ANÁLISE COMPARATIVA DE MODELOS DE APRENDIZADO DE MÁQUINA NA DETECÇÃO DE DIABETES EM ESTÁGIO INICIAL

Lucas Gentil Carlos[1]
Luan Ornelas de Souza[1]
José Guilherme Picolo[2]
Universidade São Francisco
luan.ornelas@mail.usf.edu.br

[1]Aluno do Curso de Engenharia de Computação, Universidade São Francisco; Campus Campinas
[2]Professor Orientador José Guilherme Picolo, Curso de Engenharia de Computação, Universidade São Francisco; Campus Campinas.

Resumo. Diabete Mellitus é uma doença que atinge a humanidade de maneira global sendo uma das doenças com mais casos nos últimos anos. Em casos de diagnósticos tardios, juntamente com a falta do correto tratamento, a doença pode apresentar riscos à saúde devido às possíveis complicações micro e macro vasculares. Com o intuito de realizar o diagnóstico prévio da Diabetes, este estudo utiliza algoritmos de *Machine Learning,* a partir de um conjunto de dados com atributos de sintomas, como sede excessiva, fraqueza e retardo da cicatrização, evitando a necessidade da realização de exames específicos. Para isso, os modelos avaliados foram *Random Forest*, *XGBoost*, *Support Vector Machine* e *Naive Bayes*. Para análise dos resultados, métricas como *Precision, Recall, Accuracy, F1-Score e Balanced Accuracy* foram utilizadas, tendo o *Random Forest* como algoritmo mais eficiente em todas as métricas, atingindo excelentes valores como 99.3% de *F1-Score*.

Palavras-chave: *Machine Learning*, Diabetes, Classificadores, Predição.

Introdução

Atualmente a diabetes se tornou uma das doenças com mais casos ao redor do mundo. De acordo com a *International Diabetes Federation*, projetam-se que 643 milhões de pessoas serão diagnosticadas com diabetes em 2030 (INTERNATIONAL DIABETES FEDERATION, 2023). Cientificamente classificado como diabete mellitus, esta doença está diretamente relacionada com a intolerância à glicose, que pode estar atrelada com a falta de insulina no sangue ou na capacidade da insulina agir no corpo (INTERNATIONAL DIABETES FEDERATION, 2006). A consequência desse efeito é a inefetividade da utilização do açúcar pelas células do corpo humano, alcançando a hiperglicemia, caracterizada pela alta glicose no sangue. Com os altos níveis de açúcar no sangue esta doença pode promover grandes riscos à saúde, podendo gerar problemas micro e macro vasculares (GUYTON; HALL, 2011).

Contudo, além da diminuição da expectativa de vida de quem tem a doença, as complicações nos quadros de diabetes podem se estender a outros problemas, sendo a principal causa de problemas de visão nos países considerados desenvolvidos e está atrelada à alta quantidade de amputações de membros inferiores anualmente (INTERNATIONAL DIABETES FEDERATION, 2006) (BARAKAT, N.; BRADLEY; BARAKAT, M., 2010). Todavia, ao longo dos anos, estudos apontam que através de um bom gerenciamento dos casos de diabetes e seu diagnóstico prévio, podemos estabelecer quadros preventivos, evitando o aparecimento de sintomas agravados (MARINHO et al., 2013).

Diante do apresentado, o uso da tecnologia se apresenta como uma importante solução. Na medicina, algoritmos de *Machine Learning* (ML) são construídos usando grandes

conjuntos de dados médicos, com o intuito de encontrar padrões e realizar predições (BEAM; KOHANE, 2018). Logo, o projeto de pesquisa se propõe a realizar um estudo comparativo entre alguns modelos de ML, além de mostrar que eles podem ser utilizados com a finalidade de realizar diagnósticos prévios da diabetes. Para isso, foi utilizado um conjunto de dados disponível pela *University of California, Irvine (UCI) Machine Learning Repository*, contendo como atributos sintomas da doença. Esse tipo de base de dados foi escolhido levando em consideração a acessibilidade do uso do algoritmo, pois a coleta de dados não depende de exames médicos prévios, apenas os sintomas relatados pelo paciente são necessários, o que reduz custo e auxilia os médicos a terem um primeiro diagnóstico e um direcionamento de forma rápida dos próximos procedimentos a serem feitos.

Levantamento Bibliográfico

A inteligência artificial pode ser definida de diferentes modos, nos quais, atores diferentes descrevem e determinam o assunto como a capacidade de reproduzir o comportamento humano em máquinas, tornando possível atuarem e pensarem de maneira humana. Em outras abordagens a inteligência artificial é citada como a capacidade de máquinas agirem e pensarem racionalmente (RUSSEL, 2009). Poole, Goebel, Mackworth (1998) pontuam a inteligência computacional como a capacidade de realizar estudos de estruturação de máquinas inteligentes.

Com o avanço da tecnologia e a alta quantidade de dados gerados pela humanidade, o campo da inteligência artificial se difundiu em meio aos grandes avanços computacionais, dando espaços a novos campos como o *machine learning*. Através da utilização da alta gama de dados disponíveis foi notado que era possível automatizar as predições realizadas manualmente e utilizar informações obtidas através da internet. Atualmente o campo de *machine learning* encontra-se no cotidiano de grande parte da sociedade como por exemplo na utilização de anúncios em websites, recomendações de compras, pesquisas em sites de buscas entre outros serviços (RASCHKA, 2015).

Contudo para aplicar estratégias baseadas em dados é necessário fornecer às máquinas estes elementos com foco em realizar a atividade de maneira analítica semelhante aos seres humanos, utilizando uma sequência de dados para efetuar um levantamento estatístico, em que, busca encontrar padrões para realizar o *machine learning* com soluções efetivas e automatizadas. Através deste manuseio de informações são criados algoritmos com funções de prover a máquina uma visão racional, sendo possível realizar programações adaptativas devido às técnicas de *machine learning* existentes e com isto conseguir descobrir padrões e realizar processos de melhorias ao ingerir novas informações aprimorando o processamento de forma contínua. Os algoritmos ao longo dos anos expandiram-se e estabeleceram diferentes finalidades, sendo necessário separar em diferentes categorias, em que, normalmente são organizados em dois grandes grupos, algoritmos de aprendizado supervisionado e não supervisionado (NASTESKI, 2017).

