REVISTA Mundi
Engenharia, Tecnologia e Gestão
INSTITUTO FEDERAL Paraná

CLASSIFICAÇÃO DA SITUAÇÃO EMPREGATÍCIA EM DADOS SOBRE INCLUSÃO FINANCEIRA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

CLASSIFICATION OF EMPLOYMENT SITUATION IN FINANCIAL INCLUSION DATA USING MACHINE LEARNING ALGORITHMS

Ana Maria Teixeira Fossa[1]

Nayara Manzutti Ikuta[2]

Ronaldo Octaviano Diniz Junqueira Filho[3]

Bruno Samways dos Santos[4]

Rafael Henrique Palma Lima[5]

Resumo: Para uma instituição financeira, compreender o perfil de seus clientes é de suma importância a tomada de decisões estratégicas. Dados como alta taxa de inadimplência, pouca idade e desemprego mostram à instituição qual é o perfil de seus clientes, e isso pode mudar a direção de futuras tomadas de decisão. Neste contexto, este artigo apresenta o uso de técnicas de *machine learning* para classificar a situação empregatícia com base no banco de dados disponibilizado pelo Banco Mundial sobre a inclusão financeira (*Global Findex*). Após a limpeza dos dados, realizou-se um comparativo entre as técnicas de Naïve Bayes, Regressão logística, Floresta aleatória e *k*-vizinhos mais próximos (*k-nearest neighbors* - KNN). Os resultados foram analisados pelas métricas de acurácia, precisão e *f-score* e validadas com 10-fold. Os resultados mostraram que a acurácia e *f-score* foi a melhor para a técnica KNN, enquanto Naïve Bayes atingiu a melhor precisão, mas um baixo *f-score*, sugerindo uma baixa sensibilidade. Variáveis como a idade, nível educacional, recebimento de salários e pensões e quantia salarial foram destacadas como boas variáveis preditoras de acordo com a Floresta aleatória. Neste contexto, sugeriu-se investigar estas variáveis posteriormente para uma análise mais profunda sobre o perfil dos cidadãos empregados e desempregados a partir destes preditores.

Palavras-chave: Aprendizado de máquina. Inclusão financeira. Emprego.

[1] Graduada em Engenharia de Produção, Universidade Tecnológica Federal do Paraná, Campus Londrina, anatexe@hotmail.com.

[2] Graduada em Engenharia de Produção, Universidade Tecnológica Federal do Paraná, Campus Londrina, nayara_ikuta@hotmail.com.

[3] Graduado em Engenharia de Produção, Universidade Tecnológica Federal do Paraná, Campus Londrina, rodjunqueira@gmail.com.

[4] Doutor em Engenharia de Produção e Sistemas, Universidade Tecnológica Federal do Paraná, Campus Londrina, brunosantos@utfpr.edu.br.

[5] Doutor em Engenharia de Produção, Universidade Tecnológica Federal do Paraná, Campus Londrina, rafaelhlima@utfpr.edu.br.

Abstract: For a financial institution, understanding the profile of its clients is of utmost importance for making strategic decisions. Data such as high default rates, young age, and unemployment show the institution what its clients' profile is, and this can change the direction of future decision-making. In this context, this article presents the use of machine learning techniques to classify employment status based on the database provided by the World Bank on financial inclusion (Global Findex). After data cleaning, a comparison was made between the techniques of Naïve Bayes, Logistic Regression, Random Forest, and k-Nearest Neighbors (KNN). The results were analyzed by accuracy, precision, and f-score metrics and validated with 10-fold cross validation. The results showed that accuracy and f-score were best for the KNN technique, while Naïve Bayes achieved the best precision but a low f-score, suggesting low sensitivity. Variables such as age, educational level, receipt of salaries and pensions, and salary amount were highlighted as good predictor variables according to the Random Forest. In this context, it was suggested to investigate these variables later for a more in-depth analysis of the profile of employed and unemployed citizens based on these predictors.

Keywords: Machine Learning. Financial inclusion. Employment.

1 INTRODUÇÃO

A mineração de dados e a extração de informação buscam coletar, gerenciar, processar, analisar e visualizar uma grande quantidade de dados (estruturados ou não estruturados), com um interesse maior no século XXI (LIU et al., 2019).

Atualmente, todos os sistemas automatizados acabam gerando dados, seja para análise ou diagnóstico, podendo até chegar a uma armazenagem na ordem de petabytes ou ainda exabytes (AGGARWAL, 2015).

Com a alta volatilidade tecnológica e a quantidade de informações que podem ser extraídas, é imprescindível o uso de classificadores automáticos para a obtenção de informações assertivas como auxílio na tomada de decisões estratégicas (NETA et al., 2018).

Estes classificadores pertencem a uma grande área das tarefas da mineração de dados conhecida como “preditiva”, mais especificamente na “classificação” (BENIWAL; ARORA, 2012). A classificação é uma tarefa que requer o aprendizado supervisionado, ou seja, precisa de um rótulo pré-definido na etapa de treinamento, buscando predizer os rótulos de instâncias ainda não vistos (RASCHKA, 2015). Muitos métodos de aprendizado são relevantes e úteis em várias áreas, incluindo disciplinas acadêmicas ou não, indo além da estatística tradicional (JAMES et al., 2017).

Existem várias aplicações hoje de técnicas de mineração de dados, incluindo as áreas de diagnóstico médico, desenvolvimento de produto, carga energética, fraudes em cartões de crédito, imagens de satélites, entre outros (BRAMER, 2016).

Para o setor financeiro, uma quantidade significativa de trabalhos têm sido publicada na literatura internacional, principalmente no mercado de ações (HENRIQUE; SOBREIRO; KIMURA, 2019) e análise de crédito (FENERICH et al., 2020).

O setor financeiro é considerado um item importante ao sistema econômico, sendo que uma das áreas que pode ser explorada neste componente é a inclusão financeira. Esta inclusão faz parte da estratégia de várias economias globais para desenvolver os setores econômicos e financeiro,

providenciando o acesso a estes serviços, porém elas ainda têm dificuldades em aumentar esta inclusão (MORGAN; PONTINES, 2018).

