Credit Card Customer: Descoberta de Subgrupos Utilizando Beam-Search

Amanda Mendes Pinho[1] **Gabriel Tonioni Duarte**[2]
João Vítor Fernandes Dias[2] **Larissa Duarte Santana**[2]

[1]

Abstract. *This study applies concepts of Descriptive Learning using the Beam Search algorithm to discover interpretable subgroups in a dataset containing behavioral variables of credit card customers. The dataset includes anonymized information from approximately 9,000 users over a six-month period. The analysis enabled the identification of distinct customer profiles based on usage frequency, purchasing patterns, and financial behavior, demonstrating the effectiveness of Beam Search in identifying relevant patterns sensitive to search and evaluation parameters.*

Resumo. *Este artigo aplica conceitos de Aprendizado Descritivo utilizando o algoritmo Beam Search para descoberta de subgrupos interpretáveis em um dataset com variáveis comportamentais de clientes de cartão de crédito. O conjunto de dados contém informações anonimizadas de cerca de 9000 usuários em um período de 6 meses. A análise permitiu a identificação de perfis distintos de clientes com base na frequência de uso, padrão de compras e comportamento financeiro, demonstrando a eficácia do Beam Search na identificação de padrões relevantes sensíveis aos parâmetros de busca e avaliação.*

1. Introdução

O crescimento exponencial na geração de dados tem impulsionado o desenvolvimento de métodos analíticos capazes de extrair *insights* relevantes de forma eficiente, interpretável e útil para a tomada de decisão. Nesse cenário, a mineração de dados descritiva tem ganhado destaque, especialmente por meio da tarefa de descoberta de subgrupos, voltada à identificação de padrões locais relevantes em relação a uma variável-alvo. Segundo [Proença and outros 2022], essa técnica é robusta a ruídos e variações estatísticas, sendo especialmente indicada para domínios onde a interpretabilidade e a consistência dos resultados são essenciais, como na medicina, marketing e ciências sociais.

A descoberta de subgrupos posiciona-se entre abordagens descritivas, como análise exploratória e regras de associação, e métodos supervisionados preditivos. Seu objetivo é identificar subconjuntos que apresentem distribuições significativamente distintas da população geral. Como discutido por [Gamberger and Lavrač 2002], essa tarefa busca gerar descrições compreensíveis e acionáveis. O algoritmo *Beam Search*, segundo [Atzmueller 2015], é uma estratégia eficaz para explorar espaços de busca de alta dimensionalidade, permitindo extrair subgrupos com elevado valor explicativo, mesmo em bases de dados complexas.

No contexto financeiro, essa abordagem revela-se estratégica diante da sensibilidade dos dados envolvidos. A capacidade de isolar perfis de clientes ou comportamentos atípicos é útil para aplicações como análise de crédito, detecção de fraudes e segmentação. Além de facilitar a geração de regras interpretáveis, contribui para decisões mais informadas e precisas. Estudos como [Maina et al. 2019] demonstram a preservação de subgrupos relevantes após anonimização de dados financeiros, enquanto [Dubowski and Mokryn 2021] destacam seu uso para detectar vieses em modelos de risco. Ainda, Revisões como a de [Helal et al. 2016] apontam aplicações práticas em finanças, e [Herrera et al. 2010] ampliam a discussão para marketing e comportamento do consumidor, ambos fortemente conectados ao setor.

Neste trabalho, utilizou-se uma base de dados financeira em conjunto com o algoritmo SD proposto por [Gamberger and Lavrač 2002], que aplica *Beam Search* para construir regras descritivas com base em uma função de qualidade parametrizada. Essa abordagem mostrou-se eficaz na identificação de perfis de comportamento entre clientes de cartão de crédito, revelando padrões associados à frequência de uso, volume de compras e níveis de endividamento.

2. Metodologia

O projeto foi desenvolvido sobre a base de dados *Credit Card Dataset for Clustering*, disponível no Kaggle. E para os experimentos realizados, foi utilizado o algoritmo Beam Search, implementado pela biblioteca Pysubgroup, que facilita a tarefa de descoberta de subgrupos ao fornecer métodos e ferramentas de implementação.

2.1. Base de Dados

O conjunto de dados resume o comportamento de uso de cerca de 9.000 titulares ativos de cartão de crédito num período de 6 meses. O arquivo está no nível do cliente, com 18 variáveis comportamentais. O que permite analisar perfis comportamentais de clientes, tais como clientes mais ou menos endividados (`BALANCE`, `CASH_ADVANCE`), padrões de consumo (`PURCHASES`, `PURCHASES_FREQUENCY`), capacidade de pagamento e uso do limite (`PAYMENTS`, `CREDIT_LIMIT`) e perfis de risco ou fidelização (`TENURE`, frequência de uso). Nesta seção, apresentam-se todas as variáveis disponíveis no conjunto de dados, acompanhadas de uma descrição sucinta que esclarece sua função e importância para a análise comportamental dos clientes.

