Universidade Estadual de Londrina

MURILO AUGUSTO MAESTRO

ADAPTAÇÃO DO MODELO PREDICTOUR: SERÁ POSSÍVEL PREVER COMPORTAMENTO HUMANO DE MANEIRA VERSÁTIL EM DIFERENTES CONTEXTOS?

LONDRINA
2023

MURILO AUGUSTO MAESTRO

ADAPTAÇÃO DO MODELO PREDICTOUR: SERÁ POSSÍVEL PREVER COMPORTAMENTO HUMANO DE MANEIRA VERSÁTIL EM DIFERENTES CONTEXTOS?

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Dra. Helen C. de Mattos Senefonte

LONDRINA
2023

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

Sobrenome, Nome.
Título do Trabalho : Subtitulo do Trabalho / Nome Sobrenome. - Londrina, 2017.
100 f. : il.

Orientador: Nome do Orientador Sobrenome do Orientador.
Coorientador: Nome Coorientador Sobrenome Coorientador.
Dissertação (Mestrado em Ciência da Computação) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Programa de Pós-Graduação em Ciência da Computação, 2017.
Inclui bibliografia.

1. Assunto 1 - Tese. 2. Assunto 2 - Tese. 3. Assunto 3 - Tese. 4. Assunto 4 - Tese. I. Sobrenome do Orientador, Nome do Orientador. II. Sobrenome Coorientador, Nome Coorientador. III. Universidade Estadual de Londrina. Centro de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação. IV. Título.

Dedico esse trabalho a todos os eus que escolheram os caminhos que me trouxeram até aqui, e também aos meu futuro eu, a quem envio este presente.

AGRADECIMENTOS

A quem devo agradecer? Ao reflexo de quem são as pessoas em seus comportamentos, ou às pessoas que permitiram serem refletidas em suas ações? Agradeço a todos que, com suas atitudes e escolhas, compartilharam o melhor de si e me permitiram crescer e evoluir.

Nesta jornada, cada interação, cada lição aprendida com os outros, torna-se parte de mim e, por sua vez, ecoa adiante, formando um legado eterno. Portanto, agradeço a todos que, ao revelarem suas melhores versões, permitiram que eu me tornasse uma pessoa melhor.

Que continuemos a compartilhar e refletir o melhor de nós, em um ciclo contínuo de aprendizado e crescimento, e que cada um de nós transmita o que há de melhor para as gerações futuras.

"O comportamento é um espelho em que cada um observa a sua própria imagem."
(Johann Wolfgang von Goethe - 1906)

MAESTRO, M. A.. **Adaptação do Modelo PredicTour: Será Possível Prever Comportamento Humano de Maneira Versátil em Diferentes Contextos?**. 2023. 46f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Estadual de Londrina, Londrina, 2023.

RESUMO

A previsão do comportamento humano é de grande interesse para diversas áreas, e uma ampla gama de técnicas e modelos com diferentes estruturas e métodos são empregados para estimar esses comportamentos com precisão. Um exemplo notável é o *PredicTour*, um modelo desenvolvido por Senefonte et al.[1], que visa prever a mobilidade urbana de turistas utilizando técnicas de classificação pouco convencionais no campo. Este estudo se propõe a avaliar a capacidade preditiva do modelo *PredicTour*, analisando se a adaptação direta de sua estrutura mantém sua eficácia e versatilidade. Os resultados obtidos indicam que o modelo consegue realizar predições satisfatórias, mesmo diante dos desafios impostos pela qualidade dos dados utilizados no treinamento. Essa descoberta destaca a potencial aplicabilidade do *PredicTour* em contextos variados e sua capacidade de lidar com situações complexas de previsão de comportamento humano.

Palavras-chave: PredicTour; Comportamento humano; Previsão de comportamento; Adaptação de modelo; Análise de dados; Aprendizado de máquina;

MAESTRO, M. A.. **Adapting the PredicTour Model: Can Versatile Human Behavior Prediction Be Achieved in Different Contexts?**. 2023. 46p. Final Project (Bachelor of Science in Computer Science) – State University of Londrina, Londrina, 2023.

ABSTRACT

Predicting human behavior is of great interest to many fields, and a wide range of techniques and models with different structures and methods are employed to accurately estimate these behaviors. A notable example is the *PredicTour*, a model developed by Senefonte et al.[1], which aims to predict the urban mobility of tourists using unconventional classification techniques in the field. This study proposes to evaluate the predictive capacity of the *PredicTour* model, analyzing whether the direct adaptation of its structure maintains its effectiveness and versatility. The results obtained indicate that the model is capable of making satisfactory predictions, even in the face of challenges imposed by the quality of the data used in training. This finding highlights the potential applicability of *PredicTour* in various contexts and its ability to deal with complex situations of predicting human behavior.

Keywords: PredicTour; Human behavior; Behavior prediction; Model adaptation; Data analysis; machine learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ALGORÍTMOS

SUMÁRIO

1 INTRODUÇÃO

A imprevisibilidade humana surge das diferenças individuais em contextos, experiências e características genéticas[2]. O desenvolvimento de inteligências artificiais tem como um dos principais objetivos alcançar resultados semelhantes ou idênticos aos humanos. Essas inteligências envolvem uma ampla variedade de modelos, estilos, campos e aplicações, com características distintas que determinam seu desempenho nas tarefas designadas. No entanto, muitos não atingem um nível satisfatório de respostas aceitáveis, enquanto outros excedem a capacidade humana, o que pode gerar críticas negativas.

Prever o comportamento humano é uma tarefa complexa e valiosa. Se uma máquina pudesse prever com precisão o comportamento humano, poderia simular escolhas que um indivíduo faria em diferentes situações. As aplicações potenciais dessa habilidade são vastas, incluindo previsão de criminalidade, mobilidade urbana, padrões de consumo, risco de suicídio e resultados de jogos esportivos e eletrônicos, entre outras áreas envolvendo decisões e ações humanas.

A previsibilidade do comportamento humano é um tema explorado em diversos filmes, séries de TV, livros e outras mídias, como no filme *Minority Report* [3]. O filme apresenta o dilema da previsibilidade, levantando questões sobre a possibilidade de previsões futuras divergirem. Essa questão nos faz refletir se todos os modelos, devidamente adaptados para diferentes problemas, podem convergir para a mesma resposta dada uma entrada de dados específica ou se sempre haverá modelos que apresentam melhor desempenho e são mais adequados para problemas específicos.

Dentre diversos modelos, propostos por diversos cientistas, um que se destaca é o *PredicTour*, proposto por Senefonte et al.[1], que visa fazer a predição da mobilidade urbana de turistas. Sua maneria não convencional de usar modelos de classificação para realizar predições de comportamento amostra uma capacidade interessante de realizar predições de comportamento. Sua versatilidade e estrutura que contem modelos clássicos levanta uma questão: seria possível adaptar este modelo para lidar com outros tipos de dados em outros contextos? Nos capítulos deste trabalho será descrito, discutido e trabalhado os conceitos principais do *PredicTour* e suas capacidades, além de observado como ele, apesar de uma alta dificuldade de treinamento, ainda obteve estimativas razoáveis.

2 FUNDAMENTAÇÃO

Ao longo dos anos, inúmeros estudos foram realizados visando prever o comportamento humano. Essa tarefa é complexa, uma vez que os indivíduos podem apresentar comportamentos pontuais que divergem de suas ações históricas [4][5]. Diversos modelos de inteligência artificial foram propostos e treinados com diferentes conjuntos de dados, buscando desenvolver redes neurais capazes de prever, ou simular aproximadamente, respostas que seriam dadas por seres humanos. Esses modelos variam em termos de eficácia e estrutura, sendo classificados e agrupados com base em suas habilidades de previsão em contextos específicos, bem como na capacidade de generalização para outras situações. Na seção Seção 2.1 são apresentados, superficialmente, alguns dos trabalhos, técnicas e modelos e seus resultados obtidos.

Dentre os diversos trabalhos levantados durantes o estudo, vários são recentes e abordam temas modernos. Um destes trabalhos propõe o modelo *PredicTour*[1], onde a sua função primordial é realizar a previsão da mobilidade de turistas em seus destinos de viagem. Este trabalho recorre a técnicas clássicas de aprendizado de máquina para construir um algorítimo capaz de classificar indivíduos e estimar comportamentos. E esta característica de usar técnicas conhecidas em conjunção no contexto moderno de turismo, cria um destaque para os resultados obtidos do trabalho. Por estes motivos o trabalho é usado como base de estudo e implementado em um contexto diferente, a fim de avaliar sua capacidade de predição, outros contextos e conjunto de dados. Na Seção 2.5 é descrito o modelo, suas técnicas, e como sua estrutura consegue prever a mobilidade urbana.

2.1 Estado de arte

Um levantamento de trabalhos, com uma temática abrangente, sobre predição de eventos, comportamentos, situações, identificação de perfis, entre outras palavras-chave, revela diversas técnicas, algumas inéditas, outras adaptações, e outras ainda uso sem modificações de modelos clássicos de aprendizado de máquina. A seguir, Tabela 1 é apresentada uma tabela que lista de trabalhos, seu objetivo central, e técnicas e modelos principais usados.

Os trabalhos listados demonstram certas tendências de uso de modelos para predição. Em especial *Decision Trees*[10], *Random Forest*[12], *Support Vector Machines*[12] e *Naïve Bayes*[26], pois suas estruturas tem uma facilidade para lidar com objetivos de predição. Enquanto outros trabalhos recorrem a modelos não diretamente capazes de realizar predição de eventos, ou decisões, como o trabalho de Senefonte et al.[1]. Não cabe ao escopo deste trabalho a descrição completa de todos os modelos na tabela, nem dos

Tabela 1 – Trabalhos, suas temáticas e modelos utilizados

Trabalho	Descrição de objetivo	Modelos usados
2014 Chen et al.[6]	Predizer a bolsa de valores por meio de mídias sociais	FM[7]
2019 Kumar et al.[8]	Prever inteção de recompra de produtos	ABC[9], Árvore de Decisão[10], Impulsionamento Adaptativo[11], Floresta Aleatória[12], Redes Neurais.
2019 Chen et al.[13]	Realizar a previsão de qualquer tipo de dado de série temporal	FM[7]
2020 McIlroy-Young et al.[14]	Prever o próximo movimento de um jogador no xadrez	CNN[15], Deep Learning[16]
2020 Mens et al.[17]	Antecipar comportamentos nocivos a própria saúde	Regressão Logística[18], Árvore de Classificação[10], Floresta Aleatória[12], KNN[19], GBM[20] e SVM[21]
2021 Grendas et al.[22]	Estimar chances de comportamentos nocivos a própria saúde	Regressão Logística[18], Floresta Aleatória para Análise de Sobrevivência[23]
2021 Gong et al.[24]	Predizer o nível de necessidade de substâcias em usuário dependentes de drogas	GBM[20]
2021 Alam et al.[25]	Estimar o vício em drogas	Regressão Logística[18], Árvore de Decisão[10], Floresta Aleatória[12], Naive Bayes[26] e SVM[21]
2022 Safara[27]	Estimar comportamento de compras durante pandemia	SVM[21], Árvore de Decisão[10], Otimização Sequencial Mínima, Redes Neurais, Naive Bayes[26]
2022 Senefonte et al.[1]	Predição da mobilidade urbana de turistas	SOM[28], FCM[29]

trabalhos apresentados em si. No entanto, é descrito nas outras sessões deste capitulo, alguns destes modelos de interesse para este trabalho.

Um modelo que chama atenção, tanto pelo objetivo, quando pelos modelos usados é o de Senefonte et al.[1]. As técnicas usadas estão normalmente presentes em trabalhos de criação de perfis, classificação e identificação de padrões. Neste, a autora, cria um sistema onde, mediante perfis e estimativas, é construído um vetor de dados médio de acordo com valores iniciais, e então feito uma predição de comportamento. Os resultados obtidos são apontam que as referências usadas de comparação.

2.2 *k*-Nearest Neighboors

O k-Nearest Neighbors (kNN) é um modelo clássico de aprendizado de máquina, conhecido por sua simplicidade, aplicabilidade ampla e abordagem intuitiva para problemas de classificação [19]. Este método baseia-se no pressuposto de que instâncias semelhantes compartilham características similares, portanto, podem ser agrupadas na mesma categoria.

