Modelos de Aprendizado de Máquina na Predição de Diabetes Tipo 1 na Gestação usando Dados do Sistema Único de Saúde

Jorge R. H. Moreira[1], **Heder S. Bernardino**[1], **Helio J. C. Barbosa**[1,2], **Alex B. Vieira**[1]

[1] Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora (UFJF)

[2]Laboratório Nacional de Computação Científica (LNCC)

{jorge.moreira,heder}@ice.ufjf.br, hcbm@lncc.br, alex.borges@ufjf.br

Abstract. *Pre-existing or developed diabetes mellitus during pregnancy can bring serious health risks to the pregnant woman and the baby, throughout the pregnancy-puerperal cycle. In this sense, predicting the presence of one of the types of diabetes, even before its first symptoms, can generate positive impacts on public health systems. This work generates a classification model, using data from the outpatient production of the Unified Health System, which can predict the presence of one of the types of diabetes (type 1), according to the characteristics and history of monitoring the pregnant patient. The proposed classifier is capable of separating pregnant women in the health system according to their predisposition to the disease, making it possible to generate an alert in the system and direct attention to the monitoring of pregnant women in the context of this health condition. The results obtained were relevant, presenting sensitivity and precision greater than 90%. Thus, it is believed that the proposed model can be another resource for improving the system.*

Resumo. *A diabetes mellitus pré-existente ou desenvolvida na gestação pode trazer sérios riscos de saúde à gestante e ao bebê, durante todo o ciclo gravídico-puerperal. Nesse sentido, predizer a presença de um dos tipos de diabetes, mesmo antes de seus primeiros sintomas, pode gerar impactos positivos nos sistemas de saúde pública. Este trabalho buscou gerar um modelo de classificação, utilizando dados da produção ambulatorial do Sistema Único de Saúde, que possa predizer a presença de um dos tipos de diabetes (tipo 1), conforme as características e o histórico de acompanhamento da paciente gestante. O classificador proposto é capaz de separar as gestantes no sistema de saúde conforme a predisposição à doença, possibilitando gerar um alerta ao sistema, e com isso, direcionar a atenção ao acompanhamento da gestante no âmbito dessa condição de saúde. Os resultados obtidos foram relevantes, apresentando sensibilidade e precisão superiores a 90%. Assim, acredita-se que o modelo proposto pode ser mais um recurso para aprimoramento do sistema.*

1. Introdução

Diabetes Mellitus (DM) é considerada um grupo de doenças metabólicas que possui como característica a hiperglicemia (excesso de glicose no sangue) associada a complicações, disfunções e insuficiência de vários órgãos [Ministério da Saúde 2006]. Esses distúrbios ocorrem quando o pâncreas não produz insulina suficiente ou quando

o corpo não consegue usar de forma eficaz esse hormônio [WHO 2020a, IDF 2020]. Os sinais e sintomas clássicos são: poliúria (aumento da frequência urinária), polidipsia (sede frequente), perda inexplicada de peso, polifagia (fome excessiva), além de outros sintomas menos específicos [Ministério da Saúde 2006].

Conforme a Organização Mundial de Saúde [WHO 2020a], o número de pessoas com diabetes mellitus passou de 108 milhões em 1980, para 422 milhões em 2014. A prevalência global de diabetes em adultos com mais de 18 anos de idade aumentou de 4,7%, para 8,5% nesse mesmo período. A OMS também expõe que entre 2000 e 2016, houve um aumento de 5% na mortalidade prematura por diabetes. Além disso, a prevalência dessa patologia tem aumentado mais rapidamente em países de baixa e média renda.

De acordo com [IDF 2015], o Brasil é o quarto país com maiores taxas de DM na população adulta, com um gasto anual estimado de pelo menos US$ 21,8 bilhões. Em relação à Diabetes Mellitus tipo 1, apresenta-se como a 3º maior do mundo, com mais de 30 mil portadores da doença, gerando um custo estimado em torno de US$ 2,9 bilhões ao ano [Cavalcante et al. 2020]. Esse cenário é preocupante, carecendo de uma maior atenção a essa patologia, principalmente em grupos vulneráveis como gestantes. Com isso, sistemas de saúde pública precisam estar sempre preparados para atender à demanda, com recursos que melhorem a gestão e evitem a sobrecarga do sistema.

No Brasil, vigora o Sistema Único de Saúde (SUS) que é considerado um dos maiores sistemas de saúde pública do mundo [Albuquerque et al. 2011]. Criado em 1988 pela Constituição Federal Brasileira, é regido pelos princípios da universalidade, integridade e equidade nos serviços e ações de saúde, abrangendo as três esferas de governo. Ele vem buscando atuar cada vez mais na atenção básica da saúde, sendo esta a porta de entrada no Sistema. Essa camada de atuação opera na prevenção e tratamento de doenças antes de direcionar os casos mais graves para níveis de atendimento mais complexos.

Em suma, neste trabalho, são gerados modelos de classificação a partir de características da paciente gestante (altura, índice de massa corporal e histórico de diabetes na família) e da sua região de atendimento. Esses modelos auxiliam na detecção de diabetes mellitus do tipo 1 em gestantes sem diagnóstico prévio da doença, podendo gerar um alerta num sistema de saúde para a atenção à paciente.