2.1.1. Variáveis de Saldo e Limite

- BALANCE (Saldo atual): Valor total restante na conta do cliente, disponível para compras.
- BALANCE_FREQUENCY (Frequência de atualização do saldo): Valores próximos de 1 indicam atualização constante; próximo de 0, atualização esporádica.
- CREDIT_LIMIT (Limite de crédito): Valor máximo de crédito que o cliente possui no cartão.

2.1.2. Variáveis de Compras

- PURCHASES (Valor total de compras): Soma de todas as compras realizadas.
- ONEOFF_PURCHASES (Compras únicas): Compras feitas de uma só vez (à vista ou em parcela única).
- INSTALLMENTS_PURCHASES (Compras parceladas): Valor total de compras realizadas em parcelas.
- PURCHASES_FREQUENCY (Frequência de compras): 1 indica compras muito frequentes, 0 indica compras raras.
- ONEOFFPURCHASESFREQUENCY (Frequência de compras únicas): 1 indica alta frequência de compras únicas; 0, baixa frequência.
- PURCHASESINSTALLMENTSFREQUENCY (Frequência de compras parceladas): 1 indica alta frequência de parcelamentos.
- PURCHASES_TRX (Número de transações de compra): Quantidade total de compras feitas.

2.1.3. Variáveis de Adiantamento de Dinheiro

- CASH_ADVANCE (Valor de adiantamentos): Valor total de dinheiro adiantado (saques) pelo cliente.
- CASHADVANCEFREQUENCY (Frequência de adiantamentos): Frequência com que o cliente solicita adiantamentos.
- CASHADVANCETRX (Número de transações de adiantamento): Quantidade de vezes que o cliente realizou adiantamentos.

2.1.4. Variáveis de Pagamento

- PAYMENTS (Valor total pago): Valor total pago pelo cliente.
- MINIMUM_PAYMENTS (Pagamento mínimo): Valor mínimo que o cliente pagou em seus extratos.
- PRCFULLPAYMENT (Porcentagem de pagamento total): Porcentagem do valor total da fatura que foi paga. Valores próximos de 1 indicam pagamento completo frequente.

2.1.5. Variáveis de Tempo de Relacionamento

- TENURE (Tempo de relacionamento): Tempo, em meses, que o cliente possui o cartão de crédito.

2.2. Pré-Processamento de dados

O pré-processamento dos dados, é fundamental para garantir a consistência e a qualidade das análises. Esse processo envolveu o tratamento de valores nulos e a discretização de variáveis numéricas, considerando as características específicas da base utilizada. As decisões adotadas nessa fase visaram mitigar o impacto de dados ausentes, adaptar a representação das variáveis ao contexto da análise e possibilitar a extração de padrões mais interpretáveis e estatisticamente relevantes nos experimentos subsequentes.

2.2.1. Tratamento de Valores Nulos

A base de dados apresentava alguns valores nulos, o que poderia comprometer a execução do algoritmo. Assim, adotaram-se duas estratégias distintas:

- **Remoção de valores nulos**, utilizada no experimento da seção 3.1.2;
- **Substituição pela mediana da coluna**, utilizada no experimento da seção 3.1.2 e na análise da seção 4. A mediana é uma métrica que apresenta maior robustez à presença de outliers, presentes em grande volume nos dados.

2.2.2. Discretização dos Dados

A discretização é uma etapa importante no processo de descoberta de subgrupos, pois seletores categóricos geralmente tornam as regras mais interpretáveis para os analistas. Neste trabalho, foi utilizada a discretização por quartis (`qcut`) para transformar variáveis numéricas em categorias qualitativas. No entanto, nem todas as variáveis apresentaram uma distribuição apropriada para esse tipo de transformação. Para priorizar resultados mais imediatos e consistentes, apenas as colunas com número suficiente de valores distintos foram discretizadas. As demais foram mantidas em sua forma contínua, garantindo a preservação de suas propriedades numéricas para análise posterior. Esse processo está exemplificado no código do Listing 1.