Os métodos não supervisionados são extremamente úteis para análises em que o rótulo é desconhecido e não são passados os parâmetros de classe através dos algoritmos, atribuindo à máquina o objetivo de realizar tal rotulação. Desta forma, é necessário realizar procedimentos para descobrir as características semelhantes de classe, analisando os dados presentes para verificar similaridades, criando grupos distintos de elementos tornando possível a execução do algoritmo de maneira eficaz. Os grupos classificados igualmente são definidos como clusters, e através deles é possível obter casos de usos diferentes, como por exemplo, em análise de imagens médicas por meio dos clusters definidos pela máquina, os algoritmos conseguem utilizar das similaridades presentes nas imagens dos exames médicas para classificar uma pessoa com certa doença ou não (NASTESKI, 2017).

Por sua vez, os métodos supervisionados de *machine learning* são definidos como os algoritmos que necessitam de ação humana antecipadamente para determinarem as classes dos dados existentes, rotulando e classificando em grupos para prosseguir com a aplicação dos algoritmos. Com isto, a principal atribuição da máquina é buscar possíveis padrões e efetuar cálculos matemáticos (NASTESKI, 2017). Este tipo de aprendizagem é extremamente utilizado em problemas de classificação pois o foco é garantir que a máquina aprenda de acordo com as classes estabelecidas. Contudo, este método é considerado ideal para problemas que é possível determinar os rótulos de maneira fácil ou até mesmo classificar mediante ao raciocínio lógico (AYODELE, 2010). Outras questões devem ser levadas em consideração ao utilizar métodos de classificação, como em casos de a base de dados utilizada para treinar a máquina possuir valores que estão incompletos, valores que não condizem com a realidade e até mesmo características que não possuem relevância para realizar a classificação desejada. Para solucionar tais problemas podem ser realizadas técnicas de pré-tratamento dos dados, em que, através da utilização de softwares com objetivo de realizar a substituição de informações consideradas improváveis. Em casos que as devidas trocas não sejam viáveis, os dados necessitam ser removidos para não influenciarem na classificação. Valores ausentes na base dados acabam se tornando outro desafio frequente neste ramo, visto que, diversos fatores podem influenciar na falta de dados, como dados ignorados ou extraviados no momento de composição da base de dados. Para contornar a situação, pesquisadores utilizam diferentes métodos para completar estes campos manualmente (MAGLOGIANNIS et al., 2007).

Nesse contexto, muitas pesquisas na área da saúde que utilizam dados médicos disponíveis com foco na predição de doenças, como a diabetes, os quais utilizam diferentes algoritmos de *Machine Learning*. Por exemplo, Faruque et al. (2019), realizou um estudo comparativo para detecção da diabetes utilizando um conjunto de dados contendo atributos semelhantes a base de dados deste estudo como idade, gênero, sede excessiva, mas também atributos que precisam de algum dispositivo para ser coletado como pressão arterial. Para predição utilizou *Support Vector Machine* (SVM), *Naive Bayes* (NB), K-*Nearest Neighbors* (KNN) e *Decision Tree* (DT), tendo como resultados precisão e recall de 72% e 74%, respectivamente. Utilizando dados de 768 pacientes mulheres e nove atributos como concentração de glicose plasmática, massa corporal, idade, linhagem de diabetes e insulina, o autor Sarwar et al. (2018) realizou um estudo de predição de diabetes, comparando algoritmos e levando em consideração a acurácia, no qual, KNN e SVM destacaram-se atingindo 77% ambos. Focado para a predição de diabete com técnicas de ML Zou et al. (2018), utilizou dois conjuntos amplos de dados, sendo o segundo semelhante ao utilizado por Sarwar et al. (2018). Aplicando algoritmos como *Random Forest* (RF), J48, *DT* e *Neural Network (NN)*. Os resultados obtidos foram próximos entre si, sendo o RF o algoritmo com melhores resultados, atingindo a acurácia de 80.84% no primeiro conjunto e 77.21% no segundo conjunto. Ma (2020) utilizou o mesmo conjunto de dados deste estudo para predição da diabetes em estágio inicial e utilizou *Logistic Regression* (LR), SVM, DT, RF, *Boosting* e NN, chegando a 96.2% de acurácia nos dados de teste, sendo um resultado menor em relação ao que foi obtido neste trabalho.

No artigo apresentado por Tr et al. (2022), foram aplicadas técnicas de ML para predições de doenças cardíacas utilizando algoritmos como DT, NB, RF, KNN, SVM e LR. O conjunto de dados utilizado neste estudo foi composto por 303 pacientes e 76 atributos. Quando comparados, o algoritmo que obteve melhor performance para predição de doenças cardíacas foi o KNN, apresentando 90.8% para F1-Score. Com intuito de realizar predições utilizando técnicas de ML para a doença de Alzheimer, Khan et al. (2021), exerceu um estudo classificando os pacientes entre dementes e não dementes. Foram utilizados SVM, LR, DT e RF. Para avaliar os modelos utilizaram métricas de acurácia, *recall*, precisão, AUC (*Area under the curve*) e *F1-Score*. Contudo o SVM foi o algoritmo com melhor resultado de

maneira geral, apresentando 92% de acurácia e 91,9% para as demais métricas. Com foco na detecção de doenças no fígado, Rahman et al. (2019), realizou um estudo comparativo utilizando RF, DT, SVM, KNN, NB e LR com acurácia de 75%. R et al. (2019), comparou os modelos LR, DT, NB e chegou a 99.2% de acurácia utilizando SVM na detecção de câncer de pulmão.

Materiais e Métodos

Linguagem e ambiente de desenvolvimento

Neste estudo, o *Jupyter Notebook* foi o ambiente utilizado para realização dos experimentos, tendo o Python 3.9 como linguagem de programação.