Com intuito de acompanhar a progressão desta evolução e facilitar futuras tomadas de decisão, o Banco Mundial construiu uma base de dados que reúne um conjunto abrangente de informações sobre como as pessoas em 144 países utilizam seus serviços financeiros. Esta base possui informações com mais de 5.000 projetos em países em desenvolvimento, de média e baixa renda. Envolve setores como “transporte, energia, telecomunicações, água e saneamento, em seus aspectos contratuais, fontes e fluxos de investimentos e principais investidores”. Ela é baseada nas informações de uma perspectiva “micro” e possui a limitação de ser publicada a cada três anos (MALIK et al., 2021).

Neste contexto, o presente artigo tem como objetivo realizar uma comparação entre cinco técnicas de mineração de dados utilizados para classificar a situação de pessoas quanto à situação empregatícia, a partir de um conjunto de dados de inclusão financeira. Os algoritmos de *machine learning* aplicados foram Naïve Bayes, *k*-vizinhos mais próximos (*k-nearest neighbors* – KNN), Regressão logística e Floresta aleatória.

Após esta seção introdutória, o restante do artigo está dividido em mais quatro seções. A Seção 2 discorre sobre os algoritmos de *machine learning*, a validação *k-fold* e as métricas de classificação. Na seção 3, tem-se os passos realizados para o desenvolvimento da pesquisa, bem como a descrição do conjunto de dados utilizado. A Seção 4 mostra os resultados obtidos e as discussões, enquanto a Seção 5 faz uma breve conclusão, indicando trabalhos futuros na linha pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, são descritos os conceitos básicos de mineração de dados, as tarefas mais comuns, as técnicas de mineração de dados e as métricas utilizadas na avaliação.

2.1 Mineração de dados

A mineração de dados é uma etapa fundamental do Processo de Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases* – KDD), proposto por (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), sendo que o propósito é encontrar informações não triviais em bancos de dados (BRAMER, 2016).

A partir do objetivo a ser atingido, verifica-se então qual é o tipo de aprendizado a ser aplicado no conjunto de dados disponível. Comumente, utilizam-se dois tipos de aprendizado: não supervisionado, ou supervisionado (SHALEV-SHWARTZ; BEN-DAVID, 2014).

A aprendizagem não supervisionada busca entender as relações existentes entre as instâncias, baseadas em um conjunto de dados de entrada, sem uma variável (ou rótulo) específico. As tarefas de associação e de agrupamento são bastante comuns neste contexto (TAN et al., 2018).

Para os algoritmos que são aplicados no aprendizado supervisionado, é necessário que exista uma variável de interesse, identificado como “rótulo”. Este rótulo pode ser uma variável do tipo categórica nominal ou ordinal (tarefa de classificação) ou uma saída numérica, real ou discreta, a qual pode ser estimada após aplicação do algoritmo (tarefa de regressão, ou estimação) (TAN et al., 2018).

Para o presente trabalho, foi utilizada a tarefa de classificação, que exige que os algoritmos aplicados tenham uma referência de rótulo para que se possa fazer o treinamento, sendo testado em seguida em um outro conjunto com rótulos desconhecidos (teste). Esta pesquisa fez o uso das técnicas Naïve Bayes, KNN, Regressão logística e Floresta aleatória.

2.2 Mineração de dados

Atualmente, a tarefa de classificação é uma das mais utilizadas seja para a classificação baseada em dados de corte transversal, imagens, ou texto (RASCHKA, 2015). Alguns exemplos desta tarefa são: (i) classificação de *e-mails* como *spam* ou não; (ii) tipos de tumor utilizando imagens (benigno ou maligno); (iii) presença ou não de coronavírus em pessoas; (iv) classificação de produto em conformidade ou não.

2.2.1 Naïve Bayes

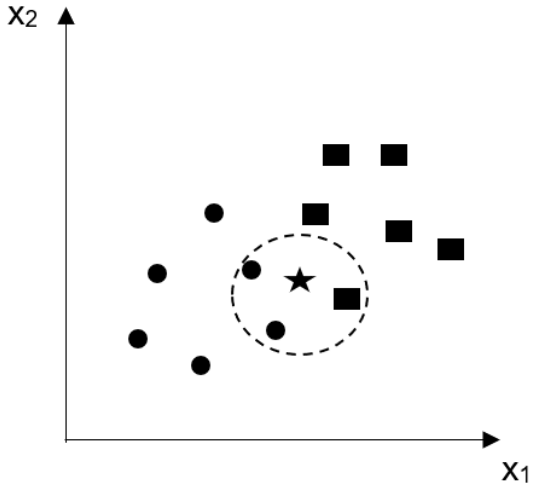
Naïve Bayes é amplamente utilizado para a tarefa de classificação e é baseado no Teorema de Bayes, ou seja, assume a independência entre todos os atributos e todos tem a mesma contribuição para a classe (MODU et al., 2017). A Equação (1) mostra a probabilidade em torno do teorema, onde P(A) e P(B) são probabilidades dos eventos A e B, de forma independente. P(A|B) é a probabilidade da condicional A sobre B e P(B|A) é a probabilidade da condicional B sobre A. A letra A indica as classes do atributo de saída, enquanto B são os atributos preditores (BERRAR, 2018).

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

2.2.1 *k*-vizinhos mais próximos

A técnica k-vizinhos mais próximos (KNN) é considerada um *lazy learner*, ou seja, ela não treina na etapa em que geralmente se treina um algoritmo de aprendizado supervisionado. O funcionamento dele está baseado nos *k*-vizinhos mais próximos de uma determinada instância de teste avaliada. Esta distância é calculada para todas as instâncias de treino em relação ao elemento de teste, classificando este elemento com a classe majoritária (OLIVEIRA et al., 2017). A Figura 1 ilustra o KNN, com um parâmetro $k = 3$.

Figura 1 - Representação da área de busca do KNN para um *k* = 3



Fonte: Autores (2023)

Nota-se na Figura 1 que, para uma nova instância “estrela” ainda não vista, ela possui uma frequência da classe dos seus vizinhos mais próximos sendo do tipo “círculo”, maior do que “quadrado”. Assim, a instância nova de teste deve ser também classificada como um “círculo”.