Os classificadores foram avaliados utilizando dados reais obtidos a partir do Sistema de Informações Ambulatoriais do SUS que incluem atendimentos e procedimentos de baixa, média e alta complexidade. Os resultados indicam que os classificadores desenvolvidos podem ser utilizados para segmentar as gestantes no sistema de saúde conforme a predisposição a DM1 com sensibilidade e precisão superiores a 90%. No âmbito desse grupo de indivíduos e da DM1, isso melhoraria a relação entre os atendimentos e procedimentos realizados na atenção básica com os realizados em outro nível de complexidade.

2. Tipos da Diabetes Mellitus

Conforme [Ministério da Saúde 2006, Taser 2021, IDF 2020], a DM pode ser dividida em três tipos principais: tipo 1, tipo 2 e a diabetes mellitus gestacional. O tipo 1 (DM1), também chamada de insulino-dependente, acomete principalmente crianças e adolescentes, entretanto pode se manifestar em pessoas de qualquer idade [IDF 2020]. Geralmente é classificada como autoimune [Sociedade Brasileira de Diabetes 2019], não comum a presença de sobrepeso [Ministério da Saúde 2006, IDF 2020]. Entretanto, o ga-

nho de peso associado ao tratamento da doença, pode resultar em um problema clínico [Moulin et al. 2003]. Conforme [Acharjee et al. 2013], estudos também mostram que paciente com histórico familiar de diabetes mellitus tem maior risco de desenvolver a doença. Ademais, de acordo com [IDF 2020, Beeck and Eizirik 2016], a exposição a algumas infecções virais também foi associada ao risco de desenvolver o distúrbio. Já o tipo 2 (DM2), conhecida como insulino-independente, tem início insidioso e sintomas mais brandos. A maioria dos casos apresenta excesso de peso ou deposição central de gordura, havendo uma deficiência relativa de insulina [Ministério da Saúde 2006].

Além dos dois tipos mencionados, existe a Diabetes Mellitus Gestacional (DMG), na qual há um estado de hiperglicemia detectado pela primeira vez na gravidez, sendo menos severo que o tipo 1 e 2 [Ministério da Saúde 2006, OPAS 2016]. Essa patologia se dá por alterações hormonais durante o período da gravidez, que acabam elevando o nível de glicose no sangue. Pode ser transitória, se resolvendo no período pós-parto, ou pode persistir após o mesmo [Sociedade Brasileira de Diabetes 2019]. Os fatores de risco estão associados à idade materna avançada, sobrepeso/obesidade, índice de massa corporal (IMC) elevado, dentre outros [OPAS 2016, Sociedade Brasileira de Diabetes 2019].

Conforme mencionado, o DM possui como característica a hiperglicemia. Considerando essa condição na gravidez, pode-se classificá-la em três tipos [OPAS 2016]:

• Diabetes pré-gestacional: casos diagnosticados antes da gravidez, que são DM1 e DM2.
• Diabetes mellitus diagnosticado na gestação: sem diagnóstico prévio de DM, com níveis glicêmicos sanguíneos que atingem os critérios da Organização Mundial de Saúde (OMS) para a DM na ausência de gestação.
• Diabetes mellitus gestacional: conforme citado, consiste de hiperglicemia detectada pela primeira vez durante a gravidez, com níveis glicêmicos sanguíneos que não atingem os critérios diagnósticos para DM.

Além disso, destaca-se que a presença de DM na gestação está associada a elevado risco de complicações, consumindo recursos de saúde e atenção especializada [Ministério da Saúde 2006]. Além disso, a prevalência do diabetes pré-gestacional e do gestacional tem aumentando nos últimos anos e pode ser justificada pela epidemia de obesidade conforme [OPAS 2016, Feitosa and Ávila 2016] e aumento da idade materna [Feitosa and Ávila 2016]. Salienta-se também que a hiperglicemia durante a gravidez constitui um importante problema por haver riscos de maus desfechos perinatais e doenças futuras. [OPAS 2016] cita que essa patologia pode afetar os filhos das gestantes, aumentando os riscos destas crianças desenvolverem obesidade e diabetes na vida futura.

Diante do exposto, reforça-se que o Diabetes Mellitus é um grupo de patologias de grande importância, inclusive em gestantes, as quais constituem uma classe especial e que necessita de atenção devido aos riscos à mulher e ao bebê. Considera-se também que, o Diabetes 1 é um tipo relevante e de acompanhamento complexo, porém carecendo de pesquisas. Nesse sentido, destacamos que o presente estudo foi realizado englobando dados de gestantes que tiveram o diagnóstico de Diabetes tipo 1 pela primeira vez na gravidez, ou seja, se enquadrando no tipo Diabetes mellitus diagnosticado na gestação.

3. Trabalhos Relacionados

Nos últimos anos, tem-se visto um crescente interesse na aplicação de técnicas de aprendizado de máquina na área da saúde. Tais técnicas ajudam os médicos e profissio-

nais de saúde a melhorar diagnósticos clínicos e tratamentos de pacientes com doenças crônicas, como a diabetes [Mainenti et al. 2020].

