https://doi.org/10.47456/bjpe.v8i1.37019

NÚCLEO DE PESQUISA EM GESTÃO DE SISTEMAS DA PRODUÇÃO - NP GSP

Brazilian Journal of Production Engineering
BJPE - Revista Brasileira de Engenharia de Produção
Disponível em: http://periodicos.ufes.br/BJPE/index
ISSN: 2447-5580

Campus São Mateus
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

Pró-Reitoria de Pesquisa e Pós-Graduação
Universidade Federal do Espírito Santo

ARTIGO ORIGINAL OPEN ACCESS

UTILIZAÇÃO DE TÉCNICAS DE CLASSIFICAÇÃO EM CONJUNTO DE DADOS SOBRE INCLUSÃO FINANCEIRA: UM ESTUDO BASEADO EM PAÍSES LATINOAMERICANOS

USE OF CLASSIFICATION TECHNIQUES IN FINANCIAL INCLUSION DATASET: A STUDY BASED IN LATIN AMERICAN COUNTRIES

Pâmela Rodrigues Venturini de Souza [1], Bruno Gigioli Tomazi [2], & Bruno Samways dos Santos [3*]

[1 2 3] Universidade Tecnológica Federal do Paraná - Campus Londrina.

[1] pamsou@alunos.utfpr.edu.br [2] brunotomazi@alunos.utfpr.edu.br [3*] brunosantos@utfpr.edu.br

ARTIGO INFO.

Recebido em: 06.11.2021

Aprovado em: 02.02.2022

Disponibilizado em: 14.02.2022

PALAVRAS-CHAVE:
Mineração de dados; Classificação; Inclusão financeira; América Latina.

KEYWORDS:
Data mining; Classification; Financial inclusion; Latin America.

***Autor Correspondente**: Santos, B. S., dos.

RESUMO

A inclusão financeira é importante para reduzir a pobreza e proporcionar um crescimento econômico inclusivo, principalmente comparando grupos com grande desigualdade social. Este artigo utilizou a pesquisa *Global Financial Inclusion* (*Global Findex*) da *World Bank Group* para comparar técnicas de aprendizado de máquina na classificação de homens e mulheres quanto ao uso de serviços financeiros. Para isso, utilizou-se os classificadores Árvore de decisão, *k*-vizinhos mais próximos, Naïve Bayes e Floresta randômica, e avaliadas as métricas de acurácia, precisão, sensibilidade, *F1-score* e área sob a curva *Receiver Operating Characteristic* (ROC). Verificou-se que todas as técnicas (exceto por Naïve Bayes) obtiveram uma acurácia próxima a 70%, sensibilidade próxima a 88% e precisão acima dos 72% na maioria dos parâmetros investigados. Quanto à área sob a curva ROC, a Floresta randômica atingiu 0,77, superando as outras técnicas nesta avaliação.

ABSTRACT

Financial inclusion is important to reduce poverty and provide inclusive economic growth, mainly comparing groups with great social inequality. This article used the Global Financial Inclusion (Global Findex) research of the World Bank Group to compare machine learning techniques in classifying men and women when making use of financial services. The classifiers Decision tree, k-nearest neighbors, Naïve Bayes, and Random forest were implemented and evaluated with accuracy metrics, precision, sensibility, F1-score, and area under the Receiver Operating Characteristics (ROC) curve. It was verified that all techniques (except for Naïve Bayes) reached an accuracy near to 70% and 88% for recall, and precision over 72% in most of the parameters studied. Concerning the area under the ROC curve, the Random forest reached 0,77, outperforming other techniques in this assessment.

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

1. INTRODUÇÃO

Os setores econômicos e financeiros de vários países são desenvolvidos pela estratégia baseada em acesso aos serviços financeiros formais, porém muitas nações ainda têm dificuldades em aumentar esta inclusão (Morgan & Pontines, 2018). A inclusão financeira em conjunto com um uso formal dos serviços financeiros é potencialmente benéfica, pois podem resolver parcialmente os problemas de informação associados àquelas pessoas que anteriormente não possuíam uma conta bancária.

Também, destaca-se que esta inclusão pode aumentar potenciais econômicos a partir do encorajamento de projetos para aumento da produtividade com o crédito de pequenas e jovens empresas, assim como de grupos excluídos (Marcelin, Egbendewe, Oloufade, & Sun, 2021). Desta forma, gestores da política pública têm desenvolvido estratégias para garantir que a inclusão financeira seja um tema prioritário, principalmente em países emergentes, devido à defasagem ao acesso a serviços financeiros quando comparados às economias mais avançadas (Abdul Razak & Asutay, 2022). A importância pode ser estendida ao debate de gênero, uma vez que ainda há uma lacuna entre homens e mulheres quanto ao acesso a uma conta bancária (Robino *et al*., 2018) e, com a geração e consumo de um volume significativo de dados (Rabelo, de Campos, & Silva, 2021), surge a oportunidade de se trabalhar com técnicas mais sofisticadas, além da estatística tradicional, como as baseadas em aprendizado de máquina.

O aprendizado de máquina (do termo inglês "*machine learning*") é um campo importante da grande área da inteligência artificial, no qual os algoritmos aprendem de acordo com as experiências e buscam predizer eventos futuros (Dogan & Birant, 2021). Com um crescimento significativo no século XXI, estas técnicas são geralmente aplicadas na fase da mineração de dados, na qual se busca coletar, gerenciar, processar, analisar e visualizar uma grande quantidade de dados estruturados ou não estruturados (Liu *et al*., 2019).

Muitos trabalhos têm sido feitos com técnicas de mineração de dados e, no lado financeiro, muitos são voltados aos investimentos, como nos exemplos de aplicação levantados na revisão de Henrique, Sobreiro e Kimura (2019), ou na concessão de crédito, como sumarizado por (Fenerich *et al.*, 2020). Entretanto, não se verificou trabalhos de classificação no contexto da inclusão financeira, principalmente com foco na realidade latino-americana.

Neste escopo, este artigo teve como objetivo comparar as técnicas de aprendizado de máquina, pesquisando sobre os países da américa latina, com o intuito de compreender o comportamento de homens e mulheres com a utilização dos serviços financeiros. Toda a pesquisa foi realizada a partir do conjunto de dados disponibilizado pelo *World Bank Group* a partir do *Global Findex*, discutido posteriormente neste artigo.

Após esta seção introdutória, o restante do artigo foi organizado em mais cinco seções. A segunda seção apresenta conceitos importantes sobre a mineração de dados e as técnicas aplicadas. A terceira seção descreve o conjunto de dados e, na quarta seção, os métodos e materiais utilizados no desenvolvimento. A quinta seção apresenta e discute os resultados, por meio de tabelas e visualização para uma melhor interpretação das comparações realizadas. Por fim, a sexta seção foi elaborada para conclusão e direções para trabalhos futuros.

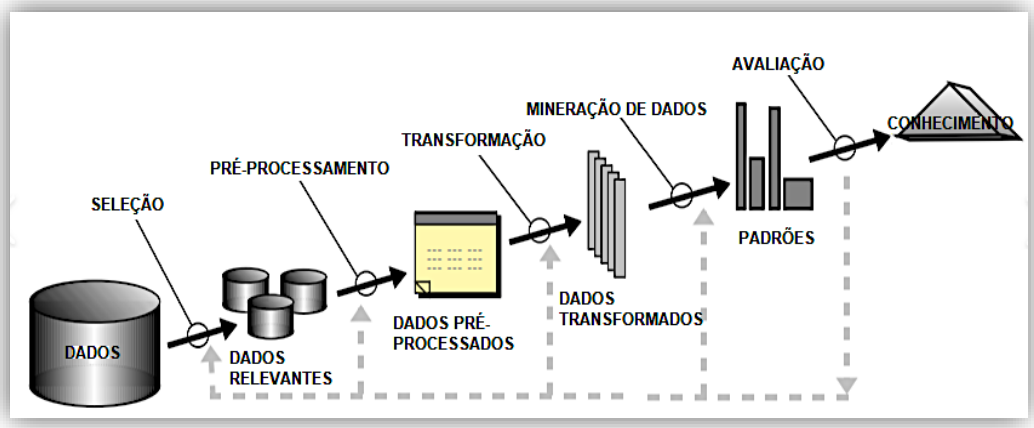
Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

A mineração de dados é parte de um processo de descoberta de conhecimento em base de dados (*Knowledge Discovery in Databases* – KDD). Fayyad, Piatetsky-Shapiro e Smyth (1996) define que a extração de conhecimento de base de dados é a identificação de padrões válidos, úteis e compreensíveis embutidos nos dados e inclui cinco etapas principais: (i) seleção; (ii) pré-processamento; (iii) transformação; (iv) mineração de dados, e; (v) avaliação de resultados (Figura 1).