Ao receber uma entrada, o algoritmo kNN consulta seu conjunto de dados previamente treinados, identificando os “k” pontos mais próximos à instância de teste. O critério para determinar a proximidade entre os pontos pode variar dependendo do problema em questão, sendo a distância euclidiana um dos métodos mais comuns. Outras métricas de distância, como a distância de Manhattan e a distância de Minkowski, também podem ser utilizadas conforme a necessidade.

Uma vez identificados os “k” vizinhos mais próximos, o algoritmo atribui à instância de teste a classe majoritária entre os vizinhos selecionados. Dessa forma, o kNN consegue classificar a entrada conforme as informações fornecidas pelo conjunto de dados de treinamento. É importante notar que a escolha do valor “k” é crucial para o desempenho do algoritmo, uma vez que valores inadequados podem levar a resultados insatisfatórios. Técnicas como validação cruzada e busca em grade são comumente empregadas para encontrar o valor ótimo de “k”.

Em resumo, o kNN é um modelo de aprendizado de máquina eficaz e versátil, capaz de lidar com problemas de classificação de maneira simples e intuitiva, valendo-se da premissa de que dados similares compartilham características semelhantes.

2.3 Self Organzing Maps

Proposto pro Kohonen[28], o modelo *Self Organizing Maps*(SOMs) é muito utilizado para análise, visualização e redução de dimensionalidade de dados, sendo frequente o uso para agrupamento e identificação de padrões. Uma das características notáveis do modelo é sua capacidade de preservar a topologia dos dados de entrada, garantindo que as relações espaciais inerentes sejam mantidas, evitando distorções no processamento dos dados para se adequar ao modelo.

O modelo é constituído por uma grade, chamada de mapa, de neurônios, atualizados iterativamente no treinamento. Estes ajustes visam modificar o mapa para ele representar melhor as características dos dados que ele trabalha. O treinamento é feito interativamente, onde cada dado do treinamento é apresentado ao modelo, um de cada vez, sendo é calculado a distância de cada neurônio com os valores do dado, sendo encontrado o ponto mais próximo (normalmente usa-se a distância euclidiana), este então é atualizado para ficar mais próximo ainda do dado, e também seus vizinhos, mas em grau menor. A interação é então reiniciada, com exceção do caso de alcançado um critério alvo para indicar convergência do modelo. Este processo é descrito pelo Algorítmo 1

Esta é uma técnica não supervisionada, evitando a necessidade de classificações prévias dos dados, o próprio mapa é capaz identificar características conforme os valores apresentados a ele. Em resumo, é uma técnica versátil e prática para produzir visualizações de dados complexos e criar categorizações de acordo os valores amostrados.

Algorítimo 1 Processo de atualização de neurônios do SOM

1: $t = 1$
2: **enquanto** $parada = FALSO$ **faça**
3: **para** $k = 1 : n$ **faça**
4: $w = argmin_k|x - w_k|$
5: **para** $k \in h_{k,w}(t)$ **faça**
6: $w_k(t+1) = w_k(t) + h_{k,w}(t)(x(t) - w_k(t))$
7: $\sigma(t) = \sigma_0 exp(\frac{-t}{\pi})$
8: $h_{k,w}(t) = exp(\frac{||w_k(t) - w_w(t)||^2}{2\delta^2(t)})$
9: $t = t + 1$

2.4 Fuzzy C Means

Proposto por Dunn[30] e otimizado por Bezdek[29], o algorítimo *Fuzzy C Means*(FCM) tem como princípio que dados podem ser classificados em grupos diferentes, considerando que podem pertencer em mais de um grupo, diferente do kNN, descrito na Seção 2.2, que classifica os dados em somente um.

O processo de determinar em quais grupos é feito por uma função, com seus parâmetros ajustados via treinamento, para minimizar seu valor objetivo. Esta função considera a distância entre os pontos de dados e os centros dos grupos, ponderada pelos graus de pertinência. Dado um dataset de n elementos, onde $X = \{x_1, \ldots, x_n\}$, com a $i^{ésima}$ amostra definida por $x_i = (x_{i1}, \ldots, x_{id})^T$, o algorítimo inicializa para cada amostra x_i pesos aleatórios $u_{ij}, i = \{1, \ldots, n\}, j = \{1, \ldots, C\}$ sendo j um grupo do total de C grupos. Depois o algoritimo computa os centros c_j para o $j^{ésimo}$ grupo conforme a Equação 2.1. E subsequentemente os graus de pertencimentos são atualizados de acordo com Equação 2.2

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m} \tag{2.1}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} (\frac{||x_i - c_j||}{||x_i - c_k||})^{\frac{2}{m-1}}} \tag{2.2}$$

Por fim, podemos concluir que o FCM é uma maneira de criar classificações de graus diferentes para dados de acordo com múltiplos critérios. Sua versatilidade e conceito de categorização múltipla permitem sua aplicação onde o kNN seria insuficiente para o trabalho de classificar grupos.

2.5 PredicTour

Uma das indústrias de mais bem sucedidas é o turismo. Todos os anos turistas viajam o mundo todo, gerando valor através de suas visita a países e localidades, criando trabalhos direta e indiretamente. Com isso, pode ser interessante uma forma de prever a taxa de convergência de turistas em localizações para, por exemplo, melhorar acessos

Algorítimo 2 Predição da mobilidade de um turista

1: Obtenha a origem e destino do turista, respectivamente, v_{origin} e v_{dest}
2: **se** turista tem histórico de outras viajens anteriores **então**
3: $\quad \bar{m}_t = \sum_{l=1}^{L_p} \frac{m_t^{(l)}}{L_p}$ ▷ média de dados históricos
4: **se não**
5: $\quad \bar{m}_t = \sum_{u=1}^{U_t} \frac{m_u}{U_t}$ ▷ média de outros turistas com mesma origem e destino
6: Calcule o v_{class} do turista
7: Concatene todos os valores em $\tilde{d}_t = (\bar{m}_t | v_{class} | v_{origin} | v_{dest})$
8: Informe $\tilde{d}_t$ para o SOM treinado e obtenha o neurônio mais próximo da entrada
9: Identifique o perfil p do neurônio obtido
10: $c_p = \frac{\sum_{i=1}^{N_p} u_{ip} m_i}{\sum_{i=1}^{N_p} u_{ip}}$
11: Calcule o resultado final $\hat{m}_t^{(dest)} = \frac{e^{\log L_p + 1} \bar{m}_t + c_p}{2}$

a esses pontos, torná-los mais atrativos, ou até reformar locais de pouco interesse para incentivar o uso deles.

PredicTour é um modelo de aprendizado de máquina proposto por Senefonte et al.[1]. Usando bases de dados do Foursquare, são criados descritores de mobilidade, que representam o deslocamento total que um turista fez em uma viagem. Esses descritores são usados para treinar o modelo a classificar turistas em perfis, além de permitir uma estimativa da mobilidade urbana.

Uma das características interessantes do modelo é que seu treinamento é não supervisionado. O modelo consegue decidir por si como classificar cada em indivíduo em seu devido perfil, necessitando somente a quantidade otimizada de perfis, sendo decidido de forma arbitrária, e no trabalho referenciado é escolhido de forma exploratória. E, mesmo que um turista não possua nenhuma informação relevante, desde que ele tenha o pais de origem e de destino, é possível criar uma expectativa de mobilidade em sua viagem.

O uso modelo constitui-se de três partes, criação de descritores de mobilidade, extração de perfis e por fim predição de mobilidade. Os descritores de mobilidade compreendem a modelagem de informações de um registro de dados de uma pessoa, sua identificação única no conjunto de amostras, seu conjunto de visitas a quais lugares, países de origem e destino e classificação entre turista "retornante" e "explorador". A criação dos mesmos é feita a partir de amostras de turismos.

Na parte de extração de perfil é usado o SOM em conjunto com FCM, descritos respectivamente nas subseções 2.3 e 2.4. Na primeira fase, é feito o treinamento do SOM, iterando por cada descritor de mobilidade, ajustando os pesos dos nerônios do mapa conforme o algorítimo do trabalho. Após o treinamento, na segunda fase, os pesos do mapa são usados para treinar o algorítimo FCM, que assim cria uma classificação de graus de perfil para cada neurônio do mapa.

A parte final é onde são feitas as predições das mobilidades dos usuários. Os usuários alvo a terem suas mobilidades preditas, tem primeiro identificado o perfil individual normalmente associado a ele, constrói-se um descritor de mobilidade médio e enfim realiza a previsão de mobilidade. Esta tarefa segue o processo descrito pelo Algorítmo 2, onde $\hat{m}_t^{(dest)}$ é vetor final de predição da mobilidade.

Em resumo, o modelo *PredicTour* usa técnicas clássicas, normalmente usadas para classificação, criando estimativas de acordo com variações finas entre categorizações de perfis. O método consegue processar, inclusive, entradas de dados sem conhecimento prévio de outras iterações do turista. O vetor final é condizente em parte com a mobilidade realizada, ganhando com erros menores, de outros modelos usados como comparação.

3 ANÁLISE EXPLORATÓRIA DO DATASET DATA MINING CUP 2013.

Todos os anos, desde 2003, é realizada uma competição internacional *Data Mining Cup*[31], onde equipes e indivíduos, de diversas universidades do mundo, enviam modelos de inteligência artificial treinados, que visam ser o melhor a resolver o problema enunciado na competição do ano vigente. Em cada edição do evento, *datasets* de treinamento são disponibilizados juntos a descrição e o enunciado de alguns desafios, onde os modelos enviados devem buscar resolver. Para este trabalho, foi-se selecionado um dos *datasets* disponíveis no site oficial da competição. O critério de escolha baseou-se na simplicidade de estrutura e quantidade de amostras, e a característica de serem dados não sintéticos. Desta forma, os dados utilizados na competição de 2013 [32] foram escolhidos como *dataset* deste trabalho.

Na Seção 3.1 é descrito a estrutura, arquitetura e variáveis, dos dados. Adiante, a Seção 3.2 e suas respectivas subseções apresentam algumas análises do dataset, assim como algumas conjecturas de relações entre variáveis. É relevante ressaltar que análise elucidará relações intrínsecas dos dados, demonstrando como certas características são estatisticamente previsíveis, assim como pontos de atenção que devem ser considerados no treinamento e na avaliação dos modelos.

3.1 Estrutura de dados do dataset Data Mining Cup 2013

O *dataset* disponível em [32] modela a interação de usuários em site de compras. A cada intervalo, aparentemente arbitrário, é gravado o estado atualizado da interação de um usuário no site. Cada acesso tem uma chave de identificação única, usada para identificar quais gravações pertencem a qual acesso ao site. Os dados estão dispostos em uma tabela, em que cada linha representa uma captura de informações de uma sessão e do usuário identificado, ou não identificado. Cada uma dessas entradas contem um conjunto de variáveis, podendo ter algumas delas com valores nulos, sendo estes chamados de **campos opcionais**, e os outros de **campos não opcionais**. A Tabela 2 apresenta o nome dessas variáveis[1], uma descrição curta de cada uma, os tipos de valores que a variável pode assumir e se a variável em si é opcional.

As gravações são armazenadas em uma tabela, onde cada coluna representa cada variável, e cada linha representa uma gravação de sessão. As linhas podem ser agrupados pela identificação única de sessão (a variável *SessionId*), criando uma série temporal, que descreve a jornada de um usuário ao usar o site. Esse agrupamento denota como

[1] Alguns nomes de variáveis foram alterados do dataset original para facilidade de leitura

Tabela 2 – Descrição de variáveis do Dataset.

