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, available on Kaggle, 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, disponível no Kaggle, 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 aumento exponencial na geração de dados em diversas áreas tem impulsionado a busca por métodos avançados de análise capazes de extrair insights relevantes de forma eficiente, interpretável e útil para a tomada de decisão. Nesse contexto, técnicas de mineração de dados descritiva têm ganhado destaque por sua capacidade de revelar padrões e estruturas complexas em grandes volumes de dados, especialmente por meio da tarefa de descoberta de subgrupos, voltada à identificação de padrões locais que se destacam em relação a uma variável de interesse. Segundo [**?**], essa técnica apresenta vantagens robustas frente a ruídos e variações estatísticas, sendo particularmente adequada para domínios nos quais a interpretabilidade dos resultados e a consistência das descobertas são essenciais, como na medicina, no marketing e nas ciências sociais.

Conceitualmente, a descoberta de subgrupos ocupa uma posição intermediária entre abordagens puramente descritivas, como a análise exploratória de dados e as regras de associação, e métodos preditivos baseados em modelos supervisionados de classificação.

Seu objetivo principal é identificar, dentro de uma população, subconjuntos que apresentem padrões estatisticamente relevantes com relação a uma variável-alvo, ou seja, grupos cuja distribuição difere de maneira significativa da distribuição global observada nos dados. Como apresentado por [**?**], a tarefa visa gerar descrições compreensíveis e acionáveis desses segmentos, contribuindo para a explicação de fenômenos complexos e para o apoio à tomada de decisão em ambientes que exigem interpretação humana qualificada.

No setor financeiro, essa técnica adquire papel estratégico, dada a complexidade e sensibilidade dos dados envolvidos. A possibilidade de isolar perfis específicos de clientes ou comportamentos atípicos torna essa abordagem particularmente útil para aplicações como análise de crédito, detecção de fraudes e segmentação de consumidores. Ao possibilitar a geração de regras compreensíveis e orientadas por dados, a técnica viabiliza decisões mais informadas e direcionadas por parte das instituições financeiras. A literatura respalda a efetividade dessa estratégia em contextos aplicados: [**?**] destaca sua importância em cenários que demandam transparência e clareza interpretativa, enquanto [**?**] enfatiza o papel do algoritmo *Beam Search* como uma estratégia eficaz para a exploração de espaços de busca de alta dimensionalidade. Tal abordagem tem se mostrado capaz de extrair subgrupos descritivos com elevado potencial explicativo, mesmo diante de conjuntos de dados complexos e heterogêneos.

Diversos estudos têm demonstrado o potencial da descoberta de subgrupos em contextos financeiros, seja para análise de comportamento de clientes ou para garantir a equidade e interpretabilidade de modelos preditivos. Por exemplo, [**?**] exploraram como subgrupos relevantes são preservados após processos de anonimização de dados financeiros, evidenciando a importância de manter estruturas interpretáveis mesmo em cenários de privacidade. Já [**?**] utilizaram técnicas de descoberta de subgrupos para identificar vieses e potenciais danos preditivos em sistemas de pontuação de risco, com foco em grupos vulneráveis dentro de bases financeiras.

Revisões mais amplas, como a de [**?**], destacam casos práticos do uso de descoberta de subgrupos em finanças, reforçando seu papel como ferramenta analítica robusta. Ainda, [**?**] discutem amplamente sua aplicação em marketing e comportamento do consumidor, áreas intrinsecamente ligadas às estratégias financeiras de segmentação, análise de crédito e fidelização.

Neste estudo, empregou-se uma base de dados do domínio financeiro em conjunto com o algoritmo SD, proposto por [**?**], o qual adota a estratégia heurística de busca em feixe (*Beam Search*) para a construção iterativa de regras descritivas. A técnica gera subgrupos por meio de uma função de qualidade parametrizada pela fórmula $q_g = \frac{TP}{FP+g}$, em que TP e FP representam, respectivamente, os verdadeiros e falsos positivos, e g é um parâmetro de generalização ajustável. Essa metodologia demonstrou-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. Entendimento da 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`), perfis de risco ou fidelização (`TENURE`, frequência de uso), oferecendo um contexto real sobre os dados dos clientes, o que permite a extração de insights relevantes em estudos de segmentação e comportamento financeiro.

