

# Uso de Machine Learning para a Detecção Precoce de Diabetes: Análise de Modelos de Classificação Binária a partir de Dados Comportamentais

André Menezes; Davi Dias; Marco Noronha; Matheus Rangel; Paulo Sendas

Pontifícia Universidade Católica de Minas Gerais

andre.menezes@sga.pucminas.br; davi.magalhaes@sga.pucminas.br; marco.noronha@sga.pucminas.br;  
matheus.figueiredo.1275135@sga.pucminas.br paulo.resende@sga.pucminas.br

## Abstract

A diabetes é a doença não transmissível com maior número de enfermos. Hábitos como sedentarismo e consumo excessivo de açúcares são considerados como causadores de diabetes. Estudar o comportamento da população através dos dados gerados através do uso de produtos digitais pode oferecer informações valiosas para detecção precoce de pré-diabetes. O uso de machine learning para a construção de sistemas de análise preditiva orientados para a detecção de doenças constitui uma poderosa ferramenta para a saúde preventiva. Este estudo analisa a eficácia de uso de modelos de classificação binária para a detecção de diabetes através de dados comportamentais.

## Keywords

Machine Learning; Classificação Binária; Diabetes; Epidemiologia.

## ACM Reference format

André Menezes, Davi Dias, Marco Noronha, Matheus Rangel and Paulo Sendas. 2024. Detecção de Diabetes In Belo Horizonte '24: Trabalho Prático de IA, Belo Horizonte, MG.

## 1. INTRODUÇÃO

A Diabetes é uma doença crônica que está relacionada ao processamento e produção de insulina pelo corpo humano. Essa doença apresenta acelerada tendência de crescimento e tem impacto global. A Federação Internacional de Diabetes (IDF) estima que há mais de 400 milhões de pessoas com diabetes e que este valor deverá ultrapassar os 700 milhões até 2045.[1] Portanto, existe uma necessidade de lidar com uma doença que cresce rapidamente através do uso das tecnologias disponíveis para prevenção e tratamento.

O uso de Machine Learning para diagnóstico de doenças é amplamente difundido pelo mundo e é uma importante

ferramenta nos esforços de prevenção de doenças.[2] No entanto, muitos estudos se baseiam em dados clínicos, que demandam maior esforço para serem obtidos. O que se traduz em uma necessidade de se abordar esse problema de outra forma, observando outras variáveis que estejam mais prontamente disponíveis. Com a disponibilidade de dados pessoais aumentando devido à soluções de Big Data que coletam esses dados associados à maior exposição de indivíduos, em redes sociais, por exemplo, o uso de dados relacionados ao estilo de vida se torna mais acessível. [4]

Este trabalho propõe avaliar a eficácia do uso de modelos de classificação de aprendizado de máquina para prever casos de diabetes a partir de características do estilo de vida da população.

## 2. DESCRIÇÃO DA BASE DE DADOS

A base de dados utilizada é a Pesquisa Nacional de Saúde (PNS) 2019. Esta pesquisa é realizada pela Fiocruz. Esta pesquisa consiste em uma pesquisa domiciliar em que a amostra é escolhida de acordo a representar os segmentos da população brasileira. Os dados pertinentes a este estudo são pertencentes à parte 4 do PNS, chamado de “Questionário do Morador Selecionado (Para pessoas de 15 anos ou mais de idade)”, e consiste dos seguintes módulos do PNS:

Módulo M - Características do trabalho e apoio social

Módulo N - Percepção do estado de saúde

Módulo P - Estilos de vida

Módulo Q - Doenças crônicas

Outros dados serão incorporados do “Módulo C - Características gerais dos moradores,” que contém informações mais genéricas dos entrevistados, como sexo e idade. A base de dados dispõe apenas dos dados fornecidos pelos respondentes. Portanto, utilizaremos os dados dos entrevistados que realizaram a entrevista completa. Dessa forma, podemos obter dados mais consistentes. Pois, estes dados já foram tratados e verificados pela Fiocruz, além do fato de que uma base de dados com poucos dados faltantes permite uma análise e modelagem mais efetivas.

