Graduação em Sistemas de Informação

Lucia Helena Aparecida Rissi Salvi

Aplicação do aprendizado de máquina na predição de diabetes mellitus tipo 2

São Caetano do Sul
2021

Graduação em Sistemas de Informação

Lucia Helena Aparecida Rissi Salvi

Aplicação do aprendizado de máquina na predição de diabetes mellitus tipo 2

Trabalho de Conclusão do Curso de Sistemas de Informação da Universidade Municipal de São Caetano do Sul como requisito à obtenção de título de bacharela em Sistemas de Informação.

São Caetano do Sul
2021

SALVI, Lucia Helena Aparecida Rissi. **Aplicação do aprendizado de máquina na predição de diabetes mellitus tipo 2**. Trabalho de Conclusão do Curso de Sistemas de Informação da Universidade Municipal de São Caetano do Sul apresentado como requisito à obtenção de título de bacharela em Sistemas de Informação. São Caetano do Sul, 2021.

Data da aprovação ____/_____/ 2021.

Banca Examinadora:

_____
Prof. Dr. Ivan Carlos Alcântara de Oliveira (Orientador).
Universidade Municipal de São Caetano do Sul

_____
Profa. Ma. Cilene Aparecida Mainente (Membro).
USCS – Universidade Municipal de São Caetano do Sul

_____
Prof. Dr. Cláudio Cura Júnior (Membro).
USCS – Universidade Municipal de São Caetano do Sul

Nada é pequeno, se feito com amor”
Santa Terezinha do Menino Jesus

DEDICATÓRIA

Como flores que são levadas ao altar, oferto este trabalho a Deus e à Nossa Senhora, pois o que tenho e o que sou é pela graça de Deus.

Dedico também, à minha família, apoiadores em todos os momentos mesmo quando eu não era útil, ou quando reaprendia a viver. Uma especial menção a meu esposo, Wagner, meu companheiro de vida.

AGRADECIMENTOS

À Universidade Municipal de São Caetano do Sul, lugar que escolhi para minha graduação e fui muito feliz nesta escolha. A meus professores, cada um plantou sua semente para frutificar o conhecimento.

Um especial agradecimento a meu orientador, Professor Ivan, por toda sua dedicação para o desenvolvimento deste trabalho, pois, ainda que sem plenos recursos, sempre incentivou o capricho em todos os detalhes.

RESUMO

A diabetes mellitus é a terceira maior causa de mortes no Brasil. Ocupam as primeiras posições as neoplasias (câncer) e as doenças cardiovasculares, com a ressalva de que ambas podem ser originadas pela diabetes. É fundamental diagnosticar precocemente para que o tratamento seja iniciado tão logo for possível e sejam afastadas eventuais complicações. Nesse sentido, este trabalho tem o objetivo de criar um modelo preditivo para diagnosticar diabetes mellitus a ser obtido pelo uso de aprendizado de máquina e base de dados pública, além da construção de uma aplicação analítica, denominada “Flocos”, como protótipo de software que possa ser utilizado para realizar predições de novos casos de diabetes. Para suprir a fonte de dados, foi pesquisada uma gama de conjuntos de dados aptos à utilização e selecionado os dados do programa *National Health and Nutrition Examination Survey* (NHANES) de estudos elaborados para avaliar a saúde e o estado nutricional de adultos e crianças norte-americanas. Realizados os estudos e experimentos com técnicas de aprendizado de máquina selecionadas para obter a mais assertiva na geração do modelo preditivo. Por fim, foi criada uma aplicação analítica fazendo uso do modelo preditivo obtido com a técnica selecionada *AdaBoosting* e acurácia de 94,15%.

Palavras-Chaves: Ciência de Dados; aprendizado de máquina; análise de dados; saúde; diabetes; aplicação analítica.

ABSTRACT

Diabetes mellitus is the third leading cause of death in Brazil, occupying the top positions as neoplasms (cancer) and cardiovascular diseases, with the exception that these can be caused by diabetes. It is essential to diagnose early so that treatment is started as soon as possible and possible complications are delayed. In this sense, this work aims to create a predictive model to diagnose diabetes mellitus supply by using machine processing and a public database, in addition to building an analytical application, called “Flocs”, as a software prototype that can be used to make predictions of new cases of diabetes. To supply the data source, a range of usable data sets were searched and data from the National Health and Nutrition Examination Survey (NHANES) program of studies designed to assess the health and nutritional status of adults were selected. and American children. Conducted studies and experiments with selected machine collection techniques to obtain the most assertive in the generation of the predictive model. Finally, an analytical application was made using the specified predictive model with the provided AdaBoosting technique and 94.15% accuracy.

Keywords: Data Science; Machine Learning; Data analysis; Health; Diabetes; Analytical Application.

ILUSTRAÇÕES

Quadros

Tabelas

Gráficos

Figuras

ABREVIATURAS E SIGLAS

AVC – Acidente vascular cerebral

Bagging – *Boosting Bootstrapped Aggregation*

CART – *Classification and Regression Tree*

CDC – Centro de Controle e Prevenção de Doenças

CHAID – *Chi-squared Automatic Interaction Detection*

DHANES – Divisão de Pesquisas de Exame de Saúde e Nutrição

DM1 – *Diabetes Mellitus* Tipo 1

DM2 – *Diabetes Mellitus* Tipo 2

GBM – *Gradient Boosting Machines*

GBRT – *Gradient Boosted Regression Tree*

ID3 – *Iterative Dichotomiser 3*

IDF – *International Diabetes Federation* – Federação Internacional de Diabetes

IMC – Índice de Massa Corporal

IoT – Internet das coisas

KNN – K – *Nearest Neighbors* – Algoritmo K – vizinho mais próximo

LADA – *Latent autoimmune diabetes in adults* – Diabetes latente autoimune do adulto

MODY – *Maturity-onset diabetes of the young* – Diabetes juvenil de início tardio

NCHS – Centro Nacional de Estatísticas de Saúde

NHANES – *National Health and Nutrition Examination* – Exame Nacional de Saúde e Nutrição

PEP – Prontuário eletrônico do paciente

RFE – Recursive Feature Elimination

SBD – Sociedade Brasileira de Diabetes

SCIELO – *Scientific Eletronic Library Online*

SEQN – Número sequencial

SOMP – Síndrome dos Ovários Micropolicístico

SOP – Síndrome dos Ovários Policísticos

SVM – Support Vector Machine

SVM – *Support Vector Machines* – Máquina de Vetor de Suporte

UML – *Unified Modeling Language*

UTI – Unidade de terapia intensiva

VN – Verdadeiro negativo

VP – Verdadeiro positivo

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

1.1.1 Contexto geral

A produção massiva de dados, consoante à variedade, volume e velocidade, tem a capacidade de gerar novas e mais complexas análises, como apontam Fernandes (et al. 2019), auxiliando “em processos de tomada de decisão, servir de ferramenta para a análise estatística e gerar conhecimento para subsidiar ações” que ofereçam oportunidades e amplas possibilidades de melhorias que afetem os vários campos da vida humana.

Por isso, em voga a mineração de dados em que, analogamente ao trabalho da mina, no qual se utiliza a peneira para extrair minérios e pedras preciosas, aqui se utilizam ferramentas para extrair dados relevantes do montante gerado diariamente. É preciso aproveitar os dados como recurso precioso, como matéria-prima alocada em diversos tipos de repositórios, transformá-los em conhecimento, em utilidade, o que se faz pela análise dos dados.

Amaral (2016) diz que “em computação, análise de dados é a aplicação de algum tipo de transformação nos dados em busca de conhecimento. O autor destaca ainda que, com o *Big Data*, a análise é realizada sobre grande concentração de dados. No campo da medicina e saúde, na hipótese de doenças, Ribeiro (2009) aponta que “métodos computacionais podem transformar eficientemente estes dados em informação útil para o diagnóstico das doenças, prevenção e controle”.

Em seu vídeo mensal componente da Rede Mundial de Oração do Papa, Francisco (2020) nos convida a refletir sobre o papel da Inteligência Artificial e a condução do desenvolvimento fundado na dignidade de pessoa humana, que também é princípio fundamental do direito brasileiro:

> A inteligência artificial está na raiz da mudança de época que estamos a viver. A robótica pode tornar possível um mundo melhor se estiver unida ao bem comum. Porque, se o progresso tecnológico aumenta as desigualdades, não é um progresso real.

> Os avanços futuros devem estar orientados para o respeito pela dignidade da pessoa e da Criação. Rezemos para que o progresso da robótica e da inteligência artificial esteja sempre a serviço do ser humano… podemos dizer, que “seja humano”. (FRANCISCO, 2020).

A inteligência artificial é a força motriz do desenvolvimento, assim como é inegável como ela é potencialmente benéfica. Mas, como é possível acrescentar sentido a esse desenvolvimento? Direcionando-o ao serviço do ser humano, por exemplo, à coleta “de informação no âmbito da saúde para se conseguir melhores diagnósticos e tratamentos”. (FRANCISCO, 2020).

Nesse contexto, um emprego viável é a geração de modelos preditivos, criados a partir da alimentação de dados históricos – por isso a importância do *Big Data* – particionados em dados de treino e dados de teste. O modelo preditivo, pode ser aplicado a novas instâncias, por exemplo, a novos dados de pacientes para gerar a classificação concernente ao diagnóstico de enfermidades.

1.1.2 Contexto da diabetes

É um cenário preocupante. O quadro 1 traz os dados epidemiológicos do diabetes no Brasil (SBD, 2019), referentes às maiores causas de mortes nos anos de 1990 e 2017 sinalizadas com cores: rosa, para doenças crônicas e não transmissíveis, como neoplasias; azul, para doenças transmissíveis, maternas, neonatais e nutricionais como enterite infecciosa; róseo, causas externas como acidentes automobilísticos.

Quadro 1. Posições das maiores causas de mortes no país, considerando as mortes por 100 mil habitantes e de ambos os sexos, nos marcos de 1990 e 2017.

1990			2017	
1	Doenças cardiovasculares		1	Doenças cardiovasculares
2	Neoplasias		2	Neoplasias
3	Afecções maternas e neonatais		3	Diabetes e doenças do rim
4	Infecções respiratórias e tuberculose		4	Infecções respiratórias e tuberculose
5	Autolesão e violência		5	Transtornos neurológicos
6	Acidente transporte		6	Autolesão e violência
7	Enterite infecciosa		7	Respiratórias crônicas
8	Respiratórias crônicas		9	Acidente automobilístico
11	Diabetes e doença do rim		12	Afecções maternas e neonatais
13	Transtornos neurológicos		17	Enterite infecciosa
Legenda:		Doenças crônicas não transmissíveis		
		Doenças transmissíveis, maternas, neonatais e nutricionais.		
		Causas externas		

Fonte: autoria própria com base em (SBD, 2019).

Em menos de 30 anos, a diabetes saltou da 11ª posição para a 3ª no ranking brasileiro da SBD (2019). Trata-se de doença que pode ser tipificada em diabetes mellitus tipo 1, *Latent autoimmune diabetes in adults* (LADA), *Maturity-onset diabetes of the young* (MODY), gestacional e diabetes mellitus tipo 2, que é alvo desta pesquisa.

A diabetes mellitus tipo 2, ou diabetes não insulinodependente, “é assintomática” (MILECH, 2017), “é o tipo mais comum de diabetes e é responsável por 90% dos casos de diabetes em todo o mundo.” (IDF, 2019, tradução nossa).

É preciso ressaltar que a diabetes mellitus tipo 2 é causa de complicações graves como doenças cardiovasculares, à exemplo do acidente vascular cerebral, que ocupa a primeira posição no período de avaliação do ranking.

1.2 MOTIVAÇÃO E JUSTIFICATIVA

O diagnóstico precoce se revela de fundamental importância porque, segundo Gross (2002), permite a adoção de medidas terapêuticas para tratamento e para evitar complicações que afetam a qualidade de vida pessoal e laboral do indivíduo.

Tanto é que “as intervenções precoces voltadas para a criação de mudanças no estilo de vida, com ou sem terapias farmacológicas associadas, têm se mostrado eficazes em retardar ou prevenir o diabetes tipo 2 e suas complicações” (OLIVERA, 2016). Pessoas podem ter seu curso de vida alterado e suas preciosas vidas poupadas.

Um modelo preditivo para diagnóstico de diabetes mellitus tipo 2 pode promover melhores e mais céleres diagnósticos.

Já se vislumbra a implementação do modelo que utilize o *dataset* de um hospital ou operadora de saúde. A alternativa é válida porquanto a base de dados de um hospital, por exemplo, contém número expressivo de registros e a predição é abastecida por grande quantidade de dados, o que torna o modelo melhor ajustado.

Assim como, ao aplicar a predição em suas bases de dados, operadoras de plano de saúde ou hospitais poderão indicar ao paciente a melhor conduta para reversão do quadro, visto que a patologia não tem cura, mas tem remissão e a qualidade de vida do paciente pode ser imensamente melhorada com o tratamento correto e precoce.

A conduta médica adequada infere a diminuição das ocorrências de complicações e, com elas, redução dos custos operacionais ao evitar internações advindas das gravosas complicações da diabetes mellitus tipo 2 não tratada. Por exemplo: em caso de cetoacidose diabética complicada com instabilidade hemodinâmica, que é advinda de complicação aguda, faz-se necessária a internação em unidade de terapia intensiva (UTI).

Certamente, a diária de um quarto de UTI é maior do que a de um quarto padrão, pela disponibilidade de recursos humanos (médicos, enfermeiras), aparelhos e monitoramento contínuo do paciente.

Os resultados da predição podem ser gravados no prontuário do paciente e disponibilizados na próxima consulta ou direcionados à equipe especializada. A análise das informações emergentes da predição pode servir de aporte para decisões como a criação de um programa de saúde voltado para a patologia.

1.2.1 DM2 infanto-juvenil

A diabetes mellitus tipo 2 alcança, inclusive, crianças e adolescentes, que “correm o risco de complicações nos primeiros anos de vida adulta, o que tem impacto significativo no indivíduo, na família e na sociedade” (IDF, 2019, tradução nossa).

Considerando as mudanças por que a sociedade passa desde os anos 1920, a crescente população urbana e o êxodo rural, a troca da alimentação doméstica e balanceada para a adoção dos estilos alimentares *fast-food* e alimentos ultra processados com excesso de gorduras, sal e açúcar associada ao sedentarismo, desde a tenra idade, elas podem contribuir para o indicativo de aumentos da DM2 em crianças e adolescentes, doença de ocorrência predominante em adultos.

O indicativo é um alarme estridente para que sejam tomadas medidas para a contenção da epidemia que é a diabetes mellitus tipo 2, associada ao estilo de vida. Estes sinais precisam ser revertidos por ações das famílias ou dos tutores e curadores, escolas, assim como de políticas públicas direcionadas para tanto, que podem se valer da atividade preditiva.

Dada sua notoriedade, para melhor ilustrar o contexto da diabetes, são elencados números publicados na edição 2019 do Atlas da Diabetes da *International Diabetes Federation* (IDF) e dados epidemiológicos da Sociedade Brasileira de Diabetes (SBD) e que formam um retrato da diabetes em sentido amplo.

1.2.2 Panorama global da diabetes

Segundo a IDF (2019, tradução nossa), “estima-se que haverá 4,2 milhões de mortes por diabetes e suas complicações durante 2019.” E “se a tendência continuar, 700 milhões de adultos terão diabetes até 2045. Os maiores aumentos ocorrerão onde as economias passarem da situação de renda baixa para média.” Esses dados indicam a curva crescente no número de diabéticos em todo o mundo e que “quase metade (46,2%) de mortes associadas à diabetes entre o grupo entre 20 e 79 anos ocorrem em pessoas abaixo de 60, ou seja, a faixa etária ativa.” (IDF, 2019, tradução nossa).

Enquanto os dados de 2018 do Banco Mundial (WORLD BANK, 2019) inferem o aumento da expectativa de vida, como exemplo, o Brasil, superior a 75 anos; Estados Unidos, superior a 78 anos; e Espanha, superior a 83 anos, as estimativas do panorama global da diabetes poderão mitigar esse aumento, com a diminuição da expectativa de vida e aumento dos índices de morte precoce causadas pela epidemia de diabetes.

Revela-se um impacto econômico negativo referido como “custos indiretos” que chegam ao total de US$ 90 bilhões devido ao diabetes, incluídos o prejuízo que mortes prematuras causam à economia norte-americana. (IDF, 2019, tradução nossa).

Notadamente, é fato que atinge o indivíduo diretamente – que precisa do tratamento e lidar com a patologia na esfera pessoal e laboral, mas que extrapola

esse limite e compromete estruturas sociais, referenciando nos custos indiretos, a economia.

1.2.3 Panorama da diabetes na América Central, América do Sul e Brasil

Em se tratando da América Central e Sul, “a prevalência de diabetes não diagnosticada na faixa etária de 20 a 79 anos é de 41,9%, o que equivale a 13,3 milhões de pessoas.” (IDF, 2019, tradução e grifo nossos), confirmando o panorama global de números crescentes da incidência da doença e que a falta de diagnóstico leva à falta de tratamento adequado.

No Brasil, em 2019, “o número de adultos com diabetes (20 – 79 anos) é de 16.780,8 milhões, que representa 11,4% da população” (IDF, 2019, tradução nossa).

Tendo por base a população das cidades brasileiras com mais de 1 milhão de habitantes, a partir da estimativa do IBGE (2020), que apresenta os números de algumas das 17 cidades mais populosas do Brasil. No presente, considerando a estimativa populacional das cidades de São Paulo, Belo Horizonte e Curitiba, conforme a tabela 1, fica mais perceptível que a soma resulta no número equivalente de diabéticos no Brasil, citado no parágrafo anterior.

Tabela 1. Estimativa populacional nas cidades de São Paulo, Curitiba e Belo Horizonte.

Ordem	UF	Cidade	População	Total
1º	SP	São Paulo	12.325.232	
4º	MG	Belo Horizonte	2.521.564	14.846.796
6º	PR	Curitiba	1.948.626	16.795.422

Fonte: autoria própria baseado em IBGE (2020).

A projeção publicada pelo Atlas de Diabetes é de que, em 2045, a população com diabetes seja de 26 milhões de pessoas, fazendo que o Brasil ocupe o 5º lugar dentre os 10 primeiros países ou territórios por quantidade de adultos (de 20 a 79 anos) com diabetes em 2019, 2030 e 2045. A tabela 2 apresenta os números desta projeção em referência ao Brasil.

Tabela 2. Prognóstico por quantidade de adultos (de 20 a 79 anos) com diabetes em 2019, 2030 e 2045 no Brasil.

Ano	Incidência	Intervalos de confiança de 95%
2019	16,8 milhões	(15,0 – 18,7)
2030	21,5 milhões	(19,3 – 24,0)
2045	26 milhões	(23,2 – 28,7)

Fonte: autoria própria baseado em IDF (2019, tradução nossa).

1.2.4 Impactos na previdência social brasileira

As faixas etárias até os 70 anos referem-se à população economicamente ativa, a que está inserida no mercado de trabalho para o provisionamento próprio e família, recolhendo contribuições que sustentam o sistema previdenciário.

Através do Instituto Nacional de Seguridade Social (INSS), a Previdência Social é a concedente do auxílio-doença, destinado a suprir a renda do trabalhador segurado ou em período de graça quando, por mais de 15 dias, ficar incapacitado para seu trabalho ou para sua atividade habitual que provisione sua renda de acordo com o artigo 59 da Lei 8.213, de 24 de julho de 1991 (BRASIL, 1991).

No caso de incapacidade decorrente do acidente de trabalho, é concedido o auxílio-doença acidentário (B91) e, quando não há nexo causal, o auxílio-doença previdenciário (B31).

Jakobi (et al. 2013), observam que as maiores ocorrências são de patologias osteomusculares, a exemplo de dorsopatias (37,7) e transtornos dos tecidos moles e músculos (17,0), parecer confirmado nos dados da Previdência Social.

No entanto, o relatório de acompanhamento mensal do benefício de auxílio-doença previdenciário concedido segundo os Códigos da Classificação Internacional de Doenças CID – 10, de janeiro a dezembro de 2019 (BRASIL, 2020) denota uma porcentagem baixa em relação ao total, mas presente e crescente, pode-se ver na tabela 3, que aponta a quantidade de benefícios de auxílio-doença concedidos, mês a mês, no ano de 2019.

Tabela 3. Benefícios concedidos de acordo com CID relacionados à diabetes, de janeiro a dezembro de 2019.

CID -10	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
E10	433	524	497	489	519	447	546	621	483	530	430	388
E11	226	296	252	243	251	240	230	246	218	257	182	159
E14	113	184	139	134	180	154	154	145	139	147	127	109
Total	**772**	**1004**	**888**	**866**	**950**	**841**	**930**	**1012**	**840**	**934**	**739**	**656**
E10	Diabetes mellitus insulino dependente											
E11	Diabetes mellitus não insulino dependente											
E14	Diabetes mellitus não especificado											

Fonte: autoria própria com base em BRASIL (2020).

Ainda que temporária, a diabetes tem sido causa de afastamento do trabalho ou da atividade de que o segurado aufira renda e possa prover o sustento próprio e da família.

O relatório de acompanhamento (BRASIL, 2018) aponta o número dos benefícios concedidos em 2016, elencados na tabela 4, aponta o número menor de benefícios pagos pelo INSS o que denota seu crescimento em relação à tabela anterior.

Tabela 4. Benefícios concedidos de acordo com CID relacionados à diabetes, de janeiro a dezembro de 2016.

CID -10	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
E10	0	0	0	1	1	0	0	0	0	1	0	1
E11	1	0	0	0	1	1	0	1	1	1	1	0
E14	0	0	0	0	0	0	1	1	0	1	0	0
Total	0	0	0	1	1	0	0	0	0	1	0	1
E10	Diabetes mellitus insulino dependente											
E11	Diabetes mellitus não insulino dependente											
E14	Diabetes mellitus não especificado											

Fonte: autoria própria com base em BRASIL (2018).

Talvez a constatação mais inesperada surge da verificação do relatório de acompanhamento de benefícios acidentários em 2019, na tabela 5, com causa referente à diabetes.

Tabela 5. Benefícios concedidos de acordo com CID relacionados à diabetes, de janeiro a dezembro de 2019.

CID -10	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
E10	1	4	1	5	1	4	3	4	5	5	5	4
E11	1	2	3	2	1	5	2	3	1	2	5	2
E14	2	1	1	1	0	2	1	2	0	3	1	1
Total	1	4	1	5	1	4	3	4	5	5	5	4
E10	Diabetes mellitus insulino dependente											
E11	Diabetes mellitus não insulino dependente											
E14	Diabetes mellitus não especificado											

Fonte: autoria própria com base em BRASIL (2020).

Para a concessão, o benefício acidentário (B91) prescinde do nexo causal entre a doença e o trabalho. Significa que a incapacidade causada por diabetes – nos códigos referenciados – guarda nexo causal com o trabalho ou acidente de trabalho.

1.3 OBJETIVO GERAL

A construção de uma aplicação analítica na forma de protótipo de software para predição de diabetes mellitus tipo 2 fazendo uso de repositório de dados públicos e da técnica de aprendizado de máquina mais bem avaliada na geração do seu modelo preditivo.

Para isso, foi feita uma investigação e análise das bases de dados em repositórios públicos, estudo e investigação das técnicas de aprendizado de máquina nas pesquisas bibliográficas, além de testes de avaliação com a finalidade de identificar aquelas com potencial para criar um modelo preditivo com melhor assertividade.

1.4 QUESTÕES DE PESQUISA

Foram elaboradas duas questões de pesquisas que nortearam todo o processo deste trabalho e cujas respostas têm o potencial de alcançar o proposto no objetivo geral, sendo elas.

- Qual a técnica de aprendizado de máquina é a mais assertiva na realização de diagnósticos preventivos de diabetes mellitus tipo 2?
- Qual melhor conjunto de atributos para treinar e gerar um modelo preditivo com boa precisão envolvendo o diagnóstico preventivo de diabetes mellitus tipo 2?

1.5 MÉTODO DO TRABALHO

A pesquisa está concentrada em dois âmbitos principais: na primeira fase, pesquisa não experimental consolidada na coleta de informações em livros e artigos científicos, obtidas em bases de trabalhos acadêmicos reconhecidos, acrescida de artefatos de desenvolvimento de software (algoritmos, diagramas, entre outros); a análise de repositórios públicos, investigação das técnicas de geração de modelo preditivo assim como a identificação dos atributos que sejam relevantes para gerar um modelo preditivo genérico.

Por conseguinte, foi eleita a melhor técnica, por meio da experimentação dos algoritmos na base escolhida. Além disso, foram realizados testes, experimentos e a aplicação de técnicas para melhoria do desempenho.

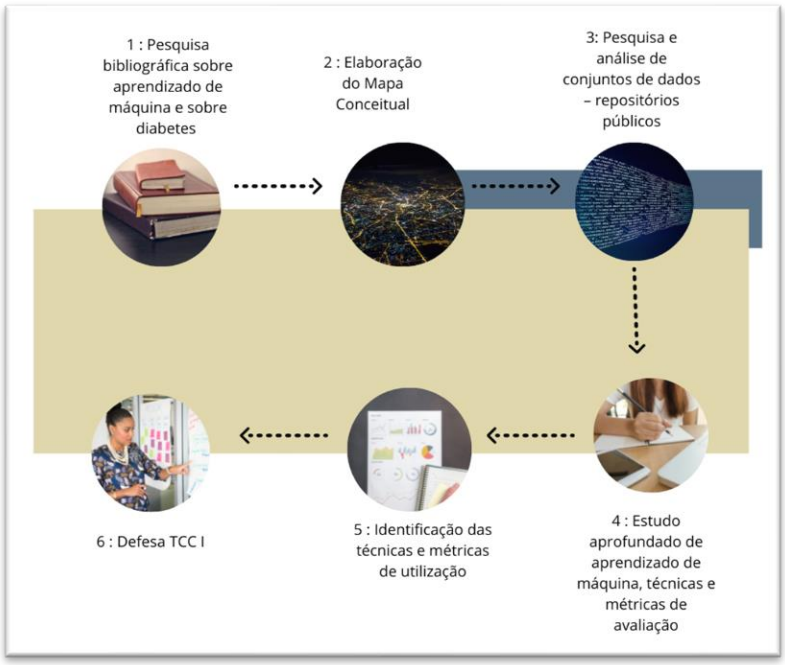
No curso desta pesquisa, foi possível oferecer respostas para as questões de pesquisa que foram tomadas como base para a construção da aplicação analítica, um protótipo, que é a representação visual dos resultados da investigação.

A aplicação teve como entrada os dados de um paciente, segundo os atributos selecionados de um *dataset.* A partir desse conjunto de dados, foi gerado um modelo preditivo, utilizando a técnica mais assertiva para tanto.

De forma mais específica, seguem delineadas as etapas desta pesquisa na figura 1, consolidando os marcos de 1 a 6, do início à defesa do TCC I e, na figura 2, os marcos de 7 a 13, culminando com a entrega da monografia.

1. Pesquisa bibliográfica sobre aprendizado de máquina e sobre diabetes, em plataformas disponíveis na internet como *Scielo* e *Google Scholar*, livros da biblioteca digital da Universidade, com a apresentação dos resultados das pesquisas e conteúdo no texto desta monografia.

Figura 1. Marcos da pesquisa de 1 a 6.



Fonte: autoria própria.

2. Elaboração do Mapa Conceitual: determina o ponto inicial visto que o mapa conceitual passa por transformações ao longo do trabalho com o aprofundamento do conhecimento. Para essa tarefa, foi utilizado o *"CMAP Tools*[1]*"*, uma ferramenta gratuita, em que é possível o desenvolvimento de mapas conceituais.
3. Pesquisa e análise de conjuntos de dados – repositórios públicos: Para a construção do modelo preditivo é preciso dispor de matéria prima (dados). Nesta pesquisa, foram utilizados e avaliados repositórios públicos, pela sua disponibilidade, uma vez que conjuntos de dados privados (como os de propriedade de hospitais ou operadoras de planos de saúde) são de difícil e demorada obtenção. Seguindo o mesmo modo de operação descrito no item 1, passaram por este crivo bases de dados, materiais disponíveis em sites, correlacionando os dados sobre estes conjuntos com síntese em quadros.

   Dentre os *datasets* apurados até a fase 6, todos com menos de 1000

[1] CMAP Tools. Disponível em https://cmap.ihmc.us/. Data da consulta: 6 de junho de 2021.

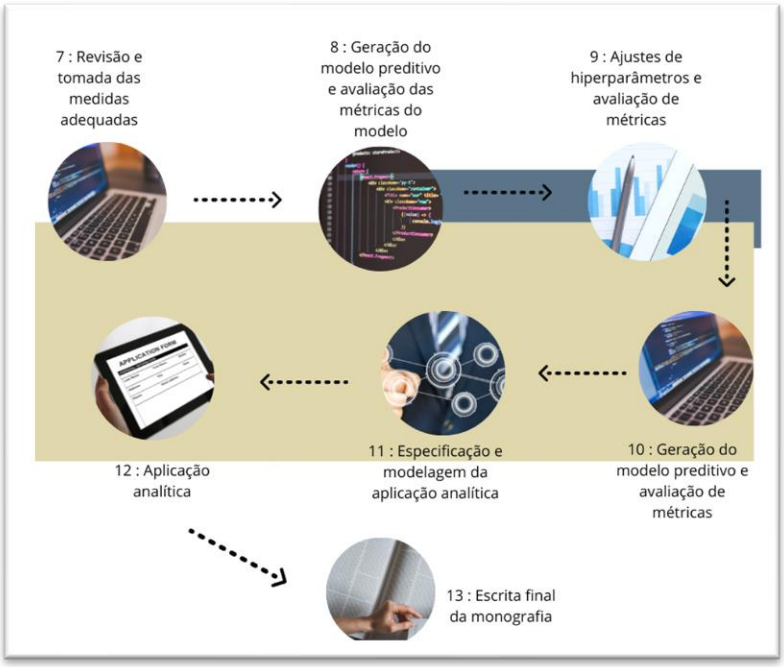
registros, o que mais se destacou foi o *Pima Indians*, além de conter o maior número de registros (768) também é mencionado em trabalhos acadêmicos ou publicações de profissionais de informática quando se trata de diabetes mellitus tipo 2.

Considerando que o número de registros abaixo da marca de 1000 poderia comprometer o desempenho do modelo preditivo, buscaram-se outros conjuntos candidatos para análise.

4. Estudo aprofundado de aprendizado de máquina, técnicas e métricas de avaliação, com base na bibliografia obtida, para adquirir o conhecimento necessário para as fases seguintes.
5. Identificação das técnicas e métricas de utilização: com base no conhecimento obtido na fase anterior, foram identificadas as técnicas e métricas de aprendizado de máquina para a realização da análise nas bases de dados públicos obtidas e definição dos passos a serem seguidos.
6. Defesa do TCC I – elaboração e apresentação do conteúdo trabalhado no segundo semestre de 2020.
7. Revisão e tomada das medidas adequadas: foi realizada a avaliação do trabalho até o ponto atingido à luz dos comentários e observações feitas pela banca examinadora.
8. Geração do modelo preditivo e avaliação das métricas do modelo.
9. Com relação aos dados, conforme citado na fase 3, foram apurados novos *datasets* em que se mostrou promissora a base NHANES, considerando o número de registros superior à 1.000. Por ser constituído de numerosos arquivos, foram necessários mais recursos para manipular os dados e obter uma tabela unificada.

