Análise e predição de diabetes com Random Forest

Pedro Evaristo Dantas – 1282116206

Hugo Da Costa Gomes – 128217256

Vítor Dantas Costa – 1282326249

João Davi Fernandes – 1282321236

Lucas Martins Santos – 1282016408

Resumo

A diabetes é uma condição crônica que afeta muitas pessoas em todo o mundo, tornando-se uma preocupação global de saúde. A aplicação de modelos de inteligência artificial como o Random Forest tem se mostrado promissora para a predição e monitoramento da doença. Neste estudo, propomos a implementação de dois modelos Random Forest para realizar a predição da diabetes utilizando o BRFSS dataset. O objetivo é analisar os dados e comparar a performance do Random Forest de acordo com a quantidade de árvores de decisão utilizadas. Obtivemos no primeiro modelo 85,80% de acurácia, enquanto o segundo modelo obteve 84,03% de acurácia. Os métodos propostos obtiveram resultados expressivos, porém, o primeiro modelo se mostrou mais eficaz na predição de diabetes dentro do dataset abordado.

Palavras-chaves: Diabetes. Random Forest. Predição.

1 Introdução

A Diabetes é uma condição crônica que afeta a forma como o corpo regula o nível de açúcar (glicose) no sangue. De acordo com Gautam, Bhatta e Aryal (2015), aproximadamente 285 milhões de pessoas no mundo possuíam diabetes em 2010. Este número tende a aumentar, impulsionado pelo envelhecimento da população e pelo aumento do sedentarismo. Além disso, os autores destacam a existência de características regionais e nacionais que influenciam a prevalência da diabetes na população.

Essas características distintas entre regiões e costumes próprios abrem espaço para a aplicação de modelos de inteligência artificial (IA) voltados para a predição da diabetes. A utilização de modelos de IA preditivos na predição de diabetes mostra-se uma abordagem promissora, pois são capazes de auxiliar os médicos realizando previsões futuras de possíveis casos de diabetes ou pré-diabetes. Isso pode levar a uma redução potencial do número de diabetes em uma escala regional ou mundial.

Nesse cenário, a aplicação do Random Forest (Floresta Aleatória) tem se destacado, pois este é um modelo capaz de lidar com conjuntos de dados complexos e diversificados. O funcionamento do modelo Random Forest se dá pela construção de múltiplas árvores de decisão durante o treinamento do modelo, que utilizam subconjuntos aleatórios dos dados de treinamento e suas características (BREIMAN, 2001). Dessa forma, o modelo Random Forest tornou-se uma ferramenta valiosa para a predição de diabetes, especialmente quando há variações significativas nos dados.

Diante disto, em nosso estudo, propomos a implementação de dois modelos Random Forest para realizar a predição da diabetes no Behavioral Risk Factor Surveillance System (BRFSS) dataset, que possui mais de 70000 amostras coletadas no ano de 2015.

O primeiro modelo implementado contém 100 (cem) árvores de decisões, enquanto o segundo modelo possui apenas 5 (cinco). Dessa forma, o objetivo do nosso estudo é levantar a análise dos dados do BRFSS dataset e comparar a performance do Random Forest de acordo com sua quantidade de árvores de decisão utilizadas na construção do modelo. A arquitetura dos dois modelos criados foi inspirada no modelo original proposto por Breiman (2001). A construção do modelo foi realizada utilizando a linguagem de programação Python e a biblioteca scikit-learning.

2 Referencial Teórico

Os estudos foram realizados pelos autores VijiyaKumar et al. (2019), Xu et al. (2017) e Benbelkacem e Atmani (2019) foram de extrema importância para o desenvolvimento de modelos de IA voltados para a predição de diabetes. Esses autores utilizaram o algoritmo original Random Forest, proposto por Breiman (2001), para criar seus modelos. Suas pesquisas destacam a eficácia desse modelo em cenários de classificação e regressão.