O subconjunto de dados utilizado, que fora mencionado anteriormente, consiste de 90846 registros. (PNS 2019 Tabela 6) Foram selecionadas 23 variáveis preditoras, relacionadas ao estilo de vida, sintomas que podem estar

relacionados à diabetes e algumas características físicas e uma variável dependente, que é o diagnóstico de diabetes realizado por um médico. A variável resposta apresenta a característica de representar se o entrevistado recebeu um diagnóstico da doença diabetes. Isso significa que a pessoa pode não ter diabetes de fato, por um conjunto de fatores, como um diagnóstico incorreto ou um erro na resposta do entrevistado. As variáveis independentes, são todas observadas da resposta do entrevistado, o que pode gerar erros derivados da interpretação da pergunta ou de uma percepção incorreta do entrevistado. Todas as variáveis utilizadas foram aferidas através de método entrevista e assumidas como corretas, exceto peso e altura, que foram medidas com balança eletrônica e estadiômetro portáteis. (PNS 2019 p.25) Algumas variáveis auxiliares foram utilizadas para corrigir o indicador de diagnóstico da diabetes, foram deduzidos os diagnósticos realizados durante a gestação. Isto é, classificações positivas se tornaram negativas quando associadas à gestação. Pois, o diagnóstico de diabetes deste estudo não inclui diabetes gestacional. (PNS 2019 p.59) Esta correção permite que uma variação da diabetes que surge em condições muito específicas, gestação, não influencie na capacidade do modelo de captar informações que levem à classificação correta do diagnóstico.

O conjunto de dados, quando dividido pelo diagnóstico referido de diabetes, cria dois subconjuntos desbalanceados. Isso cria a necessidade de balancear os conjuntos de forma artificial ou de ponderar o tamanho dos subconjuntos. A Tabela 2 contém os tamanhos de cada classe após a remoção de outliers e outras entradas indesejadas para o ajuste dos modelos. As medidas tomadas para contornar o problema do desbalanceamento serão discutidas adiante, na seção 3.2.

## 2.1 Pré processamento dos dados

Os dados foram tratados utilizando técnicas comuns de processamento de dados para machine learning. Como as variáveis são uma combinação de variáveis contínuas e categóricas. Para as variáveis contínuas, aplicou-se o método de normalização L2, que transforma os valores para frações que, quando somadas totalizam 1 ao dividir os valores do conjunto pela norma deles. Nas variáveis categóricas foi aplicado o One-Hot Encoding, que cria variáveis ‘dummy’ binárias que quando combinadas representam a presença de um valor como a combinação destas variáveis binárias.

A base de dados apresentou dados faltantes apenas na classificação de diagnóstico referido de diabetes e nas aferições de peso e altura, estas foram preenchidas através da imputação pela média, que foi considerada como uma boa métrica para este tipo de variáveis, que são normalmente distribuídas e que apresentavam 0,9% de ausência em relação ao tamanho do dataset. As entradas em que não havia valor indicando se o entrevistado havia recebido ou não o diagnóstico de diabetes foram removidas. Pois se trata da ausência da própria variável resposta. A Tabela 1 contém as variáveis, sua contagem e proporção em relação ao tamanho do dataset.

A correlação das variáveis foi estudada, e verificamos que todas as variáveis têm baixa correlação. Isto é, estão entre (-0,5; 0,5), conforme podemos observar na Imagem 1. As

variáveis de peso e altura foram utilizadas para calcular o IMC, uma métrica que é capaz de unir o significado das duas variáveis anteriores, o nível de nutrição, e reduzir a dimensionalidade do modelo.

Na Tabela 3, estão dispostas as variáveis que foram utilizadas nos modelos e seus intervalos observados. Note que as variáveis que indicam a frequência semanal foram consideradas como numéricas e as outras variáveis que representam frequências foram tratadas como variáveis categóricas, devido à forma como foram registradas no PNS 2019.