Listing 1. Discretização automática com quartis para variáveis numéricas.

```
df_binned = df.drop(columns=['CUST_ID'], errors='ignore')
df_binned = df.fillna(df.median(numeric_only=True))

numeric_cols = df_binned.select_dtypes(include=[np.number]).columns.tolist()

for col in numeric_cols:
   bin_col = f'{col}_BIN'
   try:
      df_binned[bin_col] = pd.qcut(df[col], q=4, labels=['low', 'med_low', 'med_high', '
          high'])
   except ValueError:
      print(f"Skipping␣{col}␣-␣too␣few␣unique␣values␣to␣bin")
```

Durante o processo, as seguintes variáveis foram identificadas como inadequadas para discretização automática por apresentarem poucos valores únicos: `BALANCE_FREQUENCY`, `ONEOFF_PURCHASES`, `INSTALLMENTS_PURCHASES`, `CASH_ADVANCE`, `ONEOFF_PURCHASES_FREQUENCY`, `PURCHASES_INSTALLMENTS_FREQUENCY`, `CASH_ADVANCE_FREQUENCY`, `CASH_ADVANCE_TRX`, `PRC_FULL_PAYMENT` e `TENURE`. Dessa forma, foi possível aplicar a discretização apenas onde havia suporte estatístico, mantendo a integridade e a interpretabilidade dos dados ao longo das análises.

2.3. Descoberta de Subgrupos com Beam Search

A descoberta de subgrupos é uma tarefa descritiva voltada à identificação de subconjuntos de dados que se destacam em relação a uma variável-alvo, sendo especialmente útil em contextos que exigem interpretabilidade e personalização das regras extraídas. Nesse cenário, a biblioteca *pysubgroup* se destaca por oferecer uma estrutura flexível e eficiente para a definição de funções de qualidade, configuração das condições de descoberta e avaliação sistemática dos padrões gerados, apresentando compatibilidade com

diferentes tipos de dados e objetivos analíticos [Lemmerich and Becker 2018]. Entre os algoritmos disponíveis, o *Beam Search* configura-se como uma estratégia heurística e controlada que visa encontrar subgrupos descritivos por meio de um processo composto por cinco etapas principais: **inicialização**, **expansão**, **avaliação**, **otimização (feixe)** e **iteração**.

Na fase de **inicialização**, o algoritmo parte de subgrupos simples, geralmente construídos a partir de descritores básicos. A **expansão** ocorre com o enriquecimento desses subgrupos por meio da adição de novos descritores, aumentando sua especificidade. Em seguida, a **avaliação** de cada subgrupo é realizada com base em uma função de qualidade, que define sua relevância conforme critérios estatísticos ou heurísticos previamente definidos. Durante a etapa de **seleção do feixe**, apenas os k subgrupos com melhor desempenho são mantidos, garantindo que o algoritmo concentre esforços em regiões promissoras do espaço de busca. Por fim, a fase de **iteração** repete esse ciclo até que se atinja a profundidade máxima estabelecida ou os critérios de parada definidos sejam satisfeitos.

Portanto, no *Beam Search*, os subgrupos gerados competem entre si por posições no feixe, sendo mantidos apenas aqueles com melhor desempenho de acordo com a função de qualidade. Isso garante uma busca eficiente por padrões relevantes e interpretáveis, mesmo em espaços de busca com múltiplos descritores.

2.3.1. Configurações Experimentais

Nesta etapa, foram definidos os principais parâmetros utilizados nos experimentos de descoberta de subgrupos. Como variáveis-alvo, selecionaram-se atributos diretamente relacionados ao comportamento de consumo dos clientes, a saber: `PURCHASES_FREQUENCY`, `PURCHASES_TRX`, `ONEOFF_PURCHASES` e `BALANCE`. Essas variáveis foram escolhidas por refletirem aspectos relevantes da frequência de uso, volume de transações e situação financeira dos indivíduos analisados.

O espaço de busca foi explorado de duas maneiras: de forma completa, abrangendo todos os descritores disponíveis, e de forma segmentada, com restrições específicas para diferentes categorias de análise. Para a avaliação dos subgrupos gerados, foram empregadas distintas funções de qualidade, incluindo as medidas `stdQF` e `stdQFTscore`, bem como a métrica `WRAcc` (*Weighted Relative Accuracy*), definidas conforme exemplificado no código da Listing 2. Esta última considera simultaneamente a precisão relativa do subgrupo e seu tamanho em relação ao conjunto total, o que permite equilibrar relevância estatística e representatividade [Vimieiro].

Listing 2. Função auxiliar para seleção da métrica de qualidade utilizada em cada experimento.

```
def get_quality_function(name='stdQF', a=0.5, centroid='mean'):
   gen_quality_func = {
      'stdQF': ps.StandardQFNumeric(a, centroid=centroid),
      'stdQFTscore': ps.StandardQFNumericTscore(),
      'WRAcc': ps.WRAccQF(),
   }
   return gen_quality_func.get(name, gen_quality_func['stdQF'])
```

Adicionalmente, a criação das tarefas de busca foi realizada segundo o código ilustrado na Listing 3 e o parâmetro de recompensa por tamanho do subgrupo foi variado

entre os valores 0.0, 0.3, 0.5 e 1.0, a fim de investigar o impacto do tamanho das regras na qualidade das descobertas. Também foram adotados valores fixos para as demais configurações experimentais, sendo definido o tamanho mínimo dos subgrupos como 10 instâncias, e o número máximo de descritores por subgrupo variando entre 3 e 8. Tais parâmetros impactam diretamente a complexidade das regras geradas, bem como a interpretabilidade dos padrões descobertos, especialmente em um contexto de análise comportamental com múltiplas variáveis interdependentes.