No entanto, outros autores, como Soni e Varma (2020), começaram a explorar a aplicação de diferentes modelos de IA para atingir o mesmo objetivo. Embora as abordagens Adaboost (XU et al., 2017) e Linear Rate (LR) (SONI; VARMA, 2020) demonstrem eficácia na predição de diabetes, a literatura reforça que o modelo Random Forest continua sendo o mais eficaz.

Além disso, Probst e Boulesteix (2017) conduziram um estudo para examinar a relação entre a quantidade de árvores usadas na construção do modelo Random Forest e seu desempenho. Essas discussões adicionais são de grande relevância para aprimorar a aplicação desse modelo, como para as suas futuras implementações, como abordados no nosso estudo.

Embora a literatura aponte uma maior eficácia para o modelo Random Forest atualmente, é importante destacar que a escolha do melhor modelo de IA depende do contexto particular de cada análise. Dessa forma, torna-se crucial a aplicação do modelo em novas populações e condições clínicas distintas para a predição de diabetes.

Em suma, as pesquisas em andamento nessa área são cruciais para avançar no desenvolvimento de modelos de IA precisos e confiáveis para a detecção e prevenção de doenças crônicas, como diabetes, o que pode resultar em melhores resultados de saúde para os pacientes e uma prática médica mais eficiente, como apontado por Kaul e Kumar (2020).

3 Metodologia

3.1 Dataset

Neste estudo utilizamos um dataset derivado do Behavioral Risk Factor Surveillance System (BRFSS) de 2015. O BRFSS possui uma visão abrangente dos comportamentos relacionados à saúde nos Estados Unidos, e é composto por cerca de 70000 amostras. A versão do BRFSS utilizada neste estudo já apresenta um pré-processamento inicial, realizado a partir da limpeza dos dados. Dessa forma, o BRFSS utilizado apresenta 22 colunas com as principais variáveis relacionadas aos fatores provenientes a causa do diabetes

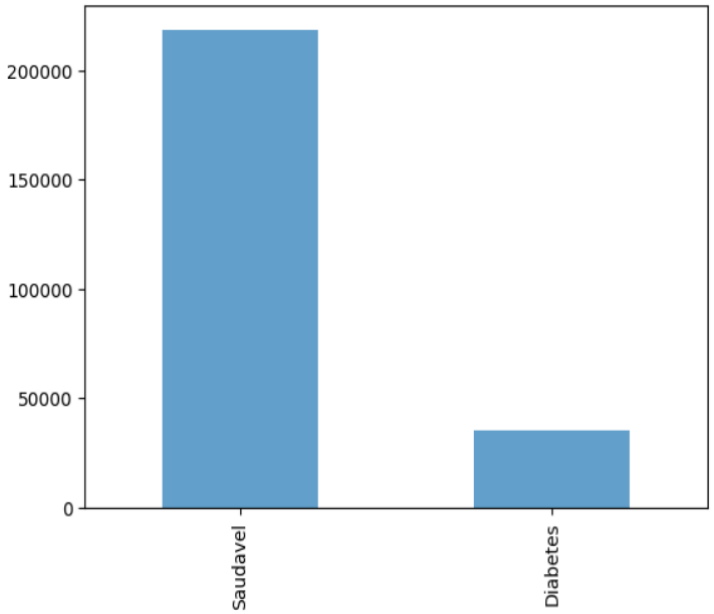

Figura 1 - Relação entre os pacientes saudáveis e pacientes diabéticos e pré-diabéticos.

A versão alterada do dataset se encontra em: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>.

O dataset apresenta três potenciais csvs a serem utilizados e analisados, dois apresentam características binárias, ou seja, apresentam apenas duas classes, denominadas: pacientes com diabetes ou pré-diabetes e pacientes sem diabetes. Já o terceiro csv, possui os dados divididos em três classes, respectivamente: diabetes, pré-diabetes e saúde.

Neste estudo foi utilizado apenas os dados binários, pois eles correspondem a uma distribuição equilibrada entre as duas classes correspondentes de pacientes, conforme ilustra a Figura 1.

3.2 Análise dos Dados

O BRFSS já possui um pré-processamento em seus dados, porém, realizamos uma verificação inicial para analisar a necessidade da aplicação de novas técnicas de pré-processamento nos dados para que tornasse possível a construção dos modelos de IA abordados neste estudo.