Figura 1. Modelo tradicional de descoberta de conhecimento em base de dados



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996)

Primeiramente, é feita a seleção dos dados (também chamada de redução), que compreende, em essência, identificar quais informações devem ser efetivamente consideradas no processo de descoberta de conhecimento (Goldschmidt, Passos, & Bezerra, 2015). A base de dados bruta pode conter dados que não são alvos da análise, e dados redundantes.

Em seguida é necessário realizar um pré-processamento e limpeza a fim de assegurar a completude, veracidade e integridade dos dados. Informações ausentes, errôneas e inconsistentes devem ser corrigidas de forma a não comprometer a qualidade dos modelos (Goldschmidt, Passos, & Bezerra, 2015).

A transformação deve ser feita para quando houver necessidade de transformar dados numéricos em dados categóricos, ou transformar dados categóricos em dados numéricos. Camilo e Silva (2009) citam que diversas técnicas podem ser empregadas conforme a necessidade. Por exemplo, a suavização é utilizada para remover valores errados, enquanto que o agrupamento reúne valores em faixas sumarizadas. Já a generalização converte valores específicos para valores genéricos e a normalização padroniza as variáveis em uma mesma escala. Por fim, também pode ser feita a criação de atributos, quando necessário. Para esta pesquisa, foi realizada a transformação por binarização, um processo amplamente necessário para tratamento que precede tarefas de mineração de dados, convertendo cada um dos valores nominais de um determinado atributo categórico em valores booleanos (Masmoudi, Turkay, & Chabchoub, 2013).

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

A mineração precede a avaliação, onde os dados preparados são repassados a um algoritmo para produzir uma determinada saída em forma de regras ou outros tipos de "padrões" (Bramer, 2016). Esta etapa será discutida com maior profundidade na Subseção 2.2 a seguir.

Como última fase do KDD, tem-se a avaliação do modelo e dos resultados, em que os especialistas de dados interpretam e geram conhecimento potencialmente útil. Diversas ferramentas gráficas podem ser utilizadas para auxiliar na visualização e é importante validar o modelo com o apoio de indicadores como acurácia e sensibilidade, visando obter a confiabilidade na análise (Camilo & Silva, 2009).

2.2 MINERAÇÃO DE DADOS

A mineração de dados é um campo interdisciplinar que combina *machine learning*, reconhecimento de padrões, estatística, base de dados e visualização, com objetivo de extrair informação de grandes conjuntos de dados. Está relacionada à aplicação de tarefas, que pode ser: agrupamento, descrição, regressão, previsão, classificação e associação, e devem ser definidas de acordo com o objetivo e as características dos dados (Larose & Larose, 2014). Estas tarefas comumente podem ser baseadas em técnicas de aprendizado supervisionado ou não supervisionado.

No caso das tarefas com técnicas de aprendizado supervisionado, existe uma classe (rótulo) ou atributo que se pode comparar e validar o resultado, como por exemplo a tarefa de classificação. A tarefa não supervisionada é aquela que não existe uma classe ou rótulo prévio, como por exemplo a tarefa de agrupamento (Amaral, 2016).

É importante destacar que a tarefa está de acordo com o objetivo da mineração de dados, entendendo-se quais são as características dos dados e verificar qual é o tipo de aprendizado a ser aplicado. A partir do conhecimento da tarefa, e consequentemente do aprendizado, entende-se que um algoritmo que possa ser aplicado para determinado tipo de tarefa e aprendizado é o meio para ser obter a informação ou padrões. Neste contexto, todas as diferentes técnicas para uma mesma tarefa têm o mesmo objetivo, por meio de algoritmos diferentes (Amaral, 2016).

2.3 TAREFA DE CLASSIFICAÇÃO

A classificação é uma tarefa que ocorre frequentemente no cotidiano (Bramer, 2016) e é a mais utilizada na mineração de dados, sendo também a que possui maior quantidade de algoritmos aplicáveis. Tem como objetivo descobrir a relação que um atributo específico (rótulo) tem em relação a outros atributos (Amaral, 2016).

Tan *et al.* (2019) define que, para uma tarefa de classificação, os dados são um conjunto de instâncias caracterizadas por um valor de x e y, onde x é o valor do registro, e y é o rótulo do atributo. O conjunto x pode conter qualquer tipo de valor, enquanto y deve ser categórico, caracterizando a tarefa de classificação. Há diversas formas de representar o modelo, como uma árvore, uma tabela de probabilidade, ou apenas um vetor. Pode-se expressar matematicamente como uma função em que a entrada é x e produz uma saída correspondente ao atributo. Se $f(x) = y$, conclui-se que o modelo classificou uma instância (x, y) corretamente.

Existem várias técnicas comuns para a classificação de registros, como as Árvores de decisão, Redes neurais artificiais, k-vizinhos mais próximos, Naïve Bayes, entre outros. Para esta

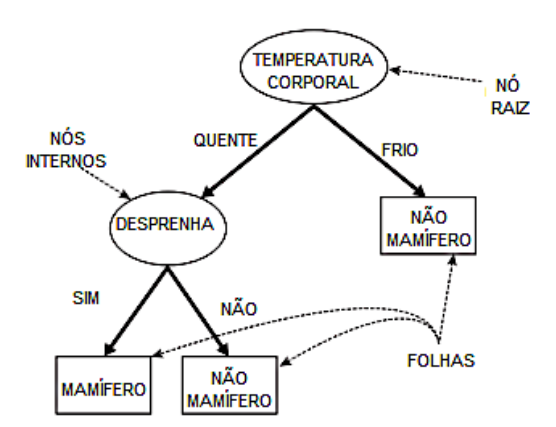
Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

pesquisa, foram utilizadas a Árvores de decisão, k-vizinhos mais próximos, Naïve Bayes e Floresta randômica, explicadas brevemente na sequência.

2.3.1 ÁRVORE DE DECISÃO

Tan, Steinbach, Karpatne e Kumar (2019) explicam que o método funciona como um fluxograma em forma de árvore, em que as folhas indicam o atributo ao qual o registro pertence, os nós internos e o nó raiz indicam um teste feito sobre o valor, e as ligações representam valores possíveis do teste anterior. Destaca-se que este algoritmo é bastante versátil, que podem executar tarefas de classificação, regressão e multiclasses, sendo capazes de se adaptar com conjuntos complexos de dados (Géron, 2019). Um exemplo de árvore criada está ilustrado na Figura 2.

Figura 2. Modelo de Árvore de decisão para classificar mamíferos



Fonte: Adaptado de Tan, Steinbach, Karpatne e Kumar (2019)

Para o caso desta pesquisa, será utilizado o algoritmo Árvore de Classificação e Regressão (*Classification and Regression Trees*, CART). Para este algoritmo, primeiramente é dividido o conjunto de treinamento em dois subconjuntos utilizando um único atributo k e um limiar t_k, por exemplo, "temperatura corporal = frio". Para definir k e t_k, busca-se pelo par (k, t_k) que produz subconjuntos mais puros (ou seja, que pertença totalmente ou parcialmente a uma única classe), ponderado pelo tamanho. Esta ideia é aplicada para os subconjuntos restantes, de forma recursiva (Géron, 2019).

2.3.2 *K*-VIZINHOS MAIS PRÓXIMOS

Esta técnica é de simples visualização e ela se encaixa no conceito de aprendizado baseado em instâncias, mais especificamente *lazy learners*. Esta característica se dá ao algoritmo que em que o treinamento é postergado até o último passo da classificação (Aggarwal, 2015).

