,0000
Polidipsia	5,4523	0,9318	5,8513	0,0000
Infecções genitais	1,1256	0,5671	1,9848	0,0472
Coceira	-2,9914	0,6140	-4,8718	0,0000
Irritabilidade	3,0889	0,6738	4,5841	0,0000
Paresia parcial	1,1904	0,5236	2,2736	0,0230
Faixa etária 36 a 45 anos	1,5807	0,5946	2,6586	0,0078

Fonte: Dados originais da pesquisa

O valor do LogLik obtido foi de -65,3. Com essas estimativas dos coeficientes, a equação final do modelo de Regressão Logística Binária, que calcula a probabilidade estimada da ocorrência do evento positivo, ou seja, positivo para diabetes, ficará conforme abaixo:

$$P_i = \frac{e^{z_i}}{1+e^{z_i}} = \frac{1}{1+e^{-z_i}} \qquad (1)$$

Sendo o logito Z_i igual a:

$$Z_i = 1{,}2473 - 5{,}0362.Sexo_{Masculino} + 4{,}1032.Poliúria + 5{,}4523.Polidipsia + 1{,}1256.InfecçõesGenitais - 2{,}9914.Coceira + 3{,}0889.Irritabilidade + 1{,}1904.ParesiaParcial + 1{,}5807.FaixaEtária_{36a45} \quad (2)$$

Tabela 2. Matriz de confusão para a regressão logística aplicados às bases de treino e teste

	Base de treino		Base de teste	
	Evento	**Não evento**	**Evento**	**Não evento**
Classificado como evento	213	16	89	21
Classificado como não evento	13	122	4	41

Fonte: Resultados originais da pesquisa

Tabela 3. Estatísticas de avaliação do modelo de regressão logística aplicados às bases de treino e teste

Estatísticas	**Base de treino**	**Base de teste**
Acurácia(eficiência global)	0,9203	0,8387
Intervalo de confiança (95% CI)	(0,8876 ; 0,9460)	(0,7712; 0,8928)
Sensitividade	0,9425	0,9570
Especificidade	0,8841	0,6613

Fonte: Resultados originais da pesquisa

Avaliando a área sob a curva ROC do presente trabalho, foi obtido o valor de 97,7% para o treino e 93,4 % para o teste.

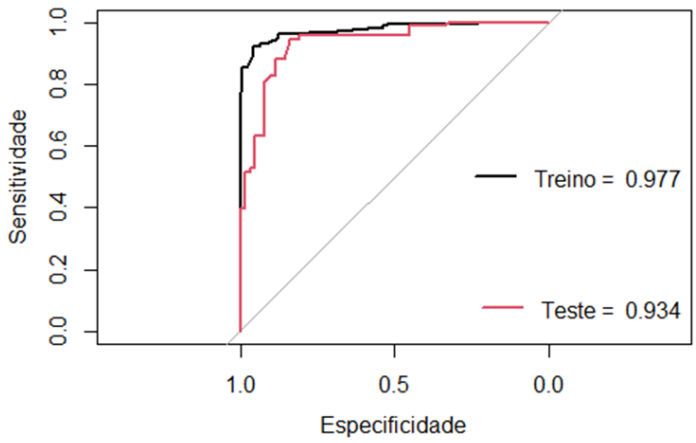

Figura 1. Área sob a curva ROC para a regressão logística
Fonte: Resultados originais da pesquisa

Regressão Logística Binária Multinível

Nível 1: indivíduo; Nível 2: faixa etária

Os seguintes atributos foram obtidos como resultado do modelo com interceptos aleatórios, considerando os indivíduos no nível 1 e o atributo faixa etária no nível 2:

Tabela 4. Atributos dos componentes de efeitos fixos obtidos como resultado da Regressão Logística

Efeitos Fixos	Estimativa	Erro padrão	Valor z	p
(Intercepto)	1,7918	0,6976	2,569	$1{,}10.10^{-2}$
Poliúria	4,3346	0,6941	6,245	$4{,}24.10^{-10}$
Polidipsia	5,0750	0,8679	5,848	$4{,}98.10^{-9}$
Coceira	-2,7604	0,6421	-4,299	$1{,}72.10^{-6}$
Irritabilidade	3,0337	0,6776	4,477	$7{,}56.10^{-6}$
Sexo Masculino	-4,6172	0,7405	-6,235	$4{,}52.10^{-10}$

