9. ORANGE DATA MINING: UMA FERRAMENTA PARA INSERÇÃO DE INTELIGÊNCIA ARTIFICIAL NA PESQUISA CIENTÍFICA

Caio Saraiva Coneglian
Henrique Leal Tavares
Diego José Macedo
Milton Shintaku

9.1 INTRODUÇÃO

A presença da tecnologia nas atividades humanas remonta à aurora da espécie, com a criação de ferramentas que apoiam atividades diárias. Com a especialização das atividades, ferramentas foram criadas exclusivamente para algumas profissões, mas que muitas vezes são adaptadas para outras. Em outros casos, algumas ferramentas já nascem fadadas a serem generalistas, podendo ser utilizadas em uma grande gama de atividades e profissões. Esse último tipo é o caso dos computadores, que por serem programáveis, são flexíveis para serem utilizados em quase todas as atividades humanas.

Dentre as atividades que podem fazer uso de computadores está a pesquisa científica, principalmente as que requerem o processamento automatizado de dados e informações. Tanto que, desde os primórdios da computação, as universidades e institutos de pesquisa possuíam computadores para serem utilizados pelos seus pesquisadores. Em alguns casos, universidades construíram os seus próprios computadores, em projetos de pesquisa ousados e inovadores como no caso do conhecido "Patinho Feio", primeiro computador brasileiro desenvolvido 100% pelo Laboratório de Sistemas Digitais (LSD) do Departamento de Engenharia Politécnica da Universidade de São Paulo (USP), lançado em julho de 1972 (Cardi; Barreto, 2012).

Inicialmente os computadores, como são conhecidos por grande parte da população atual, nasceram para processar dados numéricos e estruturados,

quase como uma calculadora programável. Tanto que, muitas vezes os computadores nas pesquisas eram utilizados para realizar cálculos repetitivos com grande quantidade de números, nem sempre complexos pelas limitações tecnológicas da época. Por isso, grande parte dos estudos que utilizam computadores nos primórdios da computação era voltada para as ciências rígidas, engenharia e estatística.

Com a evolução tecnológica, os computadores adotaram novas formas de atuação, com processamento de textos, imagens, áudio e vídeo. A criação dos chamados gerenciadores de banco de dados e linguagens de programação modernas e flexíveis contribuiu para o que é denominado de dados pudessem transcender tipos e formatos. Nesse caminho, o processamento de dados pode atender a todo o tipo de objeto digital, ou seja, tudo que pode ser codificado em formato digital.

Com a mudança de século e a popularização da *internet* e *web* novas possibilidades de pesquisa científica atendem a todas as áreas de conhecimento, em praticamente todas as etapas dos estudos. Pode-se afirmar que os computadores e seu ambiente digital estão presentes desde a criação de propostas de estudos, até as publicações e uso dos resultados. Com isso, pesquisas tendem a ser efetuadas de forma mais rápida e eficaz, assim como a disseminação dos seus resultados.

O uso da computação nas ciências se tornou tão comum, que já há consenso sobre a chamada ciência virtual em contraposição às ciências naturais. Realidade virtual, projeções e simulações são comuns nas pesquisas científicas, facilitando resolver problemas. Tanto que, por meio de pesquisas é possível indicar tendências que as organizações e instituições vão seguir na computação, incluindo as ciências para um futuro próximo, *A Gartner Group*, empresa de consultoria especializada em computação, prevê que a partir de 2023, o uso da inteligência artificial será cada vez mais constante, principalmente com as chamadas aplicações em inteligência artificial adaptáveis.

9.2 INTELIGÊNCIA ARTIFICIAL NO PROCESSO DA PESQUISA CIENTÍFICA

A pesquisa científica, ao longo de décadas, tem sido conduzida com base em métodos tradicionais que envolvem coleta manual de dados, revisões literárias extensas e análises estatísticas complexas. Em especial, no âmbito das Ciências Sociais Aplicadas, há uma série de métodos de pesquisa que podem ser utilizados para o desenvolvimento de pesquisas aplicadas.

Com a evolução da tecnologia, tendo como destaque a Inteligência Artificial, foram desenvolvidas novas técnicas que podem apoiar o desenvolvimento de pesquisas, em especial utilizando dados. Técnicas como *Machine Learning*, *Text Learning* e *Data Mining* têm desempenhado um papel fundamental na transformação da pesquisa científica, proporcionando ferramentas poderosas para extrair conhecimento a partir de grandes conjuntos de dados.

O campo da Inteligência Artificial que mais tem impactado a sociedade é *Machine Learning*. Esse campo tem a capacidade de revolucionar a pesquisa científica ao permitir que computadores aprendam com dados e façam previsões mais precisas. Ademais, tal capacidade de identificar padrões complexos e gerar insights tem aplicações em diversas áreas, desde a medicina, onde auxilia na identificação de diagnósticos precisos, até a física, onde ajuda a entender fenômenos complexos.

Uma definição de *Machine Learning* é dada por Jordan e Mitchell (2015, p. 255, tradução dos autores)[97]:

> O aprendizado de máquina é uma disciplina focada em duas questões inter-relacionadas: Como construir sistemas de computador que melhoram automaticamente com a experiência? e Quais são as leis fundamentais da teoria estatística da informação computacional que governam todos

97 **Trecho original:** *Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations? The study of machine learning is important both for addressing these fundamental scientific and engineering questions and for the highly practical computer software it has produced and fielded across many applications.*

> os sistemas de aprendizagem, incluindo computadores, seres humanos e organizações? O estudo do aprendizado de máquina é importante tanto para abordar essas questões científicas e de engenharia fundamentais, quanto para o *software* de computador altamente prático que ele produziu e utilizou em vários aplicativos.

Outra vertente, que está vinculada ao *Machine Learning*, mas com foco em tratamento de texto é o *Text Learning*. Esse campo concentra-se na análise de texto escrito, tornando possível a extração de informações valiosas a partir de documentos científicos extensos. Essa abordagem é especialmente relevante em um mundo inundado de informações, onde cientistas precisam navegar por vastos repositórios de literatura científica para manter-se atualizados e encontrar pistas para suas pesquisas.

Tommasel e Godoy (2019, p. 1, tradução dos autores)[98] apontam que *Text mining*:

> [...] refere-se a um processo de descoberta de conhecimento que visa a extração de padrões interessantes e não triviais da linguagem natural. Este processo inclui múltiplas áreas, como análise de texto, processamento de linguagem natural e recuperação de informação, entre outras.

Para o autor, o *Text Mining* é um guarda chuva que se relaciona a outras técnicas voltadas a processamento de texto para extração de informação.