2.2.3 Regressão Logística

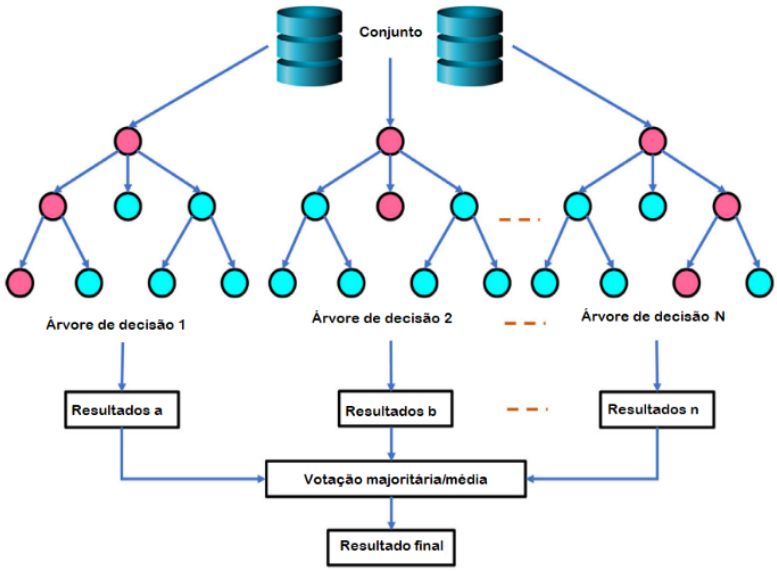
Esta técnica é considerada de bom desempenho em conjuntos de dados linearmente separáveis (RASCHKA, 2015). Diferentemente da regressão linear, que é utilizado para estimar valores contínuos, a regressão logística é um algoritmo típico de classificação que usa a função sigmóide para transformar uma saída em duas classes discretas, 0 ou 1 (LIU et al., 2021). A função sigmóide (ou logística), pode ser definida de acordo com a Equação (2).

$$logistic(y) = \frac{1}{1 + e^{-y}} \quad (2)$$

2.2.4 Floresta Aleatória

A Floresta aleatória (ou, do termo em inglês *Random forest*) foi proposta por Breiman (2001) e é uma evolução das árvores de decisão. Esta técnica é considerada um classificador *ensemble*, pois utiliza mais que um classificador para um modelo, construindo várias árvores de decisão com diferentes subconjuntos em cada árvore (KASTHURIRATHNE et al., 2016). Uma representação da Floresta aleatória pode ser vista na Figura 2.

Figura 2 – Representação de uma Floresta aleatória



Fonte: Liu et al. (2021, tradução livre)

2.2.5 Métricas de avaliação

As métricas de avaliação para a classificação geralmente são derivadas de uma matriz conhecida como “Matriz de confusão”.

A Tabela 1 mostra a característica de uma matriz de confusão, enquanto as Equações (3) a (6) define o cálculo das métricas de acuracidade, sensibilidade (ou *recall*), precisão e *f-score*, respectivamente.

Tabela 1 – Matriz de confusão

Classe real	**Classe predita**	
	SIM	NÃO
SIM	*VP*	*FN*
NÃO	*FP*	*VN*

Fonte: Adaptado de Witten e Frank (2005)

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN} \quad (3)$$

$$Sensibilidade = \frac{VP}{VP + FN} \quad (4)$$

$$Precisão = \frac{VP}{VP + FP} \quad (5)$$

$$f\text{-}score = \frac{2 \times recall \times precisão}{recall + precisão} \quad (6)$$

Onde *VP* diz respeito aos verdadeiros positivos, *VN* aos verdadeiros negativos, *FP* se refere aos falsos positivos e, finalmente, *FN* aos falsos negativos.

3 MATERIAIS E MÉTODOS

Nesta seção, serão apresentas as descrições sobre o conjunto de dados pesquisado, as ferramentas utilizadas para a implementação dos algoritmos, bem como a sequência das etapas implementadas no desenvolvimento.

3.1 Descrição do conjunto de dados

O conjunto de dados *World - Global Financial Inclusion* (2017) originalmente possui aproximadamente 150.000 instâncias. Mais de 200 atributos como país, sexo, idade, situação empregatícia e outros indicadores sobre contas, pagamentos, economias, crédito e resiliência financeira, foram coletados por meio de um questionário aplicado em mais de 144 países. A versão mais recente do estudo é de 2017 e está disponível no *website* do Banco Mundial (BANCO MUNDIAL, 2017).

Para esta pesquisa, foram utilizadas as variáveis que dizem respeito às questões de sociodemográficas e indicadores de inclusão econômica, totalizando 31 variáveis (após o pré-processamento), incluindo a classe de interesse da situação empregatícia, onde 1 = “empregado”, enquanto 0 = “desempregado”.

3.2 Etapas para o desenvolvimento da pesquisa

As etapas realizadas para o desenvolvimento desta pesquisa podem ser resumidas da seguinte forma:

i. Exclusão de instâncias e variáveis com dados faltantes;
ii. Transformação e padronização (utilizando o *z-score* para este último caso);
iii. Verificação do balanceamento de classes;
iv. Aplicação dos algoritmos de classificação;
v. Análise das métricas de classificação.

Para a validação da classificação, em todos os algoritmos testados teve a divisão dos conjuntos baseados na validação cruzada *k-fold*, com um valor de $k = 10$. Esta validação visa fazer ajustes de parâmetros, e, também, verificar a estabilidade dos modelos quando treinados e testados com conjuntos diferentes. Informações sobre esta validação pode ser vista em Berrar (2019).

Para a manipulação do conjunto de dados, visualização e aplicação dos modelos, todas as etapas foram desenvolvidas em linguagem de programação

Python usando as bibliotecas *scikit-learn* (*machine learning* e pré-processamento*)*, *pandas* (manipulação de conjunto de dados chamados de *dataframes*) e *seaborn* (visualização de dados). Para os parâmetros dos modelos, após algumas análises empíricas, optou-se por utilizar o padrão do Naïve Bayes, Regressão logística com 500 iterações, o valor de *k* o KNN foi de 50 vizinhos, e 100 árvores de decisão foram implementadas para a Floresta aleatória.

