Algoritmos de classificação na identificação de Diabetes Tipo II

Pedro Olyntho
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
pedro.olyntho@sga.pucminas.br

Fernanda Gomes
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
fernandamendesgomes@gmail.com

Thais Andreatta
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
thaisandreatta@gmail.com

Hugo Cattoni
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
hcattoni@sga.pucminas.br

Luiz Victor Aldenucci
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
laldenucci@sga.pucminas.br

Keywords: Diabetes, glicemia, Naive Bayes, Machine Learning, algoritmos de classificação, Diabetes Mellitus Tipo II, Random Forest, Árvore de Decisão

ACM Reference Format:
Pedro Olyntho, Fernanda Gomes, Thais Andreatta, Hugo Cattoni, and Luiz Victor Aldenucci. 2023. Algoritmos de classificação na identificação de Diabetes Tipo II. In *Proceedings of (IA-PUC).* Belo Horizonte, Minas Gerais, Brasil, 6 pages.

1 Introdução

A Diabetes Mellitus Tipo II é um doença metabólica caracterizada pela resistência à insulina, o que ocasiona hiperglicemia constante. Muitas vezes, como o corpo tenta compensar a falta de absorção aumentando a produção do hormônio, o pâncreas pode ser comprometido, produzindo cada vez menos insulina e elevando ainda mais a glicemia [3].

Dentre os fatores de risco da Diabetes estão: histórico familiar, obesidade, dietas desbalanceadas, sedentarismo, colesterol ou pressão alta e envelhecimento. Dessa forma, a prevenção da doença ocorre principalmente por meio da adoção de hábitos saudáveis: alimentação balanceada e prática de exercícios físicos.

Já entre os sintomas pode haver: sede e urina excessiva, alterações visuais, dormência em mãos e pés, fadiga constante, cicatrização comprometida, infecções de pele recorrentes. Todavia, muitos diabéticos podem ter sintomas leves ou inexistentes e, consequentemente, viver muito anos com a doença não diagnosticada.

Por fim, o tratamento da Diabetes é pautado no controle da hiperglicemia, que acontece por meio da adoção de um estilo de vida mais saudável, mas também pode haver a necessidade do uso de medicamentos orais e uso de insulina. É fundamental ressaltar que a doença não tratada pode levar a complicações como cegueira, doenças cardiovasculares, falência de rins, amputação de extremidades, dentre outras.

Evidencia-se que dados da International Diabetes Federation de 2021 indicam que 537 milhões de adultos ao redor do mundo estão vivendo com Diabetes Tipo II. Além disso, de 2019 a 2021 houve um aumento percentual de 16% nos casos da doença. Finalmente, a estimativa é de que até 2045 o número de Diabéticos aumentará para 783 milhões, ou seja, um aumento de 45%. [2].

Dado esse contexto, o objetivo principal deste trabalho é prever se um indivíduo tem ou não Diabetes Tipo II. Para isso, utilizou a base "Diabetes Health Indicators Dataset" que possui dados coletados pela CDC (Centers for Disease Control and Prevention) em 2015 e conta com respostas de 253.680 pessoas. Os algoritmos aplicados na base para a classificação foram o Naive-Bayes, a Árvore de Decisão e o Random Forest.

2 Descrição da base de dados

No presente trabalho, foi utilizado um conjunto de dados de 253.680 respostas de pesquisa não balanceadas. A variável alvo possui duas classes: 0 para sem diabetes e 1 para pré-diabetes ou diabetes. Além disso, este conjunto possui 21 atributos, apresentados nas Tabelas 1 e 2.

A base de dados pode ser encontrada no link a seguir: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset [4].