Base de dados

Para este experimento foi utilizado um conjunto de dados público, disponível na *UCI Machine Learning Repository*. A base de dados contém 520 pacientes, sendo 63% homens e 37% mulheres, como apresentado pela Figura 1.

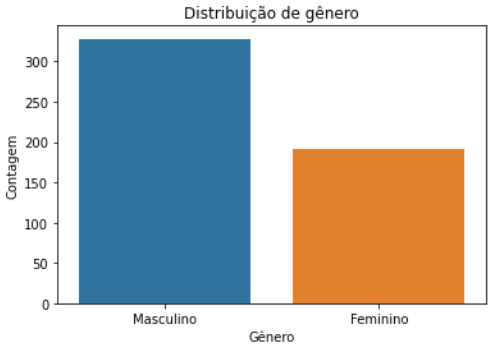

Figura 1 - Distribuição de pacientes por gênero. Fonte: Elaborada pelo autor.

Esse conjunto de dados tem 17 atributos, os quais todos são dados categóricos, isto é, contém valores qualitativos e não numéricos, com exceção da idade. Dezesseis deles estão divididos entre dados demográficos e sintomas do paciente e um atributo representa o diagnóstico positivo ou negativo para a diabetes. Dos 520 pacientes, 320 foram diagnosticados com a doença, ou seja, pertencem a classe positiva e 200 a classe negativa, conforme a Figura 2, sendo necessário realizar um balanceamento das classes antes de iniciar os experimentos.

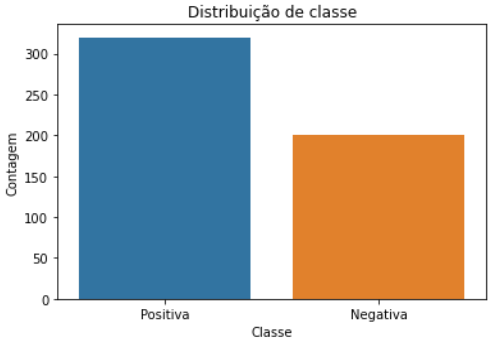

Figura 2 - Distribuição das classes. Fonte: Elaborada pelo autor.

Embora as nomenclaturas de alguns atributos sejam relativamente complexas, o significado é simples. Foi realizada a descrição das colunas na Tabela 1.

Nome da coluna	Descrição	Range de valores
Idade	Idade do paciente	16 - 90 anos
Gênero	Gênero biológico do paciente	Masculino ou Feminino
Poliúria	Necessidade excessiva de urinar	Sim ou Não
Polidipsia	Sede excessiva	Sim ou Não
Perda repentina de peso	-	Sim ou Não
Fraqueza	Sensação de cansaço constante	Sim ou Não
Polifagia	Fome excessiva	Sim ou Não
Candidíase genital	Infecção por fungos na genital	Sim ou Não
Visão borrada	-	Sim ou Não
Coceira	-	Sim ou Não
Irritabilidade	-	Sim ou Não
Cicatrização retardada	-	Sim ou Não
Paresia parcial	Grau leve a moderado de fraqueza muscular	Sim ou Não
Rigidez muscular	-	Sim ou Não
Alopecia	Queda de cabelos ou de pelos	Sim ou Não
Obesidade	Excesso de gordura corporal	Sim ou Não
Classe	Classificação positiva ou negativa para diabetes	Positiva ou Negativa

Tabela 1: Descrição do conjunto de dados. Fonte: Elaborada pelo autor.

Para a predição, todas as variáveis foram consideradas e análises exploratórias foram realizadas para entender a relação entre esses dados e as classes. Na Figura 3 é possível ver que as colunas Poliúria e Polidipsia têm a maior correlação com a classe a ser predita. Correlação é uma medida estatística que ajuda a entender como duas variáveis estão

relacionadas entre si, se essa relação é positiva ou negativa e quão forte ela é. Quanto mais próximo de 1 o valor da correlação for, maior será ela positivamente e quanto mais próximo de -1 o valor for, maior será a correlação, mas de forma negativa.