Inicialmente verificamos o tipo de dados e a quantidade total de linhas e colunas presentes no dataset. Constatou-se que o BRFSS possui 253680 linhas e 21 colunas, todas as suas variáveis são do tipo float64. Além disso, analisamos as colunas, a presença de elementos nulos (NaN) dentro do dataset e a estatística descritiva de suas variáveis para analisarmos os padrões de comportamento dos dados.

O processo de análise dos dados foi feito com o levantamento da quantidade total de dados referentes a pacientes com diabetes ou pré-diabetes e os considerados pacientes saudáveis. Adotamos uma abordagem para examinar os dados através do perfil de gênero dos pacientes, fornecendo insights sobre a distribuição de casos de diabetes entre homens e mulheres. A análise incluiu a determinação da porcentagem de mulheres afetadas por diabetes, categorização da faixa etária em relação a recorrência de diabetes e a porção de casos de diabetes entre homens e mulheres.

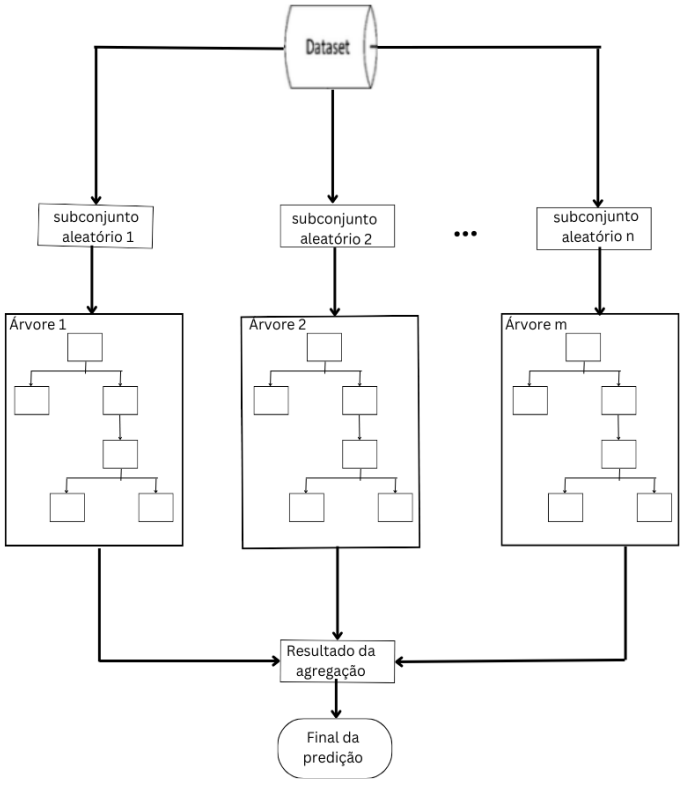

Figura 2 – Representação simplificada do modelo Random Forest.

3.3 Modelo Random Forest

A Figura 2 ilustra a arquitetura do modelo Random Forest empregada neste estudo. Que consiste na aplicação padrão da rede abordado por Breiman (2001).

Para analisarmos o desempenho da arquitetura Random Forest, criamos dois modelos utilizando a arquitetura scikit-learning do Python, conforme a sua documentação [1]

No primeiro modelo, empregamos o classificador padrão do scikit-learn, configurado com 100 (cem) árvores de decisão. No segundo modelo, optamos por utilizar o classificador com apenas 5 (cinco) árvores. Ambos os modelos foram concebidos com o propósito final de prever a incidência de diabetes

4 Resultados e Discussão

O mapa de calor do BRFSS dataset, conforme ilustrado na Figura 3, apresenta a inter-relação entre todas as variáveis, destacando a distribuição dos valores ao longo do dataset. Isso nos permitiu identificar as áreas de concentração e os padrões