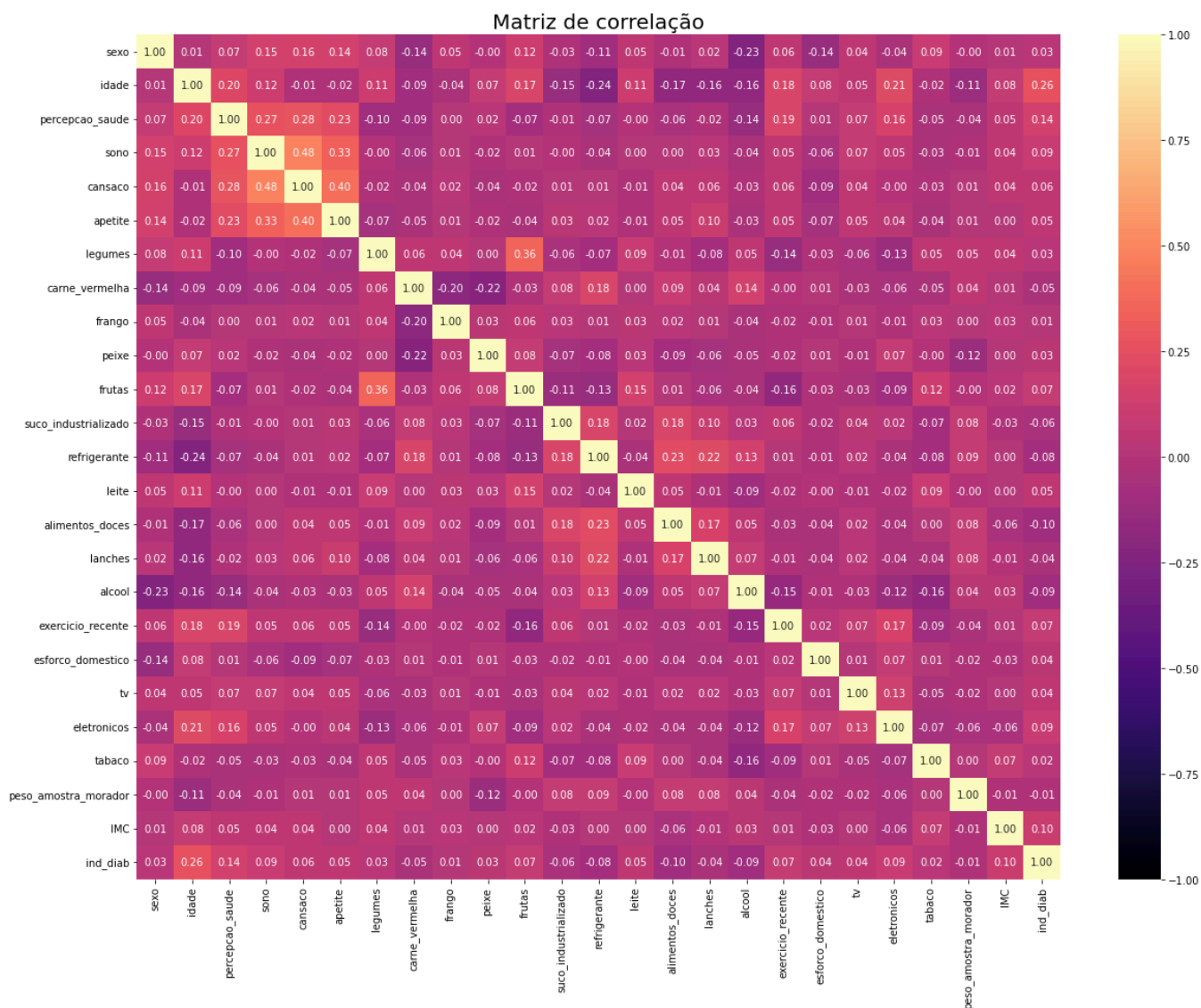
A remoção de outliers se deu de duas formas. O segundo, foi considerando o z-score do IMC e idade dos indivíduos, ambos com tolerância de 3 desvios padrões da média, de forma a remover dados muito díspares. Dessa forma o número de registros que foram utilizados de fato no ajuste do modelo é **83273**.

**Tabela 1- Variáveis com dados faltantes**

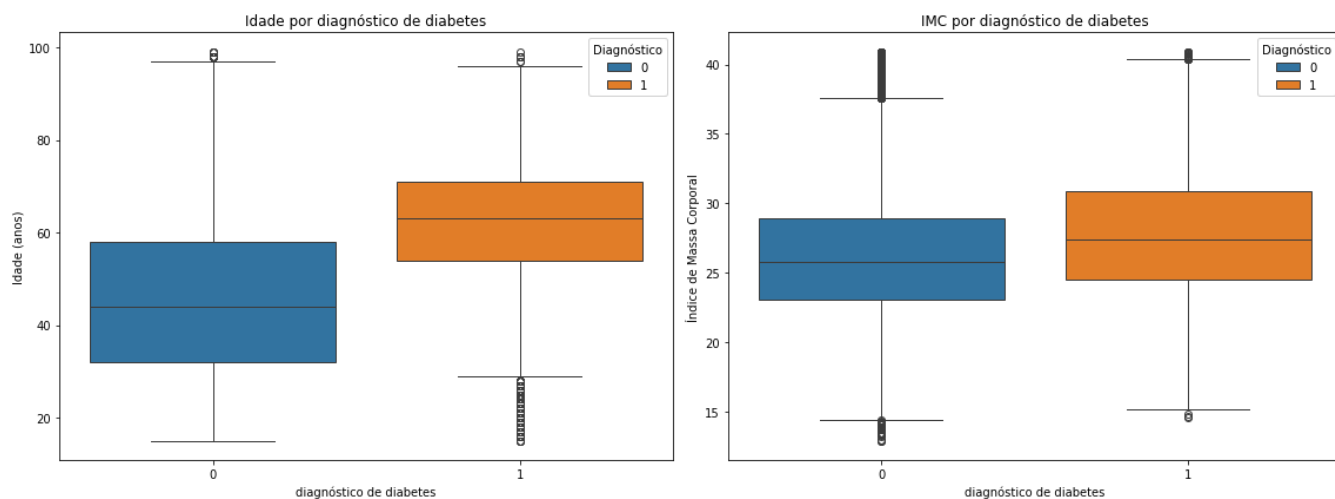
Variável	Número de dados faltantes	Proporção
‘peso’	892	0,98%
‘altura’	892	0,98%
‘diagnostico’	6773	7,46%