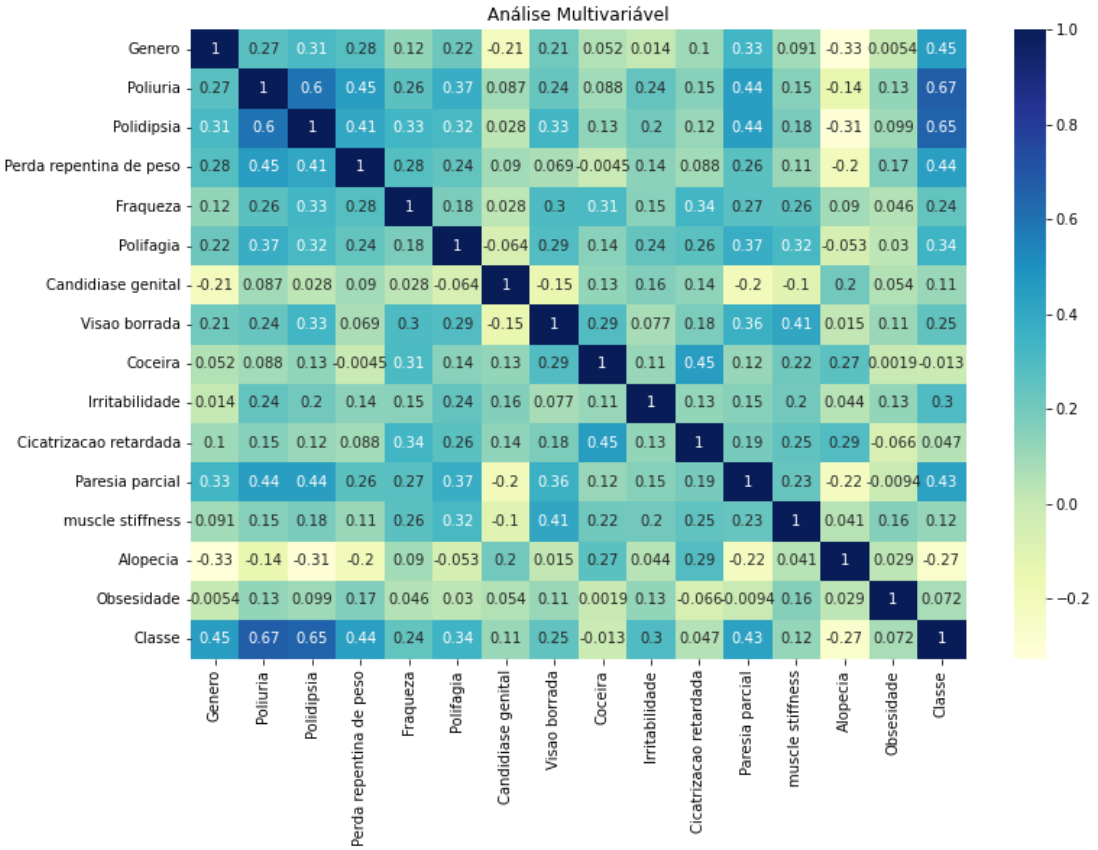

Figura 3 - Correlação entre atributos e classe. Fonte: Elaborada pelo autor.

Pré Processamento dos dados

Antes da criação dos modelos de ML, como a maior parte dos dados são categóricos com valores como “Não” ou “Sim”, foi necessário convertê-los para o tipo numérico como 0 e 1, respectivamente. A coluna idade varia com valores entre 16 a 90 anos, conforme Figura 4. Logo esses valores foram normalizados para que a distribuição ficasse entre 0 e 1 como os demais atributos. A normalização dos dados é importante, pois possibilita que todos os dados estejam na mesma escala e, por consequência, tenham a mesma importância no treinamento do modelo. Além disso, o conjunto de treinamento foi balanceado, por conta das classes não terem a mesma proporção, como visto na Figura 2. O balanceamento foi feito aumentando os dados da classe menor, gerando dados sintéticos. Para isso, foi utilizada a função *SMOTETomek,* que gera essas amostras sintéticas fazendo a interpolação dos dados da classe já existentes. Dessa forma, o número de amostras de ambas as classes é igualado.

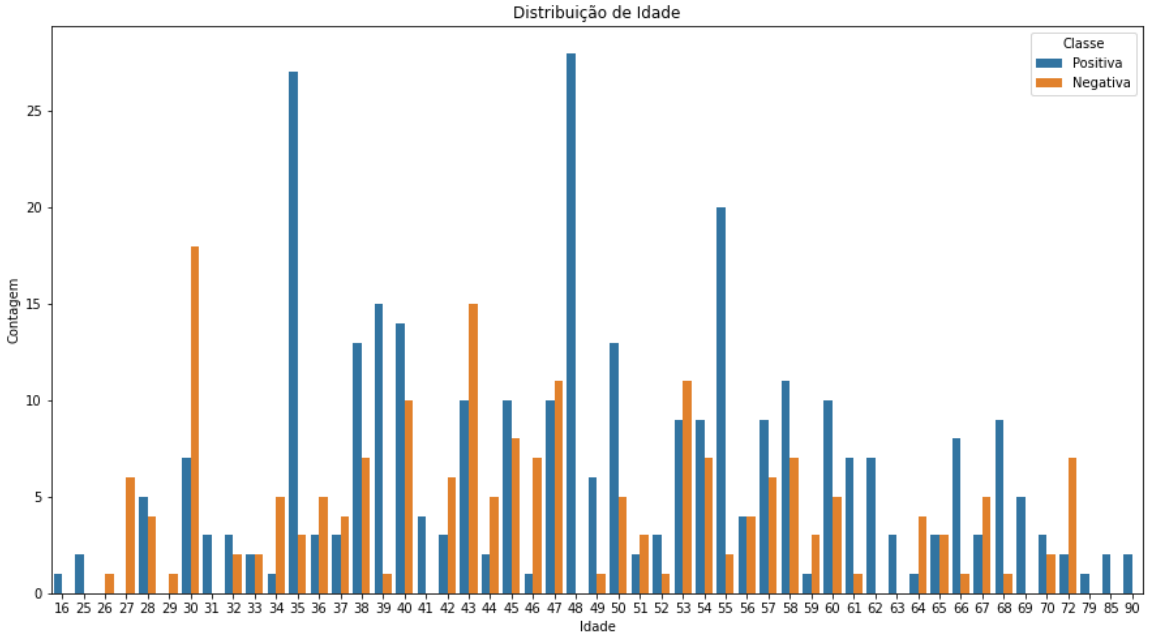

Figura 4 - Distribuição de idade entre as classes. Fonte: Elaborada pelo autor.

Algoritmos de Machine Learning

Random Forest

Random Forest (RF) é um método de aprendizado em conjunto (*ensemble*), no qual combina várias estruturas aleatórias de *Decision Tree* (DT). Cada árvore recebe uma amostra do conjunto de dados de treino através do *Bootstrap,* que é um método de reamostragem cujas amostras selecionadas podem ser reutilizadas. A partir disso, algumas das variáveis são selecionadas de forma aleatória e então cada DT realiza a classificação. Por fim, é realizada uma votação majoritária, cuja classe mais escolhida entre cada DT será a saída do algoritmo, conforme Eq. 1.

$$H(x) = arg\,max_Y \sum_{i=1}^{k} I(h_i(x) = Y) \quad (1)$$

Onde H(x) é a combinação do modelo de classificação, h_i é um modelo de DT único, Y é a variável de saída, I é a função indicadora, conforme exemplo na Figura 5. *Random Forest* é usado amplamente em muitas áreas, como bioinformática, medicina e economia. Este modelo é executado de forma rápida, sem gerar *overfitting* e consegue lidar tanto com variáveis contínuas quanto categóricas (LIU; WANG; ZHANG, 2012). *Overfitting* ocorre quando o modelo de aprendizagem de máquina se ajusta muito bem aos dados de treinamento, porém não consegue generalizar para dados novos, ou seja, o modelo se especializou nos padrões dos dados de treinamento, no entanto não foi capaz de aplicar o que aprendeu a outros conjuntos de dados.