   Escolhidos os parâmetros, na fase 5, foram realizados testes com conjuntos de dados escolhidos para geração de modelos preditivos, incluindo a aplicação de técnicas, como a de balanceamento de dados (*NearMiss*).
10. Ajuste de hiperparâmetros e avaliação das métricas. Uma das ferramentas para obtenção de melhora dos índices de desempenho é a aplicação de ajuste de hiperparâmetros. Outras técnicas foram executadas como a seleção de variáveis e limpeza de dados.

Figura 2. Marcos da pesquisa de 7 a 13.



Fonte: autoria própria.

11. Geração do modelo preditivo e avaliação das métricas. Concluídos os experimentos, foi gerado um modelo preditivo e foram obtidas as métricas que foram avaliadas.
12. Especificação e modelagem da aplicação analítica: como medida prévia à aplicação analítica foram realizadas as especificações funcional ou não funcional e sua modelagem por meio de diagramas *Unified Modeling Language* (UML), como o diagrama de caso de uso.
13. Aplicação analítica: desenvolvimento e testes da aplicação fazendo uso do modelo preditivo obtido em etapa anterior.
14. Escrita final da monografia. Refere-se a finalização do processo de escrita da monografia e revisão dos conteúdos para a entrega.

1.6 CONTRIBUIÇÕES

As contribuições envolvem:

- **o estudo e análise de bases de dados públicos**, a saber: a) *Early stage diabetes risk prediction dataset;* b) *Pima Indians;* c) *Sample: Diabetes (Azure);* d) *NHANES – National Health and Nutrition Examination.*

De plano, foram angariados e analisados três conjuntos (itens a, b e c), todos com número de registros inferior a 1000. Considerando o maior número de

registros em conjunto com o referenciado nos artigos contidos na bibliografia, foi eleito o *dataset Pima Indians*.

Foram apurados novos *datasets* candidatos e o resultado apontou para os dados do NHANES, base de dados do programa de estudos que avalia a saúde e nutrição de crianças e adultos norte-americanos.

Ao se realizar experimentos no NHANES, restaram dez variáveis das quais nove são variáveis de entrada: RIAGENDR [sexo biológico], RIDAGEYR [Idade], LBDSGLSI [glicose, soro refrigerado (mmol/L)], LBDSTRSI [triglicerídeos, soro refrigerado (mmol/L)], LBXGH [hemoglobina glicada], LBDINSI [insulina (pmol/L)], BMXWT [Peso (kg)], BMXHT [altura em pé (cm)], BMXWAIST [circunferência da cintura (cm)], e CLASSE como alvo.

Salienta-se que, para uniformização do *dataset*, foram utilizadas as variáveis das unidades de concentração do sistema internacional, como o milimol por litro (mmol/L).

- **a avaliação das técnicas de aprendizado de máquina, a seleção da mais** *assertiva* **e a geração do modelo preditivo**: a partir da base de dados selecionada, NHANES, foram consideradas as técnicas de classificação do aprendizado de máquina percorrendo o mapa conceitual *scikit-learn*[1].

Em termos de algoritmos de classificação, apreciados os estudos contidos na bibliografia, foram investigadas as técnicas KNN (*K-Nearest Neighbors*), Árvore de decisão, Regressão logística, *Naïve Bayes*, SVM (*Support Vector Machine*), além de métodos ensemble – *AdaBoosting* e *Random Forest*.

Os experimentos realizados apontaram para a técnica *AdaBoosting*. O modelo preditivo gerado obteve acurácia de 94,148% e matriz de confusão: verdadeiro positivo (VP): 559 (93,9%) e verdadeiro negativo (VN): 551 (94,3%), com valor médio de 94,1%. Ao realizar o ajuste de hiperparâmetros, os resultados obtidos com relação à acurácia foram de 94,063%. Este valor menor, provavelmente, foi devido aos parâmetros selecionados e ao intervalo dos valores

[1] Choosing the right estimator. Disponível em: encurtador.com.br/korLS. Data da Consulta: 4 de junho de 2021.

utilizados, sendo necessário aumentar a quantidade e variação para obter um resultado de acurácia igual ou superior.

- **A aplicação analítica desenvolvida:** denominada Flocos, permite prever se o paciente está com diabetes mellitus tipo 2 ou não, e sua probabilidade, a partir dos atributos escolhidos e dos valores informados na entrada com a acurácia de 91,67%.

1.7 ESTRUTURA E ORGANIZAÇÃO DO TEXTO

Esta monografia está subdividida em seções e subseções.

Na seção 1, encontra-se a introdução, na qual são apresentadas a contextualização, motivações e justificativas, somada a estatísticas referente aos panoramas globais e regionais da diabetes e outros impactos. Segue enunciando o objetivo geral, as questões alvo desta pesquisa assim como o método de trabalho com a definição de todas as etapas percorridas e as contribuições obtidas.

Na seção 2, encontra-se a revisão da literatura, traz o mapa conceitual, a pesquisa bibliográfica, que contém os resultados das pesquisas nas plataformas de busca desde os passos iniciais. A seção ainda elenca a fundamentação teórica, que trata conceitos como ciência de dados, *Big Data*, Aprendizado de máquina e algoritmos de aprendizagem supervisionada; trata sobre as métricas de desempenho e avaliação de dados desbalanceados, algoritmos e métricas de desempenho encontradas nos artigos da bibliografia; técnicas utilizadas na preparação de dados como a redução de dimensionalidade e balanceamento *NearMiss;* as tecnologia utilizadas; aporte sobre diabetes mellitus tipo 2, com conceitos, complicações e inferências sociais; além de tratar sobre os conjuntos de dados pesquisados, os encontrados nos artigos da bibliografia desta pesquisa e sobre o conjunto NHANES, que foi utilizado como matéria prima para a geração do modelo preditivo.

Na seção 3, são apresentados os experimentos com subconjuntos de dados com resultados de avaliação em referência ao conjunto de algoritmos selecionados, verificando os efeitos do balanceamento de dados, como o método *NearMiss*. Trata também de questões sobre dados ausentes, da preparação dos dados, da seleção de variáveis para escolha dos atributos que geraram o modelo preditivo.

Na seção 4, tem-se a arquitetura, modelagem e implementação do protótipo da aplicação analítica (Flocos), contendo especificação de requisitos, modelagem, interface gráfica e testes de validação.

Na seção 5, são realizadas as considerações finais e proposta de trabalhos futuros.

A pesquisa ainda enumera alguns apêndices para auxiliar o melhor entendimento e aprofundamentos do texto.

2 REVISÃO DA LITERATURA

Neste marco são elencados os conceitos utilizados na monografia, sem abordagem de suas minúcias, assim como um voo em cruzeiro verificando seus contornos e limites de uma área.

2.1 MAPA CONCEITUAL

Para todo o percurso, foi considerado como guia um mapa conceitual (figura 4), que traz em seu bojo os conceitos que estão relacionados e interligados ao tema desta pesquisa.

O mapa conceitual sinaliza, por meio de sua paleta de cores, os conceitos “Foco”, a “Aplicação analítica para predição de diabetes mellitus tipo 2”, que consolidará todos os conhecimentos e é o ponto final deste percurso. Ademais, os conceitos “Fundamentais”, assim como a expressão latina “*sine qua non*”, sem o qual não subsiste. É imprescindível para alcançar o resultado que se conheça, para construção e compreensão do foco, conceitos como *machine learning* e diabetes mellitus tipo 2.

Também delineados os conceitos “Relevantes”, necessários para a compreensão do conteúdo, como classificação e algoritmos de classificação e os “Associados”, aqueles que possuem ligação forte com o tema deste trabalho, como complicações da diabetes.

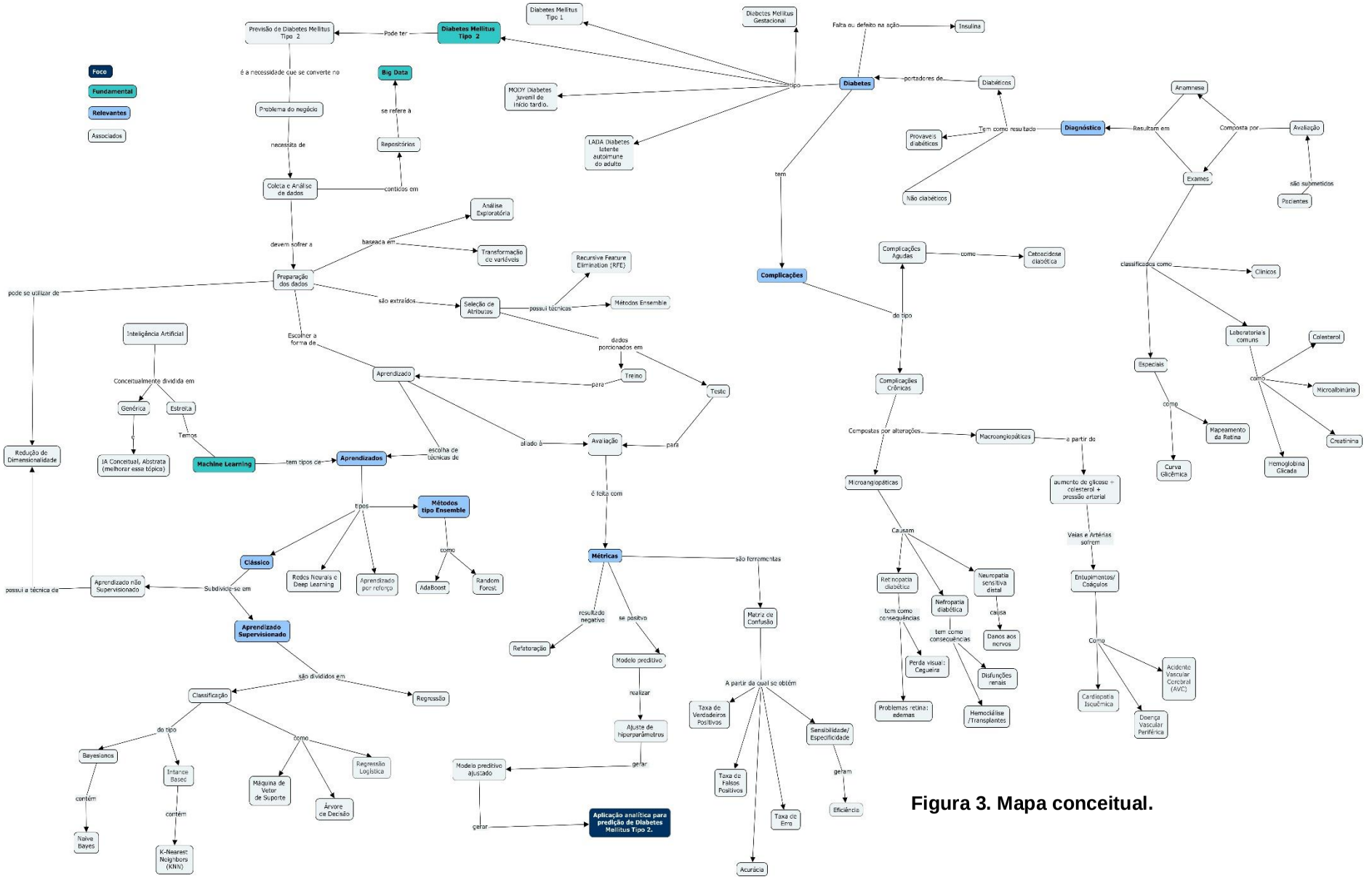


Figura 3. Mapa conceitual.

2.2 PESQUISA BIBLIOGRÁFICA

Esta pesquisa apresenta como preocupação pessoal sua realização de forma consistente e que tenha a capacidade de fazer a diferença na vida das pessoas. Então, por onde começar?

Pacientes, incluso os acompanhantes, enfrentam as dores de um diagnóstico. Cite-se não apenas sobre a dor de receber uma notícia ruim – e entender que pode ser uma queda e o quão é importante perguntar o que, então, se pode fazer.

Também as dores do longo tempo que pode demorar para o fechamento de um diagnóstico – consultas, exames, esperas. Até o início das atuações corretivas, que precisam ser rápidas para maior chance de efetividade.

Adicionado, leituras, conhecimentos obtidos em sala de aula e experiências pessoais, despontou o interesse em trabalhar com o diagnosticar de uma doença, como a diabetes mellitus tipo 2.

2.2.1 Passos iniciais

A pesquisa teve seu início no primeiro semestre de 2020 quando, no curso da matéria “Metodologia de pesquisa científica”, a professora doutora Helena Degreas propôs o desenvolvimento dos passos iniciais do trabalho de conclusão de curso como atividade prática referente à matéria que lecionava.

No sentido de colaborar com a confecção do material, ela ofereceu, gentilmente, um artigo científico (Petry, 2015), sobre *Big Data* e registros eletrônicos de saúde, vislumbrando as possibilidades de utilização.

Na mesma época, uma reportagem da SBMFS (2020) sob o título “Modelos de previsão de risco para diabetes mellitus tipo 2” apontava para um artigo de Noble (et al., 2011). Para sintetizar as descobertas iniciais, segue o quadro 2, que mostra a origem dos artigos, quantidades e citações.

Quadro 2. Síntese dos artigos encontrados no 1º Semestre/2020.

Origem	*Artigo*	*Quant.*
Indicação de artigo	Petry (2015)	1
	SBMFC (2020). Noble et al. (2011)	2
	Total (A)	3

Fonte: autoria própria.

No gráfico 1, os artigos referenciados no quadro 2, foram classificados por ano e os resultados foram graficamente representados.

Gráfico 1. Síntese dos artigos encontrados no 1º Semestre/2020 por ano de publicação.

Fonte: autoria própria.

É de importância realizar busca por artigos científicos nos meios adequados e com o estabelecimento de algumas premissas de avaliação, a saber: a) aderência do artigo por meio do seu título; b) em caso de dúvida, reavaliação com leitura do resumo e palavras-chave.

Cabe ressaltar que, ano de publicação não foi considerado um fator de corte, haja vista a existência de bons textos atemporais, porém devido ao caráter recente do tema desta proposta foi dada preferência a textos mais recentes.

As bases de trabalhos acadêmicos utilizadas foram Scielo e Google Scholar. Os resultados obtidos com as respectivas palavras-chaves são apresentados nas subseções a seguir.

2.2.2 Pesquisas em Scielo (scielo.org).

As pesquisas no sítio scielo.org, além das premissas acima explicitadas, utilizaram os termos apresentados na tabela 6, seguidos dos números referentes

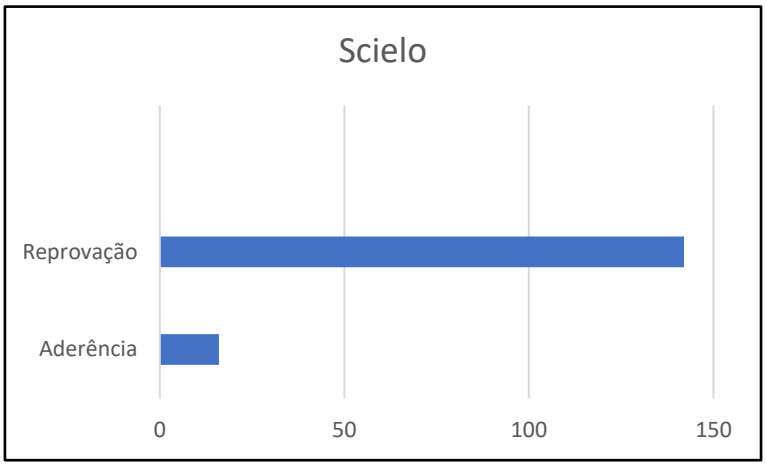
ao processo de seleção. Uma síntese da quantidade de textos aderentes e selecionados estão representados no gráfico 2.

Tabela 6. Resultados das pesquisas em Scielo.

	Quant.	Aderência do artigo ao termo de pesquisa			Reavaliados	
		Aprovados (Sim)	Reprovados (Não)	Reavaliação (Parcial)	Aprovados	Reprovados
Modelo preditivo e diabetes	10	2	6	2	1	1
Machine learning e diabetes	2	1	1	0	0	0
Machine learning e análises preditivas e saúde	5	5	0	0	0	0
Mineração de dados e saúde	49	3	46	0	0	0
Inteligência artificial e medicina	18	0	14	4	1	3
Big data e saúde	74	3	71	0	0	0
Total	158	14	138	6	2	4

Fonte: autoria própria.

Gráfico 2. Síntese dos resultados da Pesquisas em Scielo.

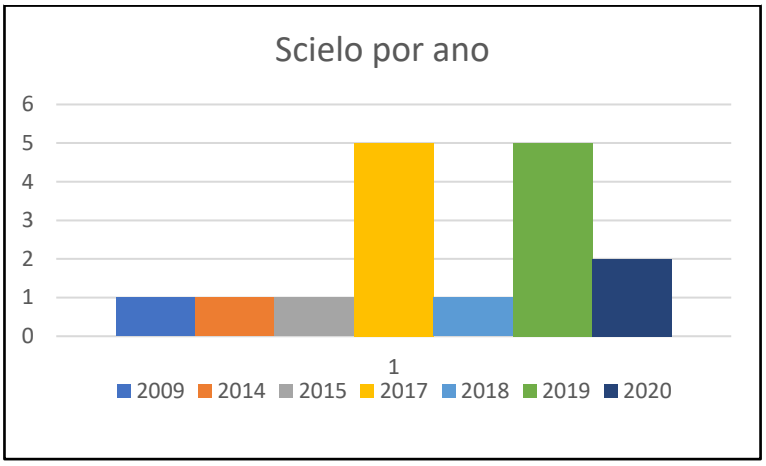


Fonte: autoria própria.

Baseado nos artigos aprovados, pode se verificar o agrupamento por ano de publicação no gráfico 3, que destaca com números mais expressivos os anos de 2017 e 2019, além de uma menção importante ao ano de 2020.

Figura 6.

Gráfico 3. Total de trabalhos aprovados por ano de publicação na base Scielo.



Fonte: autoria própria.

2.2.3 Pesquisas em Google Scholar (scholar.google.com.br/).

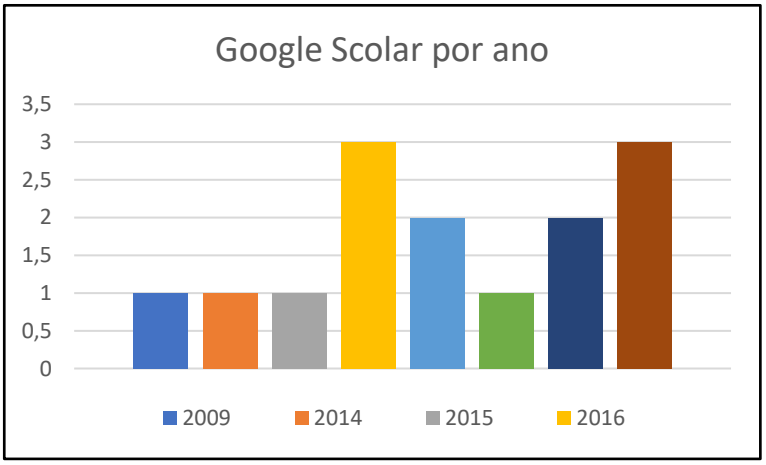
Devido ao volume de dados apresentados, foram analisados os primeiros 20, que perfazem as duas primeiras páginas de apresentação e utilizaram como termo de pesquisa “Machine learning”, “algoritmos” e “diabetes”. Os resultados encontram-se na tabela 7 e uma síntese no gráfico 4.

Tabela 7. Resultados nas Pesquisas em Google Scholar.

Número de resultados	Aderência do artigo ao termo de pesquisa			Reavaliados	
	Aprovados (Sim)	Reprovados (Não)	Reavaliação (Parcial)	Aprovados	Reprovados
20	12	4	4	1	3

Fonte: autoria própria

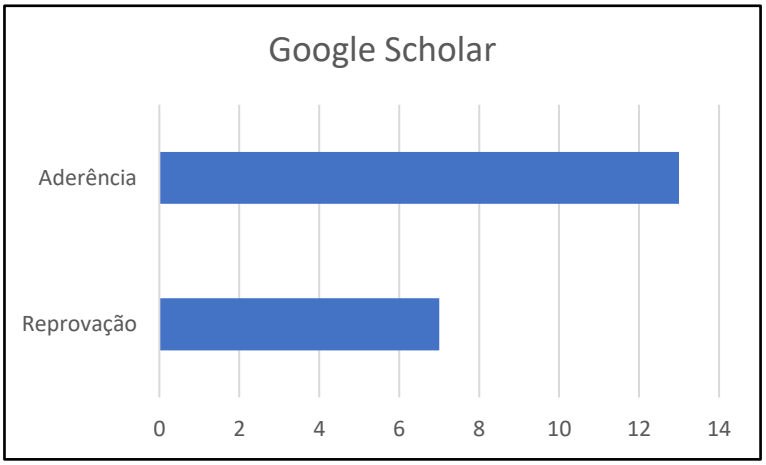
Gráfico 4. Síntese dos resultados das Pesquisas em Google Scholar por ano de publicação.



Fonte: autoria própria.

Do total de artigos, 13 foram aprovados e os demais, reprovados, o que é representado no gráfico 5, destacando-se que o maior número foi aderente aos critérios estipulados.

Gráfico 5. Total de trabalhos aprovados e reprovados na base Google Scholar.



Fonte: autoria própria.

2.2.4 Total de artigos no Scielo e Google Scholar

Dente as averiguações, foram retiradas repetições referentes às versões em duas línguas de um mesmo artigo ou atualizações, formando um conjunto de 14. No sentido de sintetizar esta segunda fase, segue o quadro 3, que demonstra a origem, quantidade total e a citação dos artigos descobertos neste ínterim.

Quadro 3. Síntese dos artigos encontrados nas plataformas Google Scholar e Scielo.

Artigos	
Basso (et al., 2014)	Lazcano-Ortiz (et al., 2009)
Carvalho (et al., 2015)	Lobo (2017)
Castrilejo (2020)	Olivera (et al. 2017)
Chiavegatto Filho (2015)	Ribeiro (2009)
Fernandes et al. (2019)	Santos (2018)
Gaytán-Hernandez (et al., 2017)	Santos (2019)
Holsbach (et al., 2014)	Sousa (2019)
Total (B)	**14**

Fonte: autoria própria.

2.2.5 Pesquisas adjacentes

Outras pesquisas se mostraram necessárias. Por exemplo, busca de dados epidemiológicos e aporte teórico da patologia em órgãos como a *International Diabetes Federation* (IDF) ou nas referências bibliográficas do material

já pesquisado que estão sinteticamente apresentados e formam o total de 34 artigos, e seus resultados são apresentados no quadro 4.

Quadro 4. Síntese dos artigos encontrados pesquisas decorrentes da pesquisa inicial.

Origem	Artigos		Quant
Referências contidas	Delagassa (2009)	Mitchell (1997)	8
	Escalera (2020)	Scheffel et al. (2014)	
	Gross et al. (2009)	Silva (2020)	
	Maione (2020)	Villaroel (2020)	
Livros TEC I	Amaral (2016)	Grus (2016)	3
	Castro (2016)		
Pessoais	Bari et al. (2019)	Francisco (2020)	2
Orientador	Carvalho (2011)	Maneo (2020)	2
Aporte teórico e dados	Austin (2019)	Moura (2017)	11
	Bandeira et al. (2015)	World Bank (2019)	
	IBGE (2020)	Pontes (et al., 2012)	
	Freitas (et al., 2018)	Stavis (2019)	
	Jakobi (et al., 2013)	Teede (et al., 2006)	
	Milech (2014)		
	ADA (2020): Avaliação de Risco		8
	BRASIL (2001): Aporte teórico		
	BRASIL (2020): Auxílio-Doença Acidentário (2019)		
	BRASIL (2020): Auxílio-Doença Previdenciário (2019)		
	BRASIL (2008): Auxílio-Doença Previdenciário (2006)		
	IDF (2019): Atlas da Diabetes		
	SBD (2019): Dados		
	SBD (2020): Aporte teórico		
	Total (C)		**34**

Fonte: autoria própria.

Por fim, segue a consolidação de todos os resultados obtidos com a somatória de todos os números apresentados no quadro 9.

Quadro 5. Número de artigos consolidados.

Total de artigos por fase	
Total (A)	3
Total (B)	14
Total (C)	34
	51

Fonte: autoria própria.

2.3 FUNDAMENTAÇÃO TEÓRICA

2.3.1 Ciência de dados

No contexto da Tecnologia da Informação, tem sido destaque e é uma área de estudo interdisciplinar. Para Amaral (2016), ela pode ser definida “como os processos, modelos e tecnologias que estudam os dados durante o seu ciclo de vida: da produção ao descarte”.

Os dados são produzidos – hoje maciçamente – por diversas fontes, com o destacado papel dos dispositivos móveis e a Internet das Coisas (IoT). Por conseguinte, são armazenados, à exemplo de bancos de dados e planilhas. A ciência de dados incorpora o armazenamento analítico (como em *Data Warehouses*, um conceito que abrange as *Data Marts*), a análise e a visualização dos dados (por exemplo: *dashboards*).

2.3.2 *Big data*

Qual a quantidade de dados produzida diariamente? Com a computação cada vez mais onipresente na vida diária, talvez seja uma surpresa fazer um relatório do consumo e produção diária de dados.

São inúmeros os exemplos: uma rápida olhada nas redes sociais, uma curtida sem qualquer ambição; uso de serviço de geolocalização para ir a um compromisso; enquanto os passageiros desfrutam do voo, a aeronave produz milhares de dados, como o espaço aéreo e indicadores; seu dispositivo vestível – relógio, conhecido como *smartwatch* – capta a velocidade dos batimentos cardíacos e pode ser capaz de detectar níveis de estresse e um infarto do miocárdio. Com poucos exemplos pode ser notada a gama de dados produzidos diariamente.

Tratar sobre *Big Data* é tratar de um grande volume de dados, como o próprio significado da expressão. No entanto, é preciso ir além, pois o conjunto de dados é caracterizado por 5V´s, sendo três os principais (volume, velocidade e variedade) e os demais, veracidade e valor, tão importantes quanto.

Chiavegatto Filho (2015) aponta o crescimento da utilização de *Big Data* na área da saúde com três áreas prósperas, ao qual são mencionadas duas:

Medicina de precisão: para o autor, saber quais os pacientes para os quais determinado medicamento não funciona: “Em vez de prescrever o mesmo anticoagulante oral para todos os pacientes, espera-se que um dia seja possível indicá-lo apenas para indivíduos para os quais o medicamento verdadeiramente funcione”. (CHIAVEGATTO FILHO, 2015).

Prontuários eletrônicos do paciente: apesar do grande volume de dados de cada paciente, muitos não têm seus dados em apenas uma base, mas

espalhados por hospitais e clínicas diagnósticas públicas e privadas. É o que pode acontecer com pacientes da rede pública que, com a oferta de serviços de análises clínicas de laboratórios com o rótulo “popular” ou mais acessível, realizam seus exames a preço de custo para tornar mais célere o tratamento não dependendo de vagas longínquas. Por maior que seja o desejo de que os repositórios de dados dos pacientes fossem universais, não só para análises, mas para o melhor acesso dos dados pela equipe médica, ainda não foi alcançada esta situação. Por certo:

> Uma solução é o uso integrado do prontuário eletrônico do paciente (PEP), que permitiria o uso remoto do mesmo prontuário por todos os estabelecimentos de saúde. Alguns dos benefícios do uso integrado do PEP são o ganho de tempo no preenchimento, a diminuição do viés de memória/esquecimentos, a completitude das informações e o seu potencial para uso em pesquisas científicas. (CHIAVEGATTO FILHO, 2015).

Ainda mais:

> Dados de pacientes, como idade, sexo, etnia, local de residência, antecedentes pessoais e familiares, sintomas e sinais apresentados, exames realizados ou obtidos por meios eletrônicos (wearable devices), diagnósticos feitos, tratamento e evolução coletados, permitiriam estabelecer uma base de dados e aprimorar condutas estabelecidas. (LOBO, 2017).

2.3.3 Aprendizado de máquina

Advindo da Inteligência Artificial, Mitchell (1997) sinaliza que o aprendizado de máquina é definido como a área de pesquisa que visa desenvolver programas computacionais capazes de melhorar automaticamente seu desempenho por meio da experiência.

“Na área de análise de dados, isso significa a elaboração de algoritmos que respondam e se adaptem automaticamente aos dados sem a necessidade de intervenção humana contínua” (CHIAVEGATTO FILHO, 2015).

O aprendizado de máquina, ou conhecidamente *machine learning*, utiliza dados já existentes e algoritmos para a criação de modelos preditivos, ou seja, modelos que podem prever os resultados a partir de novos dados de entrada.

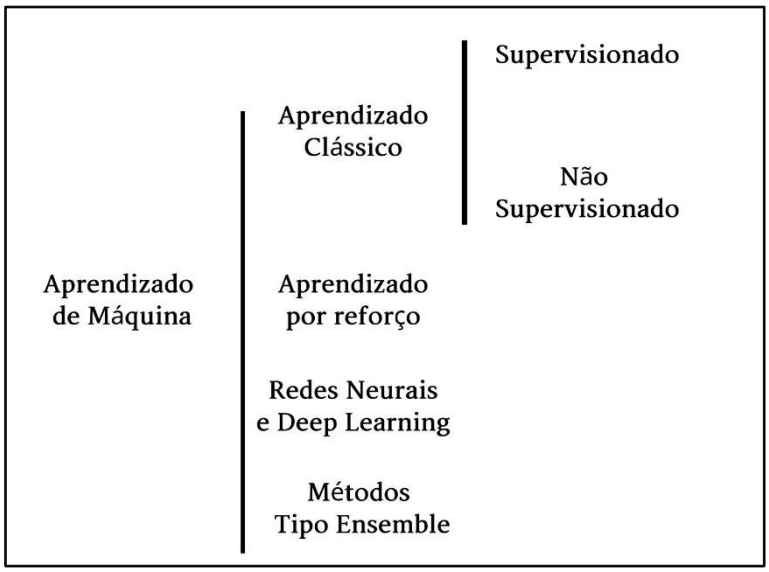
> Um algoritmo de AM preditivo é uma função que, dado um conjunto de exemplos rotulados, constrói um estimador. O rótulo ou etiqueta toma valores num domínio conhecido. Se esse domínio for um conjunto de valores nominais, tem-se um problema de classificação, também conhecido como aprendizado de conceitos, e o estimador gerado´e um classificador. (CARVALHO et al., 2011).

Pode ser utilizado, por exemplo, na previsão de fraude de uma transação, na classificação de um e-mail como spam, ou mesmo na predição dos resultados da produção agrária.

2.3.4 Tipos de aprendizado de máquina

O desenho do cenário dos tipos de aprendizado de máquina apresenta como principais tipos: a) aprendizado de máquina clássico; b) aprendizado por reforço; c) métodos tipo *Ensemble*; e d) redes neurais e aprendizado profundo (*deep learning*); na figura 4, podem ser vistos os tipos de aprendizado de máquina e a localização da classificação, linha adotada nesta pesquisa.

Figura 4. Modelo conceitual de tipos aprendizado de máquina.



Fonte: autoria própria.