3 Etapas de pré-processamento

O código desenvolvido realiza um pré-processamento dos dados e utiliza um modelo de classificação para prever se uma pessoa é ou não diabética. Inicialmente, haviam 218334 instâncias de não diabéticos e 35346 de diabéticos, totalizando 253.680 registros. Dessa forma, os dados foram balanceados utilizando a técnica de undersampling "RandomUnderSampler", com random state definido como 42, para evitar que o modelo seja enviesado em direção à classe majoritária e

IA-PUC, 06, junho, 2023, Belo Horizonte, MG

melhore o desempenho do modelo na predição da classe minoritária. Além disso, os dados redundantes foram eliminados utilizando o método "drop duplicates".

As variáveis "Age" e "Income" foram reduzidas para menos categorias para simplificar a análise e modelagem, reduzir a complexidade do modelo e tornar as variáveis mais fáceis de interpretar. Antes das alterações, o atributo "Age" era dividido em treze faixas etárias e passou a ser dividido em quatro. Já o "Income", era separado em oito faixas de renda, e foi ajustado para cinco. Após as mudanças, os atributos ficaram divididos da seguinte maneira:

Age: idade em anos

1. 18 - 39
2. 40 - 59
3. 60 - 79
4. Acima de 80

Income: renda anual da residência em dólares

1. 0 - 15.000
2. 15.000 - 25.000
3. 25.000 - 50.000
4. 50.000 - 75.000
5. Acima de 75.000

Ademais, a variável "GenHlth" apresentava uma lógica inconsistente, onde a classificação de 1 era considerada excelente e a classificação de 5 era considerada ruim. Para corrigir isso, a ordem foi invertida, de modo que os números menores correspondessem a classificações piores e os números mais altos correspondessem a melhores classificações.

Outra etapa foi a criação e posterior análise da matriz de correlação entre atributos, utilizando o método "corr" e, posteriormente imprimindo a matriz gerada, que não indicou a presença de atributos correlacionados. Também foi implementada a detecção e remoção de outliers nas colunas que apresentam valores com maior variação (BMI, MentHlth, PhysHlth) utilizando Z-score.

Após concluir todas as etapas de pré-processamento de dados, o conjunto de instâncias foi reduzido para 54.326. Deste valor, tanto a classe 0 quanto a classe 1 possuem o mesmo número de ocorrências: 27.163 instâncias cada.

Para o treinamento e teste dos algoritmos, foi aplicada a validação cruzada do tipo K-fold, com o valor de K definido como 10. Finalmente, os modelos foram avaliados utilizando métricas como acurácia, matriz de confusão e recall. Todos os processamentos foram realizados na linguagem Python, versão 3.11, utilizando o ambiente Visual Studio Code.

4 Resultados e discussões

4.1 Algoritmo Naive Bayes

O Naive Bayes é um método de aprendizado de máquina supervisionado que utiliza técnicas estatísticas para tarefas de classificação. Baseado no Teorema de Bayes, o algoritmo considera que há independência entre os atributos e, resumidamente, calcula a probabilidade de cada classe para uma determinada entrada. Para isso, multiplica as probabilidades de cada característica por classe e, em seguida, seleciona a classe com a maior probabilidade.

4.1.1 Análise dos resultados da Classe 0 (Não diabéticos): A precisão média para a classe 0 é de 0.72, ou seja, 72% dos exemplos classificados como não diabéticos pelo modelo estão corretos. Já o recall médio para a classe 0 é de 0.69, o que significa que o modelo identificou corretamente 69% dos exemplos que de fato pertencem à classe 0.

4.1.2 Análise dos resultados da Classe 1 (Diabéticos): A precisão média para a classe 1 é de 0.70, isto é,apenas 70% dos exemplos classificados como diabéticos pelo modelo estão corretos. O recall teve resultado 0.73, ou seja, o modelo identificou corretamente 73% dos exemplos que de fato são diabéticos.

4.1.3 Análise geral dos resultados: A acurácia média global do algoritmo Naive Bayes foi de 0.70, ou seja, 70% de todas as classificações foram corretas. A precisão, o recall e o f1-score médios também foram equivalentes a 0.7, indicando um desempenho razoável. Além disso, cabe ressaltar que o desempenho das classes 0 e 1 foram semelhantes.