O seu funcionamento é de acordo com a distância calculada entre a instância teste e todas as instâncias de treino, classificando o elemento de teste de acordo com a classe mais frequente dos k-vizinhos mais próximos (Oliveira, Faria, Gaio, & Reis, 2017). Supondo que todas as amostras da fase de treinamento sejam armazenadas como pontos em um espaço n-dimensional, em que n é o número de atributos, o algoritmo deve classificar uma nova amostra como pertencente ao grupo dos k vizinhos mais próximos, em que k é um parâmetro da quantidade de vizinhos analisados. Cada amostra da fase de treino corresponde a um vetor $X =$

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

$(x_1, x_2, x_3, \ldots, x_n, C)$, que pertence à classe C, e $Y = (y_1, y_2, y_3, \ldots, y_n)$ sendo um novo vetor ainda não classificado, calculam-se as distâncias dos vetores de treinamento mais próximos de Y, de acordo com a distância euclidiana (normalmente utilizada) na equação 1:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

Assim, Y é classificado dentro da classe em que os vizinhos estão mais próximos, ou seja, dentro da classe em que C que aparece com maior frequência (Kumar & Verma, 2012; Almeida & Faceroli, 2014).

2.3.3 NAÏVE BAYES

Este algoritmo é identificado como "ingênuo" (ou do inglês *naïve*) e assume que todas as variáveis são independentes considerando o valor de classe. A técnica é baseada no Teorema de Bayes e usa a probabilidade para encontrar a maior probabilidade das classificações possíveis (Aggarwal, 2015).

Supondo que $P(A)$ é a probabilidade de ocorrência do evento A, enquanto que $P(B)$ é a probabilidade de ocorrência do evento B, pode-se formular que as classes de um atributo de saída (A) e os atributos preditores (B) são dependentes de acordo com a seguinte forma (equação 2):

$$P(A|B) = \frac{P(A) \times P(B \vee A)}{P(B)} \quad (2)$$

Onde $P(A|B)$ é a probabilidade condicional da ocorrência de A dado B, identificando assim o Teorema de Bayes (Berrar, 2018).

De acordo com Bramer (2016), este algoritmo fornece uma combinação de probabilidade a *priori* e probabilidades condicionais em uma fórmula única, calculando-se a probabilidade de cada uma das classificações possíveis. Após isso, escolhe-se a classificação com o maior valor de probabilidade.

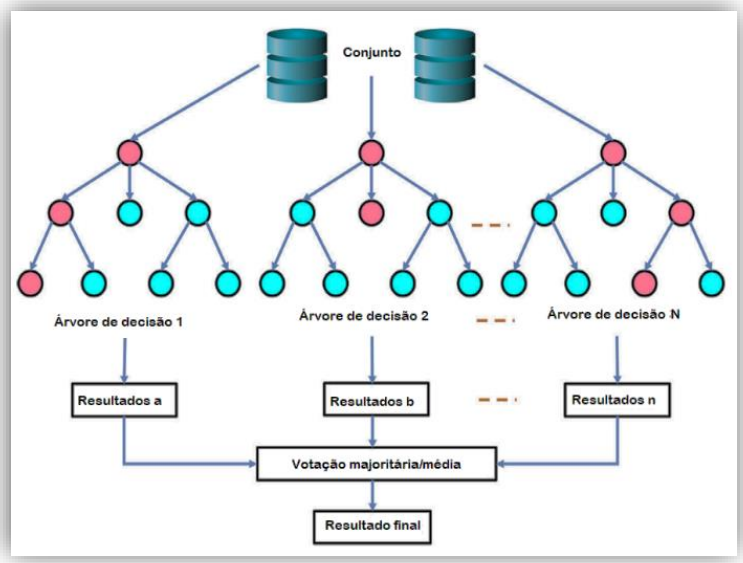
2.3.4 FLORESTA RANDÔMICA

O algoritmo de Floresta randômica (ou do inglês, *Random forest*) é uma extensão importante das Árvores de decisão. A técnica pode ser considerada um tipo de classificador *ensemble*, ou seja, ele faz o uso de vários classificadores considerados fracos (*weak learner*) em um forte (*strong learner*) (Breiman, 2001).

Esta técnica constrói uma grande coleção de Árvores de decisão não correlacionadas entre si, combinando-as de forma paralela a partir da votação majoritária de uma classe (quando a tarefa é de classificação) (Rodriguez-Galiano, Luque-Espinar, Chica-Olmo, & Mendes, 2018) (Figura 3).

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

Figura 3. Exemplo de uma Floresta randômica para a classificação



Fonte: Liu, Esan, Pan e An (2021, tradução livre)

Como mostrado na Figura 3, cada árvore cresce de forma independente e este fator favorece à redução de sobreajustes (*overfitting*) do modelo classificador (Gómez-Flores, Garza-Saldaña, & Varela-Fuentes, 2019). O treinamento de uma Floresta randômica é chamado de *bagging* (*bootstrap aggregation*) devido justamente à sua reamostragem com reposição para o treinamento de cada árvore, que são construídas paralelamente, ou seja, não há interferência ou peso de umas sobre as outras (Géron, 2019).

3. DESCRIÇÃO DO CONJUNTO DE DADOS

O Grupo do Banco Mundial (*World Bank Group*) em conjunto com a Fundação Bill & Melinda Gates, realizou a pesquisa identificada como "*Global Financial Inclusion (Global Findex)*", sobre como os adultos economizam, emprestam, pagam e gerenciam riscos financeiros. Esta pesquisa vem sendo realizada desde 2011 e publicada trienalmente, abrangendo mais de 150.000 adultos em mais de 140 economias no mundo, e é a maior nesta área de indicadores do uso de serviços e tecnologias financeiras (*World Bank Group*, 2021).

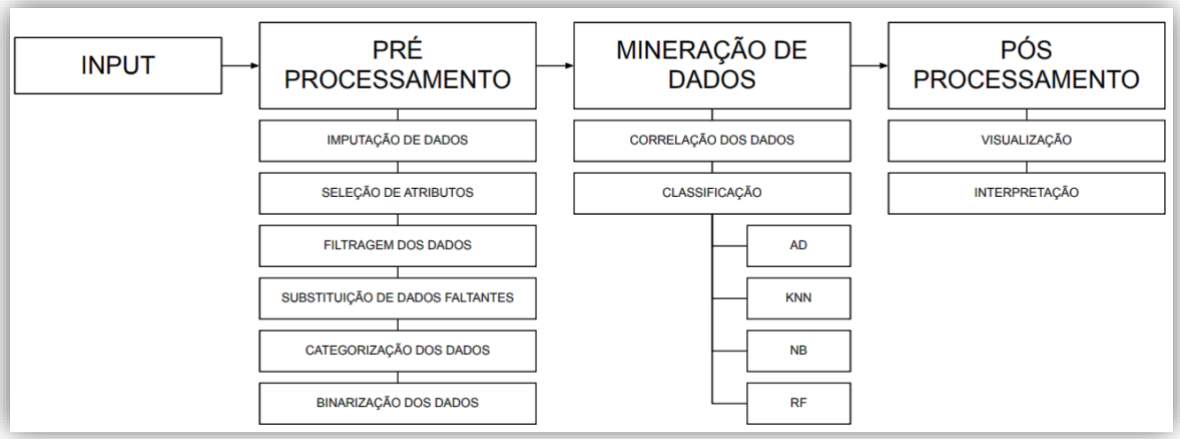
A base completa contém informações de mais de 5.000 projetos, incluindo setores como transporte, energia, telecomunicações, água e saneamento, entre outros. A partir dos mais de 150 mil registros iniciais e 105 variáveis existentes, um total de 32 atributos (incluindo a classe de interesse, neste caso a variável "*female*"), e um pouco mais de 15.504 instâncias foram utilizadas. As questões (variáveis) utilizadas e suas respectivas descrições estão no material em Anexo.

4. MATERIAL E MÉTODOS

Os dados foram pré-processados e analisados utilizando a linguagem de programação Python, versão 3.8, a partir da IDE *Jupyter*. As etapas de implementação das fases executadas estão descritas na sequência e representadas no fluxograma da Figura 4.