No contexto do aprendizado de máquina clássico, encontram-se os tipos:

Aprendizado não supervisionado: “é baseado apenas nos objetos de base, cujos rótulos são desconhecidos. Basicamente, o algoritmo deve aprender a ‘categorizar’ ou rotular objetos” (CASTRO et al., 2016).

Considerando uma criança que recebe uma caixa de formas geométricas para que as agrupe pela semelhança sem que saiba o que são ou representam. Se a caixa contém esferas e retângulos, espera-se que a criança consiga separar em dois grupos. O agrupamento é um exemplo de técnica não supervisionada.

Aprendizado supervisionado: “é baseado em um conjunto de objetivos para os quais as saídas desejadas são conhecidas, ou algum outro tipo de informação que represente o comportamento que deve ser apresentado pelo sistema”. (CASTRO et al., 2016).

Após uma criança aprender sobre três cores (azul, amarelo e vermelho), seu professor retira de uma caixa uma esfera e pergunta qual a cor daquela esfera. É de se esperar que, ao retirar uma esfera azul, a criança resposta “azul”.

O que é classificar? “Classificar um objeto significar atribuir a ele um rótulo, chamado classe, de acordo com a categoria à qual ele pertence.” (CASTRO et al., 2016).

Vamos agora conhecer um pouco sobre algoritmos de classificação.

2.3.4.1 *KNN – K-Nearest Neighbors*

O algoritmo *K-Nearest Neighbors* – KNN, ou K – vizinhos mais próximos, é um dos principais algoritmos utilizados por ser simples e apresentar boa acurácia preditiva em vários conjuntos de dados.

É um algoritmo baseado em distâncias, que:

> consideram a proximidade entre os dados na realização de predições. A hipótese base é que dados similares tendem a estar concentrados em uma mesma região no espaço de entrada. De maneira alternativa, dados que não são similares estão distantes entre si. (CARVALHO, 2011).

Segundo o raciocínio de Grus (2016), o comportamento de um indivíduo está associado ao comportamento dos vizinhos mais próximos, deixando de considerar todo o conjunto dos vizinhos.

> Este algoritmo opera de forma que, dado um objeto x_0 cuja classe se deseja inferir, encontram-se os k objetos x_i, i = 1, …, k da base que estejam mais próximos a x_0 e, depois, se classifica o

> objeto x_0 como pertencente à classe da maioria dos k vizinhos. (CASTRO et al., 2016).

Onde k é o número dos vizinhos mais próximos que são considerados, evitando-se grandes valores para a melhor definição de fronteiras, o que se faz por heurística ou tentativa e erro.

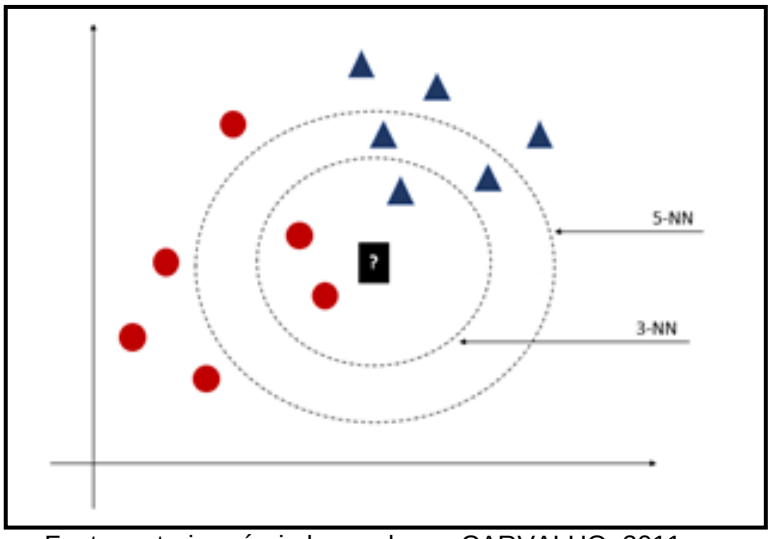
Trata-se, portanto, de algoritmo baseado em instâncias pois, a partir da classe de outros objetos (instâncias) ele determina a classe de um objeto, o que se coaduna no pensamento de que o K-NN

> representa cada elemento em função das variáveis que o definem em um espaço multidimensional e seleciona um subconjunto de K-elementos próximos entre si gerando os denominados 'bairros'. Desta forma, ao introduzir um novo valor, este será definido pela proximidade como o bairro mais próximo. (ALONSO CASTRILEJO, 2020, tradução nossa).

Segundo Carvalho (2011), de forma geral, "o valor de k é pequeno e ímpar: k = 3, 5, … Em problemas de classificação, não é usual utilizar k = 2 ou valores pares, para evitar empates".

Por exemplo: são eleitos os valores k = {3, 5}. As bolinhas vermelhas são pessoas classificadas como portadores de diabetes mellitus tipo 2. Os triângulos azuis, pessoas não portadoras de diabetes mellitus tipo 2. Abaixo, na figura 10, está ilustrado a situação deste contexto.

Figura 5. Exemplo de KNN.



Fonte: autoria própria baseado em CARVALHO, 2011.

No centro, encontra-se um elemento com o sinal de interrogação. Trata-se de elemento ainda não classificado. A menor distância é k = 3. Dentro deste círculo, a classe predominante é a de portadores de diabetes mellitus tipo 2 e assim será classificado o item.

São representante dos algoritmos *Instance based*, o KNN é acompanhado pelos algoritmos *Learning Vextor Quantization* (LVQ), *Self-Organizing Map* (SOM) e *Localli Weighted Learning* (LWL).

2.3.4.2 *Naïve Bayes*

Classificadores bayesianos, como o *Naïve Bayes*, “são classificadores estatísticos fundamentados no Teorema de *Bayes*, conforme eq. (1) (e usados para predizer a probabilidade de pertinência de um objeto e determinada classe)” (CASTRO et al., 2016).

$$P(A|B) = \frac{P\ (B|A)\ P(A)}{P\ (B)} \qquad (1)$$

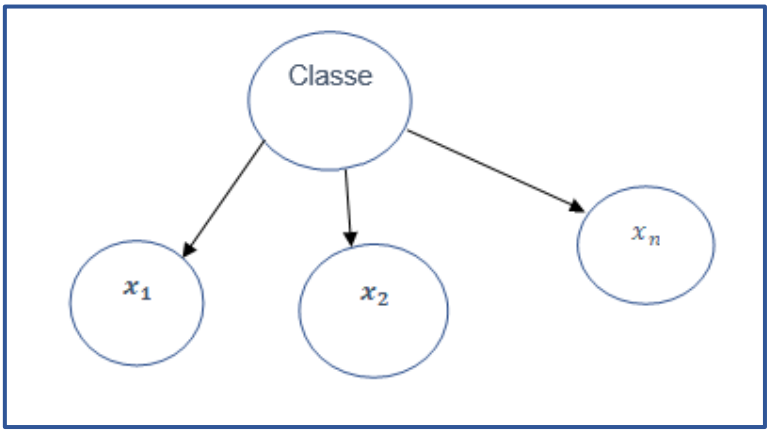
O Teorema de *Bayes* calcula a probabilidade de um evento (A) baseado na ocorrência de um evento anterior (B) e sua probabilidade (P(B)), onde:

P(A|B): probabilidade de um evento A dado que um evento B ocorreu;
P(B|A):probabilidade de um evento B dado que A ocorreu;
P(A): probabilidade de A ocorrer;
P(B): probabilidade de B ocorrer.

É importante destacar que o *Naïve Bayes* assume a independência dos valores de cada atributo, o que objetiva a simplificação dos cálculos pelo algoritmo, a permissa denominada independência condicional de classe é o que denomina como *naïve*. Um exemplo: a nectarina é fruto de coloração avermelhada, de casca lisa e formato de coração. O algoritmo irá considerar a probabilidade de cada característica de forma individual.

Como exemplifica Olivera (2016), “no grafo representado na estrutura, as únicas conexões que existem saem da variável-alvo em direção a todas as variáveis preditoras”. Nesse caso, ele não permite a dependência de atributos o que o faz o mais restrito. Veja o exemplo de um grafo bayesiano na figura 6.

Figura 6. Exemplo grafo Naïve Bayes.



Fonte: autoria própria.

Dentre os algoritmos bayesianos acompanham o Naïve Bayes, o Averaged One-Dependence Estimators (AODE), Bayesian Belief Network(BBN), Gaussian Naïve Bayes, Multinomial Nayve Bayes e Bayesian Newtwork (BN).

2.3.4.3 *Árvores de decisão*

Este tópico remete cada leitor a uma imagem conhecida de árvore. Toda árvore tem uma raiz, ramos e folhas. Sob essa aparência, representa uma tabela de decisão no formato de árvore.

> Uma árvore de decisão (*decision tree*) é uma estrutura em forma de árvore na qual cada nó interno corresponde a um teste de um atributo, cada ramo representa um resultado do teste e os nós folhas representam classes ou distribuições de classes. O nó mais elevado da árvore é conhecido como o nó raiz e cada caminho da raiz até um nó folha corresponde a uma regra de classificação. (CASTRO et al., 2016).

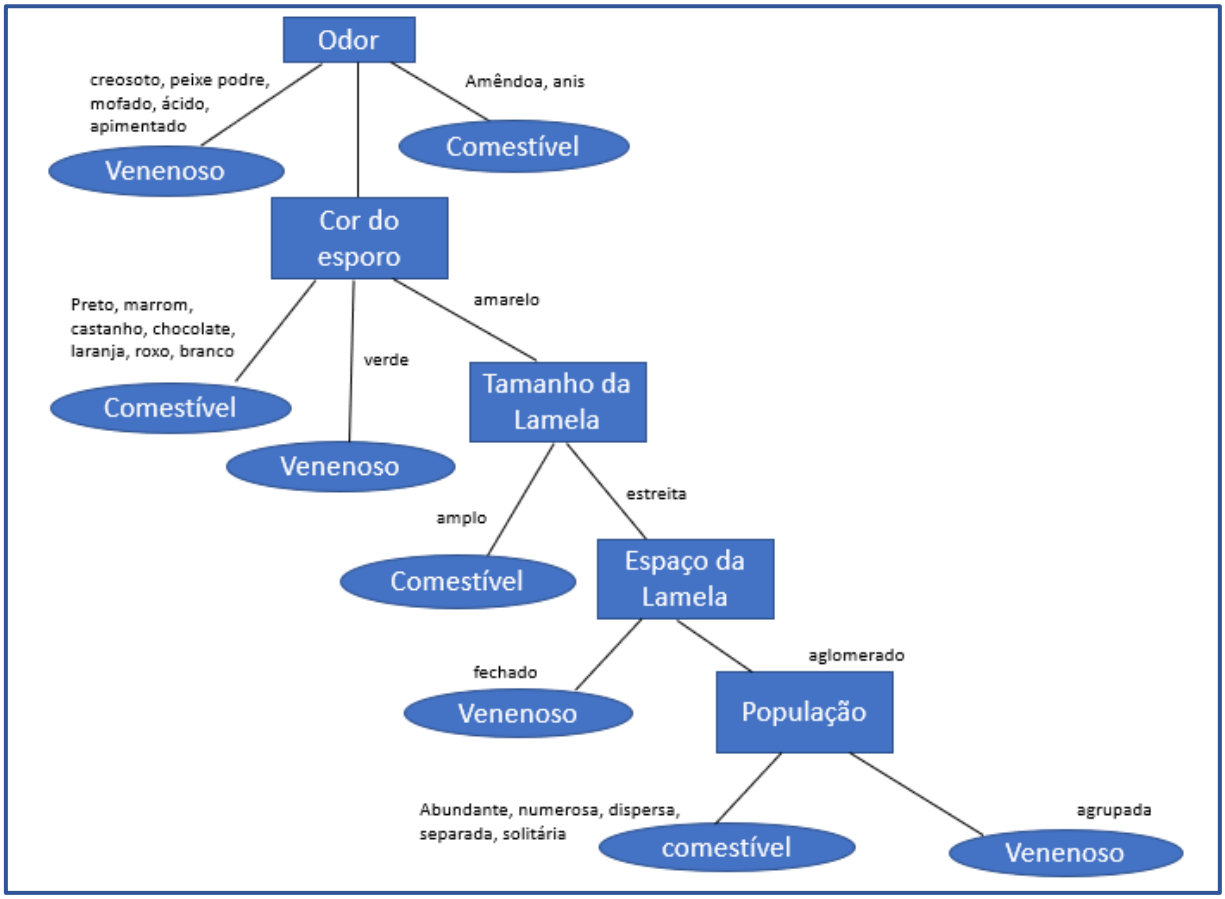
Pode-se utilizar o algoritmo para a classificação de uma classe não conhecida. É necessário que os atributos sejam testados percorrendo-a a partir da raiz até atingir o nó folha.

Por exemplo, um cogumelo possui as seguintes propriedades: odor terroso, cor do esporo amarelo e tamanho da lamela amplo. Na figura 7, observa-se um exemplo de árvore de decisão.

Dependendo dos dados, os testes podem ser feitos da seguinte forma:

1. Odor terroso: não pode ser classificado, prossegue na árvore com a cor do esporo.
2. Cor do esporo amarelo: não pode ser classificado, prossegue na árvore com o tamanho da lamela.
3. Tamanho da lamela amplo: atingiu o nó folha e pode ser classificado como comestível.

Figura 7. Exemplo de árvore de decisão sobre cogumelos baseado em CASTRO (et al., 2016) referente a um conjunto de treinamento de base de dados de Cogumelos.



Fonte: autoria própria baseado em Castro et al., 2016.

Dentre os algoritmos de árvore de decisão é possível citar: Classification and Regression Tree (CART), Iterative Dichotomiser 3 (ID3), C4.5, C5.0, Chi-squared Automatic Interaction Detection (CHAID), Decision Stump, Conditional Decision Trees e M5.

2.3.4.4 *Máquina de vetor de suporte*

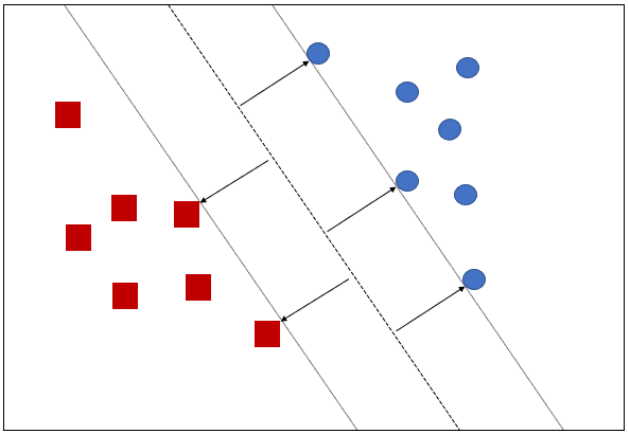
Este algoritmo de classificação utiliza uma função matemática que "atribui a novos elementos de dados uma das categorias rotuladas" e assim prever o rótulo do elemento desconhecido. (BARI et al., 2019).

A máquina de vetor de suporte (SVM) constrói

> um hiperplano com uma superfície de decisão em que a margem de separação entre as duas classes é maximizada” para “mapear os dados em função do espaço de características e em seguida tentar usar a hiper esfera para descrever os dados e para a inserção de dados” em que “os pontos do conjunto de treinamento que estão mais próximos da superfície de decisão são chamados de vetores de suporte. (RIBEIRO, 2009).

Na figura 8, as linhas pontilhadas são as margens máximas e os itens que tocam estas linhas são os vetores de suporte. A linha tracejada é o hiperplano separador ótimo.

Figura 8. Exemplo de um vetor de suporte.



Fonte: autoria própria.

Trata-se de uma forma de classificar os elementos de acordo com as extremidades.

> A máquina de vetor de suporte maximiza a margem entre as instâncias mais próximas, criando um vetor otimizado para classificá-las. Os vetores nas extremidades são as margens otimizadas, o vetor do centro é a referência para classificar novas instâncias. (AMARAL, 2016, p. 103).

2.3.4.5 *Métodos ensemble*

Conhecidos em algoritmos de ordenação, como o *Merge Sort*, Bari (et al. 2019) lembram que os métodos *ensemble* também vislumbram a premissa de “dividir para conquistar” em que diversos modelos podem ser combinados de diferentes maneiras para fazer predições, o que pode inferir uma melhor acurácia.

Conforme Oliveira (2016), alguns métodos *ensemble* são:

- ***AdaBoost (Adaptive Boosting*)**: os modelos são construídos iterativamente de modo que a construção de um modelo influencia a construção dos modelos que serão criados nas iterações posteriores. […] Dessa forma, um modelo complementa o outro. A classificação de uma nova instância é realizada de forma ponderada, levando em consideração o erro que cada modelo teve para classificar os dados de treinamento.
- **Random Forest**: a ideia é construir uma floresta de *Random Decision Trees* e utilizar todas as árvores para classificar uma nova instância. Cada árvore é gerada utilizando um subconjunto das variáveis disponíveis escolhidas de forma aleatória, além de um diferente conjunto de dados, gerado pela técnica de amostragem *bootstrap*. O algoritmo ainda pode ser utilizado para classificação de relevância de variáveis.

Dentre os métodos ensemble tem-se os algoritmos: Random Forest, Gradient Boosting Machines(GBM), Boosting Bootstrapped Aggregation (Bagging), AdaBoost *Stacked Generalization* (*Blending*), *Gradient Boosted Regression Trees* (GBRT).

2.3.4.6 *Regressão logística*

Santos (2018) diz que a regressão logística é:

> um modelo linear para problemas de classificação, cujo objetivo é predizer a probabilidade de que cada observação pertença a uma das classes da resposta de interesse. Para que essa probabilidade possa ser estimada, a variável resposta do conjunto de treinamento é modelada por meio da distribuição binomial, que tem como parâmetro, p, a probabilidade de ocorrência de uma classe específica. (SANTOS, 2018).

É um algoritmo que modela a relação entre a probabilidade de determinada resposta e o conjunto de preditores. A partir da função logística, e "conforme sua entrada fica grande e positiva, ela se aproxima cada vez mais de 1. Conforme sua entrada fica grande e negativa, se aproxima de 0." (GRUS, 2016), o que pode ser verificado na tabela 10 para cada valor assumido por X. A

probabilidade estimada de determinado evento deve estar limitada ao intervalo [0,1].

Tabela 10. Exemplo de função logística.

X	
-5	0,006693
0	0,5
2	0,880797
5	0,993307
10	0,999955

Fonte: autoria própria baseada em GRUS (2016).

Para Olivera (2016), “a regressão logística é uma técnica de classificação bem estabelecida e amplamente utilizada em estudos epidemiológicos”. A opção pela utilização advém da observação de que “geralmente usado como referência, em comparação com outras técnicas de análise de dados médicos”. Ademais:

> O princípio da regressão logística tem base na técnica de regressão linear onde uma variável quantitativa y é definida como uma relação linear das variáveis preditoras X de acordo com a seguinte fórmula: $h_{x=} w_0 + w_1 x_1 . + \dots w_k x_k$ O conjunto de pesos W é encontrado através da minimização de uma função de custo baseada nos erros quadráticos dos dados de treinamento. Na regressão logística, utiliza-se uma função sigmoide (logística) que define a equação linear fornecendo respostas entre 0 e 1. A saída dessa função significa a probabilidade condicional de uma instância pertencer a uma determinada classe dado um conjunto de variáveis quantitativas preditoras. (OLIVERA, 2016).

Um exemplo citado por Amaral (2016, p.107):

> dados de eleições passadas contém, de cada candidato, o quanto ele investiu em cada campanha, e se ele foi candidato eleito ou não. Dessa forma, dados de novos candidatos são carregados, com orçamento de campanha de cada um. A regressão logística mostra a probabilidade de um candidato ser eleito baseado no orçamento de campanha.

2.3.5 Métricas de desempenho

Quando se trata de qualidade de uma atividade, em sentido aberto, há um ciclo importante para ser percorrido a fim de que haja bons resultados. A partir

do conhecimento do tema, definir métricas, realizar medições, controlá-las e gerenciá-las são medidas desse ciclo.

William Edwards Deming (1900-1993) tem uma célebre frase que corrobora a afirmativa acima: “não se gerencia o que não se mede, não se mede o que não se define, não se define o que não se entende, e não há sucesso no que não se gerencia”. (MOURA, 2017).

Tão importante quanto a construção do modelo preditivo é realizar a avaliação do desempenho, o que pode ser feito com a utilização de métricas no ciclo de qualidade.

Um exemplo: a partir de um conjunto de dados, foram escolhidos dois algoritmos de classificação: Naïve Bayes e KNN. Para a avaliação, a métrica de acurácia. Executadas as atividades, os testes resultaram os valores de acurácia de 89% e 93%, respectivamente. Diante dos resultados, foi escolhido o algoritmo KNN para a geração do modelo preditivo.

Assim, houve a medição dos resultados e o controle – a comparação dos números resultantes com os paradigmas definidos – que é uma informação para a tomada de decisão. Pode ser que os números sejam satisfatórios ou não, e isso implicará na refatoração do processo, assim como tomar outras medidas como o ajuste de hiperparâmetros – que visam melhorar o desempenho do algoritmo.

Para problemas de classificação binária, em que duas respostas são possíveis, como o caso desta pesquisa, pode-se dizer que “existe uma classe alvo, ou seja, a classe cujo valor se deseja predizer” (CASTRO et al., 2016). Considerando o contexto, será classe positiva para “diabético tipo 2”, ou negativa, para “não diabético tipo 2”.

Segundo o mesmo autor tem-se:

- VP (verdadeiro positivo):objeto da classe positiva classificado como positivo.
- VN (Verdadeiro negativo): objeto da classe negativa classificado como negativo […]
- FP (Falso positivo): objeto da classe negativa classificado como positivo. […] É também conhecido como alarme falso ou erro tipo 1.

- FN (Falso negativo): objeto da classe positiva classificado como negativo […] É também conhecido como erro do tipo 2. (CASTRO et al., 2016).

Subsumindo à pesquisa:

- VP: “diabético tipo 2” classificado como “diabético tipo 2”.
- VN: “não diabético tipo 2” classificado como “não diabético tipo 2”. Ou seja, as duas ocorrências são caso de classificação correta dos dados.
- FP: “não diabético tipo 2” classificado como “diabético tipo 2”;
- FN: “diabético tipo 2” classificado como “não diabético tipo 2”. Ou seja, neste caso, os dados não foram classificados de forma correta.

Para apresentar o desempenho do algoritmo de classificação binária pode se construir uma matriz de confusão, ou matriz de contingência ou ainda, matriz de erro, conforme segue no quadro 6.

Quadro 6. Matriz de confusão .

		Classe predita	
		Positiva	Negativa
Classe original	Positiva	VP	FN
	Negativa	FP	VN

Fonte: autoria própria baseado em Castro (et al. 2016).

A seguir, observa-se no quadro 7, um exemplo de matriz de confusão com dados fictícios dos resultados de um algoritmo de predição de diabetes mellitus tipo 2. A predição apontada é para “diabético tipo 2” e “não diabético tipo 2” como alvo.

Quadro 7. Exemplo de matriz de confusão.

		Classe predita	
		Diabético tipo 2	Não diabético tipo 2
Classe original	Diabético Tipo 2	300	20
	Não diabético tipo 2	12	168

Fonte: autoria própria baseado em Castro (et al., 2016).

Com a matriz de confusão ou contingência, podem ser realizados cálculos importantes para mensuração do desempenho, dentre os quais:

- TVP: Taxa de verdadeiros positivos (eq. (2)) é o percentual de objetos positivos classificados corretamente. Também designada pela literatura como sensibilidade ou revocação. Com os dados da matriz de confusão, resultou em 94%.

$$TVP = \frac{Verdadeiros\ positivos}{Total\ de\ positivos} = \frac{VP}{VP\ + F\ N} \quad (2)$$

$$TVP = \frac{300}{300 + 20} = \frac{300}{320} = 0{,}9375 \quad (2a)$$

- TFP: Taxa de falsos positivos (eq. (3)) é o percentual de objetos negativos classificados como positivos. Com os dados da matriz de confusão, resultou em 11%.

$$TFP = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{VN}} \quad (3)$$

$$TFP = \frac{20}{20 + 168} = \frac{20}{188} = 0{,}1063 \quad (3a)$$

- Acurácia (ACC): taxa de sucesso do algoritmo (eq. (4)) é o, é o número de classificações corretas dividido pelo número total de classificações. Com os dados da matriz de confusão, resultou em 94% de acurácia, o que é satisfatório quando se considera a criticidade de um diagnóstico.

Nesta fórmula, observa-se que VP e VN, o numerador da fração, são as predições corretas, positivas ou negativas. O denominador é formado pelo conjunto de todas as predições, verdadeiras ou falsas. Ou seja, o número de

$$ACC = \frac{Verdadeiros\ positivos}{Total\ de\ resultados} = \frac{VP + FN}{VP\ + FP + VN + \ FN} \quad (4)$$

$$ACC = \frac{300 + 168}{300 + 12 + 168 + 20} = \frac{468}{500} = 0{,}936 \quad (4a)$$

- $E = \ 1 - ACC$Taxa de erro (eq. (5)) é o, que utiliza a acurácia como variável e, neste caso, perfaz 6,4%

$$Erro = \ 1 - \ 0{,}936 = 0{,}064 \quad (5)$$

Oliveira (2016) aponta também outras métricas para avaliação do desempenho dos algoritmos dentre as quais cabe salientar a sensibilidade (eq. (2)) e a especificidade (eq. (6)), que:

> é a razão entre os casos classificados corretamente como negativos e todos os casos negativos.” O autor salienta ainda que, “através de alterações de parâmetros do algoritmo, é possível variar entre essas duas métricas, ou seja, aumentar a sensibilidade ao preço da especificidade.

Para os números apresentados do exemplo, tem-se uma taxa de 89%.

$$Especificidade = \frac{\text{VN}}{\text{VN} + \text{FN}} \quad (6)$$

$$Especificidade = \frac{168}{168 + 20} = \frac{168}{188} = 0{,}8936 \quad (6a)$$

Por fim, Clésio (2014) traz a eficiência, que é “a média aritmética da sensibilidade e Especificidade. Na prática, a sensibilidade e a especificidade variam em direções opostas”. Por isso, é necessário atingir um balanço através da eq. (7).

$$Eficiência = \frac{\text{Sensibilidade} + \text{Especificidade}}{2} \quad (7)$$

2.3.5.1 *Avaliação de dados desbalanceados*

Quando se trata de dados desbalanceados, a acurácia não é a medida apropriada para realizar a avaliação de desempenho. Conforme Maione (2020),

> uma vez que grandes números de verdadeiros positivos tenderão a acobertar grandes números de falsos positivos e vice-versa, de maneira que um classificador que rotula muito bem amostras da classe majoritária e rotula de maneira mediana amostras da classe minoritária ainda apresentará um bom valor de acurácia.

Dessa forma, a acurácia, nesses casos, implicará em resultados que não poderão ser corroborados. E “para ser aceitável, a acurácia preditiva de um classificador para um conjunto de dados desbalanceados deve ser maior que a acurácia obtida atribuindo todo novo objeto à classe majoritária”. (CARVALHO, 2011).

Quais métricas podem ser utilizadas, nesses casos?

> A precisão é uma boa métrica para determinar quando os custos do falso positivo são altos, em outras palavras, ela indica qual a proporção de identificações positivas estava realmente correta;
>
> A revocação é uma boa métrica quando há um alto custo associado ao falso negativo, em outras palavras, ela indica qual a proporção de positivos reais foi identificada corretamente, ou seja, quão bom o modelo é para prever positivos, sendo positivo entendido como a classe que se quer prever. (VILLAROEL, 2020).

Outra métrica apontada é a Matriz de confusão ou contingência.

> Como já indicado, VP e VN registram classificações corretas. FP indica registros que não pertencem a uma classe e classificados como pertencentes a ela, e FN indica os pertencentes àquela classe, porém classificados como não pertencentes. A depender do que se quer classificar, escolher entre classificadores que apresentam valores baixos de falsos negativos ou falsos positivos é essencial. Num problema associado à saúde, por exemplo, usualmente há menos dano quando se peca pelo excesso (classificar um paciente não enfermo como doente e prosseguir a investigação) do que pela falta (deixar de classificar um paciente como enfermo quando ele o é); nessa situação de exemplo, é melhor escolher um classificador que apresente o menor número de falsos negativos. (GENTILLE, 2017, p53).

Carvalho (2011) aponta que, com dados desbalanceados, o algoritmo favorecerá a classificação dos novos dados na classe majoritária. O conjunto pode ser naturalmente balanceado com a inserção de novos dados da classe minoritária. Mas também existem técnicas de balanceamento artificial porquanto não podem ser inseridos mais dados, como a técnica *NearMiss*, utilizada nesta pesquisa. E assim obter uma solução para esse problema.

2.3.5.2 *Algoritmos encontrados na bibliografia e métricas de desempenho*

Determinados artigos contidos na bibliografia desta pesquisa contêm referências à experimentos dos autores. Sousa (2019) não traz métricas de desempenho e o artigo exemplifica a técnica com aplicações. O quadro 8 oferta uma visão geral sobre este conteúdo a partir dos autores, citando as técnicas (algoritmos), relacionando aos *datasets* e os resultados obtidos.

Quadro 8. Síntese dos algoritmos encontrados na bibliografia.

Autor	***Técnica estudada***	***Dataset***	***Desempenho***	
Sousa (2019)	K-Means (Clustering)	Índios Pima	Estudo de técnica	
Basso (et al., 2014). Ribeiro (2009)	Redes Neurais (Backpropagation)	Pima Indians	Taxa de acerto	
			187 acertos	81.31%,
			43 erros	18,69%
Ribeiro (2009). Carvalho (et al., 2015), Delagassa (2009)	Máquina de vetor de suporte (SVM – One Class)	Pima Indians	Acurácia	98,91%
			Sensibilidade	99,43%
			Especificidade	97,87%
Carvalho (et al., 2015), Delagassa (2009). Olivera (et al., 2017)	KNN; Naïve Bayes; Arvores de decisão; Redes Neurais; Classificação baseada em agrupamento Escolha do autor: Algoritmo de árvore de decisão – J48	Base de dados de uma operadora de plano de saúde do estado do Paraná com total bruto de 43.375 registros	Matriz de confusão	
			Taxa de acertos sem indicativo	96%
			Taxa de acerto com indicativo	42%
			Taxa média de acerto	88.9%
Olivera (et al., 2017)	Regressão logística Perceptron/ Backpropagation (Redes Neurais Artificiais). Multi-Layer/ Naïve Bayes/ KNN Quinlan's C5/ RIPPER Escolha do autor: Regressão logística	Dados do Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil) com 12.247 registros	Sensibilidade	68%
			Especificidade	67,2%

Fonte: autoria própria

Cumpre assinalar que, foram selecionadas como métricas de avaliação a Matriz de confusão ou contingência, que é apontada por GENTILLE (2017) como uma alternativa à dados desbalanceados, ocorrência comum entre os conjuntos de dados, e que também é utilizada por Carvalho (et al., 2015) e Delagassa (2009), conforme o quadro 9.

É de interesse da autora, utilizar a acurácia como medida para verificação, referenciada nos trabalhos de Ribeiro (2009) para avaliação de técnica de Máquina de Vetor de Suporte (SVM), um dos algoritmos que selecionados para os testes desta pesquisa.