Listing 3. Criação da tarefa de descoberta de subgrupos

```
def get_task(df, target, qf='stdQF', a=0.5, subgroup_size=10, desc_size=3):
   task = ps.SubgroupDiscoveryTask(
      df,
      targets[target],
      search_spaces[target],
      get_quality_function(qf, a),
      result_set_size=subgroup_size,
      depth=desc_size
   )
   return task
```

2.3.2. Função de Avaliação

Nos experimentos envolvendo a variável-alvo PURCHASES_TRX, buscou-se responder à pergunta: *"quem utiliza muito o cartão de crédito?"*. Para isso, foi empregada uma função de avaliação baseada na diferença da média ponderada, definida conforme a seguinte equação:

$$\text{score} = \alpha \times \text{instances}_{\text{subgroup}} \times (\text{mean}_{\text{sg}} - \text{mean}_{\text{dataset}}) \tag{1}$$

Nessa fórmula, o parâmetro α atua como um parâmetro de ajuste fino, permitindo controlar o equilíbrio entre a representatividade e a distinção dos subgrupos. Valores menores de α tendem a favorecer subgrupos mais compactos e diferenciados da média global, enquanto valores maiores priorizam a generalização e o tamanho dos grupos descobertos.

Para os experimentos com a variável-alvo BALANCE, a mesma função foi utilizada, com o objetivo de responder à pergunta: *"qual o perfil de clientes mais endividados?"*. Dado que se trata também de um atributo numérico e contínuo, a função mostrou-se igualmente adequada, permitindo capturar subgrupos significativamente distintos em relação ao saldo devedor médio observado no conjunto de dados.

3. Experimentos e Discussões

Neste estudo, foram realizados dois tipos de experimentos distintos. O primeiro utilizou como variável-alvo o atributo PURCHASES_TRX, com o objetivo principal de compreender o comportamento das variáveis ao se analisar descritores da área financeira, além de testar diferentes parâmetros das funções de qualidade. O segundo experimento teve como variável-alvo o atributo BALANCE, possuindo um caráter mais descritivo e interpretativo, voltado à análise dos padrões de consumo dos clientes. O código está disponível no GitHub

3.1. Análise do Atributo `PURCHASES_TRX`

O primeiro experimento teve como objetivo primário compreender os perfis de clientes que apresentam maior frequência de uso do cartão de crédito, e foi posteriormente direcionado à análise do impacto dos parâmetros da função de qualidade `StandardQFNumeric` sobre os subgrupos identificados pelo algoritmo *Beam Search*.

3.1.1. Análise de Padrões

Entre todos os experimentos, destacou-se o perfil de cliente que compra com alta frequência, sendo a variável mais discriminativa para um alto número de transações (`PURCHASES_TRX`) a frequência de compras (`PURCHASES_FREQUENCY`). Clientes com frequência ≥ 1 realizam, em média, 69 transações (bem acima da média geral de 15). Apesar de esperado, o resultado reforça a validade da abordagem, mostrando que o algoritmo identifica padrões coerentes com o domínio do problema.

3.1.2. Análise de Parâmetros

A seguir, são apresentados os experimentos conduzidos com diferentes valores do parâmetro α, evidenciando como essa variação permite refinar a análise e revelar nuances relevantes sobre este e outros perfis de clientes.

Experimento 1

Neste experimento, foi adotado o valor de $\alpha = 0.5$, o que direcionou o algoritmo a priorizar regras gerais, abrangendo uma quantidade expressiva de clientes. Os resultados do experimento estão ilustrados na figura 1.

Regra principal: `PURCHASES_FREQUENCY` ≥ 1.

- Tamanho do subgrupo: 2.190 clientes
- Média de transações: 68.96

Insight: A frequência de compras demonstrou ser a regra mais simples e, ao mesmo tempo, a mais preditiva para caracterizar clientes com elevado número de transações. Essa configuração se mostrou adequada para identificar os principais *drivers* de comportamento presentes na base de dados.

Experimento 2

Neste experimento, foi utilizado o valor de $\alpha = 0.3$, o que alterou significativamente o comportamento do algoritmo. Com essa configuração, o *Beam Search* passou a priorizar subgrupos menores e mais distintos da média global, favorecendo padrões mais extremos e específicos. Os resultados do experimento estão ilustrados na figura 3.