Dentre os diversos trabalhos propostos, podemos citar o [Rodríguez-Rodríguez et al. 2021], onde os autores obtiveram características de 25 pacientes com diabetes do tipo 1, durante o acompanhamento de 14 dias. Foram coletadas características como quantidade de insulina, horários de alimentação, frequência cardíaca, exercícios realizados, dentre outros. O objetivo foi aplicar diferentes técnicas de seleção de variáveis, combinando-as com diferentes algoritmos de classificação. Os autores concluíram que o método *Random Forest* teve destaque para a predição da quantidade de insulina que deveria ser aplicada ao paciente.

Taser et al. [Taser 2021] utilizaram a técnica conhecida como comitê, para combinar seis classificadores baseados em árvore de decisão: *C4.5*, *random tree*, *REPTree*, *decision stump*, *Hoeffding tree*, e *NBTree*. Utilizaram o conjunto de dados de previsão de risco de diabetes em estágio inicial disponível em [UCI 2020] e concluíram que a aplicação de abordagens de *boosting* e *bagging* nos modelos superaram os resultados de precisão e acurácia dos que foram baseados apenas em classificadores individuais.

O modelo proposto em [Javad et al. 2019] abordou a técnica de aprendizado por reforço. O objetivo foi desenvolver e validar uma estrutura geral de aprendizagem por reforço para o tratamento personalizado da diabetes tipo 1 utilizando dez anos de dados clínicos de pacientes tratados no *Mass General Hospital* [MGH 2020]. O estudo demonstrou que um algoritmo de aprendizagem por reforço pode ser usado para recomendar doses personalizadas de insulina para atingir o controle glicêmico adequado em pacientes com DM1, apesar de afirmar ser necessário mais investigação utilizando uma amostra maior de pacientes para confirmar os achados.

Em [Chaves and Marques 2021], os autores utilizaram o conjunto de dados disponível em [UCI 2020]. Eles aplicaram os métodos *Naive Bayes*, Redes Neurais, *AdaBoost*, KNN, *Random Forest* e SVM com o objetivo de diagnosticar precocemente a presença de DM, utilizando 520 observações com 16 atributos de pacientes. O modelo utilizando redes neurais com 31 camadas ocultas, utilizando a função de ativação ReLu, otimizador L-BFGS-B e alfa 0,001, obteve os melhores resultados dentre todos os algoritmos propostos, utilizando as métricas AUC, acurácia, precisão, *recall* e *f1-score*.

No presente trabalho, foi adotada uma nova abordagem, utilizando dados reais de um sistema de informação de saúde brasileiro, tendo como foco mulheres gestantes. Além disso, foram criados modelos de classificação utilizando as características da gestante, como altura, índice de massa corporal e histórico de diabetes na família, além de sua região de atendimento. Esses modelos são utilizados para identificar precocemente diabetes DM1 nas pacientes gestantes.

4. Coleta, Pré-processamento e Análise dos dados

Nesta seção será apresentado como o trabalho foi conduzido desde a escolha e coleta dos dados, até a fase de análise. Inicialmente, destacamos algumas informações do conjunto de dados utilizado no trabalho que é disponibilizado, de forma aberta, pelo Departamento de Informática do Sistema Único de Saúde (DATASUS)[1].

[1]http://www2.datasus.gov.br/DATASUS/index.php?area=0901

Dentre os vários sistemas mantidos pelo DATASUS, destaca-se o Sistema de Informações Ambulatoriais (SIA), que mantém dados de atendimentos e procedimentos realizados a nível ambulatorial no SUS, possuindo como fontes de coleta o Boletim de Produção Ambulatorial Individualizado (BPA-I), além de registros de medicamentos e laudos diversos de Autorização de Procedimento de Alta Complexidade (APAC), sendo as fontes de dados utilizadas neste trabalho.

Além do vasto volume de dados de procedimentos, medicamentos e tratamentos realizados pelos pacientes do SUS no âmbito ambulatorial, essas bases possuem, para cada registro, o código de Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde, mais conhecido como CID. Isso possibilitou realizar a busca por diagnósticos de outras doenças que as pacientes gestantes tiveram no passado, no âmbito do sistema ambulatorial do SUS.

Os arquivos são disponibilizados pelo Datasus[2] de forma compactada com extensão DBC, rotulados com a sigla de referência do tipo de arquivo de dados, a sigla do estado, além do ano e mês de registro da produção ambulatorial. Os dados foram coletados entre 01/2008 e 07/2020, totalizando 12.231 arquivos e somando aproximadamente 117GB de arquivos compactados. A Tabela 1 resume as quantidades e os tamanhos totais dos arquivos de cada base de dados utilizados neste estudo, em *gigabytes*.