Por fim, *Data Mining* é uma técnica que busca padrões ocultos em grandes conjuntos de dados, oferecendo uma abordagem que apoia na identificação de *insights* e tendências nas mais diversas áreas. Garcia, Luengo e Herrera (2015, p. 1, tradução nossa)[99] apontam que

> *DM* [*Data Mining*] trata, de modo geral, de resolver problemas por meio da análise de dados presentes em bancos de dados reais. Hoje em dia,

98 Trecho original: [...] *refers to a knowledge discovery process aiming at the extraction of interesting and non-trivial patterns from natural language. This process includes multiple fields, such as text analysis, natural language processing and information retrieval, amongst others.*

99 Trecho original: *DM is, roughly speaking, about solving problems by analyzing data present in real databases. Nowadays, it is qualified as science and technology for exploring data to discover already present unknown patterns.*

> qualifica-se como ciência e tecnologia para explorar dados para descobrir padrões desconhecidos já presentes.

Partindo desses três campos, *Machine Learning*, *Text Learning* e *Data Mining*, identifica-se que a pesquisa científica nas ciências sociais aplicadas pode usufruir de novas técnicas, em especial no âmbito de pesquisas aplicadas. Assim, entra-se em detalhes sobre como essas técnicas nas próximas subseções.

9.2.1 MACHINE LEARNING NA PESQUISA CIENTÍFICA

Machine Learning é uma subárea da inteligência artificial que se concentra no desenvolvimento de *algoritmos* e modelos que permitem que os sistemas computacionais aprendam e melhorem com a experiência. No contexto das Ciências Sociais Aplicadas, *Machine Learning* oferece uma abordagem capaz de analisar dados e compreender o comportamento humano em uma variedade de contextos.

Em especial, *Machine Learning* envolve a capacidade de computadores aprenderem com dados históricos, identificando padrões e fazendo previsões ou tomando decisões com base nesses padrões. Isso é feito por meio do treinamento de *algoritmos* em conjuntos de dados que contêm exemplos passados e resultados conhecidos. À medida que o *algoritmo* é exposto a mais dados, este ajusta seus parâmetros internos para melhorar seu desempenho, tornando-se mais preciso e eficiente em tarefas específicas.

As técnicas de *Machine Learning* podem ser usadas nos seguintes processos de Pesquisa nas Ciências Sociais Aplicadas:

- **Análise de Dados Complexos:** Em pesquisas que envolvem grandes volumes de dados, como pesquisas de opinião pública, o *Machine Learning* pode ser usado para identificar tendências, padrões de comportamento e *insights* ocultos que seriam difíceis de extrair por meio de métodos tradicionais. (Alinejad-Rokny; Sadroddiny; Scaria, 2018).

- **Previsão de Tendências:** Os *algoritmos* de *Machine Learning* podem ser aplicados para prever tendências sociais, econômicas e políticas.

Isso é particularmente necessário em previsões de eleições, análise de mercado de trabalho e estimativas de demanda por produtos e serviços (Lim; Zohren, 2021).

- **Segmentação de Público-alvo:** Em *marketing* e pesquisa de mercado, o *Machine Learning* é usado para segmentar o público-alvo com base em características demográficas, comportamentais e de consumo. Isso ajuda a direcionar campanhas publicitárias de maneira mais eficaz (Ernawati; Baharin; Kasmin, 2021).

- **Análise de Sentimento e Opinião:** Em análises de mídia social e pesquisas de opinião, o *Machine Learning* é aplicado para analisar o sentimento do público em relação a produtos, serviços ou questões políticas. Isso fornece uma análise das opiniões públicas e das tendências de opinião (Wankhade; Rao; Kulkarni, 2022).

- **Detecção de Fraudes e Anomalias:** Em finanças e segurança, o *Machine Learning* é usado para detectar fraudes e comportamentos anômalos em transações financeiras e atividades online. Tal aspecto ajuda sistemas e recursos contra atividades fraudulentas (Pourhabibi *et al.*, 2020).

Destaca-se que o *Machine Learning* oferece uma nova abordagem para análise de dados e pesquisa nas Ciências Sociais Aplicadas. Essa tecnologia permite que os pesquisadores extraiam insights mais profundos, façam previsões mais precisas e compreendam melhor o comportamento humano em uma ampla variedade de contextos, enriquecendo a pesquisa nessa área de estudo.

9.2.2 TEXT LEARNING: DA LITERATURA CIENTÍFICA AO CONHECIMENTO AVANÇADO

Text Learning, ou aprendizado de texto, é um dos campos vinculados à *Machine Learning* que se concentra na análise e na interpretação de texto escrito. Nas Ciências Sociais Aplicadas, essa abordagem tem se tornado mais relevante, proporcionando uma nova dimensão na pesquisa e compreensão de uma variedade de tópicos. Explora-se a seguir, como o *Text Learning* pode ser aplicado no processo de pesquisa.

Text Learning envolve a utilização de *algoritmos* e técnicas de Processamento de Linguagem Natural (PLN) para extrair informações, padrões e *insights* de documentos de texto. A capacidade do *Text Learning* de compreender a linguagem humana se torna uma ferramenta importante na análise de vastos conjuntos de dados textuais encontrados em documentos, redes sociais, entrevistas, discursos e muito mais.

Nesse sentido, o *Text Learning* pode ser utilizado para:

- **Análise de Opiniões e Sentimentos:** Em pesquisas de mercado e estudos de opinião pública, o *Text Learning* é usado para analisar opiniões e sentimentos expressos em redes sociais, comentários de clientes e pesquisas online. A partir desta análise, é possível ter uma compreensão mais profunda da percepção pública sobre produtos, serviços ou questões sociais (Wankhade; Rao; Kulkarni, 2022).

- **Revisão de Literatura Automatizada:** Na revisão de literatura em Ciências Sociais, o *Text Learning* pode ser aplicado para identificar e resumir automaticamente artigos científicos relevantes em uma área específica. Por meio desta revisão, é possível economizar tempo e ajudar os pesquisadores a manter-se atualizados com os avanços em seu campo (Portenoy; West, 2020).

- **Extração de Conceitos e Relações:** Em estudos das Ciências Sociais Aplicadas, o *Text Learning* pode ser usado para identificar conceitos-chave e relações entre entidades em documentos políticos, discursos políticos ou registros de reuniões, contribuindo para uma análise mais profunda de eventos políticos e sociais.

- **Análise de Texto Qualitativo:** Em pesquisas qualitativas, o *Text Learning* pode auxiliar na categorização e organização de dados de entrevistas, permitindo que os pesquisadores identifiquem tendências e padrões em narrativas humanas. (Rutkowski, 2022).

- **Detecção de Discurso de Ódio e Preconceito:** Em estudos sobre discriminação e preconceito, o *Text Learning* é utilizado para identificar e classificar discurso de ódio em textos *online*, contribuindo para uma análise mais aprofundada das dinâmicas sociais (Poletto *et al.*, 2021).