**Tabela 2 - Proporção entre classes de diagnóstico**

Classe	Número de registros
Classe 0 (diagnóstico negativo)	76699
Classe 1 (diagnóstico positivo)	7374



**Figura 1 - Matriz de correlação entre as variáveis.**



**Figura 2 - Distribuição de Idade e IMC.**

**Tabela 3 - Variáveis e seus intervalos**

Atributo	Descrição	Valores
‘sexo’	Sexo	1 - Homem 2 - Mulher
‘idade’	Idade (anos)	15 - 130
‘percepcao_saude’	Observação pessoal da própria saúde em geral	1 - Muito boa 2 - Boa 3 - Regular 4 - Ruim 5 - Muito ruim
‘IMC’	Índice de Massa Corporal (kg/m <sup>2</sup> )	1 - 100
‘sono’	Problemas no sono (falta/excesso) nas últimas 2 semanas	1 - Nenhum dia 2 - Menos da metade dos dias 3 - Mais da metade dos dias 4 - Quase todos dias
‘cansaco’	Cansaço em excesso nas 2 últimas semanas	1 - Nenhum dia 2 - Menos da metade dos dias 3 - Mais da metade dos dias 4 - Quase todos dias 9 - Ignorado
‘apetite’	Problemas de alimentação (falta/excesso) nas 2 últimas semanas	1 - Nenhum dia 2 - Menos da metade dos dias 3 - Mais da metade dos dias 4 - Quase todos dias
‘legumes’	Quantas vezes por semana come legumes?	0 - 7
‘carne_vermelha’	Quantas vezes por semana come carne vermelha?	0 - 7
‘frango’	Quantas vezes por semana come carne de ave?	0 - 7
‘suco_ind’	Quantas vezes toma suco industrializado?	0 - 7
‘fruta’	Quantas vezes na semana come frutas?	0 - 7
‘refrigerante’	Quantas vezes na semana toma refrigerante?	0 - 7
‘leite’	Quantas vezes na semana toma leite?	0 - 7

‘alimentos_doces’	Quantas vezes por semana come doces?	0 - 7
‘lanches’	Quantas vezes por semana substitui uma refeição por lanches rápidos?	Entre 0 e 7
‘alcool’	Frequência mensal de consumo de álcool	1 - Não bebo nunca 2 - Menos de uma vez por mês 3 - Uma vez ou mais por mês
‘exercicio’	Quantas vezes por semana pratica esportes?	Entre 0 e 7
‘esforco’	Faz esforço físico intenso em casa?	1 - Sim 2 - Não
‘tv’	Uso diário de TV	1 - Menos de uma hora 2 - De uma hora a menos de duas horas 3 - De duas horas a menos de três horas 4 - De três horas a menos de seis horas 5 - Seis horas ou mais 6 - Não assiste televisão
‘eletronicos’	Uso diário de PC, Tablet ou celular	1 - Menos de uma hora 2 - De uma hora a menos de duas horas 3 - De duas horas a menos de três horas 4 - De três horas a menos de seis horas 5 - Seis horas ou mais 6 - Não usa eletrônicos
‘tabaco’	Frequência de uso de tabaco	1 - Sim, diariamente 2 - Sim, menos que diariamente 3 - Não fumo atualmente
‘exame_sangue’	Exame de sangue mais recente	1 - Menos de 6 meses 2 - Entre 6 meses e 1 ano 3 - Entre 1 ano e 2 anos 4 - Entre 2 anos e 3 anos 5 - 3 anos ou mais 6 - Nunca fez
‘diagnostico’	Foi diagnosticado com diabetes?	1 - Sim 2 - Não

### 3. RESULTADOS E DISCUSSÕES

Nesta seção discutimos os métodos e técnicas utilizados para modelar a previsão de diabetes através do diagnóstico referido de diabetes e indicadores do estilo de vida da população.