Nome de variável	Descrição	Tipo de valor	Opcional
sessionId	Identificador único de sessão	Inteiro	Não
startHour	Hora de início da sessão	Inteiro no intervalo de 0 a 24	Não
startWeekday	Dia da semana de início de sessão	Inteiro no intervalo de 0 a 7	Não
duration	Duração da sessão	Ponto flutuante	Não
ordered	Indica que a sessão resultou em compra	Booleano	Não
cCount	Contagem de cliques da sessão	Inteiro	Não
bCount	Contagem de itens no carrinho de compra	Inteiro	Não
userId	Identificador único de usuário	Inteiro	Sim
cMinPrice	Menor de valor de item clicado	Ponto flutuante	Sim
cMaxPrice	Maior de valor de item clicado	Ponto flutuante	Sim
cSumPrice	Soma de todos os valores de items clicados	Ponto flutuante	Sim
bMinPrice	Menor valor de item no carrinho de compra	Ponto flutuante	Sim
bMaxPrice	Maior valor de item no carrinho de compra	Ponto flutuante	Sim
bSumPrice	Soma de todos os valores no carrinho de compra	Ponto flutuante	Sim
bStep	Estado de pagamento do carrinho	Inteiro no intervalo de 1 a 5	Sim
orderable	Disponibilidade para o usuário realizar a compra	Texto	Sim
userScore	Pontuação dada pela loja ao usuário	Inteiro	Sim
pronoum	Pronome do usuário (1-Sr., 2-Sra., 3-Cia.)	Inteiro no intervalo de 1 a 3	Sim
onlineStatus	Se usuário estava online no momento de captura	Booleano	Sim
maxSpendVal	Valor máximo permitido ao um usuário gastar	Ponto flutuante	Sim
paymentCount	Contagem de pagamentos já feitos pelo usuário	Inteiro	Sim
age	Idade do usuário	Inteiro	Sim
daysSinceLastOrder	Dias desde a última compra	Inteiro	Sim

certas varáveis permanecem iguais do início ao fim de uma sessão, sem mudar de valor em qualquer momento durante a sessão. Estas variáveis de valores imutados serão chamadas de **variáveis estáticas**, enquanto as que pelo menos em alguma sessão possui alguma variação, são de **variáveis dinâmicas**. A Tabela 3 identifica quais variáveis possuem essa qualidade apresentando seu nome e se seus valores são estáticos ou não.

Uma sessão não modela somente a jornada de um usuário, mas também o histórico de um usuário se este estiver identificado. Algumas variáveis estão somente preenchidas se

Tabela 3 – Característica valores estáticos de variáveis do Dataset.

Variáveis estáticas	**Variáveis dinâmicas**
sessionId	duration
startHour	cCount
startWeekday	bCount
ordered	cMinPrice
userId	cMaxPrice
userScore	cSumPrice
pronoum	bMinPrice
onlineStatus	bMaxPrice
age	bSumPrice
paymentCount	bStep
daysSinceLastOrder	orderable
-	maxSpendVal

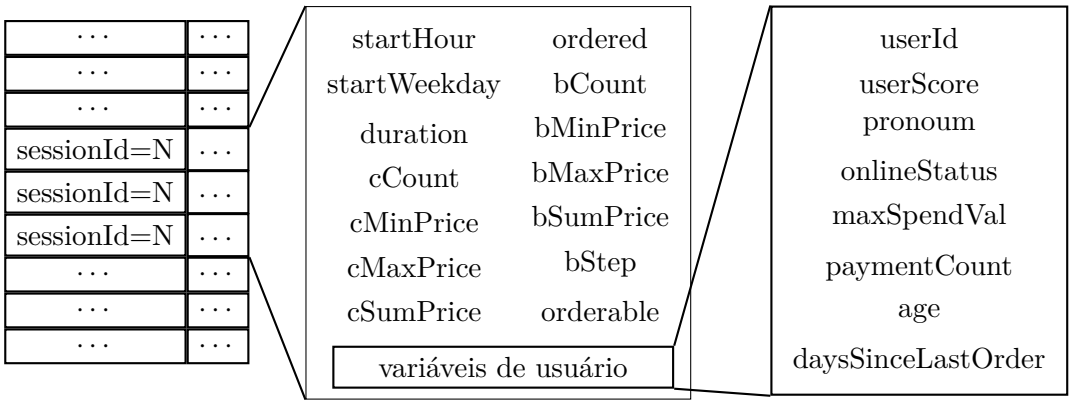


Figura 1 – Ilustração de estrutura de dados do dataset

o usuário está identificado, nomeadamente são, ***userScore***, ***pronoum***, ***onlineStatus***, ***maxSpendVal***, ***paymentCount***, ***age*** e ***daysSinceLastOrder***. A Figura 1 demonstra visualmente como esses dados podem ser separados, listando nomes das variáveis em espaços diferentes para identificar suas relações.

Em resumo, os dados do *dataset* são registros de momentos de uma sessão, dispostos em uma tabela em que cada coluna representa uma variável. Essas informações de uma sessão podem ser agrupadas formando uma série temporal, descrevendo o histórico de ações de um usuário. Destas variáveis, algumas podem ser ausentes, com valores nulos, sendo denominadas de **variáveis opcionais**, enquanto outras que estão sempre com valores informados são chamadas de variáveis **não opcionais**. Também existem variáveis qpue não mudam do início de uma sessão até o fim, essas são **variáveis estáticas**, enquanto as que possuem valores diferentes do primeiro registro são chamadas de **variáveis dinâmicas**.

3.2 Análise exploratória de variáveis, valores e interações do dataset Data Mining Cup 2013.

Um ponto crucial para o treinamento de inteligências artificiais é a qualidade dos dados de treinamento. Dados com valores descontínuos, amostras incompletas, valores corrompidos ou incorretos e entre outras qualidades negativas podem influenciar drasticamente na qualidade do treinamento, capacidade e validade de muitos modelos. Essas características agravantes, quando identificadas, podem, em alguns casos, ser tradas, contornadas e até incorporadas no treinamento. A fim de entender melhor os dados disponíveis no dataset descrito na Seção 3.1, é feita uma exploração a fim de levantar características, relações de variáveis, problemas, pontos de atenção e qualidades do dataset. Também é objetivo desse estudo sobre os dados, realizar análises e criar hipóteses, para serem comparadas com os resultados obtidos do treinamento do modelo.

Nas próximas subseções, são apresentados processos de tratamento de dados, agrupamento de entradas por variáveis, filtragem de valores relevantes e outras funções que permitem calcular, e também visualizar as interações entre as variáveis. Em cada uma delas é disposto um gráfico que demonstra os resultados obtidos, bem como observações pertinentes ao contexto da análise dos dados.

3.2.1 Análise Interna de Variáveis e Amostras

Identificar o tipo de uma variável, como numérica, texto ou ponto flutuante, não define o comportamento, a distribuição e a consistência de seus valores. No estudo proposto, analisamos as variáveis em diversas categorias, o que permitiu extrair múltiplas conclusões sobre os dados. Essas análises baseiam-se nos primeiros e últimos momentos de uma sessão, isto é, na primeira e última amostra de cada sessão, e apenas em variáveis com valores numéricos que podem ser controlados ou influenciados pelo usuário. Assim, variáveis como **sessionId**, **userId** e **orderable** não são avaliadas nessas métricas. A seguir, apresentamos as categorias analisadas e suas motivações, e, posteriormente, nesta subseção, exibimos tabelas com os valores calculados, acompanhados de comentários e análises dos resultados.

- Tendências de centrais: média e mediana foram obtidas de todas as variáveis, e usadas para analisar qual o centro dos valores possíveis.
- Dispersão de valores: variância, desvio padrão e intervalo interquartil foram calculados para determinar valores comuns, e também entender o quão variável são os valores
- Correlação de variáveis: as indicações das relações entre as variáveis, podem ser obtidas pelo cálculo de Correlações De Pearson, indicando se elas interagem de

Tabela 4 – Métricas de Tendências Centrais de Variáveis Estáticas

Variável	Valor Mínimo	Valor Máximo	Média	Mediana	Moda
startHour	0	23	14.5003	15.0	17
startWeekday	5	7	5.9542	6.0	6
ordered	0	1	0.6481	1.0	1
userScore	0.0	638.0	482.4446	518.0	70.0
accountLifetime	0.0	600.0	131.145	103.0	223.0
paymentCount	0.0	868.0	13.2014	8.0	1.0
age	17.0	99.0	44.728	44.0	46.0
pronoum	1	3	1.7238	2.0	2
daysSinceLastOrder	3.0	738.0	80.1137	34.0	10.0
bStep	1.0	4.0	1.4558	1.0	1.0

maneiras positivas, ou negativas entre seus valores.

Tabela 5 – Métricas de Tendências Centrais de Variáveis Dinâmicas

Variável	Valor Mínimo	Valor Máximo	Média	Mediana	Moda
duration	0.0	21367.546	471.9487	81.246	0.0
	0.0	21553.323	1661.5976	860.906	0.0
cCount	1	198	7.3934	4.0	3
	1	200	24.0778	13.0	3
cMinPrice	0.0	2799.99	61.6101	19.99	9.99
	0.0	2799.99	46.2597	10.95	9.99
cMaxPrice	0.0	6999.99	100.9592	29.99	19.99
	0.0	6999.99	128.283	49.99	29.99
cSumPrice	0.0	71329.21	486.5777	104.95	39.98
	0.0	76239.34	998.3965	358.0	39.98
bCount	0	19	1.1687	1.0	1
	0	108	3.8742	3.0	1
bMinPrice	0.0	6999.99	73.7187	24.99	19.99
	0.0	6999.99	57.0904	14.99	9.99
bMaxPrice	0.0	6999.99	75.3145	24.99	19.99
	0.0	6999.99	90.0423	34.99	29.99
bSumPrice	0.0	6999.99	78.969	27.85	19.99
	0.0	10429.83	168.9316	79.95	29.99
maxSpendVal	0.0	50000.0	2314.0942	1300.0	600.0
	1.0	5.0	3.2444	4.0	5.0
onlineStatus	-1.0	1.0	0.9949	1.0	1.0
	-1.0	1.0	0.9817	1.0	1.0

3.2.1.1 Tendências Centrais de Variáveis

Os valores mínimos, máximos, médios, medianos e de moda de cada variável dos dados, de início e fim de sessão, são calculados separadamente. Os resultados são então

dispostos em duas tabelas, uma para as variáveis estáticas na Tabela 4, e outra para as variáveis dinâmicas na Tabela 5, nesta tendo duas linhas para cada variável, uma para o início da sessão, e outra para o final dela. A diferenciação entre variáveis estáticas e dinâmicas são descritas na Seção 3.1, e listadas na Tabela 3

A partir dos dados obtidos várias suposições podem ser feitas. Começando com as variáveis dinâmicas, observamos haver sessões que duram zero segundos, e outras que possuem o primeiro registro muito tempo depois do início, e, que sempre são mais comuns sessões que possuem duração zero. De forma análoga, as variáveis, **cCount**, **cMinPrice**, **cMaxPrice**, **cSumPrice**, **bMinPrice**, **bMaxPrice** e **bSumPrice**, possuem algum registro com valores próximos aos valores máximos dos últimos registros de cada sessão.

Nas variáveis estáticas, assim como nas dinâmicas, podemos ver como não é possível assumir valores médios e medianas de certas variáveis reais representações de amostras médias, pois, a moda de certas variáveis são muito diferentes de suas médias e medianas, notadamente, as variáveis **startHour**, **userScore**, **accountLifetime**, **paymentCount**, **daysSinceLastOrder**, **duration**, **cCount**, **cMinPrice** (somente registros de início), **cMaxPrice**, **cSumPrice**, **bCount** (somente registros finais), **bMinPrice**, **bSumPrice** e **maxSpendVal**. Outras ponderações sobre os valores obtidos serão discorridas e avaliadas e em contexto com outras métricas na Subseção 3.2.2

3.2.1.2 Dispersão de Valores de Variáveis

Continuando a avaliação valores possíveis das variáveis, calcula-se os valores de variância, desvio padrão e intervalo interquartil. Essas métricas visam elucidar a amplitude de amostragem, e permitir uma inspeção de quais valores podem estar contribuindo para médias e medianas serem calculadas incorretamente. Os resultados foram separados em duas tabelas, uma para variáveis estáticas, a Tabela 6, e outra, para variáveis dinâmicas, a Tabela 7. A tabela de variáveis dinâmica tem dois valores por linha de variável, pois o primeiro representa as amostras de início de sessão, e o segundo representa a última informação gravada de uma sessão.

Os dados demonstrados apontam uma alta dispersão de valores, e forte indícios de muitos *outliers* estarem presente. No entanto, algumas variáveis apresentam pouca varição em relação a sua média, estas são: **paymentCount**, **cCount**, **cMinPrice**, **cMaxPrice**, **cSumPrice**, **bCount**, **bMinPrice**, **bMaxPrice** e **bSumPrice**.