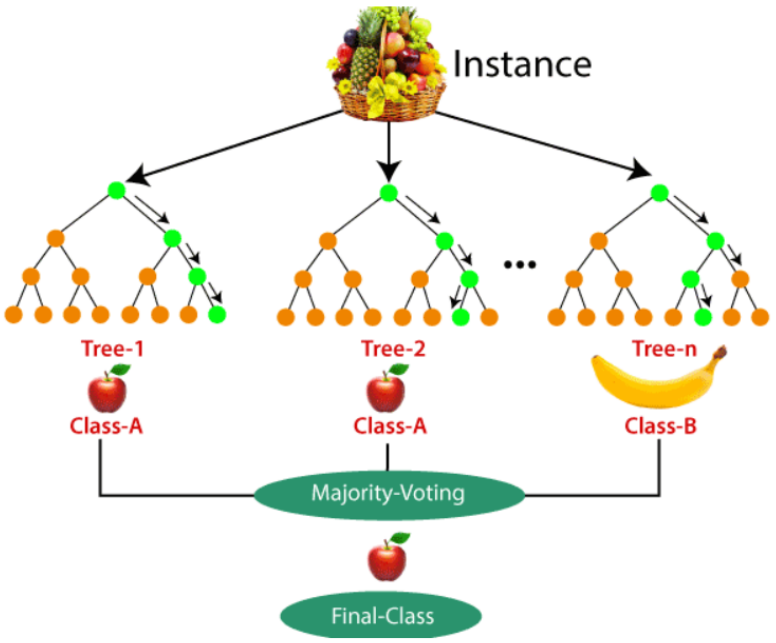

Figura 5 - Exemplo de um algoritmo RF. A partir de um conjunto de dados, árvores de decisão são geradas. A classe que aparece na maioria dos resultados dessas DT é usada como saída do RF. Fonte: (BOSE, 2023).

Extreme Gradient Boosting – XGboost

XGboost é outro método de aprendizado *ensemble*. Ele é baseado em *Decision Tree* com *Gradient Boosting*, que ao invés de ajustar o peso dos dados a cada interação, esse método ajusta um novo classificador aos erros residuais do classificador anterior, ou seja, o preditor ou DT aprende nos exemplos que seu predecessor errou, minimizando a perda. Conforme Eq. 2.

$$\widehat{y_i} = \phi(x_i) = \sum_{k=1}^{k} f_k(x_i),\ f_k \in F, \tag{2}$$

Onde F = $\{f(x) = w_{q(x)}\}(q : R^m \rightarrow T, w \in R^T)$ é o espaço das árvores de regressão. Aqui q representa a estrutura de cada árvore que mapeia um exemplo para o correspondente index da folha. T é o número de folhas na árvore. Cada f_k corresponde a uma estrutura de árvore q e os pesos de folhas w. Então, são usadas regras de decisão nas árvores (dadas pelo q) para classificar nas folhas e calcular a predição final somando o *score* de cada folha (dado por w). Para aprender o conjunto de funções usadas no modelo, é minimizado o seguinte objetivo regularizado.

$$\widehat{y_i} = \mathcal{L}(\phi) = \sum_i l(\widehat{y_i}, y_i) + \sum_k \Omega(f_k) \tag{3}$$

$$\text{onde } \Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$$

Aqui l é a função de *loss* convexa diferenciável que mede a diferença entre a predição $\widehat{y_i}$ e o alvo y_i. O segundo termo Ω penaliza a complexidade do modelo. O termo de regularização adicional ajuda a suavizar os pesos aprendidos evitando o *overfitting*. XGboost

tem resultados efetivos em uma ampla gama de problemas, tendo como maiores fatores para o seu sucesso a escalabilidade em todos os cenários e a velocidade de execução 10 vezes maior que as soluções populares (CHEN; GUESTRIN, 2016).

Support Vector Machine

O *Support Vector Machine* SVM é um algoritmo de aprendizado de máquina supervisionado desenvolvido para realizar a criação de um hiperplano responsável por separar duas classes, produzindo uma margem entre os parâmetros presentes, demarcando pontos que estarão dentro desta margem ou fora, nos quais, os pertencentes ao campo da margem são denominados pontos ideais e fora denominados como vetores de suporte (MEYER, 2023). De acordo com (CAMPBELL; YIMING, 2011) a função da reta para determinar o hiperplano central é representada por:

$$w \times x + b = 0 \quad (4)$$

Onde x são os dados localizados com o hiperplano, b o desvio do hiperplano e w o peso que determina a orientação da reta. Este hiperplano é definido pelos pontos máximos de cada classe localizados em cada lado do mapeamento, em que, são traçados dois hiperplanos para separar as classes e delimitar a margem, sendo obtidos através das equações 5 e 6:

$$yi = +1(w \times x + b \geqslant 0) \quad (5)$$
$$yi = -1(w \times x + b < 0) \quad (6)$$

Com o auxílio dos vetores de suporte podemos maximizar o hiperplano, garantindo uma maior margem e possibilitando diminuir o erro de futuras classificações (CAMPBELL; YIMING, 2011). Na figura 6 é possível visualizar a utilização do algoritmo SVM, em que duas classes são divididas e separadas através de dois hiperplanos. Os hiperplanos demonstrados na cor azul e verde são responsáveis por formar a margem entre as classes com objetivo de atingir a máxima distância possível, sendo o hiperplano central destacado em vermelho, a margem máxima. A grandeza entre os dois hiperplanos é equivalente a $\frac{2}{||w||}$ (ELSAYED et al., 2019).

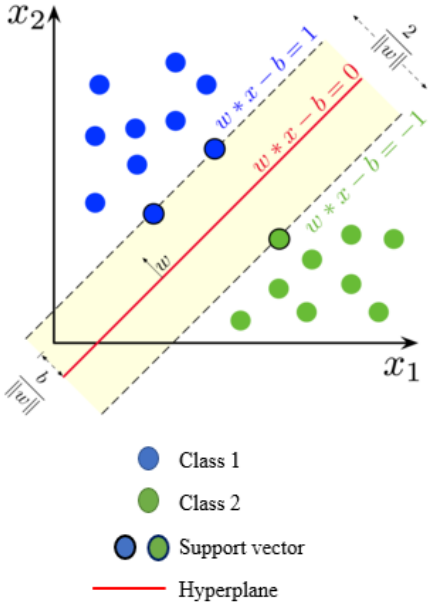

Figura 6 - Exemplo de um hiperplano traçado utilizando o algoritmo SVM. É possível visualizar o hiperplano central em conjunto com os demais definidos através das classes exemplificadas, determinando a margem para a classificação. Fonte: (ELSAYED et al., 2019).