#### 3.1. Modelos

Os modelos utilizados são todos modelos de classificação e se diferem pelo método pelo qual realizam a atribuição das classes. Os modelos selecionados foram:

- Gaussian Naive Bayes
- K Nearest Neighbors
- Random Forest Classifier
- Gradient Boosting Classifier
- eXtreme Gradient Boosting Classifier
- Support Vector Machine Classifier
- Multi-Layer Perceptron Classifier

A seguir, uma breve explicação de cada modelo.

Naive Bayes é um modelo que calcula a probabilidade de ocorrência de cada classe dado os preditores que são associados à cada classe. Para classificar novos dados o mesmo cálculo é aplicado. Esse modelo assume independência entre as variáveis, por isso o nome de “Naive” (ingênuo). Este é um modelo simples que tem dificuldades em modelar relacionamentos complexos entre os dados. Utilizamos a versão gaussiana do modelo para ajuste.

K Nearest Neighbors utiliza um cálculo de distância entre os valores dos dados e os agrupa conforme proximidade utilizando os K mais próximos. No caso da classificação, a classe mais presente dentre os K mais próximos é a associada com o registro a ser classificado. Este modelo é propenso a capturar features menos importantes devido a forma pela qual associa as classes quando as variáveis não têm pesos adequados. Neste estudo, foi utilizado K=9 para o modelo.

Random Forest Classifier cria divisões no conjunto de dados de forma que as separações são baseadas nas probabilidades de um registro pertencer a um dos subconjuntos criados nas partições feitas pelo algoritmo. Este modelo tem pouca sensibilidade à correlação, mas é comum fazer sobreajustes nos dados. Utilizamos diversos hiperparâmetros para gerar um bom modelo, com 200 árvores e 100 níveis oferecendo o melhor resultado. O método de grid search foi utilizado para buscar os melhores hiperparâmetros.

Gradient Boosting Classifier é um modelo mais complexo, que combina o ajuste de vários modelos simples para ajustar os seus parâmetros e obter um modelo mais robusto. Esse modelo utiliza a minimização dos erros dos ajustes como forma de otimizar os parâmetros. Por combinar modelos mais simples para obter uma modelagem através da otimização desses modelos, o Gradient Boosting têm melhor desempenho em geral e é menos propenso à vieses e sobreajuste dos modelos. No entanto, é mais custoso computacionalmente. Este modelo utilizou os seguintes hiperparâmetros: taxa de aprendizado = 0.5, nível máximo de árvores = 2 e 100 árvores de decisão, função logística e erro log-loss.

O eXtreme Gradient Boosting Classifier é uma versão melhorada do Gradient Boosting mencionado acima. Ele utiliza técnicas como “poda de árvores”, normalização e subamostragem para refinar o desempenho do Gradient Boosting, evitando o overfitting e melhorando o desempenho

computacional, foram utilizados os mesmos hiperparâmetros do Gradient Boosting.

O Support Vector Machine Classifier busca a melhor separação entre as classes ao encontrar o hiperplano que tem a maior margem de diferença entre essas classes. Tem capacidade de modelar relações complexas sem overfitting. Os hiperparâmetros utilizados foram: núcleo como função de base radial e sem fatores de regularização.

O Multi-Layer Perceptron é um tipo de rede neural artificial que minimiza o erro gerado durante as iterações do algoritmo ao recalculas as funções de ativação e de erro. Esse modelo é capaz de obter resultados robustos, sendo capaz de modelar relacionamentos complexos entre as variáveis. No entanto é custoso e de difícil interpretação. Os hiperparâmetros utilizados foram: 2 camadas ocultas com 64 e 32 nós, respectivamente, resolvidor ‘Adam’, ativação linear retificada e taxa de aprendizado de 0,1%.

#### 3.2. Amostragem e Reamostragem

Nesta subseção destacamos os métodos de amostragem e reamostragem utilizados para fazer o ajuste dos modelos. Considere que os métodos apresentados são os que foram utilizados para todos os modelos e são a referência. Outras técnicas foram utilizadas em um primeiro momento mas foram desconsideradas devido à baixa performance.

Os dados foram divididos em subconjuntos de treino e teste segundo a proporção 70-30. Como existe um forte desbalanceamento entre as classes 0 (negativo) e 1 (positivo), o uso de técnicas de reamostragem foram utilizadas para que os modelos fossem capazes de distinguir melhor entre as classes. O uso do modelo de Reamostragem Sintética Adaptativa (ADASYN) foi utilizado para criar uma sobreamostra da classe 1. Ao sobrerepresentar a classe com menos registros, os modelos conseguem incorporar melhor as características relevantes dos preditores da classe.

