Uso de Algoritmos de Classificação para Predição de Novos Casos de Diabetes Mellitus

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

31 de Outubro, 2024

Abstract

Este estudo apresenta a aplicação de algoritmos de aprendizagem Máquina, focada em RandomForestClassifier, para previsão de novos casos de diabetes em mulheres. O objetivo é demonstrar como treinar, avaliar e usar esses modelos para prever a variável dependente diabetes (dicotômica) com base em novas entradas de dados clínicos e demográficos. Usando verificação cruzada através do K-Fold, procuramos identificar os modelos mais eficazes e validar a sua aplicação prática em cenários reais de saúde pública.

1 Introdução

A capacidade de prever novos casos de diabetes com base em dados clínicos é uma ferramenta poderosa para a saúde pública e a medicina personalizada. A diabetes é uma doença crónica associada a graves riscos de complicações e morte, sendo particularmente prevalente nas mulheres devido a fatores específicos como a diabetes gestacional e alterações hormonais.

À medida que a disponibilidade de dados clínicos continua a aumentar, os algoritmos de aprendizagem automática destacam-se pela sua capacidade de identificar padrões complexos e fazer previsões precisas. Este estudo explora a aplicação de algoritmos de classificação, com foco em *RandomForestClassifier*, para prever novos casos de diabetes em mulheres com base em suas características clínicas e demográficas.

O foco principal deste trabalho é demonstrar como modelos treinados e validados podem ser usados para prever a presença de diabetes em novas

linhas de dados, fornecendo uma abordagem prática e aplicável à tomada de decisão clínica.

2 Metodologia

2.1 Conjunto de Dados

O conjunto de dados utilizado foi do Kaggle e continha informações sobre 768 pacientes do sexo feminino. Cada registro é composto por 8 variáveis clínicas e demográficas e uma variável alvo (*Diabetic*) que indica a presença ou ausência de diabetes.

Table 1: Dicionário de Dados

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

2.2 Algoritmo de Classificação: RandomForestClassifier

Neste trabalho, utilizamos *RandomForestClassifier*, um modelo baseado em árvore de decisão que combina múltiplas árvores para gerar previsões robustas. Sua capacidade de lidar com variáveis complexas e detectar interações não lineares o torna ideal para resolver tais problemas.

2.3 Processo de Validação

A validação do modelo foi realizada utilizando a técnica de validação cruzada K-fold (5 vezes). Esta abordagem:

- Garante que o modelo seja avaliado em relação a todos os dados disponíveis;
- Reduz o risco de overfitting;
- Fornece métricas de desempenho mais consistentes e confiáveis.

2.4 Critério para Seleção do Melhor Modelo

Ao final do processo de validação cruzada, o modelo com maior precisão média em *folds* é selecionado e salvo em disco usando a biblioteca `pickle`. O modelo é posteriormente usado para prever novas entradas de dados.

3 Uso do Modelo Treinado para Predição de Novos Casos

Depois de selecionar o melhor modelo, usamos o *RandomForestClassifier* salvo para fazer previsões sobre novos dados. Este processo demonstra a aplicabilidade prática do modelo em cenários reais.

3.1 Exemplo de Aplicação

Considere os dados de um novo paciente:

```
novo_paciente = [[5, 176, 72, 17, 24.6, 0.387, 34]]
```

Os passos para realizar a previsão são:

1. Carregar o modelo salvo utilizando a biblioteca `pickle`.
2. Passar os dados do novo paciente ao modelo para gerar a predição.
3. Interpretar o resultado: 1 (diabético) ou 0 (não diabético).

```
# Carregar o modelo salvo
with open("melhor_modelo_random_forest.pkl", "rb") as f:
    modelo = pickle.load(f)

# Prever o novo paciente
```

```
predicao = modelo.predict(novo_paciente)

# Resultado
resultado = "Diabético" if predicao[0] == 1 else "Não diabético"
print(f"Previsão: {resultado}")
```

4 Resultados

Os resultados da validação cruzada mostram que *RandomForestClassifier* tem uma precisão média de 85% nos dados de teste, destacando sua eficiência na previsão de novas situações. O modelo demonstra robustez e consistência globais.

5 Conclusão

Este estudo demonstra a eficácia do uso de algoritmos de classificação (com ênfase em *RandomForestClassifier*) para prever novos casos de diabetes em mulheres. As aplicações práticas do modelo, exemplificadas pelas previsões sobre novos dados, destacam o potencial destas ferramentas para apoiar a tomada de decisões clínicas e estratégias de saúde pública. Melhorias futuras incluem a adição de mais dados e o uso de técnicas de otimização de hiperparâmetros para melhorar ainda mais o desempenho do modelo.