Destaca-se que o *Text Learning* é uma técnica que pode ser utilizada em uma série de ferramentas nas Ciências Sociais Aplicadas, permitindo que os pesquisadores analisem e compreendam melhor os conjuntos de dados textuais que permeiam as áreas de estudo. Essa abordagem está transformando a maneira como a pesquisa é conduzida, oferecendo novas maneiras de extrair *insights* significativos de documentos escritos e enriquecendo o conhecimento nas Ciências Sociais.

9.2.3 DATA MINING: DESCOBRINDO CONHECIMENTO EM DADOS MASSIVOS

Data Mining, ou mineração de dados, é uma técnica poderosa que desvenda padrões, relações e informações ocultas em grandes conjuntos de dados. Nas Ciências Sociais Aplicadas, essa abordagem se tornou uma ferramenta valiosa para compreender fenômenos sociais complexos e informar a pesquisa. A seguir, explora-se o *Data Mining* e como esta técnica pode ser aplicada no processo de pesquisa nas Ciências Sociais Aplicadas.

Data Mining envolve a análise sistemática de dados para identificar tendências, padrões, correlações e informações relevantes que não seriam facilmente percebidas com métodos convencionais.

Nesse contexto, pode-se fazer uso do *Data Mining* para:

- **Análise de Redes Sociais:** Em sociologia e estudos sociais, o *Data Mining* é usado para analisar redes sociais e interações humanas. Ele pode identificar influenciadores, comunidades e dinâmicas de grupo, fornecendo uma compreensão mais profunda das relações sociais e suas implicações (Serrat, 2017).

- **Previsão de Tendências Sociais:** O *Data Mining* é aplicado para prever tendências sociais, econômicas e políticas. Isso é útil em estudos de políticas públicas, onde os resultados podem informar decisões governamentais e alocar recursos de forma mais eficaz (Serrat, 2017).

- **Identificação de Fatores de Risco:** Em pesquisas de saúde pública e epidemiologia, o *Data Mining* ajuda a identificar fatores de risco em grandes conjuntos de dados de saúde, contribuindo para a prevenção e o controle de doenças.

- **Detecção de Anomalias:** Em segurança cibernética e na detecção de fraudes, o *Data Mining* é utilizado para identificar comportamentos anômalos em transações financeiras e atividades *online*, protegendo sistemas contra ameaças (Pang *et al.*, 2021).

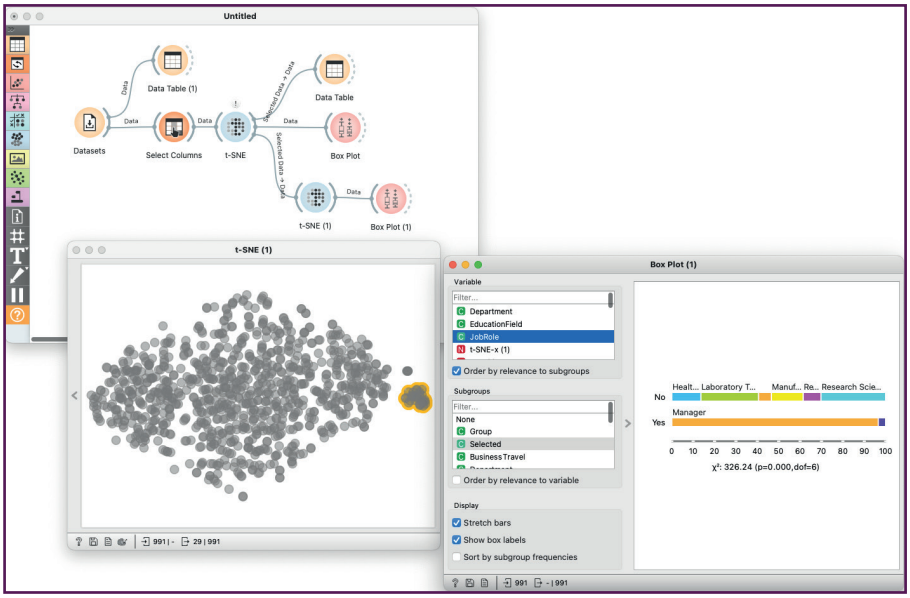
- **Modelagem de Comportamento:** Em psicologia e estudos comportamentais, o *Data Mining* pode ser aplicado para modelar o comportamento humano em diferentes contextos, ajudando a entender e prever respostas a estímulos específicos.

O *Data Mining* oferece uma nova perspectiva na pesquisa nas Ciências Sociais Aplicadas, permitindo que os pesquisadores descubram conhecimento valioso e padrões ocultos em grandes conjuntos de dados. Essa abordagem está transformando a maneira como a pesquisa é conduzida nessas áreas, fornecendo uma base sólida para a tomada de decisões informadas e o avanço do conhecimento nas Ciências Sociais.

9.3 ORANGE DATA MINING

O *Orange Data Mining*, também conhecido como *Orange*, é uma ferramenta visual projetada para ajudar os profissionais de ciência de dados e pesquisadores a explorarem e analisarem os dados de forma eficiente. Tal ferramenta é de código aberto e oferece uma ampla gama de recursos para tarefas de mineração de dados, análise exploratória, modelagem de aprendizado de máquina e visualização de resultados, tornando-a uma escolha popular na comunidade de análise de dados. A Figura 1 apresenta algumas telas da ferramenta.

Figura 1 - Telas da ferramenta *Orange Data Mining*



Fonte: Orange Data Mining (2023)[100]

O *Orange* é conhecido por sua interface intuitiva de arrastar e soltar, que permite aos usuários criar fluxos de trabalho de análise de dados sem a necessidade de codificação extensiva. A ferramenta combina a facilidade de uso com a flexibilidade de personalização, tornando-o adequado tanto para iniciantes quanto para usuários avançados. Além disso, suporta a linguagem de programação *Python*, o que significa que você pode integrar facilmente *scripts Python* personalizados em seus projetos *Orange*.