#### 3.3. Resultados

Os resultados encontrados para os 7 modelos são pouco robustos. Embora a precisão de alguns modelos seja alta, isso se deve ao fato de que há desbalanceamento das classes. Isto é, os diagnósticos negativos são muito mais presentes no conjunto que os diagnósticos positivos. O uso de sobreamostragem, discutido no subtópico anterior, consegue elevar a capacidade dos modelos de prever a classe sub representada. Sem o uso de sobreamostragem ADASYN, ou com outros métodos de balanceamento, o desempenho dos modelos ao tentar prever a classe 1 é muito baixo e os valores altos de acurácia derivam desse desbalanceamento. Como métrica de avaliação para a captura do relacionamento entre as variáveis e a distinção de classes, utilizamos a Curva Característica de Operação do Receptor (ROC) que compara a proporção de classificações corretas de verdadeiros positivos com os falsos positivos classificados incorretamente. Ou seja, a proporção entre sensibilidade (recall) e o complemento da especificidade. Quanto mais próximo a 100%, melhor o desempenho do modelo segundo esta métrica.

O melhor modelo, em geral, é o Support Vector Machine, que obteve 83,86% de acurácia e 64,39% de ROC. O Naive Bayes Gaussiano obteve 72,02% de ROC e 66,99% de acurácia. Isso indica que ele foi capaz de fazer uma distinção entre as classes, mas com pouca acurácia geral. Pois, a precisão

da classe 1 é de apenas 17%. No entanto, a sensibilidade da classe 1 foi a mais alta dentre todos os modelos com 78%. O extreme gradient boosting obteve a melhor acurácia geral, com 89,42%, mas um baixo ROC, de 58,36%. Isso indica que o modelo não conseguiu perceber uma distinção entre a classe 0 e a classe 1. Embora tenha uma precisão e revocação para a classe 0 de 93% e 94%, respectivamente. Isso indica que os dados utilizados são bons preditores da classe 0 mas não da classe 1, mesmo com o uso de sobreamostragem da classe 1. Os outros modelos tiveram desempenho semelhante ao do Extreme Gradient Boosting. Observamos que em todos os modelos, houve uma precisão alta para a classe 0 e baixa previsão para a classe 1. Modelos com valores de precisão mais baixos, obtiveram melhores valores de revocação para a classe 1. Modelos como Naive Bayes e K Neighbors produzem um elevado número de Falsos Positivos (FP). Isso indica que estes modelos priorizam a classificação de registros com a Classe 1. Indicando maior revocação da Classe 1. Os demais modelos apresentam valores de Falsos Positivos e Falsos Negativos em proporções mais baixas em relação ao tamanho do conjunto. O que indica que estes modelos são otimizados buscando a acurácia e precisão da classe 0, que é mais distinguível para estes modelos. Pois, como a classe 0 é predominante, prevê-la corretamente melhora o desempenho geral do modelo às custas de não capturar com clareza a classe 1.

As métricas de desempenho dos modelos utilizados neste trabalho se encontram na Tabela 5. A Tabela 4 apresenta os valores das matrizes de confusão para cada modelo.

### 3.4. Discussão

Observamos que o desempenho mediano dos modelos independente do uso de técnicas que mitiguem os problemas de desbalanceamento, multicolinearidade e correlacionamento das variáveis pode ser decorrente do fato de que a variável resposta, que é o diagnóstico referido de diabetes do entrevistado, tenha influência nas variáveis preditoras. De forma que, quando uma pessoa é diagnosticada, ela passa a ter hábitos mais saudáveis, condizentes com o tratamento da doença. Ademais, pessoas que já apresentam hábitos saudáveis têm menor predisposição a desenvolver diabetes. Conseguimos obter modelagens que captam até certo ponto o relacionamento entre diabetes e estilo de vida, pois diversas variáveis preditoras são relacionadas à comportamentos pregressos ou contínuos, que influenciam na saúde dos entrevistados, como o uso de tabaco e álcool. O teste de glicose, entre outros hábitos, como evitar doces e refrigerantes, pode ser um indicativo da presença da doença. Pois indica um comportamento relacionado à medidas de tratamento não medicamentoso da diabetes e de pré-diabetes.