Contudo, o algoritmo de SVM de maneira geral apresenta bons resultados e uma teoria extremamente embasada. O SVM também é capaz de tratar casos de regressões de maneira numérica, diferenciando-se dos métodos categóricos, classificados entre sim e não (BOSWELL, 2002).

Naive Bayes Classifier

Nai*ve Bayes Classifier* NB é um algoritmo de aprendizado de máquina supervisionado, o qual é fundamentado no Teorema de Bayes, sendo um dos algoritmos mais conhecidos para classificação. O teorema consiste em uma análise de probabilidade que pretende calcular as chances de um evento acontecer após outro evento, representado pelas variáveis A e B respectivamente na Eq. 7 (KANNAN et al., 2022):

$$P\,(A \mid B) \;=\; P(A) \;\times\; \frac{P(B|A)}{P(B)} \tag{7}$$

Os quais P indica a probabilidade a ser calculada, A o evento a ocorrer e B o evento ocorrido. Na utilização para *Machine Learning*, a abordagem segue outro viés, podendo ser representada através da Eq. 8:

$$Posterior \;=\; \frac{Probabilidade \times Anterior}{Normalização} \tag{8}$$

Em que o termo Posterior é considerado a probabilidade de ocorrência do resultado de P(A|B) com a condição que o atributo B do conjunto de dados esteja sendo atendido, o termo Probabilidade definido como a probabilidade de um atributo atender o seguinte parâmetro, o recurso A presente na base de dados ser verdadeiro relacionando se ao atributo B, o termo Anterior considerado a probabilidade de um evento ocorrer antes dos dados serem constatados e normalização que refere-se a probabilidade de cada atributo contido no conjunto de dados (KANNAN et al., 2022).

Resultados e Discussão

Métricas

Neste estudo foram utilizados 4 algoritmos de ML para realizar a predição de diabetes em estágio inicial, sendo estes, SVM, NB, RF e *XGboost*. Como métricas utilizadas para comparações e análises de desempenho de cada algoritmo foram aplicados seis diferentes critérios denominados como precision, recall, accuracy, F1-Score e balanced accuracy.

Na utilização de modelos de classificação, tradicionalmente é realizada a criação de uma matriz de confusão, responsável por separar os resultados dos dados de teste entre, verdadeiros positivos (VP), falsos positivos (FP), falsos negativos (FN) e verdadeiros negativos (VN). Os dados classificados como VP acontecem quando o modelo classifica um paciente que possui diabetes como uma pessoa que possui diabetes. Os valores que resultam na classe de VN são atribuídos quando uma pessoa que não possui a doença é classificada como não diabética. Para os dados que resultarem na classe de FP classificamos as pessoas que não possuem diabetes e que são classificadas como diabéticas.. Os atribuídos na classe FN são referentes às pessoas que possuem diabetes e são classificadas como saudáveis.

Através da figura 7 é possível visualizar a matriz de confusão e suas respectivas divisões entre as classes citadas, demonstrando a diagonal principal formada pelos acertos de classificação do algoritmo, sendo estes, os dados categorizados como VP e VN, enquanto a diagonal secundária é composta pelos erros de classificação dos dados especificados como FN e FP.

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 7 - Exemplo de uma matriz de confusão utilizada para classificação. Fonte: (RODRIGUES, 2019).

Com isto é possível obter os parâmetros necessários para realizar o cálculo das métricas que variam os resultados entre os valores de 0 a 1, sendo os valores mais próximos de 1 os melhores resultados. Tais métricas são representadas pelas equações abaixo de acordo com (VAROQUAUX; COLLIOT, 2023):

Precisão é considerada a quantidade dados que realmente são positivos, classificados como positivos conforme demonstrado na equação 9.

$$Precision = \frac{VP}{VP + FP} \quad (9)$$

Recall é considerada a quantidade de dados que almeja-se retornar ao valor positivo e são de fato positivos conforme demonstrado na equação 10.

$$Recall = \frac{VP}{VP + FN} \quad (10)$$

A acurácia é considerada a quantidade de dados classificados de maneira correta conforme demonstrado na equação 11.

$$Accuracy = \frac{VP + VN}{VP + FP + VN + FN} \quad (11)$$

F1-Score realiza o cálculo da média entre a precisão e o recall conforme demonstrado na equação 12.

$$F1\ Score = \frac{2 \times VP}{2 \times VP + FP + FN} \quad (12)$$

Especificidade é o termo referente aos dados classificados como negativos e de fato são negativos conforme demonstrado na equação 13.

$$Specificity = \frac{VN}{VN + FP} \quad (13)$$

A acurácia balanceada é considerada a quantidade de dados classificados de maneira correta, porém através dela é possível obter melhores resultados,visto que, por este cálculo os dados não balanceados não influem no resultado conforme demonstrado na equação 14.

$$Balanced\ Accuracy = \frac{Recall + Specificity}{2} \quad (14)$$

A curva ROC (Característica de Operação do Receptor) serve para representar a relação entre a taxa de VP ou sensibilidade e a taxa de FP ou especificidade. Para o cálculo da área sob a curva ROC é necessário realizar o cálculo de probabilidade de ocorrência de um dado positivo ser maior do que um dado negativo, sendo considerado o valor 1 o valor ideal para classificação e o valor de 0.5 como uma classificação aleatória.

