Previsão de Diabetes

Luca Escopelli

Junho de 2023

1 Introdução

A diabetes é uma doença relacionada à insulina, hormônio que regula a glicose no sangue e garante a produção de energia para o corpo. Pode ocorrer pela produção insuficiente ou má utilização do hormônio. Em ambos os casos o paciente deve estar atento e tomar as providências adequadas a fim de evitar outras complicações de saúde.

O tratamento da doença pode variar a depender do caso. Na maioria deles, o controle da alimentação, prática de atividades físicas e acompanhamentos de rotina são suficientes para o paciente viver tranquilamente. Em outros casos, é necessário acompanhamento diário do nível de insulina e aplicação de doses do hormônio. Para esse segundo caso, é ainda mais importante a identificação do problema com antecedência, tendo em vista que os riscos podem ser maiores e também envolvem o setor de saúde como um todo na provisão das doses necessárias.

Tendo em vista que essa é uma doença séria, mas que não apresenta muitos riscos se tratada corretamente, é de extrema importância que se consiga identificar se o paciente possui ou não diabetes. Com isso em mente, esse trabalho propõe a criação de um modelo que preveja a existência da doença de acordo com os dados do paciente.

Para isso, estamos utilizando um conjunto de dados de 100.000 pacientes que foram avaliado para verificar a presença da doença. Entre os dados temos: o sexo, a idade, se tem problema de hipertensão, se tem problema cardíaco, se fuma ou já fumou, o índice de massa corporal (IMC), o nível da hemoglobina A1c, o nível de glicose no sangue e se o paciente possui diabetes.

Pretendemos, com isso, contribuir para o setor da saúde facilitando a indetificação de pacientes com diabetes. Dessa forma, podemos trabalhar com antecedência no tratamento e melhorar a qualidade de vida dos indivíduos além de reforçar a qualidade do serviço público de saúde.

2 Métodos

O objetivo do nosso modelo é prever se determinado indivíduo possui diabetes ou não. Para isso, optamos por modelar utilizando regressão logística (como apresentado em [1], dado que é um bom método para classificação de variáveis binárias. Além disso, esse método indica as probabilidades de classificação, o que permite melhorias para esse problema, como a reavaliação dos pacientes que se encontram em uma situação intermediária.

Nosso conjunto de dados possui as seguintes informações:

- gender: sexo biológico do indivíduo (male/female/other)
- age: idade
- hypertension: se possui hipertensão (0 ou 1)
- heart_disease: se possui problema cardíaco (0 ou 1)
- smoking_history: se fuma ou já fumou (current/not current/never/ever/former/No Info)
- bmi: Índice de massa corporal (IMC)
- HbA1c_level: nível da hemoglobina A1c no sangue
- blood_glucose_level: nível de glicose no sangue
- diabetes: se possui diabetes (0 ou 1)

Analisando os dados, percebemos que a grande maioria das pessoas avalidas não possui diabetes. Com isso em mente, devemos ter alguns cuidados ao realizar o modelo.

Para começar, dividimos os dados entre treino e teste. Nessa divisão, optamos por um percentual equilibrado para evitar que algum grupo apresentasse uma quantidade muito pequena de pacientes diabéticos, o que poderia influenciar no modelo. Portanto, optamos por uma divisão de 60% dos dados para treino e 40% para teste.

Além disso, devido à disproporção entre os dados, a acurácia pode não ser a métrica mais interesante (Note que um modelo que simplesmente classique todos como não diabéticos teria acurácia alta). Ao invés disso, vamos avaliar, principalmente, o percentual de registros Falso Positivos. Queremos que esse percentual seja baixo, dado o perigo que seria para um portador da doença ser classificado como saudável.

Com isso em mente, optamos por classificar como positivo (possui diabetes) todos pacientes que o modelo apontasse terem mais de 20% de chance de possuir a doença. Posteriormente, podemos fazer uma reavaliação desse percentual de acordo com o interesse na classificação correta dos pacientes com diabetes a custo de avaliar incorretamente pacientes saudáveis.