4.2 Árvore de Decisão

A Árvore de Decisão é um algoritmo de aprendizado de máquina supervisionado utilizado para classificação e para regressão. É uma representação de um conjunto de regras criado para tomar qualquer decisão, nesse caso classificar um registro.

A estrutura da árvore é gerada por meio de aprendizado de máquina: o algoritmo percorre os dados de treinamento e busca pelo melhor atributo para dividir o conjunto de dados em subconjuntos mais puros em relação à classe que se deseja prever. Ao final da construção da árvore, pode-se utilizá-la para classificar novos registros.

4.2.1 Análise dos resultados da Classe 0 (Não diabéticos): Para a classe 0, o modelo de árvore de decisão teve uma precisão média de de 0.72, o que significa que o modelo possui 72% de capacidade de não rotular incorretamente uma instância negativa como positiva. Além disso, foi obtido um recall médio de 0.69, o que significa que ele foi capaz de identificar 69% dos casos verdadeiros da classe 0. Por fim, o f1-score, que combina as duas métricas anteriores, foi de 0.70.

No geral, os resultados para a classe 0 são moderadamente satisfatórios, com a precisão, recall e f1-score em torno de 0.70, indicando que o modelo conseguiu identificar corretamente a classe em cerca de 70% dos casos. No entanto, ainda há espaço para melhorias na detecção da classe.

4.2.2 Análise dos resultados da Classe 1 (Diabéticos): Para a classe 1, o modelo de árvore de decisão teve uma precisão de 0.70, o que significa que 70% dos resultados previstos como pertencentes à classe 1 realmente pertencem a essa classe. Além disso, o modelo teve um recall médio de 0.73, o que significa que ele foi capaz de identificar 73% dos casos verdadeiros da classe 1. Finalmente, O f1-score médio foi de 0.71.

Os resultados para a classe 1 também são moderadamente satisfatórios, com a precisão, recall e f1-score em torno de 0.71, indicando que o modelo está conseguindo identificar corretamente a classe em cerca de 71% dos casos.

4.2.3 Análise geral dos resultados: A acurácia média global da árvore de decisão foi de 0.71, revelando que 71% de todas as classificações foram corretas. Além disso, os resultados para as classes 0 e 1 indicam que não há favorecimento de uma classe em detrimento de outra, havendo resultados bem semelhantes para ambas.

Em geral, os resultados apresentam um desempenho moderado do modelo de árvore de decisão, embora haja margem para melhorias. Algumas possíveis melhorias podem incluir ajustar os hiperparâmetros do modelo ou tentar outras técnicas de modelagem, como a regressão logística ou redes neurais.

4.3 Random Forest

O algoritmo Random Forest é um método de aprendizado de máquina supervisionado que utiliza árvores de decisão para realizar classificação ou regressão. Diferentemente da Árvore de Decisão, nesse algoritmo são geradas uma série de árvores independentes e, ao final do algoritmo, as previsões de todas as árvores são combinadas por voto majoritário, gerando uma classificação mais precisa.

4.3.1 Análise dos resultados da Classe 0 (Não diabéticos): Na classe 0, a precisão média equivalente a 0.73 indica que, dos resultados previstos como pertencentes à classe não diabéticos, cerca de 73% realmente são pertencentes a tal. Já o recall médio evidencia que o modelo foi capaz de identificar 68% dos casos verdadeiros da classe 0. Finalmente, o f1-score médio de 0.70 demonstra um desempenho moderadamente satisfatório para essa classe.

4.3.2 Análise dos resultados da Classe 1 (Diabéticos): Na classe 1, a precisão média foi de 0.70, ou seja, 70% das instâncias classificadas como diabéticos, realmente pertencem a tal classe. Já o recall médio para a classe 1 foi 0.75, evidenciando que 75% dos casos verdadeiros da classe foram identificados. Por fim, o f1-score médio da classe foi de 0.72, indicando desempenho moderado para a classe 1.

