Trabalho 2 da disciplina de Aprendizado de máquina - K means

Giovani Da Silva[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)

{giovani.silva}@inf.ufrgs.br

1. Objetivo

O objetivo do trabalho consiste em utilizar o algoritmo de agrupamento K-means para identificar padrões em um conjunto de dados e, consequentemente, classificar o risco de crédito para clientes bancários. O processo teve início com a análise exploratória dos dados, a fim de compreender as suas características e distribuição. Em seguida, foi aplicado o algoritmo K-means, método de agrupamento não supervisionado, para encontrar grupos de clientes que apresentem características similares em relação ao risco de crédito.

2. Informações sobre a implementação

O trabalho foi desenvolvido em um Jupyter Notebook disponibilizado no Google Colab e contou com o uso de diversas bibliotecas, tais como Pandas, Numpy e Scikit-learn.

3. Atributos e pré-processamento

Os atributos contidos no dataset são os seguintes:

- **Age**: idade do cliente (numérico).
- **Credit amount**: valor do crédito solicitado (numérico).
- **Duration**: duração do crédito em meses (numérico).
- **Sex**: gênero do cliente (categórico: masculino, feminino).
- **Job**: nível de emprego do cliente (categórico: 0 - não qualificado e não residente, 1 - não qualificado e residente, 2 - qualificado, 3 - altamente qualificado).
- **Housing**: tipo de imóvel em que o cliente reside (categórico: próprio, alugado ou gratuito).
- **Saving Account**: nível de poupança do cliente (categórico: pequena, moderada, bastante rico, rico).
- **Checking Account**: nível de saldo da conta corrente do cliente (categórico: pequena, moderada, bastante rico, rico).
- **Purpose**: finalidade do crédito solicitado pelo cliente (categórico: carro, móveis/equipamentos, rádio/TV, eletrodomésticos, reparos, educação, negócios, férias/outros).

Percebe-se que há tanto atributos categóricos como numéricos, A análise de agrupamento será feita através dos atributos numéricos: Age, Credit Amount, Duration.

4. K means

O K-means é um algoritmo de aprendizado de máquina não supervisionado utilizado para análise de agrupamentos. O objetivo do algoritmo é dividir um conjunto de dados em k grupos diferentes, onde cada grupo representa uma classe específica. O processo é realizado através de iterações, onde cada ponto de dado é atribuído ao grupo mais próximo do centróide, que é um ponto médio de cada grupo. Em seguida, o centróide é recalculado para representar o novo grupo e o processo é repetido até que a convergência seja alcançada. O K-means é amplamente utilizado em tarefas de mineração de dados, classificação de imagens e processamento de linguagem natural.

Para testar o K means foi utilizada a biblioteca do sci kit learn, o número ótimo de K's(agrupamentos) foi feito utilizando dois métodos: Método do cotovelo e o método da silhueta.

4.1. Método do cotovelo

O método do cotovelo é um dos métodos utilziados para se buscar um valor ideal de K(agrupamentos) no K means. Sua ideia é avaliar o gráfico da soma dos erros quadráticos em relação ao número de agrupmanetos e identificar quando a adição de cluster começa a interferir menos, em outras palavras, tem um ganho menor, em relação ao redução do erro.

Esse gráfico é conhecido como cotovelo pois de certa forma lembra o cotovelo humano, onde o ponto de inflexão seria o representante do ponto de k ideal.

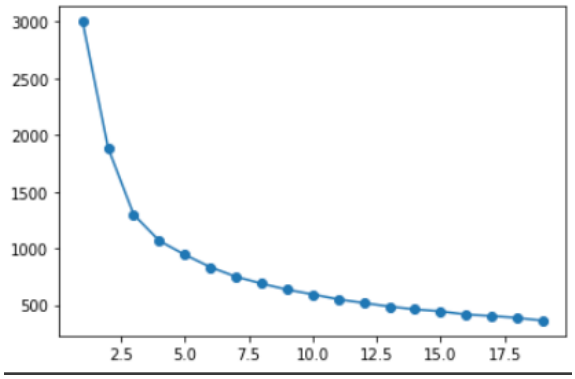


Figura 1. Plot do gráfico do cotovelo para esse dataset

Percebe-se pelo gráfico que o ponto ideal para K seria entre 3 e 5, isso será discutido futuramente (**??**).