Fonte: Dados originais da pesquisa

Tabela 5. Atributos do componente de efeitos aleatórios obtidos como resultado da Regressão

Efeitos Aleatórios	Variância	Desvio padrão
Faixa Etária (Intercepto)	0,7326	0,8559

Fonte: Dados originais da pesquisa

Os interceptos aleatórios u_0j para obtidos para cada uma das faixas etárias são exibidos abaixo:

Tabela 6. Interceptos aleatórios para cada uma das faixas etárias definidas

Faixa Etária	Interceptos Aleatórios u_{0j}
20 a 35 anos	-0,4635937
36 a 45 anos	1,0014989
46 a 55 anos	0,3359659
56 a 65 anos	-0,2731166
Acima de 65 anos	-0,6007545

Fonte: Dados originais da pesquisa

O valor do LogLik obtido foi de -70,4. A equação do modelo será:

$$P_{ij} = \frac{1}{1+e^{-z_{ij}}} \quad (3)$$

$$Z_{ij} = \gamma_{00} + \gamma_{10}.Poliúria + \gamma_{20}.Polidpsia + \gamma_{30}.Coceira + \gamma_{40}.Irritabilidade + \gamma_{60}.Sexo_{Masculino} + u_{0j} \quad (4)$$

Substituindo os valores dos coeficientes no logito, temos a seguinte equação final:

$$Z_{ij} = 1{,}7918 + 4{,}3346.Poliúria + 5{,}075.Polidpsia - 2{,}7604.Coceira + 3{,}0337.Irritabilidade - 4{,}6172 Sexo_{Masculino} + u_{0j} \quad (5)$$

Analisando o resultado da matriz de confusão, para um cutoff de 0,5, foi obtido como resultado a eficiência global de 91,76%, sensitividade de 92,04% e especificidade de 91,30%. O valor de LogLik na modelagem foi de -70,4.

Na generalização com o auxílio da base teste, mantendo o cutoff de 0,5, obteve-se uma eficiência global de 92,90%, sensitividade de 95,70% e especificidade de 88,71%.

Tabela 7. Matriz de confusão para a regressão logística multinível (indivíduo – faixa etária) aplicados às bases de treino e teste

	Base de treino		**Base de teste**	
	Evento	**Não evento**	**Evento**	**Não evento**
Classificado como evento	208	12	89	7
Classificado como não evento	18	126	4	55

Fonte: Resultados originais da pesquisa

Tabela 8. Estatísticas de avaliação do modelo de regressão logística multinível (indivíduo – faixa etária) aplicados às bases de treino e teste

Estatísticas	**Base de treino**	**Base de teste**
Acurácia (eficiência global)	0,9176	0,9290
Intervalo de confiança (95% CI)	(0,8844; 0,9437)	(0,8766; 0,9640)
Sensitividade	0,9204	0,9570
Especificidade	0,9130	0,8871

Fonte: Resultados originais da pesquisa

Sobre a área sob a curva ROC, foram obtidos os valores de 97,7% para a base treino e 98,1% para a base teste.

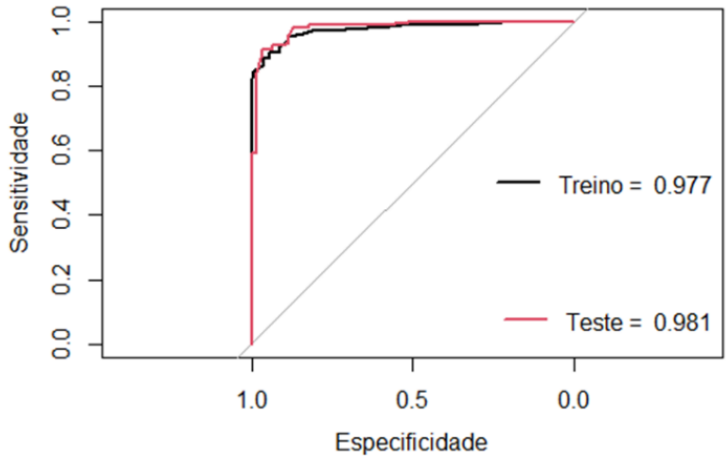

Figura 2. Área sob a curva ROC para a regressão logística multinível (indivíduo – faixa etária)
Fonte: Resultados originais da pesquisa