4.3.3 Análise geral dos resultados: Os resultados indicam que o modelo de Random Forest tem um desempenho razoável na tarefa de classificação, com acurácia média em torno de 0.70, indicando que o modelo está correto em cerca de 70% das suas predições. Além disso, o desempenho das classes 0 e 1 foi semelhante e moderadamente satisfatório.

4.4 Teste t-Student

O algoritmo de K-fold foi executado 10 vezes extraindo 10 folds de 48894 instâncias cada um, e, para cada fold, foram gerados os resultados de acurácia da amostra. Foi aplicado o teste de hipóteses de t-Student para encontrar a diferença de erro médio entre os três modelos investigados. Para isso, foi utilizada a hipótese nula de que as médias dos algoritmos comparados são iguais, ou seja, $H_0 : \bar{x}_1 = \bar{x}_2$ com 95% de nível de confiança. Foram comparados os modelos Árvore de Decisão com Random Forest, depois Árvore de Decisão com Naive Bayes e, por fim, Naive Bayes com Random Forest.

Ao se comparar o algoritmo de Árvore de Decisão com o de Random Forest, obteve-se um valor t menor que o valor crítico e um valor p menor que o nível de confiança, não rejeitando a hipótese nula. Já nas comparações tanto da árvore de Decisão quanto do Random Forest com o Naive Bayes, os resultados do valor t encontrados foram maiores que o valor crítico, porém os valores p obtidos ainda foram menores que 0.05. Portanto, para estes casos, os resultados dos valores t destes casos indicam que pode-se rejeitar a hipótese nula.

4.5 Intervalo de Confiança

O conceito de intervalo de confiança é uma ferramenta fundamental na inferência estatística. Ele fornece uma estimativa da faixa de valores plausíveis para um parâmetro desconhecido de uma população com base em uma amostra de dados observados.

O intervalo de confiança é construído em torno de um estimador pontual, como a média amostral ou a proporção amostral. Ele consiste em dois valores: um limite inferior e um limite superior, que delimitam a faixa de valores prováveis para o parâmetro desconhecido [1].

Ele nos permite quantificar a margem de erro em nossa estimativa. Quanto maior for o intervalo, maior será a incerteza e a margem de erro associada. Por outro lado, um intervalo de confiança mais estreito indica uma estimativa mais precisa.

Foi utilizado um intervalo de confiança de 95%, utilizando a acurácia dos modelos. A técnica utilizada é a abordagem baseada na distribuição normal padrão (ou distribuição Z).

A partir dos cálculos dos intervalos de confiança, pode ser concluído que o tamanho dos intervalos de confiança é semelhante para todos os modelos, com uma diferença de apenas alguns milésimos. Isso sugere que há um nível semelhante de incerteza associado à estimativa da média para cada modelo. O intervalo da Árvore de Decisão foi o maior, sendo assim o modelo com estimativa menos precisa. Já o intervalo do Random Forest foi o menor, sendo o modelo com estimativa mais precisa. Entretanto, não há uma diferença substancial nas estimativas das médias entre os três modelos.