[5]

**Tabela 4 - Matrizes de Confusão de cada modelo**

Modelo	Valor previsto	Valor Verdadeiro	
		Classe 0	Classe 1
Gaussian Naive Bayes	Classe 0	15125 (TN)	7794 (FP)
	Classe 1	453 (FN)	1610 (TP)
K Neighbors Classifier (K=9)	Classe 0	14755 (TN)	8164 (FP)
	Classe 1	727 (FN)	1336 (TP)
Random Forest Classifier	Classe 0	21541 (TN)	1378 (FP)
	Classe 1	1508 (FN)	555 (TP)
Gradient Boosting Classifier	Classe 0	21486 (TN)	1433 (FP)
	Classe 1	1499 (FN)	564 (TP)
eXtreme Gradient Boosting Classifier	Classe 0	21896 (TN)	1023 (FP)
	Classe 1	1619 (FN)	444 (TP)
Support Vector Machine Classifier	Classe 0	20105 (TN)	2814 (FP)
	Classe 1	1216 (FN)	847 (TP)
Multi-Layer Perceptron Classifier	Classe 0	19061 (TN)	3858 (FP)
	Classe 1	1170 (FN)	893 (TP)

Tabela 5 - Métricas de desempenho dos modelos ajustados.

Modelo	Acurácia	Precisão (0)	Recall (0)	F1-score (0)	Precisão (1)	Recall (1)	F1-score (1)	ROC AUC
Gaussian Naive Bayes	0.6699	0.97	0.66	0.79	0.17	0.78	0.28	0.7202
K Neighbors Classifier (K=9)	0.6441	0.95	0.69	0.80	0.15	0.60	0.24	0.6456
Random Forest Classifier	0.8826	0.93	0.94	0.94	0.28	0.27	0.28	0.6054
Gradient Boosting Classifier	0.8844	0.93	0.94	0.94	0.29	0.27	0.28	0.6044
eXtreme Gradient Boosting Classifier	0.8942	0.93	0.96	0.94	0.30	0.22	0.25	0.5853
Support Vector Machine Classifier	0.8387	0.94	0.88	0.91	0.23	0.41	0.30	0.6439
Multi-Layer Perceptron Classifier	0.7987	0.94	0.83	0.88	0.19	0.43	0.26	0.6323

### 3.5. Conclusão


O presente trabalho evidencia que existe valor em utilizar dados pessoais de comportamento diário para extrair informações relacionadas à detecção de doenças, especialmente a diabetes. No entanto, deve-se observar a natureza dos dados e o contexto em que foram gerados. Na base do PNS, indicadores de classificação da presença de doenças são obtidos conforme entrevista que busca observar o histórico do entrevistado segundo o histórico médico e hábitos referidos pelo respondente com critérios que possuem influência com ponderações enfraquecidas devido à referenciais adotados que buscam simplificar a coleta de informações. Um exemplo são os hábitos de consumo de alimentos com alto açúcar. Esta variável considera frequência semanal e, o sedentarismo, outra variável considerada relacionada ao desenvolvimento de diabetes tem referência trimestral. A Figura 3 evidencia a relação entre comportamentos saudáveis e diagnóstico de diabetes. Uma exceção é a prática de exercícios físicos. Isso ocorre devido ao fato de que a diabetes é mais comum em pessoas de mais idade e que, portanto, têm menos disposição para atividades físicas.