2.3.6 Técnicas de Preparação de Dados

No contexto da preparação de dados, há técnicas prévias ao processamento que são utilizadas. Sua aplicação visa dar maior conformidade para que sejam utilizados pelos algoritmos de aprendizado de máquina. Algumas delas serão resumidas a seguir.

2.3.6.1 *Redução de dimensionalidade*

A primeira análise sobre todos os dados disponíveis no NHANES formou um conjunto expressivo de variáveis, entre 37 e 39. Não se diminui a importância que cada uma delas tem na atividade preditora, entretanto, é interessante diminuir a dimensão de variáveis para gerar um modelo genérico e uma aplicação analítica.

Uma nova análise selecionou 22 variáveis e outras foram realizadas até alcançar o número de 10 variáveis. Mas qual o benefício da técnica?

> Em muitos algoritmos de aprendizado de máquina, para que dados com um número elevado de atributos possam ser utilizados, a quantidade de atributos precisa ser reduzida. A redução do número de atributos pode ainda melhorar o desempenho do modelo induzido, reduzir seu custo computacional e tornar os resultados obtidos mais compreensíveis. (CARVALHO, 2011).

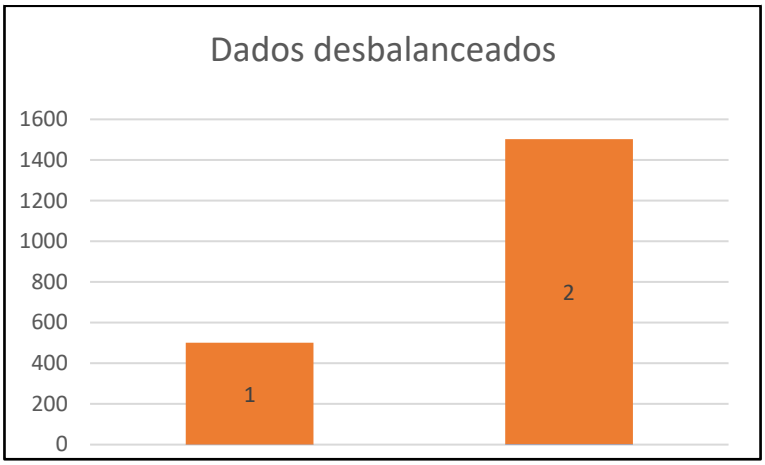
2.3.6.2 *Balanceamento NearMiss*

O balanceamento dos dados implica na subamostragem da classe majoritária que “tem seu tamanho reduzido, visando diminuir os efeitos do desbalanceamento das bases de dados para melhorar a classificação. Alguns exemplos deste tipo de técnica são: aleatória, cluster centroides e NearMiss”. (SILVA, 2020).

São “regras heurísticas para escolher as instâncias a serem eliminadas na classe dominante. São oferecidas três versões do algoritmo que implementam heurísticas diferentes, todas baseadas em vizinhos mais próximos”, ou seja, KNN. (ESCALERA, 2020).

Considerando que os dados se encontravam desbalanceados, foi empregada a técnica de balanceamento *NearMiss*. Em linhas gerais, supõe-se que um conjunto de dados tenha 2.000 registros, dos quais, 500 pertencem à classe 1 e 1500 pertencem à classe 2, como demonstra o gráfico 6.

Gráfico 6. Gráfico que exemplifica conjunto de dados desbalanceados.



Fonte: autoria própria.

Com uma técnica de balanceamento, a classe majoritária terá seu tamanho reduzido tornando iguais o número de amostras de cada classe, no caso, formar o conjunto de 1000 registros com 500 em cada. E assim ocorre para melhorar o processo de classificação não fazendo com que o aprendizado seja majoritário em apenas uma classe.

Conforme aponta Silva (2020), há três versões de aplicação da técnica de balanceamento *NearMiss*, o que se verifica no quadro 6.

Quadro 9. Versões da técnica de balanceamento NearMiss.

Versão	***Modo de operação***
NearMiss1	Seleciona as amostras da classe majoritária que estão mais próximas das demais classes para a subamostragem. é calculada a distância e os elementos mais próximos são selecionados.
NearMiss2	Seleciona as amostras com maior distância média aos N vizinhos das demais classes
NearMiss3	Inicialmente para cada amostra das classes minoritárias, os elementos da classe majoritária mais próximos são selecionados. Em seguida, as amostras selecionadas são aquelas para as quais a distância média até o N vizinho mais próximo é a maior.

Fonte: autoria própria baseada em Silva (2020).

Em seus trabalhos, Escalera (2020) selecionou a terceira versão do NearMiss por ter entendido “ser a menos suscetível a ruído” e uma característica é permitir “selecionar o número de instâncias que as classes majoritárias terão uma vez reduzida”. E isso justifica a escolha por esse método para aplicação.

2.3.7 Tecnologias utilizadas

Para a construção do protótipo foi utilizada a linguagem **Python 3**. A literatura a aponta como uma linguagem poderosa vistos os seus recursos e bibliotecas, os quais pode-se ver adiante e que têm forte apego à endentação.

Como plataforma para desenvolvimento, o aplicativo web **Jupyter Notebook** tem suporte a diversas linguagens de programação. É capaz de criar e compartilhar documentos com código ativo e pode produzir saída iterativas. Disponível no sítio https://jupyter.org/.

2.3.7.1 *Bibliotecas*

Uma das mais populares bibliotecas para a visualização de dados e geração de gráficos é a **MatPlotLib** (https://matplotlib.org/). Com ela, é possível criar gráficos, histogramas e outras visualizações que são úteis para a visualização dos dados.

Uma outra biblioteca utilizada para visualização estatística de dados é a **Seaborn** (https://seaborn.pydata.org/) que constrói gráficos atraentes e informativos.

Um pacote fundamental para computação matemática em Python, a NumericalPython, mais conhecida como **NumPy** (www.numpy.org) oferece as bases matemáticas necessárias para a construção de aplicações de Inteligência Artificial além de fornecer suporte para arranjos simples ou bidimensionais (matrizes) e suas operações.

Para manipular estruturas de dados de forma rápida e expressiva, a **Pandas** (https://pandas.pydata.org/) ou PythonData AnalysisLibrary, muito utilizada para análise de dados.

Por fim, a **Scikit-Learn** (https://scikit-learn.org/), é referente à aprendizado de máquina de código aberto com exemplos e algoritmos para os diversos tipos de aprendizado. Contém também amostras de dados.

2.3.7.2 *Aplicação*

Flask é um framework Python que ajuda a construir aplicativos web a partir de design leve e modular, possuindo diversos recursos prontos para uso como servidor de desenvolvimento integrado e depurador rápido. Em conjunto, a utilização do **Jinja**, linguagem programação de modelos para Python, modelada a partir dos modelos (templates) do Django.

Para a construção da interface gráfica, **HTML 5** (Linguagem de Marcação de Hipertexto) é utilizado para construir a estrutura da página e seu conteúdo. Em conjunto, é utilizado **CSS 3** (*Cascading Style Sheets* ou Folhas de Estilo em Cascata) que descreve a aparência dos elementos, como cor de fundo; e **JavaScript**, conhecido como JS, é uma linguagem interpretada e baseada em funções. No caso em tela, utilizada na validação de dados de entrada da aplicação.

2.3.7.3 *Manipulação de arquivos*

Para manipulação dos arquivos do conjunto NHANES, após o download no site, foi utilizado o “SAS.UniViewr” que permite salvar em arquivo nos formatos .csv e .xml. Para visualizar os dados também é possível utilizar o “XPT(SAS)viewr”.

Cada arquivo foi salvo com extensão .xml porquanto se salvo em .csv implicaria em erros consoante aos dados de conteúdo decimal, gerando colunas sobressalentes por causa da separação por vírgula.

Com o arquivo .xml, deve ser aberto o Microsoft Excel. No menu “Desenvolvedor”, clicar em “Importar” para completar. Salvar em “CSV (MS-DADOS) (.csv)”. Esse formato preserva os números decimais pois utilizar ponto-e-vírgula (;) para separação dos dados.

Os arquivos foram importados para o SGBD Microsoft SQL Server (v 18.5.1) para a continuidade da manipulação dos dados.

A partir dos dados da tabela de prescrição de medicamentos e da lista do CID-10 contida na documentação do *dataset*, foi gerada a tabela “DIAGNOSTICO” formada pelo SEQN (número sequencial) contida em todas as tabelas e a CLASSE, sendo 1 para diabetes mellitus tipo 2.

Para tanto, foram utilizados comandos SQL no SGBD SQL Server:

- Para inserir ‘0’ em todos os registros:

```
USE DiabetesMellitus
UPDATE DIAGNOSTICO
SET CLASSE = 0
```

- Para inserir ‘1’ em todos os registros que contém o diagnóstico de diabetes mellitus tipo 2:

```
USE DiabetesMellitus
UPDATE DIAGNOSTICO
SET CLASSE = 1
WHERE RXDRSC1 LIKE '%E11%' OR
RXDRSC2 LIKE '%E11%' OR
RXDRSC3 LIKE '%E11%';
```

Com todas as tabelas no banco de dados, a partir do comando JOIN foi gerado um arquivo único (.csv) lançando seu uso para a geração do modelo preditivo.

2.4 DIABETES

A diabetes mellitus, conhecida como diabetes,

> é uma síndrome de etiologia múltipla, decorrente da falta de insulina e/ou da incapacidade de a insulina exercer adequadamente seus efeitos. Caracteriza-se por hiperglicemia crônica com distúrbios do metabolismo dos carboidratos, lipídeos e proteínas. (BRASIL, 2001).

Trata-se de patologia que é classificada como crônica e não transmissível, não tem cura, mas tem tratamento que contribui para a melhor qualidade de vida do paciente. É classificada em: TIPO 1 (insulinodependente), TIPO 2 (não insulinodependente), como as mais conhecidas, gestacional, a diabetes latente autoimune do adulto (LADA) e diabetes juvenil de início tardio (MODY).

2.4.1.1 *Diabetes Mellitus Tipo 2*

Dentre os tipos de diabetes, a diabetes mellitus tipo 2:

> resulta, em geral, de graus variáveis de resistência à insulina e de deficiência relativa de secreção de insulina (…) Sua

> incidência é 90% dos pacientes diabéticos. (…) Denomina-se resistência à insulina o estado no qual ocorre menor captação de glicose por tecidos periféricos (especialmente muscular e hepático), em resposta à ação da insulina. As demais ações do hormônio estão mantidas ou mesmo acentuadas. (BRASIL, 2001).

Também se faz necessário alertar que:

> Como grande parte dos indivíduos com DM2 é assintomática, o diagnóstico é feito tardiamente, o que leva, não raramente, à presença de complicações já na detecção inicial. Como consequência dessas complicações, os pacientes com diabetes apresentam elevada morbidade, redução na expectativa de vida e mortalidade 2 a 3 vezes maior do que aqueles não afetados. (MILECH, 2017).

2.4.2 Complicações

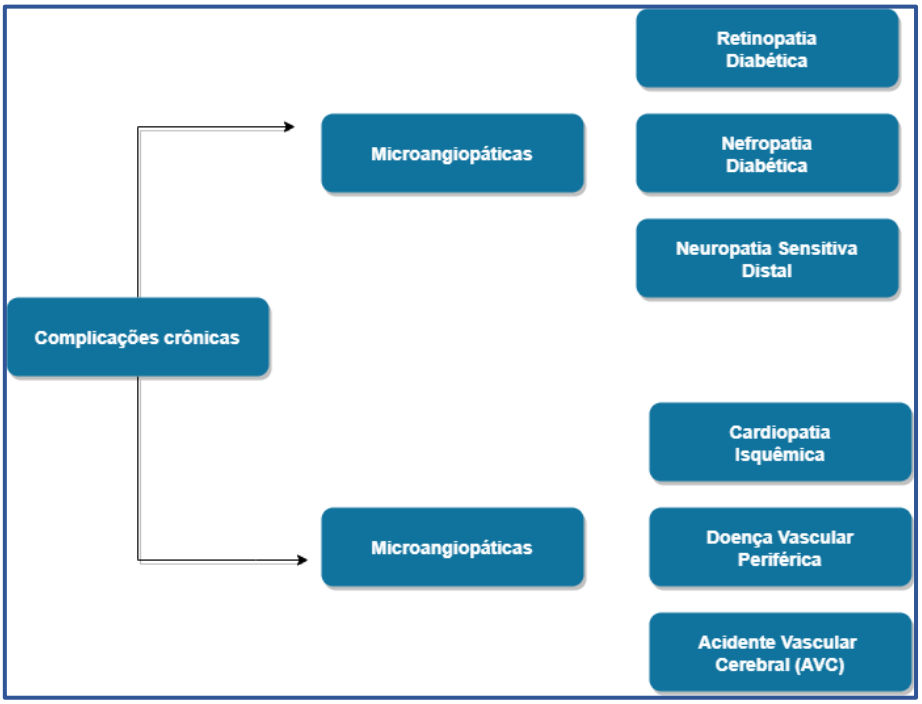
É de fundamental importância que haja o diagnóstico da patologia tão cedo for possível, para que o paciente não sofra ou evite as complicações que a doença induz, assim como para inibir a redução de expectativa de vida ou a maior mortalidade dos afetados. Num quadro geral,

> Sua principal manifestação, a hiperglicemia, lesa vários órgãos e sistemas do corpo, principalmente olhos, coração, rins e sistema nervoso, com potencial incapacidade causada pela diminuição da funcionalidade física, psicológica e social da pessoa, o que leva ao seu desajuste. Muitas vezes, o DM2 está associado a fatores de risco como hiperlipidemia caracterizada por níveis elevados de triglicerídeos e colesterol que por sua vez desencadeiam problemas cardiovasculares como hipertensão, aterosclerose, angina coronária, infarto do miocárdio, fatores que afetam a adaptação pessoa e sua vida produtiva. Por esse motivo, o controle glicêmico é fundamental para a pessoa com DM2 para reduzir o risco de complicações em longo prazo. (LAZCANO-ORTIZ, 2009).

Explicitando um pouco mais as complicações, é pertinente anotar que para Scheffel (2004), SBD (2019) e IDF (2019), a diabetes mellitus tipo 2 está associada ao aparecimento de dois tipos de complicações crônicas, as macros e

as microangiopáticas. Tais complicações estão demonstradas na figura 9, que apresenta os tipos e exemplos.

Figura 9. Quadro geral das complicações crônicas decorrentes da diabetes mellitus tipo 2.



Fonte: autoria própria baseada em Scheffel (2004).

Complicações crônicas podem ser classificadas como:

a) Macroangiopáticas, a exemplo de:

- Cardiopatia isquêmica
- Doença vascular periférica
- Acidente vascular cerebral, conhecido pela sigla AVC.

O aumento de glicose no sangue aliado a colesterol e pressão arterial altos promovem a formação de placas de colesterol que entopem artérias e coágulos. Atingem o coração na obstrução nas artérias coronárias, causando cardiopatia isquêmica; quando estreita ou obstrui os vasos sanguíneos que levam o sangue às extremidades, causando doença vascular periférica; ou quando a obstrução da artéria impede o fluxo de sangue ao cérebro, causando acidente vascular cerebral (AVC).

b) Microangiopáticas, a exemplo de:

- Retinopatia diabética

Designa todas os problemas de retina causados pelo diabetes. Em caso de edema macular (micro vasos do globo ocular), a visão embaça e pode causar a incapacidade visual parcial ou total (cegueira).

- Nefropatia diabética

A hiperglicemia prolongada causa lesões e disfunções renais, afetando a capacidade de filtragem dos rins pela sobrecarga e perda de moléculas de proteínas pela urina (albumina). Pode gerar a perda de capacidade de filtragem (insuficiência renal crônica) assim como a necessidade de hemodiálise e transplantes.

- Neuropatia sensitiva distal

Causa danos aos nervos. Dentre os sintomas: dor como ardência ou picada, dormência nos pés e pernas, fraqueza e perda de sensibilidade no pé, dificultando a percepção de calor, frio e mesmo de algum machucado. Sem enfaixamento, a lesão pode se infeccionada. Ela é responsável por cerca de dois terços das amputações não traumáticas (que não são causadas por acidentes e fatores externos).

Dentre as complicações agudas, temos cetoacidose diabética que é “um distúrbio metabólico complexo que ocorre quando o fígado degrada a gordura em uma taxa excessiva. O subproduto desse processo, os corpos cetônicos, pode tornar o sangue perigosamente ácido”. (IDF, 2010, tradução nossa).

2.4.3 Considerações sobre os *Pima Indians* e inferências sociais

Uma constante nas pesquisas é o conjunto de dados relativo aos índios Pima, nativos americanos e que formam um grupo homogêneo. Ocorre que, conforme aponta Ribeiro (2009), os índios Pima tiveram grande aumento da prevalência de diabetes que tem como hipotético resultado “da interação da predisposição genética com uma súbita mudança da dieta tradicional de produtos agrícolas para alimentos industrializados do século passado”.

Isso se mostra ainda mais forte quando a autora traz a comparação dos Pima, que são geneticamente semelhantes e vivem no México que, ao contrário dos norte-americanos, tem taxa zero de diabetes mellitus tipo 2.

Denota-se que, sendo geneticamente semelhantes, poderiam ter igual ou maior índice de diabetes mellitus tipo 2. No entanto, não é o que se verifica. E este é um fator importante: a alimentação.

Ribeiro (2009) cita “mudança da dieta tradicional de produtos agrícolas para alimentos industrializados”. É uma constatação da importância do fator ambiental no controle da doença que tem sido agravado pela própria evolução, com alto consumo de alimentos formados por excesso de gorduras, açúcares, sal.

Se nossos genes não podem ser alterados, resta-nos fazer o que nos cabe: agir no fator ambiental. Nesse ponto, notamos a atitude do indivíduo, na sua luta pessoal, mas também que esse deve ser alvo de políticas públicas consistentes para evitar outras epidemias de doenças crônicas.

Sabe-se que:

> A identificação precoce dos casos e o estabelecimento do vínculo entre os portadores da doença e as unidades de saúde são elementos imprescindíveis para o sucesso do controle desses agravos. O acompanhamento e o controle do diabetes mellitus no âmbito da atenção básica pode evitar o surgimento e a progressão das complicações, reduzindo o número de internações hospitalares, bem como a mortalidade devido a esses agravos.
>
> Também é de conhecimento que a diabetes é uma doença crônica, uma das causas de morte prematura, como visto nas motivações desta pesquisa. É de fundamental importância o diagnóstico precoce para que seja iniciado o tratamento, trazendo maior qualidade de vida para o paciente ao evitar complicações uma vez que “o diabetes também acarreta um prejuízo social, já que é responsável pelo aumento da mortalidade precoce e por muitas incapacitações”. (RIBEIRO, 2009).

Maneo (2020) estampa a marca de 1 milhão de mortos de COVID, marca superior ao número de vítimas de guerras, como a Guerra do Paraguai (1864-1870) e infere dados importantes: em 2006, 1.600.000 pessoas morreram devido à diabetes, números apresentados na tabela 8, acompanhado dos números de outras endemias e pandemias como a que ficou conhecida como gripe espanhola.

Tabela 8. Números da covid e outras doenças que causam grande número de mortes.

	Número de mortes	Período	Acumulado	Período acumulado
Pandemia de gripe espanhola			50.000.00	1918-1919
Câncer	9.600.000	2019		
Doenças cardíacas	9.400.000	2016		
Diabetes	1.600.000	2016		
Tuberculose	1.500.000	2018		
Pandemia de coronavírus	1.000.000	9.jan – 28.set (8 meses)		
HIV/Aids	690.000	2019	33.000.000	Desde 1990
Malária	405.000	2019		
Pandemia de H1N1			18.449	2009-2010
Cólera	5.654	2017	Entre 21 mil e 143 mil/ano	Desde 1989
Ebola	2.287	2018-2020	13.597	Desde 2014

Fonte: autoria própria baseada em Maneo (2020).

A mesma reportagem ainda mostra que a pandemia do Coronavírus, como as guerras, atingem os mais pobres: de fato, as orientações da quarentena implementada pelas autoridades determinam medidas de distanciamento social e de higiene. Infelizmente, a disseminação da covid mostrou as mazelas da administração pública, pois os mais pobres vivem em casas minúsculas, muitas vezes sem água potável e saneamento básico.

Também atinge os mais pobres pela baixa oferta alimentação mais variada, com menos frutas e verduras frescas, e mais industrializada, características que favorecem a ocorrência de diabetes. Essa e outras características dos processos de urbanização estão associadas ao efeitos sobre fatores de risco da diabetes mellitus tipo 2, o que pode ser observado no quadro 7.

Quadro 7. Características da urbanização que favorecem a ocorrência de diabetes tipo 2 em pessoas geneticamente predispostas.

Processos de urbanização	***Efeito sobre fatores de risco de diabetes tipo 2***
Melhores condições socioeconômicas, melhor higiene, acesso a serviços médicos (acompanhamento da gravidez, vacinações etc.)	Aumento da expectativa de vida
Alimentação mais acessível e variada	Aumento da expectativa de vida
Alimentação mais “industrializada”, com excesso de gordura, açúcares e sal	Aumento do sobrepeso e da obesidade
Redução da atividade física: no trabalho, nos meios de transporte, nas opções de lazer	Aumento do sobrepeso e da obesidade. Sedentarismo.

Fonte: autoria própria baseada em Bandeira (et al., 2015).

Pode-se verificar que o aumento de sobrepeso e obesidade significam o aumento do IMC que, conforme a avaliação de risco da *American Diabetes Federation*, aumenta o risco de diabetes.

Além dos direitos básicos elencados no artigo 5º da Constituição Federal de 1988, a Carta Magna traz no artigo 196 o preceito sobre a saúde que fundamenta o questionamento sobre o papel do Estado, nas suas diversas esferas, de prover medidas para evitar que tantas vidas sejas ceifadas ante sua omissão:

> Art. 196. A saúde é direito de todos e dever do Estado, garantido mediante políticas sociais e econômicas que visem à redução do risco de doença e de outros agravos e ao acesso universal e igualitário às ações e serviços para sua promoção, proteção e recuperação.

Os estudos mostram que “A maioria das pessoas com diabetes tipo 2 vivem em países de renda baixa e média. Nesses e outros países, deve-se priorizar os esforços de colaboração para a prevenção primária do diabetes tipo 2 e outras doenças não transmissíveis a nível social.” (IDF, 2010, tradução nossa) além de direcionar recursos em investigações epidemiológicas

Assim como para Noble (et al., 2011), os autores afirmam que “os planejadores e comissários usam padrões de risco para direcionar recursos para cuidados de saúde preventivos para certos subgrupos”

Em suma, Gaytán-Hernández (et al., 2017), apontam que:

> Considerem a necessidade de reestruturar as políticas de saúde com o objetivo de tornar eficaz sua aplicação, gerando ações e estratégias que garantam a prevenção entre os diferentes setores (educacional, econômico e social.) e sensibilizar as famílias para que tenham estilos de vida saudáveis.

2.4.4 Conjuntos de dados

2.4.4.1 *Datasets candidatos à utilização*

Para a criação de um modelo preditivo é imprescindível a utilização de um conjunto de dados que será particionado para o treinamento e, posteriormente,

para o teste do modelo. Para obtenção dos *datasets*, foram pesquisados nos sites *Data.world*, *UCI Machine Learning* e *Google* com o termo “Diabetes”.

Do total verificado, após leitura mais cuidadosa, foram selecionados os conjuntos com maior proximidade de acordo com a necessidade do modelo, que é o diagnóstico da diabetes mellitus tipo 2.

2.4.4.2 *Early Stage diabetes risk prediction dataset*

O quadro 8 apresenta o conjunto de dados de previsão de risco de diabetes em estágio inicial (ISLAM et al., 2020, tradução nossa). Esse conjunto é composto por 520 registros, contendo 16 variáveis de entrada e uma classe alvo. Denota-se, além dos dados idade e gênero, as demais variáveis são questionamentos sobre sintomas, por exemplo, a presença ou ausência de alopecia, obesidade entre outras.

Quadro 8. Conjunto de dados de previsão de risco de diabetes em estágio inicial (ISLAM et al., 2020, tradução nossa).

Variável	***Descrição***	***Valor***
Age	Idade	(20-65)
Gender	Gênero	Male (Masculino)
		Female (Feminino)
Polyuria	Poliúria	Sim, Não
Polydipsia	Polidipsia	
Sudden weight loss	Perda súbita de peso	
Weakness	Fraqueza	
Polyphagia	Polifagia	
Genital thrush	Tordo genital	
Visual blurring	Embaçamento visual	
Itching	Comichão	
Irritability	Irritabilidade	
Delayed healing	Cura retardada	
Partial paresis	Paresia parcial	
Muscle stiffness	Rigidez muscular	
Alopecia	Alopecia	
Obesity	Obesidade	
Class	Classe	Positive (Positivo)
		Negative (Negativo)

Fonte: autoria própria a partir de Islam (et al., 2020, tradução nossa).

2.4.4.3 *Pima Indians*

A quadro 9 apresenta o conjunto de dados de diabetes dos índios Pima (SIGILITTO, 1990, tradução nossa). O conjunto é composto por 768 registros, dos quais 500 são não diabéticos e 268, diabéticos, contendo 8 variáveis de entrada e uma classe alvo.

É composto por dados invasivos, obtidos por análise laboratorial de coleta sanguínea, como insulina. Outros dados são não invasivos, como obtenção do IMC, resultante da equação de altura e peso, além do número de gestações.

Quadro 9. Conjunto de dados de diabetes dos índios Pima.

Variável	*Significado*	*Tipo*
preg	Número de vezes que engravidou	Numérico
plas	Concentração de glicose plasmática a 2 horas em um teste oral de tolerância à glicose	Numérico
pres	Pressão arterial diastólica (mm Hg)	Numérico
skin	Espessura da dobra da pele do tríceps (mm)	Numérico
insu	Insulina sérica de 2 horas (mu U / ml)	Numérico
mass	Índice de massa corporal	flutuante
pedi	Função de pedigree de diabetes	Ponto flutuante
age	Anos de idade	Numérico
class	Classe	Texto

Fonte: autoria própria a partir de Sigilitto (1990, tradução nossa).

2.4.4.4 *Sample: Diabetes (Azure).*

Outro conjunto de dados é apresentado no quadro 10, o conjunto de amostras de aprendizado de máquina: diabetes (AZURE, 2020, tradução nossa). É composto por 442 registros, contendo 10 variáveis de entrada e uma medida quantitativa de progressão de doença um ano após a linha de base.

As variáveis transitam entre idade, gênero, medidas corporais e dados invasivos, resultantes de exames laboratoriais como a mensuração de lipoproteínas e hormônios.

Quadro 10. Conjunto de amostras de aprendizado de máquina: diabetes.

Atributo	*Significado*	*Tipo*
AGE	Idade em anos	Numérico
SEX	Sexo	
BMI	índice de massa corporal	Ponto flutuante
BP	pressão arterial média	Numérico
S1	s1 tc, células T (um tipo de células brancas do sangue)	
S2	ldl, lipoproteínas de baixa densidade	Ponto flutuante
S3	hdl, lipoproteínas de alta densidade	Numérico
S4	tch, hormônio estimulador da tireoide	
S5	ltg, lamotrigina	Ponto flutuante
S6	glu, nível de açúcar no sangue	Numérico
Y	medida quantitativa da progressão da doença um ano após a linha de base	

Fonte: autoria própria a partir de Azure (2020, tradução nossa).

De forma geral, a tabela 9 sintetiza os *datasets* e o números de registros de cada um, observando que nenhum deles alcança a marca de mil registros.

Tabela 9. Síntese de *Datasets* constando o número de entradas e registros.

Dataset	Entradas	Registros
Early stage diabetes risk prediction dataset (ISLAM et al., 2020)	16	520
Pima Indians Diabetes Database (SIGILITTO, 1990)	8	768
Machine Learning Samples: Diabetes (AZURE, 2020)	10	442

Fonte: autoria própria.

Nesse cenário, a melhor escolha se faz optando pelo *dataset* dos índios Pima, considerando o maior número de registros entre os itens da tabela 9. Notadamente, o número baixo de registros impacta significativamente no desempenho do modelo preditivo.

Em consideração à toda a importância de conjunto de dados, visto ser recorrentemente citado e no interesse da demonstração do desempenho, à frente serão explicitados resultados obtidos com o *dataset* índios Pima.

2.4.4.5 *NHANES – National Health and Nutrition Examination*

Outro conjunto de dados, o NHANES, Exame Nacional de Saúde e Nutrição, em CDC (2017, tradução nossa), é um programa de pesquisa que avalia a saúde de adultos e crianças norte-americanas. O Centro Nacional de Estatísticas de Saúde (NCHS), Divisão de Pesquisas de Exame de Saúde e Nutrição (DHANES), parte dos Centro de Controle e Prevenção de Doenças (CDC). A seleção de respondentes inclui fases pela seleção de condados, de municípios, investigação aprofundada do domicílio selecionado e a seleção aleatória de algoritmos em que a participação não é obrigatória.

Organização e estrutura dos dados:

- Dados demográficos: como a identificação por idade e sexo do respondente.
- Dados dietéticos: destacando os hábitos alimentares.
- Dados de exame, como medidas corporais, exames de imagem e radiológicos.
- Dados de laboratório, com análises de material coletado (sangue, urina).
- Dados do questionário, um amplo conjunto de perguntas sobre a saúde em diversos aspectos.
- Dados de acesso limitado; resultados de conteúdo sensível.

Os dados do período 2019-2020 ainda não estavam disponíveis na ocasião da coleta, por isso, os dados abaixo são do conjunto 2017-2018 e, ainda que disponíveis, a pandemia da covid certamente gerou influência nos dados obtidos.

Em continuação, são elencados os dados já selecionados dentro de seus respectivos grupos como acima especificados. A seleção dos atributos ocorreu a partir dos conhecimentos pessoais e os acumulados durante esta pesquisa.

Por exemplo, seleção de idade e sexo, dois atributos que compõem a avaliação de risco da American Diabetes Associates, que se encontram no Apêndice A, assim como indicadores como hemoglobina glicada, pressão arterial, colesterol e a exclusão de itens como código postal ou identificadores.

Em sua origem, os dados estão separados em diversos arquivos e foram manuseados e observadas suas características, como o número sequencial denominado SEQN, um elo entre as tabelas, utilizando os conceitos de chave primária e chave estrangeira para criação de índices que possibilitam o processamento de comandos de união.

Os quadros seguintes elencam as informações de variável utilizada, sua descrição, o conjunto de códigos ou valores, e a descrição dos valores previstos.

Dados demográficos: nesta primeira seção, o quadro 10 mostra as variáveis selecionadas do arquivo “Variáveis demográficas e pesos de amostra”, sendo o sexo biológico do participante e a idade em anos.

Quadro 10. Dados selecionados de Dados demográficos: Variáveis demográficas (NHANES).

Variável	*Descrição*	*Código ou Valor*	*Descrição do valor*
RIAGENDR	Sexo	1	Masculino
		2	Feminino
			Ausência
RIDAGEYR	Idade	0 a 79	Faixa de valores
		80	80 anos de idade e mais
			Ausência

Fonte: autoria própria com base em CDC, 2017, tradução nossa.

Dados de Exame: na seção de dados de exames são compostos de dados de exame físico, com medições não invasivas. Foram selecionadas duas subseções: pressão arterial e medidas corporais.

