# COMPARAÇÃO DE ALGORITMOS DE MACHINE LEARNING PARA DETECÇÃO DE DIABETES

**Keven Kauê Gonçalves Pinto**[1]

[1]*Universidade Federal do Pará, Tucuruí, Brasil (keven.pinto@tucurui.ufpa.br)*

**Área: Multidisciplinar**

*Resumo: O Diabetes Mellitus (DM) é uma doença metabólica causada pela falta de insulina, resultando em hiperglicemia e complicações como problemas cardíacos e renais. Este estudo usa algoritmos de Machine Learning (ML) para prever a existência do diabetes, avaliando métricas como o f-measure. O modelo Gradient Boosting (GB) obteve os melhores resultados.*

*Palavras-chave*: diabetes; classificação; machine learning; F-Measure.

## INTRODUÇÃO

O crescimento exponencial do diabetes em meio a população brasileira salienta a necessidade de ações preventivas, haja vista, as complicações decorrentes de seu aparecimento. O aumento dessa incidência está atrelado a diversas razões socioecônomicas, fruto de dietas desrreguladas e inatividades físicas (Grillo e Gorini, 1997).

Bertoldi et al. (2013), destaca que apenas no ano de 2013 cerca de 12 milhões de brasileiros apresentavam um quadro diabético, exprimindo a magnitude que essa doença crônica tomou.

Para evitar os desdobramentos dessa desrregulação do organismo é necessário uma identificação precoce da mesma, nesse sentido o presente trabalho realiza uma análise de dados tabelados extraídos de pacientes em conjunto à métodos de *Machine Learning* (ML), comparando resultados e destacando o melhor modelo, com o objetivo de obter os melhores resultados na classificação da existência ou não de diabetes, assim sendo mais uma ferramenta de auxílio para essa área da saúde.

## MATERIAL E MÉTODOS

O dataset utilizado é proveniente do repositório de ML da Universidade da Califórnia (UC Irvine) com informações oriundas do *Centers for Disease and Control* (CDC), um coletor de dados anuais do Estados Unidos. O dataset é organizado em 21 colunas contendo certas caracterícas de análise e contando com 253.680 amostras de pacientes.

Para o ambiente de execução foi escolhido do *Google Colab*, por sua integração ao *Google Drive* além de seu poder de processamento de 12GB de CPU RAM, sendo essencial para compilar o programa sem imprevistos.

Os algoritmos utilizados nesse estudo variam desde modelos de decisão baseados em árvores, métodos ensemble até modelos de regressão logística. Dentre os algoritmos

escolhidos estão: *Extra Trees* (EXT), *Gradient Boosting* (GB), *K-Nearest Neighbors* (KNN), *Logistic Regression* (LR), *Random Forest* (RF) e por fim Linear *Support Vector Machine* (LinearSVC).

Para garantir que houvesse um melhor processamento o primeiro passo empregado foi verificar quais colunas eram mais significativas para o trabalho, garantindo que as colunas mais relevantes sejam utilizadas para o treino e teste dos modelos de ML. Assim, a abordagem utilizada baseia-se no método *SelectKBest* utilizando os parâmetros k=10, para pegar as 10 mais relevantes características, e $chi^2$ para selecionar features. O funcionamento desse segundo parâmetro está fundamentado nas relações das features e característica alvo por meio da comparação dos resultados do $chi^2$ com a de um $p-valor$, oriundo de cálculos relacionados à probabilidades entre as suas predições e o real valor dos dados (Wainer et al., 2007). O $chi^2$ pode ser representado matemáticamente como:

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad (1)$$

Onde $x^2$ representa o resultado de $chi^2$, $O_i$ o valor observado e $E_i$ o valor esperado. Supondo que seja obtido um valor alto significa uma forte relação, em contraponto valores baixos indicam uma relação significativamente menor.

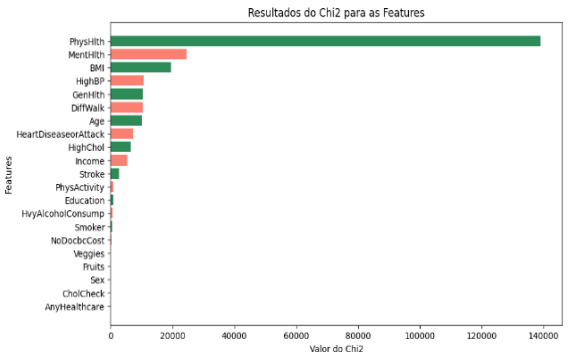


Figura 1. Resultados individuais para as features.