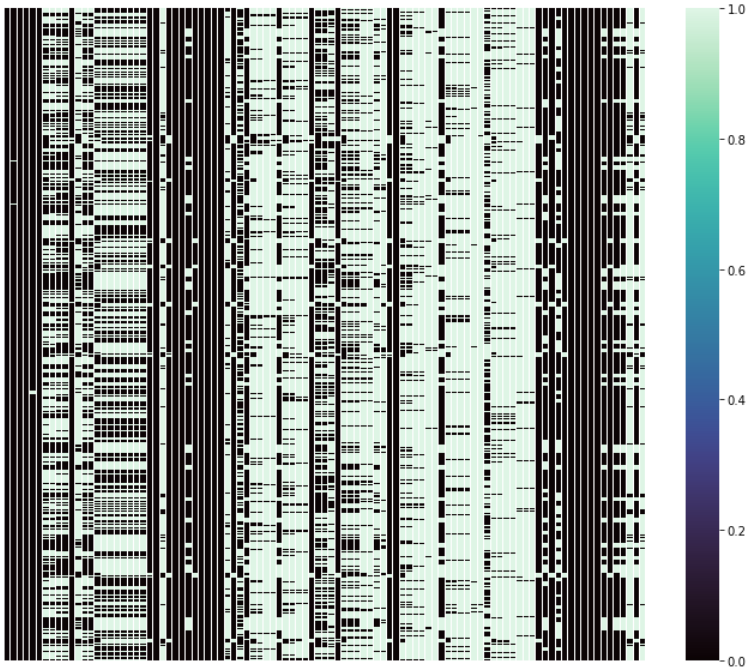
4 RESULTADOS E DISCUSSÃO

Nesta seção, são descritos os resultados obtidos, desde o pré-processamento até as métricas de avaliação. Na sequência, tem-se as discussões acerca dos resultados encontrados e suas implicações.

4.1 Etapas para o desenvolvimento da pesquisa

Para a classificação, primeiramente foi feita uma análise sobre os dados faltantes, como ilustrado na Figura 3.

Figura 3 - Frequência de dados faltantes com o conjunto de dados original



Fonte: Autores (2023)

A Figura 3 representa a matriz de dados, no qual as linhas são as instâncias e as colunas os atributos. Para melhorar a qualidade da imagem, optou-se por retirar qualquer nome de coluna e identificação de linha.

As cores em tom “azul claro” indicam que o valor é 1 e a condição é “Verdadeira”, ou seja, esta célula é um dado faltante do conjunto. Onde há a cor “preta” na matriz denotam dados válidos. O total de dados faltantes deste conjunto original foi de 7.222.617 dados, correspondendo a 47,1% da matriz. Como a maioria dos dados faltantes pertenciam a colunas (variáveis) específicas, então os dados que possuíam até 10% de dados faltantes foram excluídos. Na sequência, foram retiradas as instâncias com qualquer dado faltante, resultando em 144.593 linhas e 31 colunas. Quanto ao balanceamento das classes, a Tabela 2 mostra a distribuição daqueles que estão empregados e dos desempregados.

Tabela 2 – Frequência de instâncias para as classes de empregados e desempregados

Classes	Frequência	Total
Empregados	91.003	144.593
Desempregados	53.590	

Fonte: Autores (2023)

Verifica-se pela Tabela 2 que a frequência relativa de pessoas empregadas e desempregadas são, respectivamente, 62,9% e 37,1%. Como existem mais de 140 mil instâncias no conjunto completo, não se viu uma necessidade de balancear o conjunto de dados para equilibrar as duas classes.

4.2 Avaliação das métricas de classificação

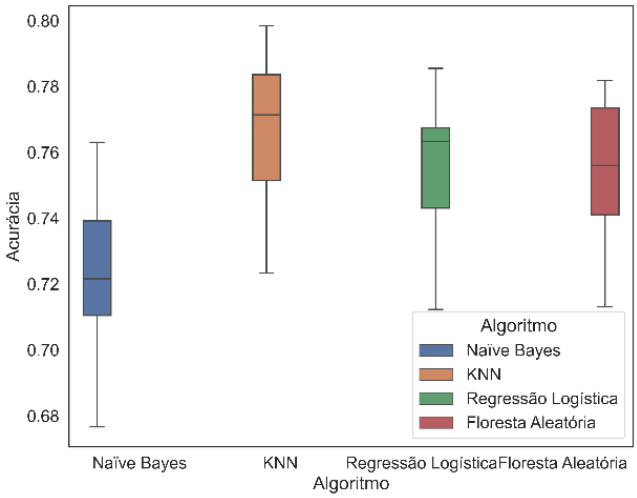
Os resultados das métricas, com citado anteriormente, foram a acurácia, precisão e *f-score*. A acurácia mostra o total de acertos do classificador, sem ter alguma categoria de interesse.

Já a precisão indica o quão bom o modelo é para classificar uma categoria de interesse a partir dos seus falsos positivos e verdadeiros positivos. Para este caso, verificou-se o quão bom o modelo é para classificar aqueles

que estão empregados em relação a todas as instâncias que foram classificadas pelo modelo como empregados.

Por fim, o *f-score* é a média harmônica entre a precisão e a sensibilidade (mostra o quão bom o modelo é para classificar os empregados). Isto é importante pois estabelece um equilíbrio entre estas duas métricas que podem levar a alguma conclusão equivocada sobre uma classe de interesse. As Figuras 4, 5 e 6 mostram os resultados de cada uma das métricas, estratificadas por técnica.

