

Por:

Kensley Alves de Oliveira – 32411BSI037

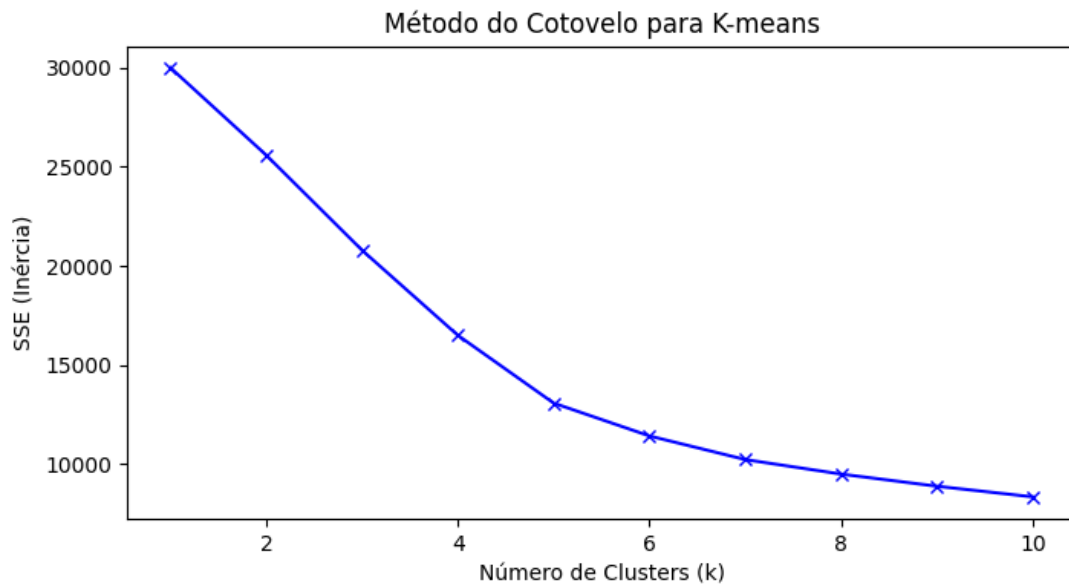
Pedro Paulo Cunha

## 1. Tendência de Agrupamento

A clusterização não tem a meta de prever um valor, mas sim descobrir "grupos naturais" ou segmentos onde os objetos dentro de um grupo sejam muito similares entre si (alta coesão) e muito diferentes dos objetos de outros grupos (alta separação). Com este conceito em mente, um resultado da estatística de Hopkins de 0.9828 é excelente. Como valores acima de 0.5 indicam que os dados não são aleatórios, este índice quase próximo a 1.0 confirma que existem estruturas e grupos naturais muito fortes nos seus dados de crédito.

## 2. Análise do K-Means

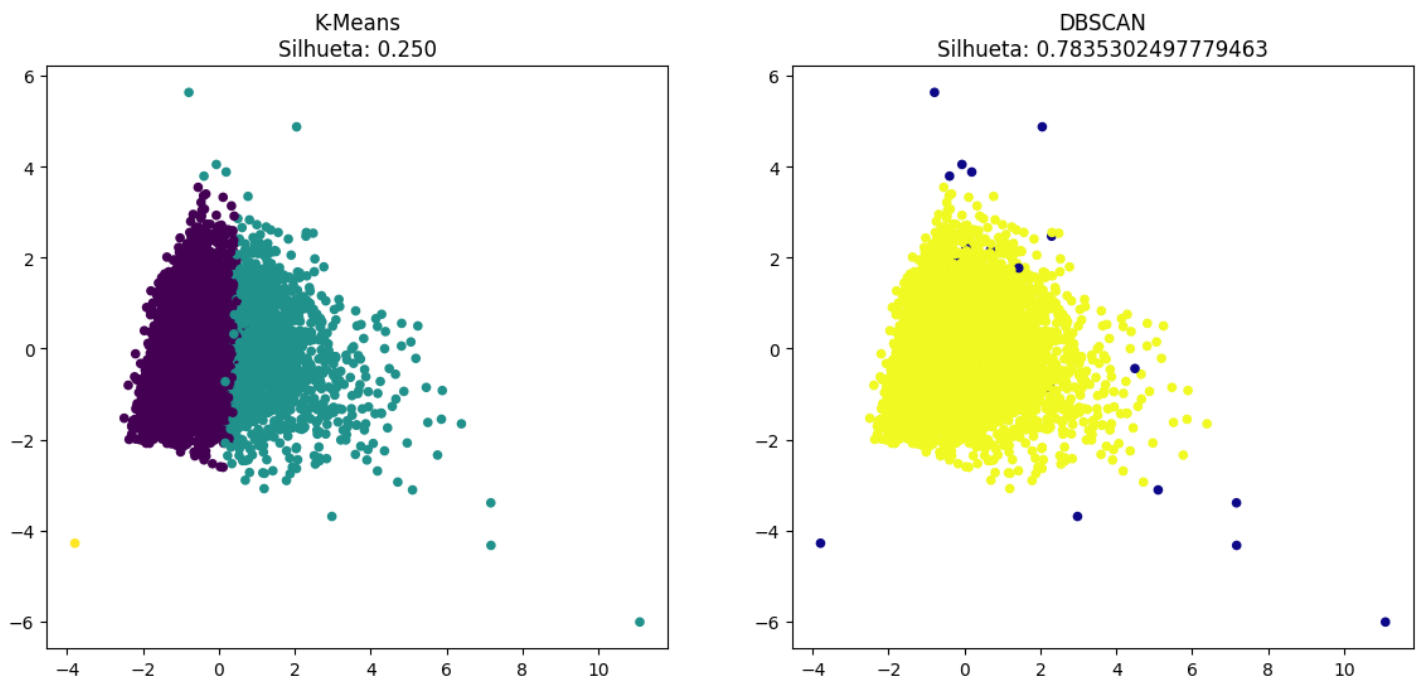
O K-Means tentou forçar uma partição esférica nos dados. A figura 01 mostra que no método do cotovelo há uma queda suave na inércia (SSE), com uma leve quebra em torno de  $k=3$  ou  $k=4$ .