9.3.1 FUNCIONAMENTO DA FERRAMENTA

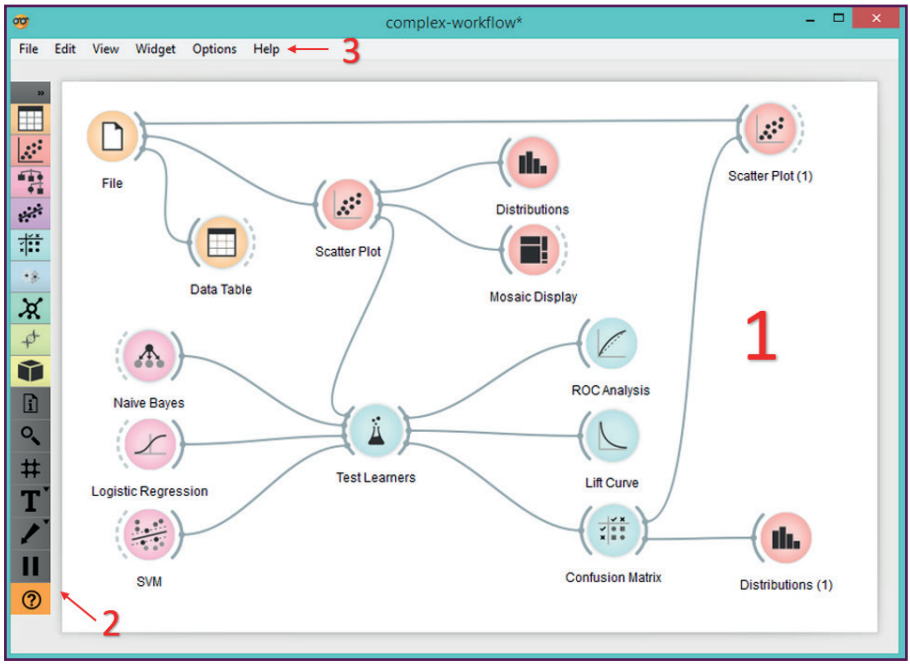
Antes de apresentar o potencial da ferramenta, apresenta-se os principais componentes da interface do *Orange*, buscando demonstrar a navegação e a organização da ferramenta.

100 Disponível em: https://orangedatamining.com/. Acesso em: 28 set. 2023.

9.3.1.1 A INTERFACE ORANGE

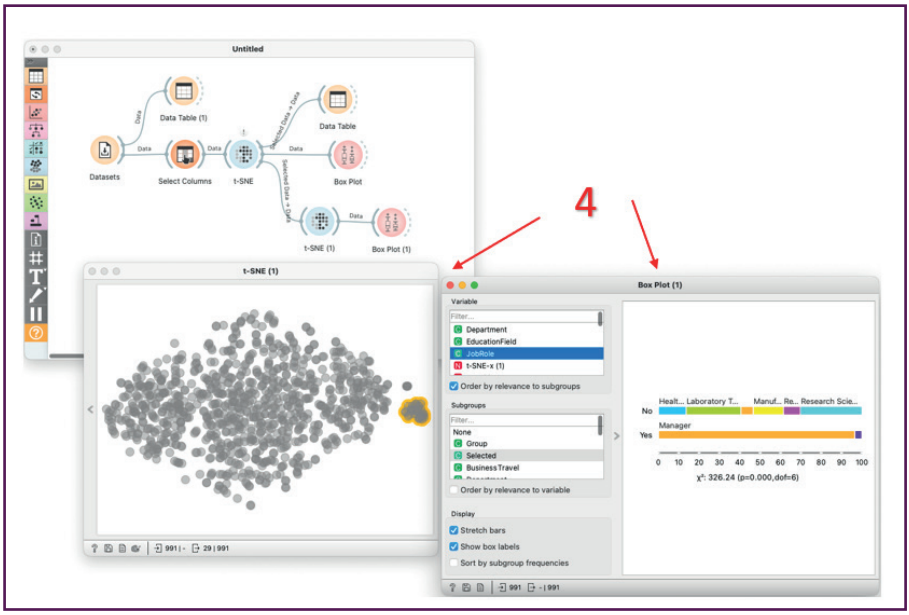
Ao iniciar o *Orange*, é possível ter controle de uma interface amigável composta por várias janelas e painéis. Neste texto, serão identificados numericamente os elementos principais, como ilustrado nas Figuras 2 e 3.

Figura 2 - Tela principal do *Orange Data Mining*



Fonte: Criado pelos autores (2023).

Figura 3 - Continuação da apresentação da tela principal do *Orange Data Mining*



Fonte: Criado pelos autores (2023).

- **Área de Canvas Principal:**
    - Esta é a zona central na qual são construídos e visualizados os fluxos de trabalho. É neste espaço que o usuário arrasta e solta os componentes para criar as suas análises.

- **Painel de Componentes:**
    - À esquerda da interface, o usuário encontra uma variedade de componentes que podem ser usados em seus fluxos de trabalho, como fontes de dados, *algoritmos* de aprendizado de máquina, visualizações e muito mais. Para utilizar tais componentes, basta arrastá-lo para a área de canvas para começar a construir seu fluxo de trabalho.

- **Barra de Ferramentas:**
    - A parte superior da interface contém uma barra de ferramentas com comandos como abrir, salvar e executar fluxos de trabalho, bem como outras funcionalidades essenciais.
- **Janelas de Visualização:**
    - O *Orange* oferece várias janelas de visualização para inspecionar seus dados e resultados, como gráficos, tabelas e painéis de pré-visualização de dados. Essas janelas são abertas automaticamente quando você executa componentes relevantes.

O *Orange Data Mining*, ainda, permite a personalização da interface de acordo com as preferências do usuário. Ademais, é possível ajustar a disposição das janelas, escolher um esquema de cores e até mesmo criar atalhos para as tarefas frequentes. Essa flexibilidade torna a experiência de uso do *Orange* altamente adaptável às suas necessidades específicas.

9.3.2 MANIPULAÇÃO DE DADOS

A manipulação de dados é uma etapa fundamental em qualquer projeto de análise de dados. O *Orange Data Mining* oferece uma série de recursos para importar, preparar e transformar dados, permitindo trabalhar de forma eficiente com conjuntos de dados de diferentes origens e formatos. Neste subcapítulo, explora-se em detalhes como o pesquisador pode realizar a manipulação de dados no *Orange*.

9.3.2.1 IMPORTAÇÃO DE CONJUNTO DE DADOS

A primeira etapa para qualquer análise de dados, é a importação dos conjuntos de dados para o *Orange*. A ferramenta suporta uma ampla variedade de formatos de arquivo, incluindo *CSV* (Comma-separated values), *Excel*, *SQL* (Structured Query Language), e até mesmo a conexão direta com fontes de dados *online*.

Carregar um conjunto de dados no *Orange* é uma tarefa simples, tendo que clicar no *widget 'File'* e selecionar o conjunto de dados na pasta em que ele está armazenado, tornando o processo semelhante ao abrir um arquivo no *Excel*. Aqui estão os passos básicos para importar um conjunto de dados:

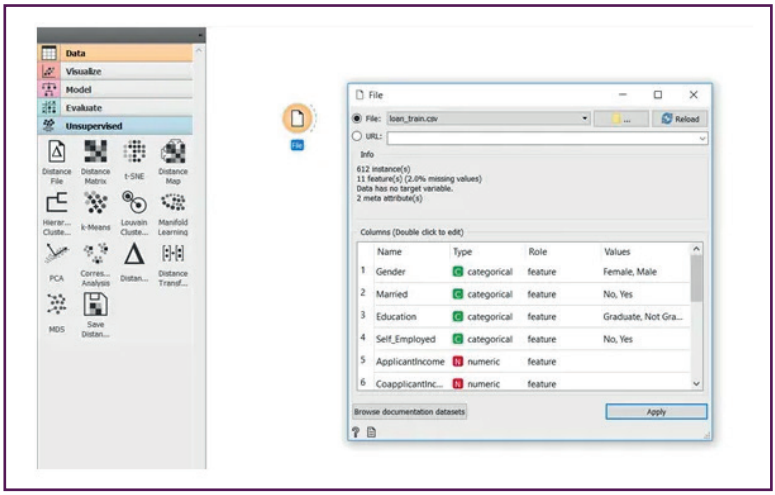
1. **Abrindo um Conjunto de Dados**: No painel de Componentes, tem um componente chamado "Arquivo" (File), sendo necessário arrastá-lo para a área de *Canvas*.