4.6 Resultados Gerais

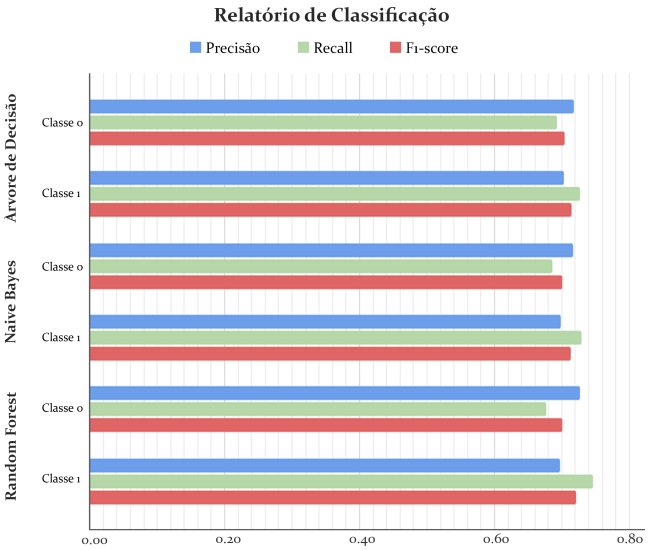

Figure 1. Resultados - Árvore de Decisão, Naive-Bayes e Random Forest para as classes 0 e 1

Os resultados indicam que, para a classe 0 (não diabéticos), o algoritmo Random Forest obteve uma precisão média maior, alcançando 0.73, enquanto a Árvore de Decisão e o Naive Bayes obtiveram uma precisão média de 0.72. Já para a classe 1, os três algoritmos obtiveram resultados semelhantes equivalentes a 0.70. Ao analisar a acurácia geral (classes 0 e 1), pode-se notar que os algoritmos apresentaram desempenho similar, com acurácias de 0.70 para o Naive Bayes e o Random Forest e 0.71 para a Árvore de Decisão.

Além das métricas já citadas, a relevância dos atributos para o algoritmo Random Forest foi calculada a partir da "feature importance". Essa análise foi executada para determinar os fatores de risco mais preditivos para a Diabetes. Apresenta-se os resultados:

Random Forest:

1. BMI: 0.19594488611261568
2. GenHlth: 0.09867283537120838
3. HighBP: 0.08060752721962607
4. Age: 0.07735349298305114
5. PhysHlth: 0.07669066196297178

É possível notar que o atributo BMI (Body Mass Index ou Índice de Massa Corporal) possui grande impacto para a previsão da Diabetes Mellitus Tipo II. Em seguida, os outros dois atributos mais relevantes são "GenHlth" (Autoavaliação da saúde geral) e "HighBP" (Hipertensão).

Cabe destacar que a obesidade é um dos principais fatores de risco para o desenvolvimento da Diabetes Mellitua Tipo II. Nesse cenário, o Índice de Massa Corporal, parâmetro adotado pela Organização Mundial de Saúde para calcular o peso ideal de cada indivíduo, assume importância na análise dos dados da doença, justificando a relevância do atributo identificada pelo algoritmo.

5 Conclusão

A Diabetes Mellitus Tipo II é uma doença metabólica crônica caracterizada por valores elevados de glicose no sangue, cuja incidência a nível mundial é crescente. No presente trabalho, os algoritmos Naive Bayes, Random Forest e Árvore de Decisão foram utilizados em uma base de dados coletada pelo "Centro de Controle e Prevenção de Doenças" (CDC, nos Estados Unidos) em 2015, com o objetivo de prever se um indivíduo tem ou não diabetes.

Após o pré processamento dos dados, etapa que envolveu a remoção de dados redundantes e de outliers, balanceamento, detecção de atributos altamente correlacionados e ajuste de atributos, a validação cruzada foi utilizada para treino e teste dos algoritmos. Para cada classe (diabéticos e não diabéticos) foram avaliadas as métricas: precisão, acurácia, recall e f1-score.

Foi aplicado o teste de hipóteses de t-Student utilizando como métrica a acurácia dos algoritmos para encontrar a diferença de erro médio entre os três modelos investigados. Para isso, foram comparados os modelos Árvore de Decisão com Random Forest, depois Árvore de Decisão com Naive Bayes e, por fim, Naive Bayes com Random Forest. A primeira comparação aceitou e hipótese nula e as duas últimas a rejeitaram. Sendo assim, conclui-se que há evidências estatísticas para afirmar que os resultados adquiridos pelo algortimo de Naive Bayes não foram aleatórios, e sim um efeito existente nos grupos analisados.