Outras variáveis que possuem valores baixos, ou parecidos com as que são identificadas como pouco dispersas, não são consideradas agrupadas devido à definição de valores possíveis, por exemplo, a variável **ordered**, que possui somente valores 0 (zero) e 1 (um), que torna os resultados, de certa forma, inválidos, pois não existem valores além dos presentes. Desta forma, as variáveis, **ordered**, **pronoum**, **onlineStatus** e **bStep**, são valores discretos e limitados, que faz suas métricas serem imprecisas e inconclusivas

Tabela 6 – Métricas de Dispersão de Valores de Variáveis Estáticas

Variável	**Variância**	**Desvio Padrão**	**Intervalo Interquartil**
startHour	18.8603	4.3428	7.0
startWeekday	0.6306	0.7941	2.0
ordered	0.2281	0.4776	1.0
userScore	17612.9446	132.7138	76.0
accountLifetime	11719.3579	108.256	177.0
payments	728.2247	26.9856	11.0
age	147.8923	12.1611	17.0
pronoum	0.2021	0.4495	1.0
daysSinceLastOrder	12893.7662	113.5507	73.0
onlineStatus	0.0101	0.1007	0.0

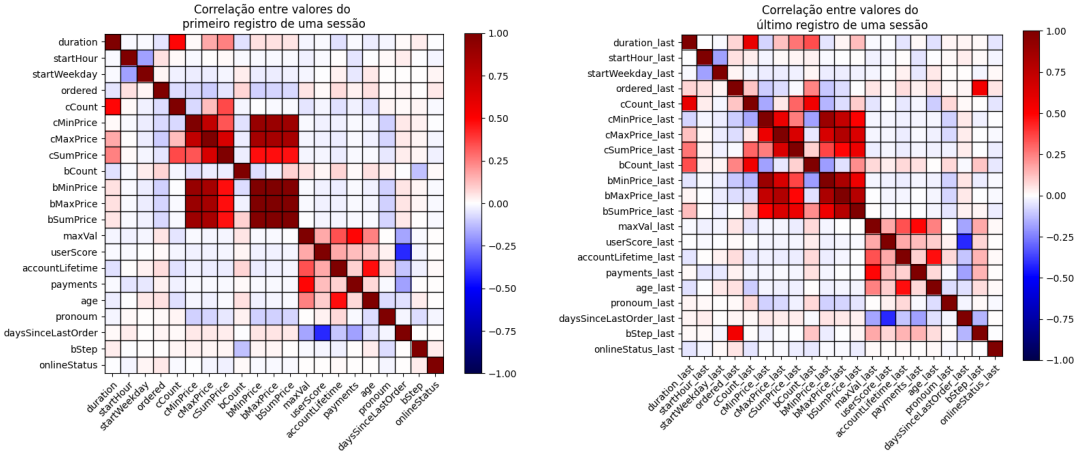
Tabela 7 – Métricas de Dispersão de Valores de Variáveis Dinâmicas

Variável	**Variância**	**Desvio Padrão**	**Intervalo Interquartil**
duration	1763998.99 5693964.76	1328.16 2386.2	345.64 1647.07
cCount	123.29 938.6	11.1 30.64	6.0 25.0
cMinPrice	20768.31 16798.97	144.11 129.61	30.0 18.99
cMaxPrice	48817.31 65020.97	220.95 254.99	52.0 75.0
cSumPrice	3712011.58 6903012.99	1926.66 2627.36	250.93 793.72
bCount	1.04 15.76	1.02 3.97	0.0 4.0
bMinPrice	28748.65 24147.15	169.55 155.39	37.0 21.0
bMaxPrice	28826.89 35003.33	169.78 187.09	35.0 47.23
bSumPrice	29627.73 108504.08	172.13 329.4	44.96 135.09
bStep	0.52 2.48	0.72 1.58	1.0 3.0

sobre os resultados. Das outras variáveis não mencionadas aqui, é possível concluir que a dispersão de seus valores é grande, e a presença de amostras de valores extremos prejudica a capacidade de terminar um valor comum para a variável.

3.2.1.3 Correlação de Variáveis

A medida que valores de uma variável alteram, é esperado que outra variável também altere de acordo com essa variação, por exemplo a medida que **duration** aumenta,



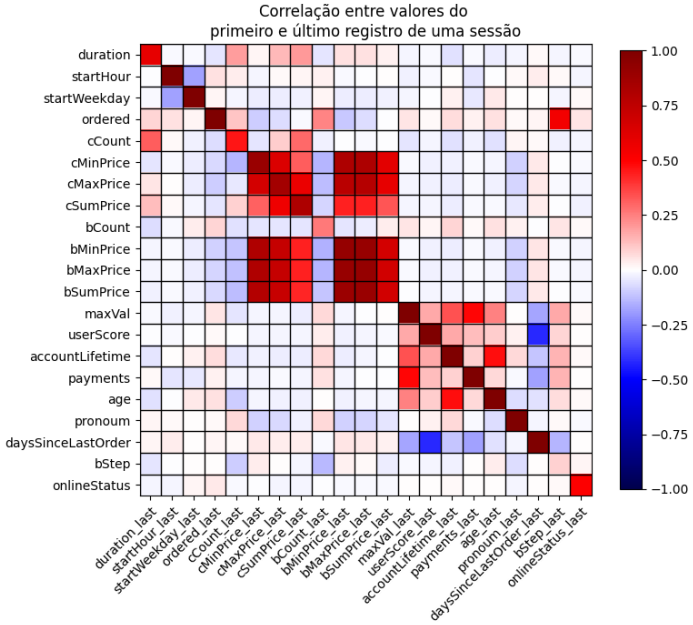


Figura 2 – A direita, correlações de variáveis somente para amostras de início de sessão

Figura 3 – A esquerda, correlações de variáveis somente para amostras de fim de sessão

Figura 4 – A baixo, correlações de variáveis de início e fim de sessão

é esperado que **bCount** também aumente. No entanto, nem sempre as relações implícitas entre variáveis são fáceis de detectar, nem a magnetude de sua interação. A fim de obter algum indicativo de como variáveis interagem, é feito o cálculo de Correlação de Pearson entre cada par de variável. Além disso, essas métricas são realizadas de três maneiras distintas, uma fazendo uso somente de amostras de início de sessão, outra somente para as de fim de sessão, e a última é a interação mista entre cada valor de variável e o fim de sua sessão. O valores obtidos são então tratados, concatena-se às variáveis de fim de sessão o sufixo "_last", e então dispostos visualmente nas Figuras 2, 3 e 4.

Estes resultados elucidam certas relações entre variáveis, que em retrospecto das sessões 3.2.1.1 e 3.2.1.2, não eram óbvias. As correlações de amostras de início de sessão apresentam uma forte conexão entre as variáveis **cMinPrice**, **cMaxPrice**, **cSumPrice**, **bMinPrice**, **bMaxPrice**, **bSumPrice**, sendo possivelmente resultado de valores muito parecidos, pois registros iniciais podem ter poucos ou nenhuma interação. Outro ponto de interesse é a inversa relação entre os valores de **userScore** e **daysSinceLastOrder**, estas, claramente reagem inversamente, potencialmente sendo um indicativo que usuários que não realizaram compras recentemente tem **userScore** gradualmente reduzido. Existe um indício de que a variável **cCount** tem uma relação com **duration**, que pode ser equivocado, pois durações curtas menos cliques, que podem contribuir uma percepção de valores identicos. Por fim, há um princípio de relações entre os pares de variáveis **age** e **accountLifeTime**, e o par **paymentsCount** e **maxVal**, a primeira provavelmente indica existem contas de longa data no site, e a segunda relação possivelmente representa uma quantidade maior de crédito liberado para compra de usuários que possuem mais históricos de compra.

Agora, analisando as correlações para as amostras de fim de sessão, temos similaridades com as de amostras de início de sessão, com algumas variações pontuais e novas. As variáveis **cMinPrice**, **cMaxPrice**, **cSumPrice**, **bMinPrice**, **bMaxPrice**, **bSumPrice** tem suas relações reduzidas, muito provavelmente devido a sessões terminarem com estes valores muito discrepantes. Outras relações permanecem muito próximas a de início, exceto as novas, de **bCount** com **duration**, **bCount** com **cCount** e **bStep** com **ordered**. Os motivos dessas relações aparecerem estão relacionados ao fato que algumas variáveis compartilham partes de seus valores, como **cCount**, sendo sempre menor ou igual a **bCount**. A exceção é a relação de **bStep**, que aqui demonstra a possibilidade de indicar se um usuário terminou o processo de compra, algo que não fica claro no enunciado do *dataset*[32].

Enfim, as relações cruzadas das amostras não apresentam relações novas significativas, somente apresentam em lugares diferentes outras relações das amostras e início e fim. Com isso, considerando as correlações de amostras de início e fim de sessão, podemos definir que as variáveis são pouco relacionadas em alguns campos, e também pouco iterativas entre si, com alguns casos em exceção.

3.2.2 Análises Específicas de Variáveis e Suas Interações

Na Subseção 3.2.1 foram apresentados cálculos e métricas para levantar características do dataset. No entanto, algumas interações entre variáveis ficam inconclusivas, ou, não se relacionaram a qualquer outra variável, e ainda, algumas nem são comentadas por não apresentarem valores marcantes. Nas subseções a seguir é feito alguns levantamentos específicos, usando técnicas simples de estatística, cálculos e representações visuais dos

dados como ferramenta de análise.

3.2.2.1 Distribuição de sessões que resultaram em compra

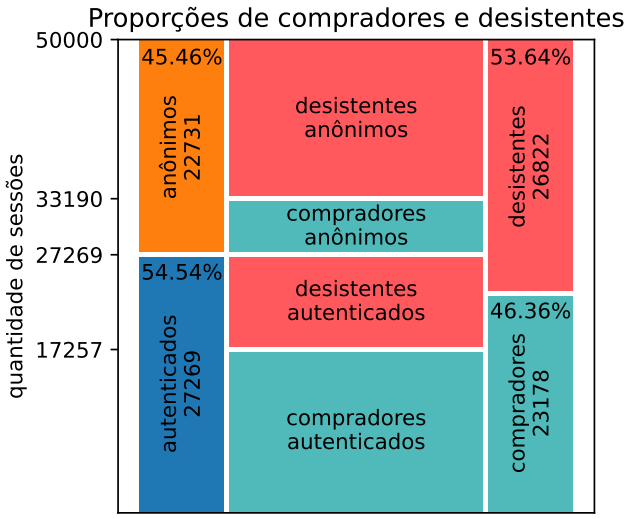


Figura 5 – Proporções de usuários e anônimos que compraram e desistiram

No *dataset*, os valores para a variável **ordered**, descrita na Tabela 2, indica se a sessão teve uma compra realizada, ou se não concluiu o processo de compra. Como uma sessão não necessariamente possui um usuário autenticado, é de interesse observar como estão distribuídos as sessões que há compra ou não. Aqui é referenciado como **comprador** sessões em que um usuário realizou uma compra, esse valor é contado por sessão, e não por identificação de usuário. Em contrapartida, é nomeado **desistente** qualquer sessão que não houve compra. Para realizar essa análise, é necessário agrupar o *dataset* pela variável *sessionId*, e então separar os grupos em duas categorias: usuários autenticados e usuários anônimos. Por fim, é contada a quantidade em cada categoria, em cada parte do processo.

Na Figura 5 temos a porcentagem de usuários anônimos e identificado, bem como suas participações na contagem de sessões de compradores e desistências. Estes dados exemplificam que existem mais chances de um usuário identificado realizar uma compra, do que usuários anônimos. Essa qualidade demonstra existir uma forte relação entre o usuário ter uma conta no site e ter interesse em comprar, bem como permite uma identificação de padrões e comportamentos de compra de cada indivíduo. Mas, deve-se atentar ao fato que a maioria das sessões de usuários anônimos podem colateralmente divergir algumas partes do treinamento.