Nível 1: indivíduo; Nível 2: sexo

Os seguintes atributos foram obtidos como resultado do modelo com interceptos aleatórios, considerando os indivíduos no nível 1 e o atributo sexo no nível 2:

Tabela 9. Atributos dos componentes de efeitos fixos obtidos como resultado da Regressão Logística Multinível (indivíduo – sexo)

Efeitos Fixos	Estimativa	Erro padrão	Valor z	p
(Intercepto)	-0,8104	2,6386	-0,307	0,75874
Poliúria	4,2901	0,6580	6,520	$7{,}02.10^{-11}$
Polidipsia	5,2438	0,8764	5,983	$2{,}19.10^{-9}$
Coceira	-2,7135	0,5877	-4,617	$3{,}89.10^{-6}$
Irritabilidade	3,1024	0,6702	4,629	$3{,}67.10^{-6}$
Faixa Etária (36 a 45)	1,5363	0,5507	2,790	0,00527

Fonte: Dados originais da pesquisa

Tabela 10. Atributos do componente de efeitos aleatórios obtidos como resultado da Regressão Logística Multinível (indivíduo – sexo)

Efeitos Aleatórios	Variância	Desvio padrão
Sexo (Intercepto)	13,61	3,689

Fonte: Dados originais da pesquisa

Os seguintes interceptos aleatórios u_0j foram obtidos no modelo, para os sexos masculino e feminino:

Tabela 11. Interceptos aleatórios para o sexo do indivíduo

Sexo	**Intercepto aleatório u_{0j}**
Feminino	2,43606
Masculino	-2,43606

Fonte: Dados originais da pesquisa

O valor do LogLik obtido para o modelo foi de -70,9. A equação do modelo é exibida abaixo:

$$P_{ij} = \frac{1}{1 + e^{-Z_{ij}}} \quad (6)$$

$$Z_{ij} = \gamma_{00} + \gamma_{10}.Poliúria + \gamma_{20}.Polidpsia + \gamma_{30}.Coceira + \gamma_{40}.Irritabilidade + \gamma_{60}.FaixaEtária_{36a45} + u_{0j} \quad (7)$$

Substituindo os valores dos coeficientes no logito, temos a seguinte equação final:

$$Z_{ij} = -0{,}8104 + 4{,}2901.Poliúria + 5{,}2438.Polidpsia - 2{,}7135.Coceira + 3{,}1024.Irritabilidade + 1{,}5363.FaixaEtária_{36a45} + u_{0j} \quad (8)$$

Analisando o resultado da matriz de confusão para a base treino, para cutoff de 0,5, obteve-se como resultado a eficiência global de 92,31%, sensitividade de 92,04% e especificidade de 92,75%. O valor de LogLik da modelagem foi de -70,9.

Para a base de teste, obteve-se uma eficiência global de 94,19%, sensitividade de 95,70% e especificidade de 91,94%.

Tabela 12. Matriz de confusão para a regressão logística multinível (indivíduo – sexo) aplicados às bases de treino e teste

	Base de treino		**Base de teste**	
	Evento	**Não evento**	**Evento**	**Não evento**
Classificado como evento	208	10	89	5
Classificado como não evento	18	128	4	57

Fonte: Resultados originais da pesquisa

Tabela 13. Estatísticas de avaliação do modelo de regressão logística multinível (indivíduo – sexo) aplicados às bases de treino e teste

Estatísticas	Base de treino	Base de teste
Acurácia (eficiência global)	0,9231	0,9419
Intervalo de confiança (95% CI)	(0,8907 ; 0,9484)	(0,8926; 0,9731)
Sensitividade	0,9204	0,9570
Especificidade	0,9275	0,9194

Fonte: Resultados originais da pesquisa

Avaliando a área sob a curva ROC, tem-se 97,5% para a base treino e 97,9% para a base teste.