Regra principal: `BALANCE_FREQUENCY` $\geq$ 1.0, `INSTALLMENTS_PURCHASES` $\geq$ 620.84, `ONEOFF_PURCHASES` $\geq$ 834.8, `ONEOFF_PURCHASE_FREQUENCY` $\geq$ 0.42 e `PAYMENTS` $\geq$ 2373.9.

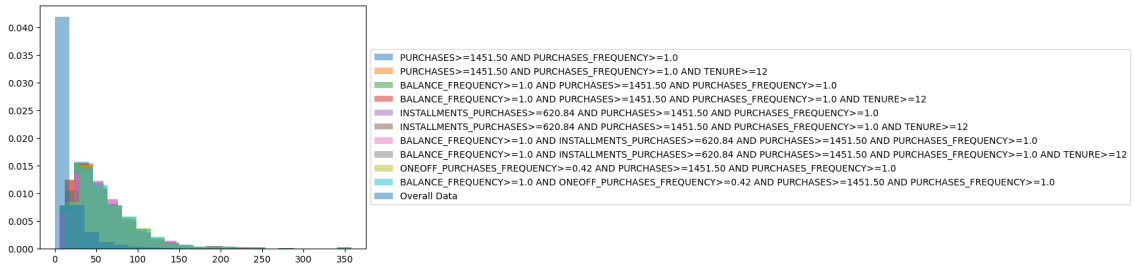

Figure 1. Distribuição dos subgrupos para o experimento realizado com `depth` = 8, α = 0.5, e competição livre entre todos os seletores.

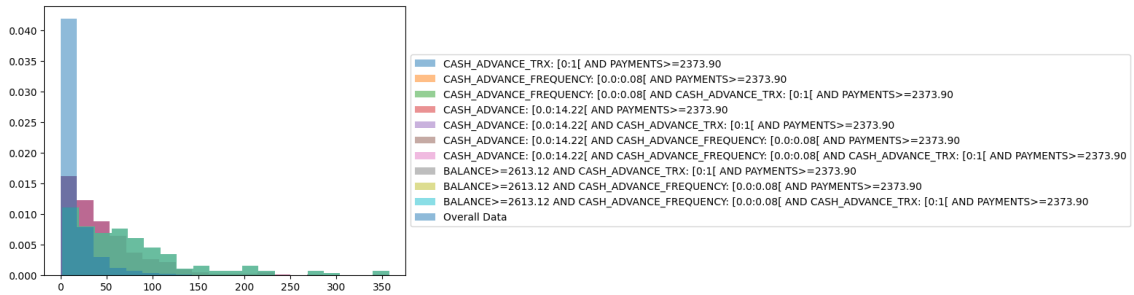

Figure 2. Distribuição dos subgrupos definidos por múltiplos descritores financeiros, com $\alpha = 0.3$, profundidade máxima igual a 8 e competição restrita a seletores do domínio financeiro.

- Tamanho do subgrupo: 632 clientes
- Média de transações: 107.03

Insight: Essa configuração revelou um perfil mais exclusivo de clientes que, além de apresentarem alta frequência de uso, também realizam transações de valores significativamente mais altos. O uso de α reduzido permitiu capturar nuances comportamentais mais específicas e distantes da média global, destacando subgrupos de maior valor transacional.

Figure 3. Distribuição dos subgrupos gerados com $\alpha = 0.3$, profundidade máxima igual a 8 e top-10 com competição entre todos os seletores.

3.1.3. Análise Crítica da Variação de α

O parâmetro α é um hiperparâmetro de ajuste fino que permite controlar o *trade-off* entre encontrar subgrupos estatisticamente muito distintos (com grande diferença em relação à média global) e subgrupos que são representativos e acionáveis (com grande número de instâncias).

Valores mais baixos de α (como 0.3) reduzem a qualidade dos subgrupos, mas revelam padrões menos óbvios, como pode ser visto nas figuras 2 e 4. Embora nem

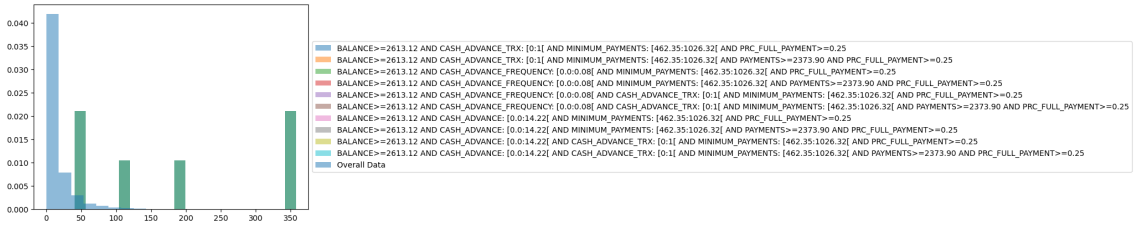

Figure 4. Distribuição dos subgrupos gerados com $\alpha = 0.3$, profundidade máxima igual a 8 e top-10 com competição restrita a seletores financeiros.

sempre interessantes, tais padrões trazem uma quantificação que pode ser interessante conforme o contexto. Regras longas podem ser redundantes e difíceis de interpretar, por isso segmentar os seletores pode melhorar a análise. Não há α ideal: para uma visão geral, $\alpha = 0.5$ é adequado; para perfis extremos, $\alpha = 0.3$ costuma ser melhor.