Na subseção pressão arterial, no quadro 11, infere medidas como a leitura da sistólica e diastólica; na subseção medidas corporais, no quadro 12, infere medidas como peso e altura, necessários ao cálculo de Índice de massa corpórea (IMC), uma das medidas mais consideradas para estado de saúde física.

Quadro 11. Dados selecionados de Dados de exame: Pressão arterial (NHANES).

Variável	*Descrição*	*Código ou Valor*	*Descrição do valor*
BPACSZ	Tamanho codificado do manguito: Tamanho do manguito (cm) (largura X comprimento).	1	Infantil(6x12)
		2	Criança (9 x 17)
		3	Adulto (12 x 22)
		4	Grande (15 x 32)
		5	Coxa (18 X 35)
		.	Ausência
BPXSY1	Sistólica: pressão arterial (primeira leitura) mm Hg	72 a 228	Faixa de Valores
		.	Ausência
BPXDI1	Diastólica: pressão arterial (primeira leitura) mm Hg	0 a 136	Faixa de Valores
		.	Ausência

Fonte: autoria própria com base em CDC, 2017, tradução nossa.

Quadro 12. Dados selecionados de Dados de exame: Medidas corporais (NHANES).

Variável	*Descrição*	*Código ou Valor*	*Descrição do valor*
BMXWT	Peso (kg)	3,2 a 242,6	Faixa de Valores
		.	Ausência
BMXHT	Altura em pé (cm).	78,3 a 197,7	Faixa de Valores
		.	Ausência
BMXBMI	Índice de Massa Corporal	12,3 a 86,2	Faixa de Valores
		.	Ausência
BMXARMC	Circunferência do braço (cm).	11,2 a 56,3	Faixa de Valores
		.	Ausência
BMXWAIST	Circunferência da cintura (cm)	40 a 169,5	Faixa de Valores
		.	Ausência

Fonte: autoria própria com base em CDC, 2017, tradução nossa.

No tocante à circunferência da cintura, em Freitas (et al., 2018), não há um consenso sobre os valores de corte, mas é controverso que o aumento da circunferência abdominal é causa de riscos à saúde. Na tabela 9 estão referenciadas as medidas adequadas e inadequadas de circunferência da cintura para homens e mulheres segundo organizações como o IDF e NCEP.

Tabela 10. Medidas de circunferência abdominal segundo a Federação Internacional de Diabetes (IDF) e Programa Nacional de Educação sobre o Colesterol dos Estados Unidos (NCEP).

	Adequado		Inadequado	
	Homens	Mulheres	Homens	Mulheres
IDF	< 90cm	< 80cm	≥ 90cm	≥ 80cm
NCEP	< 88cm	< 102cm	≥ 88cm	≥ 102cm

Fonte: autoria própria baseada em Freitas (et al., 2018).

Dados de Laboratório: nesta seção, são apresentados dados invasivos, a partir de coleta de materiais (sangue e urina) do participante. Dentre as subseções, quadro 13, são elencados os dados como Albumina e Creatinina a partir de coleta de urina, que tem relação com o funcionamento renal, órgão que sofre com a falta de tratamento da DM2.

Quadro 13. Dados selecionados de Dados de laboratório: Albumina e creatinina – urina (NHANES).

Variável	*Descrição*	*Código ou Valor*	*Descrição do valor*
URXUMA	Albumina, urina (ug/mL)	0,21 a 14040	Faixa de Valores
		.	Ausência
URXUMS	Albumina, urina (mg/L)	0,21 a 14040	Faixa de Valores
		.	Ausência
URXUCR	Creatinina, urina (mg/dL)	3,54 a 621	Faixa de Valores
		.	Ausência
URXCRS	Creatinina, urina (umol/ L)	312,9 a 54896,4	Faixa de Valores
		.	Ausência
URDACT	Razão de creatinina albumina (mg / g)	0,27 a 11.676,92	Faixa de Valores
		.	Ausência
	URXUMA / URXUCR x 100, arredondado para 0,01		

Fonte: autoria própria com base em CDC (2017, tradução nossa).

Nesta subseção, no quadro 14, são elencados os dados das frações HDL do Colesterol. Popularmente denominado colesterol bom, ajuda a não formação de acúmulo de gorduras nas artérias.

Quadro 14. Dados selecionados de Dados de laboratório: Colesterol – Lipoproteína de alta densidade (HDL) (NHANES).

Variável	Descrição	Código ou Valor	Descrição do valor
LBDHDD	Colesterol HDL direto (mg / dL)	10 a 189	Faixa de Valores
		.	Ausência
LBDHDDSI	Colesterol HDL direto (mmol / L)	0,26 a 4,89	Faixa de Valores
		.	Ausência

Fonte: autoria própria com base em CDC (2017, tradução nossa).

Nesta subseção, no quadro 15, são elencados os dados das frações triglicerídeos do Colesterol, que compõem o perfil lipídico.

Quadro 15. Dados selecionados de Dados de laboratório: Triglicerídeos (NHANES).

Variável	*Descrição*	*Código ou Valor*	*Descrição do valor*
LBXTR	Triglicerídeo (mg / dL)	10 a 2.684	Faixa de Valores
		.	Ausência
LBDTRSI	Triglicerídeo (mmol / L)	0,113 a 30,302	Faixa de Valores
		.	Ausência

Fonte: autoria própria com base em CDC (2020, tradução nossa).

Nesta subseção, no quadro 16, são elencados os dados totais de Colesterol, composto pela soma de todos os tipos de colesterol.

Quadro 16. Dados selecionados de Dados de laboratório: Colesterol – Total (NHANES).

Variável	*Descrição*	*Código ou Valor*	*Descrição do valor*
LBXTC	Colesterol Total (mg/dL).	76 a 446	Faixa de Valores
		.	Ausência
LBDTCSI	Colesterol Total (mmol/L)	1,97 a 11,53	Faixa de Valores
		.	Ausência

Fonte: autoria própria com base em CDC (2017, tradução nossa).

Nesta seguinte subseção, no quadro 17, são elencados os dados de ferritina presentes. A ferritina é um indicador da síntese do ferro no corpo além de ter sido observada em investigações em trabalhos ligados à obesidade.

Quadro 17. Dados selecionados de Dados de laboratório: Ferritina (NHANES).

Variável	Descrição	Código ou Valor	Descrição do valor
LBXFER	Ferritina (ng / mL)	1,04 a 5190	Faixa de Valores
		.	Ausência
LBDFERSI	Ferritina (ug / L)	1,04 a 5190	Faixa de Valores
			Ausência

Fonte: autoria própria com base em CDC (2017, tradução nossa).

Nesta subseção, no quadro 18, são elencados os percentuais de hemoglobina glicada, encontrada na literatura médica com os nomes glicohemoglobina, hemoglobina glicosilada, hemoglobina A1c ou HbA1c, entre outros.

A hemoglobina glicada é uma forma de medir a glicose sanguínea em um período de 2 ou 3 meses. Por diversos fatores, como a utilização de medicamento ou a realização de atividade física vigorosa no dia anterior à coleta, a dosagem de glicemia em jejum pode ser comprometida.

Nos casos em que a hiperglicemia ou alta dosagem de glicose sanguínea não for inequívoca, são solicitados a análise de alteração em exames

como curva glicêmica ou hemoglobina glicada, o que é mais bem aceito pelo paciente por não requerer preparo além do jejum mínimo de 8 horas.

Quadro 18. Dados selecionados de Dados de laboratório: Hemoglobina glicada (HbA1c) (NHANES).

Variável	*Descrição*	*Código ou Valor*	*Descrição do valor*
LBXGH	HbA1c (%).	3,8 a 16,2	Faixa de Valores
		.	Ausência

Fonte: autoria própria com base em CDC (2017, tradução nossa).

Informações extraídas de laudos laboratoriais da autora apontam que, para valores inferiores à 5,7%, considera-se normal; o intervalo de 5,7 a 5,4%, significa risco aumentado para diabetes mellitus; igual ou superior à 6,5%, representa caso de diabetes. E que a ADA recomenda como meta para o tratamento de pacientes diabéticos os resultados de HbA1c igual ou superior à 7%

Nesta subseção, no quadro 19, são elencados os percentuais presentes de insulina, hormônio produzido pelo pâncreas. O aumento de glicose no sangue é um estimulador da produção de insulina.

Quadro 19. Dados selecionados laboratório: Insulina (NHANES).

Variável	Descrição	Código ou Valor	Descrição do valor
LBXIN	Insulina (uU/mL)	0,71 a 485,1	Faixa de Valores
		.	Ausência
LBDINSI	Insulina (pmol/L)	4,26 a 2910,6	Faixa de Valores
		.	Ausência

Fonte: autoria própria com base em CDC (2017, tradução nossa).

Nesta subseção, no quadro 20, são elencados os percentuais presentes de glicose em coleta sanguínea após jejum. Observa-se que o período mínimo de 8 horas e máximo de 12 horas.

Quadro 20. Dados selecionados de laboratório: Glicose plasmática de jejum (NHANES).

Variável	*Descrição*	*Código ou Valor*	*Descrição do valor*
LBXGLU	Glicose de jejum (mg/dL)	47 a 451	Faixa de Valores
		.	Ausência
LBDGLUSI	Glicose em jejum (mmol/L)	2,61 a 25	Faixa de Valores
		.	Ausência

Fonte: autoria própria com base em CDC (2017).

Nesta seguinte, no quadro 21, são elencados os valores de albumina, creatinina, triglicérides, entre outros, de material sangue, referentes ao perfil bioquímico.

Quadro 21. Dados selecionados de Dados de laboratório: Perfil padrão de bioquímica (NHANES).

Variável	Descrição	Código ou Valor	Descrição do valor
LBXSAL	Albumina, soro refrigerado (g / dL)	2,1 a 5,4	Faixa de Valores
		.	Ausência
LBDSALSI	Albumina, soro refrigerado (g / L)	21 a 54	Faixa de Valores
		.	Ausência
LBXSCR	Creatinina, soro refrigerado (mg / dL)	0,25 a 12,74	Faixa de Valores
		.	Ausência
LBDSCRSI	Creatinina, soro refrigerado (umol / L)	22,1 a 1126,22	Faixa de Valores
		.	Ausência
LBXSGL	Glicose, soro refrigerado (mg / dL)	47 a 626	Faixa de Valores
		.	Ausência
LBDSGLSI	Glicose, soro refrigerado (mmol / L)	2,61 a 34,75	Faixa de Valores
		.	Ausência
LBXSCH	Colesterol, soro refrigerado (mg / dL)	77 a 438	Faixa de Valores
		.	Ausência
LBDSCHSI	Colesterol, soro refrig (mmol / L)	1.991 a 11.327	Faixa de Valores
		.	Ausência
LBXSTR	Triglicerídeos, soro refrig (mg / dL)	25 a 2923	Faixa de Valores
		.	Ausência
LBDSTRSI	Triglicerídeos, soro refrig (mmol / L)	0,282 a 33,001	Faixa de Valores
		.	Ausência

Fonte: autoria própria com base em CDC (2017, tradução nossa).

No quadro 22 são elencados os medicamentos e diagnósticos de cada respondente.

Quadro 22. Dados selecionados laboratório: Prescrição de medicamentos.

Variável	Código ou Valor	Descrição do valor
RXDRSC1	Código 1 CID-10-CM	Valor foi registrado
	55555	Desconhecido
	77777	Recusou
	99999	Não sei
	<em branco>	Ausência
RXDRSC2	Código 2 CID-10-CM	Valor foi registrado
	<em branco>	Ausência
RXDRSC3	Código 3 CID-10-CM	Valor foi registrado
	<em branco>	Ausência
RXDRSD1	Descrição do código 1 CID-10-CM	Valor foi registrado
	<em branco>	Ausência
RXDRSD2	Descrição do código 2 ICD-10-CM	Valor foi registrado
	<em branco>	Ausência
RXDRSD3	Descrição do código 3 CID-10-CM	Valor foi registrado
	<em branco>	Ausência

É importante a leitura desta tabela com os códigos (Código Internacional de Doenças) CID-10 abaixo:

- E11 – Diabetes mellitus não-insulino-dependente
- E11.0 – Diabetes mellitus não-insulino-dependente – com coma

- E11.1 – Diabetes mellitus não insulino-dependente – com cetoacidose
- E11.2 – Diabetes mellitus não insulino-dependente – com complicações renais
- E11.3 – Diabetes mellitus não insulino-dependente – com complicações oftálmicas
- E11.4 – Diabetes mellitus não insulino-dependente – com complicações neurológicas
- E11.5 – Diabetes mellitus não insulino-dependente – com complicações circulatórias periféricas
- E11.6 – Diabetes mellitus não insulino-dependente – com outras complicações especificadas
- E11.7 – Diabetes mellitus não insulino-dependente – com complicações múltiplas
- E11.8 – Diabetes mellitus não insulino-dependente – com complicações não especificadas
- E11.9 – Diabetes mellitus não insulino-dependente – sem complicações

As informações do quadro 22, juntamente da lista dos códigos CID-10 são necessárias para estabelecimento do atributo CLASSE, referente à variável preditiva. Mais informações estão dispostas no item 2.3.7 Tecnologias utilizadas.

2.4.4.6 *Conjunto de dados encontrados nos artigos selecionados para Bibliografia*

Em Ribeiro (2009), foi realizada a classificação de pacientes em diabéticos e não diabéticos do tipo 2. O quadro 23 explicita a base de índios Pima que é citada por Basso (et al., 2014) e Souza (2019).

Quadro 23. Conjunto *Pima Indians* constante dos artigos.

Tipo	Descrição
Invasivas	Concentração de glicose no sangue(mg/dl)
	Quantidade de insulina em 2 hs de jejum (um U/ml);
Não invasivas	Número de vezes que a mulher engravidou;
	Pressão arterial diastólica (mmHg);
	IMC (Índice de massa corporal)
	Função de ocorrência de casos da doença na família;
	Idade (anos);
	Medida de espessura da dobra cutânea do tríceps (mm)

Fonte: autoria própria baseado em Ribeiro (2009), Basso (et al. 2014) e Souza (2019).

Em Carvalho (et al. 2015), a busca a identificação de beneficiários com potencial de desenvolvimento de doença crônica (Diabetes Mellitus tipo 2) é realizada utilizando base de dados de uma operadora de plano de saúde do Estado do Paraná com total bruto de 43.375 registros, os quais estão explicitas suas variáveis e descrições no quadro 24, e que foram divididos na proporção: dados de treino (66%) e teste (34%).

O período de avaliação para todas as solicitações é de 6 anos.

Quadro 24. Base de dados de uma operadora de plano de saúde do Estado do Paraná.

Tipo	Variável	Descrição
Exames especiais	Qt_exa_map_ret	Quantidade total de exames de mapeamento de retina
	Qt_exa_cur_gli	Quantidade total de exames de curva glicêmica
Exames laboratoriais	Qt_exa_creatinina	Quantidade total de exames de creatinina
	Qt_exa_glicose	Quantidade total de exames de glicose(glicemia)
	Qt_exa_microal	Quantidade total de exames de microalbuminuria
	Qt_exa_coles	Quantidade total de exames de colesterol total
Consultas especialidades	Qt_cons_nefro	Quantidade total de consultas realizadas com nefrologista
	Qt_cons_oftal	Quantidade total de consultas realizadas com oftalmologista
	Qt_cons_endo	Quantidade total de consultas realizadas com endocrinologista
	Qt_cons_cardio	Quantidade total de consultas realizadas com cardiologista
Outras	sexo	Sexo do beneficiário (M ou F)
	idade	Idade do beneficiário
	St_cid_obesidade	Quantidade total de atendimentos realizados com o código CID de obesidade (E66)
	Classe	Atributo meta: indicativo, com indicativo e com forte indicativo.

Fonte: autoria própria com base em Carvalho (et al. 2015) e Delagassa (2009).

Olivera (et al., 2017) apresentam um estudo sobre várias técnicas e utilizou um conjunto de dados com 12.447 registros, com divisão de 70% para treino (8.738) e 30% para testes (3.709). No quadro 22, as variáveis categóricas e no quadro 25, as variáveis quantitativas.

Quadro 26.

Quadro 25. Variáveis categóricas na base de dados utilizada por Olivera (et al., 2017).

Variável	Descrição	Valores
a_ativfisica	Atividade física no lazer	1 Fraca 2 Moderada 3 Forte
a_bebexcessivo	Bebedor excessivo	0 = Não 1 = Sim
a_binge	Bebedor excessivo esporádico	0 = Não 1 = Sim
a_chdhard	Doença Coronariana grave autorreferida	0 = Não 1 = Sim
a_chdlight	Doença Coronariana leve autorreferida	0 = Não 1 = Sim
a_consdiafrutas	Consumo diário de frutas	0 = Não 1 = Sim
a_consdiaverduras	Consumo diário de verduras e legumes	0 = Não 1 = Sim
a_escolar	Escolaridade	1 = Fundamental incompleto 2 = Ensino médio incompleto 3 = Ensino médio completo 4 = Superior completo
a_fumante	Fumante	0 = Nunca fumou 1 = Ex fumante 2 = Fumante
a_gidade	Grupo de idade do participante	1 = 35 a 44 2 = 45 a 54 3 = 55 a 64 4 = 65 a 74
a_imc2	Índice de Massa Corporal com 4 categorias	1 = Magreza, 2 = Eutrofia 3 = Sobrepeso 4 = Obesidade
a_medanthipert	Uso de medicamento antihipertensivo	0 = Não, 1 = Sim
a_medoutahip	Uso de outros antihipertensivos	0 = Não 1 = Sim
a_medredlip	Uso de hipolipemiantes	0 = Não usa, 1 = Uso de estatina 2 = Uso de outro 3 = Uso de mais de um tipo
a_prvdcc	Prevalência de doença coronariana	0 = Não 1 = Sim
a_sfhfprem	Insuficiência Cardíaca (<50 anos) autorreferida	0 = Não 1 = Sim
a_sfmiprem	Infarto do Miocárdio (<50 anos) autorreferido	0 = Não 1 = Sim
a_sfrevprem	Revascularização (<50 anos) autorreferida	0 = Não 1 = Sim
a_sfstkprem	Derrame (<50 anos) autorreferido	0 = Não 1 = Sim
a_sintsono	Qualidade do sono	0 = Não 1 = Sim
a_sitconj	Situação conjugal	1 = Casado 2 = Divorciado 3 = Solteiro 4 = Viúvo 5 = Outro
claa2	Dor/desconforto penas quando anda(Q2)	0 = Não 1 = Sim
diea133	Consome café(Q133)	0 = Não, 1 = Sim, com cafeína, 2 = Sim, descafeinado
hfda07	Hipertensão na família(Q7)	0 = Não 1 = Sim
hfda11	Diabetes na família (Q11)	0 = Não 1 = Sim
hmpa08	Colesterol alto (Q8)	0 = Não 1 = Sim
rcta8	Sexo	1 = Masculino 2 = Feminino
a_dm	Indicação se o indivíduo tem ou não diabetes mellitus tipo 2	0 = Não 1 = Sim

Fonte: autoria própria baseado em Olivera (et al., 2017).

Quadro 26. Variáveis quantitativas na base de dados utilizada por Olivera (et al. 2017).

Variável	Descrição
a_cint	Circunferência da cintura(cm)
a_cons_est_nacl	Consumo diário de sal estimado
a_imc1	Índice de massa corporal(kg/m2)
a_rcq	Relação cintura-quadril
a_rendapercapita	Renda per capita
a_volalc	Quantidade de álcool ingerida por semana(ml)
afia7	Uso de bicicleta para transporte(dias/semana)

Fonte: autoria própria baseado em OLIVERA (et al., 2017).

Em síntese, o quadro 27 apresenta os artigos de acordo com os autores. Nota-se que o conjunto de dados dos índios Pima se apresenta três vezes. Isso porque, quando se busca um *dataset* relacionado à Diabetes Mellitus tipo 2, é um dos resultados mais incidentes.

Quadro 27. Síntese dos Conjuntos de dados encontrados na bibliografia.

Artigo	Dataset	Origem	Registros
RIBEIRO (2009), BASSO (et al., 2014), SOUSA (2019)	Pima Indians	Público	768
CARVALHO et al. (2015), e DELAGASSA (2009)	Base de dados de uma operadora de plano de saúde do estado do Paraná	Privado	Total bruto de 43.375 registros
OLIVERA (et al., 2017)	Dados do Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil)	Privado	12.247 registros

Fonte: autoria própria.

Assim como a base de dados da operadora de saúde, os dados do Estudo Longitudinal de Saúde do Adulto (ELSA -Brasil) não estão disponíveis, por serem dados privados.

3 EXPERIMENTOS

A partir da base de dados, foram selecionadas técnicas de aprendizado de máquina.

Para esta pesquisa, será predita uma categoria (diabetes mellitus tipo 2) e os dados estão rotulados, o que visualizamos com a classe alvo (0: não portador de diabetes mellitus tipo 2 e 1: portador de diabetes mellitus tipo 2).

Os algoritmos selecionados para investigação foram: KNN (K-Nearest Neighbors), Árvore de decisão, Naïve Bayes, utilizados nos trabalhos de Carvalho (et al., 2015) e Delagassa (2009); SVM (Support Vector Machine), utilizada por Ribeiro (2009); Regressão logística, presente em Olivera (et al., 2017); além do particular interesse em diligenciar as técnicas dos métodos Ensemble – AdaBoosting e Random Forest.

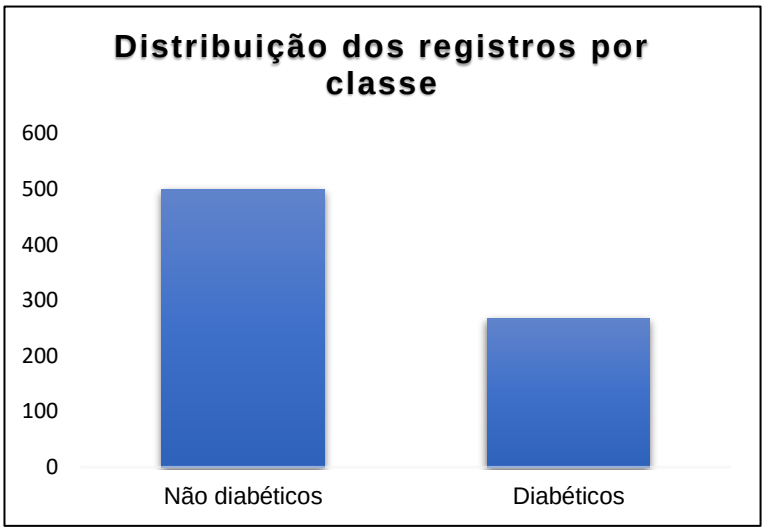
No sentido de avaliar o desempenho, foram escolhidas como métricas a acurácia e a matriz de confusão, conforme já explicitadas as razões ao final do item 2.3.5 Métricas de desempenho.

3.1 PIMA INDIANS

Em atenção a sua importância, citado em diversos trabalhos acadêmicos relacionados ao tema da pesquisa, este dataset é tomado para análise de desempenho.

Como característica dos dados, dos 768 registros, temos 500 registros sem diabetes mellitus tipo 2 e 268, com diabetes mellitus tipo 2, perfazendo respectivamente, 65,1% e 34,9%. São, portanto, dados desbalanceados e sua distribuição está apresentada no gráfico 7.

Gráfico 7. Gráfico da distribuição dos dados Pima Indians.



Fonte: autoria própria.

Submetendo o conjunto de dados à criação de um modelo preditivo, sem preparação de dados, utilizando os algoritmos selecionados, a tabela 11 apresenta a acurácia, a principal medida adotada, ordenada do maior para o menor valor resultante.

Tabela 11. Acurácia obtida a partir do dataset *Pima Indians*.

Algoritmo	Acurácia
Regressão Logística	78,07
SVM	76,95
Random Forest	75,84
Naïve Bayes	74,72
AdaBoosting	73,98
KNN	69,89
Árvore de decisão	69,14

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 12.

Tabela 12.Matriz de Confusão dos algoritmos relacionados na tabela 8.

Regressão Logística		Classe predita				
		Positiva	Negativa	Total	%	%
Classe original	Positiva	152	22	174	87,4	12,6
	Negativa	37	58	95	38,9	61,1
Support Vector Machine (SVM)		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	159	15	174	91,4	8,6
	Negativa	47	48	95	49,5	50,5
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	149	25	174	85,6	14,4
	Negativa	37	58	95	38,9	61,1
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	140	34	174	80,5	19,5
	Negativa	34	61	95	35,8	64,2
da Boosting		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	143	31	174	82,2	17,8
	Negativa	39	56	95	41,1	58,9
KNN		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	139	35	174	79,9	20,1
	Negativa	46	49	95	48,4	51,6
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	133	41	174	76,4	23,6
	Negativa	42	53	95	44,2	55,8

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, erro, sensibilidade e eficiência, que estão abaixo descritas na tabela 13.

Tabela 13.TVP, TFP, Erro, Sensibilidade e Eficiência referente a cada matriz dos resultados da tabela 9.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
Regressão Logística	87,36	38,95	21,93	61,05	74,205
SVM	91,38	49,47	23,05	50,53	70,955
Random Forest	85,63	38,95	24,16	61,05	73,34
Naïve Bayes	80,46	35,79	25,28	64,21	72,335
AdaBoosting	82,18	41,05	26,02	58,95	70,565
KNN	79,89	48,42	30,11	51,58	65,735
Árvore de decisão	76,44	44,21	30,86	55,79	66,115

Fonte: autoria própria.

Após o balanceamento pelo *NearMiss*, o conjunto de dados possui 268 registros para cada classe e a tabela 14 apresenta a acurácia, a principal medida adotada, ordenada do maior para o menor valor resultante.

Tabela 14. Acurácia obtida a partir do dataset *Pima Indians* após balanceamento pelo NearMiss:

Algoritmo	Acurácia (%)
Regressão Logística	73,45
Naïve Bayes	71,19
Random Forest	70,06
SVM	67,8
AdaBoosting	67,8
Árvore de Decisão	64,97
KNN	64,41

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 15.

Tabela 15. Matriz de Confusão dos algoritmos relacionados na tabela 11.

Regressão Logística		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	65	22	87	74,7	25,3
	Negativa	25	65	90	27,8	72,2
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	71	16	87	81,6	18,4
	Negativa	35	55	90	38.9	61.1
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	63	24	87	72,4	27,6
	Negativa	29	61	90	32,2	67,8
SVM		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	69	18	87	79,3	20,7
	Negativa	39	51	90	43,3	56,7
AdaBoosting		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	64	23	87	73,6	26,4
	Negativa	34	56	90	37,8	62,2
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	57	30	87	65,5	34,5
	Negativa	32	58	90	35,6	64,4
KNN		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	62	25	87	71,3	28,7
	Negativa	38	52	90	42,2	57,8

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, Sensibilidade, especificidade e erro, que estão abaixo descritas na tabela 16.

Tabela 16. TVP, TFP, Erro, Sensibilidade e Eficiência referente a cada matriz dos resultados da tabela 12.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
Regressão Logística	73,81	27,78	21,93	72,22	73,02
Naïve Bayes	81,61	38,89	23,05	61,11	71,36
Random Forest	72,41	32,22	24,16	67,78	70,1
SVM	79,31	43,33	25,28	56,67	67,99
AdaBoosting	72,62	37,78	26,02	62,22	67,42
Árvore de Decisão	65,52	35,56	30,11	64,44	64,98
KNN	67,53	42,22	30,86	57,78	62,66

Fonte: autoria própria.

3.2 NHANES – NATIONAL HEALTH AND NUTRITION EXAMINATION

Continuando os experimentos, foram gerados alguns conjuntos a partir dos dados do NHANES. A estes subconjuntos, foram aplicadas as técnicas selecionadas para obtenção do desempenho. Na tabela 17, estão elencados os subconjuntos de dados contendo o número de registros e características como a quantidade de registros por classe e por sexo biológico.

Tabela 17.Listagem de subconjuntos de dados experimentais (NHANES).

Dataset	Nº	Var.	Sexo biológico		Classe	
			Masc.	Fem.	Negativa	Positiva
NHANES 2017-2018 Sem nulos	2490	39	1.120	1.270	2.233	257
NHANES 2017-2018 Sem nulos e balanceamento (NearMiss)	514	39	------	----	257	257
NHANES 2017-2018 Substituição de Nulos	9.920	39	4.904	5.016	9.210	710
NHANES 2017-2018 Substituição de Nulos e balanceamento (NearMiss)	1.420	39	719	701	710	710
NHANES 2017-2018 e NHANES 2015-2016 Substituição de Nulos	20.505	39	10.116	10.389	19.135	1.370
NHANES 2017-2018 e NHANES 2015-2016 Substituição de Nulos e balanceamento (NearMiss)	2.740	39	1.390	1.350	1.370	1.370
JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS Balanceamento Natural	21.105	37	10.404	10.701	19.135	1.970
JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS Balanceamento (NearMiss)	3.940	37	1.990	1.941	1970	1970
JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS REDUÇÃO (Redução de Variáveis)	21105	22	10.404	10.701	19,135	1.970
JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS REDUÇÃO Balanceamento (NearMiss)	3.940	22	2.025	1.915	1.970	1.970

Nº: número de registros; Var.: quantidade de variáveis; Masc.: sexo biológico masculino; Fem.: sexo biológico feminino.

Fonte: autoria própria.

Com tais subconjuntos, foram executados experimentos e seus resultados estão delineados na tabela 18, salientando que foram tomados os dois melhores resultados obtidos segundo a acurácia, onde: B.: ocorrência de balanceamento

Tabela 18. Resultados dos experimentos com os subconjuntos da tabela 16.

Dataset	B	Algoritmo	Acurácia %	Matriz de confusão Média	Positivo	Negativo	Eficiência
NHANES 2017-2018 Sem nulos	N	Regressão Logística	90,02	55,4	98,6	12,2	55,43
		SVM	90,02	50	100	0	50
NHANES 2017-2018 Sem nulos e balanceamento (NearMiss)	S	KNN	59,41	60,3	80,2	40,4	60,35
		Regressão Logística	58,82	59,4	71,6	47,2	59,4
NHANES 2017-2018 Substituição de Nulos	N	SVM	92,33	50	100	0	50
		KNN	92,15	53,9	99,8	8	49,9
NHANES 2017-2018 Substituição de Nulos e balanceamento (NearMiss)	S	Regressão Logística	89,55	87,75	85,4	94,1	89,79
		Naïve Bayes	88,7	88,8	86,2	91,4	88,84
NHANES 2017-2018 e NHANES 2015-2016 Substituição de Nulos	N	SVM	93,11	50	100	0	50
		KNN	93,92	50,1	99,8	0,4	50,1
NHANES 2017-2018 e NHANES 2015-2016 Substituição de Nulos e balanceamento (NearMiss)	S	Regressão Logística	89,06	89,05	85	93,1	89,08
		Naïve Bayes	88,51	88,55	84,6	92,5	88,52
JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS (Balanceamento Natural)	N	KNN	92,95	60,9	93	28,8	64,28
		AdaBoosting	92,53	71,65	97,5	45,8	71,64
JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS Balanceamento (NearMiss)	S	AdaBoosting	92,95	84	79	89	84,03
		Regressão Logística	92,53	83,5	78,4	88,6	83,5
JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS REDUÇÃO (Redução de variáveis)	N	AdaBoosting	92,78	71,15	97,9	44,4	71,17
		KNN	92,33	65,1	98,8	31,4	65,08
JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS REDUÇÃO Balanceamento (NearMiss)	S	AdaBoosting	76,33	76,3	78,6	74	76,28
		Naïve Bayes	72,79	72	90	54	72,45
NHANES 2017-2018 e NHANES 2015-2016 PLUS Preenchimento mediana	N	AdaBoosting	92,74	71,9	97,7	46,1	71,89
		KNN	92,33	65,4	98,7	32,1	65,42

Fonte: autoria própria.