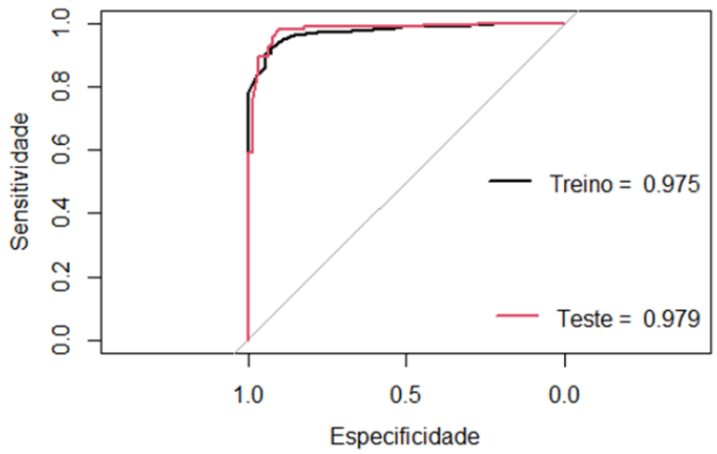

Figura 3. Área sob a curva ROC para a regressão logística multinível (indivíduo – sexo)
Fonte: Resultados originais da pesquisa

Avaliação e comparação dos modelos

O primeiro item a ser analisado é a necessidade de uma abordagem multinível para tal problema. A abordagem não deve ser apenas teoricamente defendida, mas sim baseada em coeficientes que demonstrem as diferenças existentes entre os níveis utilizados na modelagem e essa avaliação pode ser realizada através da correlação intra-classe (KREFT; LEEUW, 1998), que nos indica a proporção da variância total dos termos de erro que é devido ao segundo nível e que varia entre 0 e 1. Se essa correlação intra-classe for igual a zero ou muito próxima de zero, a utilização da modelagem multinível não se justifica.

Para a regressão logística multinível a correlação intra-classe rho pode ser calculada da seguinte forma, onde τ_{00} é a variância no segundo nível (FÁVERO; BELFIORE, 2021):

$$rho = \frac{\tau_{00}}{\tau_{00} + \frac{\pi^2}{3}} \quad (9)$$

Calculando rho para o modelo hierárquico indivíduo – faixa etária temos que rho=0,18, ou seja, indica que aproximadamente 18% da variância total dos termos de erro é devido a alteração do comportamento da variável dependente entre as faixas etárias; para o modelo hierárquico indivíduo – sexo temos que rho=0,80, ou seja, 80% da variância total dos termos de erro é devido a alteração do comportamento da variável dependente entre os sexos.

Dando continuidade às avaliações, para facilitar a visualização dos resultados dos modelos e sua comparação, o valor de LogLik, área sob a curva ROC, acurácia, sensitividade e especificidade encontram-se consolidados nos quadros abaixo, tanto para o treino do modelo, quanto para a sua generalização com a base de teste.

Tabela 14. Consolidação dos resultados de treino para todos os modelos (aplicação à base de treino)

Modelos	LogLik	Área ROC	Acurácia	Sensitividade	Especificidade
Logística Binária	-65,3	0,977	0,9203	0,9425	0,8841
Logística Multinível (indivíduo-faixa etária)	-70,4	0,977	0,9176	0,9204	0,9130
Logística Multinível (indivíduo – sexo)	-70,9	0,975	0,9231	0,9204	0,9275

Fonte: Resultados originais da pesquisa

Para o treino do modelo todos os modelos os valores de LogLik obtidos foram muito próximos, tanto para as regressões logísticas hierárquicas quanto para a regressão logística convencional. Como a estimação dos critérios da equação do modelo pela máxima verossimilhança sempre busca a maximização do LogLik (GONZALEZ, 2018), temos que a regressão logística convencional apresentou um resultado um pouco superior às modelagens hierárquicas, mas ainda assim pouco expressivo.

Para a generalização do modelo, aplicando-o a base de teste, os resultados consolidados são apresentados abaixo:

Tabela 15. Consolidação dos resultados da generalização para todos os modelos (aplicação à base de teste)

Modelos	Área ROC	Acurácia	Sensitividade	Especificidade
Logística Binária	0,934	0,8387	0,9570	0,6613
Logística Multinível (indivíduo – faixa etária)	0,981	0,9290	0,9570	0,8871
Logística Multinível (indivíduo – sexo)	0,979	0,9419	0,9570	0,9194

Fonte: Resultados originais da pesquisa

4. DISCUSSÃO

O estudo de Islam *et al.* (2020) que foi utilizado como inspiração para o presente trabalho, obteve o resultado final de acurácia para a regressão logística, ou seja, percentual de acerto de classificação do modelo, de 91,0%. A sensitividade obtida, percentual de acerto de classificação de observações que de fato são eventos, foi de 94,7%. A especificidade, percentual de acerto de observações que não são eventos, foi de 86,0%. Vale ressaltar que o mesmo não deixa claro qual foi a seleção final das variáveis após a avaliação da significância estatística de dos atributos, nem o cutoff utilizado para chegar a tal resultado.

