**TECHNICAL REPORT**

|  |
| --- |
| Aluno: José Davi Araujo Gomes |

1. **Introdução**

**Dataset Stroke Prediction**

***Descrição:*** De acordo com a Organização Mundial da Saúde (OMS), o AVC é a segunda principal causa de morte em todo o mundo, responsável por aproximadamente 11% do total de mortes.   
Este conjunto de dados é usado para prever se um paciente é susceptível de obter acidente vascular cerebral com base nos parâmetros de entrada como sexo, idade, várias doenças e tabagismo. Cada linha dos dados fornece informações relevantes sobre o paciente.

|  |  |
| --- | --- |
| **Tópico** | **Descrição** |
| id | identificador único; |
| gender | "Masculino", "Feminino" ou "Outros"; |
| age | idade do paciente; |
| hypertension | 0 se o paciente não tiver hipertensão, 1 se o paciente tiver hipertensão; |
| heart\_disease | 0 se o paciente não tem nenhuma doença cardíaca, 1 se o paciente tiver uma doença cardíaca; |
| ever\_married | "Não" ou "Sim"; |
| work\_type | "crianças", "Govt\_jov", "Never\_worked", "Private" ou "Auto-empregados"; |
| Residence\_type | "Rural" ou "Urban"; |
| avg\_glicose\_nível | nível médio de glicose no sangue; |
| bmi | índice de massa corporal; |
| smoking\_status | "anteriormente fumado", "nunca fumou", "smokes" ou "Unknown"; |
| stroke(AVC) | 1 se o paciente teve um acidente vascular cerebral ou 0 se não. |

Nota: "*Unknown*" em “*smoking\_status”* significa que a informação não está disponível para este paciente.

Os dados identificadores das colunas do dataset estão definidos da seguinte forma:

|  |  |  |
| --- | --- | --- |
| **Dados Categóricos** | **Dados Contínuos** | **Dados Binários** |
| gender | age | hypertension |
| ever\_married | avg\_glucose\_level | heart\_disease |
| work\_type | bmi | stroke |
| Residence\_type | **-** | **-** |
| smoking\_status | **-** | **-** |

1. **Observações**

Na questão 3 tive problemas inicialmente para gerar o scatteplot pois os centroides estavam se acumulando em um único ponto devido as colunas alvo do meu dataset serem compostas por dados em 0 e 1.

1. **Resultados e discussão**

***Questão 1***

Começarei carregando e analisando o dataset para verificar as colunas e o tipo de dados que temos. O dataset contém colunas com dados demográficos e médicos, com a variável-alvo stroke (1 para ocorrência de AVC e 0 para ausência).

O objetivo é determinar o melhor valor de **k** e a métrica de distância mais adequada, garantindo um melhor desempenho do modelo.

O dataset foi carregado e pré-processado conforme descrito a seguir:

* **Remoção de colunas irrelevantes:** A coluna id foi excluída por não contribuir para a classificação.
* **Tratamento de valores ausentes:** A coluna bmi continha valores nulos, que foram substituídos pela média da coluna.
* **Codificação de variáveis categóricas:** As colunas categóricas (gender, ever\_married, work\_type, Residence\_type, smoking\_status) foram convertidas para valores numéricos utilizando LabelEncoder.
* **Normalização dos dados:** Foi utilizada a técnica de **StandardScaler** para normalizar as variáveis, garantindo uma escala uniforme.

Para encontrar a melhor configuração do KNN, utilizei **GridSearchCV** com validação cruzada 5-fold. Os parâmetros testados foram:

* **Número de vizinhos (k):** [1, 3, 5, 7, 9, 11, 13, 15]
* **Métrica de distância:** [‘euclidean’, ‘manhattan’, ‘minkowski’]
* **Peso dos vizinhos:** [‘uniform’, ‘distance’]

Os melhores parâmetros encontrados foram:

|  |  |
| --- | --- |
| **Melhores Parâmetros Encontrados** | |
| Melhor Métrica | Manhattan |
| Melhor n° vizinho | 'n\_neighbors': 5 |
| Melhor Acurácia Média | 0.954257181590469 |
| Acurácia no conjunto de teste | 0.9403131115459883 |