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

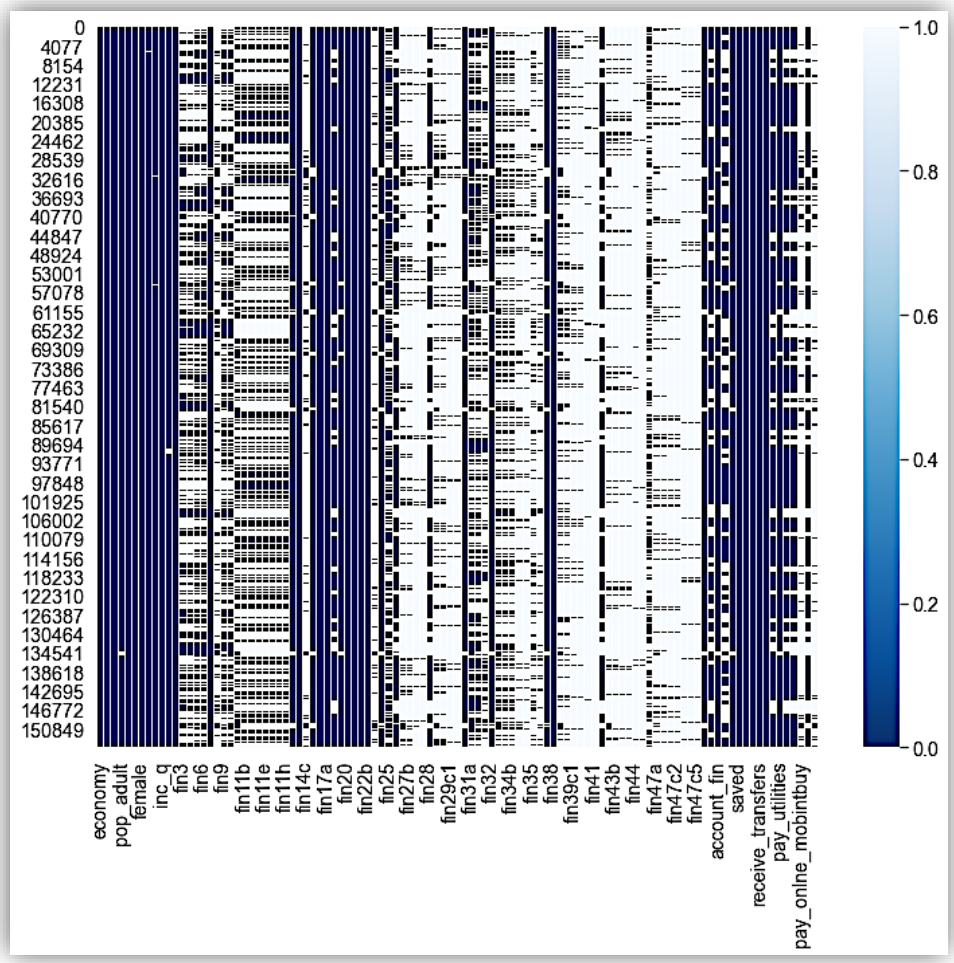
Figura 4. Etapas de implementação



Fonte: Autores (2021)

A base de dados bruta foi importada no programa e o filtro inicial foi realizado. Para esta primeira etapa, houve a análise de dados faltantes nas mais de 150 mil instâncias e mais de 40 atributos. A Figura 5 ilustra a ocorrência de dados faltantes, representada pela cor azul clara, enquanto que existe determinado dado onde há o azul escuro.

Figura 5. Análise de dados faltantes na base bruta



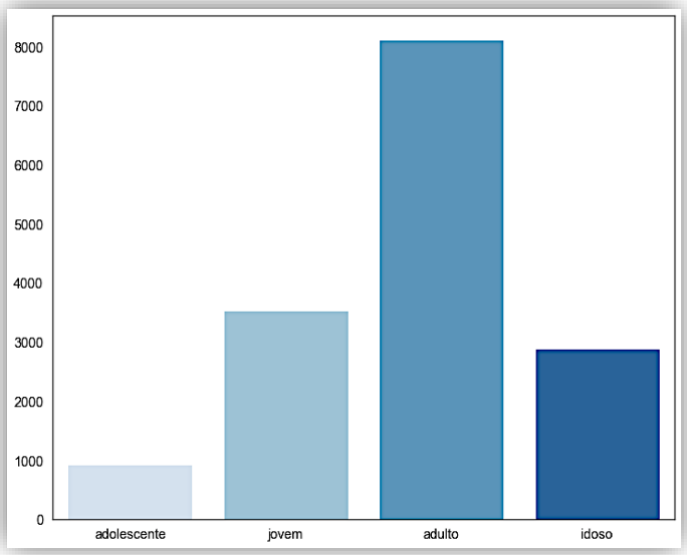
Fonte: Autores (2021)

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

Percebe-se na Figura 5 que há variáveis com inúmeros dados faltantes, precisamente 7.223.617 valores faltantes, correspondendo a 44,41% de toda a base. Para o pré-processamento, primeiramente foi realizada a filtragem dos atributos, onde as variáveis foram selecionadas pelo critério dos autores como as mais relevantes ou não derivadas de outras existentes, e também viáveis (com poucos dados faltantes). Um filtro foi realizado para obter apenas as instâncias que pertenciam a países latino-americanos (considerados de baixa ou média renda), incluindo o Brasil, obtendo-se um conjunto de 15.504 instâncias, com 25 atributos e apenas 0,26% dos dados faltantes. Os atributos "*age*" e "*account_mob*" foram completados com as suas respectivas medianas. Os dados cujo os entrevistados não quiseram responder a pesquisa, ou afirmaram não saber a resposta, foram removidos.

Na sequência, foi feita a categorização dos atributos "*age*" (resultando na Figura 6), em que as idades foram separadas nos seguintes intervalos e suas respectivas categorias: de 0 a 17 anos = "adolescente"; de 18 a 27 anos = "jovem"; de 28 a 59 anos = "adulto"; e, de 60 a 100 anos = "idoso".

Figura 6. Frequência do atributo *age* categorizado

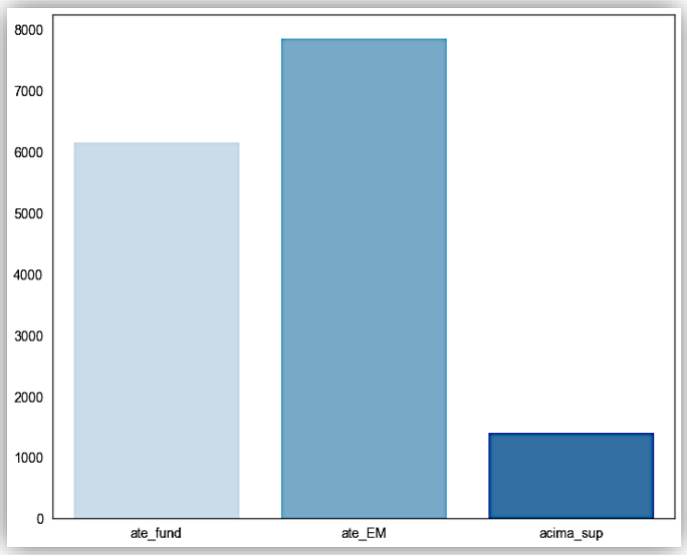


Fonte: Autores (2021)

Para a categorização do atributo *educ* (Figura 7), foi discretizada em três categorias, sendo "ate_fund" aqueles que cursaram até o ensino fundamental completo, "ate_EM" para aqueles que cursaram até o ensino médio completo e superior incompleto, e "acima_sup" para os que possuíam ensino superior completo e qualquer outro título acima deste.

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

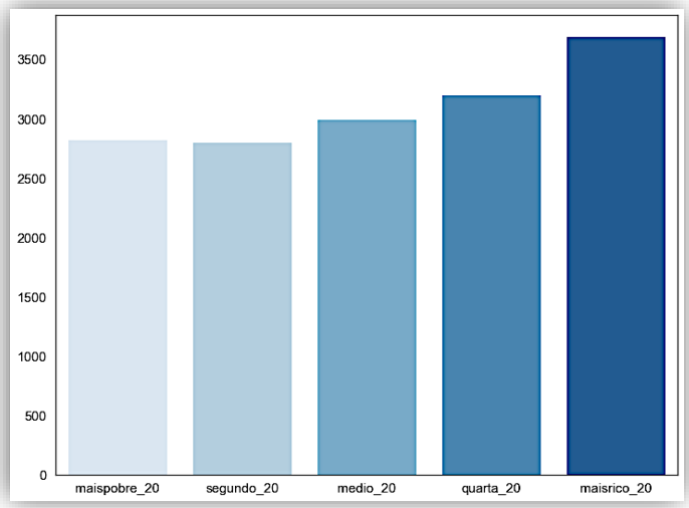
Figura 7. Frequência do atributo *educ* categorizado



Fonte: Autores (2021)

Por fim, a categorização do atributo *inc_q* (Figura 8), podendo ser categorizado em cinco classes: “maispobre_20” = mais pobre do quintil de renda da população; “segundo_20” = segundo quintil mais pobre; “médio_20” = terceiro quintil; “quarta_20” = quarto quintil; e, “maisrico_20” = 20% mais rico do quintil de renda da população.