Tabela 1. Quantidade e tamanho dos arquivos de dados

	Qtd. Arquivos (DBC/DBF)	Arquivos DBC (GB)	Arquivos DBF (GB)
BPA-I Boletim de Produção Individual	4077	58,4	317,5
APAC Medicamentos	4225	9,7	49,6
APAC Laudos Diversos	4075	2,1	9,6
Total	12377	70,2	376,7

A Figura 1 resume o processo de coleta e preparação dos dados, mostrando a sequencia de passos utilizada até chegar à base unificada contendo as características e histórico de DM1 em gestantes. As três fontes de dados usadas aqui compartilham o atributo Cartão Nacional de Saúde (CNS), que é o identificador único do paciente, sendo o atributo utilizado para realizar a ligação entre as bases e criar um cenário de histórico das pacientes gestantes. A base APAC de medicamentos, além desse atributo em comum, possui os três atributos que seguem e que são utilizados neste trabalho contendo características das pacientes: peso, altura e um atributo sinalizando se a paciente é gestante.

Em relação aos códigos de doenças usados aqui, é importante ressaltar que a OMS utiliza duas classificações de referência para as doenças e descrição das condições de saúde de um paciente: a Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde (CID) e a Classificação Internacional de Funcionalidade, Incapacidade e Saúde (CIF)[WHO 2020b, Di Nubila and Buchalla 2008]. Este trabalho foi baseado na CID versão 10 (CID-10) [WHO 2019], que divide os mais variados tipos de doenças e condições de saúde em códigos separados por grupos. As doenças endócrinas, como a diabetes mellitus, estão no grupo E10-E14 do Capítulo IV. A DM tipo 1 (insulino-dependente), foco desse estudo, encontra-se no subgrupo E100-E109.

[2]ftp://ftp.datasus.gov.br/dissemin/publicos/SIASUS/200801_/Dados/

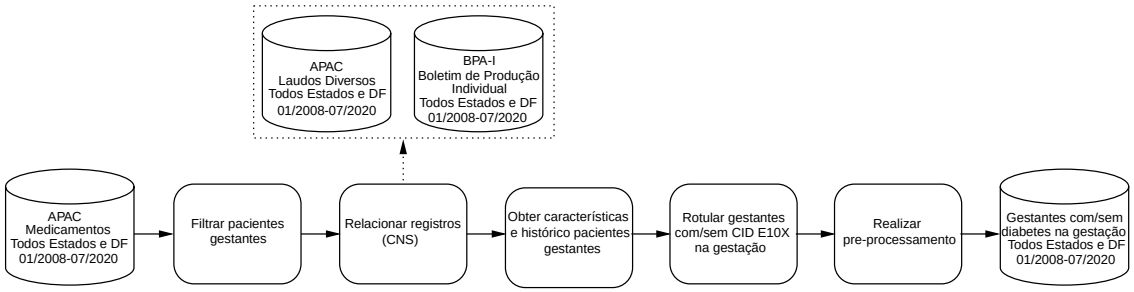


Figura 1. Fluxo da coleta e preparação dos dados

Visando a integração dos conjuntos de dados, inicialmente, a base APAC de Medicamentos foi separada em gestantes e não-gestantes, sendo que dentro do conjunto de gestantes foi realizado um filtro pelas que apresentavam diagnóstico de diabetes mellitus do tipo 1, conforme o CID da doença (rótulo E10X).

A outra parte da base (não-gestantes), foi utilizada para realizar uma varredura em busca da identificação de outros possíveis diagnósticos de diabetes das pacientes antes da gestação. Dessa forma, pode-se identificar se a paciente gestante já tinha passado por algum outro procedimento ou tratamento relacionado à diabetes que revelaria que ela já sabia que era portadora da enfermidade antes da gravidez e que, provavelmente, já realizava algum tratamento prévio em relação à doença.

Em um segundo momento, foi utilizada a base APAC de laudos diversos para realizar o mesmo procedimento - buscar possíveis registros das pacientes gestantes da base de medicamentos que pudessem ter algum histórico da doença. Por último, foi utilizada a base de Boletim de Produção Individualizado para também verificar o histórico familiar de diabetes mellitus (CID Z83.3) e outras doenças associadas a essa patologia, como infecções virais (CID B34X)[IDF 2020].

A Tabela 2, apresenta um resumo das informações de cada base de dados utilizada no trabalho e que foram disponibilizadas[3]. Com isso, diante dos dados compilados em um único conjunto de dados, foi feito um tratamento dos atributos de interesse.

Tabela 2. Informações das Bases

		Qtd. Registros	Qtd. Pacientes (CNS)
APAC Medicamentos	Gestantes	206.060	11.424
	Não-gestantes	99.566.881	41.917.459
APAC Laudos Diversos	Mulheres	19.319.766	8.569.189
BPA-I Boletim de Produção Ambulatorial Individual	Mulheres	913.700.906	253.350.826
Base Compilada	Gestantes com DM1 início da gestação e sem histórico da doença	886	81
	Gestantes sem DM1 e sem histórico da doença	67098	4977

Os atributos adotados nesse trabalho, selecionados a partir do processo de combinação das três bases de dados descritas anteriormente, são: idade, altura, raça,

[3]http://netlab.ice.ufjf.br/index.php/dm1inpregnancydata/

Tabela 3. Atributos - Conjunto de dados final

Atributo	Descrição	Tipo de dado
Idade	Idade em anos	Quantitativo discreto
Altura	Altura da paciente gestante	Quantitativo contínuo
IMC	Índice de massa corporal	Quantitativo contínuo
Raça	Raça da paciente gestante	Binário
Região	Região que foi realizado o atendimento/ procedimento registrado no SIA	Binário
DM	Paciente gestante diagnosticada ou não com diabetes mellitus tipo 1	Binário

estado onde foi realizado o atendimento/procedimento e o código CID relacionado ao diagnóstico de DM1. Esses dados foram tratados como segue:

• Criação de uma variável para adicionar o índice de massa corporal (IMC), calculado a partir da altura e peso da paciente, e depois transformando-o em variável categórica, conforme a tabela de adequação de peso durante a gestação apresentada em [Ministério da Saúde 2012]. Essa tabela divide o IMC em 4 categorias: (i) Baixo peso: IMC $\leq$ 19,9; (ii) Adequado: 20 $\leq$ IMC $\leq$ 24,9; (iii) Sobrepeso: 25 $\leq$ IMC $\leq$ 30; e (iv) Obesidade: IMC $\geq$ 30,1;
• Tendo em vista as diferenças sociais e econômicas entre as regiões do país que possam refletir na predisposição a doença, foi utilizado o atributo que continha o estado onde foi realizado o procedimento ou tratamento da gestante para criar o atributo região, conforme divisão regional proposta pelo IBGE[4];
• Padronização dos dados dos atributos idade, altura e peso;
• Transformação das variáveis raça e região em variáveis binárias.

Após realizar o pré-processamento dos dados, foi realizado um teste de correlação de Pearson entre as variáveis altura, peso e IMC, para verificar a possível correlação entre esses atributos. Correlação entre atributos, caso utilizados em conjunto, podem influenciar no desempenho dos modelos de classificação. Para verificar possíveis correlações foi empregada uma função estatística da biblioteca *Scipy*[5], que confirmou tais correlações. Neste cenário, foi optado pela retirada apenas da variável peso do conjunto de dados.

Há um desbalanceamento na quantidade de registros de presença ou não da doença em pacientes gestantes quando agrupados por região, conforme apresentado na Figura 2. Acredita-se que tal cenário é frequente quando o foco de estudos está relacionado a um evento incomum, como a detecção de uma doença específica dentro de uma população e onde a classe positiva (que tem a doença) acaba sendo muito superada em números pela classe negativa. Além disso, conforme mostra a Figura 2, a região Norte não possui nenhum registro que contenha a classe que marca a presença da doença. Portanto, essa região foi removida do estudo. Por fim, a Tabela 3 apresenta os atributos utilizados neste trabalho e descreve seus tipos de dados.

5. Métodos de Classificação

Os métodos de aprendizado de máquina são classificados em duas abordagens: supervisionados ou não-supervisionados. A primeira trata do aprendizado utilizando rótulos pré-existentes das variáveis dependentes (uma classe, no caso de classificação).

[4]https://www.ibge.gov.br/geociencias/cartas-e-mapas/mapas-regionais/10861-mapas-regionais.html?=&t=sobre

[5]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html

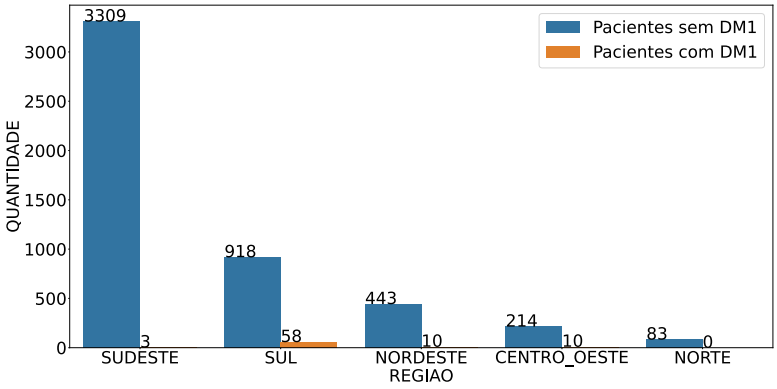


Figura 2. Número de gestantes com e sem diabetes mellitus do tipo 1 por região.

A segunda abordagem, é o aprendizado onde não são apresentados os rótulos para as instâncias. Neste estudo, é utilizada a abordagem supervisionada, em especial, de classificação binária. Aqui há a presença da variável dependente que marca a presença ou não da DM1 nas pacientes. No restante desta seção, é apresentada uma breve descrição dos métodos de classificação utilizados neste trabalho, estratégias para tratar o desbalanceamento (como foi visto, há menos instâncias de pacientes com DM1), e o procedimento para treinar e avaliar os modelos que foi adotado.

Inicialmente, destacamos o *Naive Bayes*, que é um método probabilístico baseado no teorema de Bayes. Ele calcula a probabilidade de um evento a partir de um conhecimento *a priori* de fatores observados nos dados e associados à ocorrência do evento.

Outro método muito utilizado em problemas de classificação, que também foi utilizado neste trabalho, é a regressão logística. Esse método adapta técnicas de regressão linear para medir a relação entre a variável dependente, geralmente binária, e outras variáveis independentes, usando uma função logística para estimar as probabilidades.