3.2. Análise do Atributo `BALANCE`

O segundo experimento possui um caráter mais descritivo e tem como objetivo principal identificar os tipos de clientes com comportamento mais impulsivo, que apresentam um saldo devedor acima da média global. O atributo `BALANCE` representa o saldo devedor, ou seja, o valor total que ainda precisa ser pago na fatura do cartão de crédito.

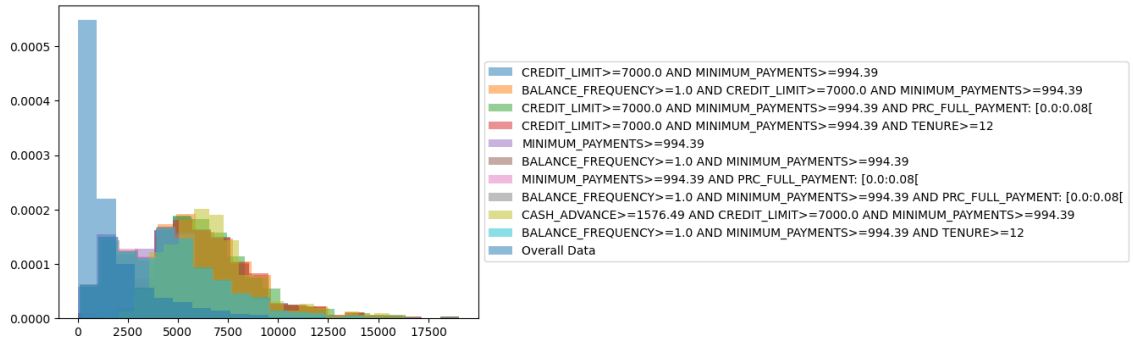

Figure 5. Distribuição do saldo devedor (`BALANCE`) para os subgrupos identificados com maior nível de endividamento, com destaque para os descritores `MINIMUM_PAYMENTS`, `CREDIT_LIMIT`, `PRC_FULL_PAYMENT` e `CASH_ADVANCE`.

De acordo com os resultados dos experimentos, observados na Figura 5 , as características que mais se repetem e definem os grupos com saldo devedor mais alto são descritas a seguir:

- **Pagamento Mínimo Elevado**: A condição `MINIMUM_PAYMENTS` ≥ 994.39 aparece em quase todas as regras de alta qualidade. Este é um indicativo direto de que o saldo devedor (`BALANCE`) é elevado, já que o pagamento mínimo geralmente é uma fração desse saldo. Trata-se do principal descritor encontrado, embora revele uma relação direta e esperada por conta da correlação entre as variáveis.
- **Limite de Crédito Alto**: A condição `CREDIT_LIMIT` ≥ 7000.0 foi a segunda mais forte nos subgrupos identificados. Clientes com limites de crédito elevados tendem a acumular saldos devedores maiores.

- **Não Pagamento do Valor Total**: A condição PRC_FULL_PAYMENT $\in$ [0.0, 0.08[indica que o cliente paga entre 0% e 8% da fatura, caracterizando o comportamento conhecido como ”rotativo”. Esses clientes financiam o saldo de um mês para o outro, acumulando juros e, acabam aumentando seu saldo devedor.
- **Frequência de Atualização do Saldo**: A condição BALANCE_FREQUENCY $\geq$ 1.0 indica que o saldo desses clientes é atualizado com frequência, sugerindo que são usuários ativos do cartão.

Com base nesses resultados, foram identificados dois perfis distintos que definem a categoria de clientes com maior saldo devedor, descritos a seguir.

3.2.1. Análise de Padrões

O subgrupo com maior média de saldo devedor (R$ 6.715,71) é caracterizado pela combinação das seguintes condições: CREDIT_LIMIT $\geq$ 7000.0, MINIMUM_PAYMENTS $\geq$ 994.39 e CASH_ADVANCE $\geq$ 1576.49. Esse padrão sugere que o comportamento de realizar saques em dinheiro é um acelerador significativo do saldo devedor para clientes que já apresentam alto nível de endividamento.