3.2.2.2 Relação entre interação com o site e duração de uma sessão

Conforme as conclusões da Subsubseção 3.2.1.3, supõe-se que quanto maior o tempo de sessão, maior a quantidade de interação do usuário com o site. Esta hipótese é

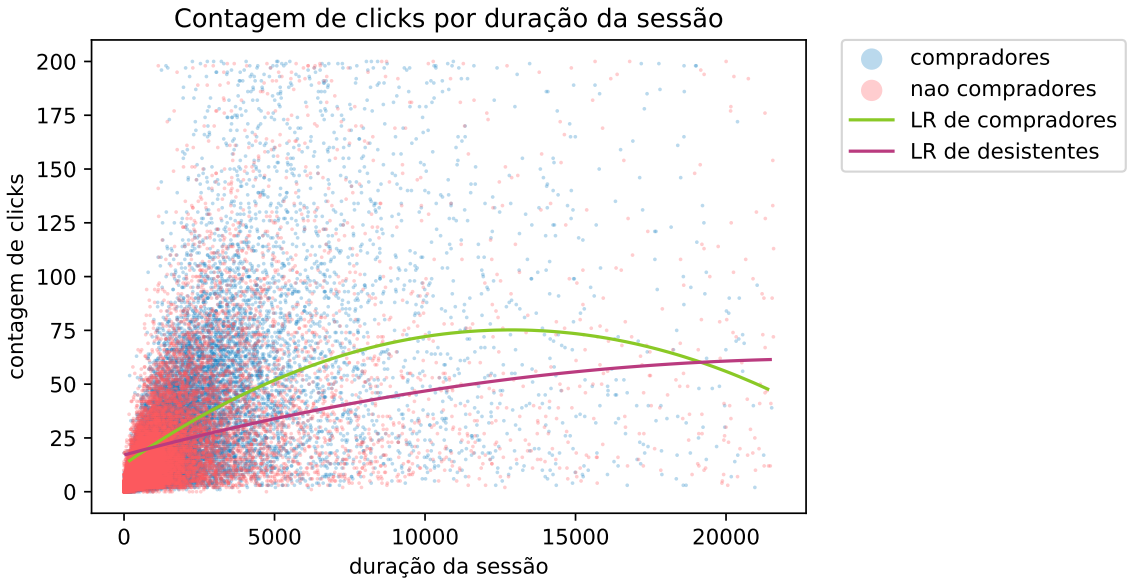


Figura 6 – Relação entre duração de sessão e cliques

avaliada na Figura 6. Nela temos a distribuição da relação entre duração de uma sessão e sua respectiva contagem de cliques, tanto para compradores quando para desistentes. Estes dados foram obtidos ao agrupar os dados por *sessionId*, e depois separar cada um nas categorias de compradores, e desistentes. Depois deste processo, é extraído o último registro de cada sessão, então realiza-se contagem de cliques, obtendo a dispersão da duração pela interação máxima de cada sessão. Como existem muitas entradas de dados nessa avaliação, foi aplicada uma regressão linear de grau dois sobre os resultados obtidos. O modelo treinado produz uma estimativa de interação para os grupos de compradores e desistentes, a curva de regressão é indicada por *LR de compradores* e *LR de desistentes* na Figura 6.

Enquanto a dispersão mostra onde há a maior taxa de amostragem, sendo na primeira uma hora e trinta minutos com até setenta e cinco interações, a regressão demonstra uma curva suave, indicando que compradores interagem mais com o site, em média, do que outros usuários que não efetuam compras. Esses resultados serão muito relevantes para a avaliação final de modelos, pois, pelo modelo treinado, possivelmente usuários que compram terão mais peso por interagirem com o site com mais frequência.

3.2.2.3 Disposição de grupos de idade e chances de compra

Expandindo a análise da Subsubseção 3.2.2.1, é realizado um estudo dos grupos de pronomes de usuários e sua relação com as características de idade e as chances de um usuário possuir pelo menos uma sessão com compra. Para realizar esta avaliação, os registros são agrupados por **age**, então aplicado um filtro, onde são aceitos somente os dados mais recentes de cada usuário. Além disto, é coletado se o usuário possui ao menos uma compra em alguma entrada. Esses grupos de idade então são usados para contar a

quantidade de usuários para cada valor de pronome, definidos na variável **pronoum**, e quantos indivíduos de cada pronome realizaram alguma compra.

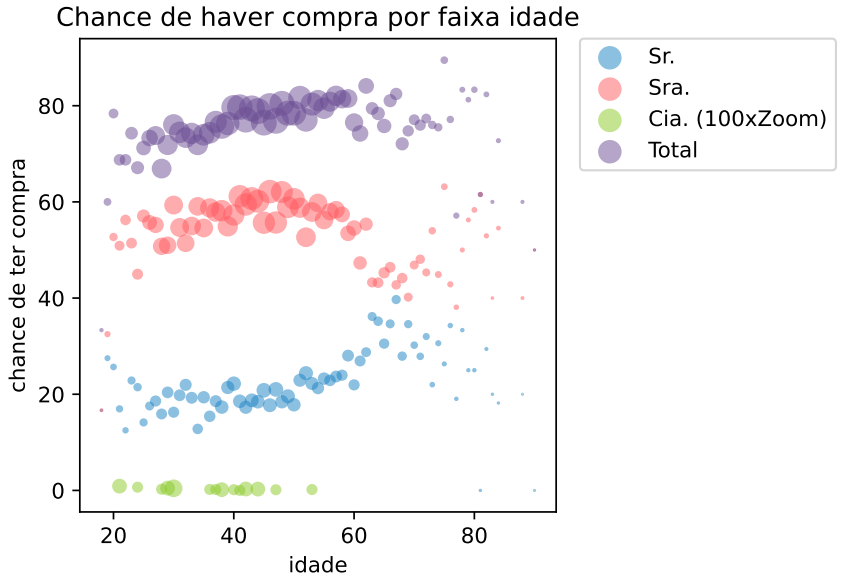


Figura 7 – Relação entre idade e gênero com chances de ser comprador

Os dados agora processados, são usados para calcular em cada faixa de idade, num grupo de pronomes, as chances que individuo que pertence a esse conjunto ser um comprador. Os resultados são apresentados graficamente na Figura 7, onde cada ponto indica a relação entre quatro variáveis: idade, tamanho de grupo, pronome e chances de algum usuário que pertence ao grupo ser um comprador, representados respectivamente por, o deslocamento horizontal do ponto, o tamanho do ponto (escalados para o maior ponto pertencer ao maior grupo), a cor, e o deslocamento vertical do ponto.

Avaliando o gráfico, é possível observar que a distribuição de idade se concentra nas faixas de trita a sessenta anos, assim como as chances de um usuário de pronome "Sra." ser um comprador é, aproximadamente, de sessenta porcento. Também constatado graficamente que usuários de pronome "Cia." tem uma representação quase insignificativa para as suas faixas de idade, pois o tamanho de seus pontos, que representam a quantidade de usuários que pertencem ao grupo, necessitam de uma ampliação de cem vezes o seu tamanho para serem visíveis. Esse gráfico amostra como cada grupo de idade possui uma participação ativa de compras, pois, em média, a chance total de uma pessoa do grupo ser um comprador é de oitenta porcento. Uma hipótese que deve ser considerada para o treinamento é que os modelos podem ter uma facilidade maior para prever o comportamento de usuários identificados como "Sra." do que outros pronomes.

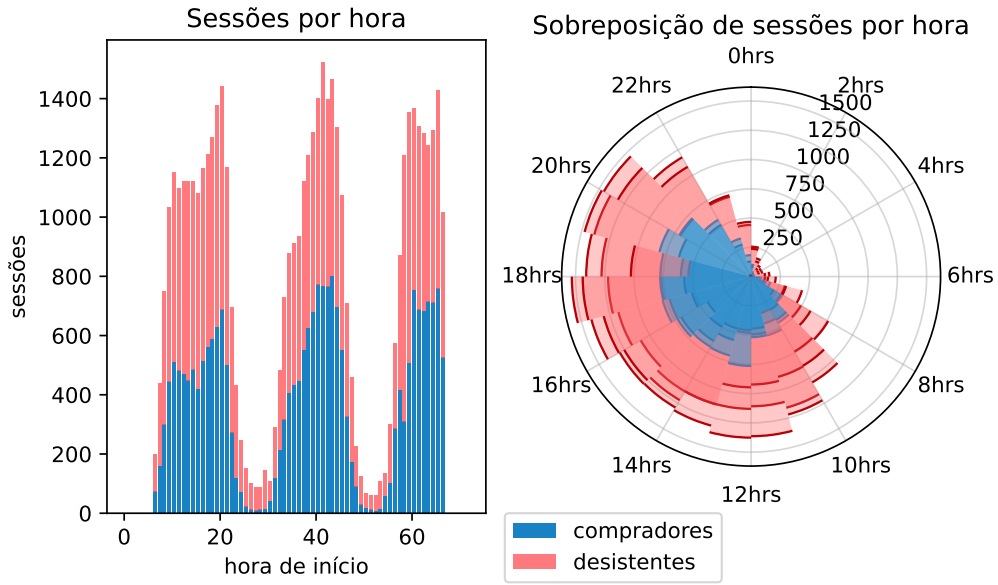


Figura 8 – Relação entre hora de início e quantidade de sessões de compradores e desistentes

3.2.2.4 Interação entre horários e comportamentos de compras

A fim de delimitar melhor como usuários se comportam em horários ao longo dos dias, os dados processados da Subsubseção 3.2.2.1, são reagrupados e estudados novamente para estimar quais padrões de compra podem existir. Para obter os valores relevantes a quantidade de sessões por hora, os grupos totais de compradores e desistentes foram reorganizados internamente sob as variáveis de **startWeekday** e **startHour**, criando um histórico de acessos ao longo dos dias.

Na figura Figura 8, o primeiro gráfico demonstra como há uma oscilação de quantidade de acessos durante cada dia, e por consequência, a contagem de sessões de compradores oscilam também. No segundo gráfico fica claro que das 21 horas até as 10 horas da manhã possuem poucas amostras de sessão em comparação com a média dos outros horários. Isso clarifica a característica que o conjunto de dados por ser insuficiente para treinar modelos para prever chances de compra à noite, e principalmente de madrugada. No entanto, essa relação de dimensões pode ser claramente usada para fazer a previsão da quantidade de acessos para um horário, por haver um padrão de horas de pico e queda de visitas. Com isso, podemos concluir que usar a hora de acesso, ou o dia de acesso, possivelmente não terá relevância no treinamento.

3.2.2.5 Impacto da duração de uma sessão com o valor de compra

Os valores da variável **bSumPrice** do dataset, representam o custo total de um carrinho, tanto em sessões de compradores, quanto de desistentes de compras. A fim de identificar quando um usuário desiste de completar o processo de compra, é necessário

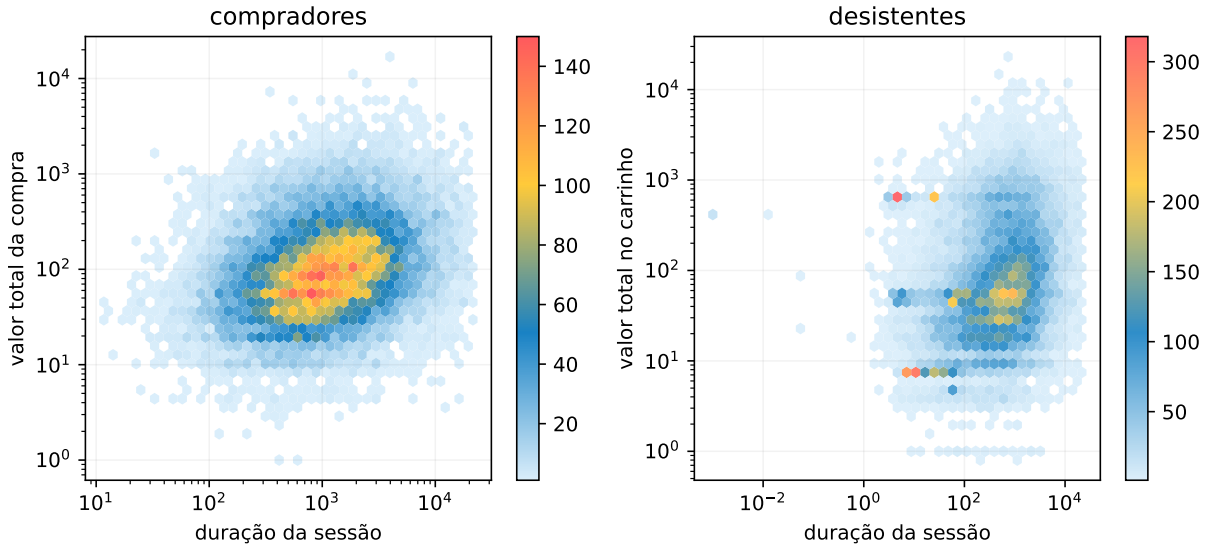


Figura 9 – Relação entre duração de sessão e valor de carrinho para compradores e desistentes

avaliar o quanto que a duração de uma sessão se relaciona com o valor total do carrinho. Para isso todos os usuários foram ordenados por duração de sessão, pela variável **duration**, e depois agrupados por **sessionId**, e então tiveram os últimos registros de cada sessão extraídos, resultando na duração máxima da sessão por quanto o valor máximo do carrinho estava ao final da sessão.