3.3 TRABALHANDO COM DADOS AUSENTES

Há métodos para trabalhar com dados ausente. Carvalho (2011) aponta que podemos eliminar os objetos com dados faltantes ou nulos, definir e preencher

de forma manual, utilizar um método ou heurística assim como utilizar algoritmos de aprendizado de máquina que possam realizar esta tarefa.

O conjunto de dados NHANES contém dados ausentes. Os testes utilizaram como alternativa a retirada de todos os dados nulos. Em alguns casos nesta pesquisa, os dados foram eliminados. Também, trocados por um valor neutro (-1) ou preenchidos com a moda, mediana ou média.

A linguagem Python possibilita verificar a quantidade de dados nulos ou não existentes em cada variável. Sendo “diabetes” o nome da variável para manipulação dos dados a verificação se faz pelo comando diabetes.isnull().sum().

Outra forma de verificar é com o comando diabetes.isnull() que resulta uma matriz em que os valores são dados boleanos. Dada a tabela dos dados, como uma matriz, se o dado for faltante, então é nulo, apresentará valor *True*. Do contrário, apresentará valor False, conforme quadro 10.

Quadro 28. Matriz de verificação de dados ausentes.

	SEQN5	RIAGENDR	RIDAGEYR	LBDSALSI
0	False	False	False	False
1	False	False	False	False
2	False	False	False	False
3	False	False	False	False
4	False	False	False	False
5	False	False	False	False
6	False	False	False	False

Fonte: autoria própria.

Uma forma de trabalhar com dados faltante é a substituição pelo valor de média ou a moda. Carvalho (2011) adverte que, em alguns casos, “a imputação de valores ausentes pela média pode levar a inconsistências”.

Algumas informações da documentação da linguagem Python[1] podem nos ajudar neste processo decisório. Por exemplo: a moda (função *mode* ()), ou valor mais comum, é destinado a trabalhar com valores discretos ou nominais. Diferente de parte dos dados que estão alocados no *dataset*.

[1] Statistics — Funções estatísticas. Disponível em: encurtador.com.br/jkvw2. Acesso em 6 jun. 2021.

> A média é fortemente afetada por valores discrepantes e pode não ser um estimador robusto para localização central: a média não é necessariamente uma representação típica dos dados. Para medidas mais robustas de centralidade, veja *median()* e *mode()*.

Eis um exemplo de como realizar o preenchimento dos dados através do comando:

diabetes['LBXSAL'].fillna(diabetes['LBXSAL'].median(), inplace = True).

3.4 PREPARAÇÃO PARA GERAÇÃO DO MODELO PREDITIVO

Como outrora citado, utilizamos os dados de pesquisas do NHANES, uma divisão de pesquisas de exame de saúde e nutrição que é parte do Centro de Controle e Prevenção de Doenças (CDC).

Abaixo, na tabela 19 estão consignados os períodos e a quantidade de registros obtidos, salientando as quais classes os dados são pertencentes.

Tabela 19. Total de dados dos conjuntos NHANES conforme a distribuição de classes.

Período	Total de Registros	Classe 0	Classe 1
2017-2018	9.920	9.210	710
2015-2016	10.585	9.925	660
2013-2014	600		600
	21.105	19.135	1.970

Fonte: autoria própria.

De início, foram utilizados somente os dados do período 2017-2018. Para obter mais dados, foram analisados e utilizados também o período de 2015-2016. No que se refere ao período 2013-2014, este se refere ao balanceamento natural, por isso, foram tomados apenas os referentes à classe 1 para que houvesse um maior equilíbrio entre as classes. O conjunto totalizou 21.105 registros, com 39 atributos e uma classe alvo.

Foram realizados procedimentos para:

Eliminação manual de atributos: SEQN. O atributo SEQN é um número sequencial e é um atributo muito importante para a coerência, afinal, ele teve importante função para trabalhar com os dados sendo utilizado como chave identificador. No entanto, ele se revela irrelevante para a atividade preditiva por não fornecer contribuição para tanto.

Dados redundantes (objetos): os objetos redundantes participam mais de uma vez no processo e, por isso, podem formar uma tendência nos resultados. É preciso a identificação e eliminação destas ocorrências, o que foi feito.

Dados redundantes (atributos): Um atributo é considerado redundante se seu valor puder ser estimado a partir de pelo menos um dos seus atributos. Isso ocorre quando dois ou mais atributos têm a mesma informação preditiva.

BMXBMI (Índice de Massa Corporal). O IMC é obtido pela seguinte fórmula:

$$IMC = \frac{peso\ (kg)}{altura^2} \tag{7}$$

De tal maneira, peso (kg) e altura (cm), respectivamente BMXWT e BMXHT, são os itens de composição da fórmula. Portanto, o atributo BMXBMI deve ser retirado para não supervalorizar o atributo e criar destaque na predição.

a) URXUMA e URXUMS: ao verificar a tabela do conjunto de dados, as duas variáveis se referem ao mesmo item (Albumina) e as duas colunas têm os mesmos valores. Portanto, a opção foi retirar uma das colunas.
b) URDACT (Razão de creatinina albumina) é também um dado obtido após cálculos. Como o objetivo não é privilegiar um ou outro atributo, é importante proceder a exclusão dele.

Por fim, no tocante ao colesterol total, este é a soma dos níveis de colesterol presentes no sangue como HDL, LDL. No entanto, neste caso, eles permanecerão como já retirados. Nesta pesquisa, a medição de LDL é calculada a partir de valores de colesterol total (LBXTC), triglicerídeos (LBXTR) e HDL-C (LBDHDD) utilizando a equação de Friedewald; a equação de Martin-Hopkins; e a Equação 2 de NIH.

Com o mesmo fim já enunciado, foram mantidas as coletas referentes à colesterol total (LBXTC), triglicerídeos (LBXTR) e HDL-C (LBDHDD).

3.5 SELEÇÃO DE VARIÁVEIS

Para a seleção de variáveis, foi tomado o *dataset* completo, com 21105 registros em 36 variáveis de entrada e uma variável alvo, com classe majoritária pertencente à não diabéticos tipo 2 (0: 19135) e com valores próximos de pessoas do sexo feminino (10701) e masculino.

Ressalte-se que, os dados faltantes foram preenchidos com o valor (-1) para viabilizar a execução do algoritmo de seleção. Dentre as técnicas, a *Recursive Feature Elimination* (RFE), que recursivamente remove os atributos e constrói o modelo com os atributos remanescentes, utilizando a acurácia para identificação dos atributos que mais contribuem para a predição.

Nesse caso em tela, as cinco que mais contribuem são, na ordem de apresentação:

RIDAGEYR (Sexo biológico), LBDSGLSI (Glicose), LBXSGL (Glicose), LBXGH (Hemoglobina Glicada), BMXWT (Peso).

Também precisam ser comentados os métodos ensemble para seleção, com a utilização do *Bagged Decision Trees*, estimam a importância de cada atributo. A lógica é: o método retorna um escore de cada atributo e quanto maior o escore, maios a importância que ele tem. Segue os 20 primeiros atributos, na tabela 20, denotando a importância do atributo LBXGH (Hemoglobina Glicada) presente em ambos os processamentos

Tabela 20. Resultados da aplicação de seleção de variáveis utilizando métodos Ensemble.

VARIÁVEL	VALOR
LBXGH	0.171969
LBDSGLSI	0.091433
LBXSGL	0.082977
RIDAGEYR	0.063715
LBDGLUSI	0.037660
URXUMA	0.033767
BMXWAIST	0.030242
BMXWT	0.028564
BMXHT	0.027723
BMXARMC	0.025283
BPXSY1	0.024264
URXCRS	0.024130
LBXGLU	0.024074
LBXSTR	0.023582
URXUCR	0.023458
LBDSTRSI	0.022942
LBXSCH	0.021349
LBXSCR	0.021005
LBDSCRSI	0.020526
LBDSCHSI	0.020481

Fonte: autoria própria.

3.6 A GERAÇÃO DO MODELO PREDITIVO

Considerando o conjunto de dados do programa NHANES, foram selecionados atributos considerando e convergindo:

a) Conhecimentos colacionados nesta pesquisa, como a análise de risco da ADA (Apêndice A);
b) Seleção de atributos obtidos com as técnicas de *Recursive Feature Elimination* RFE e métodos Ensemble;
c) Anamneses concernentes às experiências pessoais da autora para avaliação de diabetes mellitus.

Do todo, restaram dez variáveis às quais são 9 variáveis de entrada e uma classe alvo que estão explicitadas a seguir, com o nome de cada atributo e sua descrição no quadro 29.

Quadro 29. Consolidação dos atributos para geração do modelo preditivo.

Variável	Conteúdo
RIAGENDR	Sexo biológico
RIDAGEYR	Idade
LBDSGLSI	Glicose, soro refrigerado (mmol / L)
LBDSTRSI	Triglicerídeos, soro refrigerado (mmol / L)
LBXGH	Hemoglobina Glicada
LBDINSI	Insulina (pmol / L)
BMXWT	Peso (kg)
BMXHT	Altura em pé (cm)
BMXWAIST	Circunferência da cintura (cm)
CLASSE	Alvo

Fonte: autoria própria.

Salientamos que, para uniformização do *dataset*, foram utilizadas as variáveis das unidades de concentração do sistema internacional, como a milimol por litro (mmol/L). Do conjunto inicial, após limpeza e balanceamento, resultaram 3.940 registros, que compunham 1970 registros para cada classe. Desse conjunto, foram retirados dez registros para testes de validação do modelo.

De tal forma, foi gerado o modelo preditivo utilizando o classificador *AdaBoosting*. Além de ter sido apontado como o melhor resultado em testes realizados, este classificador possui vantagens como corrigir de forma iterativa os erros dos classificadores fracos e melhora a precisão combinando estimadores fracos. Também, não é propenso à ocorrência de *overfitting*, ou quando um modelo apresenta 100% dos valores reais (observados) encaixados na sua função.

Para a avalição de desempenho formulou-se a adoção da acurácia e da matriz de confusão ou contingência como métricas principais. No decurso, também foram estudadas outras métricas que foram elencadas de maneira acessória, mas úteis ao processo de conhecimento.

Em se tratando de um protótipo que a predição é o diagnóstico de uma patologia, a acurácia mínima foi fixada em 90%. No entanto, no início dos trabalhos, o conjunto elegível era o *dataset* Índios Pima. Com 768 registros, a atividade preditiva restaria prejudicada ao não ter atingido o corte de 1.000 registros.

Importante salientar os resultados da avaliação após da geração do modelo preditivo, com acurácia de 94,15%; a tabela 21 apresenta a matriz de confusão do modelo gerado.

Tabela 21. Matriz de confusão do modelo preditivo gerado.

		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	559	36	595	**93,9**	6,1
	Negativa	33	551	584	5,7	**94,3**

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 22.

Tabela 22. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 21.

TVP	TFP	ERRO	SENSIBILIDADE	EFICIÊNCIA
93,95	5,65	5,937	94,35	94,15

Fonte: autoria própria.

Após a submissão à ajuste de hiper parâmetros, apresentou acurácia de 94.063% além dos resultados, delineados na tabela 23, que se refere à matriz de confusão gerada.

Tabela 23. Matriz de confusão do modelo preditivo gerado após hiper parâmetros.

		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	559	36	595	**93,9**	6,1
	Negativa	34	550	584	5,8	**94,2**

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 24.

Tabela 24. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 23.

TVP	TFP	ERRO	SENSIBILIDADE	EFICIÊNCIA
93,95	5,82	5,937	94,18	94,07

Fonte: autoria própria.

Com o *dataset* NHANES, a acurácia mínima foi revista e determinada em acima de 90%. De fato, pode ser observado que o modelo gerado alcançou 94,063% o que pode ser observada a matriz de contingência.

4 PROTÓTIPO: ARQUITETURA, MODELAGEM E IMPLEMENTAÇÃO

De todo lido e analisado, neste capítulo pretende-se apresentar a arquitetura, modelagem e a implementação da Aplicação Analítica nomeada Flocos.

4.1 ARQUITETURA FLOCOS

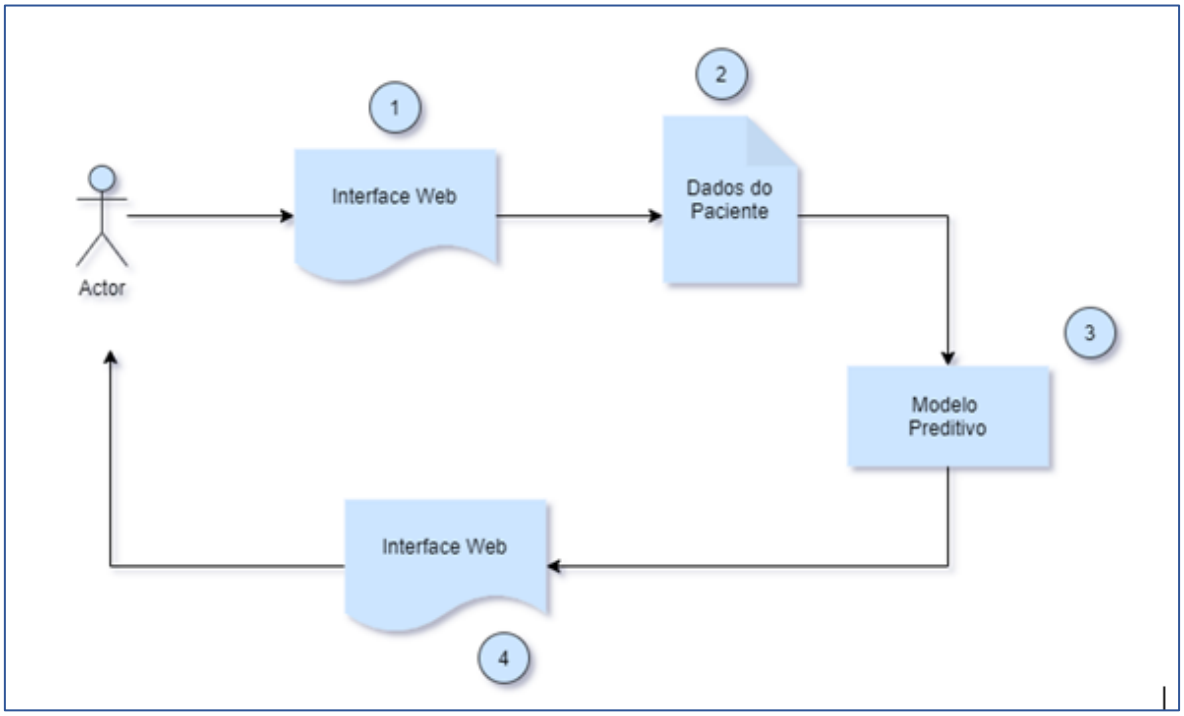
Ante todo o exposto, a construção de um protótipo pode tornar palpável a abordagem desta pesquisa. De forma sintética, utiliza-se um protótipo para demonstrar os conceitos e experimentar opções do projeto; é uma versão inicial de um sistema de software.

Por isso, foi criado Flocos, o protótipo de predição de diabetes mellitus tipo 2. Para conhecimento, Flocos é uma referência ao animal de estimação da autora, um coelho albino Fuzzy Loop e que foi a inspiração para nomear o protótipo.

A arquitetura da solução, conforme a figura 10, é composta pelos elementos: interface web para entrada de dados, modelo preditivo *AdaBoosting* e interface web para apresentação dos resultados do processamento.

O usuário acessa a interface web (1) e preenche os campos solicitados com os dados do paciente. Estes dados (2) são enviados ao modelo preditivo (3) que faz o processamento das informações e a apresenta o resultado na interface web (4) que será acessada pelo usuário.

Figura 10. Fluxograma da arquitetura Flocos.



Fonte: autoria própria.

Para o desenvolvimento do protótipo, além da linguagem de programação *Python* e suas bibliotecas (*NumPy*, *Pandas*, *Seaborn*, *Sickit-Learn*), foi utilizada a *Interface Development Environment* (IDE) *Jupyter Notebook*, para elaboração do código fonte e testes.

A aplicação utiliza Flask e Jinja. Para a criação da página Web, HTML 5, CSS 3 para sua estilização e *JavaScript*, na tarefa de validação de dados.

4.2 ESPECIFICAÇÃO DE REQUISITOS

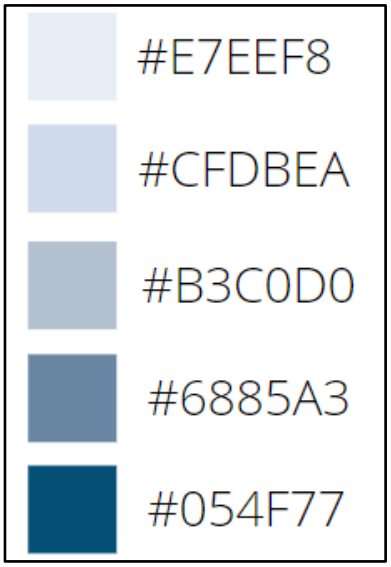
4.2.1 Requisitos não funcionais

O protótipo tem como requisitos, a usabilidade, concernente à facilidade na utilização da interface gráfica a partir do conceito de design limpo e prático, com design responsivo e adaptável.

Para atingir a acessibilidade, utiliza-se de paleta de cores que obtenha nível de contraste adequado segundo a ferramenta “*contrast-ratio*” (https://contrast-ratio.com/) para maior acessibilidade à pessoa com menor acuidade visual.

No sentido de ser figurativo, a barra de progresso utilizada para a probabilidade de diagnóstico obedecerá à paleta de cores, conforme a figura 11, a fim de indicar cenário.

Figura 11. Paleta de cores da barra de progresso de probabilidade do diagnóstico.



Fonte: autoria própria.

Em conformidade a Lei Geral de Proteção de Dados, o protótipo não fará o armazenamento de dados e as informações solicitadas são com a finalidade de obter o resultado para predição de acordo com a geração do modelo preditivo, o que deve ser realizado com a anuência do paciente.

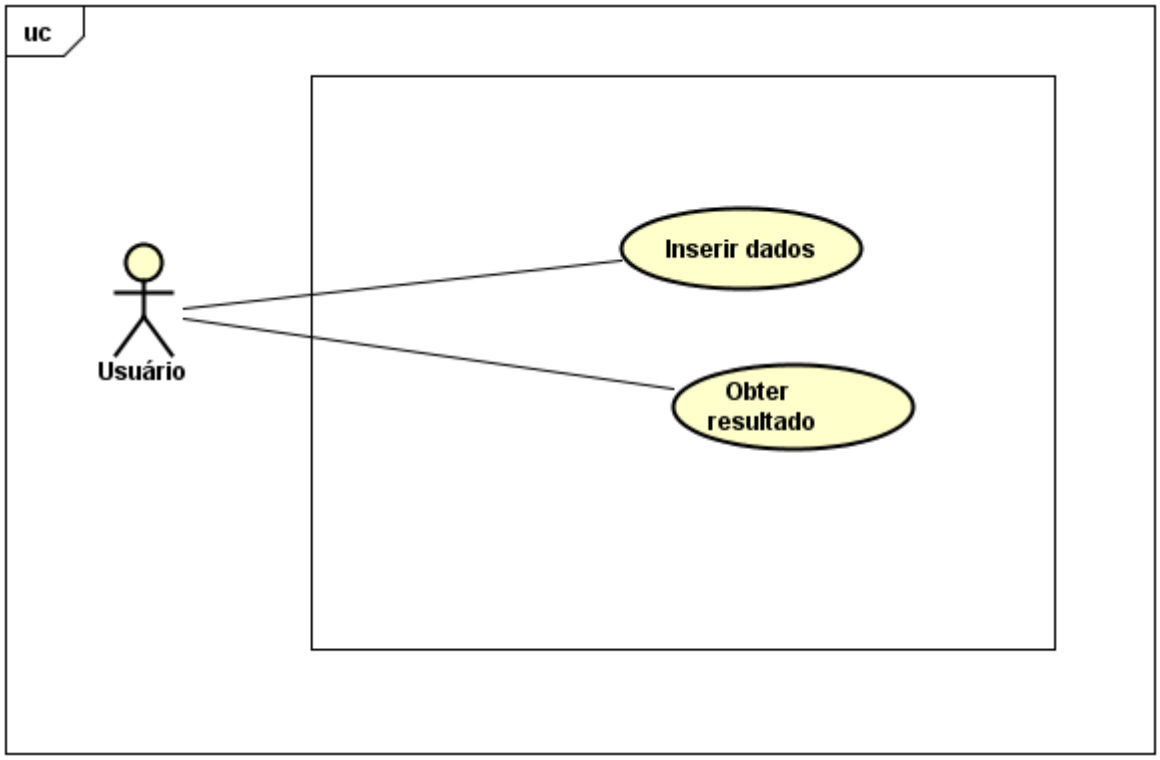
4.2.2 Requisitos funcionais

- O usuário deve inserir os dados de entrada do paciente, conforme as especificações, utilizando apenas ponto (.), excluindo vírgula (,) e caracteres especiais (@, $, entre outros).
- O sistema não deve aceitar campos com valor vazio.
- O sistema deve processar os dados de entrada após o usuário clicar no botão “Nome do botão”, utilizando o modelo preditivo descrito.
- O sistema deve apresentar o resultado, quais sejam: os dados informados do paciente e o valor percentual de probabilidade de o paciente ser portador de diabetes mellitus tipo 2.
- O sistema deve voltar a tela inicial caso o usuário clique no botão “VOLTAR”.

4.3 MODELAGEM DA APLICAÇÃO

Para explicitar a modelagem da aplicação analítica, podemos iniciar com as funcionalidades. Na figura 12, é ilustrado o diagrama de caso de uso, considerando a abordagem UML – Unified Modeling Language.

Figura 12. Diagrama de caso de uso.



Fonte: autoria própria.

A descrição textual do caso de uso é apresentada a seguir:

1 – Inserir dados e obter resultado

Ator: Usuário do sistema. Cabe salientar que, neste caso, podemos abranger a persona entrevistador ou mesmo o paciente.

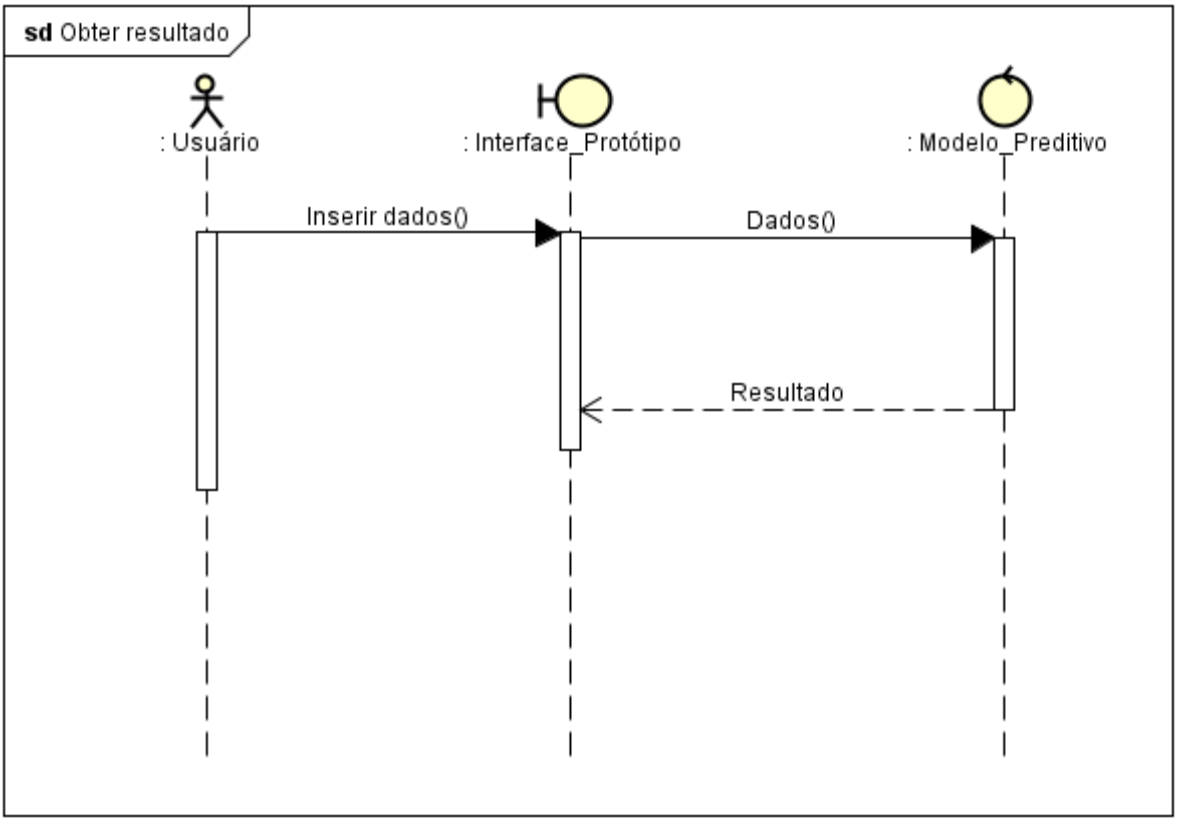
Precondições: Possuir dados médicos solicitados

Cenário Principal de Sucesso:

- Inserir dados e obter resultado
- O caso de uso inicia quando o usuário acessa à interface do sistema. Na tela inicial, o sistema apresenta um título e os campos de dados solicitados, os quais deverão ser preenchidos sem a utilização de vírgula (,) ou caracteres especiais (tais como: @, #, $).
- O usuário deve clicar no botão “nome do botão”.
- O sistema processa os dados de entrada.
- O sistema exibe na tela as informações os dados informados e o resultado do processamento, qual seja, a predição acerca da diabetes mellitus tipo 2.
- O caso de uso é finalizado.

A sequência da funcionalidade, ilustrada com a figura 13, pode ser assim descrita no diagrama de sequência da função: o entrevistador insere os dados do paciente. A aplicação analítica é a interface do protótipo, analogamente a um formulário de dados. Os dados serão processados e o resultado será apresentado na interface da aplicação.

Figura 13. Diagrama de sequência da função predizer diagnóstico.



Como premissa, a aplicação, por exemplo, não persistirá dados e, por isso, não será necessário a utilização de um SGBD para o armazenamento e que apoie a aplicação.

4.4 AVALIAÇÃO DE DESEMPENHO

No sentido de ilustrar os dados selecionados, segue o quadro 30, que apresenta os códigos referentes à cada atributo utilizado no *dataset* de geração do modelo preditivo.

Quadro 30. Atributos referenciados código da tabela de teste.

Código	Variável
A1	RIAGENDR
A2	RIDAGEYR
A3	LBDSGLSI
A4	LBDSTRSI
A5	LBXGH
A6	LBDINSI
A7	BMXWT
A8	BMXHT
A9	BMXWAIST
A10	CLASSE

Fonte: autoria própria.

A princípio, do total de registros do *dataset* eleito para a geração do modelo preditivo, foram separados dez registros para realizar os testes de validação. Com os referidos registros e o quadro 30 foi confeccionada a tabela 25, que contém os registros selecionados acrescidos de dois outros.

Tabela 25. Dados segregados para testes de validação do modelo.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
1	2	60	4.77	1.276	5.9	82.56	120.2	166.7	124.1	0
2	2	56	4.88	0.824	5.2	28.98	48.8	157.5	71.4	0
3	1	36	4.61	3.026	5.1	40.08	57.9	179.1	76	0
4	2	13	5.44	1.321	5.5	32.94	69.9	163.4	94.2	0
5	2	77	5	1.05	5.5	80.94	84	158	101.1	0
6	2	48	17.15	5.408	11.5	124.86	95.4	161.7	125.7	1
7	1	63	9.66	1.355	6.9	78.06	78.4	174.2	100	1
8	2	68	5.16	1.694	6.2	130.32	82.6	167.2	116	1
9	2	63	7.49	1.276	8.5	184.02	68.1	146.1	100.1	1
10	1	60	8.72	1.694	8	41.64	84.2	171.4	107	1
11	1	38	5.22	1.560	5.8	0	92	167	0	0
12	2	34	0	1.9	6.0	0	100	165	0	1

Fonte: autoria própria.

Observe-se que, considera-se 1, para o sexo masculino, e 2 para o feminino, quanto ao sexo biológico.

A partir de então, com os dados de entrada acima determinados, foram anotados os resultados para cada predição. Verifica-se na tabela 26, para a sequência de dados, os resultados esperados, os resultados obtidos e o percentual do resultado obtido.

Tabela 26. Resultados esperados e obtidos nos testes de validação do modelo preditivo.

	Resultado esperado		Resultado obtido	Percentual
1	0	Não diabético	Não diabético	49.27
2	0	Não diabético	Não diabético	45.55
3	0	Não diabético	Não diabético	44.14
4	0	Não diabético	Não diabético	38.49
5	0	Não diabético	Não diabético	49.57
6	1	Diabético	Diabético	51.23
7	1	Diabético	Diabético	51.84
8	1	Diabético	Diabético	50.44
9	1	Diabético	Diabético	52.93
10	1	Diabético	Diabético	52.68
11	0	Não Diabético	Diabético	76.31
12	1	Diabético	Diabético	86.29

Fonte: autoria própria.

Com os resultados, foi possível formular a matriz de contingência como consolidação das ocorrências da relação entre a classe predita e a classe original, na tabela 27.

Tabela 27. Matriz de contingência do modelo preditivo (fase de testes de validação).

		Classe predita		
		Positiva	Negativa	Total
Classe original	Positiva	6	0	6
	Negativa	1	5	6

Fonte: autoria própria.

Em acréscimo, segue o cálculo da acurácia (eq. (7) – item 2.3.5 Métricas de desempenho) do modelo preditivo gerado que se comporta dentro dos parâmetros estabelecidos.

$$ACC = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{FP} + \mathrm{VN} + \mathrm{FN}} \tag{8}$$

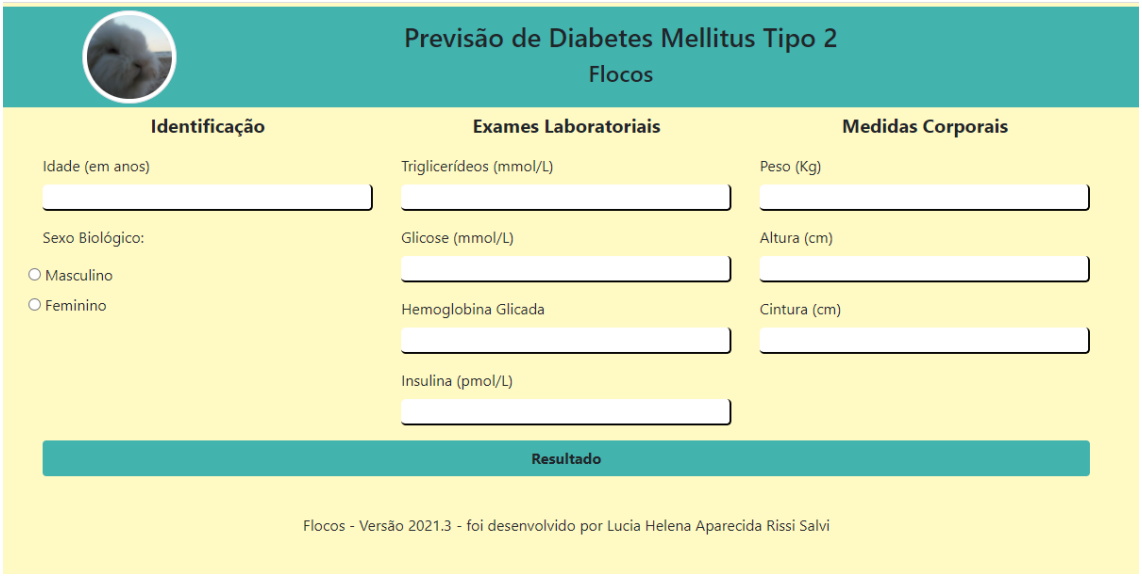
$$ACC = \frac{6 + 5}{6 + 1 + 5 + 0} = \frac{11}{12} = 0{,}9166666666666667 * 100 = 91{,}67\% \tag{9}$$

Conclui-se, portanto, que os resultados da aplicação analítica são aderentes aos resultados do modelo preditivo.