Análise dos resultados

Neste estudo, foram realizadas diferentes análises para avaliar os quatro modelos de classificação para a predição de diabetes em estágio inicial. Em todos os resultados RF se saiu melhor que os demais, tendo uma acurácia de 99% e uma acurácia balanceada de 99.3%. Com uma precisão de 100% e um *recall* de 98.6%, seu F1-score foi de 99.3% nos dados de teste. Em contrapartida, o que obteve o menor desempenho foi o *Naive Bayes,* tendo uma acurácia de 91.3% e uma acurácia balanceada de 89.6%. Sua precisão e *recall* foram de 93.1% e 94.4%, respectivamente e, consequentemente, seu F1-Score foi de 93.7%. De acordo com os critérios utilizados para as comparações dos experimentos realizados, o classificador *Random*

Forest é mais efetivo que os demais na predição prévia da diabetes. A Tabela 2 apresenta os resultados dos quatro modelos de aprendizagem de máquina e a Figura 8 mostra a curva ROC.

Model	Accuracy	Balanced Accuracy	F1_Score	Precision	Recall
Random Forest Classifier	0.99	0.993	0.993	1	0.986
XGBoost Classifier	0.962	0.972	0.971	1	0.944
SVM Linear	0.933	0.926	0.95	0.957	0.944
Naive Bayes Classifier	0.913	0.896	0.937	0.931	0.944

Tabela 2 - Resultados dos classificadores. Fonte: Elaborada pelo autor.

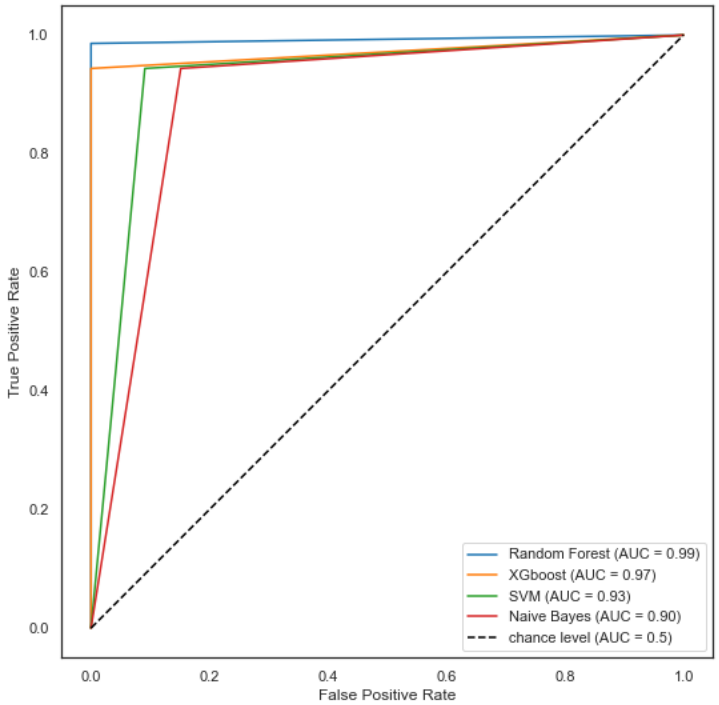

Figura 8 - Característica de Operação do Receptor (ROC). Fonte: Elaborada pelo autor.

Conclusão

Este estudo teve como objetivo investigar a adequação de algoritmos de *Machine Learning* para a identificação de diabetes em estágio inicial. Foram implementados quatro modelos distintos - *Random Forest, XGBoost, Support Vector Machine e Naive Bayes* - para avaliar a eficácia na predição dessa condição. Os resultados obtidos revelaram que o modelo *Random Forest* destacou-se significativamente em relação aos demais, apresentando uma acurácia de 99%, uma acurácia balanceada de 99.3%, uma precisão de 100%, e um *recall* de 98.6%. O F1-score deste modelo foi notável, atingindo 99.3% nos dados de teste. Em contrapartida, o *Naive Bayes* demonstrou o menor desempenho, com uma acurácia de 91.3%, acurácia balanceada de 89.6%, precisão de 93.1%, *recall* de 94.4%, e F1-Score de 93.7%.

Os resultados reforçam a conclusão de que algoritmos de *Machine Learning* são eficazes na identificação de diabetes em estágio inicial, sendo o *Random Forest* o mais efetivo entre os modelos avaliados. Essa constatação sugere a viabilidade de implementar sistemas de *Machine Learning* como ferramentas de apoio à decisão na identificação precoce da diabetes. Além disso, a alta precisão, *recall*, acurácia, F1-Score e acurácia balanceada alcançadas pelo *Random Forest* corroboram a robustez do modelo. As implicações práticas deste estudo são significativas, indicando que a aplicação de algoritmos de *Machine Learning* pode contribuir para a otimização do processo de identificação de diabetes em estágio inicial. Isso não apenas reduziria os custos associados a exames extensivos, mas também tornaria o processo mais

rápido e acessível. No entanto, é importante reconhecer as limitações inerentes à base de dados utilizada neste estudo.

Como direção para pesquisas futuras, é sugerido disponibilizar o modelo desenvolvido para testes, a fim de validar sua aplicabilidade em cenários práticos. Este passo adicional pode fornecer insights valiosos sobre a eficácia do modelo em ambientes do mundo real, contribuindo para a contínua evolução das ferramentas de diagnóstico baseadas em *Machine Learning*. Em resumo, os resultados deste estudo reforçam a promissora aplicação de algoritmos de *Machine Learning* na identificação de diabetes em estágio inicial, com implicações práticas significativas para aprimorar os métodos de diagnóstico e melhorar a eficiência dos sistemas de saúde.

Referências Bibliográficas

AYODELE, T. O. Types of Machine Learning Algorithms. In: ZHANG, Y. **New Advances in Machine Learning**. Croacia: IntechOpen, 2010. p. 19-48.

BARAKAT, N.H.; BRADLEY, A.P.; BARAKAT, M.N. Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. **IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society**, Austrália, jul. 2010. DOI: https://doi.org/10.1109/TITB.2009.2039485.

BEAM, A.L.; KOHANE, I.S. Big Data and Machine Learning in Health Care. **JAMA**, v.13, p.1317–1318, Abr. 2018. DOI: https://doi.org/10.1001/jama.2017.18391

BOSE, S. Random Forest Algorithm. Inside AIML, 2021. Disponível em: https://insideaiml.com/blog/Random-Forest-Algorithm-1029. Acesso em: 13 nov. 2023.