Abaixo estão listadas todas as variáveis presentes nos dados, com uma explicação simples sobre o que representam:

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

Antes da aplicação dos algoritmos de descoberta de subgrupos, foi necessário realizar etapas de pré-processamento dos dados, fundamentais 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 primeiro experimento;
- **Substituição pela mediana da coluna**, utilizada no segundo experimento, em razão da quantidade reduzida de dados ausentes.

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.

O código abaixo ilustra o processo de discretização com tratamento de exceções para variáveis inadequadas:

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
- 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 [**?**]. 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*). 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 [**?**].

As funções de qualidade utilizadas foram definidas conforme exemplificado no código da Listing 2, enquanto a criação das tarefas de busca foi realizada segundo o código ilustrado na **??**.

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'])
```

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
```

Adicionalmente, 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.

2.3.2. Função de Avaliação Utilizada

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:

$$score = \alpha \times instances_{subgroup} \times (mean_{sg} - mean_{dataset}) \tag{1}$$

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.

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*.

3.2. Análise de Padrões

Ao analisar todos os experimentos, um perfil de cliente se destacou de forma consistente como o que mais utiliza o cartão: aquele que compra com alta frequência. Portanto, a característica preditiva com maior capacidade discriminativa em relação a um elevado número de transações (`PURCHASES_TRX`) foi, inequivocamente, a frequência de compras (`PURCHASES_FREQUENCY`). Clientes com `PURCHASES_FREQUENCY` ≥ 1, ou seja, que realizam compras de forma consistente, apresentam uma média de 69 transações — um salto expressivo em relação à média global de 15 transações. Embora esse resultado seja esperado e relativamente evidente do ponto de vista interpretativo, ele reforça a validade da abordagem adotada, demonstrando que o algoritmo é capaz de identificar relações coerentes e alinhadas com o domínio do problema.

3.3. 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.

3.3.1. 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.

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.

3.3.2. 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.

Regra principal:

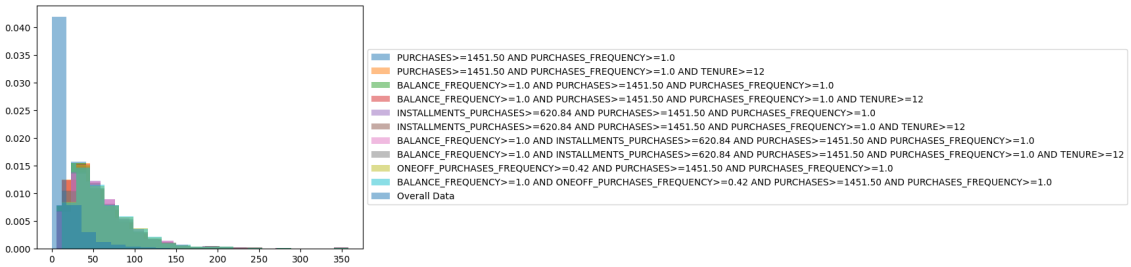

Figure 1. Distribuição de `PURCHASES_TRX` no subgrupo definido por `PURCHASES_FREQUENCY` $\geq$ 1. O experimento foi realizado com `depth` = 8, α = 0.5, e competição livre entre todos os seletores, resultando em subgrupos fortemente caracterizados pelo comportamento de compra.