Os resultados foram dispersos em um campo de agrupamento, na Figura 9, onde são delimitadas áreas que contam quantas entradas estão em cada faixa de valor. Os resultados possuem uma quantidade alta de *outliers* a média do grupo, e por isso, os gráficos foram reescalados na escala logarítmica de base dez. Os dois gráficos, à direita e à esquerda da figura, representam a dissipação de valor de carrinho, pela duração de uma sessão. O gráfico de compradores apresenta um espalhamento uniforme de entradas próximos ao núcleo de valores totais de carrinho, que está na faixa próxima aos dezesseis minutos e com carrinho de valor cem (100). Já o gráfico de desistentes claramente possuem uma tendência de se agrupar, principalmente quando o valor do carrinho ou está muito acima, ou muito abaixo, de onde se encontram a maioria das sessões, que no caso, são aproximadamente aos dezesseis minutos, assim como o de compradores, mas com valor de carrinho mais baixo, perto de trinta (30). Essas dispersões demonstram como há uma forte relação entre valor de compra, duração de sessão e chances de haver uma compra, e isso deve ser considerado ao avaliar o desempenho do treinamento de modelos.

4 PREVISÃO DE JORNADA DE USUÁRIO

O Capítulo 3 apresenta uma base de dados de um site de compras, onde cada registro é um momento gravado da jornada de um usuário. As particularidades dos dados e características descritas e exploradas no capítulo sugerem uma dificuldade de métodos estatísticos clássicos identificarem padrões, que poderiam permitir a predição de iterações no site. Em um contexto similar, de dados de relações convolutas, o trabalho de Senefonte et al.[1] desenvolve um modelo capaz de realizar a predição de acordo com classificações de usuários, nomeado *PredicTour*. Este modelo é melhor descrito em detalhes na Seção 2.5.

A fim de explorar o nível de capacidade do modelo *PredicTour*, é experimentalmente aplicado os dados mencionados, ao modelo, e então avaliada sua capacidade preditiva em relação aos valores esperados. Foram necessárias adaptações, tanto a estrutura do aprendizado de máquina proposto, assim como os dados fornecidos a ele. As modificações no conteúdo dos dados e outras alterações são descritas na Seção 4.1, enquanto ajustes no modelo para o treinamento foram discorridos na Seção 4.2. Por fim, a Seção 4.3 usa a base de dados tratada com o modelo ajustado para realizar o treinamento, e descreve como o processo é realizado.

4.1 Tratamento do *dataset* Datamining Cup 2013

Os resultados obtidos na Seção 3.2 demonstram inconsistências dos dados, assim como valores ridículos para captura inicial de sessões, e sessões que podem ser consideradas inúteis, por terem duração total nula. Outros valores absurdos de entrada foram observados ao diretamente ler registros do banco. Além de certas variáveis possuírem valores confusos, outras também apresentam valores novos para o *dataset* de validação incluso em [32]. Estas complicações são filtradas do dataset de treinamento, produzindo dados tratados, livres de distorções. A Tabela 8 apresenta uma lista de filtragem aplicadas, o total de entradas ou colunas removidas pela filtragem, sendo aplicada consecutivamente.

Os três primeiros itens lidam com amostras que em potencial podem atrapalhar a fase do treinamento. De maneira análoga, as remoção das colunas visa reduzir dimensionalidade, além evitar treinamentos enviesados, como a análise feita em Subsubseção 3.2.2.4 aponta como possibilidade. Além disso, a coluna “orderable” é removida tanto por seu pouco impacto, quanto por seus valores inconsistentes com a descrição apresentada em [31]. A remoção da coluna “onlineStatus” é devido ao caso dos valores quase sempre não mudarem entre início e fim de sessão, logo o treinamento pode acabar tendencioso, além ser uma informação intuitivamente irrelevante. Agora, a remoção de “ordered” é feita para incentivar o uso de mais perfis, pois ao usar uma variável que pode assumir somente dois

Tabela 8 – Filtragens de linhas e colunas do dataset Datamining World Cup 2013

Total de linhas e colunas iniciais	**427765 linhas, 24 colunas**
Descrição	**Total de linhas ou colunas afetadas**
Removendo sessões com duração nula	2248 linhas removidas
Removendo sessões com primeiro registro com mais de duas horas	3746 linhas removidas
Removendo sessões com zero interações no total	833 linhas removidas
Removendo sessões com um único registro	4960 linhas removidas
Removendo coluna "*startWeekday*"	1 coluna removida
Removendo coluna "*startHour*"	1 coluna removida
Removendo coluna "*orderable*"	1 coluna removida
Removendo coluna "*onlineStatus*"	1 coluna removida
Removendo coluna "*ordered*"	1 coluna removida
Total de linhas e colunas finais	**415378 linhas, 19 colunas**

valores, e que está presente em todas as entradas, ele poderá causar uma tendência de pesos ficarem mais próximos a essa variável que as outras. Ao final, temos uma redução de 23,12% (vinte três vírgula doze por cento) do tamanho total do *dataset*. O cálculo deste valor é demonstrado abaixo.

$$\%\text{ de redução } = \frac{(427765 \times 24) - (415378 \times 19)}{427765 \times 24} \times 100\% \approx 23.12\%$$

Por fim, os dados são novamente transformados e filtrados para obter vetores que presentam tanto o início de uma sessão, quanto fim, permitindo seu uso no modelo *PredicTour*. São agrupadas as entradas pelo "sessionId", e então são obtidos os primeiros e ultimos registros. É criado então um novo dataset, onde cada linha é constituída pelas variáveis do primeiro registro, criando colunas para cada uma. Então é acrescido as colunas de variáveis dinâmicas do último registro, em respeito ao "sessionId" de cada registro. As colunas finais tem acrescidas em seu nome sufixo "_last". O dataset final pode descrito como cada entrada sendo o registro de início e fim da jornada de um usuário interagindo com o site. O conjunto final de variáveis é listado na Tabela 9.

4.2 Adaptação de PredictTour

Prever a mobilidade urbana e o comportamento de um usuário em um site de compras apresentam desafios distintos devido às diferenças inerentes entre esses dois contextos. Na mobilidade urbana, é comum lidar com locais visitados por várias pessoas, padrões recorrentes e valores discretos, como o número de paradas de ônibus ou estações de metrô. Essas características exigem uma abordagem de modelagem específica para capturar adequadamente os padrões de movimento e as preferências dos indivíduos.

Tabela 9 – Variáveis de dataset tratado para treino.

Variáveis estáticas	**Variáveis dinâmicas**	**Variáveis dinâmicas com valores finais**
sessionId	duration	duration_last
userId	cCount	cCount_last
userScore	bCount	bCount_last
pronoum	cMinPrice	cMinPrice_last
age	cMaxPrice	cMaxPrice_last
paymentCount	cSumPrice	cSumPrice_last
daysSinceLastOrder	bMinPrice	bMinPrice_last
-	bMaxPrice	bMaxPrice_last
-	bSumPrice	bSumPrice_last
-	bStep	bStep_last
-	maxSpendVal	maxSpendVal_last

Por outro lado, ao analisar a interação de um usuário em um site de compras, os pesquisadores frequentemente encontram valores em espaços contínuos, como duração de sessões, preços de produtos e intervalos de tempo entre ações. Esses fatores podem variar consideravelmente entre os usuários e exigem uma abordagem diferente para prever adequadamente seus comportamentos e preferências.

Dessa forma, é essencial adaptar o modelo *PredicTour*[1] para atender às necessidades específicas do *dataset* de treinamento descrito na Seção 4.1. Estes ajustes no modelo visam remover partes específicas para contexto de mobilidade. Não é objetivo deste processo acrescentar fases ou partes ao modelo, buscando manter ele o mais intacto possível.

Resumindo o fluxo do modelo apresentado na Seção 2.5, temos três etapas: criação de descritores de mobilidade, extração de perfis e predição de mobilidade. Na primeira etapa, os descritores podem ser substituídos pelas amostras do dataset final, capturando dados de início e fim de sessão. A segunda etapa possui mais modificações, ajustando a quantidade de perfis e parâmetros de treinamento para atender ao novo contexto. Na etapa final, a classificação de "retornante" e "explorador" é omitida, e o método para obter valores médios é adaptado com base nos dados históricos ou na média de usuários com inícios de sessão similares.

Em suma, o processo de adaptação do modelo envolve ajustes em suas etapas-chave para garantir que ele seja eficaz em seu novo contexto. Essas modificações são feitas considerando as diferenças entre os domínios de aplicação e as necessidades específicas de cada caso. Ao aplicar essas mudanças, espera-se que o modelo atualizado consiga fornecer resultados precisos e significativos, demonstrando sua versatilidade e aplicabilidade em diferentes cenários.

4.3 Treinamento *PredicTour* com dados tratados

Utilizando o conjunto final de dados de treinamento e validação, conforme descrito na Seção 4.1, juntamente com o modelo adaptado da Seção 4.2, o treinamento do SOM e FCM foi realizado. Não houve separação dos conjuntos de dados nem a criação de graus de dificuldade, considerando que apenas algumas amostras dos conjuntos de treinamento e validação possuem histórico no conjunto de treinamento. A única etapa específica envolveu o descarte das colunas "sessionId" e "userId", por serem exclusivas para cada amostra, além da normalização dos valores das colunas para manter relações internas e externas.

Os parâmetros, como tamanho do mapa, quantidade de perfis, número de iterações e taxa de aprendizado, foram selecionados iterativamente, visando otimizar funções específicas para cada parte do modelo. Para o SOM utilizado no *PredicTour*, uma busca em grade foi realizada, iterando pelos parâmetros de *sigma* (distância dos vizinhos), taxa de aprendizado e número de épocas, buscando minimizar o erro de quantização do SOM. Para o FCM, não foi necessário realizar uma busca, pois os valores-base de cinco clusters, máximo de 150 (cento e cinquenta) iterações e m igual a 2 (dois) produziram resultados satisfatórios.

A previsão de jornada não apresentou melhorias significativas ao iterar parâmetros para minimizar os erros, devido à alta dimensionalidade e aos diferentes tipos de valores. Assim, valores arbitrários foram escolhidos para os parâmetros de seleção de amostras médias em cada previsão. Em suma, embora algumas limitações tenham sido encontradas durante o processo de adaptação e otimização do modelo, os resultados obtidos demonstram qualidades satisfatórias para realizar uma análise dos resultados.

5 RESULTADOS OBTIDOS

Na Seção 4.3 é descrito como o treinamento do modelo *PredicTour*[1] adaptado, foi treinado com um *dataset* normalizado, também descrito neste trabalho no Capítulo 3, e seus tratamentos especiais em Seção 4.1. Neste capítulo são apresentados os resultados obtidos em cada etapa do treinamento, bem como gráficos e ponderações, e análises discutindo possibilidades de interpretação.

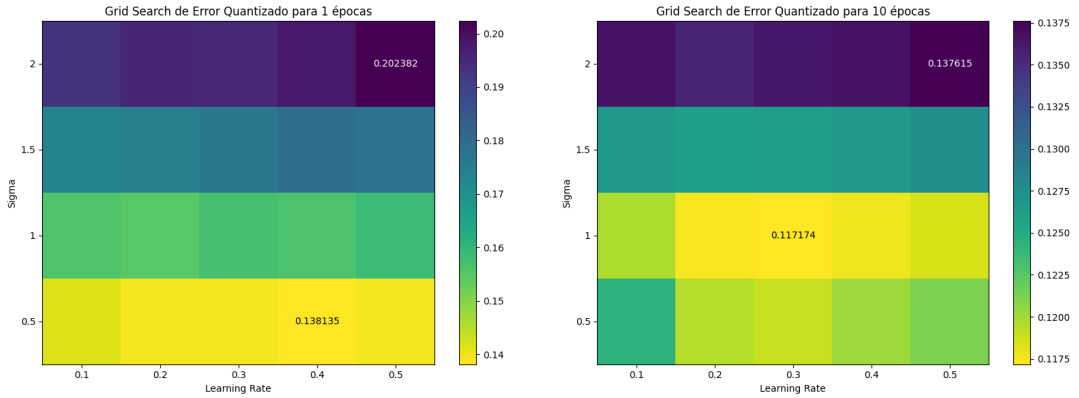
5.1 Resultados de Treinamento de SOM, erro total e busca em grade de parâmetros otimizado

Como descrito na Seção 4.3, o treinamento do SOM usado no modelo é feito iterativamente a fim de escolher o melhor conjunto de parâmetros. O conjunto de dados é o mesmo para todos os modelos, no entanto, os pesos são iniciados aleatória mente a cada variação de parâmetro. Cada conjunto de parâmetros foi treinado diversas vezes, mas, os valores de erro foram os mesmo, ou muito próximos, descartando a possibilidade de estados aleatórios de pesos inicias influenciarem majoritariamente os resultados.