Figura 8. Frequência do atributo *inc_q* categorizado



Fonte: Autores (2021)

Além da categorização dos atributos anteriores, também foi feita a criação de variáveis *dummy* (binarização, ver Masmoudi, Turkay e Chabchoub (2013)) de todos atributos que não eram dicotômicos. Foi então criado um conjunto de dados para treinamento, e um outro conjunto de dados para teste, sendo divididos em 80-20, ou seja, 80% dos dados foram para treinamento e o restante para testar o poder preditivo dos algoritmos. Para o conjunto de treinamento, foi realizado o teste de correlação dos atributos pelo método de Phi, o qual é um teste não paramétrico para medir a força da associação entre dois atributos dicotômicos (Frey, 2018).

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

Para realizar a tarefa de classificação, foram aplicadas as quatro técnicas descritas na Seção 2.3. Os modelos foram criados (treinados) a partir de uma amostra com 12.403 instâncias, e a fase de teste foi realizada com 3.101 instâncias. A avaliação do modelo foi feita com as métricas de acurácia, precisão, *recall* (ou sensibilidade), *F1-score* e a área sob a curva da característica de operação do receptor, ou do inglês *area under the curve – Receiver Operating Characteristic* (auc-ROC).

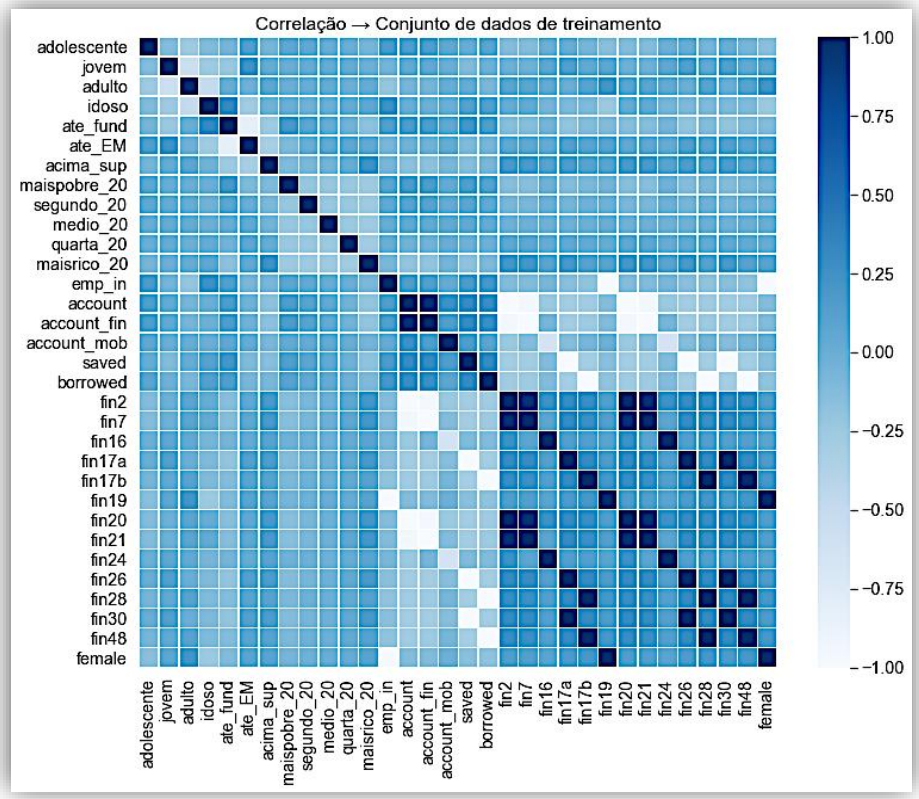
5. RESULTADOS E DISCUSSÃO

O atributo alvo (*target*) para os testes é a variável *female*, com a função de mapear entradas para saídas desejáveis, ou seja, o processo de treinamento continua até o modelo atingir um determinado nível de precisão desejável nos dados de teste. Busca-se analisar com este atributo alvo o comportamento de homens e mulheres a respeito do uso de serviços financeiros.

Primeiramente, foi verificado o balanceamento das classes para a variável de saída, e identificou-se no conjunto de treinamento um total de 8.317 mulheres (classe de interesse) e 4.086 homens (classe contrária), representando 67% de mulheres no conjunto. Desta forma, levantou-se a hipótese de que o modelo conseguiria aprender de forma mais efetiva os padrões de mulheres, por ter um maior número de exemplos na etapa de treinamento.

Para a etapa de correlação, o *heatmap* da Figura 9 ilustra tal situação para o conjunto de treinamento. Quanto mais escura a cor do quadrado na Figura 9, maior é a correlação positiva, enquanto que os quadrados mais próximos da cor branca, estes possuem uma correlação mais forte negativa.

Figura 9. Correlação de dados no conjunto de treinamento



Fonte: Autores (2021)

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

Durante a análise de correlação (Figura 9), notou-se que o atributo-alvo, *female*, possui total relação (correlação perfeita) com o atributo *fin19*, como também ser inversamente proporcional ao atributo *emp_in*, o que estava influenciando o programa na realização dos cálculos de forma previsível e incoerente, sendo assim foi removido do conjunto de entrada as seguintes variáveis: *female* (por ser o atributo-alvo), *fin19* e *emp_in*. Decidiu-se por manter todos os outros atributos, mesmo encontrando fortes correlações entre si (variáveis de entrada).

Como forma de avaliação do desempenho dos algoritmos, utilizou-se as métricas de acurácia, precisão, sensibilidade e *F1-score*, em que quanto mais próximas de um, melhor é o seu desempenho. A Tabela 1 apresenta os resultados obtidos pela técnica da Árvore de decisão, em que para ajuste do modelo, foram modificados dois parâmetros, com variação de apenas um deles:

- *max_depth*: limita a profundidade da árvore. Será testado com profundidade igual a cinco, buscando evitar possível *overfitting*;
- *min_samples_leaf*: número mínimo de observações que deve existir em uma folha da árvore. Foram testados os valores: 1, 5, 10, 20, 30.

Tabela 1. Resultados da Árvore de decisão

Técnica - AD (m=5, s=1)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 70%	Homem	0,63	0,39	0,48
	Mulher	0,73	0,88	0,79
	Macro Média	0,68	0,63	0,64
	Média Ponderada	0,69	0,70	0,68
Técnica - AD (m=5, s=5)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 70%	Homem	0,63	0,39	0,48
	Mulher	0,73	0,87	0,79
	Macro Média	0,68	0,63	0,64
	Média Ponderada	0,69	0,70	0,68
Técnica - AD (m=5, s=10)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 71%	Homem	0,63	0,39	0,48
	Mulher	0,73	0,88	0,79
	Macro Média	0,68	0,63	0.64
	Média Ponderada	0,69	0,71	0,68
Técnica - AD (m=5, s=20)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 71%	Homem	0,63	0,39	0,48
	Mulher	0,73	0,88	0,79
	Macro Média	0,68	0,63	0,64
	Média Ponderada	0,69	0,71	0,68
Técnica - AD (m=5, s=30)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 71%	Homem	0,63	0,40	0,49
	Mulher	0,73	0,87	0,79
	Macro Média	0,68	0,64	0,64
	Média Ponderada	0,69	0,71	0,69