**Figura 01:** Gráfico método cotovelo para estimativa do K.

## RELATÓRIO DE CLUSTERIZAÇÃO

Quanto ao desempenho, o coeficiente de Silhueta de 0,250 é considerado baixo. Isso indica que, embora o algoritmo tenha dividido os dados, há uma sobreposição significativa entre os grupos, e os pontos não estão tão próximos de seus centros quanto deveriam. Na figura 02, observa-se o gráfico de dispersão, as divisões verticais artificiais, típicas do K-Means quando tenta separar nuvens de dados densas e alongadas.



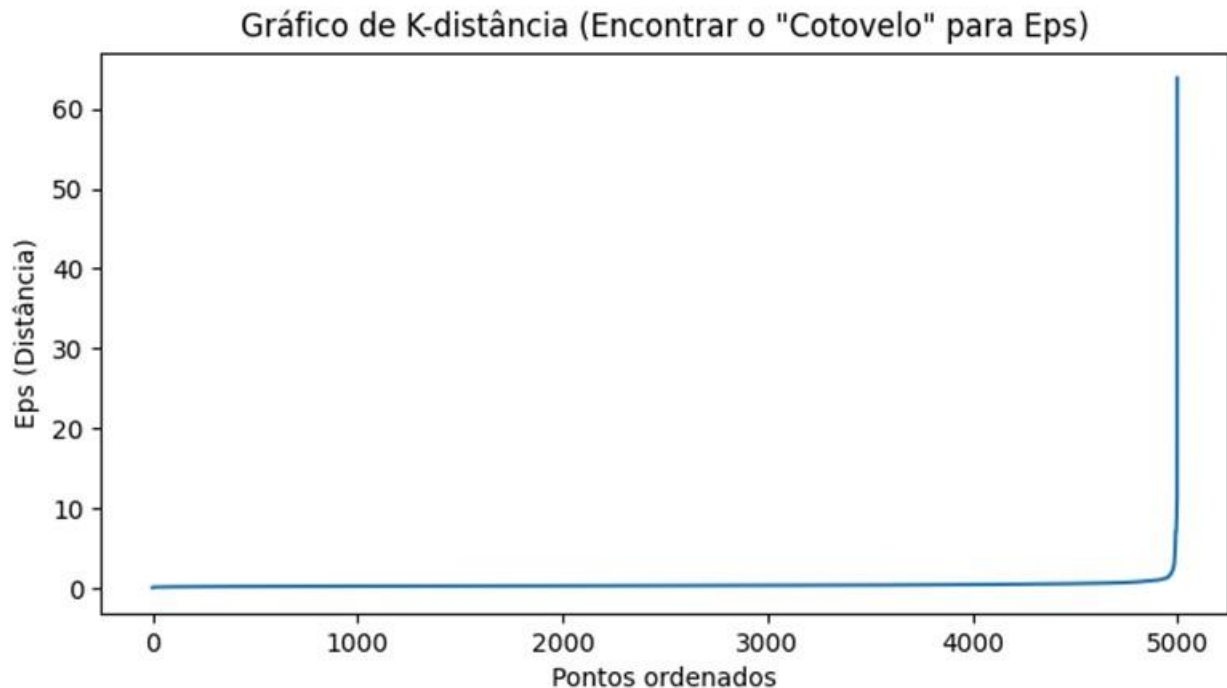
**Figura 02:** Gráfico da dispersão dos dados pelos algoritmos K-Means e DBSCAN

### 3. Análise do DBSCAN

O DBSCAN apresentou um comportamento muito mais alinhado à realidade dos dados financeiros. O Gráfico de K-distância (Figura 03) mostra um "cotovelo" muito nítido próximo ao valor de distância 1.5, validando a escolha do parâmetro Eps.