4.5 APLICAÇÃO

Com o intuito de empregar os conhecimentos dispostos nesta pesquisa, sucedeu a construção de uma aplicação. A figura 14 ilustra a interface inicial da aplicação correspondente à utilização do browser Google Chrome.

Figura 14. Interface gráfica inicial da aplicação.



Fonte: autoria própria.

Nos termos dos requisitos levantados, a tela tem característica responsiva para torná-la mais acessível, o que se pode verificar pela figura 15, que apresenta o teste para smartphone Moto G4 na disposição vertical. Do mesmo modo, a figura 25 apresenta o teste para smartphone Moto G4 na disposição horizontal.

Fonte: autoria própria.

Figura 15. Teste de tela responsiva: Smartphone Moto G4 na posição vertical.

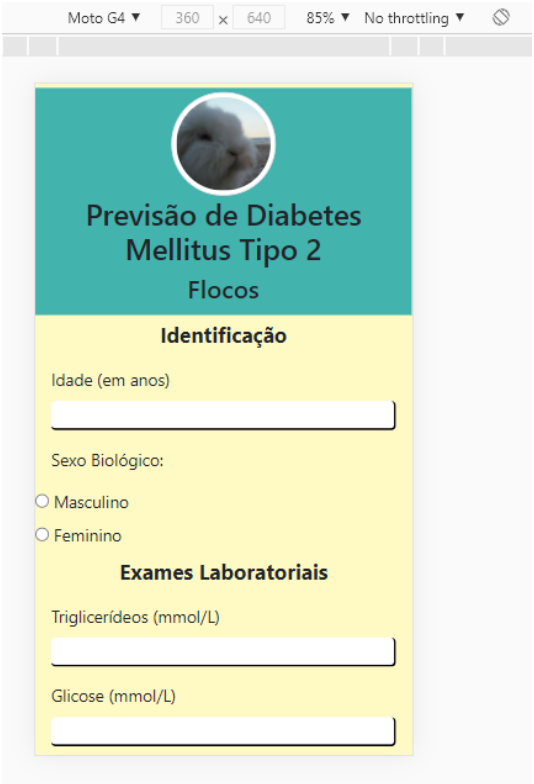


Figura 16. Teste de tela responsiva: Smartphone Moto G4 na posição Horizontal.

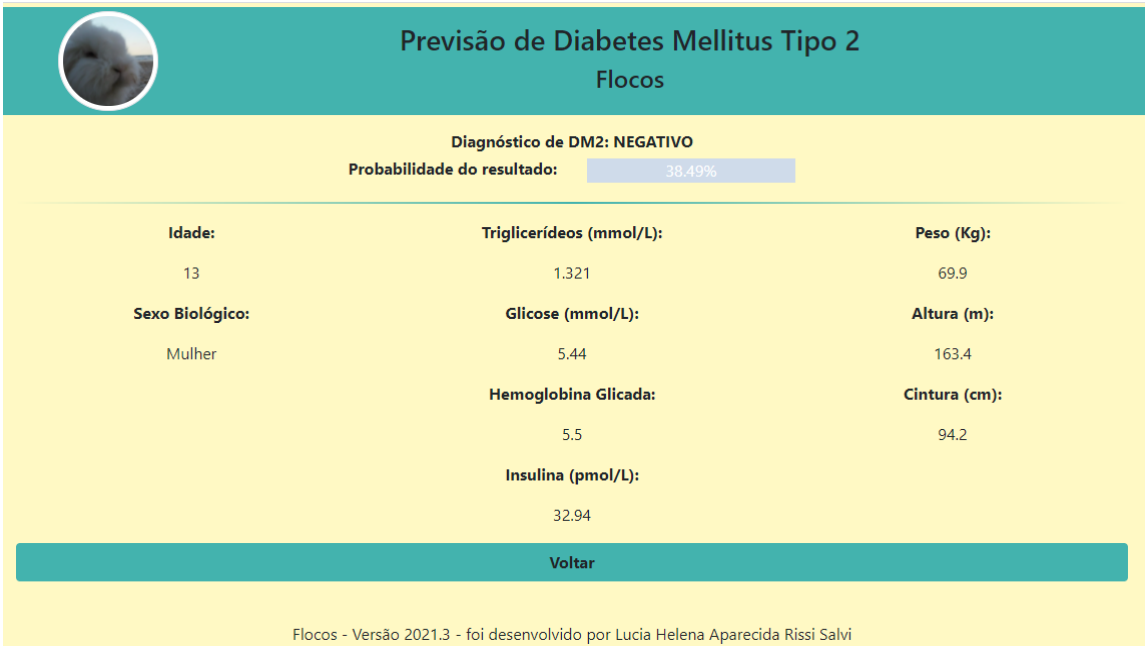
Fonte: autoria própria.

Em termos de teste de aplicação, realizou-se teste com um dos registros disponíveis para demonstrar as telas do sistema. Na figura 26, denotamos a tela inicial da interface com os dados de teste e, na figura 27, a interface que apresenta o resultado da predição.

Figura 17. Interface gráfica de inserção de dados da aplicação analítica.

Fonte: autoria própria.

Figura 18. Interface gráfica de resultado da aplicação analítica.



Fonte: autoria própria.

5 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

5.1 SÍNTESE DOS RESULTADOS

Esta pesquisa se fundou na investigação e avaliação de técnicas de classificação do aprendizado supervisionado e conjuntos de dados públicos que possibilitaram gerar o modelo preditivo que realiza o diagnóstico de diabetes mellitus tipo 2.

Por isso, dadas as circunstâncias manifestas em relação às motivações e justificativas, foram levantadas, em sede de revisão de literatura, as técnicas de aprendizado de máquina supervisionado e os conceitos correlatos, assim como as métrícas de avaliação, técnicas para preparação de dados e tecnologias aplicáveis à construção do protótipo.

Foi necessário angariar e estudar conjuntos de dados que tivessem aptidão para a tarefa. Eis que o número de dados e como eles estão dispostos, se rotulados ou não, influenciaram sobremaneira na seleção dos algoritmos assim como contribuiram para alcançar avaliação satisfatória de desempenho.

Durante este percurso, foram realizados estudos, testes e experimentos com os dados disponíveis para a tomada de decisão da técnica e do conjunto de dados eleitos para a geração do modelo preditivo.

Nesta seara, foram selecionados algoritmos KNN (*K-Nearest Neighbors*), Árvore de decisão, Regressão logística, *Naïve Bayes*, SVM (*Support Vector Machine*), além de métodos ensemble – AdaBoosting e Random Forest para testes com destaque para os AdaBoosting, Random Forest e Regressão Logística que forneceram bons resultados dos testes.

Pela sua inequívoca abrangência, esta pesquisa perpassou por conhecimentos da tecnologia da informação e da ciência médica, entrelaçando-os para caminhar em direção à construção da aplicação analítica de predição de diagnóstico.

Para tanto, foi escolhida a técnica *AdaBoosting* em que o modelo preditivo que obteve acurácia de 94,063%. A aplicação analítica permite prever se o paciente está com diabetes mellitus tipo 2 ou não, e sua probabilidade, a partir

dos atributos escolhidos e os valores informados na entrada com a acurácia de 91,67%.

5.2 TRABALHOS FUTUROS

Considerando trabalhos futuros, pode se vislumbrar que:

- O modelo preditivo seja gerado utilizando uma base de dados privada, por exemplo, de um hospital ou plano de saúde, concedida para este fim;
- A entrada de dados da aplicação seja por meio de arquivo para seu processamento, com a gravação dos resultados no banco de dados de origem,
- Viabilizar a utilização de *Business Intelligence* (BI) que é uma ferramenta importante para obtenção de *insights* que pode tornar a aplicação um produto competitivo.
- A utilização da aplicação desenvolvida na base de dados de determinada população, de forma periódica, para estabelecimento e monitoramento de políticas públicas.
- Adicionar outras técnicas ao arcabouço selecionado para experimentos prévios à geração de modelo preditivo.

REFERÊNCIAS

ADA, American Diabetes **Association. Risk Test**. 2020. Disponível em: encurtador.com.br/ptxSY. Acesso em 22 nov. 2020.

AMARAL, Fernando. **Introdução a Ciência de Dados**: mineração de dados e *Big data*. – Rio de Janeiro: Alta Books, 2016.

AUSTIN, Data. **The official City of Austin open data portal**. 2019. Disponível em encurtador.com.br/jBDM1. Acesso em 6 jan. 2021.

AZURE, Microsoft. Machine Learning Samples: Diabetes. Disponível em encurtador.com.br/gtAG9. Acesso em 27 out. 2020.

BANDEIRA, Francisco; MANCINI, Marcio; GRAF, Hans. **Endocrinologia e Diabetes**. Rio de Janeiro: MedBook Editora, 2015. Disponível em: encurtador.com.br/XZ379. Acesso em 16 nov. 2020.

BARI, Anasse; CHAOUCHI, Mohamed; JUNG, Tommy. **Análise preditiva para leigos**. Rio de Janeiro: Alta Books, 2019.

BASSO, Maik; VIEIRA, João Paulo; PARREIRA, Fábio José; SILVEIRA, Sidnei Renato; SOUZA, Adriana Sadowski de. **Sistema Inteligente para Apoio ao Diagnóstico de Diabetes Empregando Redes Neurais**. 2014. Disponível em: encurtador.com.br/KNST9. Acesso em 11 out. 2020.

BRASIL, Lei 13.709, de 14 de agosto de 2018. **Lei Geral de Proteção de Dados Pessoais** (LGPD)., DOU de 15.8.2018.

BRASIL. Constituição (1988). **Constituição da República Federativa do Brasil**. Brasília, DF: Senado, 1988.

BRASIL. Ministério da Economia. Secretaria Especial de Previdência e Trabalho. **Acompanhamento Mensal do Benefício Auxílio-Doença Acidentário Concedido Segundo os Códigos da CID-10 – Janeiro a Dezembro de 2019**. 2020. Disponível em encurtador.com.br/buD06 Acesso em 8 jan. 2020.

BRASIL. Ministério da Economia. Secretaria Especial de Previdência e Trabalho. **Acompanhamento Mensal do Benefício Auxílio-Doença Previdenciário Concedido Segundo os Códigos da CID-10 – Janeiro a Dezembro de 2019**. 2020. Disponível em encurtador.com.br/npC15. Acesso em 8 jan 2020.

BRASIL. Ministério da Economia. Secretaria Especial de Previdência e Trabalho **Acompanhamento Mensal dos Benefícios Auxílios-Doença Concedidos segundo Códigos da Classificação Internacional de Doenças** – 10ª Revisão. 2006.(CID-10) 2008. encurtador.com.br/kFL13. Acesso em 8 jan 2020.

BRASIL. Ministério da Saúde. Secretaria de Políticas de Saúde. Departamento de Ações Programáticas Estratégicas. **Plano de reorganização da atenção à hipertensão arterial e ao diabetes mellitus**: hipertensão arterial e diabetes mellitus / Departamento de Ações Programáticas Estratégicas. Brasília: Ministério da Saúde, 2001. Disponível em: encurtador.com.br/zNW49. Acesso em 26 out. 2020.

CARVALHO, André C. Ponce de Leon Ferreira de; FACELI, Katti; LORENA, Ana Carolina; GAMA. João; et al. **Inteligência Artificial** – uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

CARVALHO, Deborah Ribeiro; DALLAGASSA, Marcelo Rosano; DA SILVA, Sandra Honorato. Uso de Técnicas de Mineração de Dados para a Identificação Automática de Beneficiários Propensos ao Diabetes Mellitus Tipo 2. **Informação & Informação**,

[S.l.], v. 20, n. 3, p. 274 – 296, dez. 2015. ISSN 1981-8920. Disponível em: encurtador.com.br/jrLPY. Acesso em 11 out. 2020.

CASTRILLEJO, Sergio Alonso. **Aplicación de algoritmos de machine learning en la prediccion de la diabetes mellitus tipo II. 2020**. Disponível em: encurtador.com.br/bEHS4. Acesso em 11 out. 2020.

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes; Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações. – São Paulo: Saraiva, 2016.

CDC., Centers for Disease Control and Prevention. **NHANES 2013-2014, 2013**. Disponível em: encurtador.com.br/wDOSZ. Acesso em: 6 fev. 2021.

CDC., Centers for Disease Control and Prevention. **NHANES 2015-2016**, 2015. Disponível em: encurtador.com.br/enuO8. Acesso em: 6 fev. 2021.

CDC., Centers for Disease Control and Prevention. **NHANES 2017-2018**, 2017. Disponível em: encurtador.com.br/FXZ67. Acesso em: 6 fev. 2021.

CHIAVEGATTO FILHO, Alexandre Dias Porto. Uso de *big data* em saúde no Brasil: perspectivas para um futuro próximo. **Epidemiol**. Serv. Saúde, Brasília, v. 24, n. 2, p. 325-332, junho 2015. Disponível em encurtador.com.br/ilwW4. Acesso em 29 mar. 2020.

DELAGASSA, Marcelo Rosano. **Concepção de uma metodologia para identificação de beneficiários com indicativos de diabetes mellitus tipo 2**. Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2009. Disponível em: encurtador.com.br/fgvN8. Acesso em 22 nov. 2020.

Disponível em: encurtador.com.br/cyIL3. Acesso em 13 dez. 2020.

Disponível em: encurtador.com.br/mzR58. Acesso em 11 out. 2020.

ESCALERA, David Campos. **Predicción del impacto de regulaciones aéreas sobre los flujos de aeronaves utilizando técnicas de aprendizaje automático: um modelo a nível de vuelos**. Trabajo Fin de Máster (Máster Universitario em Inteligencia Artificial) – Universidad Politécnica de Madrid, Madrid, 2020. Disponível em: encurtador.com.br/dgBLV Acesso em: 5 fev 2021.

FERNANDES, Fernando Timoteo; CHIAVEGATTO FILHO, Alexandre Dias Porto. Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. **Rev. bras. saúde ocup**., São Paulo, v. 44, e13, 2019. . Disponível em: encurtador.com.br/dmqyC. Acesso em 29 mar. 2020. Epub Nov 4, 2019. encurtador.com.br/awBGY.

FRANCISCO, Papa. Novembro: Inteligência Artificial. **Rede Mundial de Oração do Papa**. 2020. Disponível em: encurtador.com.br/mNRXZ. Acesso em 15 nov. 2020.

GAYTÁN-HERNÁNDEZ, Darío; GUTIÉRREZ-ENRÍQUEZ, Sandra Olímpia; DÍAZ-OVIEDO, Aracely; GONZÁLEZ-ACEVEDO, Claudia Elena; MIRANDA-HERRERA, Magdalena; HERNÁNDEZ-IBARRA, Luis Eduardo. Escenario futuro de la diabetes mellitus tipo 2 estimado con un modelo de simulación dinámico predictivo. **Rev Panam Salud Publica**. 2017;41:e93. Disponível em: encurtador.com.br/kxBLP Acesso em 14 ago. 2020.

GROSS, Jorge L. et al. Diabetes Melito: Diagnóstico, Classificação e Avaliação do Controle Glicêmico. **Arq Bras Endocrinol Metab**, São Paulo, v. 46, n. 1, p. 16-26, Feb. 2002. Disponível em: encurtador.com.br/xKWZ2. Acesso em 16 nov. 2020.

GRUS, Joel. **Data Science do Zero** – Primeiras Regras com o Python. Ed. Alta Books. Rio de Janeiro. 2016.

HOLSBACH, Nicole; FOGLIATTO, Flávio Sanson; ANZANELLO, Michel Jose. Método de mineração de dados para identificação de câncer de mama baseado na seleção de variáveis. **Ciênc. saúde coletiva**, Rio de Janeiro, v. 19, n. 4, p. 1295-1304, Apr. 2014. Disponível em: encurtador.com.br/pEKPV. Acesso em 11 out. 2020.

IBGE, Instituto Brasileiro de Geografia e Estatística. IBGE divulga estimativa da população dos municípios para 2020. **Editoria Estatísticas Sociais**. 2020. Disponível em encurtador.com.br/QVZ14. Acesso em 21 dez 2020.

IDF, International Diabetes Federation. **Atlas da Diabetes da IDF**. Brussels: 2019 Disponível em: encurtador.com.br/jpyY5. Acesso em 23 nov. 2020.

ISLAM, MMF; FERDOUSI, Rahatara; RAHMAN, Sadikur; BUSHRA, Humayra Yasmin (2020) Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. Em: Gupta M., Konar D., Bhattacharyya S., Biswas S. (eds) Computer Vision and Machine Intelligence in Medical Image Analysis. **Advances in Intelligent Systems and Computing**, vol 992. Springer, Singapura. Disponível em encurtador.com.br/dktJ0. Acesso em 27 out. 2020.

JAKOBI, Heinz Roland; BARBOSA-BRANCO, Anadergh; BUENO, Luis Fernando; FERREIRA, Ricardo de Godoi Mattos; CAMARGO, Luís Marcelo Aranha. Incapacidade para o trabalho: análise dos benefícios auxílio-doença concedidos no estado de Rondônia. **Ciênc. saúde coletiva**. Rio de Janeiro, v. 18, n. 11, p. 3157-3168, nov. 2013. Available from encurtador.com.br/gsGTX. Acesso em 8 jan. 2021.

LAZCANO-ORTIZ, Margarita; SALAZAR-GONZALEZ, Bertha Cecilia. Adaptación en pacientes con diabetes Mellitus Tipo 2, según Modelo de Roy. **Aquichan**. Bogotá, v. 9, n. 3, p. 236-245, Dec. 2009. Disponível em: encurtador.com.br/jO035. Acesso em 11 out. 2020.

LOBO, Luiz Carlos. Inteligência Artificial e Medicina. **Rev. bras. educ. med.** Rio de Janeiro, v. 41, n. 2, p. 185-193, June 2017. Disponível em: encurtador.com.br/nvx58. Acesso em 11 out. 2020.

MAIONE, Camila. **Balanceamento de dados com base em oversampling em dados transformados**. Tese (Doutorado) – Universidade Federal de Goiás. Goiânia, 2020. Disponível em: encurtador.com.br/jtwW2. Acesso em 5 fev 2021.

MANEO, Adriano. Total de 1 milhão de mortos por coronavírus supera óbitos de guerras históricas. **Folha de São Paulo**, 28/09/2020. Disponível em: encurtador.com.br/fpzAD. Acesso em 7 nov. 2020.

MILECH, Adolpho. **Rotinas de Diagnóstico e Tratamento do Diabetes Mellitus**. AC Farmacêutica: Grupo GEN, 2014. 978-85-8114-270-8. Disponível em: encurtador.com.br/zABEJ. Acesso em 1 nov. 2020.

MITCHELL, Tom M. **Machine learning**. Nova York: McGraw-Hill, 1997.

MOURA, Alexandro Avila de. **Indicadores de desempenho** – 5 coisas que você precisa saber. 2017. Disponível em: encurtador.com.br/oELTY. Acesso em 20 dez 2020.

NOBLE, Douglas; MATHU, Rohini; DENT, Tom; MEADS, Catherine; GREENHALG, Trisha. **Risk models and scores for type 2 diabetes**: systematic review BMJ 2011; 343:d7163. Disponível em: encurtador.com.br/abcH6. Acesso em 14 ago. 2020.

OLIVERA, André Rodrigues et al. Comparação de algoritmos de aprendizado de máquina para construir um modelo preditivo para a detecção de diabetes não diagnosticada. **ELSA-Brasil: estudo de precisão**. São Paulo Med. J., São Paulo, v. 135, n. 3, p. 234-246, junho de 2017. Disponível em encurtador.com.br/bAMP8o. Acesso em 29 mar. 2020. encurtador.com.br/huwTZ.

PETRY, Marcio. **Big Data e registros eletrônicos de saúde**: um estudo exploratório sobre desafios e oportunidades para aplicações de saúde em hospital. 2015. Disponível em: encurtador.com.br/yBC59. Acesso em 11 abr. 2020.

PONTES, Ana Gabriela; REHME, Marta Francis Benevides; MICUSSI, Maria Thereza Albuquerque Barbosa Cabral; MARANHÃO, Técia Maria de Oliveira; PIMENTA, Walkyria de Paula; CARVALHO, Lídia Raquel de; PONTES, Anaglória. A importância do teste de tolerância à glicose oral no diagnóstico da intolerância à glicose e diabetes mellitus do tipo 2 em mulheres com síndrome dos ovários policísticos . **Revista Brasileira de Ginecologia e Obstetrícia**. Mar 2012, Volume 34 Nº 3 Páginas 128 – 132.

RIBEIRO, Áurea Celeste da Costa. **Diagnosis of diabetes type II by eficiente coding and vector machine support.** 2009. 52 f. Dissertação (Mestrado em Engenharia) – Universidade Federal do Maranhão, São Luis, 2009. Disponível em: encurtador.com.br/doyEN. Acesso em 11 out. 2020.

SANTOS, Hellen Geremias dos et al. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. **Cad. Saúde Pública**. Rio de Janeiro, v. 35, n. 7, e00050818, 2019. Disponível em: encurtador.com.br/nBILY. Acesso em 29 mar. 2020. Epub July 29, 2019. encurtador.com.br/ijHJV.

SANTOS, Hellen Geremias dos. **Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina**. 2018. Tese (Doutorado em Epidemiologia) – Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2018. doi:10.11606/T.6.2018.tde-09102018-132826. Acesso em 29 Mar. 2020.

SBD, Sociedade Brasileira de Diabetes. **Complicações do diabetes**. 2020. Disponível em: encurtador.com.br/kFUZ1. Acesso em 15 nov. 2020.

SBD, Sociedade Brasileira de Diabetes. Dados epidemiológicos do diabetes mellitus no Brasil. 2019. Disponível em: encurtador.com.br/eiyS8. Acesso em 22 nov. 2020.

SBMFC, SOCIEDADE BRASILEIRA DE MEDICINA DA FAMÍLIA E SOCIEDADE. **Modelos de previsão de risco para diabetes mellitus tipo 2**. Disponível em: encurtador.com.br/etwzL. Acesso em 5 jul. 2020.

SCHEFFEL, Rafael Selbach et al. Prevalência de complicações micro e macrovasculares e de seus fatores de risco em pacientes com diabetes melito do tipo 2 em atendimento ambulatorial. Rev. **Assoc. Med. Bras.** São Paulo, v. 50, n. 3, p. 263-267, set. 2004. Disponível em: encurtador.com.br/mrvOW. Acesso em 15 nov. 2020.

SIGILLITO, Vincent. Pima Indians Diabetes Database. National Institute of Diabetes and Digestive and Kidney Diseases. Centro de Pesquisa, Laboratório de Física Aplicada Líder do Grupo RMI, Universidade Johns Hopkins Johns Hopkins Road Laurel, **MD** 20707 (301) 953-6231 (c). Disponível em: encurtador.com.br/eiwB7. Acesso em 27 out. 2020.

SILVA, Bruno Riccelli dos Santos. **Uma análise comparativa de técnicas de subamostragem para projetos de sistemas de detecção de intrusão em redes de computadores**. Dissertação (mestrado) – Universidade Federal do Ceará, Fortaleza, 2020. Disponível em: encurtador.com.br/hxAY3. Acesso em: 5 fev 2021.

SOUSA, Maria Cristina Cordeiro. **Uma análise do algoritmo K-means como introdução ao aprendizado de máquinas**. 2019. 74 f. Monografia (Graduação) – Curso de Matemática, Universidade Federal do Tocantins, Araguaína, 2019.STAVIS, Robert L. **Recém-nascido grande para a idade gestacional (GIG)**. Manual MSD Versão para profissionais de Saúde. 2019. Disponível em: encurtador.com.br/ftLSY. Acesso em 3 jan. 2021.

TEED HJ, HUTCHISON S, ZOUNGAS S, MEYER C. Insulin resistance, the metabolic syndrome, diabetes, and cardiovascular disease risk in women with PCOS. **Endocrine**. 2006 Aug;30(1):45-53. Disponível em: encurtador.com.br/itA19. Acesso em 12 jan. 2021. encurtador.com.br/fgtBT.

VILLAROEL, Rosivaldo Gabriel. **O uso da análise de sentimentos como ferramenta de apoio à gestão acadêmica. São Paulo**. Dissertação (Mestrado em Engenharia da Computação) – Instituto de Pesquisas Tecnológicas do Estado de São Paulo, São Paulo, 2020.

WORLD BANK. Data for Brazil, United States, Mexico, Japan, Spain. **The World Bank Group at Work**. Washington D.C.: World Bank, 2019. Disponível em: encurtador.com.br/bfMR6. Acesso em 6 Jan 2021.

APÊNDICE A:

Avaliação de riscos segundo a American Diabetes Association (ADA, 2020).

A ADA possui uma ferramenta de avaliação de risco, um instrumento de acesso aberto na internet, em língua inglesa e, futuramente, o será em língua espanhola. No quadro 1, foram ordenadas as perguntas e as opções de respostas.

Quadro 1. Questões de avaliação de risco ADA (2020, tradução nossa)

Pergunta	***Opções***	
"Tem quantos anos?"	Menos de 40 anos	
	40-49 anos	
	50-59 anos	
	mais de 60 anos	
"É homem ou mulher?"	Masculino	
	Feminino	Você já foi diagnosticado com diabetes gestacional? Sim Não
"Sua mãe, pai, irmão ou irmã tem diabetes?"	Sim	
	Não	
"Alguma vez foi diagnosticado com hipertensão?"	Sim	
	Não	
"É fisicamente ativo?"	Sim	
	Não	
"Que raça ou grupo étnico o descreve melhor?"	Branco	
	Hispano latino	
	Negro afro-americano	
	Asiático	
	Indígena americano	
	Nativo do Alaska	
	Nativo do Havaí ou outra ilha do Pacífico	
	Outro	
	Prefiro não dizer	
"Qual sua altura e peso?"	Altura	
	Peso	

Fonte: autoria própria com base em (ADA, 2020).

O conteúdo da página de internet também traz pontos norteadores e que são comentados, conforme segue. A primeira pergunta é sobre a idade: quanto maior a idade maior é o risco de desenvolvimento da DM2, doença observada mais comumente a partir da idade adulta.

No que se refere ao gênero: homens são mais propensos a ter diabetes não diagnosticada fato este explicável pela falta de consultas médicas regulares. Campanhas como "Novembro Azul" direcionadas à prevenção do câncer de próstata, assim como outras, vem de encontro a esta necessidade de consultas

médicas de acompanhamento da saúde e preventivas não se restringindo às emergências médicas.

Há que se notar, no entanto, quando no papel de funcionário, homem ou mulher, podem se ver tolhidos desta necessidade para preservação de seus empregos, principalmente, nos cargos mais baixos da hierarquia da empresa. Agrava a situação os casos de empregados que se utilizam, deliberadamente, de falsos ou desnecessários atestados médicos para faltas ao trabalho.

No que se refere ao gênero, as mulheres podem sofrer com a diabetes gestacional, um tipo de diabetes que se desenvolve durante a gestação e se encerra com ela. Eis, portanto, outro fator de risco para diabetes mellitus tipo 2 assim causar macrossomia no bebê.

Segundo STAVIS (2019), há raras doenças que podem causar a macrossomia, que é o peso ao nascer superior à 4kg e “o diabetes mellitus materno é a principal causa de recém-nascidos grandes para a idade gestacional (GIG)”.

Em referência ao gênero feminino, a avaliação de risco poderá vir a considerar a ocorrência de Síndrome dos Ovários Policísticos (SOP) ou Micropolicístico (SOMP), uma patologia endócrina mais comum em mulheres em idade reprodutiva. Sua incidência afeta de 6% a 10% da população (TEEDE et al., 2006).

> Mulheres com SOP apresentam risco elevado para intolerância à glicose (IG) e diabetes mellitus do tipo 2 (DM-2). Por este motivo, o rastreamento para as alterações do metabolismo da glicose é recomendado para todas as pacientes com SOP. O DM é um problema de saúde pública, sendo que a SOP é considerada um fator predisponente para o desenvolvimento do DM. (PONTES et al., 2012).

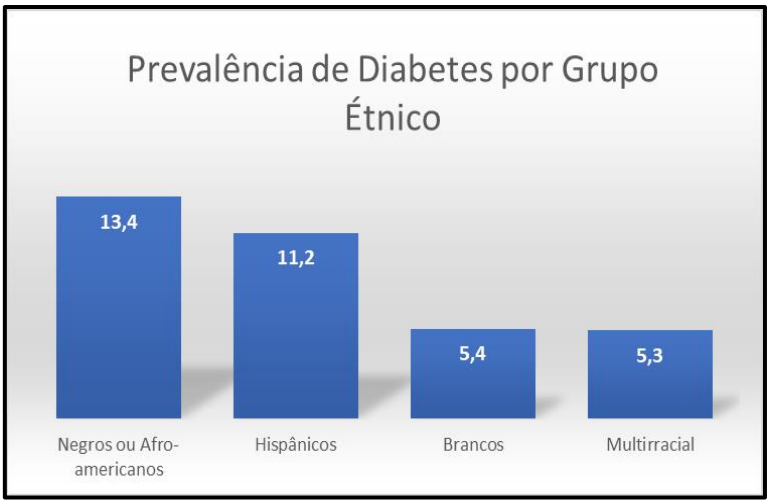
Podem contribuir para o desenvolvimento da DM2 ter antecedentes familiares, por isso, é importante o conhecimento de histórico familiar além de que ela conste na anamnese.

Representam risco o diagnóstico de hipertensão arterial assim como o pertencimento à determinados grupos raciais ou étnicos, porquanto, alguns grupos se observam maiores riscos. Ao tratar sobre os índios Pima, um grupo homogêneo

habitante da região norte-americana do Arizona e no México, nota-se que são predispostos geneticamente para a DM2.

O portal oficial de dados abertos da cidade de Austin, (AUSTIN, 2019) no Estado do Texas, Estados Unidos, para os anos combinados de 2011-2015, negros ou afro-americanos residentes têm a maior prevalência de diabetes (13,4%), em comparação com os hispânicos (11,2%), brancos (5,4%) e outros adultos de raça/multirracial (5,3%), como na figura abaixo.

Figura 1. Gráfico da prevalência de diabetes por grupo étnico baseado em, (AUSTIN, 2019).