3 Resultados

Ao analisar os resultados, devemos estar atentos a um ponto muito importante, que está relacionado à divisão dos registros dos nossos dados. Nosso conjunto de dados possui informações sobre 100.000 pessoas, das quais apenas 8.500 possuem diabetes efetivamente, isso é, apenas $8,5\%$ das pessoas avaliadas são portadoras da doença. Com isso em mente, devemos estar atentos aos nosso resultados, visto que um simples modelo que avalie todas as pessoas como "sem diabetes"teria $91,5\%$ de acurácia.

Passado esse primeiro ponto, vamos analisar as variáveis que temos para utilizar na formação nosso modelo. Verificando a matriz de correlação dos nossos dados, percebemos que nenhum par de variáveis possui correlação forte, e o que mais se aproxima disso é o par "age"e "bmi"(Verificar Figura 1). Faz sentido essa relação entre esse par, dado que ao envelhecer as pessoas mudam seus hábitos, se tornam mais sedentárias e o IMC tende a aumentar. No entanto, mesmo nesse caso nosso coeficiente está no nível de 34%, indicando que os dados possuem relação, porém não é uma relação extremamente forte que possa impactar o modelo.

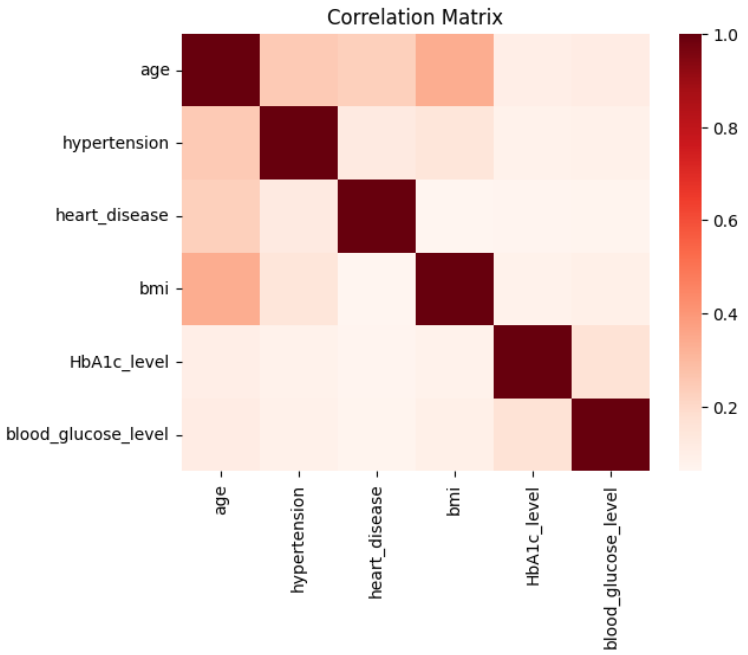

Figura 1: Matriz de correlação

Outra análise interessante é a verificação da relação das variáveis "HbA1c_level"e "blood_glucose_level"com o que queremos prever ("diabetes"). É intuitivo pensar que o nível de açúcar e glucose no sangue tenha impacto direto na presença

da doença no indivíduo. Com isso em mente, verificamos dois gráficos que indicam o valor dessa variáveis e se o sujeito possui ou não diabetes.