4.2. Método da silhueta

O método da silhueta é outra técnica utilizada para a valiar a qualidade dos agrupamentos gerados e descobrir o melhor valor de K.

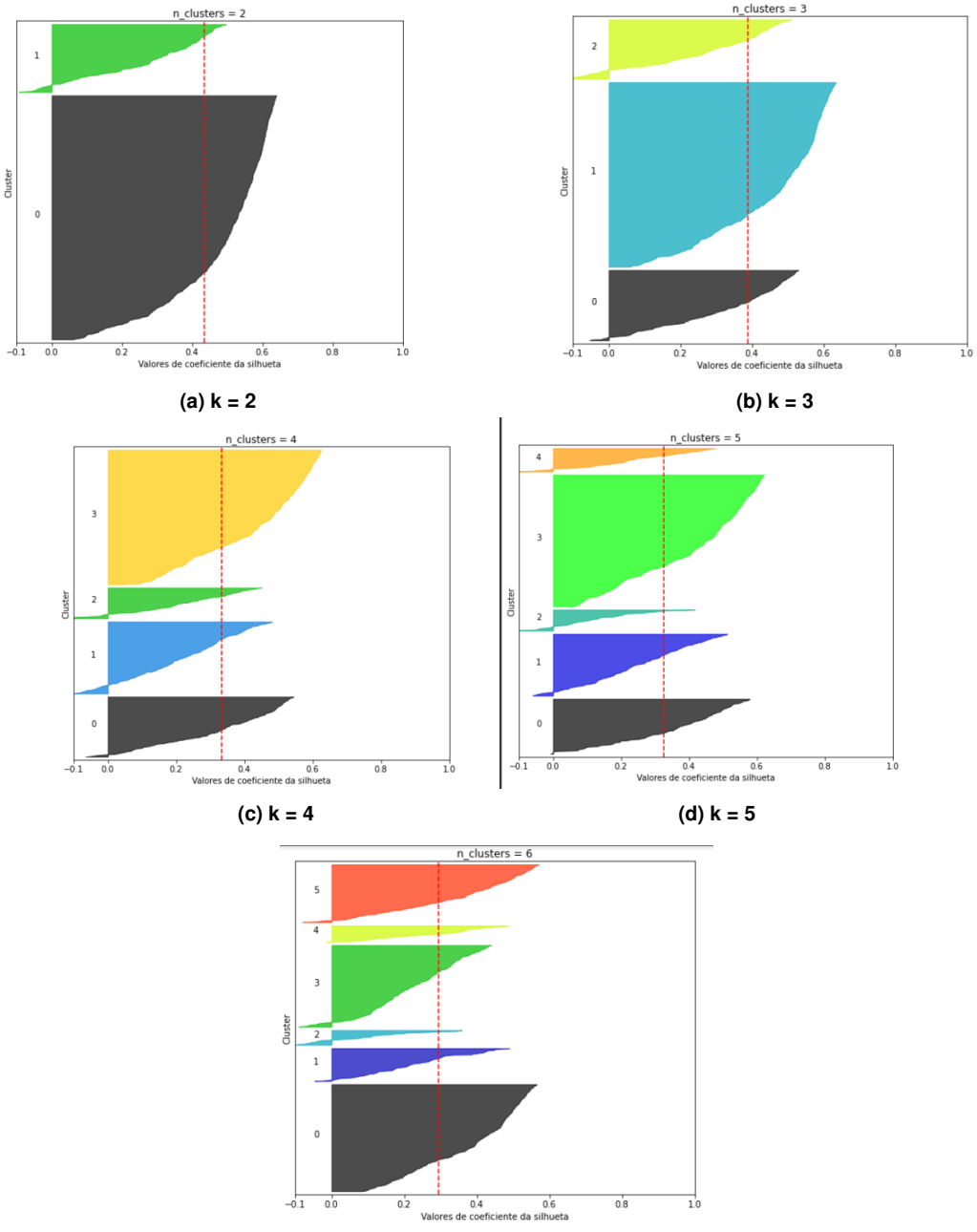
A idéia consiste em calcular quão bem o ponto está associado a aquele cluster K em relação aos outros clusters, esse cálculo nos fornece a silhueta.