2. **Configurando o Componente de Arquivo**: Após selecionar componente de arquivo, é necessário escolher o arquivo desejado e definir as configurações de importação, como o tipo de delimitador (para *CSV*) ou as credenciais de conexão (para fontes de dados *online*).

3. **Conectando Componentes**: Para continuar a análise, há a necessidade de conectar o componente de arquivo a outros componentes, como gráficos ou *algoritmos* de aprendizado de máquina.

A Figura 4 exemplifica a tela de carregar informações dentro do elemento *File*.

Figura 4 - Tela de carregamento de dados



Fonte: Batista (2019)[101]

101 Disponível em: https://acesse.dev/1kWv4. Acesso em: 28 set. 2023.

9.3.2.2 LIMPEZA E TRANSFORMAÇÃO DE DADOS

Após o processo de importação de dados, a próxima etapa é limpar e transformá-los, preparando-os para análises mais avançadas. O *Orange* oferece uma variedade de componentes para ajudar nessa tarefa:

- **Filtros:** Utilização de filtros para remover dados irrelevantes, duplicados ou outliers do seu conjunto de dados.
- **Normalização e Padronização:** Normalizar ou padronizar atributos numéricos é essencial para muitos *algoritmos* de aprendizado de máquina. O *Orange* fornece componentes para executar essas operações.
- **Transformação de Atributos:** É possível criar novos atributos ou transformar atributos existentes usando funções matemáticas, de texto ou lógicas. Tal transformação ajuda na criação de características mais significativas para sua análise.
- **Amostragem:** Para grandes conjuntos de dados, pode ser útil realizar amostragens aleatórias ou estratificadas para tornar a análise mais ágil e economizar recursos computacionais.

9.3.2.3 SELEÇÃO DE ATRIBUTOS

Destaca-se que nem todos os atributos dos dados são igualmente informativos para as análises. Desta forma, algumas vezes, é necessário selecionar um subconjunto relevante de atributos para melhorar a eficiência do seu modelo. O *Orange* oferece maneiras de fazer isso:

- **Seleção Manual:** É possível selecionar manualmente os atributos que o usuário deseja manter ou remover da análise.
- **Seleção Automática:** O *Orange* também oferece métodos de seleção automática de atributos que identificam os atributos mais importantes com base em critérios estatísticos.

- **Redução de Dimensionalidade:** Em casos de conjuntos de dados de alta dimensionalidade, a redução de dimensionalidade pode ser aplicada para preservar informações essenciais enquanto reduz a complexidade.

A manipulação de dados é uma fase crítica em qualquer projeto de análise de dados, e o *Orange Data Mining* facilita essas tarefas. Compreender como importar, limpar, transformar e selecionar dados é fundamental para preparar seus dados para análises mais avançadas, como modelagem de aprendizado de máquina.

9.3.3 ANÁLISE EXPLORATÓRIA DE DADOS

Na sequência, tem-se uma etapa crítica em qualquer projeto de análise de dados, que é a análise exploratória de dados (AED). Esta etapa envolve a exploração e compreensão inicial dos dados antes de aplicar *algoritmos* de aprendizado de máquina ou realizar análises estatísticas mais avançadas. O *Orange Data Mining* oferece um conjunto robusto de ferramentas e recursos para ajudar na AED, permitindo que os usuários investiguem os dados, identifiquem padrões, avaliem a qualidade e ganhem *insights* valiosos. Neste capítulo, explora-se as diversas maneiras de realizar uma AED eficaz com o *Orange*.

9.3.3.1 TIPOS DE EXPLORAÇÃO VISUAL DOS DADOS

Uma das características distintivas do *Orange* é sua capacidade de permitir que os usuários explorem visualmente os dados de forma mais simples e intuitiva. A seguir, apresenta-se algumas destas opções:

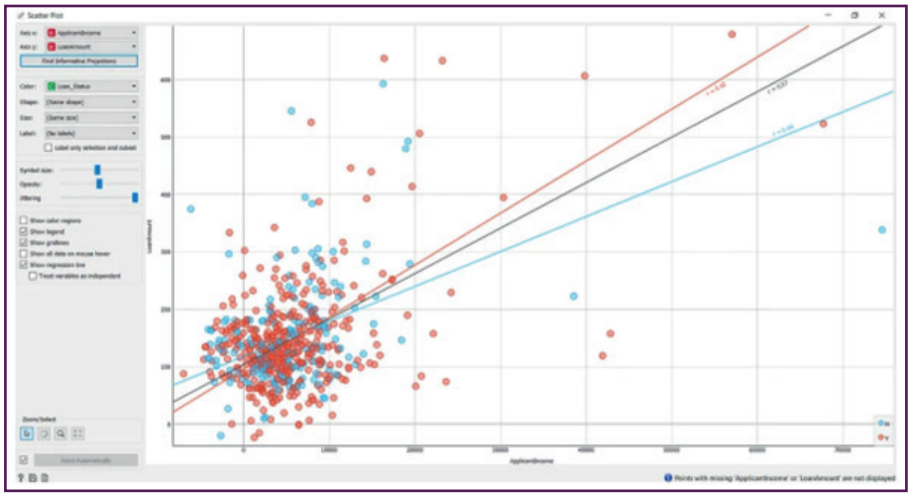
- **Gráficos de Dispersão (Scatter Plots)**: Os gráficos de dispersão são úteis para visualizar relações entre pares de atributos. No *Orange*, é possível criar gráficos de dispersão por meio dos atributos no *widget* de gráfico de dispersão.

- **Histogramas**: Os histogramas ajudam a entender a distribuição dos valores em um atributo. Eles podem ser criados com o *widget* de histograma e são uma ferramenta essencial para identificar a forma das distribuições de dados.

- **Visualizações de Dados Multidimensionais**: Com o *widget* "Projection", é possível criar visualizações de dados multidimensionais, como *PCA* (Análise de Componentes Principais) ou *t-SNE* (t-Distributed Stochastic Neighbor Embedding), para reduzir a dimensionalidade e visualizar agrupamentos e padrões em dados de alta dimensão.