BOSWELL, D. Introduction to Support Vector Machines. **Computer Science, Mathematics**, 2002. Disponível em: https://home.work.caltech.edu/~boswell/IntroToSVM.pdf. Acesso em: 15 out. 2023

CAMPBELL, C.; YIMING, Y. **Learning with Support Vector Machines**. Morgan & Claypool Publishers, 2011. 100 p.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)**. Association for Computing Machinery, New York, USA, p.785–794. Aug. 2016. DOI: https://doi.org/10.1145/2939672.2939785

ELSAYED, H. et. al. Integrating Modern Classifiers for Improved Building Extraction from Aerial Imagery and LiDAR Data. **American Journal of Geographic Information System** 213-220. Oct, 2019. Disponível em: https://www.researchgate.net/publication/336810849_Integrating_Modern_Classifiers_for_Improved_Building_Extraction_from_Aerial_Imagery_and_LiDAR_Data. Acesso em: 29 out. 2023

FARUQUE, M. F.; ASADUZZAMAN; SARKER, I. H. Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus. **International Conference on Electrical, Computer and Communication Engineering (ECCE)**, Bangladesh, p. 1-4, Feb. 2019. DOI: https://doi.org/10.1109/ECACE.2019.8679365

GUYTON, J.E.; HALL, A.C. **Tratado de Fisiologia Médica**. Elsevier, 2011. 1176 p.

INTERNATIONAL DIABETES FEDERATION. **Diabetes Atlas**. Bélgica: International Diabetes Federation, 2006, 3rd. ed. Disponível em: https://diabetesatlas.org/upload/resources/previous/files/3/Diabetes-Atlas-3rd-edition.pdf. Acesso em: 15 set. 2023.

INTERNATIONAL DIABETES FEDERATION. Facts & figures. In: Facts & figures, 2023. Disponível em: https://idf.org/about-diabetes/diabetes-facts-figures/. Acesso em: 15 set. 2023.

KANNAN, H. et. al. **Bayesian Reasoning and Gaussian Processes for Machine Learning Applications**. Chapman and Hall/CRC, 2022. 133 p.

KHAN, M. M. A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease. **Journal of Healthcare Engineering**, vol. 2021, p. 1-12, Jul. 2021. DOI: https://doi.org/10.1155/2021/9917919

LIU, Y.; WANG, Y.; ZHANG, J. New Machine Learning Algorithm: Random Forest. **Information Computing and Applications (ICICA)**, Springer, Berlin, v.7473, p.246-252. 2012, DOI: https://doi.org/10.1007/978-3-642-34062-8_32.

MA, J. Machine Learning in Predicting Diabetes in the Early Stage. **2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)**, Taiyuan, China, p. 167-172, Oct, 2020. Disponível em: https://doi.org/10.1109/MLBDBI51377.2020.00037.

MAGLOGIANNIS, I. et. al. **Emerging Artificial Intelligence Applications in Computer**: Real Word Ai Systems With Applications in Ehealth. Holanda: Ios Pr Inc, 2007. v. 160. 407 p.

MARINHO, N. B. P. et. al. Risco para diabetes mellitus tipo 2 e fatores associados. **Acta Paulista de Enfermagem**, Ceará, v. 26, n. 6, p. 569-574, jun. 2013. DOI: https://doi.org/10.1590/S0103-21002013000600010

MEYER, D. Support Vector Machines: The Interface to libsvm in package e1071. **R-News**, Austria, p. 1-8, fev, 2023. Disponível em: https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf. Acesso em: 14 out. 2023.

NASTESKI, V. An overview of the supervised machine learning methods. **Horizons**, Macedônia, v. 4, p. 51-52, Dez. 2017.

POOLE, D.; GOEBEL, R.; MACKWORTH, A.K. **Computational Intelligence**: A Logical Approach. Nova Iorque: Oxford University Press, 1998. 576 p.

R, P.R; NAIR, R.S. A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms. **IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)**, Coimbatore, India, p. 1-4, Feb. 2019. DOI: https://doi.org/10.1109/ICECCT.2019.8869001

RAHMAN, A.K.M.S. et. al. A Comparative Study on Liver Disease Prediction Using Supervised Machine Learning Algorithms. **International Journal of Scientific &**

Technology, v.8, p. 419-422, nov. 2019. Disponível em: https://www.ijstr.org/final-print/nov2019/A-Comparative-Study-On-Liver-Disease-Prediction-Using-Supervised-Machine-Learning-Algorithms.pdf. Acesso em: 28 set. 2023.

RASCHKA, S. **Python Machine Learning**. Reino Unido: Packt Publishing, 2015. 454 p.

RODRIGUES, V. Métricas de Avaliação: acurácia, precisão, recall… quais as diferenças?. **Medium**., Abr. 2019. Disponível em: https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c. Acesso em: 29 out. 2023.

RUSSEL, S.; NORVIG, P. **Artificial Intelligence**: A Modern Approach. 3. ed., Inglaterra: Pearson, 2009. 1152 p.

SARWAR, M. A. et. al. Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. **24th International Conference on Automation and Computing (ICAC)**, Newcastle Upon Tyne, UK, p. 1-6, Sep. 2018. DOI: https://doi.org/10.23919/IConAC.2018.8748992

TR, R. et. al. Predictive Analysis Of Heart Diseases With Machine Learning Approaches. **Malaysian Journal of Computer Science**, p. 132–148, Mar. 2022. DOI: https://doi.org/10.22452/mjcs.sp2022no1.10

VAROQUAUX, G.; COLLIOT, O. **Machine Learning for Brain Disorders**. Humana, 2023, p. 1047. DOI: https://doi.org/10.1007/978-1-0716-3195-9_20

ZOU, Q. et. al. Predicting Diabetes Mellitus With Machine Learning Techniques. **Front. Genet**, v. 9, p. 1-10, Nov. 2018. DOI: https://doi.org/10.3389/fgene.2018.00515