Como demonstrando na Figura 10, os parâmetros que produzem o menor erro quantizado são: taxa de aprendizado 0.2 (zero vírgula dois), *sigma* (distância para considerar com neurônio vizinho) igual a 1 e quantidade de épocas ideal é 30 (trinta). O SOM resultante, bem como o mapa de distâncias médias, e o mapa para cada variável é demonstrado nas figuras Figura 11 e Figura 12 respectivamente. E por meio deles é possível observar como cada dimensão do mapa identifica quais neurônios estão mais próximos de quais valores dos dados. Estes resultados apontam um bom treinamento, visto que, o erro quantizado do SOM ficou aproximadamente 0.1156650, sendo um valor satisfatório para continuar com treinamento do modelo.

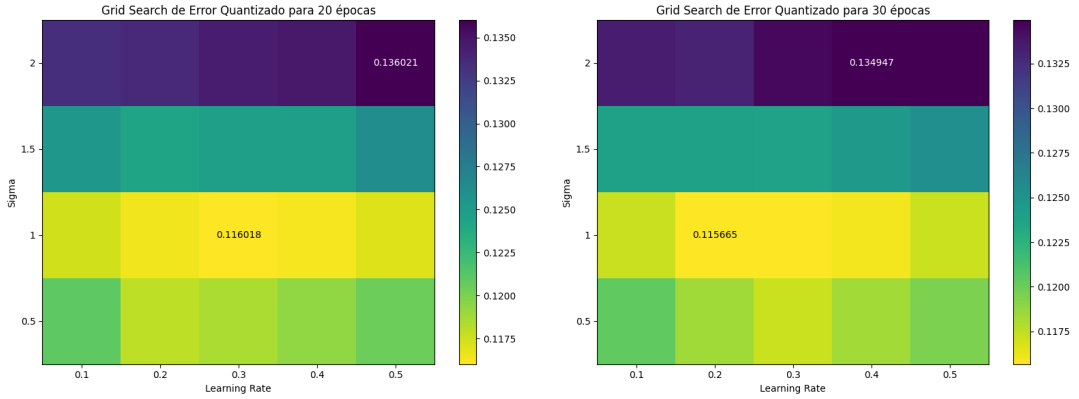
5.2 Resultados de Treinamento de FCM com mapa de pesos do SOM

Com o treinamento bem-sucedido do SOM, é viável prosseguir com o treinamento do FCM, empregando o mapa de neurônios gerado pelo SOM. Para cada neurônio, atribui-se uma classificação com múltiplos graus de pertinência a um número arbitrário de perfis. Após a classificação dos neurônios, avalia-se a qualidade do treinamento por meio de métricas específicas. Entre as métricas utilizadas estão: Silhueta, Calinski-Harabasz e Índice de Davies-Bouldin, que permitem uma análise mais aprofundada e uma compreensão do desempenho do modelo durante o treinamento. Os resultados obtidos são apresentados



(a) Resultados de *grid search* com uma época.

(b) Resultados de *grid search* com dez épocas.



(c) Resultados de *grid search* com vinte épocas.

(d) Resultados de *grid search* com trinta épocas.

Figura 10 – Resultados de *grid search* de parametros de SOM para diferentes épocas.

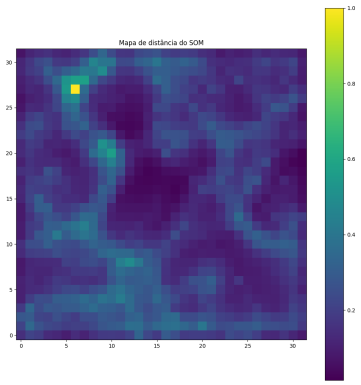


Figura 11 – Mapa de distâncias entre neurônios.

na Figura 13, onde estão separados os resultados para as três métricas separadamente.

Estes resultados apontam que o número ideal de perfis é 2 (dois). Além disso, do total de 1024 neurônio, 699 foram classificados no “perfil 1”, com a média da margem de vitória, isto é, por quanto o grau do perfil ganhou, ficou aproximadamente 0.281382, o que significa que ele ganha por quase um terço de grau quando supera o outro perfil. No entanto, deste mesmo grupo, “perfil 1” tem um total de 23 integrantes marginalmente

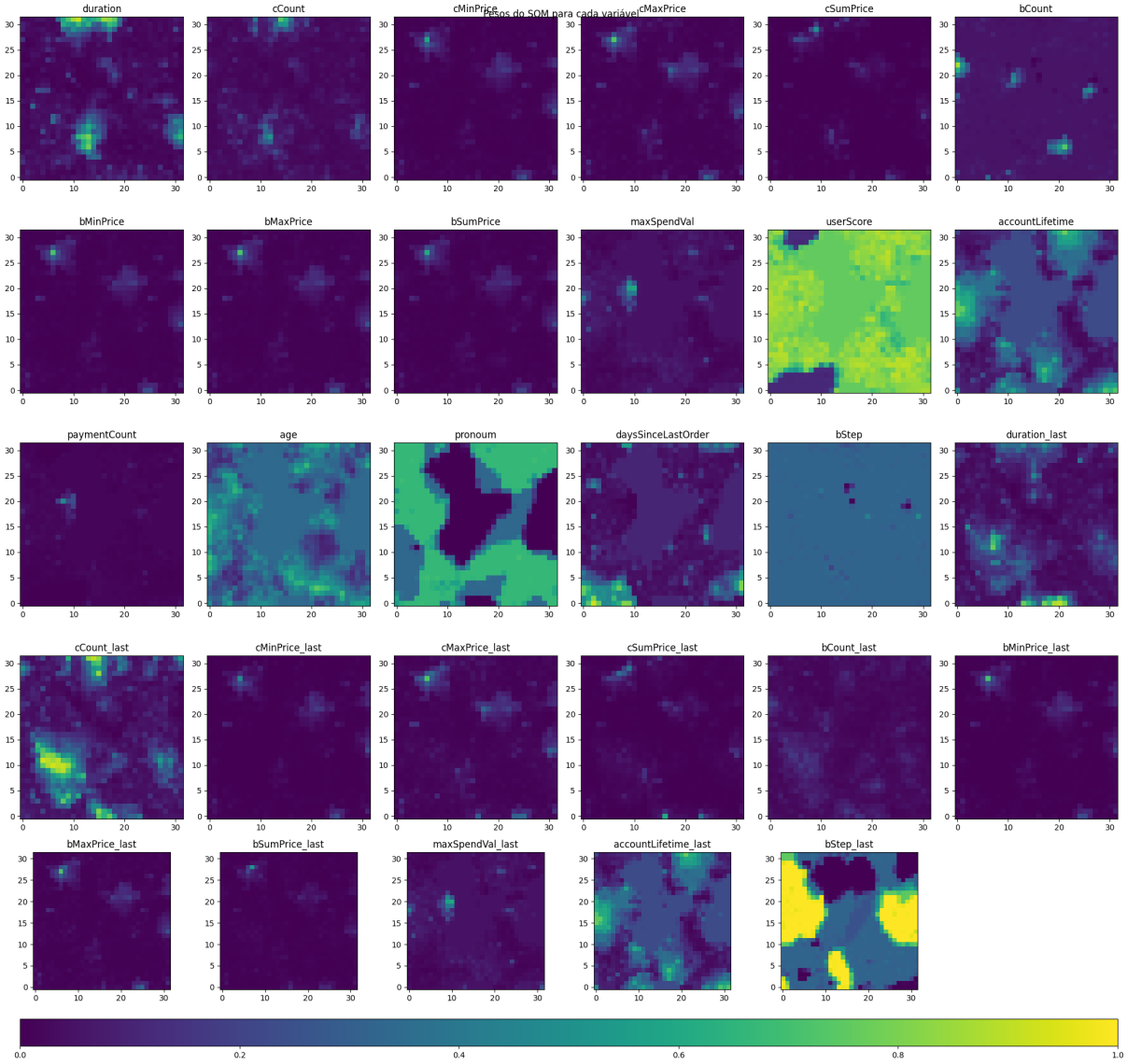


Figura 12 – Mapa de pesos do SOM para cada variável.

participando do grupo, por um total de 0.05 de diferença não são classificados em outro perfil, em contrapartida, o "perfil 2" possui somente 4 neurônios marginais.

A Figura 14 apresenta os valores médios dos pesos dos neurônios de cada perfil. A Figura 14a ilustra que o SOM treinado tende a mapear entradas com valores altos para a variável "bStep_last" em neurônios com maior grau no "perfil 2". De forma análoga, a Figura 14b mostra que o "perfil 1" é o representante majoritário para a variável "daysSinceLastOrder" com valores altos. Nas demais variáveis, os dois perfis se sobrepõem, dificultando a identificação de variáveis proeminentes em algum perfil específico.

5.3 Resultados da previsão de Jornada de Usuário

Agora, na parte final do modelo, é feita a predição da jornada de um usuário. Nesta parte, é usado o *dataset* de validação, tratado igualmente ao de treinamento, onde

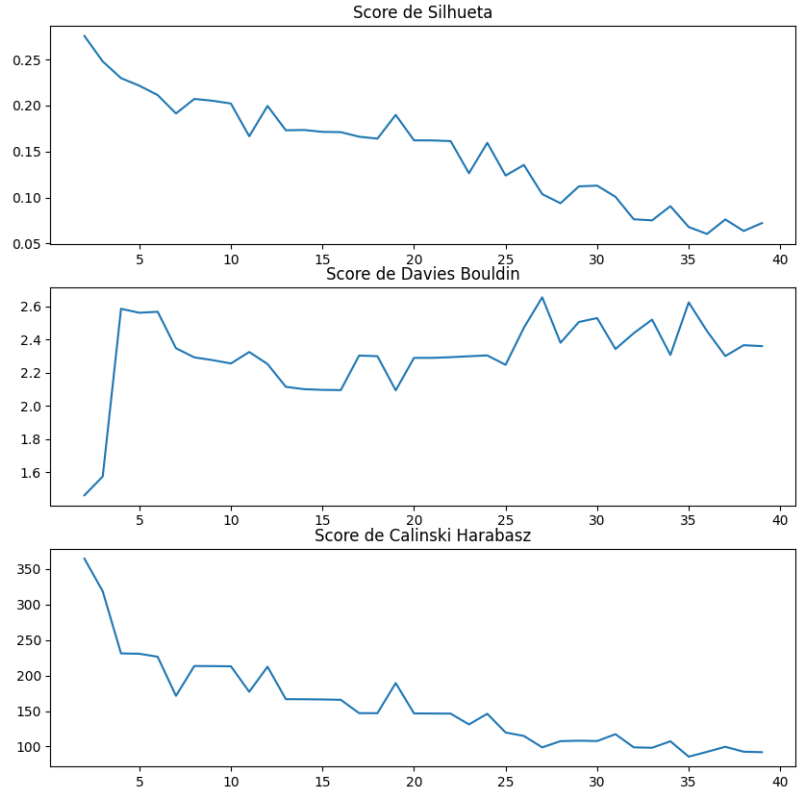


Figura 13 – Resultados de Treinamento de FCM para quantidades de perfis diferentes.