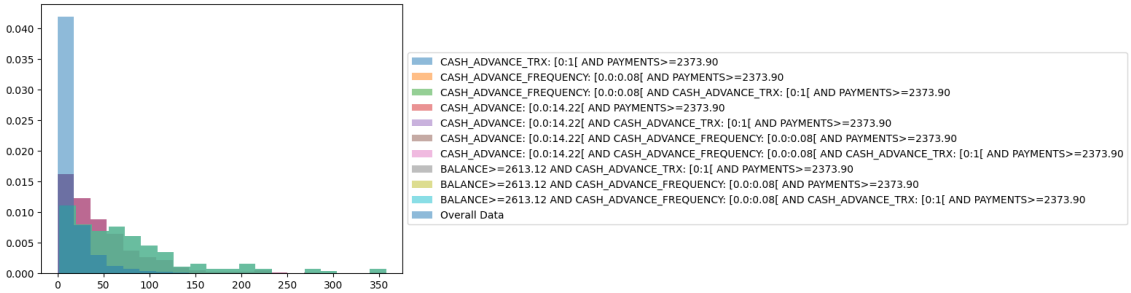

Figure 2. Distribuição de `PURCHASES_TRX` no subgrupo definido por múltiplos descritores financeiros, com $\alpha = 0{,}3$, profundidade máxima igual a 8 e competição restrita a seletores do domínio financeiro. Observa-se que as descrições financeiras apresentam representatividade reduzida na composição do subgrupo.

Figure 3. Distribuição de `PURCHASES_TRX` no subgrupo com múltiplos descritores e $\alpha = 0{,}3$

- `BALANCE_FREQUENCY` ≥ 1.0
  `INSTALLMENTS_PURCHASES` ≥ 620.84
  `ONEOFF_PURCHASES` ≥ 834.8
  `ONEOFF_PURCHASE_FREQUENCY` ≥ 0.42
  `PAYMENTS` ≥ 2373.9
  - 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.

top-10 • depth = 8 • a = 0.3 • competição entre todos os seletores Subgrupos mais deslocados. Aqui aparecem mais variáveis relacionadas à finanças.

top-10 • depth = 8 • a = 0.3 • apenas seletores de finanças Diminuir o “a” nesse caso retorna subgrupos que chegam a cobrir apenas um usuário.

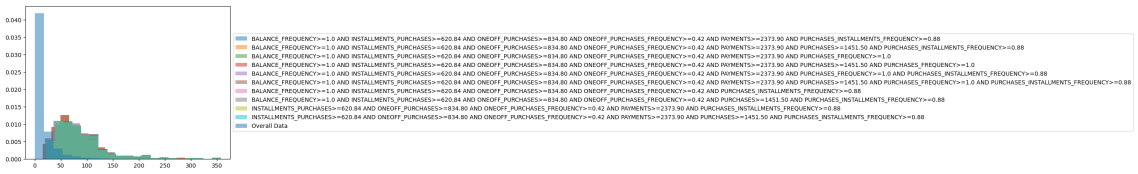

Figure 4. Distribuição de `PURCHASES_TRX` no subgrupo com múltiplos descritores e $\alpha = 0{,}3$

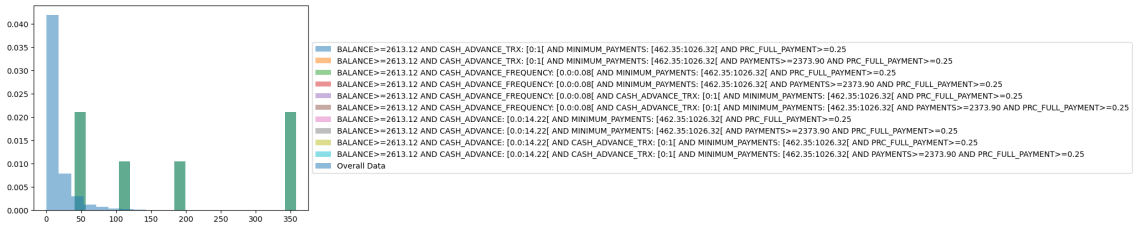

Figure 5. Distribuição de `PURCHASES_TRX` no subgrupo com múltiplos descritores e $\alpha = 0{,}3$

3.3.3. 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).