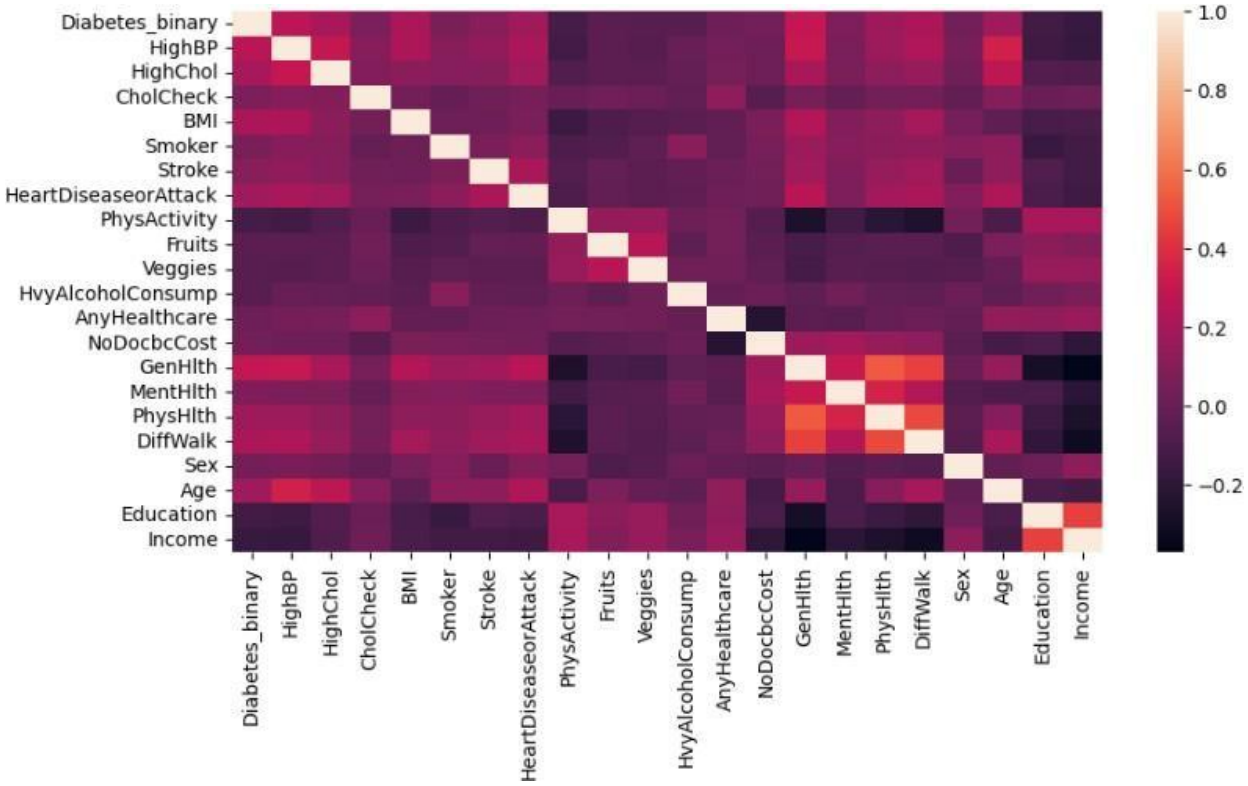

Figura 3 – Mapa de calor do BRFSS dataset.

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

Ao analisarmos as variações das cores, observamos que o diabetes e o pré-diabetes apresentam uma predominância significativa em pacientes que apresentam problemas relacionados à saúde em geral (GenHlth), saúde mental (MentHlth), saúde física (PhysHlt) e dificuldade de caminhar (DiffWalk).

Ao explorarmos os dados na pespectiva da abordagem de perfil de gênero, identificamos uma prevalência da diabetes e pré-diabetes em pacientes do sexo masculindo, representando 52,1% dos casos positivos registrados no dataset, conforme mostra a Figura 4 . Em contraste, na população feminina, apenas 13,9% apresentaram diabetes ou pré-diabetes, como ilustra a Figura 5.