Ao aplicar a equação obtida na modelagem deste trabalho à base de treino e analisar os outputs do modelo através da matriz de confusão, para um cutoff de 0,5, obteve-se como resultado a eficiência global de 92,03%. Avaliando a sensitividade obteve-se o valor de 94,25%. Já a especificidade foi de 88,41%. O valor de LogLik obtido na modelagem foi de -65,3.

Para avaliar a generalização do modelo, o mesmo foi aplicado à base de teste. Mantendo o cutoff de 0,5, obteve-se uma eficiência global de 83,37%, sensitividade de 95,7% e especificidade de 66,13%.

Avaliando também o resultado da acurácia em comparação com outros autores, podemos utilizar como referência o estudo de Rami (2020), no qual a regressão logística também foi empregada para a predição de Diabetes. No citado estudo uma base de dados com informações sobre 2000 pessoas foi utilizada, a qual contém variáveis como número de gestações, nível de glicose, pressão sanguínea, idade, etc. Tanto no treino do modelo quanto no teste, a acurácia obtida no estudo foi de 78%.

Outro estudo que pode ser comparado é o de Meng *et al.* (2013), que ao aplicar a regressão logística a uma base com dados de 1487 indivíduos, sendo eles 735 diabéticos e 752 saudáveis, e contendo variáveis como idade, histórico familiar de diabetes, nível de escolaridade, estresse no trabalho, duração do sono, consumo de café, sexo, etc, obteve para o treino do modelo acurácia, sensitividade e especificidade de 75,95%, 79,68% e 72,40%, respectivamente. Para o teste do modelo obteve acurácia, sensitividade e especificidade de 76,54%, 79,40% e 73,54%, respectivamente.

A melhoria nos parâmetros de avaliação era esperada dado que ele considera as estruturas de agrupamento ou hierarquia. Quando os dados estão estruturados de maneira hierárquica as observações de um dado nível pertencentes a um nível mais alto são raramente independentes, e isso acontece porque geralmente essas unidades possuem características semelhantes entre si, ou pertencem a um mesmo ambiente (TAMURA, 2007).

Analisando os resultados de acurácia, sensitividade e especificidade obtidos de maneira geral, tanto para o treino quanto para o teste, nota-se valores elevados para todos os tipos de modelagem. Isso indica que os modelos possuem um ótimo ajuste à sua base de treino, mas também que os mesmos também possuem um ótimo resultado de generalização (YING, 2019). Ou seja, os modelos também conseguiram se ajustar a observações que não estavam presentes na sua amostra de treinamento, o que indica que a modelagem foi bem-sucedida sem a ocorrência de overfitting.

Pelas matrizes de confusão reafirma-se que houve um bom ajuste dos modelos, dado que temos um grande acerto nas classificações de eventos que são efetivamente eventos e nas incidências de não eventos classificados pelos modelos como não eventos, com um melhor desempenho nos modelos multiníveis. Silva (2021) testou quatro modelos preditivos, sendo o que apresentou melhor acurácia no treinamento foi o svmRadialSigma com 76,48% de acurácia, em segundo o Naive Bayes com 76,23% de acurácia, em terceiro o segundo modelo KNN treinado com 73,54% de acurácia e com a pior acurácia o primeiro modelo KNN treinado com 71,94% de acurácia.

Silveira *et al.* (2021) para análise dos dados dos fatores de risco associados à hipertensão arterial aplicando a regressão lógica obteve resultados semelhantes ao do trabalho. Em seu trabalho, a acurácia do modelo foi de 81,7% e a área sob a curva (AUC) é de 82,5%, valor que, segundo Hosmer e Lemeshow (2013), indica uma excelente capacidade preditiva do modelo. Nota-se que, apesar os valores de LogLik serem muito

próximos, com uma pequena vantagem para a regressão logística convencional, os valores gerais da área sob a curva ROC (que relacionam sensitividade e especificidade) e a acurácia são melhores para as modelagens multiníveis.