A silhueta é um valor que varia de -1 a 1, quanto mais próximo de 1 melhor o ponto está associado ao seu cluster e quanto mais distante, menos associado.

A partir disso, o método da silhueta calcula a média das silhuetas de todos os ponto e determina a qualidade geral dos clusters, um valor de silhueta médio próximo de 1 indica que os clusters estão bem definidos e o contrário ocorre quando esse valor está distante de 1.

Nos melhores agrupamentos, o ideal é que todas as instâncias superem esse valor médio e o clusters sejam menos dispersos o possível.

Tendo em vista os gráficos apresentados a cima, percebmos que os melhores valores de k foram k = 3 e k = 4.



(a) k = 2

(b) k = 3

(c) k = 4

(d) k = 5

(e) k = 6

Figura 2. Método da silhueta

4.3. Valores escolhidos e resultados

Tendo em vista os melhores resultados observados tanto no método do cotovelo quanto no método da shilueta, o valor de k escolhido foi 3.

Para visualizar como estão distribuidos os cluters, vamos usar uma técnica de PCA(Principal Component Analysis) que faz uma combinação linear dos atributos originais e separa os pontos em pontos por seus respectivos clusters.Cada ponto será representado pelas coordenadas PC1, PC2 (onde PC = Principal Component) e a cor do ponto no gráfico corresponde ao seu respectivo cluster.