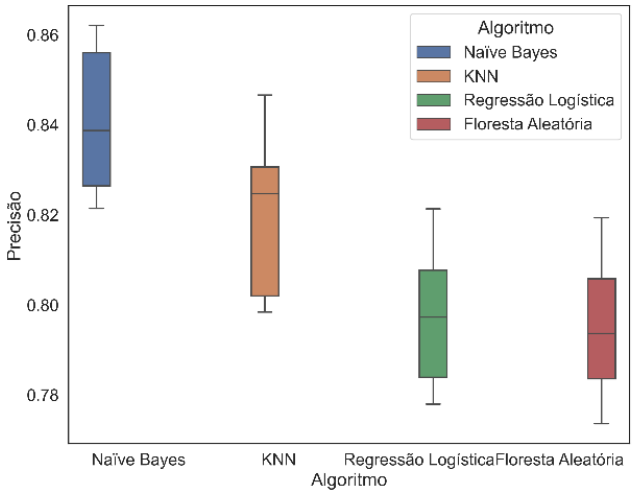
Figura 4 – Resultados dos modelos para a acurácia



Fonte: Autores (2023)

Verifica-se na Figura 4 que o KNN foi o algoritmo que obteve a melhor mediana da acurácia (77,1%), seguido de Regressão logística, Floresta aleatória e Naïve Bayes, com valores de 76,3%, 75,6% e 72,2%, respectivamente. Percebe-se que a variação entre os mínimos e máximos fica em torno de 8% para a validação utilizada, mostrando uma variação não tão baixa quanto a esperada.

Figura 5 – Resultados para a precisão

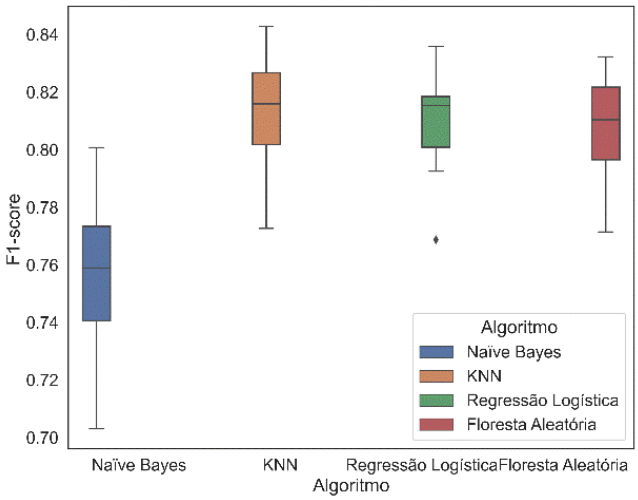


Fonte: Autores (2023)

A Figura 6 destaca que o algoritmo Naïve Bayes obteve a melhor precisão entre todos os métodos, com uma mediana de 83,9%, seguido de KNN, Regressão Logística e Floresta aleatória. Isto mostra que, de todas as instâncias classificadas como pessoas empregadas pelo modelo, obteve-se este acerto (em %). Esta métrica é importante para avaliar se o modelo não identifica muitos falsos positivos, porém não se deve fazer nenhuma conclusão apenas a tendo como referência.

Existe então o *f-score*, que mostra o equilíbrio entre a precisão e o *recall*, constata que Naïve Bayes obteve o *f-score* mais baixo (Figura 6), ou seja, uma sensibilidade bastante inferior deve ter sido alcançada neste modelo para gerar tal valor.

Figura 6 – Resultados para o *f-score*



Fonte: Autores (2023)

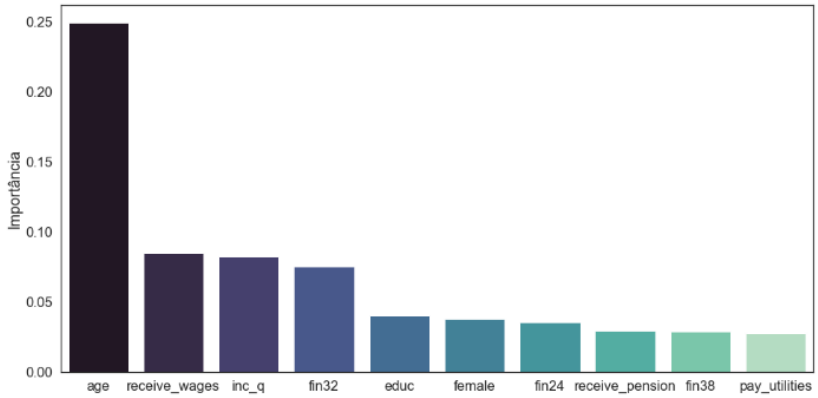
Por fim, a Figura 6 mostra que o *f-score* do KNN, Regressão logística e Floresta aleatória possuem praticamente a mesma mediana, com valores muito próximos a 81%. Isto mostra um equilíbrio entre a sensibilidade e a precisão principalmente para as duas últimas, pois obtiveram também valores muito próximos de precisão (vide Figura 5).

Quanto aos resultados dos modelos computacionais aplicados, o KNN se mostrou uma técnica promissora, com melhores acurácia e *f-score*. Estes resultados surpreendem, pois o KNN pode sofrer de *overfitting* para modelos com muitos atributos (RASCHKA, 2015). Também, esperava-se um melhor desempenho para a Floresta aleatória, pois a seleção interna de atributos, a não necessidade de presumir linearidade dos modelos e a utilização de atributos preditores que são altamente correlacionados (HE et al., 2018), tornam esta técnica bastante versátil.

Vale ressaltar que a Regressão logística possui a vantagem de estar atrelada em alguns modelos com a regularização, o que permite evitar possíveis *overfitting* nos modelos criados (AGGARWAL, 2018). Entretanto, para este estudo, Regressão logística e Floresta aleatória não foram superiores ao KNN, como esperado.

Como complemento aos resultados obtidos pela classificação, verificou-se também pela Floresta aleatória a importância das variáveis. Isto pode ser feito com a análise do decrescimento médio da impureza (RASCHKA, 2015). A Figura 7 mostra as variáveis mais importantes, na ordem decrescente.