O desempenho com o algoritmo DBSCAN apresentou um coeficiente de Silhueta (Figura 02) de 0,7835. Este é um valor alto e indica que os clusters formados são extremamente coesos e bem separados entre si.

Uma observação importante no gráfico, quanto ao tratamento de Outliers, é que diferente do K-Means, o DBSCAN identificou pontos escuros isolados (ruídos). Em risco de crédito, esses são os perfis atípicos (ex: rendas astronômicas ou dívidas fora do padrão) que não devem ser misturados com a massa de clientes comum.

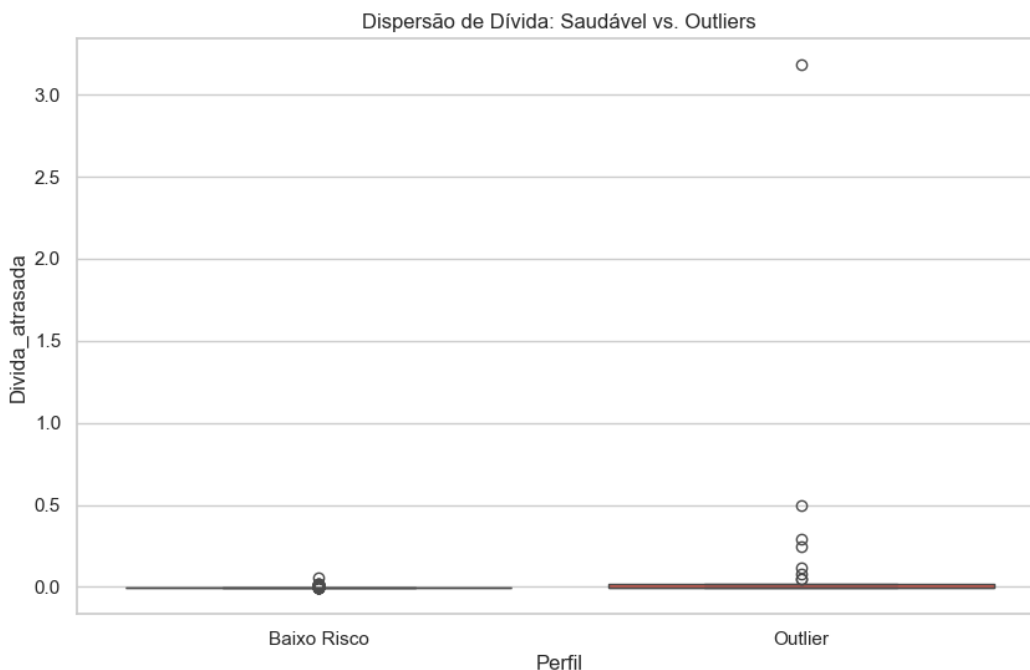


**Figura 03:** Gráfico de K-distância

### 3. Análise de Risco e Outliers

- **Dispersão de Dívida:** O gráfico Boxplot (Figura 04) revela que o grupo de Baixo Risco possui uma dívida atrasada concentrada em zero, com quase nenhuma variabilidade. Já o grupo Outlier apresenta uma dispersão drástica, com valores atingindo patamares extremos (acima de 3.0 na escala normalizada).
- **Propensão à Rejeição:** O fato de os outliers serem "infinitamente mais propensos a rejeição" confirma que o DBSCAN isolou exatamente os clientes que o banco já vinha recusando internamente ou que possuem histórico crítico no Serasa.

## RELATÓRIO DE CLUSTERIZAÇÃO



**Figura 04:** Gráfico Boxplot da dispersão da dívida entre bons pagadores (Saudáveis) e Outliers

## 4. Conclusão

A clusterização validou a hipótese de que o risco de crédito está intrinsecamente ligado ao ciclo de vida do cliente e seu histórico de relacionamento. Conforme Tan et al. (2009), grupos com alta coesão e separação (como demonstrado pelo Hopkins de 0.98) permitem a criação de políticas de crédito personalizadas.

Para o negócio, isso justifica a aplicação de um Scorecard Diferenciado:

- ✓ Para o Cluster 0, o foco deve ser retenção e *upselling* (aumento de limites).
- ✓ Para o Cluster 1, o foco deve ser monitoramento rigoroso e limites iniciais reduzidos, dada a propensão de risco 1.9x maior que a do grupo estável.

A clara distinção dos centróides corrobora o uso de modelos lineares e árvores de decisão, uma vez que a fronteira de separabilidade entre os perfis de baixo e alto risco é definida pela maturidade do proponente (GÉRON, 2021).

### REFERÊNCIAS BIBLIOGRÁFICAS

- FAWCETT, Tom; PROVOST, Foster. **Data Science para Negócios**. Rio de Janeiro: Alta Books, 2016.
- GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow**. 2. ed. Rio de Janeiro: Alta Books, 2021.
- McKINNEY, Wes. **Python para Análise de Dados**. Rio de Janeiro: Novatec, 2018.
- TAN, Pang-Ning et al. **Introdução ao Data Mining Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna, 2009.