Neste trabalho, também foram empregadas árvores de decisão. Elas apresentam uma maneira simples de visualizar um modelo de tomada de decisão por meio de uma estrutura em árvore. Sua composição básica consiste de nó raiz, nós de decisão e nós terminais, também chamados de folhas. As decisões se bifurcam nessa estrutura até que uma decisão de previsão seja feita para um determinado registro. Árvores de decisão podem ser utilizadas tanto com dados categóricos, quanto numéricos.

O *K-Nearest Neighbor* (KNN), é um método que utiliza um critério de similaridade, que é definido usualmente através de uma medida de distância no espaço multidimensional das observações. O tipo de medida de distância pode variar, sendo a mais usual a distância Euclidiana. No caso de uma nova instância, as previsões são feitas pesquisando todo o conjunto de treinamento para as K instâncias mais próximas. Em problemas de classificação, dada uma nova instância, são recuperados os k vizinhos mais próximos e é atribuída, a nova instância, a classe mais frequente entre esses k vizinhos.

Também foi utilizado o *Support Vector Machine* (SVM), que busca pelo hiperplano de separação definido por uma função kernel que maximiza sua margem em relação aos elementos de classes distintas. Adotou-se aqui o kernel de função de base radial

(RBF). Por último, foi utilizada a *Random Forest*, que é considerada um comitê. Ela cria várias árvores de decisão usando aleatoriedade na amostragem dos dados de treinamento e nos atributos que podem ser explorados. A floresta escolhe a classe prevista combinando as respostas das suas árvores de decisão.

Para todos os métodos, foi utilizada a estratégia de validação cruzada estratificada (Stratified K-Fold)[6] com k=10. A validação cruzada é um procedimento para treinar e avaliar a capacidade de generalização de um modelo. Nela, os dados são divididos em k subconjuntos (folds) e, a cada iteração, um desses conjuntos é usado para testar os modelos gerados usando os demais dados no treinamento. A variante estratificada mantém a proporção original das classes nos conjuntos (*folds*).

Os resultados foram determinados como a média de 20 execuções e as seguintes métricas foram usadas: precisão, *recall* e *f1-score*. *Recall* que consiste na proporção de previsões corretas da classe dependente (verdadeiros positivos) em relação à soma dos verdadeiros positivos com os casos de falsos negativos. Ela foi usada pois é plausível considerar que a presença de falsos negativos seja mais prejudicial que os falsos positivos em um cenário de predição de doenças. A precisão é uma métrica que identifica as proporções de previsões corretas da classe dependente (verdadeiros positivos), em relação aos casos previstos pelo modelo como sendo da classe positiva (doença). Por fim, o *f1-score* consiste da média harmônica entre a precisão e o *recall*. É utilizada para se ter um resumo da qualidade do modelo em aspectos de *trade-off* entre essas duas métricas.

Dado o alto desbalanceamento das classes, utilizamos o *Synthetic Minority Oversampling Technique* (SMOTE)[7] como técnica de sobreamostragem (*oversampling*), que consiste na criação de instâncias sintéticas da classe minoritária, e o *NearMiss*[8] para subamostragem (*undersampling*), que remove instâncias da classe majoritária para realizar o balanceamento. A biblioteca *Imbalanced-learn*[9] foi usada para ambos os casos.

6. Resultados e Discussão

Na literatura, são encontradas diversas métricas para avaliação de modelos de classificação. Dentre as principais e mais utilizadas está a acurácia, que tem como foco a avaliação do desempenho geral de um modelo. Porém, é uma métrica que para conjuntos desbalanceados não se mostra muito viável, tendo em vista a tendência de classificar sempre, com um bom desempenho, acertos da classe majoritária, ou seja, os registros de pacientes que não tem a doença. Portanto, reforçamos a opção pela utilização das métricas: precisão, *recall* (sensibilidade) e *f1-score* (média harmônica entre os dois), que são métricas mais adequadas para este tipo de cenário.

A Tabela 4 apresenta os resultados médios e os desvios padrões para cada métrica utilizada. Conforme observado, os melhores resultados em geral foram obtidos quando a técnica de sobreamostragem é adotada. Dentro desse escopo, os métodos *Random Forest* e Árvore de Decisão foram os que mais se destacaram quando avaliados via as métricas precisão e *f1-score*. Por outro lado, o melhor valor de *recall* foi obtido usando o SVM.

Observando os outros cenários, pode-se destacar a Regressão Logística com a

[6]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
[7]https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html
[8]https://imbalanced-learn.org/dev/references/generated/imblearn.under_sampling.NearMiss.html
[9]https://imbalanced-learn.org/stable/