Figura 7 – Importância das variáveis pelo método *feature importances* da Floresta aleatória

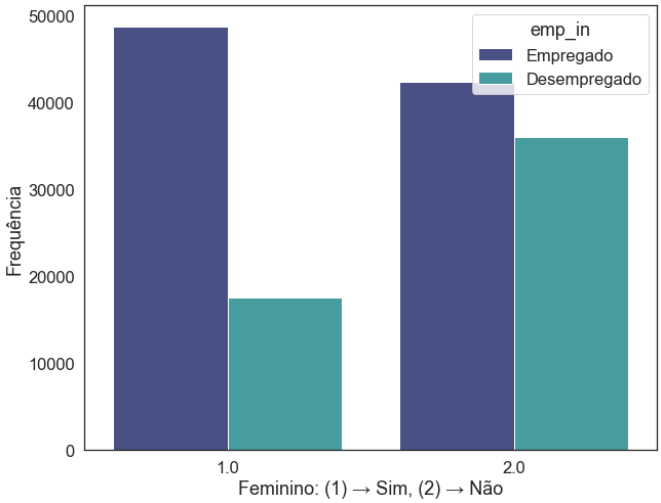


Fonte: Autores (2023)

De acordo com a Figura 7, nota-se que a idade, é o fator mais importante, seguido de recebimento de salários ou de alguma renda nos últimos 12 meses, recebimento de salários via empregador (último ano), nível educacional, e se é mulher. Muitas destas variáveis estão relacionadas diretamente com o recebimento de algum tipo de renda ou pensão e estas se mostraram boas preditoras para identificar qual é a situação empregatícia de um indivíduo.

Observando-se de forma separada variáveis demográficas, como a idade e se é do sexo feminino, pode-se notar alguns pontos pertinentes. Por exemplo, na Figura 8 nota-se que mais de 70% das mulheres estão empregadas e, este fator, pode ser um ponto decisivo do modelo entender como o sexo pode influenciar na classificação de uma instância em que a resposta é “empregado”.

Figura 8 – Quantidade de mulheres e homens quanto à situação empregatícia



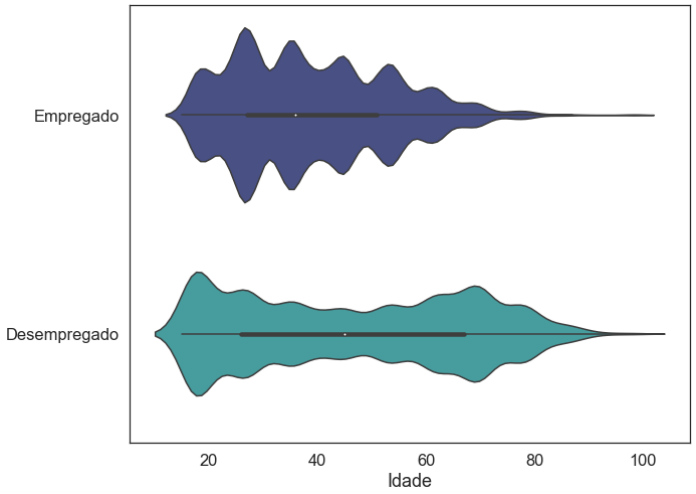
Fonte: Autores (2023)

Sob uma perspectiva mais descritiva, ainda se tem uma proporção menor de mulheres empregadas ou de força de trabalho em relação aos homens no mundo.

De acordo com o *International Labour Organization* (ILO, 2022), ainda há uma diferença de 25 pontos percentuais entre homens e mulheres, sendo que estas correspondem próximo a 47% de toda a força de trabalho. Entretanto, para esta base, verifica-se uma maior quantidade de mulheres empregadas quando comparado aos homens e isto é decisivo de acordo com o modelo computacional.

A Figura 9 ilustra a distribuição da idade em relação à situação empregatícia.

Figura 9 – Distribuição da idade estratificada de acordo com a situação empregatícia



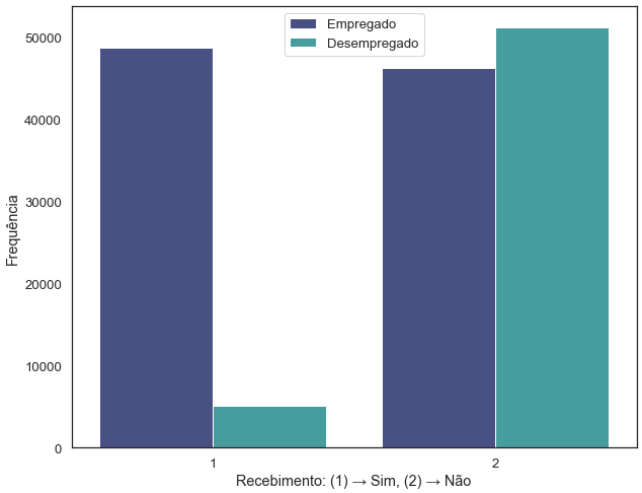
Fonte: Autores (2023)

Nota-se que, na Figura 9, existe uma concentração mais alta de empregados próximo aos 30 anos e com uma mediana próxima aos 40 anos de idade, situação comum em vários países, pois muitos estão estudando com uma idade em torno de 20 a 25 anos e entram na aposentadoria após 60 anos. Esta situação é diferente para as pessoas desempregadas, com uma concentração próxima aos 20 e depois fica mais largo novamente em torno dos 70 anos, próximo à aposentadoria.

De fato, como mostrado em um levantamento da literatura realizada por Lange et al. (2021), estudos apontaram uma correlação negativa entre o avanço da idade e a empregabilidade, corroborando com o que foi encontrado na Figura 9, especificamente no gráfico dos empregados. Entretanto, também foram identificados nos estudos que há múltiplos fatores que podem estar negativamente relacionados com a empregabilidade, porém associados com a idade, por exemplo, flexibilidade do trabalho, reconhecimento social, habilidades ocupacionais, entre outros (LANGE et al., 2021).

A Figura 10 mostra a quantidade de pessoas que receberam algum tipo de remuneração ou salário nos últimos 12 meses de algum empregador (variável “Fin32”).