A escolha de valores para o expoente α próximos de 0,3 ou inferiores tende a degradar a qualidade geral dos subgrupos retornados, mas possibilita a exploração de relações entre variáveis menos fortemente correlacionadas com a variável-alvo. Isso pode ser útil para a indução de padrões menos óbvios. Embora padrões evidentes nem sempre sejam interessantes por si só, os descritores e as estatísticas associadas aos subgrupos ajudam a quantificar sua obviedade, o que pode ter utilidade prática em determinados contextos. Paralelamente, regras excessivamente grandes podem ser difíceis de interpretar e, em alguns casos, até redundantes. Para explorar de forma mais eficaz o espaço de seletores, pode ser vantajoso segmentá-los em diferentes categorias de análise e executar tarefas de descoberta de subgrupos separadamente. Portanto, não existe um valor universalmente ótimo para α — a escolha depende diretamente do objetivo final. Para uma visão geral dos principais perfis, $\alpha = 0{,}5$ se mostra uma combinação adequada. Já para identificar clientes *premium* ou com comportamento mais extremo, $\alpha = 0{,}3$ tende a ser mais eficaz.

3.4. 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.

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

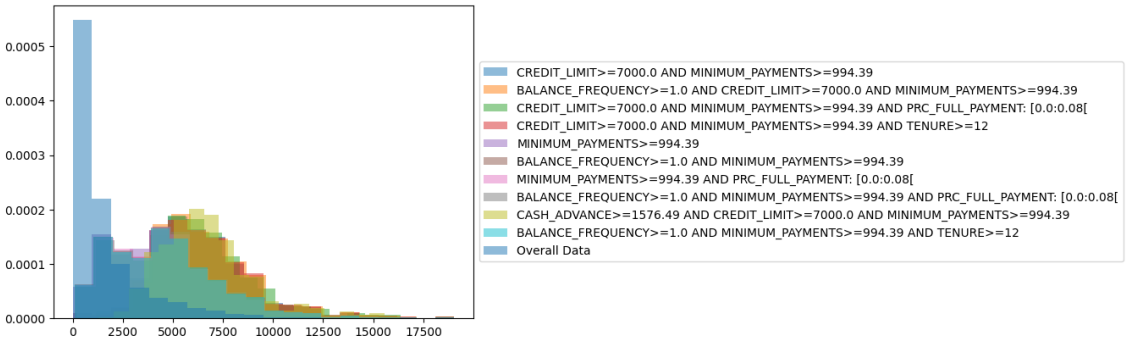

Figure 6. Distribuição do saldo devedor (`BALANCE`) para os subgrupos identificados com maior nível de endividamento, com destaque para os descritores `MINIMUM_PAYMENTS`, `CREDIT_LIMIT`, `PRC_FULL_PAYMENT` e `CASH_ADVANCE`.

- **Pagamento Mínimo Elevado**: A condição $\texttt{MINIMUM\_PAYMENTS} \geq 994.39$ 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 $\texttt{CREDIT\_LIMIT} \geq 7000.0$ 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 $\texttt{PRC\_FULL\_PAYMENT} \in [0.0, 0.08[$ 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, por consequência, aumentando seu saldo devedor.
- **Frequência de Atualização do Saldo**: A condição $\texttt{BALANCE\_FREQUENCY} \geq 1.0$ 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.

3.4.1. O Saldo Mais Elevado

O subgrupo com maior média de saldo devedor (R$ 6.715,71) é caracterizado pela combinação das seguintes condições:

- $\texttt{CREDIT\_LIMIT} \geq 7000.0$
- $\texttt{MINIMUM\_PAYMENTS} \geq 994.39$
- $\texttt{CASH\_ADVANCE} \geq 1576.49$

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.

3.4.2. Clientes Antigos e Ativos

Outro perfil recorrente nos subgrupos de maior BALANCE é o de clientes com:

- TENURE ≥ 12
- BALANCE_FREQUENCY ≥ 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. 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, buscando responder perguntas críticas de negócio de forma clara e direta. 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.