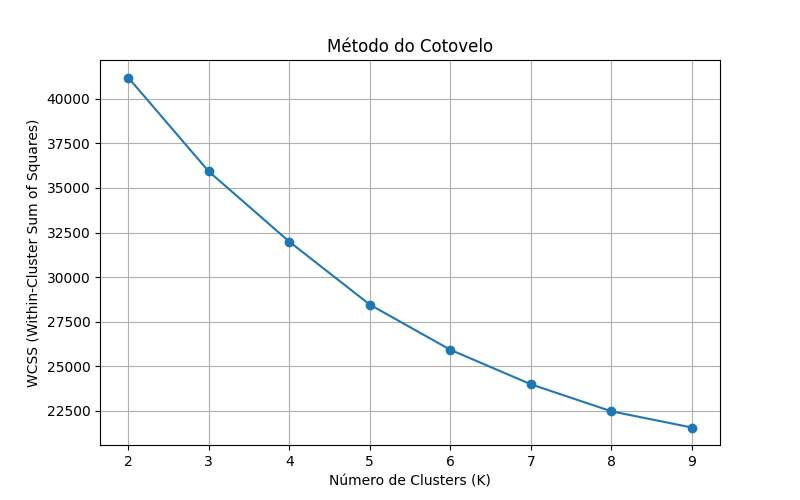
Comparando os resultados com a AV1, nessa tive um resultado diferente utilizando o **GridSearchCV** visto que lá a melhor métrica foi a Euclidiana com um número de vizinhos igual a 5.

***Questão 2***

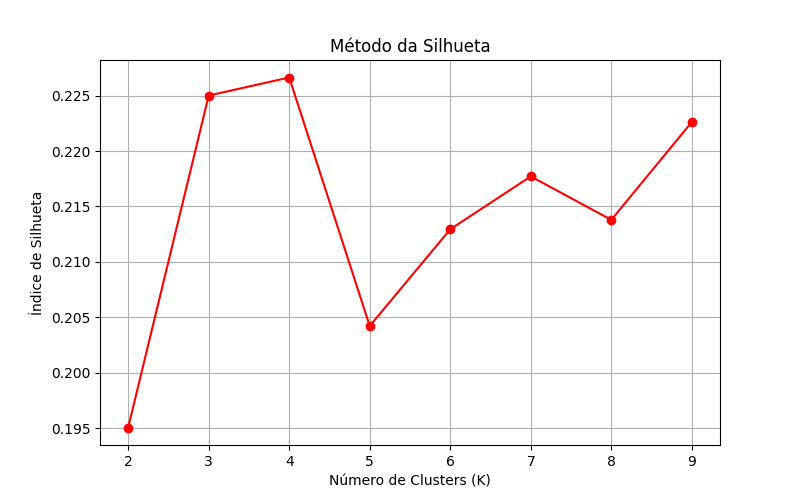
Nessa questão visa determinar a quantidade ideal de clusters para o dataset de AVC utilizando os métodos do **Cotovelo** e da **Silhueta**. Esses métodos são amplamente utilizados para definir o número ótimo de clusters em algoritmos de agrupamento como o **K-Means**.

Primeiramente inicializei o **K-Means** com o range de 2 a 10, os resultados obtidos foram:

* **Método do Cotovelo:** Indicou um valor ótimo de **k=4**.
* **Índice de Silhueta:** Sugeriu um valor de **k=3** como mais adequado.



Cotovelo



Silhueta

A diferença nos valores de k entre os dois métodos é comum, pois cada técnica foca em um aspecto diferente do agrupamento. O Cotovelo prioriza a minimização da inércia, enquanto a Silhueta avalia a qualidade da separação entre clusters.

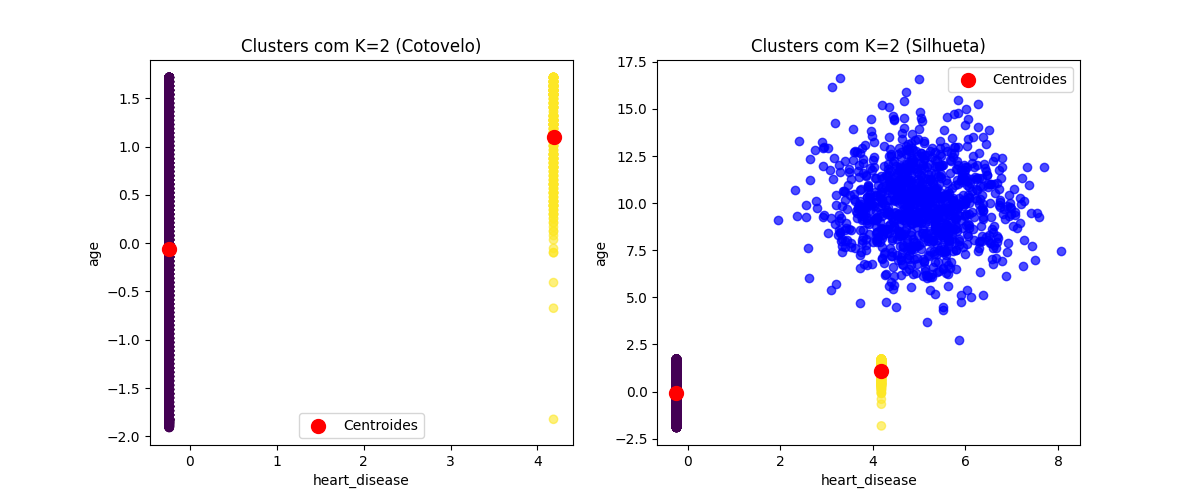
***Questão 3***

Para continuar a análise de clusterização realizada anteriormente, utilizei o método de regressão Lasso para selecionar os dois atributos mais relevantes do conjunto de dados. Com esses atributos, refiz a definição da quantidade ideal de clusters utilizando os métodos do cotovelo e da silhueta. Meu objetivo era verificar se a quantidade de clusters mudou e analisar a distribuição dos dados por meio de visualizações gráficas mais informativas.