A Figura 5 apresenta a tela de *scatter plot:*

Figura 5 - Tela de *scatter plot*



Fonte: Batista (2019)[102]

9.3.3.2 ESTATÍSTICAS DESCRITIVAS

Além da visualização, o *Orange* fornece informações estatísticas detalhadas sobre seus dados. Isso inclui:

- **Estatísticas Básicas**: É possível obter estatísticas descritivas básicas, como média, mediana, desvio padrão, mínimo e máximo para atributos numéricos.

102 Disponível em: https://encr.pw/1kWv4. Acesso: em 28 set. 2023.

- **Distribuição de Classe**: Em conjuntos de dados de classificação, é possível visualizar a distribuição de classes para entender o desequilíbrio de classes.

- **Correlação**: O *Orange* permite calcular a matriz de correlação para avaliar as relações entre os atributos numéricos.

A detecção de *outliers* e anomalias é fundamental na AED. O *Orange* oferece métodos e ferramentas para identificar pontos de dados que podem ser considerados valores atípicos:

- **Box Plots**: O *widget* "Box Plot" permite criar gráficos de caixa para identificar *outliers* em atributos numéricos.

- **Scatter Plots de Distância Mahalanobis**: Esses gráficos de dispersão ajudam a identificar pontos de dados que estão longe da média multivariada.

Vale destacar que tais análises não são válidas apenas para para dados numéricos, o *Orange* também fornece recursos para explorar dados categóricos:

- **Histogramas Categóricos**: Você pode criar histogramas para variáveis categóricas para entender a distribuição de categorias.

- **Tabelas de Contingência**: As tabelas de contingência ajudam a analisar as relações entre variáveis categóricas, destacando associações e dependências.

O *Orange Data Mining* oferece, ainda, outros recursos para análises geoespaciais, permitindo a obtenção de *insights* a partir de dados de localização. Além desta, os mapas de calor são eficazes para visualizar densidades geográficas e tendências em dados de localização. Com eles, é possível identificar áreas de alta atividade, concentração ou distribuição de eventos geográficos. O *Orange*, ainda, facilita a criação de mapas de calor interativos que destacam as áreas com maior densidade de ocorrências, tornando mais simples a identificação de padrões geoespaciais. Além disso, o *Orange* permite o mapeamento de pontos de dados diretamente em mapas interativos. Essa funcionalidade é especialmente útil para entender a distribuição espacial de eventos ou informações geográficas. Ao visualizar pontos de dados em um mapa, é possível explorar relações entre localizações e identificar *clusters* ou áreas de interesse. Essa compreensão espacial

mais profunda pode ser valiosa em uma variedade de aplicações, como análises de negócios, geografia de saúde pública ou pesquisa ambiental. Com a combinação dessas ferramentas, é possível ter uma visão para análise de dados geoespaciais de maneira eficaz, revelando informações importantes sobre padrões e tendências em informações de localização.

A análise exploratória de dados é uma etapa crucial para compreender a natureza dos seus dados e identificar padrões ou anomalias que podem orientar o restante do seu projeto de análise de dados. Com as ferramentas e recursos do *Orange*, é possível realizar uma AED eficaz de forma intuitiva e informada, estabelecendo uma base sólida para análises mais avançadas e modelagem de aprendizado de máquina.

9.3.4 MODELAGEM DE APRENDIZADO DE MÁQUINA

A modelagem de aprendizado de máquina é uma das etapas com mais potencial no contexto da análise de dados. Tal etapa permite a construção de modelos preditivos a partir de dados, o que pode ser aplicado a uma ampla variedade de cenários, desde classificação de *e-mails* como *spam* ou *não spam* até previsões de vendas de produtos. O *Orange Data Mining* oferece uma ampla gama de *algoritmos* de aprendizado de máquina e ferramentas para criar, avaliar e ajustar esses modelos. Assim, a seguir explora-se em detalhes como realizar modelagem de aprendizado de máquina no *Orange*.

9.3.4.1 ESCOLHA DE ALGORITMO

Na etapa de modelagem de aprendizado de máquina, a escolha do *algoritmo* apropriado é fundamental para o sucesso do projeto. O *Orange* oferece uma variedade de *algoritmos* de classificação, regressão, *clustering* e associação. Alguns dos *algoritmos* mais utilizados e destacados incluem:

- **Regressão Linear:** este *algoritmo* é utilizado para modelar relações lineares entre variáveis dependentes e independentes, sendo especialmente útil quando se deseja entender como uma variável afeta outra de maneira linear.

- **Árvores de Decisão**: árvores de decisão são excelentes escolhas para problemas de classificação, além de oferecerem uma interpretabilidade natural do modelo, permitindo compreender como as decisões são tomadas.

- **Random Forests**: esta é uma abordagem de *ensemble* que combina várias árvores de decisão para melhorar o desempenho do modelo. É particularmente eficaz para reduzir o *overfitting* e aumentar a precisão.

- **K-Means Clustering:** um *algoritmo* amplamente utilizado para tarefas de *clustering*, o *K-Means* segmenta os dados em grupos com base em suas similaridades. É valioso para identificar padrões e estruturas em dados não rotulados.

- **Regras de Associação:** essas regras são úteis para descobrir padrões de associação em conjuntos de dados, como análises de cestas de compras, onde se deseja entender quais itens tendem a ser comprados juntos.

A escolha do *algoritmo* adequado dependerá do tipo de problema enfrentado e dos objetivos da análise. Experimentar diferentes *algoritmos* e avaliar seu desempenho com métricas apropriadas, como mencionadas anteriormente, é uma prática comum para determinar qual *algoritmo* se adapta melhor aos seus dados e metas.

A avaliação de modelo desempenha um papel crucial na construção de sistemas de aprendizado de máquina robustos e eficazes. Esta permite medir o desempenho de um modelo e identificar áreas para melhorias. No *Orange Data Mining*, é possível encontrar um conjunto de ferramentas abrangentes para avaliar modelos de classificação, regressão, *clustering* e associação. Apresenta-se na sequência como realizar uma avaliação completa de modelo.

9.3.4.2 MÉTRICAS DE AVALIAÇÃO DE CLASSIFICAÇÃO

Para modelos de classificação, a escolha das métricas apropriadas é essencial para entender o quão bem o modelo está funcionando. Algumas das métricas de avaliação de classificação mais comuns incluem:

- **Precisão (Accuracy)**: Mede a proporção de instâncias classificadas corretamente em relação ao total de instâncias.

- **Recall (Sensibilidade)**: Calcula a proporção de instâncias positivas corretamente classificadas em relação ao total de instâncias positivas.

- **F1-Score**: Uma métrica que combina precisão e *recall*, sendo útil quando há desequilíbrio entre classes.

- **Matriz de Confusão**: Uma tabela que mostra as classificações corretas e incorretas feitas pelo modelo, fornecendo uma visão detalhada do desempenho em diferentes categorias.