5. CONSIDERAÇÕES FINAIS

Este estudo alcançou seu objetivo ao identificar o modelo mais eficaz na previsão de diabetes com base em sintomas, destacando a superioridade da regressão logística multinível sobre a regressão logística convencional. A próxima etapa envolveria a análise de conjuntos de dados mais diversificados, incorporando participantes de diferentes países, para examinar a influência de variáveis como hábitos alimentares e estilos de vida no desenvolvimento da diabetes.

REFERÊNCIAS

DICLEMENTE, R.; NOWARA, A.; SHELTON, R.; WINGOOD, G. Need for Innovation in Public Health Research. **American journal of public health**, 109(S2), S117–S120, 2019.Acesso em: 23 abr. 2022. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6383977/>

FÁVERO, L. P.; BELFIORE, P. **Manual de Análise de Dados: Estatística e Modelagem Multivariada com Excel**, SPSS e Stata. 1ed. LTC, Rio de Janeiro, 2021. 1216 p.

GONZALEZ, L. A. **Regressão Logística e suas Aplicações**. Monografia de bacharel em Ciência da Computação. Centro de Ciências Exatas e Tecnológicas, Universidade Federal do Maranhão, São Luís, Maranhão, 2018, 46 f. Acesso em: 14 mar. 2022. Disponível em: <https://monografias.ufma.br/jspui/bitstream/123456789/3572/1/LEANDRO-GONZALEZ.pdf>

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. 2 ed. Wiley, New York, EUA, 2000, 510 p.

INTERNATIONAL BUSINESS MACHINES CORPORATION [IBM]. **The value of analytics in healthcare**. 2012. Acesso em: 04 mai. 2022. Disponível em: <https://www.ibm.com/downloads/cas/NJA9K0DV>

ISLAM, M.M. F.; FERDOUSI, R.; RAHMAN, S.; BUSHRA, H. Likelihood prediction of diabetes at early stage using data mining techniques. In: **Computer Vision and Machine Intelligence in Medical Image Analysis**. Springer, Singapore. 2020, p.113-125

INTERNATIONAL DIABETES FEDERATION - IDF. **IDF Diabetes Atlas**, 9th ed. 2019. Acesso em: 24 out. 2021. Disponível em: <http://www.diabetesatlas.org>

KREFT, I .G. G.; LEEUW, J. **Introducing multilevel modeling**. Sage Publications, Londres, 1998, 160 p.

LOBO, L. C. Inteligência Artificial e Medicina. **Revista Brasileira de Educação Médica**, v. 41, n. 2, p. 185-193, 2017. Acesso em: 17 nov. 2021. Disponível em: <https://doi.org/10.1590/1981-52712015v41n2esp>

MENG, X.H.; HUANG, Y.X.; RAO, D.P.; ZHANG, Q.; LIU, Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. **The Kaohsiung Journal of Medical Sciences.** v. 9, n. 2, p. 93-99, 2013. Acesso em: 12 mai. 2022. Disponível em: < https://www.sciencedirect.com/science/article/pii/S1607551X12002173>

MERLO, J. Multilevel analytical approaches in social epidemiology: Measures of health variation compared with traditional measures of association. **Journal of Epidemiology and Community Health. v.** 57, n. 8, p. 550-552, 2003. Acesso em: 20 mar. 2022. Disponível em: <https://www.researchgate.net/publication/10642881_Multilevel_analytical_approaches_in_social_epidemiology_Measures_of_health_variation_compared_with_traditional_measures_of_association>

MINISTÉRIO DA SAÚDE. **Saúde de A a Z: Diabetes (diabetes mellitus)**. 2020. Acesso em: 23 out. 2021. Disponível em: < https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/diabetes-diabetes-mellitus)>.

NILSON, E. A. F., ANDRADE, R. C. S; BRITO, D. A.; OLIVEIRA, M. L. Custos atribuíveis a obesidade, hipertensão e diabetes no Sistema Único de Saúde, Brasil, 2018. **Revista Panamericana de Salúd Publica**. v. 44, n. 32, 2020. Acesso em: 23 out. 2021. Disponível em: <https://doi.org/10.26633/RPSP.2020.32>

OLIVEIRA, A. F.; VALENTE, J. G.; LEITE, I. C.; SCHRAMM, J. M. A.; AZEVEDO, J. M. A.; GADELHA, A. M. J. Global burden of disease attributable to diabetes mellitus in Brazil. **Cadernos de Saúde Pública.** v. 25, n. 6, p. 1234-1244, 2009. Acesso em: 25 out.2021. Disponível em: https://doi.org/10.1590/S0102-311X2009000600006

