Predição de Diabetes Mellitus em Mulheres: Desenvolvimento e Validação de um Modelo Preditivo

Pedro Jorge de Souza Colombrino
Matheus Ferreira Amaral Madeira
Guilherme Vieira Rodrigues

31 de Outubro, 2024

Abstract

BLABLABLA BLABLABLA BLABLABLA BLABLABLA BLABLABLA BLABLABLA BLABLABLA BLABLABLA BLABLABLA BLABLABLA BLABLABLA BLABLABLA BLABLABLA BLABLABLA BLABLABLA

1 Introdução

O diabetes mellitus, uma condição crônica marcada por níveis elevados de glicose no sangue, representa um desafio crescente para a saúde pública brasileira. Dados da Vigite Brasil (2023) revelam que cerca de 9 porcento da população adulta de cada estado convive com essa doença, com destaque para a capital paulista e o Distrito Federal. A prevalência entre mulheres é particularmente preocupante, sendo influenciada por fatores como diabetes gestacional, síndrome dos ovários policísticos e as alterações hormonais da menopausa, exige abordagens inovadoras para prevenção e tratamento.

A predição de pacientes femininas com maior risco de desenvolver diabetes é fundamental para a implementação de estratégias de prevenção personalizadas e para otimizar o cuidado dessas mulheres. Através da análise de dados clínicos e demográficos, é possível identificar padrões e biomarcadores que antecedem o diagnóstico e, assim, intervir de forma precoce, reduzindo as complicações da doença.

O objetivo deste projeto é desenvolver e validar um modelo preditivo capaz de identificar mulheres com maior risco de desenvolver diabetes mellitus,

com base em dados clínicos e demográficos, visando otimizar a prevenção e o tratamento dessa doença."

2 Metodologia

2.1 Conjunto de dados

O conjunto de dados, obtido do Kaggle, consiste em um conjunto de 768 registros médicos detalhados de pacientes do sexo feminino. Cada registro é caracterizado por 8 atributos clínicos relevantes, como idade, índice de massa corporal (IMC), níveis de glicose, pressão arterial, entre outros.

Table 1: Dicionário de Dados

Variável	**Descrição**	**Tipo de dado**
Gravidez	Número de vezes que o paciente esteve grávido	int64
Glicose	Concentração de glucose no plasma após um teste oral de tolerância à glucose de 2 horas	int64
PressaoSanguinea	Pressão arterial diastólica (mm Hg)	int64
EspessuraDaPele	Dobra cutânea tricipital (mm)	int64
Insulina	Insulina sérica de 2 horas (mu U/ml)	int64
IMC	Índice de massa corporal (peso em kg/(altura em m)^2)	float64
DiabetesPedigree	Uma função que representa o pedigree da diabetes do paciente (ou seja, a probabilidade de diabetes com base no historial familiar)	float64
Idade	Idade do doente (anos)	int64
Diabetico	Resultado binário (0 ou 1) em que 1 indica a presença de diabetes e 0 indica a ausência.	int64

2.2 Importância das Variáveis na Previsão de Diabetes

As variáveis presentes no dataframe têm papéis importantes na predição de futuros pacientes suspeitos de diabetes. Vamos dissecá-las para entender como elas contribuem para essa previsão:

- **Gravidez**: O número de vezes que uma mulher esteve grávida pode influenciar o risco de diabetes gestacional, que é um tipo de diabetes que ocorre durante a gravidez. Mulheres que tiveram diabetes gestacional têm maior risco de desenvolver diabetes tipo 2 no futuro. Além disso, a gravidez pode causar mudanças hormonais que afetam a sensibilidade à insulina.

- **Glicose**: A concentração de glicose no plasma é um indicador direto do nível de açúcar no sangue. Níveis elevados de glicose após um teste de tolerância à glicose podem indicar resistência à insulina ou diabetes. A glicose é uma variável crucial, pois níveis elevados de açúcar no sangue são a característica principal do diabetes.