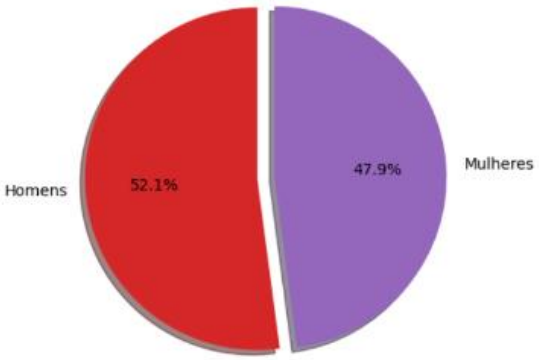

Figura 4 – Relação do diabetes com base no perfil de gênero

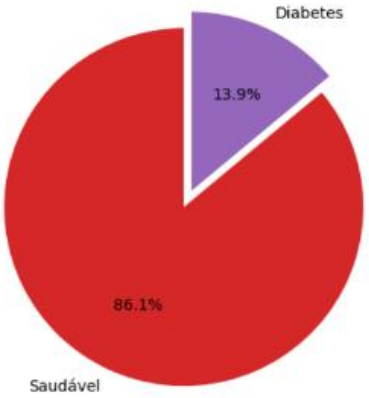

Figura 5 – Casos mulheres com diabetes e saudáveis

Além disso, notamos que a faixa etária com a maior prevalência de casos de diabetes e pré-diabetes dentro do BRFSS foi 11 anos (Figura 6).

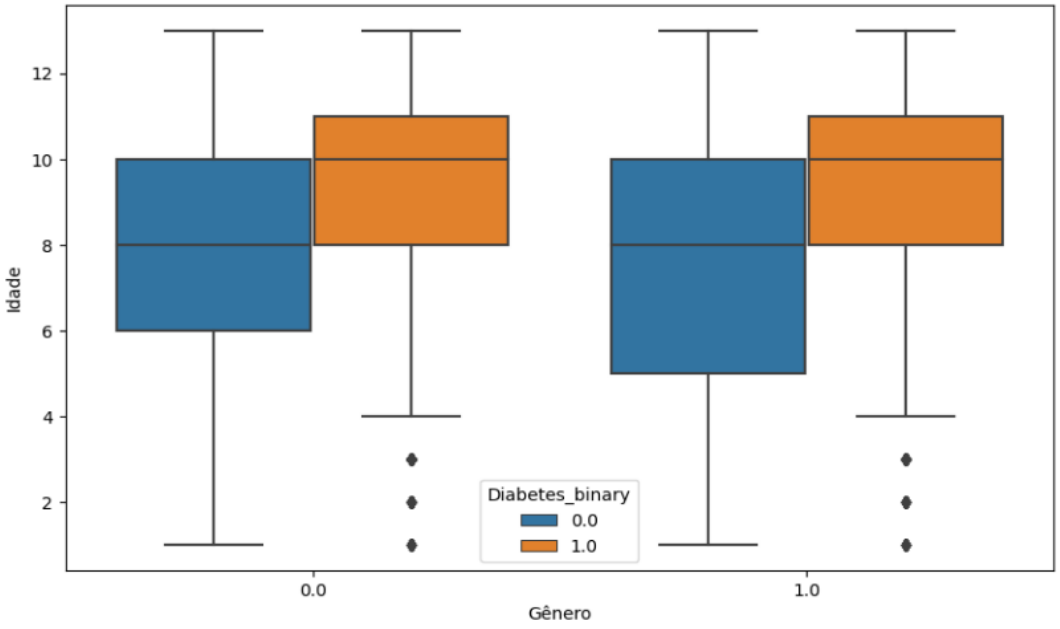

Figura 6 – Relação entre o gênero e a idade dos casos de diabetes e pré-diabetes

Ambos os modelos Random Forest foram avaliados utilizando as métricas de acurácia e relatório de classificação. O relatório de classificação abrange métricas essenciais como precisão, recall, F1 score e suporte. Todas essas métricas são calculadas através da relação dos rótulos reais e as previsões geradas pelo modelo implementado.