Outro perfil recorrente nos subgrupos de maior BALANCE é o de clientes com TENURE $\geq$ 12 e BALANCE_FREQUENCY $\geq$ 1.0. Esse comportamento indica que o endividamento tende a se consolidar ao longo do tempo, sendo mais comum entre clientes antigos e com uso contínuo do cartão.

Em resumo, a análise identificou com sucesso o perfil do cliente *“endividado de alto valor”*: alguém com tempo de uso elevado, limite de crédito alto, que não quita a fatura mensal, utiliza crédito rotativo e realiza saques, comportamento esse que pode ser tratado como um padrão de risco elevado para instituições financeiras.

4. Análise

Além dos experimentos, este estudo também realizou uma análise mais completa acerca de uma questão, descrever clientes com dificuldade de pagar a fatura. A análise utilizou uma mistura de variáveis categóricas e contínuas, como detalhado na na Tabela 1. As variáveis categóricas foram criadas a partir de uma discretização dos dados, como descrito na seção 2.2.2.

Table 1. Variáveis utilizadas na análise, separadas por tipo de representação

Variáveis Discretizadas	Variáveis Contínuas
BALANCE_BIN	BALANCE_FREQUENCY
PURCHASES_BIN	ONEOFF_PURCHASES
PURCHASES_FREQUENCY_BIN	INSTALLMENTS_PURCHASES
PURCHASES_TRX_BIN	CASH_ADVANCE
CREDIT_LIMIT_BIN	ONEOFF_PURCHASES_FREQUENCY
TENURE[1]	PURCHASES_INSTALLMENTS_FREQUENCY
	CASH_ADVANCE_FREQUENCY
	CASH_ADVANCE_TRX

[1] A variável TENURE não foi discretizada, mas é considerada categórica por natureza.

Table 2. Resultado da análise por descoberta de subgrupos de clientes com dificuldade de pagar a fatura.

	quality	subgroup
0	0.053724	BALANCE_BIN=='high'
1	0.050012	BALANCE_BIN=='high' AND BALANCE_FREQUENCY>=1.0
2	0.045886	BALANCE_FREQUENCY>=1.0 AND INSTALLMENTS_PURCHASES: [0.0:1.95[
3	0.045813	BALANCE_FREQUENCY>=1.0 AND PURCHASES_INSTALLMENTS_FREQUENCY: [0.0:0.08[
4	0.045813	BALANCE_FREQUENCY>=1.0 AND INSTALLMENTS_PURCHASES: [0.0:1.95[AND PURCHASES_INSTALLMENTS_FREQUENCY: [0.0:0.08[
5	0.044880	BALANCE_BIN=='high' AND BALANCE_FREQUENCY>=1.0 AND TENURE>=12
6	0.043976	BALANCE_BIN=='high' AND TENURE>=12
7	0.042736	BALANCE_FREQUENCY>=1.0 AND ONEOFF_PURCHASES_FREQUENCY: [0.0:0.08[
8	0.042736	BALANCE_FREQUENCY>=1.0 AND ONEOFF_PURCHASES: [0.0:0.01[
9	0.042736	BALANCE_FREQUENCY>=1.0 AND ONEOFF_PURCHASES: [0.0:0.01[AND ONEOFF_PURCHASES_FREQUENCY: [0.0:0.08[

Table 3. Detalhamento do resultado da análise por descoberta de subgrupos de clientes com dificuldade de pagar a fatura.

size_sg	size_dataset	positives_sg	positives_dataset	size_complement	relative_size_sg	relative_size_complement	coverage_sg	coverage_complement	target_share_sg	target_share_complement	target_share_dataset	lift	
0	2238	8950	1252	3084	6712	0.250056	0.749944	0.405966	0.594034	0.559428	0.272944	0.344581	1.623502
1	2050	8950	1154	3084	6900	0.229050	0.770950	0.374189	0.625811	0.562927	0.279710	0.344581	1.633656
2	2607	8950	1309	3084	6343	0.291285	0.708715	0.424449	0.575551	0.502110	0.279836	0.344581	1.457160
3	2606	8950	1308	3084	6344	0.291173	0.708827	0.424125	0.575875	0.501919	0.279950	0.344581	1.456606
4	2606	8950	1308	3084	6344	0.291173	0.708827	0.424125	0.575875	0.501919	0.279950	0.344581	1.456606
5	1896	8950	1055	3084	7054	0.211844	0.788156	0.342088	0.657912	0.556435	0.287638	0.344581	1.614815
6	1963	8950	1070	3084	6987	0.219330	0.780670	0.346952	0.653048	0.545084	0.288250	0.344581	1.581875
7	2802	8950	1348	3084	6148	0.313073	0.686927	0.437095	0.562905	0.481085	0.282368	0.344581	1.396145
8	2802	8950	1348	3084	6148	0.313073	0.686927	0.437095	0.562905	0.481085	0.282368	0.344581	1.396145
9	2802	8950	1348	3084	6148	0.313073	0.686927	0.437095	0.562905	0.481085	0.282368	0.344581	1.396145

Para guiar a análise, foi criada uma nova variável binária com o intuito de capturar a noção de dificuldade de pagar a fatura. Os valores dessa nova variável são dados pela fórmula: $\frac{MINIMUM\_PAYMENTS}{PAYMENTS} > 0.9$.