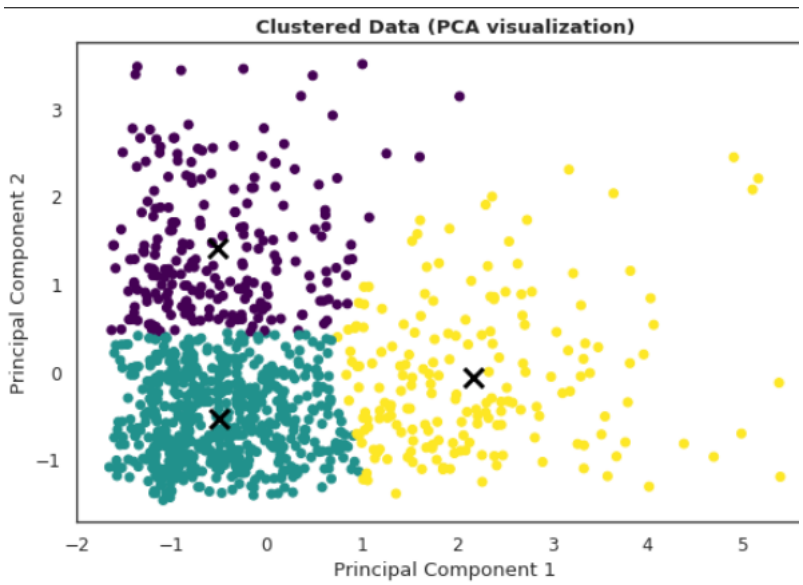
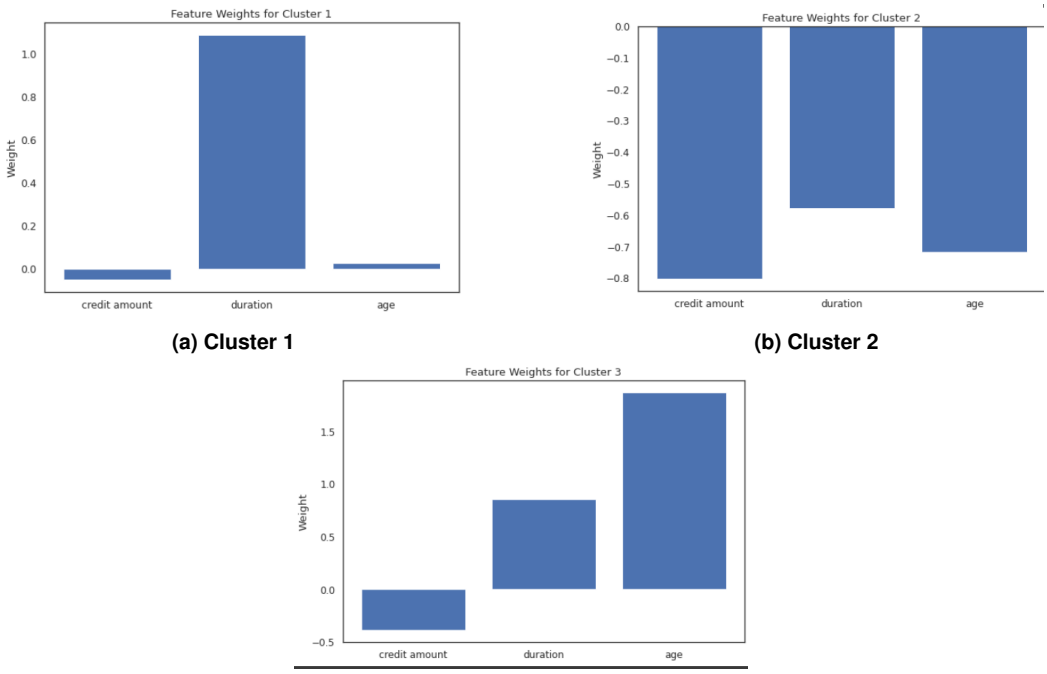


Figura 3. Resultado do gráfico gerado pela técnica de PCA

Como podemos observar, com k = 3 os agrupamentos conseguiram ser bem divididos, com centros bem definidos e pouca ou quase nenhuma intersecção entre eles.

A partir disso, vamos visualizar os resultados para cada cluster.

Começando pelo peso no qual cada atributoi numérico influencia em cada cluster:



(a) Cluster 1

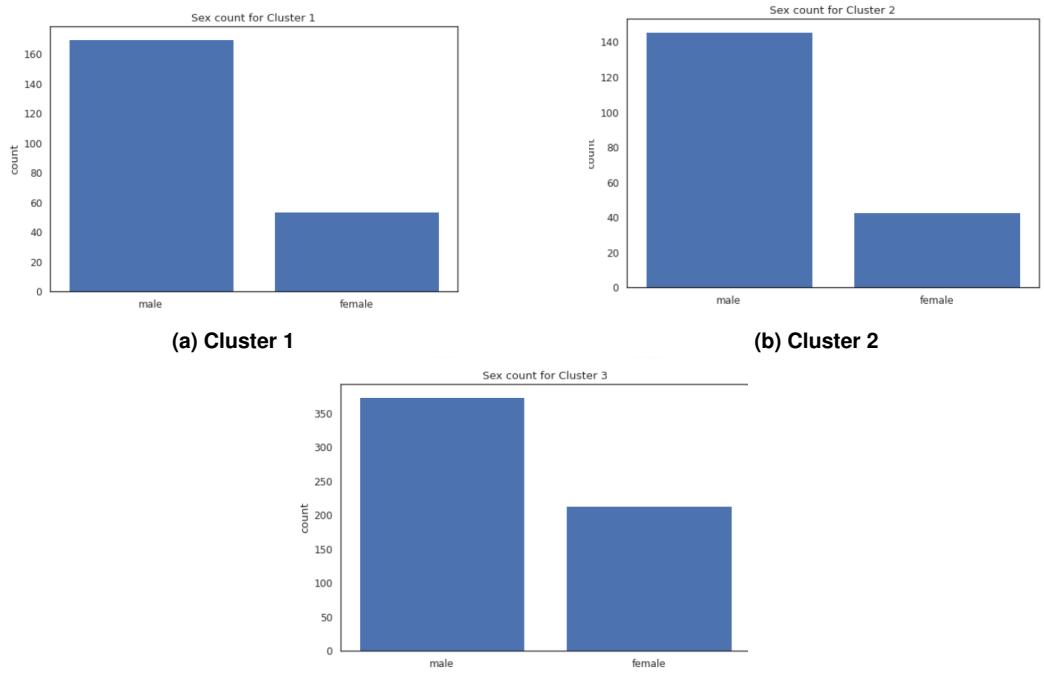
(b) Cluster 2

(c) Cluster 3

Figura 4. Pesos dos atributos numéricos em cada cluster

Percebe-se que o cluster 1 foi agrupado principalmente em função do atributo numérico duration, o cluster 2 por sua vez estava bem distribuido e o cluster 3 se agrupou tanto em função de ”duration”quanto de ”age”.

Vamos observar agora quanto ao atributo sex.



(a) Cluster 1

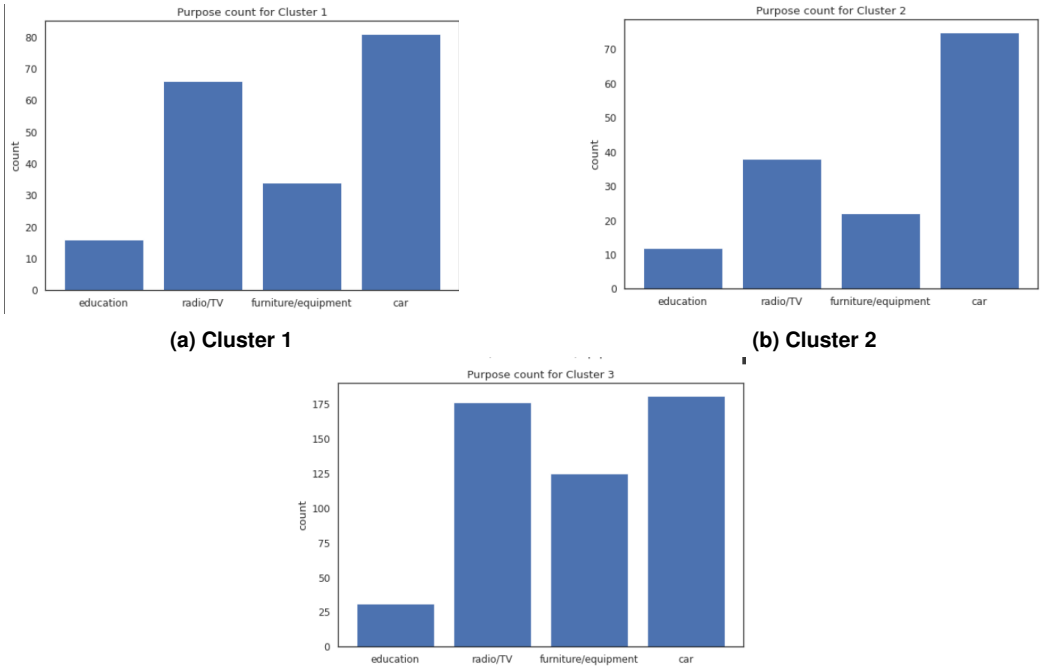
(b) Cluster 2

(c) Cluster 3

Figura 5. Contagem do atributo sexo em cada cluster

Quanto ao atributo sexo, percebemos que a sua separação foi similar nos clusters 1 e 2, com maior quantidade de homens do que mulheres, por outro lado, no cluster 3 temos uma proporção maior de mulheres.