Tabela 4. Resultados dos classificadores

		Sem Balanceamento		Oversampling		Undersampling	
		Média	**DP**	**Média**	**DP**	**Média**	**DP**
Naive Bayes	**Recall**	0.8007	0.0353	0.9406	0.0011	0.3128	0.015
	Precisão	0.0411	0.0019	0.7676	0.0006	0.97	0.0458
	F1	0.0777	0.0036	0.8452	0.0006	0.4543	0.0158
Árvores de Decisão	**Recall**	0.0677	0.024	0.9637	0.0015	0.5526	0.0279
	Precisão	0.0811	0.0339	0.9751	0.0016	0.6933	0.031
	F1	0.0659	0.0236	0.9695	0.0012	0.5909	0.0274
Random Forest	**Recall**	0.0113	0.0078	0.975	0.0007	0.6199	0.0199
	Precisão	0.0341	0.0282	0.976	0.0009	0.6796	0.0259
	F1	0.0173	0.011	0.9756	0.0007	0.624	0.0162
KNN	**Recall**	0.0333	0.0132	0.9746	0.0012	0.5616	0.0229
	Precisão	0.1206	0.0529	0.9299	0.0008	0.7584	0.0293
	F1	0.0473	0.0166	0.9517	0.0007	0.62	0.0185
Regressão Logística	**Recall**	0.9203	0.0192	0.9498	0.0005	0.5983	0.0243
	Precisão	0.0491	0.0008	0.7981	0.0003	0.7653	0.0211
	F1	0.0928	0.0014	0.8673	0.0003	0.6519	0.0202
SVM	**Recall**	0.7452	0.0204	0.9856	0.0001	0.6876	0.0245
	Precisão	0.0676	0.0015	0.8704	0.0004	0.7777	0.0135
	F1	0.1231	0.0026	0.9244	0.0002	0.7203	0.0245

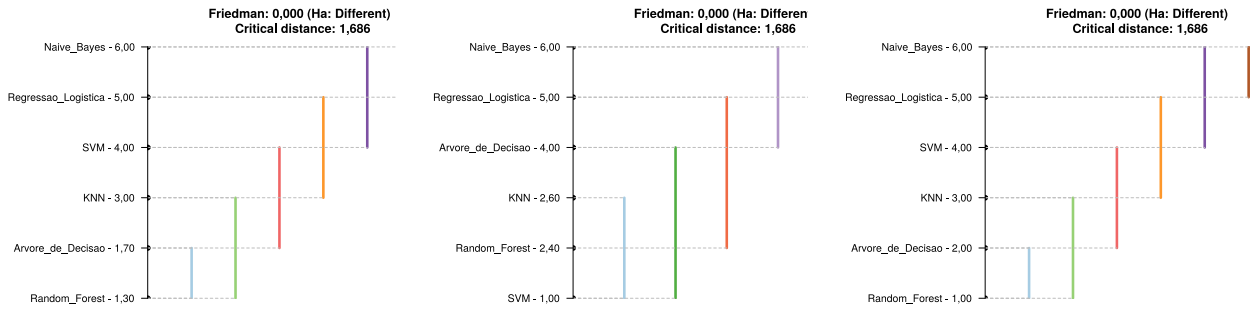
métrica *recall* no conjunto de dados que não tem aplicação de método de balanceamento. Porém, além de um desvio padrão alto, suas demais métricas não demonstram relevância para aplicação no problema em questão. Caso semelhante acontece com o *Naive Bayes* com a métrica precisão e utilizando subamostragem. Nota-se também um desvio padrão alto, revelando uma instabilidade para diferentes conjuntos de treinamento e teste.

Diante dos resultados e escopo de avaliação dos algoritmos utilizados no conjunto de dados com sobreamostragem, foi conduzido o teste estatístico de Friedman e o teste *post-hoc* de Nemenyi[10] para investigar a diferença estatística entre os algoritmos propostos. Os testes foram executados com um nível de confiança de 95% para os resultados das três métricas. Conforme observado na Figura 3, o teste de Friedman resultou em zero, mostrando a existência de diferença entre algum dos resultados. Em relação ao teste de Nemenyi, nota-se que a distância crítica foi de 1,686, indicando que as distâncias entre os resultados dos algoritmos devem ser maiores do que este valor para as diferenças serem significativas. Podemos então considerar que não existe diferença estatística entre os resultados das métricas precisão e *f1-score* da *Random Forest* e da Árvores de decisão, ou seja, os dois métodos possuem desempenho similar considerando essas métricas. Além disso, é possível considerar que os dois são melhores em relação a essas duas métricas quando comparados aos demais algoritmos. Já em relação à métrica *recall*, os algoritmos SVM, *Random Forest* e KNN possuem resultados semelhantes, com uma distância crítica inferior a 1,686. Assim, não há diferença significativa entre eles. Ademais, é possível considerar que os três são superiores no quesito *recall* em relação aos demais algoritmos.

7. Considerações finais

A diabetes mellitus tipo 1 é uma doença severa e pode trazer grandes riscos à saúde da gestante e do bebê. Nesse sentido, este trabalho buscou gerar modelos de predição

[10]https://www.rdocumentation.org/packages/tsutils/versions/0.9.2/topics/nemenyi



(a) Precisão **(b) Recall** **(c) f1-score**

Figura 3. Resultado dos testes de Friedman e Nemenyi.

utilizando aprendizado de máquina para identificar a presença de diabetes tipo 1 conforme as características e o histórico de acompanhamento da paciente gestante. Os dados foram obtidos do Sistema de Informações Ambulatoriais do SUS.