Com os atributos selecionados, utilizei novamente os métodos do cotovelo e do índice de silhueta para encontrar a quantidade ideal de clusters. Meu objetivo era verificar se essa quantidade mudaria em relação à análise inicial, feita com todas as variáveis.

|  |  |
| --- | --- |
| **Atributos mais relevantes segundo Lasso** | |
| 'heart\_disease' | 'age' |

Gerei scatterplots para visualizar a distribuição dos clusters usando os dois atributos mais relevantes.



Após aplicar a seleção de atributos, percebi que o número ideal de clusters permaneceu o mesmo segundo o índice de silhueta, mas o método do cotovelo sugeriu um valor diferente. Isso demonstra que a quantidade de informação disponível nos dados influencia diretamente a segmentação.

A visualização dos clusters com apenas dois atributos tornou a separação dos grupos mais clara. Pude observar que certos clusters estavam bem definidos, enquanto outros apresentavam uma sobreposição maior, o que indica que outros fatores podem estar contribuindo para a classificação.

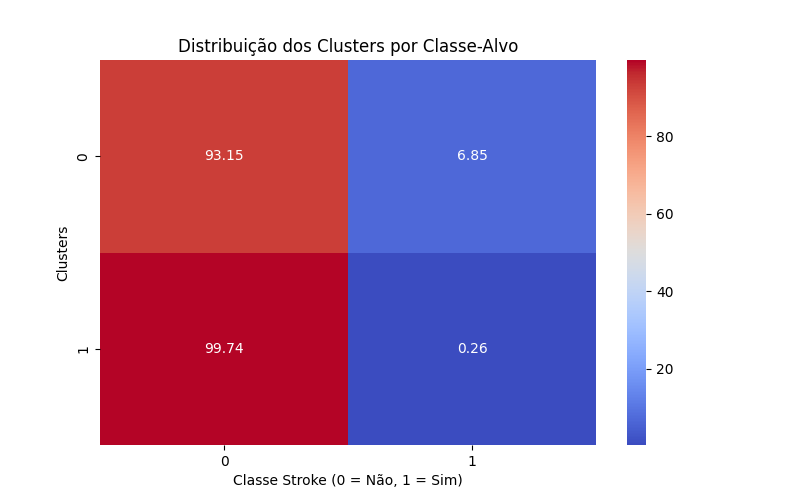
***Questão 4***

Nesta etapa da análise, utilizei o método de clusterização **K-Means**, com a quantidade de clusters definida pelo **índice de silhueta**, para examinar como os grupos foram distribuídos em relação às classes da variável alvo. Para isso, elaborei uma tabela **crosstab**, permitindo uma melhor compreensão da relação entre clusters e categorias da coluna alvo.

Os resultados da crosstab revelaram padrões interessantes na distribuição dos clusters. Percebi que:

* Um dos clusters continha uma concentração significativamente maior de pacientes que sofreram AVC.
* Os outros dois clusters pareciam agrupar pacientes sem AVC, mas com diferenças nos fatores de risco, como hipertensão e diabetes.

|  |  |  |
| --- | --- | --- |
| Distribuição de Clusters por Classe-Alvo (Stroke) | | |
| Stroke | 0 (Não) | 1 (Sim) |
| 0 | 93.15 | 6.85 |
| 1 | 99.74 | 0.26 |
|  |  |  |
| Legenda: "Stroke" = Amostragem de pessoas que sofreram AVC | | |



1. **Conclusões**

Os resultados obtidos foram, em grande parte, satisfatórios, pois consegui identificar padrões relevantes na clusterização e entender como a seleção de atributos impacta na definição dos grupos. No entanto, notei que houve divergência entre os métodos do cotovelo e do índice de silhueta na determinação do número ideal de clusters. Isso indica que os dados podem ter características que dificultam uma separação clara, ou que outros fatores, como a normalização e a escolha de atributos, podem estar influenciando a segmentação. Além disso, a visualização dos clusters mostrou que alguns grupos ficaram bem definidos, enquanto outros apresentaram sobreposição, o que sugere a necessidade de refinar a abordagem.

1. **Próximos passos**

Seria interessante explorar outras técnicas de seleção de atributos, para comparar os resultados com os obtidos pelo Lasso. Além disso, poderia experimentar algoritmos de clusterização diferentes, como o DBSCAN para avaliar se eles oferecem uma segmentação mais clara dos dados.