Em seguida, vamos avaliar em relação ao atributo purpose(propósito):



(a) Cluster 1

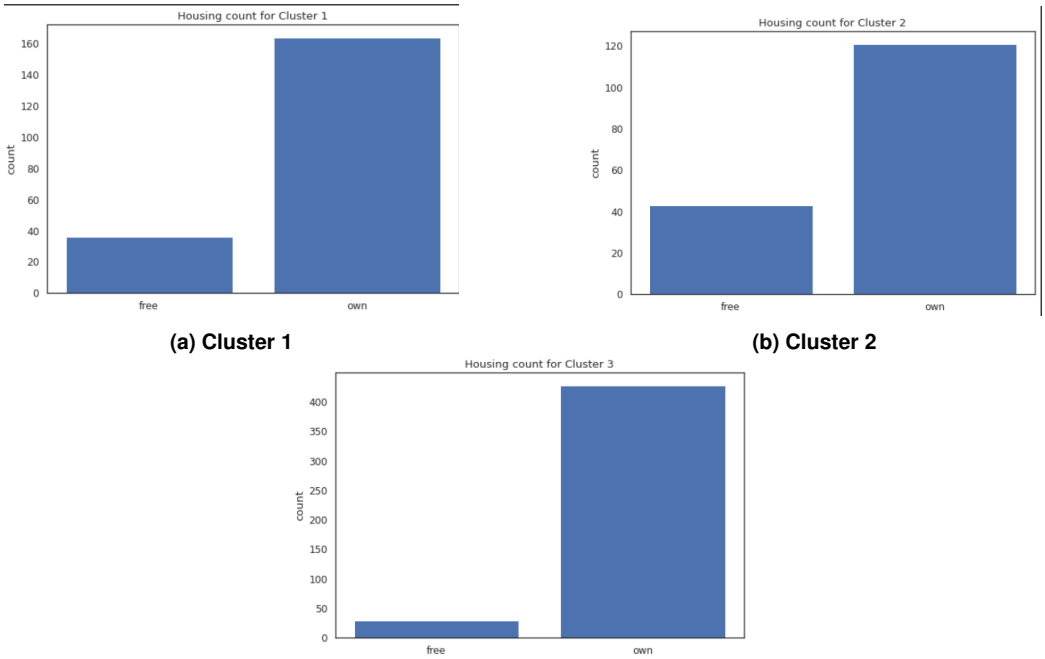
(b) Cluster 2

(c) Cluster 3

Figura 6. Contagem do atributo purpose em cada cluster

É visto que todos os clusters tem uma grande quantidade de carros como valor do atributo, enquanto no cluster 1 existe uma proporção similar entre radio/tv e furniture, no cluster 2 há uma proporção muito maior de radio/tv, no cluster 3 estão todos bem distribuidos, com um foco maior em furniture e radio/tv.

Vamos avaliar agora com relação ao atributo housing.