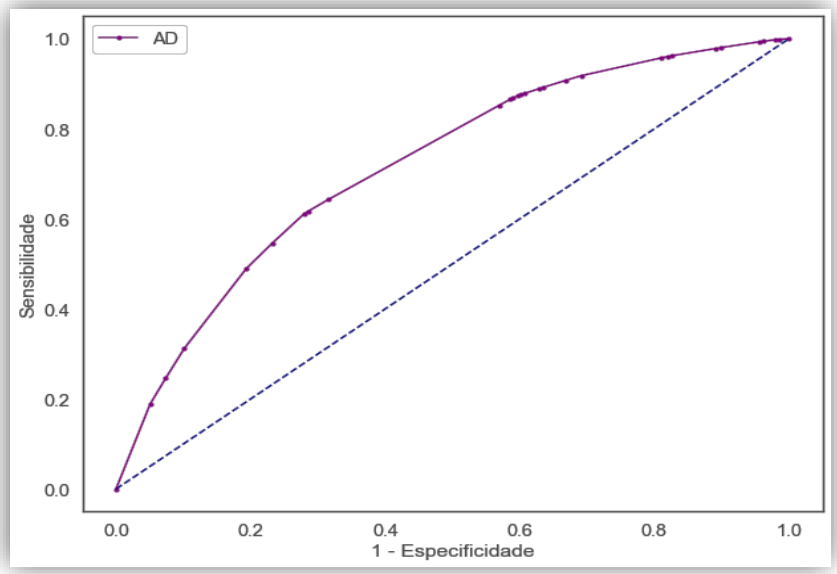
Nota-se na Tabela 1 que não se pode afirmar que um modelo foi melhor do que outro, pois todos alcançaram acurácias muito parecidas, de 70% ou 71%. Para a classificação de mulheres, o modelo atingiu 88% de sensibilidade e uma precisão de 73%, porém o melhor resultado de especificidade (ou sensibilidade para homens, que é a classe contrária), atingiu-se apenas 40%, bastante baixo quando comparado à identificação das mulheres.

A Figura 10 mostra o gráfico da curva ROC, que é obtida por meio do cálculo de probabilidade de cada observação pertencer à classe. Foi realizado o cálculo da área sob a curva ROC (auc-

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

ROC), a partir do valor obtido para esta métrica que, no caso, foi de 0,72 para todos os modelos criados de Árvores de decisão.

Figura 10. Gráficos da Curva ROC dos testes de Árvore de decisão



Fonte: Autores (2021)

A Tabela 2 aponta os resultados da técnica do *k*-vizinhos mais próximos, em que para ajuste do modelo foi analisado o número de vizinhos:

- *n_neighbors*: número de vizinhos necessários para cada amostra. Serão testados os valores: 3, 10, 15, 30 e 50.

Tabela 2. Resultados do *k*-vizinhos mais próximos (KNN)

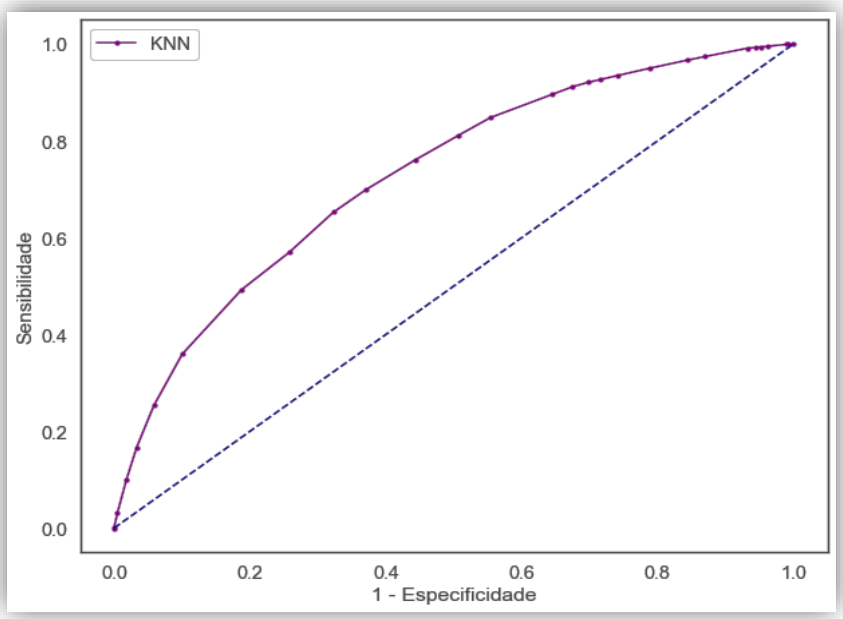
Técnica - KNN (n. vizinhos = 3)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 65%	Homem	0,51	0,47	0,49
	Mulher	0,72	0,75	0,74
	Macro Média	0,62	0,61	0,61
	Média Ponderada	0,65	0,65	0,65
Técnica - KNN (n. vizinhos = 10)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 69%	Homem	0,58	0,41	0,48
	Mulher	0,73	0,84	0,78
	Macro Média	0,65	0,63	0,63
	Média Ponderada	0,67	0,69	0,67
Técnica - KNN (n. vizinhos = 15)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 70%	Homem	0,63	0,34	0,44
	Mulher	0,71	0,89	0,79
	Macro Média	0,67	0,63	0,62
	Média Ponderada	0,68	0,70	0,67
Técnica - KNN (n. vizinhos = 30)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 69%	Homem	0,63	0,30	0,41
	Mulher	0,70	0,91	0,79
	Macro Média	0,67	0,60	0,60
	Média Ponderada	0,68	0,69	0,66
Técnica - KNN (n. vizinhos = 50)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 70%	Homem	0,65	0,30	0,41
	Mulher	0,71	0,91	0,80
	Macro Média	0,68	0,61	0,61
	Média Ponderada	0,69	0,70	0,66

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

Para o caso do *k*-vizinhos mais próximos (Tabela 2), observa-se que a acurácia com o melhor modelo também alcançou 70%, com uma sensibilidade de 91% e precisão de 71%, retornando consequentemente o melhor *F1-score* (80%), focando para o caso da classe de interesse. Estes resultados foram obtidos com 50 vizinhos, mas a sensibilidade de 30% para classificar homens nesta configuração mostra que o modelo tem maior dificuldade em classificar homens do que mostrou a Árvore de decisão. Esta situação é diferente para um número de vizinhos igual a 3 ou 10, que atingiram melhores sensibilidades para homens (ou especificidade) do que outras configurações. Entretanto, como uma métrica melhorava para uma classe, em detrimento da outra, não se consegue afirmar que um modelo se sobressai em relação aos outros.

A Figura 11 representa o gráfico da auc-ROC dos testes aplicados à técnica do *k*-vizinhos mais próximos. Foi realizado o cálculo da área, sendo que o melhor resultado alcançado foi de 0,73 a partir do parâmetro *k* com 30 vizinhos.

Figura 11. Gráficos da Curva ROC dos testes do *k*-vizinhos mais próximos



Fonte: Autores (2021)

A Tabela 3 mostra os resultados obtidos pela técnica Naïve Bayes (Bernoulli).

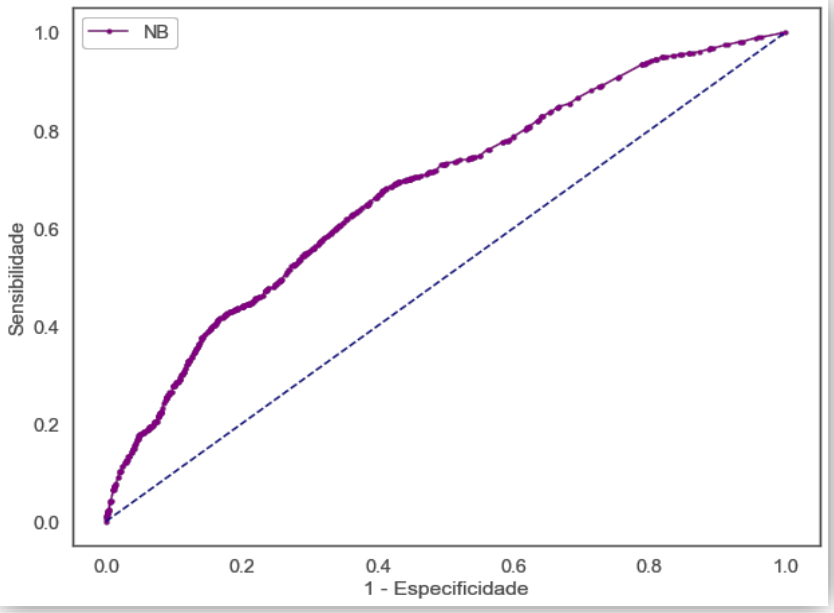
Tabela 3. Resultados do Naïve Bayes (Bernoulli)

Técnica - NB		Precisão	Sensibilidade	*F1-score*
Acurácia: 64%	Homem	0,49	0,64	0,55
	Mulher	0,64	0,64	0,70
	Macro Média	0,63	0,64	0,62
	Média Ponderada	0,67	0,64	0,65

Verifica-se pela Tabela 3 que os resultados obtidos por Naïve Bayes não foram competitivos em termos de acurácia, precisão e também as médias dos resultados (macro média e média ponderada). Por mais que a sensibilidade para se classificar homens e a precisão para se classificar mulheres tenham sido melhores do que outros métodos até então, quando é avaliada a sensibilidade para mulheres e precisão para homens, o desempenho é ruim se comparado aos outros métodos. A Figura 12 mostra o auc-ROC obtido por este algoritmo, com o valor 0,70.