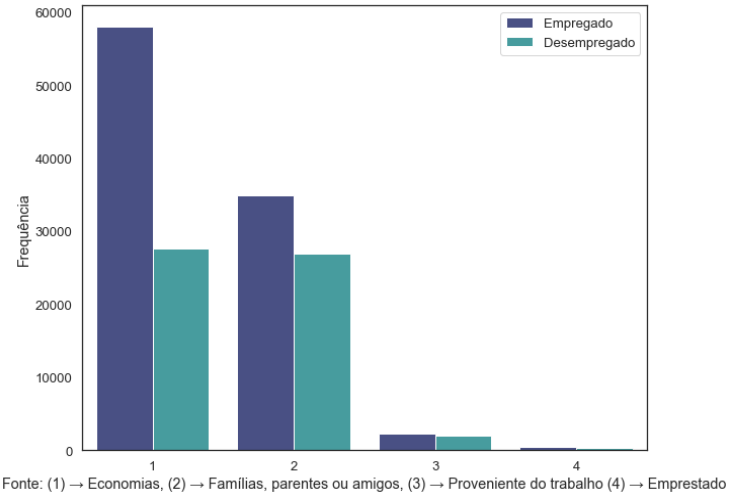
Figura 10 – Frequência do recebimento de salários ou remuneração de um empregador nos últimos 12 meses



Fonte: Autores (2023)

Percebe-se pela Figura 10 que a maioria das pessoas que receberam algum tipo de remuneração ainda estava empregada durante a aplicação do questionário, mostrando que o algoritmo provavelmente identifica categoria como essencial para conseguir entender o padrão de cidadão empregados. Já a Figura 11 ilustra a maneira a fonte de pagamento em casos de emergências.

Figura 11 – Fontes de pagamentos em casos de emergência



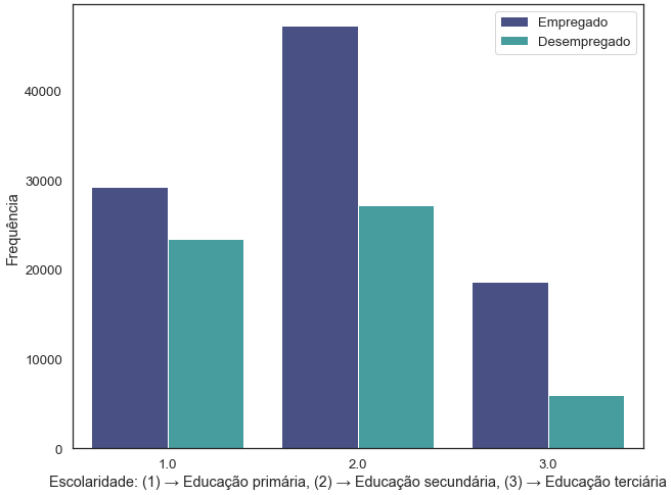
Fonte: Autores (2023)

A Figura 11 mostra que as pessoas empregas, em sua grande maioria, tem o hábito de guardar/economizar dinheiro de seus salários ou remuneração,

tendo um fundo de reserva para qualquer emergência, como geralmente consultores financeiros têm indicado atualmente.

Por fim, a Figura 12 retrata a situação empregatícia de acordo com o nível de escolaridade.

Figura 12 – Nível de escolaridade estratificado por situação empregatícia



Fonte: Autores (2023)

Percebe-se pela Figura 12 que à medida que se avança o nível de escolaridade, o modelo computacional consegue entender que há uma predominância das pessoas que estão empregadas. Desta forma, vai ao encontro à hipótese que um maior nível de escolaridade, tem-se mais oportunidades de empregos.

Para fechar a discussão e contribuições de artigo, entende-se que os modelos de *machine learning* podem contribuir de diferentes formas para dados referentes à pesquisa *Global Findex*:

- O modelo computacional consegue ter uma predição razoavelmente bem a situação empregatícia de cidadãos, principalmente analisando características demográficas, como idade, sexo e nível de escolaridade;
- Estes modelos podem ajudar a entender melhor as características relacionadas à inclusão financeira, analisando de formas distintas os cidadãos empregados ou desempregados;

- O modelo preditivo é uma contribuição empírica e prática, pois os algoritmos aqui investigados são bastante conhecidos na literatura, mas ainda com relevância em diversas áreas do conhecimento.

5 CONCLUSÕES

Esta pesquisa teve como objetivo comparar cinco técnicas de mineração de dados com o intuito de classificar se as pessoas estavam empregadas ou desempregadas. Para isso, foi utilizado um conjunto de dados sobre inclusão financeira "*Global Findex*", um estudo recorrente do Banco Mundial com mais de 150 mil pessoas envolvidas e de várias economias mundiais.

O estudo mostrou que o KNN foi a melhor técnica, seguido por Regressão logística, Floresta aleatória e *Naïve Bayes*. A mediana da acurácia do KNN chegou a 77,1%, porém com um *f-score* de aproximadamente 81% para KNN, Regressão logística e Floresta aleatória.

Também foi feito uma breve análise sobre as variáveis mais importantes utilizando a Floresta aleatória, destacando-se variáveis demográficas (idade, nível educacional e sexo), ao passo que variáveis específicas como recebimentos de salários, pensão e quantia salarial foram os fatores que o modelo destacou como as que mais contribuíam para a tarefa de classificação.

Estudos futuros podem incluir uma variação de técnicas de mineração de dados que não foram exploradas neste artigo, como Redes neurais artificiais, *Support Vector Machines* (SVM) ou outras técnicas do tipo *ensemble*, como o *Gradient Boosting Machines* (GBM) ou *XGBoost*. Estas técnicas possuem uma boa generalização e se encaixam no perfil do conjunto de dados com variáveis mistas no contexto da inclusão financeira. Outra abordagem que também poderia ser explorada seria a utilização de técnicas de seleção de variáveis a partir da estatística, como a correlação de *Pearson* (variáveis numéricas) ou coeficiente *phi* e teste qui-quadrado (variáveis dicotômicas). Estas técnicas são bastante apropriadas para a configuração desta base e podem reduzir a complexidade dos modelos computacionais, buscando melhores métricas de classificação.

REFERÊNCIAS

AGGARWAL, C. C. **Data Mining**. Cham: Springer International Publishing, 2015. DOI 10.1007/978-3-319-14142-8. Available at: http://link.springer.com/10.1007/978-3-319-14142-8_10.

AGGARWAL, C. C. **Neural Networks and Deep Learning: a Textbook**. Cham: Springer Publishing, 2018. https://doi.org/10.1007/978-3-319-94463-0

BANCO MUNDIAL. The Global Findex Database 2017. 2017**. Global Findex**. Available at: https://globalfindex.worldbank.org/. Accessed on: 20 Nov. 2020.