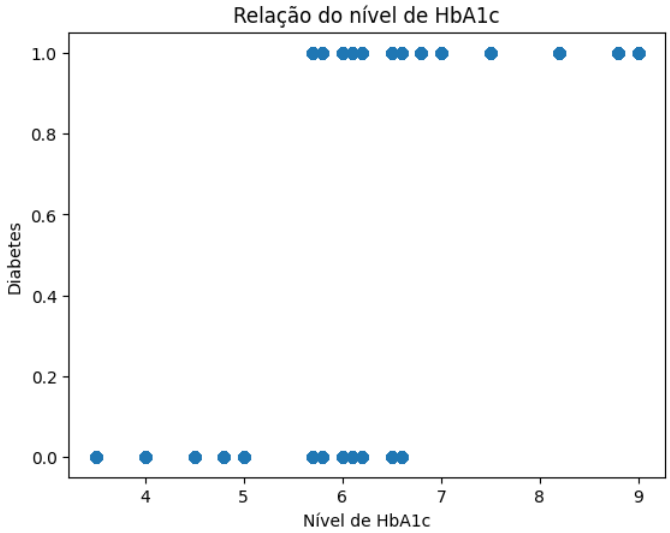

Figura 2: Nível de HbA1c

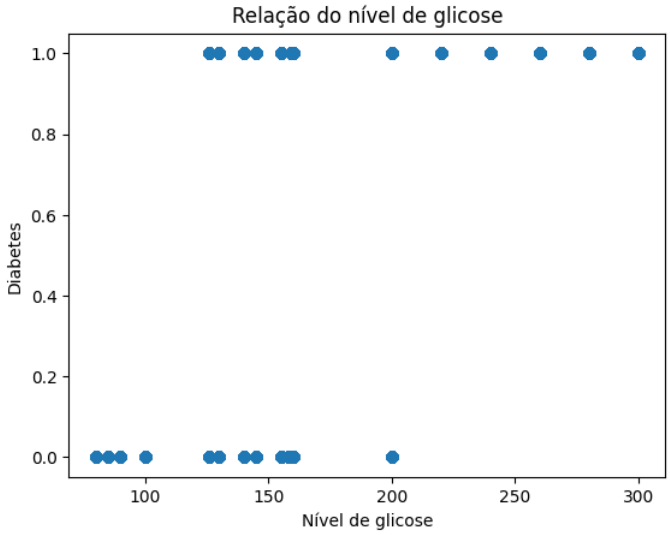

Figura 3: Nível de glicose

Podemos perceber pela Figura 2 que todos os indivíduos cujo nível de hemoglobina glicada (HbA1c) no sangue supera o valor de 7 possuem diabetes.

Assim como todos de nível inferior a 5,5 não possuem a doença.

Da mesma forma, podemos analisar a Figura 3 para concluir que todos que apresentaram nível de glicose superior a 200 são portadores de diabetes e aqueles com nível até 125 são saudáveis.

Com base nisso, nosso primeiro modelo realizou uma regressão logística com base apenas nessas duas informações para avaliar o quão relevante esses dados realmente são na previsão da existência da doença.

Avaliando esse modelo nos nossos dados de teste obtivemos a matriz de confusão da Figura 4. Conseguimos obter resultados bem satisfatórios para a quantidade de dados reduzida. O modelo apresentou acurácia superior a 92%, mas mais importante que isso, avaliou corretamente mais de 2/3 dos que possuem a doença.

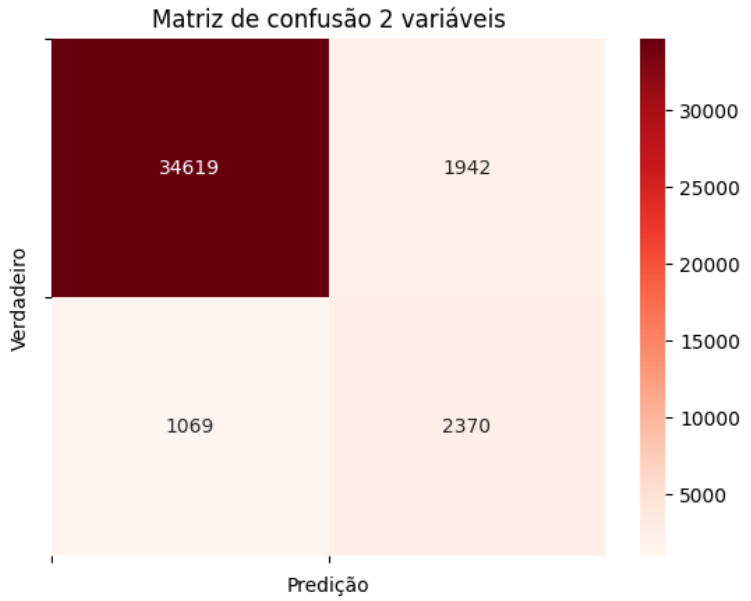

Figura 4: Modelo com apenas 2 variáveis

No entanto, possuímos mais informações do que apenas esses dois dados, então não faz sentido desperdiçá-las. Sendo assim, vamos refazer esse modelo de regressão logística, agora utilizando todos os dados. Para isso, vamos transformar os campos de "gender"e "smoking_history"em variáveis dummy referentes às categorias possíveis.