- **Pressão Sanguínea**: A pressão arterial diastólica elevada pode ser um sinal de hipertensão, que é um fator de risco para diabetes tipo 2. A hipertensão e a diabetes frequentemente ocorrem juntas e podem aumentar o risco de complicações cardiovasculares. Controlar a pressão sanguínea é essencial para prevenir complicações associadas ao diabetes.

- **Espessura da Pele**: A dobra cutânea tricipital é uma medida da gordura subcutânea. A obesidade é um fator de risco significativo para diabetes tipo 2, e a espessura da pele pode ser um indicador de excesso de gordura corporal. A gordura corporal excessiva pode levar à resistência à insulina, aumentando o risco de diabetes.

- **Insulina**: Níveis elevados de insulina sérica podem indicar resistência à insulina, uma condição em que as células do corpo não respondem adequadamente à insulina. A resistência à insulina é um precursor comum do diabetes tipo 2. Monitorar os níveis de insulina pode ajudar a identificar indivíduos em risco antes que a diabetes se desenvolva.

- **IMC (Índice de Massa Corporal)**: O IMC é uma medida do peso corporal em relação à altura. Um IMC elevado indica sobrepeso ou obesidade, que são fatores de risco importantes para o desenvolvimento de diabetes tipo 2. Manter um IMC saudável é fundamental para a prevenção do diabetes.

- **Diabetes Pedigree**: Esta variável representa a probabilidade de diabetes com base no histórico familiar. Um alto valor de pedigree de diabetes indica uma predisposição genética para a doença. Conhecer o histórico familiar pode ajudar na identificação precoce e na implementação de medidas preventivas.

- **Idade**: A idade é um fator de risco para diabetes tipo 2. O risco de desenvolver diabetes aumenta com a idade, especialmente após os 45 anos. O envelhecimento está associado a uma diminuição da função das células beta do pâncreas e a uma maior resistência à insulina.
- **Diabético**: Esta variável indica se o paciente tem diabetes (1) ou não (0). É o resultado binário que mostra a presença ou ausência da doença. Esta variável é essencial para a classificação e análise dos dados, permitindo a identificação de padrões e fatores de risco associados ao diabetes.

Cada uma dessas variáveis pode fornecer informações valiosas sobre o risco de diabetes e ajudar na identificação precoce e no manejo da doença. A combinação dessas variáveis em modelos preditivos pode melhorar a precisão na previsão de diabetes e permitir intervenções mais eficazes.

3 Análise dos Resultados

Os resultados da análise estatística revelaram associações significativas entre diversas variáveis e o diagnóstico de diabetes. A tabela abaixo resume os valores de p para cada variável testada:

Variável	Valor-p
Gravidez	0.0000000005
Glicose	0.0000000000
Pressão Sanguínea	0.0715139001
Espessura da Pele	0.0383477048
Insulina	0.0002861865
IMC	0.0000000000
Diabetes Pedigree	0.0000012546
Idade	0.0000000000

Table 2: Valores de p para cada variável.

Um valor-p menor que 0,05 indica uma associação estatisticamente significativa entre a variável e o diagnóstico de diabetes.

3.1 Discussão dos Resultados

A associação significativa entre os níveis de insulina e o diagnóstico de diabetes é consistente com a compreensão atual da fisiopatologia do diabetes

tipo 2. Níveis elevados de insulina podem indicar resistência à insulina, uma condição em que as células do corpo se tornam menos responsivas à ação da insulina, levando ao aumento dos níveis de glicose no sangue.

3.2 Possíveis Causas para o Baixo Valor-p da Insulina

- **Resistência à insulina:** Como mencionado, a resistência à insulina é um fator chave no desenvolvimento do diabetes tipo 2.
- **Disfunção das células beta:** As células beta do pâncreas são responsáveis pela produção de insulina. Uma disfunção dessas células pode levar a níveis elevados de insulina em um esforço para compensar a resistência à insulina.
- **Inflamação:** Processos inflamatórios crônicos podem contribuir para a resistência à insulina e disfunção das células beta.
- **Outros fatores genéticos e ambientais:** Vários outros fatores, como genética, estilo de vida e fatores socioeconômicos, podem influenciar os níveis de insulina e o risco de diabetes.