Torna-se importante salientar que as métricas foram obtidas por meio do cálculo padrão da biblioteca scikit-learn. As Equações abaixo resumem os métodos utilizados pela biblioteca para cada uma das métricas.

$$Acurácia = \frac{PrevisõesCorretas}{Previsões\ total}$$

$$Precisão = \frac{VerdadeirosPositivos}{VerdadieosPositivos + FalsosPositivos}$$

$$Recall = \frac{VerdadeirosPositivos}{VerdadieosPositivos\ +\ FalsosPositivossl}$$

$$F1\ Score = \frac{Precisão\ X\ Recall}{Precisão + Recall}$$

O desempenho do primeiro modelo, utilizando 100 (cem) árvores, superou o do segundo modelo, com apenas 5 (cinco) árvores. O primeiro modelo atingiu 85,80% de acurácia, enquanto o segundo modelo obteve apenas 84,03% de acurácia.

Conforme apresentado na Tabela 1, observa-se que o desempenho do segundo modelo torna-se superior na métrica Precisão para os casos negativos de diabetes ou pré-diabetes. Além disso, também é possível notar que o segundo modelo ultrapassa o primeiro no F1 score para casos confirmados de diabetes e pré-diabetes.

Tabela 1 –

Classe	Modelo	Acurácia	Precisão	Recall	F1 Score
Total	Modelo 100 árvores	0.8580	–	–	–
	Modelo 5 árvores	0.8403	–	–	–
Saudável	Modelo 100 árvores	–	0.88	0.97	0.92
	Modelo 5 árvores	–	0.88	0.94	0.91
Diabetes	Modelo 100 árvores	–	0.50	0.57	0.26
	Modelo 5 árvores	–	0.40	0.24	0.30

Métricas obtidas através dos modelos estudados.

Os modelos apresentam uma disparidade de acurácia entre si de apenas 1,77%. Além disso, ambos utilizam o mesmo critério de importância de variáveis, sendo o índice de massa corporal (BMI) considerado como variável mais importante, conforme a Figura 7.