Fonte: autoria própria baseado em Austin (2019, tradução nossa).

Duas condições importantes que finalizam as considerações e que aumentam o risco de desenvolvimento de DM2 é a inatividade física e o IMC, obtidos por uma expressão que tem como variáveis duas medidas corporais: peso e altura.

O que se corrobora pelos dados epidemiológicos providos pela SBD (2019) reveladores de que os principais fatores de risco para o diabetes mellitus tipo 2, dentre os quais se destacam nos dois primeiros lugares, o índice de massa corporal elevado (maior que 25) e baixa atividade física.

APÊNDICE B

Relato dos experimentos

Neste apêndice estão relatados experimentos realizados no curso desta pesquisa que se firmaram como aporte à avaliação de algoritmos candidatos à geração do modelo preditivo. Procurou-se padronizar a porcentagem de testes em 0.33, ou 33%.

NHANES 2017-2018 Sem nulos

Submetendo o conjunto de dados à aplicação dos algoritmos selecionados, com a exclusão de todos os valores nulos.

Saliente-se que a composição da amostra, um total de 2490 registros, com 39 variáveis de entrada ou preditoras e uma classe alvo. Do total, 2.233 pertencem à classe 0, não portador de diabetes mellitus tipo 2, e 257, à classe 1, portador de diabetes mellitus tipo 2, tornando assim, dados desbalanceados. Esta distribuição dos dados pela classe está representada na figura 1, respectivamente 89,7% e 10,3%.

Figura 1. Distribuição dos registros pela variável preditiva do *dataset* NHANES 2017-2018 Sem nulos

Fonte: autoria própria.

Além disso, 1120 registros pertencem à respondentes do sexo masculino, e 1270, do sexo feminino. Para este conjunto, a tabela 1 apresenta a acurácia, ordenada do maior para o menor valor resultante.

Tabela 1. Valores de acurácia obtida a partir do conjunto NHANES 2017-2018 Sem nulos.

Algoritmo	Acurácia (%)
Regressão Logística	90,02
SVM	90,02
KNN	89,42
Naïve Bayes	86,86
AdaBoosting	86,25
Random Forest	82,48
Árvore de Decisão	79,44

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 2, também apresentadas na mesma ordem dos resultados de acurácia

Tabela 2. Matriz de Confusão dos resultados dos algoritmos ordenados pelo valor de acurácia referente à tabela 1.

Regressão Logística		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	730	10	740	98,6	1,4
	Negativa	72	10	82	87,8	12,2
SVM		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	740	0	740	100	0
	Negativa	82	0	82	100	0
KNN		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	734	6	740	99,2	0,8
	Negativa	81	1	82	98,8	1,2
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	681	59	740	92	8
	Negativa	49	33	82	59,8	40,2
AdaBoosting		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	690	50	740	93,2	6,8
	Negativa	63	19	82	76,8	23,2
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	672	68	740	90,8	9,2
	Negativa	76	6	82	92,7	7,3
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	641	99	740	86,6	13,4
	Negativa	70	12	82	85,4	14,6

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 3.

Tabela 3. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 2.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
Regressão Logística	98,65	87,8	9,98	12,2	55,43
SVM	100	100	9,98	0	50
KNN	99,19	98,78	10,58	1,22	50,21
Naïve Bayes	92,03	59,76	13,14	40,24	66,14
AdaBoosting	93,24	76,83	13,75	23,17	58,21
Random Forest	90,81	92,68	17,52	7,32	49,07
Árvore de Decisão	86,62	85,37	20,56	14,63	50,63

Fonte: autoria própria.

NHANES 2017-2018 Sem nulos e balanceamento (*NearMiss*).

O conjunto formado pelos mesmos atributos foi testado com a aplicação da técnica de balanceamento pelo *NearMiss.*

Saliente-se que a composição da amostra, um total de 514 registros, com 39 variáveis de entrada ou preditoras e uma classe alvo. Em virtude do balanceamento, 257 pertencem à classe 0, não portador de diabetes mellitus tipo 2, e 257, à classe 1, portador de diabetes mellitus tipo 2, divididos com porcentagem igual de 50% cada, representados na figura 2.

Figura 2. Distribuição dos registros pela variável preditiva em *dataset* com dados balanceados.

Fonte: autoria própria.

Segue a tabela 4 que apresenta a acurácia dos algoritmos selecionados, ordenada do maior para o menor valor resultante.

Tabela 4. Valores de acurácia obtida a partir do conjunto NHANES 2017-2018 Sem nulos e balanceamento (*NearMiss*)

Algoritmo	Acurácia (%)
KNN	59,41
Regressão Logística	58,82
Naïve Bayes	54,71
AdaBoosting	54,12
Random Forest	52,94
SVM	48,24
Árvore de Decisão	45,88

Fonte: autoria própria.

Também as matrizes de confusão de cada algoritmo, na tabela 5, apresentadas na mesma ordem dos resultados da tabela anterior (tabela 4).

Tabela 5. Matriz de Confusão dos resultados dos algoritmos ordenados pelo valor de acurácia referente à tabela 4.

KNN		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	65	16	81	80,2	19,8
	Negativa	53	36	89	59,6	40,4
Regressão Logística		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	58	23	81	71,6	28,4
	Negativa	47	42	89	52,8	47,2
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	71	10	81	87,7	12,3
	Negativa	67	22	89	75,3	24,7
AdaBoosting		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	44	37	81	54,3	45,7
	Negativa	41	48	89	46,1	53,9
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	42	39	81	51,9	48,1
	Negativa	41	48	89	46,1	53,9
SVM		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	81	0	81	100	0
	Negativa	88	1	89	98,9	1,1
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	40	41	81	49,4	50,6
	Negativa	51	38	89	57,3	42,7

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 6.

Tabela 6. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 5.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
KNN	80,25	59,55	40,59	40,45	60,35
Regressão Logística	71,6	52,81	41,18	47,19	59,4
Naïve Bayes	87,65	75,28	45,29	24,72	56,19
AdaBoosting	54,32	46,07	45,88	53,93	54,13
Random Forest	51,85	46,07	47,06	53,93	52,89
SVM	100	98,88	51,76	1,12	50,56
Árvore de Decisão	49,38	57,3	54,12	42,7	46,04

Fonte: autoria própria.

NHANES 2017-2018 Substituição de Nulos

Além da exclusão de valores nulos, outra opção é substituir os valores nulos por um valor fixo preenchendo com valor neutro, como (-1). Submetendo o conjunto de dados à aplicação dos algoritmos selecionados, com a substituição de todos os valores nulos por (-1), a tabela 7 apresenta a acurácia, ordenada do maior para o menor valor resultante.

Saliente-se que a composição da amostra, um total de 9.920 registros, com 39 variáveis de entrada ou preditoras e uma classe alvo. Do total, 9.210 pertencem à classe 0, não portador de diabetes mellitus tipo 2, e 710, à classe 1, portador de diabetes mellitus tipo 2, tornando assim, dados desbalanceados. Trata-se, respectivamente, de 92,8% e 7,23%.

Além disso, 4.904 registros pertencem à respondentes do sexo masculino, e 5016, do sexo feminino.

Tabela 7. Valores de acurácia obtida a partir do conjunto NHANES 2017-2018 Substituição de Nulos

Algoritmo	Acurácia (%)
SVM	92,33
KNN	92,15
Regressão Logística	91,88
AdaBoosting	91,36
Random Forest	87,94
Árvore de Decisão	86,68
Naïve Bayes	77,58

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 8, também apresentadas na mesma ordem da tabela anterior.

Tabela 8. Matriz de Confusão dos resultados dos algoritmos ordenados pelo valor de acurácia referente à tabela 7.

SVM		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	3023	0	3023	100	0
	Negativa	251	0	251	100	0
KNN		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	3017	6	3023	99,8	0,2
	Negativa	251	0	251	100	0
Regressão Logística		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	2988	35	3023	98,2	1,2
	Negativa	231	20	251	92	8
AdaBoosting		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	2941	82	3023	97,3	2,7
	Negativa	201	50	251	80,1	19,9
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	2870	153	3023	94,9	5,1
	Negativa	242	9	251	96,4	3,6
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	2819	204	3023	93,3	6,7
	Negativa	232	19	251	92,4	7,6
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	2370	653	3023	78,4	21,6
	Negativa	81	170	251	32,3	67,7

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 9.

Tabela 9. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 8.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
SVM	100	100	7,67	0	50
KNN	99,8	100	7,85	0	49,9
Regressão Logística	98,84	92,03	8,12	7,97	53,41
AdaBoosting	97,29	80,08	8,64	19,92	58,61
Random Forest	94,94	96,41	12,06	3,59	49,27
Árvore de Decisão	93,25	92,43	13,32	7,57	50,41
Naïve Bayes	78,4	32,27	22,42	67,73	73,07

Fonte: autoria própria.

NHANES 2017-2018 Substituição de Nulos e balanceamento (NearMiss).

O conjunto de dados relacionado, desta vez, sofreu o balanceamento pelo *NearMiss*, e, na tabela 10, apresenta a acurácia, ordenada do maior para o menor valor resultante.

Saliente-se que a composição da amostra, um total de 1.420 registros, com 39 variáveis de entrada ou preditoras e uma classe alvo. Em virtude do balanceamento, 710 pertencem à classe 0, não portador de diabetes mellitus tipo 2, e 710, à classe 1, portador de diabetes mellitus tipo 2, divididos com porcentagem igual de 50% cada. Quanto ao sexo biológico, os números mantêm proximidade: contém 719 registros pertencem à respondentes do sexo masculino, e 701, do sexo feminino.

Tabela 10. Valores de acurácia obtida a partir do conjunto NHANES 2017-2018 Substituição de Nulos e balanceamento (NearMiss)

Algoritmo	Acurácia (%)
Regressão Logística	89,55
Naïve Bayes	88,7
KNN	88,06
AdaBoosting	87,63
Random Forest	86,99
SVM	83,16
Árvore de Decisão	82,3

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 11, também apresentadas na mesma ordem da tabela anterior.

Tabela 11. Matriz de Confusão dos resultados dos algoritmos ordenados pelo valor de acurácia referente à tabela 10.

Regressão Logística		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	211	36	247	85,4	14,6
	Negativa	13	209	222	5,9	94,1
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	213	34	247	86,2	13,8
	Negativa	19	203	222	8,6	91,4
KNN		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	212	35	247	85,8	14,2
	Negativa	21	201	222	9,5	90,5
AdaBoosting		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	211	36	247	85,4	14,6
	Negativa	22	200	222	9,9	90,1
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	204	43	247	82,6	17,4
	Negativa	18	204	222	8,1	91,9
SVM		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	218	29	247	88,3	11,7
	Negativa	50	172	222	22,5	77,5
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	205	42	247	83	17
	Negativa	41	181	222	18,5	81,5

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 12.

Tabela 12. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 11.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
Regressão Logística	85,43	5,86	100	94,14	89,79
Naïve Bayes	86,23	8,56	100	91,44	88,84
KNN	85,83	9,46	100	90,54	88,19
AdaBoosting	85,43	9,91	100	90,09	87,76
Random Forest	82,59	8,11	100	91,89	87,24
SVM	88,26	22,52	100	77,48	82,87
Árvore de Decisão	83	18,47	100	81,53	82,27

Fonte: autoria própria.

NHANES 2017-2018 e NHANES 2015-2016 Substituição de nulos

Ao conjunto já selecionado, foram adicionados os dados do período 2015-2016. Submetendo o conjunto de dados à aplicação dos algoritmos selecionados, com a substituição de todos os valores nulos por (-1), a tabela 13 apresenta a acurácia, ordenada do maior para o menor valor resultante.

Saliente-se que a composição da amostra, um total de 20.505 registros, com 39 variáveis de entrada ou preditoras e uma classe alvo. Do total, 19.135 pertencem à classe 0, não portador de diabetes mellitus tipo 2, e 1.370, à classe 1, portador de diabetes mellitus tipo 2, tornando assim, dados desbalanceados. Trata-se, respectivamente, de 93,3% e 6,7%.

Além disso, 10.116 registros pertencem à respondentes do sexo masculino, e 10.389, do sexo feminino.

Tabela 13. Valores de acurácia obtida a partir do conjunto NHANES 2017-2018 e NHANES 2015-2016 Substituição de nulos

Algoritmo	Acurácia (%)
SVM	93,11
KNN	92,92
Regressão Logística	92,7
AdaBoosting	92,37
Random Forest	88,7
Árvore de Decisão	87,75
Naïve Bayes	79,34

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 14, também apresentadas na mesma ordem da tabela anterior.

Tabela 14. Matriz de Confusão dos resultados dos algoritmos ordenados pelo valor de acurácia referente à tabela 13.

SVM		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	6301	0	6301	100	0
	Negativa	466	0	466	100	0
KNN		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6286	15	6301	99,8	0,2
	Negativa	464	2	466	99,6	0,4
Regressão Logística		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6236	65	6301	99	1
	Negativa	429	37	466	92,7	7,9
AdaBoosting		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6129	172	6301	97,3	2,7
	Negativa	344	122	466	73,8	26,2
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	5982	319	6301	94,9	5,1
	Negativa	446	20	466	95,7	4,3
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	5904	397	6301	93,7	6,3
	Negativa	432	34	466	92,7	7,3
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	5043	1258	6301	80	20
	Negativa	140	326	466	30	70

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 15.

Tabela15. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 14.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
SVM	100	100	6,89	0	50
KNN	99,76	99,57	7,08	0,43	50,1
Regressão Logística	98,97	92,06	7,3	7,94	53,46
AdaBoosting	97,27	73,82	7,63	26,18	61,73
Random Forest	94,94	95,71	11,3	4,29	49,62
Árvore de Decisão	93,7	92,7	12,25	7,3	50,5
Naïve Bayes	80,03	30,04	20,66	69,96	75

Fonte: autoria própria.

NHANES 2017-2018 e NHANES 2015-2016 Substituição de Nulos e balanceamento (NearMiss).

Ao conjunto de dados anterior, foi aplicado o balanceamento pelo *NearMiss* e os resultados da acurácia estão apresentado na tabela 16, ordenada do maior para o menor valor resultante.

Saliente-se que a composição da amostra, um total de 2.740 registros, com 39 variáveis de entrada ou preditoras e uma classe alvo. Do total, 1.370 pertencem à classe 0, não portador de diabetes mellitus tipo 2, e 1.370, à classe 1, portador de diabetes mellitus tipo 2, tornando assim, dados balanceados, divididos com porcentagem igual de 50% cada.

Do total, 1.390 registros pertencem à respondentes do sexo masculino, e 1.350, do sexo feminino.

Tabela 16. Valores de acurácia obtida a partir do conjunto NHANES 2017-2018 e NHANES 2015-2016 Substituição de Nulos e balanceamento (NearMiss).

Algoritmo	Acurácia (%)
Regressão Logística	89,06
Naïve Bayes	88,51
AdaBoosting	87,85
KNN	86,19
Random Forest	83,54
SVM	83,43
Árvore de Decisão	77,79

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 17, também apresentadas na mesma ordem da tabela anterior.

Tabela 17. Matriz de Confusão dos resultados dos algoritmos ordenados pelo valor de acurácia referente à tabela 16.

Regressão Logística		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	386	68	454	85	15
	Negativa	31	420	451	6,9	93,1
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	384	70	454	84,6	15,4
	Negativa	34	417	451	7,5	92,5
AdaBoosting		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	386	68	454	85	15
	Negativa	42	409	451	9,3	90,7
KNN		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	395	59	454	87	13
	Negativa	66	385	451	14,6	85,4
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	367	87	454	80,8	19,2
	Negativa	62	389	451	13,7	86,3
SVM		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	395	59	454	87	13
	Negativa	91	360	451	20,2	79,8
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	365	89	454	80,4	19,6
	Negativa	112	339	451	20,2	79,8

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 18.

Tabela 18. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 17.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
Regressão Logística	85,02	6,87	10,94	93,13	89,08
Naïve Bayes	84,58	7,54	11,49	92,46	88,52
AdaBoosting	85,02	9,31	12,15	90,69	87,86
KNN	87	14,63	13,81	85,37	86,19
Random Forest	80,84	13,75	16,46	86,25	83,55
SVM	87	20,18	16,57	79,82	83,41
Árvore de Decisão	80,4	24,83	22,21	75,17	77,79

Fonte: autoria própria.

JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS (Balanceamento Natural).

Ao conjunto de dados formado foi adicionada, como balanceamento natural, 600 registros resultantes do conjunto NHANES 2013-2014 referentes à

classe 1, com a substituição de todos os valores nulos por (-1). A tabela 19 apresenta os valores de acurácia, ordenada do maior para o menor valor resultante.

Saliente-se que a composição da amostra, um total de 21.105 registros, com 37 variáveis de entrada ou preditoras e uma classe alvo. Essa redução se deve à não ter sido encontrada, no referido montante, dados de exame de verificação de Ferritina.

Do total, 19.135 pertencem à classe 0, não portador de diabetes mellitus tipo 2, e 1.970, à classe 1, portador de diabetes mellitus tipo 2, tornando assim, dados desbalanceados. Trata-se, respectivamente, de 90,7% e 9,3%.

Além disso, 10.404 registros pertencem à respondentes do sexo masculino, e 10.701, do sexo feminino.

Tabela 19. Valores de acurácia obtida a partir do conjunto JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS (Balanceamento Natural).

Algoritmo	Acurácia (%)
KNN	92,95
AdaBoosting	92,53
Regressão Logística	90,75
SVM	90,44
Random Forest	89,06
Árvore de Decisão	88,13
Naïve Bayes	81,74

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 20 também apresentadas na mesma ordem da tabela anterior.

Tabela 20. Matriz de Confusão dos resultados dos algoritmos ordenados pelo valor de acurácia referente à tabela 19.

KNN		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	6282	17	6299	93	7
	Negativa	474	192	666	71,2	28,8
AdaBoosting		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6140	159	6299	97,5	2,5
	Negativa	361	305	666	54,2	45,8
Regressão Logística		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6191	108	6299	98,3	1,7
	Negativa	536	130	666	80,5	19,5
SVM		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6299	0	6299	100	0
	Negativa	666	0	666	100	0
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	5983	316	6299	95	5
	Negativa	446	220	666	67	33
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	5913	386	6299	93,9	6,1
	Negativa	441	225	666	66,2	33,8
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	5235	1064	6299	83,1	16,9
	Negativa	208	458	666	31,2	68,8

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 21.

Tabela 21. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 20.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
KNN	99,73	71,17	7,05	28,83	64,28
AdaBoosting	97,48	54,2	7,47	45,8	71,64
Regressão Logística	98,29	80,48	9,25	19,52	58,91
SVM	100	100	9,56	0	50
Random Forest	94,98	66,97	10,94	33,03	64,01
Árvore de Decisão	93,87	66,22	11,87	33,78	63,83
Naïve Bayes	83,11	31,23	18,26	68,77	75,94

Fonte: autoria própria.

JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS Balanceamento (NearMiss).

Ao conjunto de dados foi aplicado o balanceamento pelo *NearMiss* e os resultados da acurácia estão apresentado na tabela 22, ordenada do maior para o menor valor resultante.

Saliente-se que a composição da amostra, um total de 3.940 registros, com 37 variáveis de entrada ou preditoras e uma classe alvo. Do total, 1.970 pertencem à classe 0, não portador de diabetes mellitus tipo 2, e 1970, à classe 1, portador de diabetes mellitus tipo 2, tornando assim, dados balanceados, divididos com porcentagem igual de 50% cada.

Do total, 1.990 registros pertencem à respondentes do sexo masculino, e 1.941, do sexo feminino.

Tabela 22. Valores de acurácia obtida a partir do conjunto JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS Balanceamento (NearMiss)

Algoritmo	Acurácia (%)
AdaBoosting	92,95
Regressão Logística	92,53
Naïve Bayes	90,75
KNN	90,44
Random Forest	89,06
SVM	88,13
Árvore de Decisão	81,74

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 23, também apresentadas na mesma ordem da tabela anterior.

Tabela 23. Matriz de Confusão dos resultados dos algoritmos ordenados pelo valor de acurácia referente à tabela 22.

AdaBoosting		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	524	139	663	79	21
	Negativa	70	568	638	11	89
Regressão Logística		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	520	143	663	78,4	21,6
	Negativa	73	565	638	11,4	88,6
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	492	171	663	74,2	25,8
	Negativa	69	569	638	10,8	89,2
KNN		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	556	107	663	83,9	16,1
	Negativa	138	500	638	21,6	78,4
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	503	160	663	75,9	24,1
	Negativa	113	525	638	17,7	82,3
SVM		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	521	142	663	78,6	21,4
	Negativa	132	506	638	28,8	71,2
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	507	156	663	76,5	23,5
	Negativa	184	454	638	28,8	71,2

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 24.

Tabela 24. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 23.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
AdaBoosting	79,03	10,97	7,05	89,03	84,03
Regressão Logística	78,43	11,44	7,47	88,56	83,5
Naïve Bayes	74,21	10,82	9,25	89,18	81,7
KNN	83,86	21,63	9,56	78,37	81,12
Random Forest	75,87	17,71	10,94	82,29	79,08
SVM	78,58	20,69	11,87	79,31	78,95
Árvore de Decisão	76,47	28,84	18,26	71,16	73,82

Fonte: autoria própria.

Algoritmos de aprendizado de máquina podem ter problemas com muitos atributos, o que podemos verificar ao notar que os conjuntos de dados contêm entre 37 e 39 variáveis cada. Por isso, foi realizado um processo de seleção e redução de variáveis que foram elencadas no quadro 1:

Quadro 1. Descrição e classificação das variáveis selecionadas

	Descrição		
SEQN1	Número sequencial	Qualitativo	Ordinal
RIAGENDR	Sexo	Qualitativo	Ordinal
RIDAGEYR	Idade	Quantitativo	Discreto
LBXGH	Hemoglobina Glicada	Quantitativo	Continuo
LBXIN	Insulina (uU/mL)	Quantitativo	Continuo
LBDINSI	Insulina (pmol/L)	Quantitativo	Continuo
BMXWT	Peso (kg)	Quantitativo	Continuo
BMXHT	Altura em pé (cm)	Quantitativo	Continuo
BMXBMI	Índice de Massa Corporal	Quantitativo	Continuo
BMXWAIST	Circunferência da cintura (cm)	Quantitativo	Continuo
LBXSAL	Albumina, soro refrigerado (g/dL)	Quantitativo	Continuo
LBDSALSI	Albumina, soro refrigerado (g/L)	Quantitativo	Continuo
LBXSCH	Colesterol, soro refrigerado (mg/dL)	Quantitativo	Continuo
LBDSCHSI	Colesterol, soro refrig (mmol/L)	Quantitativo	Continuo
LBXSCR	Creatinina, soro refrigerado (mg/dL)	Quantitativo	Continuo
LBDSCRSI	Creatinina, soro refrigerado (umol/L)	Quantitativo	Continuo
LBXSGL	Glicose, soro refrigerado (mg/dL)	Quantitativo	Continuo
LBDSGLSI	Glicose, soro refrigerado (mmol/L)	Quantitativo	Continuo
LBXSTR	Triglicerídeos, soro refrig (mg / dL)	Quantitativo	Continuo
LBDSTRSI	Triglicerídeos, soro refrig (mmol / L)	Quantitativo	Continuo
BPXSY1	Sistólica: pressão arterial (primeira leitura) mm Hg	Quantitativo	Continuo
BPXDI1	Diastólica: pressão arterial (primeira leitura) mm Hg	Quantitativo	Continuo
CLASSE	Classe Alvo	Quantitativo	Discreto

Fonte: autoria própria.

JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS REDUÇÃO (Redução de Variáveis).

Para o conjunto de dados reduzido, a tabela 25 apresenta a acurácia, ordenada do maior para o menor valor resultante. Saliente-se que a composição da amostra, um total de 21.105 registros, com 22 variáveis de entrada ou preditoras e uma classe alvo.

Do total, 19.135 pertencem à classe 0, não portador de diabetes mellitus tipo 2, e 1.970, à classe 1, portador de diabetes mellitus tipo 2, tornando assim, dados desbalanceados. Trata-se, respectivamente, de 90,7% e 9,3%.

Além disso, 10.404 registros pertencem à respondentes do sexo masculino, e 10.701, do sexo feminino.

Tabela 25. Valores de acurácia obtida a partir do conjunto JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS REDUÇÃO (Redução de Variáveis)

Algoritmo	Acurácia (%)
AdaBoosting	92,78
KNN	92,33
Regressão Logística	90,88
SVM	90,44
Random Forest	88,99
Árvore de Decisão	87,77
Naïve Bayes	76,94

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 26, também apresentadas na mesma ordem da tabela anterior.

Tabela 26. Matriz de Confusão dos resultados dos algoritmos ordenados pelo valor de acurácia referente à tabela 25.

AdaBoosting		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	6167	132	6299	97,9	2,1
	Negativa	370	296	666	55,6	44,4
KNN		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6222	77	6299	98,8	1,2
	Negativa	457	209	666	68,6	31,4
Regressão Logística		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6188	111	6299	98,2	1,8
	Negativa	524	142	666	78,7	21,3
SVM		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6299	0	6299	100	0
	Negativa	666	0	666	100	0
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	5976	323	6299	94,9	5,1
	Negativa	444	222	666	66,7	33,3
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	5889	410	6299	93,5	6,5
	Negativa	442	224	666	66,4	33,6
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	4824	1475	6299	76,6	23,4
	Negativa	131	535	666	19,7	80,3

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 27.

Tabela 27. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 26.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
AdaBoosting	97,9	55,56	7,22	44,44	71,17
KNN	98,78	68,62	7,67	31,38	65,08
Regressão Logística	98,24	78,68	9,12	21,32	59,78
SVM	100	100	9,56	0	50
Random Forest	94,87	66,67	11,01	33,33	64,1
Árvore de Decisão	93,49	66,37	12,23	33,63	63,56
Naïve Bayes	76,58	19,67	23,06	80,33	78,46

Fonte: autoria própria.

JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS REDUÇÃO Balanceamento (NearMiss).

Ao conjunto de dados anterior, foi aplicado o balanceamento pelo NearMiss e os resultados da acurácia estão apresentado na tabela 28, ordenada do maior para o menor valor resultante.

Saliente-se que a composição da amostra, um total de 3.940 registros, com 22 variáveis de entrada ou preditoras e uma classe alvo. Do total, 1.970 pertencem à classe 0, não portador de diabetes mellitus tipo 2, e 1970, à classe 1, portador de diabetes mellitus tipo 2, tornando assim, dados balanceados, divididos com porcentagem igual de 50% cada.

Do total, 2.025 registros pertencem à respondentes do sexo masculino, e 1.915, do sexo feminino.

Tabela 28. Valores de acurácia obtida a partir do conjunto JUNÇÃO NHANES 2017-2018 e NHANES 2015-2016 PLUS REDUÇÃO Balanceamento (NearMiss)

Algoritmo	Acurácia (%)
AdaBoosting	76,33
Naïve Bayes	72,79
Regressão Logística	71,48
KNN	69,1
Random Forest	66,26
SVM	62,34
Árvore de Decisão	60,88

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 29, também apresentadas na mesma ordem da tabela anterior.

Tabela 29. Matriz de Confusão dos resultados dos algoritmos ordenados pelo valor de acurácia referente à tabela 28.

AdaBoosting		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	521	142	663	78,6	21,4
	Negativa	166	472	638	26	74
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	598	65	663	90,2	9,8
	Negativa	289	349	638	45,3	54,7
Regressão Logística		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	492	171	663	74,2	25,8
	Negativa	200	438	638	31,3	68,7
KNN		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	576	87	663	86,9	13,1
	Negativa	315	323	638	49,4	50,6
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	425	238	663	64,1	35,9
	Negativa	201	437	638	31,5	68,5
SVM		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	495	168	663	74,7	25,3
	Negativa	322	316	638	50,5	49,5
Árvore de Decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	422	241	663	63,7	36,3
	Negativa	268	370	638	42	58

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 30.

Tabela 30. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 29.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
AdaBoosting	78,58	26,02	23,67	73,98	76,28
Naïve Bayes	90,2	45,3	27,21	54,7	72,45
Regressão Logística	74,21	31,35	28,52	68,65	71,43
KNN	86,88	49,37	30,9	50,63	68,76
Random Forest	64,1	31,5	33,74	68,5	66,3
SVM	74,66	50,47	37,66	49,53	62,1
Árvore de Decisão	63,65	42,01	39,12	57,99	60,82

Fonte: autoria própria.

NHANES 2017-2018 e NHANES 2015-2016 PLUS (Preenchimento mediana).

Uma técnica para lidar com dados nulos é o preenchimento com valores como média, moda ou mediana. Pelos fatores que foram expostos anteriormente, o preenchimento dos dados nulos foi feito com a mediana.

Assim, denotamos que na tabela 31 são apresentados os percentuais de acurácia referente aos algoritmos que foram selecionados, em ordem do maior para o menor valor.

Tabela 31. Valores de acurácia obtida a partir do conjunto NHANES 2017-2018 e NHANES 2015-2016 PLUS (Preenchimento mediana).

Algoritmo	Acurácia (%)
AdaBoosting	92,74
KNN	92,33
Regressão Logística	91,17
SVM	90,44
Random Forest	89,44
Naïve Bayes	88,82
Árvore de Decisão	87,94

Fonte: autoria própria.

Seguem as matrizes de confusão de cada algoritmo, na tabela 32, também apresentadas na mesma ordem da tabela anterior.

Tabela 32. Matriz de Confusão dos resultados dos algoritmos ordenados pelo valor de acurácia referente à tabela 31.

AdaBoosting		Classe predita				
		Positiva	Negativa	Total		
Classe original	Positiva	6152	147	6299	97,7	2,3
	Negativa	359	307	666	53,9	46,1
KNN		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6217	82	6299	98,7	1,3
	Negativa	452	214	666	67,9	32,1
Regressão Logística		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6188	111	6299	98,2	1,8
	Negativa	504	162	666	75,7	24,3
SVM		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	6299	0	6299	100	0
	Negativa	666	0	666	100	0
Random Forest		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	5988	311	6299	95,1	4,9
	Negativa	450	216	666	67,6	32,4
Naïve Bayes		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	5861	438	6299	93	7
	Negativa	341	325	666	51,2	48,8
Árvore de decisão		Classe predita				
		Positiva	Negativa			
Classe original	Positiva	5896	403	6299	93,6	6,4
	Negativa	437	229	666	65,6	34,4

Fonte: autoria própria.

Com as matrizes de confusão, foi possível realizar os cálculos de TVP, TFP, sensibilidade, erro e eficiência que estão abaixo descritas na tabela 33.

Tabela 33. Percentuais de TVP, TFP, erro, sensibilidade e eficiência referente à cada resultado da matriz de confusão da tabela 32.

Algoritmo	TVP (%)	TFP (%)	Erro (%)	Sensibilidade (%)	Eficiência (%)
AdaBoosting	97,67	53,9	7,26	46,1	71,89
KNN	98,7	67,87	7,67	32,13	65,42
Regressão Logística	98,24	75,68	8,83	24,32	61,28
SVM	100	100	9,56	0	50
Random Forest	95,06	67,57	10,56	32,43	63,75
Naïve Bayes	93,05	51,2	11,18	48,8	70,93
Árvore de Decisão	93,6	65,62	12,06	34,38	63,99

Fonte: autoria própria.