Para executar a tarefa de descoberta de subgrupos o algoritmo Beam Search foi utilizado para retornar um ranking com os dez melhores subgrupos encontrados e descrições com no máximo três seletores, para facilitar a interpretação. Os resultados da execução são sintetizados na tabela 2 e 3. Esta tabela mostra que todos os subgrupos encontrados tem valor de $lift > 1$, o que mostra que a correlação entre os seletores é positiva. Também são apresentados valores da métrica de qualidade $WRAcc > 0$, o que indica que os subgrupos encontrados são interessantes do ponto de vista desta métrica de qualidade.

5. Conclusão

O mercado financeiro tem se tornado cada vez mais competitivo com o passar dos anos. Tratar todos os clientes da mesma forma torna-se, assim, uma oportunidade perdida para as empresas deste setor. Milhões de transações e dados de comportamento escondem padrões valiosos que as médias gerais não revelam. Dessa forma, utilizar a *Descoberta de Subgrupos* neste contexto é ir além da análise superficial. Ao encontrar grupos com regras compreensíveis, uma instituição financeira pode criar ações de marketing, gestão de risco e fidelização de forma cirurgicamente precisa, otimizando recursos e construindo um relacionamento mais inteligente e rentável com cada segmento de sua base de clientes.

Como resultado dos experimentos realizados, constatou-se que a manutenção dos atributos numéricos em sua forma contínua, por meio da não discretização dos dados, foi uma decisão assertiva, especialmente diante das características da base utilizada. Em um conjunto com número limitado de instâncias e atributos financeiros sensíveis, a discretização poderia comprometer a qualidade analítica, ao eliminar variações sutis, porém significativas, entre os registros, especialmente quando tais variações ocorrem em intervalos estreitos. A preservação dos valores contínuos permitiu que os algoritmos identificassem padrões mais refinados, respeitando particularidades do comportamento financeiro dos clientes. Além disso, essa abordagem é relevante para funções de qualidade

como a diferença da média ponderada, cujo desempenho depende diretamente da fidelidade dos valores numéricos originais.

A geração de insights significativos a partir dos subgrupos descobertos demonstrou depender fortemente da colaboração com especialistas da área financeira, pois embora os algoritmos possam identificar padrões estatisticamente relevantes, a validação semântica e a interpretação prática desses padrões requerem essa colaboração para a identificação de regras realmente úteis, em que é realizada a filtragem daquelas, que embora sejam interessantes estatisticamente, não seriam viáveis operacionalmente, por não refletirem comportamentos significativos no contexto da instituição.

References

[Atzmueller 2015] Atzmueller, M. (2015). Subgroup discovery. *WIREs Data Mining and Knowledge Discovery*.

[Dubowski and Mokryn 2021] Dubowski, D. and Mokryn, O. (2021). Towards assessment of subgroup harms and predictive bias in risk scoring systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):15062–15070.

[Gamberger and Lavrač 2002] Gamberger, D. and Lavrač, N. (2002). Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, 17:501–527.

[Helal et al. 2016] Helal, A., Gaber, M. M., and Hüllermeier, F. (2016). Subgroup discovery algorithms: A survey and empirical evaluation. *Journal of Computer Science and Technology*, 31(5):925–945.

[Herrera et al. 2010] Herrera, F., Carmona, C. J., González, P., and del Jesus, M. J. (2010). An overview on subgroup discovery: Foundations and applications. *Knowledge and Information Systems*, 29(3):495–525.

[Lemmerich and Becker 2018] Lemmerich, F. and Becker, M. (2018). pysubgroup: Easy-to-use subgroup discovery in python. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 658–662. Set.

[Maina et al. 2019] Maina, G., Fung, B., and Chan, K. L. (2019). Subgroup preservation in financial data anonymized by a hybrid approach. *Nature Scientific Reports*, 9(1):1–13.

[Proença and outros 2022] Proença, M. L. P. and outros (2022). Robust subgroup discovery. *Data Mining and Knowledge Discovery*, 36:1885–1970.

[Vimieiro] Vimieiro, R. Aula 09 – aprendizado descritivo supervisionado. Aula da disciplina de Aprendizado Descritivo, 2024.