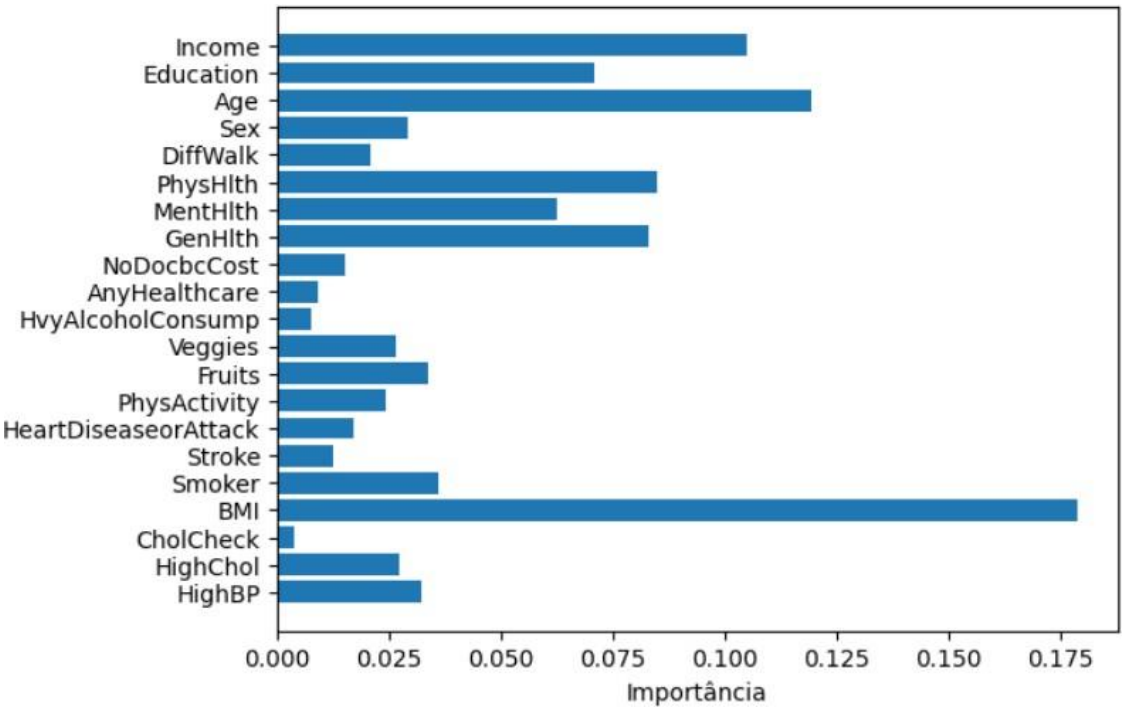

Figura 7 – Critério de importância de variáveis dos modelos

Portanto, ambos os modelos possuem viabilidade para a predição de diabetes, sendo o primeiro modelo mais indicado por ser mais robusto e eficaz, corroborando para a questão levantada por Probst e Boulesteix (2017), que afirma que quanto maior o número de árvores de decisão, melhor o modelo se comporta diante dos dados.

Entretanto, o primeiro modelo necessita de um poder computacional maior e possui uma taxa de desempenho relativamente pequena comparada ao segundo modelo. Desse modo, deve-se levar em consideração a aplicação do modelo, pois em modelos embarcados mais simples, a aplicação do modelo com apenas 5 (cinco) árvores de decisão torna-se mais interessante, devido a baixa necessidade computacional.

5 Conclusão

Neste estudo, desenvolvemos dois modelos de Random Forest para prever casos de diabetes. Os modelos utilizaram como forma de validação as métricas de acurácia, F1 score, precisão e recall. Além disso, utilizamos como parâmetro de comparação a quantidade de árvores de decisão presentes em cada um dos modelos.

O primeiro modelo empregou 100 (cem) árvores em sua construção, enquanto a segunda utilizou apenas 5 (cinco) árvores. Os modelos propostos exibiram uma discrepância de acurácia de 1,77% entre si, favorecendo o desempenho do primeiro modelo que obteve 85,80% de acurácia. Contudo, ao analisar as outras métricas abordadas, observamos que o segundo modelo obteve um desempenho superior em relação às métricas F1 score em casos confirmados de diabetes ou pré-diabetes, e na precisão em casos de pacientes saudáveis.

Apresentando uma taxa de desempenho superior maior, o primeiro modelo se mostrou ideal para melhorias e otimizações em trabalhos futuros. Porém, deve-se ser levado em consideração o tipo de aplicação desejada nos trabalhos futuros, uma vez que o segundo modelo possui uma baixa necessidade computacional e um desempenho similar ao primeiro modelo.

Além disso, pretende-se abordar em trabalhos futuros o csv que foi dividido em três classes distintas, sendo elas respectivamente, pacientes saudáveis, pacientes com diabetes e pacientes pré-diabetes. Neste caso, pretendemos explorar a eficiência do modelo Random Forest com os dados brutos (desbalanceados) e aplicando técnicas de balanceamento dos dados.

Referências

BENBELKACEM, S.; ATMANI, B. Random forests for diabetes diagnosis. In: *2019 International Conference on Computer and Information Sciences (ICCIS)*. 2019. p. 1–4.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, p. 5–32, 2001.

GAUTAM, A.; BHATTA, D. N.; ARYAL, U. R. Diabetes related health knowledge, attitude and practice among diabetic patients in nepal. *BMC endocrine disorders*, BioMed Central, v. 15, n. 1, p. 1–8, 2015.

KAUL, S.; KUMAR, Y. Artificial intelligence-based learning techniques for diabetes prediction: challenges and systematic review. *SN Computer Science*, Springer, v. 1, n. 6, p. 322, 2020.

PROBST, P.; BOULESTEIX, A.-L. *To tune or not to tune the number of trees in random forest?* 2017.

SONI, M.; VARMA, S. Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (Ijert) Volume*, v. 9, 2020.

VIJIYAKUMAR, K. et al. Random forest algorithm for the prediction of diabetes. In: IEEE. *2019 IEEE international conference on system, computation, automation and networking (ICSCAN)*. 2019. p. 1–5.

XU, W. et al. Risk prediction of type ii diabetes based on random forest model. In: *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*. 2017. p. 382–386.