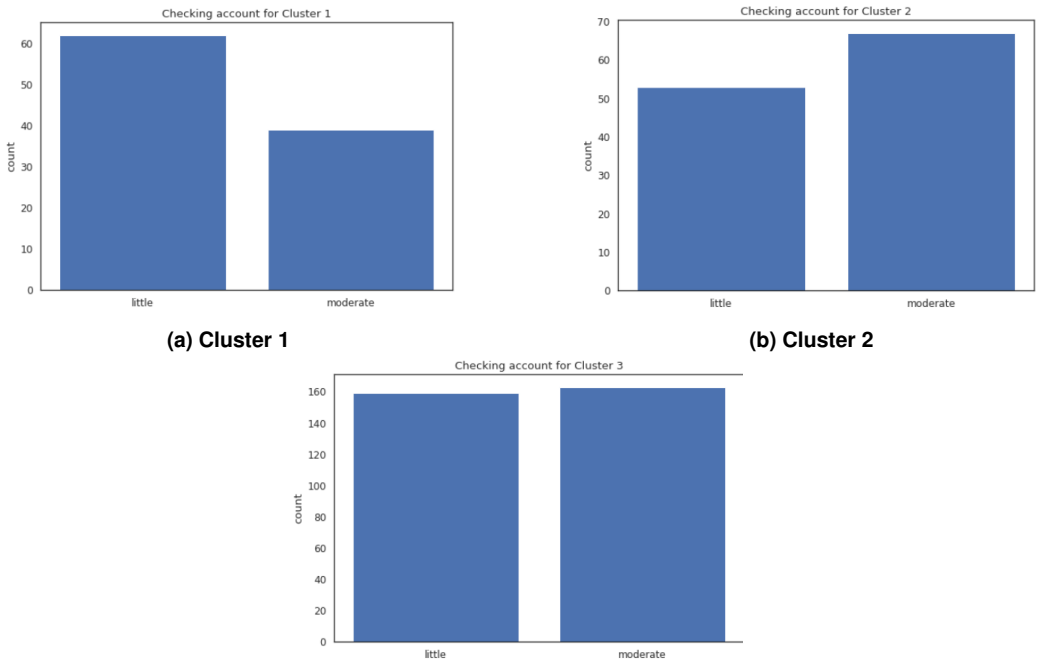
(a) Cluster 1 **(b) Cluster 2**

(c) Cluster 3

Figura 7. Contagem do atributo housing em cada cluster

Quanto ao atributo housing, percebemos que todos os clusters tem uma quantidade muito maior de ”own”do que de ”free”, entretanto, a proporção é muito maior no cluster 3 que no cluster 1 e no cluster 2.

Vamos observar agora o atributo checking account.



(a) Cluster 1

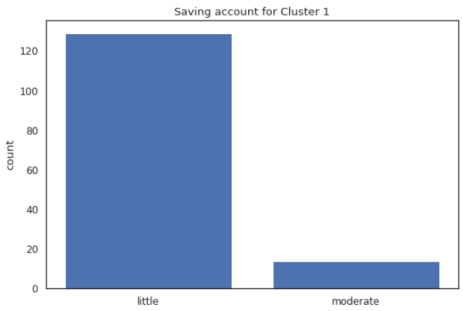
(b) Cluster 2

(c) Cluster 3

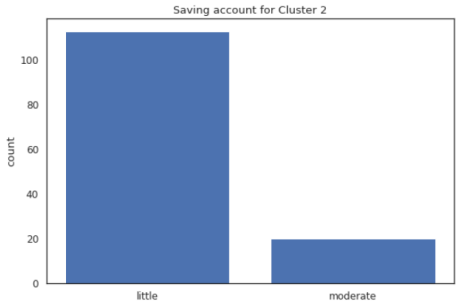
Figura 8. Contagem do atributo checking account em cada cluster

Neste atributo, percebemos que o cluster 1 tem umaq maior quantidade de little do que moderado, isso se inverte no cluster 2 e eles tem a mesma quantidade praticamente no cluster 3.

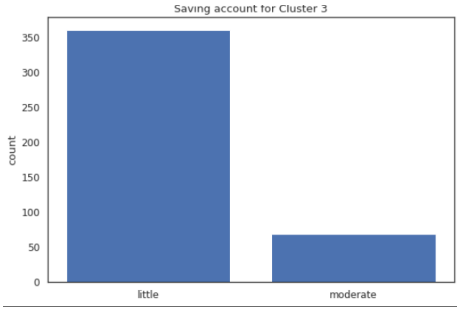
Vamos observar agora o atributo saving account.