SCHIMIDT, M. I. DUNCAN, B. B. D.; SILVA, G. A.; MENEZES, A. M.; MONTEIRO, C. A. BARRETO, S. M. CHR, D.; MENEZES, P. R. Doenças crônicas não transmissíveis no Brasil: carga e desafios atuais. **Saúde no Brasil.** v. 4, p. 61-74, 2011. Acesso em: 25 out. 2021. Disponível em: <http://www.abc.org.br/IMG/pdf/doc-574.pdf>

TAMURA, K. A. **Modelo Logístico Multinível: um enfoque em métodos de estimação e predição**. Dissertação (Mestrado em Ciências). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2007. Acesso em: 28 mar. 2022. Disponível em: <https://www.teses.usp.br/teses/disponiveis/45/45133/tde-01072007-080446/publico/KarinAyumiTamura.pdf>

PACE, A. E.; OCHOA-VIGO, K.; CALIRI, M. H. L.; FERNANDES, A. P. M. Knowledge on diabetes mellitus in the self care process. **Revista Latino-Americana de Enfermagem**. v. 14, n. 05, p. 728-734, 2016. Acesso em: 25 out. 2021. Disponível em: <https://doi.org/10.1590/S0104-11692006000500014>

RANI, K.J. Diabetes Prediction using Machine Learning Techniques. International **Journal of Scientific Research in Computer Science, Engineering and Information Technology**. v. 6, n. 4, p. 294-305, 2020. Acesso em: 08 mai. 2020. Disponível em: <https://www.researchgate.net/publication/347091823_Diabetes_Prediction_Using_Machine_Learning>

SALATHÉ, M.; BENGTSSON, L.; BODNAR, T. J.; CERVEJEIRO, D. D.; BROWNSTEIN, J. S.; BUCKEE, C.; CAMPBELL, E. M.; CATTUTO, C.; KHANDELWAL, S.; MABRY, P. L.; VESPIGNANI, A. Digital Epidemiology. **PLoS Computational Biology**. v. 8, n. 7, p. e1002616, 2012. Acesso em: 23 abr. 2022. Disponível em: <https://doi.org/10.1371/journal.pcbi.1002616>

SHMUELI, G. To Explain or to Predict? **Statistical Science**. **v.** 25, n. 3, p. 289 – 310, 2010. Acesso em 23 abr. 2022. Disponível em: <https://doi.org/10.1214/10-STS330>

SILVA, C. O. L; COSTA, D. S.; NETO G. H. Desenvolvimento de um Modelo Preditivo em Prol da Identificação de Futuros Casos de Diabetes. **Revista Eletrônica de Computação Aplicada**. v. 02. n. 2, 2021.

SILVEIRA, M. B. G.; BARBOSA, N. F. M.; PEIXOTO, A. P. B.; XAVIER, É. F. M.; XAVIER JÚNIOR, S. F. A. Aplicação da regressão logística na análise dos dados dos fatores de risco associados à hipertensão arterial. **Research, Society and Development**, v. 10, p. e20101622964, 2021.

TONACO, L. A. B.; VIEIRA, M. A. S.; GOMES, C. S.; ROCHA, F. L.; OLIVEIRA-FIGUEIREDO, D. S. T.; MALTA, D. C.; VELESQUEZ-MALENDEZ, G. Social vulnerability associated with the self-reported diagnosis of type II diabetes: a multilevel analysis. **Revista Brasileira de Epidemiologia**. v. 24, n. 1: e210010. 2021. Acesso em: 12 mai. 2022. Disponível em: < https://doi.org/10.1590/1980-549720210010.supl.1>

YING, XUE. An Overview of Overfitting and its Solutions. **Journal Physics**. Conference Series 1168, 2019. Acesso em: 28. Mar. 2022. Disponível em: < https://iopscience.iop.org/article/10.1088/1742-6596/1168/2/022022/pdf>

ZANIELLI, D. **Doença crônica baseada em disglicemia: reinterpretando o pré-diabetes**. PEBMED .2019. Acesso em: 02 out. 2021. Disponível em: < https://pebmed.com.br/doenca-cronica-baseada-em-disglicemia-reinterpretando-o-pre-diabetes/ >