Os resultados obtidos por meio do uso $chi^2$ indicam uma forte relevância de características como saúde corporal, índices de massa corporal, alta taxa de batimentos cardíacos dentre outras para a classificação do modelo.

Segundamente, após a diminuição do número de colunas foram separadas o contingente do dataset entre 80% de seu total para o treinamento dos modelos escolhidos e 20% para o teste dos mesmos.

Após essa separação foram aplicados métodos relacionados à redução da dimensionalidade do dataset, isto é, balanceá-lo, visto que, uma desrregulação na classe alvo analisada levaria o modelo a enviesar-se, prejudicando o seu desempenho. Desse modo, para evitar esse problema o primeiro método utizado chama-se *RandomUnderSampler*.

O seu funcionamento baseia-se no equilíbrio das classes-alvo por meio da retirada aleatória de valores da classe dominante (pessoas sem diabete) enquanto classe minoritária permanece inalterada, ao fim desse procedimento ambas as esferas de features terão parcelas iguais de features.

O segundo passo para o pré-processamento dos dados foi utilizar o *StandartScaler* para padronizar o restante dos dados. Esse procedimento é de suma importância para garantir o correto funcionamento de modelos pautados em escalas de dados como o KNN, *LinearSVC* e LR, evitando que escalas maiores dominem o modelo enquanto as minoritárias sejam evitadas (Raju et al., 2020). Ademais, facilitando o processamento por esses algoritmos, uma vez que, com as instâncias balanceadas os resultados convergiram mais rapidamente.

Sua perfomance baseia-se na transformação de cada feature de acordo com a seguinte função matemática:

$$z = \frac{x-\mu}{\sigma} \quad (2)$$

Na qual, $z$ significa a feature após a sua padronização, $x$ o valor originário da feature, $\mu$ a média simples da feature em meio ao seu conjunto de dados e $\sigma$ o seu desvio padrão em relação aos dados. Dessa forma, o modelo estará pronto para ser treinado, evitando o maior número de problemáticas relacionadas ao conjunto de dados.

As bases para avaliação dos modelos são pautados em modelos matemáticos atrelados a 4 alicerces da área de aprendizado de máquina: acurácia, precisão, *recall* e *f-measure*. Esses métodos de avaliação estão associados a quatro termos amplamente conhecidos durante manejo de dados binários, sendo esses, *True Positives* (TP), pessoas sem diabetes classificadas corretamente, *True Negatives* (TN), diabéticos reconhecidos corretamente, *False Positives* (FP), indivíduos saudáveis errôneamente classificados com a doença e *False Negatives* (FN), sendo esse o pior dos casos onde uma pessoa com esse acometimento é classificada como saudável. Visando reduzir ao máximo os casos de FN a métrica de escolha para avaliar os resultados obtidos de cada modelo foi o *f-measure*, para entende-lo é necessário primeirante entender o restante das métricas, iniciando pela acurácia:

$$acurácia = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Representa o desempenho geral do modelo, entretanto não enfatiza a proporção de falsos positivos e negativos, cujos são essenciais para avaliar o impacto do modelo.

$$precisão = \frac{TP}{TP+FP} \quad (4)$$

Enfatiza a taxa de reais positivos dentre todas as previsões positivas do modelo, é necessária para calcular quantas pessoas foram diagnosticadas corretamente evitando os falsos positivos.

$$recall = \frac{TP}{TP+FN} \quad (5)$$

É uma métrica de avaliação que objetiva medir a maior quantidade de verdadeiros positivos do modelo, ou seja, visa minimizar a taxa de falsos negativos.

$$f - measure = 2 * \frac{precisão*recall}{precisão+recall} \quad (6)$$

Finalmente, o *f-measure* retrata a média harmônica entre a precisão e o *recall*, basicamente funcionando um resultado geral de ambos, sendo a métrica mais essencial visando maximizar a análise de erros e acertos de cada modelo, logo, sendo a escolhida para analisar o desempenho dos algoritmos de ML já abordados.