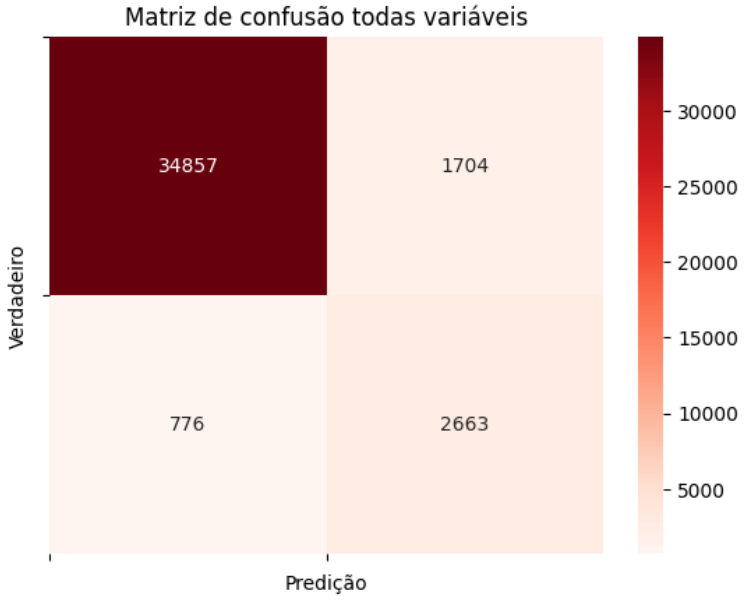

Figura 5: Modelo com todas variáveis

Podemos ver nesse novo modelo, apresentado pela Figura 5, que utilizando todas a variáveis conseguimos melhorar aquele modelo simples anterior. Agora apresentando uma acurácia de quase 94% e apontando 3/4 dos portadores da doença corretamente, conseguimos um modelo que atende às nossas necessidades.

4 Conclusão

Verificamos que é possível fazer um bom modelo de previsão de diabetes com base em algumas informações de saúde do paciente. Com isso, conseguimos melhorar a eficiência dos hospitais no tratamento da doença. Além disso, permitimos uma identificação antecipada da doença nos pacientes, o que acarreta numa melhor qualidade de vida com o tratamento.

Além disso, percebemos que há espaço para melhorias no modelo a depender da intenção de quem o utiliza. Uma das possíveis melhorias está relacionada à porcentagem limitadora da classificação. A qual podemos reduzir para melhorar a precisão na identificação dos portadores da doença, com o custo de aumentar a classificação incorreta de pessoas saudáveis.

Uma outra possível melhoria está relacionada à abordagem dos percentuais. Dado que o modelo utilizado fez uso de regressão logística e esse método nos retorna percentuais de classificação. Podemos dividir o problema em 3 classes ao invés de apenas 2, criando uma classe intermediária. Dessa forma, podemos passar mais confiança a quem for classificado como saudável e também iniciar de

imediato o tratamento de quem for classificado como diabético. Para os pacientes que se encontrarem na classe intermediária podemos requisitar mais exames a fim de obter certeza quanto a existência da doença ou não. Dessa forma, evitamos a surpresa posterior de pacientes que foram classificados incorretamente.

Com isso, podemos continuar o estudo dessa doença e a busca de maneiras mais eficientes para lidar com esse problema de saúde.

Referências

[1] Andrew Gelman, Jennifer Hill e Aki Vehtari. *Regression and Other Stories*. 2020.