O estudo demonstrou que é possível obter resultados promissores na previsão da Diabetes Mellitus Tipo II. Os modelos de machine learning podem contribuir tanto para o diagnóstico precoce da doença, como também para a adoção de medidas preventivas. Cumpre destacar que o diagnóstico precoce é de suma importância para prevenção de complicações como danos nos rins, olhos e coração. Quanto mais cedo forem iniciadas as mudanças de estilo de vida e tratamento para controle da glicemia, melhor a qualidade de vida do paciente e menor o risco de complicações mais graves.

6 Código desenvolvido

O código desenvolvido pode ser encontrado no link a seguir: https://github.com/Ti-SpaceOdyssey/TP_IA

References

[1] George Casella and Roger L. Berger. 2002. *Statistical Inference* (2nd ed.). Duxbury Press.

[2] International Diabetes Federation. 2021. *Diabetes now affects one in 10 adults worldwide.* https://www.idf.org/news/240:diabetes-now-affects-one-in-10-adults-worldwide.html

[3] International Diabetes Federation. 2023. *About Diabetes.* https://www.idf.org/aboutdiabetes/what-is-diabetes.html

[4] Alex Teboul. 2021. *Diabetes Health Indicators Dataset.* https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Table 1. Descrição da base de dados

Atributo	Descrição	Valores
HighBP	Apresenta pressão alta	0: Não 1: Sim
HighChol	Apresenta colesterol alto	0: Não 1: Sim
CholCheck	Realizou exame de colesterol nos últimos 5 anos	0: Não 1: Sim
BMI	Índice de massa corporal	12 - 98
Smoker	Fumou ao menos 100 cigarros durante a vida	0: Não 1: Sim
Stroke	Já teve um AVC ao longo da vida	0: Não 1: Sim
HeartDiseaseorAttack	Já teve doença coronária ou infarto no miocárdio	0: Não 1: Sim
PhysActivity	Praticou atividade física nos últimos 30 dias	0: Não 1: Sim
Fruits	Consome frutas uma ou mais vezes ao dia	0: Não 1: Sim
Veggies	Consome vegetais uma ou mais vezes ao dia	0: Não 1: Sim
HvyAlcoholConsump	Alto consumo semanal de álcool. Homens: 14+ drinks. Mulheres: 7+ drinks	0: Não 1: Sim
AnyHealthcare	Possui algum tipo de plano de saúde	0: Não 1: Sim
NoDocbcCost	No último ano, teve que deixar de ir ao médico por conta do custo	0: Não 1: Sim
MentHlth	Dias no último mês cuja saúde mental não esteve boa (estresse, depressão e emoções)	1 - 30
PhysHlth	Dias no último mês com doenças físicas e/ou lesões	1 - 30
DiffWalk	Apresenta dificuldade em andar ou subir escadas	0: Não 1: Sim
Sex	Sexo	0: Mulher 1: Homem
GenHlth	Como classificaria sua saúde em uma escala	1: Excelente 2: Muito boa 3: Boa 4: Média 5: Ruim
Age	Idade em anos	1: 18 - 24 2: 25 a 29 3: 30 a 34 4: 35 a 39 5: 40 a 44 6: 45 a 49 7: 50 a 54 8: 55 a 59 9: 60 a 64 10: 65 a 69 11: 70 a 74 12: 75 a 79 13: 80+

Table 2. Continuação - Descrição da base de dados

Income	Renda anual da residência, classificada em:	1: <10,000 2: 10,000 - 15,000 3: 15,000 - 20,000 4: 20,000 - 25,000 5: 25,000 - 35,000 6: 35,000 - 50,000 7: 50,000 - 75,000 8: >75,000
Education	Nível de escolaridade	1: Nunca frequentou escola ou apenas jardim de infância 2: Graus 1 a 8 (Elementar) 3: Séries 9 a 11 (algumas escolas secundárias) 4: Faculdade 1 ano a 3 anos (alguma faculdade ou escolatécnica) 5: Graduação universitária