BENIWAL, S.; ARORA, J. Classification and Feature Selection Techniques in Data Mining**. International Journal of Engineering Research and Technology in Data Mining**, vol. 1, no. 6, p. 1–6, 2012.

BERRAR, D. Bayes' Theorem and Naive Bayes Classifier. **Encyclopedia of Bioinformatics and Computational Biology**, vol. 1, p. 403–412, 2018. https://doi.org/10.1016/b978-0-12-809633-8.20473-1.

BERRAR, D. Cross-validation. **Encyclopedia of Bioinformatics and Computational Biology**, vol. 1, p. 542–545, 2019. DOI 10.1177/0146621613518094. Available at: http://arxiv.org/abs/1703.03167.

BRAMER, M. **Principles of Data Mining**. 3rd ed. London: Springer London, 2016(Undergraduate Topics in Computer Science). DOI 10.1007/978-1-4471-7307-6.

BREIMAN, L. Random forests. **Machine Learning**, vol. 45, p. 5–32, 2001. https://doi.org/10.1007/9781441993267_5.

DE LANGE, A. H.; VAN VUUREN, T.; FURUNES, T.; DE LANGE, C.; DIKKERS, J. Employable as We Age? A Systematic Review of Relationships Between Age Conceptualizations and Employability**. Frontiers in Psychology**, v. 11, 2021. https://doi.org/10.3389/fpsyg.2020.605684.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases**. AI Magazine**, vol. 17, no. 3, p. 37–53, 1996. https://doi.org/10.1609/aimag.v17i3.1230.

FENERICH, A.; STEINER, M. T. A.; STEINER NETO, P. J.; TOCHETTO, E.; TSUTSUMI, D.; ASSEF, F. M.; DOS SANTOS, B. S. Use of machine learning techniques in bank credit risk analysis**. Revista Internacional de Metodos Numericos para Calculo y Diseno en Ingenieria**, vol. 36, no. 3, p. 1–15, 2020. https://doi.org/10.23967/J.RIMNI.2020.08.003.

HE, L.; LEVINE, R. A.; FAN, J.; BEEMER, J.; STRONACH, J. Random Forest as a Predictive Analytics Alternative to Regression in institutional Research. **Practical Assessment, Research & Evaluation**, v. 23, n. 1, 2018. https://doi.org/10.7275/1WPR-M024.

HENRIQUE, B. M.; SOBREIRO, V. A.; KIMURA, H. Literature review : Machine learning techniques applied to financial market prediction. **Expert Systems With Applications**, vol. 124, p. 226–251, 2019. DOI 10.1016/j.eswa.2019.01.012. Available at: https://doi.org/10.1016/j.eswa.2019.01.012.

INTERNATIONAL LABOUR ORGANIZATION. The gender gap in employment: What's holding women back? 2022. Disponível em: < https://www.ilo.org/infostories/en-GB/Stories/Employment/barriers-women#intro>. Acesso em: 13 abril 2023.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning with applications in R**. 8 ed. New York: Springer Science, 2017. https://doi.org/10.1007/978-1-4614-7138-7.

KASTHURIRATHNE, S. N.; DIXON, B. E.; GICHOYA, J.; XU, H.; XIA, Y.; MAMLIN, B.; GRANNIS, S. J. Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection**. Journal of Biomedical Informatics**, vol. 60, p. 145–152, Apr. 2016. DOI https://doi.org/10.1016/j.jbi.2016.01.008.

LIU, J.; KONG, X.; ZHOU, X.; WANG, L.; ZHANG, D.; LEE, I.; XU, B.; XIA, F. Data Mining and Information Retrieval in the 21st century: A bibliographic review. **Computer Science Review**, vol. 34, 2019. https://doi.org/10.1016/j.cosrev.2019.100193.

LIU, Y.; ESAN, O. C.; PAN, Z.; AN, L. Machine learning for advanced energy materials. **Energy and AI**, vol. 3, 2021. https://doi.org/10.1016/j.egyai.2021.100049.

MALIK, A. H.; MD ISA, A. H. bin; JAIS, M. bin; UR REHMAN, A.; ALI KHAN, M. Financial stability of Asian Nations: Governance quality and financial inclusion. **Borsa Istanbul Review**, vol. No prelo, 2021. DOI 10.1016/j.bir.2021.05.005. Available at: https://doi.org/10.1016/j.bir.2021.05.005.

MODU, B.; POLOVINA, N.; LAN, Y.; KONUR, S.; TAUFIQ ASYHARI, A.; PENG, Y. Towards a predictive analytics-based intelligent malaria outbreakwarning system. **Applied Sciences (Switzerland)**, School of Electrical Engineering and Computer Science, University of Bradford, Bradford, United Kingdom, vol. 7, no. 8, 2017. DOI 10.3390/app7080836.

MORGAN, P. J.; PONTINES, V. Financial stability and financial inclusion: The case of SME lending. **The Singapore Economic Review**, vol. 63, no. 01, p. 111–124, 8 Mar. 2018. DOI 10.1142/S0217590818410035. Available at: https://www.worldscientific.com/doi/abs/10.1142/S0217590818410035.

OLIVEIRA, A.; FARIA, B. M.; GAIO, A. R.; REIS, L. P. Data Mining in HIV-AIDS Surveillance System: Application to Portuguese Data. **Journal of Medical Systems**, Center of Mathematics, University of Porto, Porto, Portugal, vol. 41, no. 4, 2017. DOI 10.1007/s10916-017-0697-4. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85013157852&doi=10.1007%2Fs10916-017-0697-4&partnerID=40&md5=614ecf46aaf68e64003a91805a6f0394.

RASCHKA, S. **Python Machine Learning**. 1st ed. Birmingham: Packt Publishing Ltd., 2015.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding Machine Learning: From Theory to Algorithms**. 1st ed. New York: Cambridge University Press, 2014.

TAN, P.-N.; STEINBACH, M.; KARPATNE, A.; KUMAR, V. **Introduction to Data Mining**. 2nd ed. [S. l.]: Pearson Prentice Hall, 2018.

WITTEN, I. H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. 2nd ed. San Francisco, CA, E.U.A.: Elsevier Ltd, 2005.