(a) Cluster 1



(b) Cluster 2

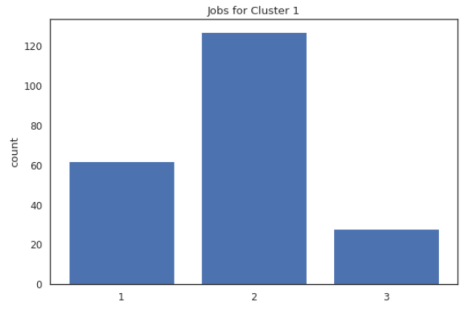


(c) Cluster 3

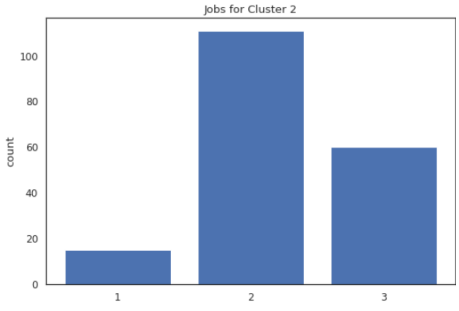
Figura 9. Contagem do atributo saving account em cada cluster

Nesse atributo saving account, percebemos uma diferença de contagem maior no cluster 3, enquanto a proporção é simlar nos clusters 1 e 2.

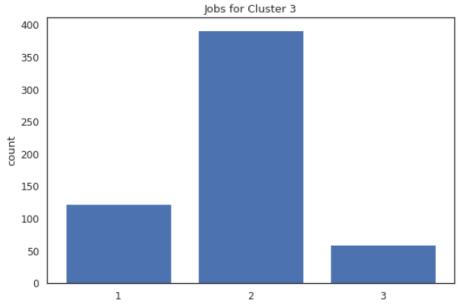
Iremos agora observar a contagem referente ao atributo jobs:



(a) Cluster 1



(b) Cluster 2



(c) Cluster 3

Figura 10. Contagem do atributo jobs em cada cluster

Percebemos que o atributo jobs em todos os clusters tem uma proporção maior de 2, entretanto o cluster 1 contém uma proporção maior de 1 job, enquanto o cluster 2 uma proporção maior de 3 jobs.O cluster 3 tem uma predominância muito forte do atributo 2.

Por fim, vamos observar a média dos atributos numéricos, estes valores estão normalizados:

```
cluster
0    1.416543
1   -0.069428
2   -0.518201
Name: Age, dtype: float64
cluster
0   -0.309353
1    1.537709
2   -0.377056
Name: Credit amount, dtype: float64
cluster
0   -0.419387
1    1.524306
2   -0.330751
Name: Duration, dtype: float64
```

Figura 11. Médias dos atributos numéricos para cada cluster

Percebemos que o cluster 1 tem uma média de idade maior, enquanto o cluster 3, uma menor.

Também percebemos que a quantidade de crédito é maior no cluster 2 e bem inferiores nos clusters 1 e 3.Algo similar ocorre quanto a duração do empréstimo.

5. Conclusão

Após, pelo método do cotovelo e da silhueta, definirmos que o número ideal de clusters para esse dataset era 3, executamos o algoritmo e observamos os resultados.

Ao analisar Os resultados obtidos quando se agrupa o dataset em 3 clusters, percebemos que o primeiro tem tendências a ser um cluster em que os empréstimos tem menor duração, a pessoa tem menor crédito, são mais velhas, majoritariamente são homens e muitas vezes não possuem imóveis, além de não possuirem empregos qualificados. Quanto ao segundo cluster, observou-se a presença de maior crédito, empréstimos de maior duração, uma proporção maior de imóveis próprios e pessoas com empregos mais qualificados. Quanto ao cluster 3, é observado uma maior proporção de mulheres, muitas vezes jovens com empréstimos de menor duração, com menos crédito, além de um nível de emprego médio.