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

Figura 12. Gráficos da Curva ROC dos testes de Naïve Bayes



Fonte: Autores (2021)

A Tabela 4 sintetiza os resultados para a Floresta randômica, variando-se o parâmetro sobre a quantidade de Árvores de decisão (estimadores) da seguinte forma:

- *n_estimators* = quantidade de Árvores de decisão para a montagem da floresta, variando este parâmetro com os valores 50, 100, 500, 1.000 e 2.000.

Tabela 4. Resultados da Floresta randômica

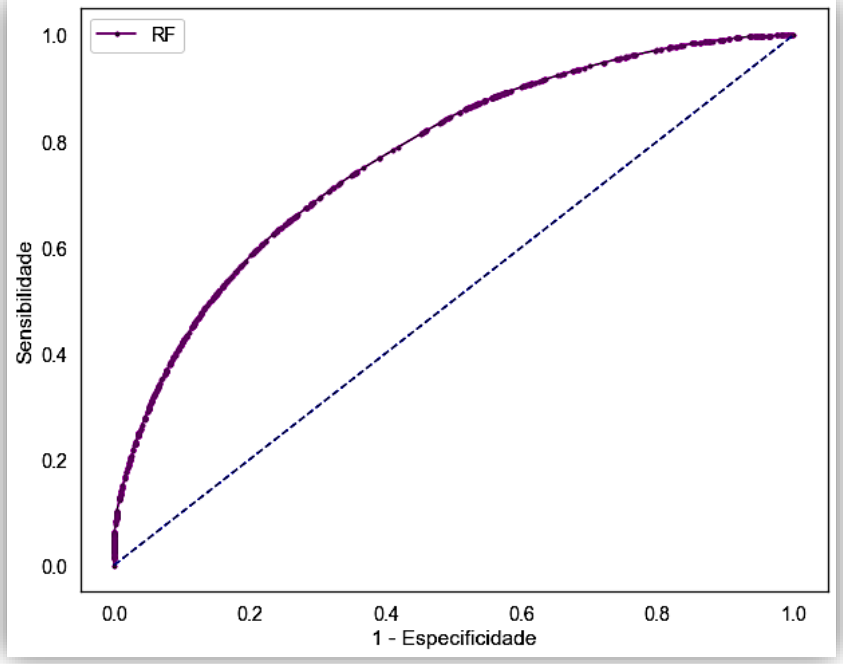
Técnica - RF (n. árvores = 50)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 69%	Homem	0,65	0,38	0,46
	Mulher	0,72	0,86	0,78
	Macro Média	0,66	0,62	0,62
	Média Ponderada	0,68	0,69	0,67
Técnica - RF (n. árvores = 100)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 69%	Homem	0,65	0,38	0,47
	Mulher	0,72	0,86	0,78
	Macro Média	0,66	0,62	0,63
	Média Ponderada	0,68	0,69	0,67
Técnica - RF (n. árvores = 500)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 69%	Homem	0,60	0,38	0,47
	Mulher	0,72	0,86	0,79
	Macro Média	0,66	0,62	0,63
	Média Ponderada	0,68	0,69	0,67
Técnica - RF (n. árvores = 1000)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 69%	Homem	0,61	0,37	0,46
	Mulher	0,72	0,87	0,79
	Macro Média	0,66	0,62	0,63
	Média Ponderada	0,68	0,69	0,67
Técnica – RF (n. árvores = 2000)		**Precisão**	**Sensibilidade**	***F1-score***
Acurácia: 69%	Homem	0,60	0,38	0,46
	Mulher	0,72	0,86	0,79
	Macro Média	0,66	0,62	0,63
	Média Ponderada	0,68	0,69	0,67

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

Analisando-se a Tabela 4, verifica-se que a Floresta randômica não conseguiu alcançar os 71% de acurácia, como houve na árvore de decisão. Entretanto, assim como a técnica do *k*-vizinhos mais próximos, obteve uma melhor sensibilidade e precisão para a classificação de mulheres, mas falhando na tentativa de se identificar os homens contidos no conjunto de dados.

A Figura 13 mostra a auc-ROC com um valor alcançado de 0,77, praticamente o mesmo para todas as configurações dos parâmetros da técnica. Esta métrica consegue identificar o poder de distinção do modelo ao tentar diferenciar, de forma probabilística, as duas classes analisadas.

Figura 13. Gráficos da Curva ROC dos testes de Floresta Randômica



Fonte: Autores (2021)

Em geral, pode se concluir que as técnicas não puderam alcançar acurácias tão significativas, com um máximo de 71% com as Árvores de decisão. Porém, muitas delas conseguiram classificar as mulheres com um alto número de classificações corretas, com até 88% na métrica de sensibilidade e, das instâncias que foram classificadas como mulheres pelos modelos, estes conseguiram predizer corretamente acima de 73% em diversos casos.

Identificou-se que as limitações dos modelos estão, basicamente, em conseguir classificar homens a partir dos atributos de dados utilizados. Ao se verificar o balanceamento das classes (início da Seção 5), notou-se que as mulheres correspondiam a 67% das instâncias, mais do que o dobro de homens. Em um primeiro momento, acreditou-se que este número de mulheres favorecia o modelo por ter mais exemplos para a etapa do treinamento.

Neste contexto, fez-se duas pequenas variações nos conjuntos de dados, de forma a tentar balancear as classes no conjunto de treinamento, com dois novos experimentos:

1. Balanceamento por subamostragem aleatória, retirando instâncias de forma aleatória da classe mais favorecida (mulheres), e;
2. Balanceamento por reamostragem aleatória, reutilizando instâncias de forma aleatória, da classe menos favorecida (homens).

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

Para isso, fez-se com que a classe menos favorecida tivesse 90% de instâncias da classe mais favorecida (parâmetro escolhido por conveniência). Assim, obteve-se dois novos conjuntos, com a seguinte frequência:

- Conjunto com subamostragem: 4.540 mulheres e 4.086 homens;
- Conjunto com reamostragem: 8.317 mulheres e 7.485 homens.

Para estes novos conjuntos, ambos foram comparados apenas para as Árvores de decisão com amostra mínimas de 1 e 30, para as métricas de acurácia, sensibilidade, precisão e *f1-score*, não sendo analisadas as macro médias e médias ponderadas (Tabela 5).

Tabela 5. Resultados das Árvores de decisão com os conjuntos original, com subamostragem e reamostragem

AD (m=5, s=1)	**Classe**	**Acurácia**	**Precisão**	**Sensibilidade**	***F1-score***
Original	Homem (4.086)	70%	0,63	0,39	0,48
	Mulher (8.317)		0,73	0,88	0,79
Subamostragem	Homem (4.086)	66%	0,51	0,69	0,59
	Mulher (4.540)		0,79	0,65	0,71
Reamostragem	Homem (7.485)	66%	0,51	0,68	0,59
	Mulher (8.317)		0,79	0,65	0,71
AD (m=5, s=30)	**Classe**	**Acurácia**	**Precisão**	**Sensibilidade**	***F1-score***
Original	Homem (4.086)	71%	0,63	0,40	0,49
	Mulher (8.317)		0,73	0,87	0,79
Subamostragem	Homem (4.086)	66%	0,51	0,69	0,59
	Mulher (4.540)		0,79	0,65	0,71
Reamostragem	Homem (7.485)	66%	0,51	0,69	0,59
	Mulher (8.317)		0,79	0,65	0,71

Percebe-se na tabela 5 que mesmo após um balanceamento entre as classes, não se conseguiu obter resultados que sejam interessantes em todas as métricas, pois por mais que a sensibilidade tenha aumentado em quase 40%, a sensibilidade para mulheres e a acurácia são prejudicadas, não se identificando desta forma uma abordagem que seja melhor em termos de preparação dos modelos.