- **Curvas ROC e AUC**: Permitem avaliar o desempenho do modelo em diferentes limiares de classificação e medir a capacidade de discriminação do modelo. A curva *ROC* (Receiver Operating Characteristic Curve) representa a taxa de verdadeiros positivos em relação à taxa de falsos positivos, enquanto a *AUC* (Area under the ROC Curve) quantifica a qualidade geral do modelo, sendo um valor entre 0 e 1, onde valores mais próximos de 1 indicam um modelo melhor.

Para modelos de regressão, as métricas de avaliação focam na precisão das previsões. Alguns exemplos de métricas de avaliação de regressão incluem:

- **Erro Quadrático Médio (RMSE)**: Mede a média dos erros quadrados das previsões em relação aos valores reais, fornecendo uma medida de quão bem as previsões se ajustam aos dados. Quanto menor o *RMSE* (Root Mean Squared Error), mais precisas são as previsões.

- **Erro Absoluto Médio (MAE)**: Calcula a média dos valores absolutos das diferenças entre as previsões e os valores reais, oferecendo uma medida direta da magnitude média dos erros de previsão. O *MAE* (Mean Absolute Error) é útil para entender o tamanho médio dos erros.

- **Coeficiente de Determinação (R^2)**: Avalia a proporção da variabilidade nos dados explicada pelo modelo. O R^2 varia de 0 a 1, onde 1 indica que o modelo explica toda a variabilidade e 0 indica que o modelo não explica nenhuma. É uma métrica importante para determinar o ajuste global do modelo aos dados.

Além destas, apresenta-se com mais detalhes, algumas técnicas que podem ser utilizadas para validação:

- **Validação Cruzada**

  - A validação cruzada é uma técnica essencial para avaliar o desempenho do modelo em dados não vistos. O *Orange* suporta diferentes estratégias de validação cruzada, como validação cruzada *k-fold* e validação cruzada estratificada. Essas abordagens auxiliam na prevenção do superajuste (overfitting) e permitem obter uma estimativa mais confiável do desempenho do modelo em dados futuros. Ao dividir o conjunto de dados em partes menores e testar o modelo em diferentes combinações de treinamento e teste, a validação cruzada fornece uma avaliação mais sólida da capacidade do modelo de generalizar para novos dados, aumentando a confiabilidade de suas conclusões e previsões.

- **Ajuste de Hiperparâmetros**

  - Os hiperparâmetros de um modelo são configurações que podem afetar significativamente seu desempenho. O *Orange* oferece ferramentas para ajustar automaticamente esses hiperparâmetros, como a busca em grade (grid search) e a otimização *bayesiana*. Isso permite encontrar a combinação ideal de configurações para o modelo, maximizando sua capacidade de generalização e precisão. O ajuste de hiperparâmetros é uma etapa crucial no desenvolvimento de modelos de aprendizado de máquina robustos e eficazes, e o *Orange* simplifica esse processo, economizando tempo e esforço dos cientistas de dados e pesquisadores.

- **Visualização de Resultados**

  - Além das métricas numéricas, o *Orange* fornece visualizações interativas para auxiliar na análise dos resultados do modelo. Por exemplo, é possível visualizar as curvas *ROC*, matrizes de confusão e gráficos de dispersão de resíduos para obter uma compreensão mais profunda do desempenho do modelo. Essas visualizações não apenas tornam os resultados mais acessíveis, mas também permitem identificar tendências,

anomalias ou áreas que podem exigir ajustes no modelo, contribuindo assim para aprimorar sua eficácia.

- **Exportação e Implantação**

    - Uma vez que se tenha construído e avaliado o modelo com sucesso no *Orange*, a ferramenta permite exportá-lo para ser utilizado em outros contextos. É possível exportar os modelos treinados em *Python* e integrá-los em aplicações ou fluxos de trabalho de produção.

Destaca-se que com as ferramentas e métricas disponíveis no *Orange Data Mining*, você pode medir, aperfeiçoar e compreender o desempenho dos seus modelos em detalhes. Esse processo é fundamental para assegurar que suas soluções de aprendizado de máquina atendam aos requisitos de qualidade e confiabilidade, garantindo que eles sejam eficazes e precisos em suas aplicações práticas.

9.4 CONSIDERAÇÕES

O capítulo ressalta a influência positiva da Inteligência Artificial (IA) nas Ciências Sociais Aplicadas, especificamente por meio das técnicas de *Machine Learning*, *Text Learning* e *Data Mining*. Essas tecnologias têm o potencial de revolucionar a pesquisa nesse campo, permitindo uma compreensão mais profunda e precisa de fenômenos sociais e comportamento humano.

Machine Learning oferece a capacidade de lidar com grandes volumes de dados, identificando padrões e fazendo previsões precisas, com aplicações em diversas áreas, como mercado de trabalho e análise de opiniões públicas. O *Text Learning* automatiza a revisão de literatura, economizando tempo na identificação de artigos relevantes, além de analisar sentimentos e opiniões expressos em textos. O *Data Mining* revela padrões em grandes conjuntos de dados, permitindo uma compreensão mais profunda das interações sociais e a previsão de tendências.

Importante destacar que essas técnicas não substituem os métodos tradicionais de pesquisa, mas complementam, sendo a experiência humana e

a ética essenciais na pesquisa científica. Questões éticas e de privacidade relacionadas ao uso de dados também devem ser consideradas.

O *Orange* é apresentado como uma ferramenta valiosa não apenas para profissionais e pesquisadores em ciência de dados, mas também para outros profissionais e pesquisadores das ciências sociais aplicadas. Ele possui uma interface intuitiva, suporte à linguagem *Python* e flexibilidade na personalização. A ferramenta abrange todas as fases do processo de análise de dados, desde a importação até a modelagem de aprendizado de máquina, com ênfase na avaliação de modelos.

Em suma, a IA, por meio das técnicas mencionadas, está enriquecendo a pesquisa nas Ciências Sociais Aplicadas, oferecendo novas abordagens para análise de dados. O *Orange Data Mining*, com sua usabilidade, flexibilidade e recursos, é uma ferramenta importante para profissionais e pesquisadores envolvidos em análise de dados, contribuindo para avanços nesse campo.

REFERÊNCIAS

ALINEJAD-ROKNY, Hamid; SADRODDINY, Esmaeil; SCARIA, Vinod. Machine learning and data mining techniques for medical complex data analysis. **Neurocomputing**, [*s. l.*], v. 276, n. 1, p. 1-2, 2018. DOI https://doi.org/10.1016/j.neucom.2017.09.027. Disponível em: https://www.sciencedirect.com/science/article/pii/S0925231217315400. Acesso em: 28 set. 2023.

CARDI, Marilza de Lourdes; BARRETO, Jorge Muniz. Primórdios da computação no Brasil. *In*: SIMPÓSIO DE HISTÓRIA DA INFORMÁTICA NA AMÉRICA LATINA E CARIBE (SHIALC), 2.; CLEI, 38., Medellin, 2012. **Anais** [...]. [*S. l.: s. n.*], 2012. Disponível em: https://www.cos.ufrj.br/shialc/2012/content/docs/shialc_2/clei2012_submission_126.pdf. Acesso em: 28 set. 2023.