Dentre os métodos avaliados aqui, pode-se destacar o modelo de *random forest* com o desbalanceamento dos dados tratado com *oversampling*, que foi capaz de atingir um *f1-score* de 97,56%. Entendemos como limitação do trabalho, a necessidade de treinar e validar os modelos propostos aplicando também técnicas de seleção de atributos e ajuste dos hiperparâmetros dos modelos.

Trabalhos futuros incluem o uso de novas fontes de dados contendo dados clínicos, como exames de glicemia e sintomas coletados na etapa de anamnese. Pretendemos também estender o estudo usando outros métodos de aprendizado de máquina, como redes neurais artificiais, e outros comitês de classificadores, como os baseados em *boosting*.

Agradecimentos

Os autores agradecem o suporte da FAPEMIG, CNPq, CAPES e UFJF.

Referências

Acharjee, S., Ghosh, B., Al-Dhubiab, B., and Nair, A. (2013). Understanding type 1 diabetes: etiology and models. *Canadian Journal of Diabetes*, 37(4):269–276.

Albuquerque, J. P. D., Vasques, E. P., and Machado, G. R. (2011). Ambivalent implications of health care information systems: A study in the Brazilian public health care system. *Revista de Administração de Empresas*, 51.

Beeck, A. and Eizirik, D. (2016). Viral infections in type 1 diabetes mellitus - why the β cells? *Nature Reviews Endocrinology*, 12:263–273.

Cavalcante, G. L., Moura, M. C. L., Lago, A. F. V., and de Oliveira Lima, J. V. (2020). Perfil farmacoepidemiológico de pacientes com diabetes mellitus tipo 1. *Research, Society and Development*, 9:184953361.

Chaves, L. and Marques, G. (2021). Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study. *Applied Sciences*, 11(5).

Di Nubila, H. B. V. and Buchalla, C. M. (2008). O papel das Classificações da OMS-CID e CIF nas definições de deficiência e incapacidade. *Revista Brasileira de Epidemiologia*, 11(2):324–335.

Feitosa, A. C. R. and Ávila, A. N. (2016). Uso do prontuário eletrônico na assistência pré-natal às portadoras de diabetes na gestação. *Revista Brasileira de Ginecologia e Obstetrícia*, 38:9 – 19.

IDF (2015). IDF Diabetes Atlas. 7th ed. Brussels: International Diabetes Federation. http://www.diabetesatlas.org, Acessado em: 24/09/2020.

IDF (2020). What is diabetes. International Diabetes Federation.

Javad, M. O. M., Agboola, S. O., Jethwani, K., Zeid, A., and Kamarthi, S. (2019). A Reinforcement Learning-Based Method for Management of Type 1 Diabetes: Exploratory Study. *JMIR Diabetes*.

Mainenti, G., Campanile, L., Marulli, F., Ricciardi, C., and Valente, A. S. (2020). Machine learning approaches for diabetes classification: Perspectives to artificial intelligence methods updating. In *5th International Conference on Internet of Things, Big Data and Security (IoTBDS 2020)*, pages 533–540.

MGH (2020). Massachusetts General Hospital. Disponível em: https://www.massgeneral.org. Acessado em: 04/09/2020.

Ministério da Saúde (2006). Caderno de atenção básica. diabetes mellitus. n.16. série a. normas e manuais técnicos - brasília.

Ministério da Saúde (2012). Departamento de Atenção Básica. Atenção ao Pré-Natal de Baixo Risco. Normas e Manuais Técnicos, Cadernos de Atenção Básica, n° 32. Brasília, DF.

Moulin, C. M., Portella, R. B., Pinheiro, V. S., Oliveira, M. M., Fuks, A. G., Cunha, E. F., and Gomes, M. B. (2003). Prevalência de sobrepeso e obesidade em pacientes com diabetes tipo 1. *Arquivos Brasileiros de Endocrinologia & Metabologia*, 47.

OPAS (2016). Organização Pan-Americana da Saúde. Ministério da Saúde. Federação Brasileira das Associações de Ginecologia e Obstetrícia. Sociedade Brasileira de Diabetes. Rastreamento e diagnóstico de diabetes mellitus gestacional no Brasil. Brasília.

Rodríguez-Rodríguez, I., Rodríguez, J.-V., Woo, W. L., Wei, B., and Pardo-Quiles, D.-J. (2021). A Comparison of Feature Selection and Forecasting Machine Learning Algorithms for Predicting Glycaemia in Type 1 Diabetes Mellitus. *Applied Sciences*, 11(4).

Sociedade Brasileira de Diabetes (2019). Diretrizes Brasileiras de Diabetes 2019-2020.

Taser, P. Y. (2021). Application of bagging and boosting approaches using decision tree-based algorithms in diabetes risk prediction. *Proceedings*, 74(1).

UCI (2020). UC Irvine Machine Learning Repository . Disponível em: https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

WHO (2019). International Statistical Classification of Diseases and Related Health Problems 10th Revision. World Health Organization.

WHO (2020a). Ficha Técnica Diabetes. World Health Organization. www.who.int/news-room/fact-sheets/detail/diabetes, Acessado em: 25/03/2021.

WHO (2020b). Int. Statistical Classification of Diseases and Related Health Problems (ICD). https://www.who.int/standards/classifications/classification-of-diseases. Acessado em: 24/09/2020.