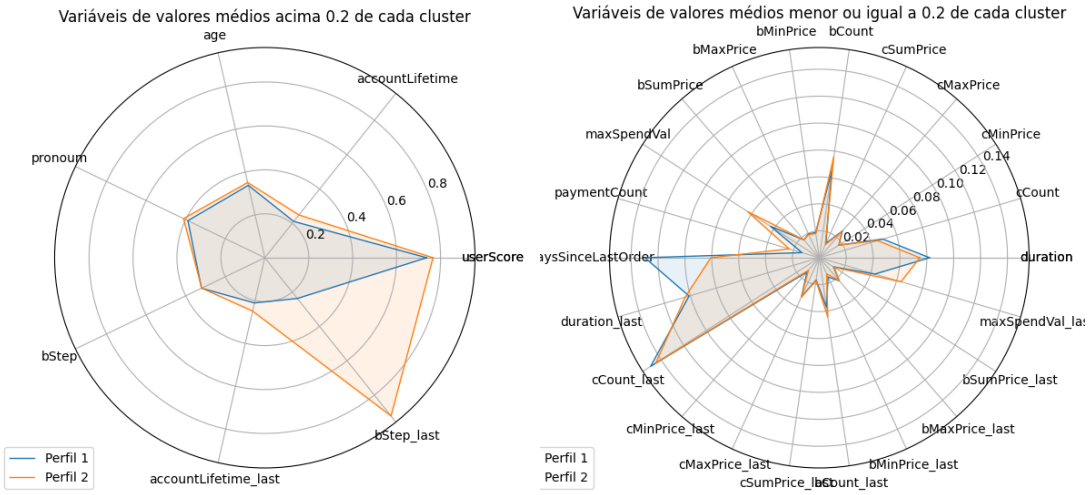
Tabela 10 – Erro médio de predição de jornada.

Nome de variável	**Média de erro**
duration_last	836.875951
cCount_last	11.739108
cMinPrice_last	34.495147
cMaxPrice_last	84.916103
cSumPrice_last	610.017100
bCount_last	2.525708
bMinPrice_last	41.129722
bMaxPrice_last	63.937220
bSumPrice_last	116.176661
maxSpendVal_last	855.226527
bStep_last	2.165703

amostras inéditas para os modelos são utilizadas para validar a capacidade preditivas do modelo. O processo segue o mesmo descrito no Algorítmo 2, com exceção de pular o passo de obter o v_{class} do usuário, como comentado na Seção 4.2, isto se deve a adaptação necessária pela mudança de contexto.

Os resultados desta etapa estão apresentados na Figura 14. A Tabela 10 mostra o erro médio da predição completa do modelo. A partir desses dados, é possível concluir que o modelo, apesar de algumas falhas de classificação aparentes, realiza previsões razoáveis para algumas variáveis, enquanto outras apresentam menor precisão.

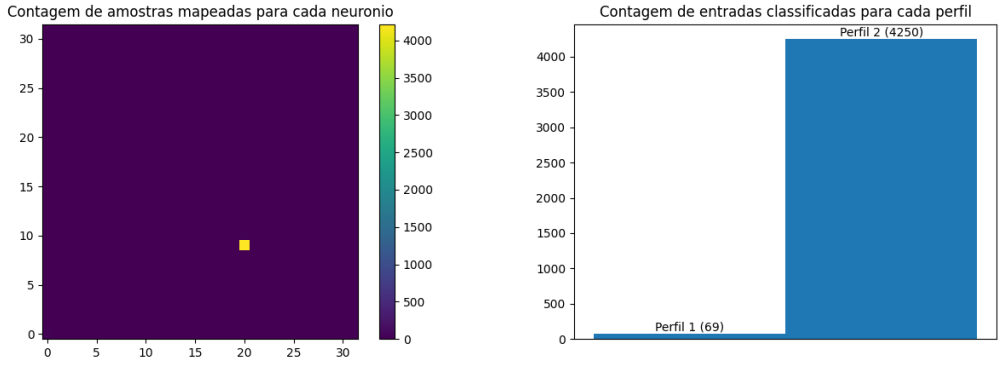
As variáveis “cCount_last”, “bCount_last” e “duration_last” exibem resultados



(a) Pesos de SOM com valores em médias acima de 0.2 separados por perfis de maior grau.

(b) Pesos de SOM com valores em médias abaixo ou igual a 0.2 separados por perfis de maior grau.

Figura 14 – Médias de pesos de SOM classificados por meio do FCM.



(a) Contagem de amostras mapeadas por neuronio.

(b) Contagem de amostras classificadas para cada perfil.

Figura 15 – Resultados de Predição de jornada.

aceitáveis, considerando a ampla variação que seus valores podem apresentar. Embora outras variáveis possam ter erros menores em comparação com a magnitude de seus valores, é importante considerar que estão diretamente relacionadas. Portanto, erros distintos nessas variáveis podem indicar dificuldades na identificação dos possíveis valores finais de carrinho e interesse.

Por fim, a variável “bStep_last”, apontada como um critério forte para classificação, apresenta erros significativos no “perfil 2”. Com erros de até 2 passos e considerando que o valor é discreto e varia entre 1 e 5, a taxa de erro é consideravelmente alta.

6 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho propôs avaliar a capacidade preditiva do modelo *PredicTour*, considerando suas técnicas incomuns para prever comportamentos. As adaptações descritas no Capítulo 4 visaram ajustar o modelo para realizar predições em um dataset com temáticas e valores distintos. Os resultados apresentados no Capítulo 5 sugerem que a adaptação foi insuficiente para lidar com os desafios e nuances do dataset, mas ainda assim obteve resultados razoáveis em algumas predições.

Uma possível justificativa para os resultados insatisfatórios pode ser a mudança drástica na problemática alvo do modelo. No trabalho original, os dados eram sempre discretos e limitados, sem valores nulos. Neste estudo, o modelo teve de lidar com casos mais complexos. Ainda que o objetivo fosse adaptar o modelo da forma mais pura possível, algumas alterações necessárias ficaram pendentes, o que poderia resultar em melhorias no treinamento e resultados mais abrangentes.

O próprio *dataset* apresenta desafios, principalmente em relação ao conteúdo. Diversas vezes, durante o estudo e coleta de informações, foi necessário retreinar os modelos devido à necessidade de ajustes nos filtros, normalizações e descarte de colunas. Além disso, algumas características podem ter afetado negativamente o desempenho, como valores de carrinho muito semelhantes e sessões claramente errôneas. Revisar o *dataset* usado para treinamento e validação pode ser fundamental para melhorar os resultados.

Outro aspecto a ser explorado é a comparação com outros modelos. A Seção 2.1 menciona vários trabalhos, alguns dos quais se encaixam perfeitamente no dataset estudado. No entanto, o escopo deste trabalho não abrange o tratamento de todos esses modelos. Uma continuação deste estudo pode envolver a análise e adaptação de outros modelos, comparando-os com os resultados obtidos aqui.

Em suma, este trabalho atingiu seu objetivo de elucidar a capacidade preditiva do modelo *PredicTour*. Os resultados apontam um bom desempenho preditivo, ressaltando as problemáticas relacionadas aos dados utilizados para treinamento e à adaptação do modelo ao novo contexto. Por meio deste estudo, é possível obter percepções sobre como o *PredicTour* utiliza classificações de comportamentos médios para identificar possíveis decisões de usuários.

REFERÊNCIAS

[1] SENEFONTE, H. C. M. et al. Predictour: Predicting mobility patterns of tourists based on social media user's profiles. *IEEE Access*, v. 10, p. 9257–9270, 2022.

[2] NETTLE, D. The evolution of personality variation in humans and other animals. *American Psychologist*, American Psychological Association, v. 61, n. 6, p. 622, 2006.

[3] MINORITY Report. Direção: Steven Spielberg. Produção: Bonnie Curtis, Jan de Bont, Gerald R. Molen e Walter F. Parkes Roteiro: Scott Frank e Jon Cohen. Intérpretes: Tom Cruise; Colin Farrell; Samantha Morton; Max von Sydow e outros. [S.l.]: 20th Century Fox, 2002. 1 filme (145 min), son., color., 27 mm.

[4] BANDURA, A. *Self-efficacy: the exercise of control*. New York: Freeman, 1997. ISBN 0-7167-2626-2 0-7167-2850-8.

[5] GAWRONSKI, B. et al. Effects of incidental emotions on moral dilemma judgments: An analysis using the CNI model. *Emotion*, American Psychological Association (APA), v. 18, n. 7, p. 989–1008, fev. 2018.

[6] CHEN, C. et al. Exploiting social media for stock market prediction with factorization machine. In: . [S.l.: s.n.], 2014. p. 142–149.

[7] RENDLE, S. Factorization machines. In: *2010 IEEE International Conference on Data Mining*. [S.l.: s.n.], 2010. p. 995–1000.

[8] KUMAR, A. et al. Combined artificial bee colony algorithm and machine learning techniques for prediction of online consumer repurchase intention. *Neural Computing and Applications*, v. 31, p. 877–890, 02 2019.

[9] KARABOğA, D. An idea based on honey bee swarm for numerical optimization. In: . [S.l.: s.n.], 2005.

[10] QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, p. 81–106, 1986.

[11] FREUND, Y.; SCHAPIRE, R. E. A desicion-theoretic generalization of on-line learning and an application to boosting. In: VITÁNYI, P. (Ed.). *Computational Learning Theory*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995. p. 23–37. ISBN 978-3-540-49195-8.

[12] KLEINBERG, E. M. Stochastic discrimination. *Annals of Mathematics and Artificial Intelligence*, v. 1, n. 1, p. 207–239, Sep 1990. ISSN 1573-7470. Disponível em: <https://doi.org/10.1007/BF01531079>.

[13] CHEN, T. et al. Sequence-aware factorization machines for temporal predictive analytics. 11 2019.

[14] MCILROY-YOUNG, R. et al. Aligning superhuman ai and human behavior: Chess as a model system. 06 2020.

[15] O'SHEA, K.; NASH, R. *An Introduction to Convolutional Neural Networks*. 2015.

[16] SCHMIDHUBER, J. Deep learning in neural networks: An overview. *Neural Networks*, Elsevier BV, v. 61, p. 85–117, jan 2015. Disponível em: <https://doi.org/10.1016j.neunet.2014.09.003>.

[17] MENS, K. et al. Predicting future suicidal behaviour in young adults, with different machine learning techniques: A population-based longitudinal study. *Journal of Affective Disorders*, v. 271, 04 2020.

[18] COX, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 20, n. 2, p. 215–232, 1958.

[19] COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27, 1967.

[20] FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189 – 1232, 2001. Disponível em: <https://doi.org/10.1214/aos/1013203451>.

[21] CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, Sep 1995. ISSN 1573-0565. Disponível em: <https://doi.org/10.1007/BF00994018>.

[22] GRENDAS, L. et al. Comparison of traditional model-based statistical methods with machine learning for the prediction of suicide behaviour. *Journal of Psychiatric Research*, v. 145, 11 2021.

[23] ISHWARAN, H. et al. Random survival forests. *The Annals of Applied Statistics*, Institute of Mathematical Statistics, v. 2, n. 3, p. 841 – 860, 2008. Disponível em: <https://doi.org/10.1214/08-AOAS169>.

[24] GONG, H. et al. Psychosocial factors predict the level of substance craving of people with drug addiction: A machine learning approach. *Int. J. Environ. Res. Public Health*, MDPI AG, v. 18, n. 22, p. 12175, nov. 2021.

[25] ALAM, K. M. et al. A comparative machine learning study to predict drug addiction in bangladesh. In: . [S.l.: s.n.], 2021. p. 1–6.

[26] JOHN, G. H.; LANGLEY, P. *Estimating Continuous Distributions in Bayesian Classifiers*. 2013.

[27] SAFARA, F. A computational model to predict consumer behaviour during covid-19 pandemic. *Computational Economics*, v. 59, p. 1–14, 04 2022.

[28] KOHONEN, T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, v. 43, n. 1, p. 59–69, Jan 1982. ISSN 1432-0770. Disponível em: <https://doi.org/10.1007/BF00337288>.

[29] BEZDEK, J. *Pattern Recognition With Fuzzy Objective Function Algorithms*. [S.l.: s.n.], 1981. ISBN 978-1-4757-0452-5.

[30] DUNN, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, Taylor & Francis, v. 3, n. 3, p. 32–57, 1973. Disponível em: <https://doi.org/10.1080/01969727308546046>.

[31] GK Software Group. *Data Mining Cup*. 2022. Disponível em: <https://www.data-mining-cup.com/>. Acesso em: 02 de novembro 2022.

[32] GK Software Group. *Data Mining Cup 2013*. 2022. Disponível em: <https://www.data-mining-cup.com/reviews/dmc-2013/>. Acesso em: 02 de novembro 2022.