ERNAWATI, E.; BAHARIN, S. S. K.; KASMIN, F. A review of data mining methods in RFM-based customer segmentation. **Journal of Physics**: Conference Series, [*s. l.*], v. 1869. p. 012085, 2021. Disponível em: https://iopscience.iop.org/article/10.1088/1742-6596/1869/1/012085/meta. Acesso em: 28 set. 2023.

GARCÍA, Salvador; LUENGO, Julián; HERRERA, Francisco. **Data preprocessing in data mining**. Cham: Springer, 2015. (Intelligent Systems Reference Library, v. 72). Disponível em: https://content.e-bookshelf.de/media/reading/L-3926777-b03bc1919c.pdf. Acesso em: 28 set. 2023.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, Washington, v. 349, n. 6245, p. 255-260, July 2015. Disponível em: https://www.cs.cmu.edu/~tom/pubs/Science-ML-2015.pdf. Acesso em: 30 ago. 2023.

LIM, Bryan; ZOHREN, Stefan. Time-series forecasting with deep learning: a survey. **Philosophical Transactions of the Royal Society A**, Londres, v. 379, n. 2194, p. 20200209, 2021. DOI https://doi.org/10.1098/rsta.2020.0209. Disponível em: https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0209. Acesso em: 30 ago. 2023.

PANG, Guansong; SHEN, Chunhua; CAO, Longbing; VAN DEN HENGEL, Anton. Deep learning for anomaly detection: A review. **ACM Computing Surveys**, [*s. l.*], v. 54, n. 2, p. 1-38, 2021. Disponível em: https://dl.acm.org/doi/10.1145/3439950. Acesso em: 30 ago. 2023.

POLETTO, Fabio; BASILE, Valerio; SANGUINETTI, Manuela; BOSCO, Cristina; PATTI, Viviana. Resources and benchmark corpora for hate speech detection: a systematic review. **Language Resources and Evaluation**, London, v. 55, p. 477-523, 2021. DOI https://doi.org/10.1007/s10579-020-09502-8 . Disponível em: https://link.springer.com/article/10.1007/s10579-020-09502-8. Acesso em: 28 set. 2023.

PORTENOY, Jason; WEST, Jevin D. Constructing and evaluating automated literature review systems. **Scientometrics**, Dordrecht, v. 125, n. 3, p. 3233-3251, 2020. DOI https://doi.org/10.1007/s11192-020-03490-w . Disponível em: https://link.springer.com/article/10.1007/s11192-020-03490-w . Acesso em: 25 set. 2023.

POURHABIBI, Tahereh; ONG, Kok-Leong; KAM, Booi H; BOO, Yee Ling. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. **Decision Support Systems**, [*s. l.*], v. 133, p. 113303, 2020. DOI https://doi.org/10.1016/j.dss.2020.113303.

Disponível em: https://www.sciencedirect.com/science/article/pii/S0167923620300580. Acesso em: 24 set. 2023.

RUTKOWSKI, Rachel A.; LEE, John D.; COLLER, Ryan J.; WERNER, Nicole E. How can text mining support qualitative data analysis?. *In*: HUMAN FACTORS AND ERGONOMICS SOCIETY INTERNATIONAL ANNUAL MEETING, 66th, Atlanta, 2022. **Proceedings** [...]. Los Angeles, CA: SAGE Publications, 2022. p. 2319-2323. Disponível em: https://journals.sagepub.com/doi/pdf/10.1177/1071181322661535. Acesso em: 1 out. 2023.

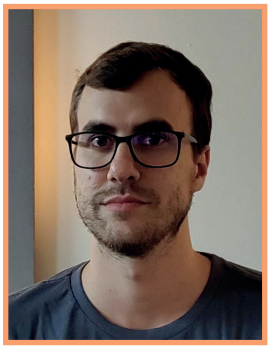
SERRAT, Olivier. Social Network Analysis. *In*: SERRAT, O. (ed.). **Knowledge Solutions**: Tools, Methods, and Approaches to Drive Organizational Performance. Singapore: Springer, 2017. p. 39-43.

TOMMASEL, Antonela; GODOY, Daniela. Short-text learning in social media: a review. **The Knowledge Engineering Review**, Cambridge, v. 34, p. e7, 2019. DOI https://doi.org/10.1017/S0269888919000018. Disponível em: https://www.cambridge.org/core/journals/knowledge-engineering-review/article/shorttext-learning-in-social-media-a-review/F2A5A1F47D512C94BD265A2D63AFF593. Acesso em: 1 out. 2023.

WANKHADE, Mayur; RAO, Annavarapu Chandra Sekhara; KULKARNI, Chaitanya. A survey on sentiment analysis methods, applications, and challenges. **Artificial Intelligence Review**, Oxford, v. 55, n. 7, p. 5731-5780, 2022. DOI https://doi.org/10.1007/s10462-022-10144-1. Disponível em: https://link.springer.com/article/10.1007/s10462-022-10144-1. Acesso em: 27 set. 2023.

DADOS DOS AUTORES:

Caio Saraiva Coneglian



Caio Saraiva Coneglian é Doutor e mestre em Ciência da Informação. Bacharel em Ciência da Computação. Docente da Universidade de Marília - UNIMAR. Pesquisador do Instituto Brasileiro de Informação em Ciência e Tecnologia- IBICT. Docente colaborador do PPGCI - UNESP.

https://orcid.org/0000-0002-6126-9113
caio.coneglian@gmail.com

Henrique Leal Tavares

Henrique Leal Tavares é Mestre em Ciência da Computação. Tecnólogo em Análise e Desenvolvimento de Sistemas. Docente da Universidade de Marília - UNIMAR.

https://orcid.org/0009-0006-8960-3386
henriquetavares@unimar.br

Diego José Macêdo

Diego José Macêdo é Mestre em Ciência da Informação pela Universidade de Brasília. Bacharel em Sistema de Informação pela Universidade Católica de Brasília. Atualmente é tecnologista do Instituto Brasileiro de Informações em Ciência e Tecnologia - Ibict.

diegomacedo@ibict.br
https://orcid.org/0000-0002-5696-0639

Milton Shintaku

Milton Shintaku é Doutor em Ciência da Informação pela Universidade de Brasília. Coordenador de Tecnologia para Informação (Cotec) do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

shintaku@ibict.br
https://orcid.org/0000-0002-6476-4953

Como referenciar o capítulo 9:

CONEGLIAN, Caio Saraiva; TAVARES, Henrique Leal; MACÊDO, Diego José; SHINTAKU, Milton. Orange Data Mining: Uma ferramenta para inserção de inteligência artificial na pesquisa científica. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 9. p. 245-273. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap9.