## RESULTADOS E DISCUSSÃO

Após a realização todo o processamento descrito os resultados obtidos com base no *f-measure* para cada um dos modelos apresentou-se da seguinte maneira, ressaltando que o melhor modelo foi destacado em negrito.

Tabela 1. Resultados individuais para cada algoritmo de ML.

| Modelos | F-measure |
|---|---|
| EXT | 70,16% |
| **GB** | **75,22%** |
| KNN | 70,49% |
| LinearSVC | 74,36% |
| LG | 74,23% |
| RF | 71,43% |

A Tabela 1 demonstra que o melhor resultado foi alcalçado pelo método ensemble *Gradient Boosting*, dentre as outras métricas sua acurácia ficou em 74,26%, precisão em 72,51% e seu *recall* em 78,13%, o último sendo o maior dentre todos os outros algoritmos. A explicação de seu ótimo funcionamento no conjunto de dados está atrelado aos seus métodos internos de mitigação de erro, utilizando pequenas árvores de decisão com altura limitada.

Para cada amostra há um cálculo entre o valor predito e o valor real chamado de resíduo, o qual é utilizado para ajustar o modelo, isso ocorre pela junção do resíduo as árvores de decisão pequenas visando minimizar os erros do modelo geral (Natekin e Knoll, 2013).

Investigando mais a fundo os resultados do GB, foi gerado uma matriz de confusão para verificar a proporção entre as classificações errôneas e verdadeiras.

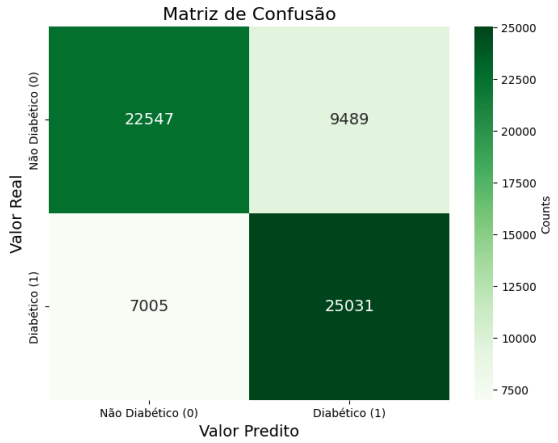


Figura 2. Matriz de confusão do Gradient Boosting.

Analisando a Figura 2 ,é possível ver que dentre as amostras separadas unicamente para teste houveram ainda sim uma grande incidência de casos onde pessoas com diabetes eram classificadas sem essa enfermidade. No entanto, é fundamental destacar que o método proposto ainda obteve um percentual de 75% de acerto geral, indicando que para cada conjunto de quatro pessoas, três seriam corretamente diagnosticadas, revelando seu potencial como ferramenta de suporte atrelada à saúde brasileira e demonstrando que simples métodos de análise em informações obtidas de pacientes podem ser utilizadas como alicerce para um rápido panorama acerca da saúde dos mesmos.

## CONCLUSÃO

Em síntese, o presente estudo alcançou o objetivo proposto de investigar métodos de *machine learning* associados à classificação da existência do diabetes mellitus, alcançando bons valores em relação ao modelo selecionado, *gradient boosting*, mostrando seu potencial de uso para classificação das amostras e seu grau de relevância para soluções de problemas da área da saúde.

## REFERÊNCIAS

BERTOLDI, Andréa D. et al. Epidemiology, management, complications and costs associated with type 2 diabetes in Brazil: a comprehensive literature review. Globalization and health, v. 9, p. 1-12, 2013.

GRILLO, Maria de Fátima Ferreira; GORINI, Maria Isabel Pinto Coelho. Caracterização de pessoas com diabetes mellitus tipo 2. Revista Brasileira de Enfermagem, v. 60, p. 49-54, 2007.

NATEKIN, Alexey; KNOLL, Alois. Gradient boosting machines, a tutorial. Frontiers in neurorobotics, v. 7, p. 21, 2013.

RAJU, VN Ganapathi et al. Study the influence of normalization/transformation process on the accuracy of supervised classification. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, p. 729-735, 2020.

WAINER, Jacques et al. Métodos de pesquisa quantitativa e qualitativa para a Ciência da Computação. Atualização em informática, v. 1, n. 221-262, p. 32-33, 2007.