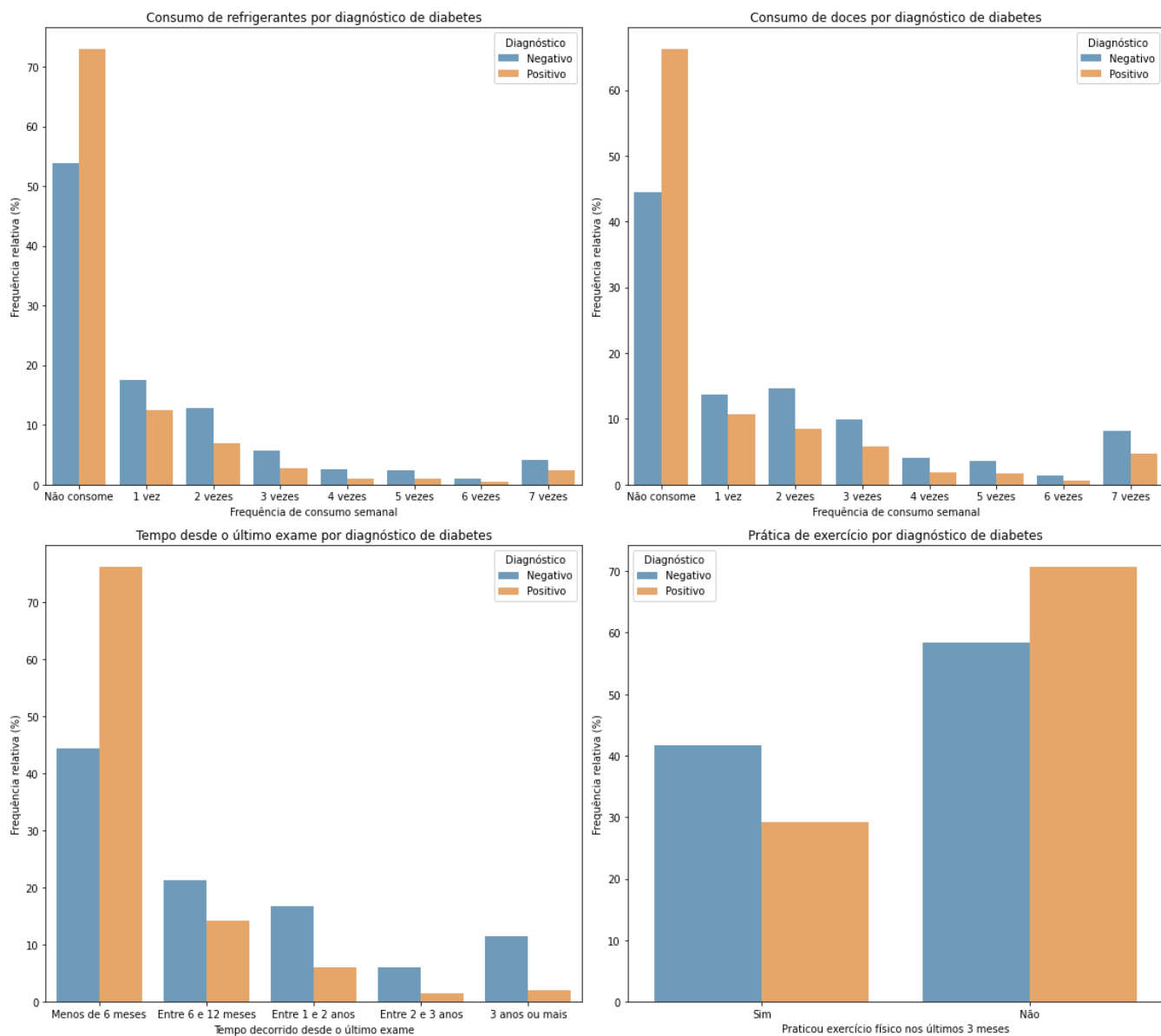
As observações feitas através da análise dos dados e do ajuste dos modelos evidencia que muito valor pode ser extraído da componente temporal dos dados. Dados que captam o comportamento recente do indivíduo modelam melhor situações em que uma pessoa tenha diabetes. Muitos dos dados que se relacionam à presença de diabetes decorrem da ciência do portador da doença. Isto é, as variáveis são condicionadas pela resposta que buscamos. Para contornar este problema, outras variáveis que modelam a presença de pré-diabetes devem ser incluídas no estudo para contribuir com a capacidade preditiva dos modelos.

## 4. REFERÊNCIAS

- Magliano, Dianna J., Edward J. Boyko, and IDF Diabetes Atlas. "What is diabetes?." *IDF DIABETES ATLAS [Internet]. 10th edition*. International Diabetes Federation, 2021. ISBN: 978-2-930229-98-0
- Chiavegatto Filho, Alexandre Dias Porto. "Uso de big data em saúde no Brasil: perspectivas para um futuro próximo." *Epidemiologia e Serviços de Saúde* 24 (2015): 325-332.
- Instituto Brasileiro de Geografia e Estatística (IBGE). Pesquisa Nacional de Saúde 2019. [acessado 2024 Maio 13]. Disponível em: <https://www.pns.icict.fiocruz.br/wp-content/uploads/2021/02/liv101764.pdf>
- Bormida, M.D. (2021), "The Big Data World: Benefits, Threats and Ethical Challenges", *Iphofen, R. and O'Mathúna, D. (Ed.) Ethical Issues in Covert, Security and Surveillance Research (Advances in Research Ethics and Integrity, Vol. 8)*, Emerald Publishing Limited, Leeds, pp. 71-91. <https://doi.org/10.1108/S2398-601820210000008007>
- Luciana Bahia, Bianca de Almeida-Pititto, Bertoluci M. Tratamento do diabetes mellitus tipo 2 no SUS. Diretriz Oficial da Sociedade Brasileira de Diabetes (2023). DOI: 10.29327/5238993.2023-11, ISBN: 978-85-5722-906-8.

## 5. CÓDIGO DESENVOLVIDO

 [trabalho\\_diabetes.ipynb](https://github.com/trabalho_diabetes.ipynb)



**Figura 3 - Frequências relativas de comportamentos agrupados por diagnóstico**