3.3 Conclusão

Os resultados deste estudo sugerem que os níveis de insulina são um importante marcador de risco para o desenvolvimento de diabetes. A compreensão dos mecanismos subjacentes à associação entre insulina e diabetes é fundamental para o desenvolvimento de novas estratégias de prevenção e tratamento.

4 Uso do RandomForestClassifier para Predição de Diabetes

Neste estudo, utilizamos o **RandomForestClassifier** para prever a variável *Diabetico*, que indica a presença (1) ou ausência (0) de diabetes. A Random Forest é um modelo de aprendizado de máquina que combina várias árvores de decisão para melhorar a precisão e a robustez das previsões. Esse método é amplamente reconhecido por sua eficiência em lidar com variáveis complexas e pela capacidade de capturar interações importantes entre diferentes características dos dados.

4.1 Como o RandomForestClassifier Funciona

O *RandomForestClassifier* funciona criando diversas árvores de decisão durante o processo de treinamento. Cada árvore é construída a partir de uma amostra aleatória do conjunto de dados, com algumas variáveis selecionadas aleatoriamente para cada divisão. No momento da previsão, o resultado final é determinado por um processo de votação das árvores (para classificação) ou pela média das predições (para regressão). Essa estratégia reduz significativamente o risco de sobreajuste (overfitting), que pode ocorrer em árvores de decisão isoladas, e melhora a capacidade do modelo de fazer previsões precisas em novos dados.

4.2 A Importância da Validação Cruzada K-Fold

Para avaliar o desempenho do modelo, utilizamos a técnica de validação cruzada K-Fold. Essa técnica divide o conjunto de dados em k partes, chamadas *folds*. Em cada iteração, um dos *folds* é separado como conjunto de teste, enquanto os outros *k-1* são usados para treinar o modelo. O processo é repetido k vezes, garantindo que cada parte dos dados seja usada para teste exatamente uma vez.

Neste estudo, escolhemos uma validação cruzada com 5 *folds* (5-fold). Essa abordagem traz várias vantagens importantes:

- **Uso de todos os dados**: Cada observação nos dados é usada tanto para treino quanto para teste, o que proporciona uma avaliação mais abrangente da performance do modelo.
- **Redução da variabilidade**: Com os resultados das várias iterações, obtemos uma média que reflete de forma mais precisa a capacidade do modelo de fazer boas previsões em dados desconhecidos.
- **Prevenção de sobreajuste**: O modelo é testado em diferentes subconjuntos, o que ajuda a identificar se ele está aprendendo de forma excessivamente específica para o conjunto de treino.

A acurácia média das predições nos diferentes *folds* é calculada para fornecer uma medida consolidada do desempenho do modelo. Isso permite uma visão clara da eficácia do *RandomForestClassifier* para identificar mulheres com risco de diabetes.

$$\text{Acurácia média} = \frac{\sum_{i=1}^{k} \text{Acurácia}_i}{k} \tag{1}$$

4.3 Resultados e Conclusões

Os resultados indicaram que o uso da Random Forest em combinação com a validação cruzada K-Fold é uma abordagem eficaz para a previsão da presença de diabetes. A média das acurácias em todas as dobras mostra a consistência do modelo e sua capacidade de generalizar para novos dados, tornando-o uma ferramenta confiável para identificar padrões e prever o risco de diabetes com base em dados clínicos e demográficos.

A metodologia adotada neste estudo proporciona uma avaliação sólida e prática da performance do modelo, garantindo que ele aproveite ao máximo os dados disponíveis e que as previsões sejam o mais precisas possível.