Por fim, pode-se levantar a hipótese também de que as mulheres da América Latina possuem um perfil muito mais parecido de consumo ou características próprias, encontrando-se um padrão para este sexo. Já os homens latino-americanos parecem ter muito mais diferenças, seja em demografia, social ou de consumo, sendo mais difícil então de se conseguir chegar a um perfil geral deste público. Neste contexto, é mais interessante proporcionar programas ou incentivos para a inclusão financeira voltada para o público feminino.

6. CONCLUSÕES E TRABALHOS FUTUROS

Este artigo apresentou o uso de quatro técnicas de mineração de dados aplicadas à tarefa de classificação, comparando estes algoritmos por meio de várias métricas de classificação. O conjunto de dados utilizado foi extraído de um conjunto público disponível eletronicamente e foram utilizadas as mais variadas etapas de pré-processamento, mineração e obtenção do conhecimento. Foram investigados diferentes parâmetros de forma a melhorar o poder preditivo da classificação, principalmente no que diz respeito às mulheres, foco deste trabalho.

Comparando-se as técnicas, foi encontrado um resultado melhor de acurácia, sensibilidade e precisão (com referência às mulheres) para aquelas baseadas em árvores de indução, ou seja, Árvores de decisão e Floresta randômica. O poder de distinção baseado na métrica auc-ROC

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

foi melhor com a Floresta randômica, a qual é implementada a partir da construção de várias Árvores de decisão. Esta métrica alcançou o valor de 0,77, maior do as outras técnicas.

Entretanto, os modelos criados não puderam identificar padrões em homens latino-americanos a partir do treinamento, mesmo quando se fez um treinamento com classes balanceadas, encontrando assim uma limitação do estudo aqui apresentado. Também é importante destacar que esta pesquisa é um ponto de partida para um melhor entendimento sobre as características quanto ao uso de serviços financeiros por mulheres e homens.

Neste contexto, para pesquisas futuras, sugere-se utilizar técnicas que possam selecionar os atributos mais relevantes para os modelos, pois técnicas como o *k*-vizinhos mais próximos e Naïve Bayes não possuem critérios de seleção de variáveis embutidos em seus algoritmos, como é o caso das Árvores de decisão e de Florestas randômicas. Também se ressalta a importância de uma quantidade maior de exemplos para a classe de homens. Desta forma, irá abrir a possibilidade de se encontrar padrões dentro desta classe masculina, podendo melhorar o poder de predição desta classe específica, como também manter uma boa classificação de mulheres, visando um aumento de acurácia no conjunto de teste, e consequentemente melhorando a generalização dos modelos.

REFERÊNCIAS BIBLIOGRÁFICAS

Abdul Razak, A., & Asutay, M. (2022). Financial inclusion and economic well-being: Evidence from Islamic Pawnbroking (Ar-Rahn) in Malaysia. Research in International Business and Finance, 59, 101557. https://doi.org/10.1016/j.ribaf.2021.101557

Aggarwal, C. C. (2015). Data Mining. In Data Mining. Springer International Publishing. https://doi.org/10.1007/978-3-319-14142-8

Almeida, R. C. de, & Faceroli, S. T. (2014). *Análise comparativa das técnicas KNN e rede neural MLP na classificação de padrões mioelétricos*. Anais Do XXIV Congresso Brasileiro de Engenharia Biomédica.

Amaral, F. (2016). *Aprenda Mineração de Dados: Teoria e Prática* (1 ed.). Alta Books.

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

Bramer, M. (2016). Principles of Data Mining (3rd ed.). Springer London. https://doi.org/10.1007/978-1-4471-7307-6

Breiman, L. (2001). Random forests. Machine Learning, *45*, 5–32. https://doi.org/10.1007/9781441993267_5

Camilo, C. O., & Silva, J. C., da. (2009). Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. Recuperado de https://rozero.webcindario.com/disciplinas/fbmg/dm/RT-INF_001-09.pdf

Dogan, A. & Birant, D. (2021). Machine learning and data mining in manufacturing. Expert Systems with Applications, *166*, 114060. https://doi.org/10.1016/j.eswa.2020.114060

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, *17*(3), 37-53. 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. (2020). Use of machine learning techniques in bank credit risk analysis. *Revista Internacional de Metodos Numericos Para Calculo y Diseno En Ingenieria*, *36*(3), 1-15. https://doi.org/10.23967/J.RIMNI.2020.08.003

Frey, B. B. (2018). Phi Correlation Coefficient. In The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation. SAGE. https://doi.org/10.4135/9781506326139

Géron, A. (2019). *Mãos à Obra: Aprendizado de Máuqina com Scikit-Learn & TensorFlow* (1 ed.). Alta Books.

Citação (APA): Souza, P. R. V., de., Tomazi, B. G., & Santos, B. S., dos. (2022). Utilização de técnicas de classificação em conjunto de dados sobre inclusão financeira: um estudo baseado em países Latinoamericanos. *Brazilian Journal of Production Engineering*, *8*(1), 73-91.

Goldschmidt, R., Passos, E., & Bezerra, E. (2015). *Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações* (2a ed.). Elsevier.

Gómez-Flores, W., Garza-Saldaña, J. J., & Varela-Fuentes, S. E. (2019). Detection of Huanglongbing disease based on intensity-invariant texture analysis of images in the visible spectrum. *Computers and Electronics in Agriculture*, *162*(2018), 825-835. https://doi.org/10.1016/j.compag.2019.05.032

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review : Machine learning techniques applied to financial market prediction R. *Expert Systems With Applications*, *124*, 226-251. https://doi.org/10.1016/j.eswa.2019.01.012

Kumar, R. & Verma, R. (2012). Classification Algorithms for Data Mining: A Survey. *International Journal of Innovations in Engineering and Technology*, *1*(2), 7-14.

Larose, D. T. & Larose, C. D. (2014). *Discovering Knowledge in Data* (2nd ed.). John Wiley & Sons, Inc.

Liu, J., Kong, X., Zhou, X., Wang, L., Zhang, D., Lee, I., Xu, B., & Xia, F. (2019). Data Mining and Information Retrieval in the 21st century: A bibliographic review. *Computer Science Review*, *34*. https://doi.org/10.1016/j.cosrev.2019.100193

Liu, Y., Esan, O. C., Pan, Z., & An, L. (2021). Machine learning for advanced energy materials. *Energy and AI*, *3*. https://doi.org/10.1016/j.egyai.2021.100049

Marcelin, I., Egbendewe, A. Y. G., Oloufade, D. K., & Sun, W. (2021). Financial inclusion, bank ownership, and economy performance: Evidence from developing countries. *Finance Research Letters*, *102322*. https://doi.org/10.1016/j.frl.2021.102322

Masmoudi, Y., Turkay, M., & Chabchoub, H. (2013). A binarization strategy for modelling mixed data in multigroup classification. *International Conference on Advanced Logistics and Transport*, 347-353. https://doi.org/10.1109/ICAdLT.2013.6568483

Morgan, P. J., & Pontines, V. (2018). Financial stability and financial inclusion: The case of SME lending. *The Singapore Economic Review*, *63*(01), 111-124. https://doi.org/10.1142/S0217590818410035

Oliveira, A., Faria, B. M., Gaio, A. R., & Reis, L. P. (2017). Data Mining in HIV-AIDS Surveillance System: Application to Portuguese Data. *Journal of Medical Systems*, *41*(4). https://doi.org/10.1007/s10916-017-0697-4

Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na Era do Big Data. *Brazilian Journal of Production Engineering*, *7*(3), 106-125. https://doi.org/10.47456/bjpe.v7i3.35743

Robino, C., Trivelli, C., Villanueva, C., Sachetti, F. C., Walbey, H., Martinez, L., & Marincioni, M. (2018). *Financial Inclusion for Women: A Way Forward*.

Rodriguez-Galiano, V. F., Luque-Espinar, J. A., Chica-Olmo, M., & Mendes, M. P. (2018). Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. *Science of the Total Environment*, *624*, 661-672. https://doi.org/10.1016/j.scitotenv.2017.12.152

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson Prentice Hall.