Análise de Diferentes GMMs e o seu Impacto numa RBFN

Vasco Pereira, 103368

Agenda

01	Objetivos do projeto
02	Dataset
03	Regressão Logística
04	GMMs, covariâncias e valor de silhueta
05	GMMs - Resultados
06	RBFN - Conceito
07	RBFN - Resultado
08	Conclusões
09	Agradecimento & Questões

NTT DATA

TÉCNICO LISBOA

O projeto tem como objetivo principal, além de 5 secundários, criar o modelo mais eficaz em classificar corretamente o dataset

Explorar o uso de Gaussian Mixture Models (GMMs) para determinar o número de neurónios ideais em RBFNs.

Comparar diferentes configurações de covariância (full vs. tied) e avaliar o seu impacto na precisão do modelo.

Investigar o efeito da normalização e redução de dimensionalidade (PCA) no desempenho dos modelos.

Validar e medir a accuracy da RBFN em relação a outros modelos de referência (p.ex., Regressão Logística).

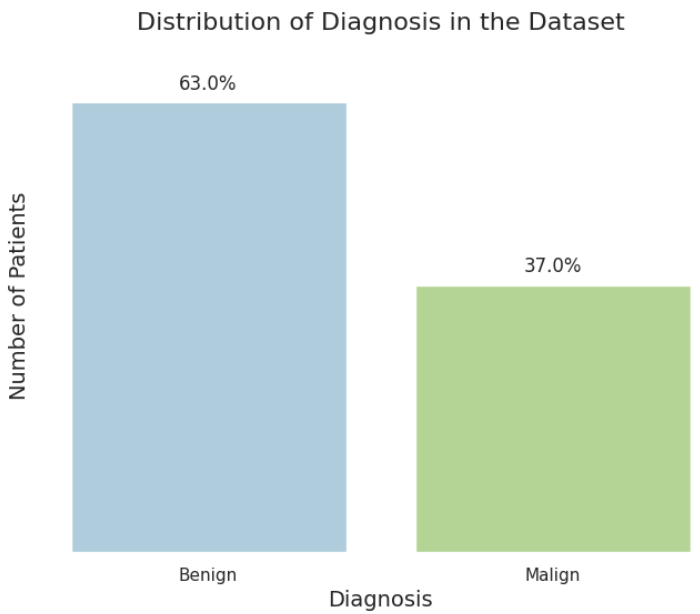
Comparar os clusters ideais para o valor de silhueta e para a accuracy.

Agenda

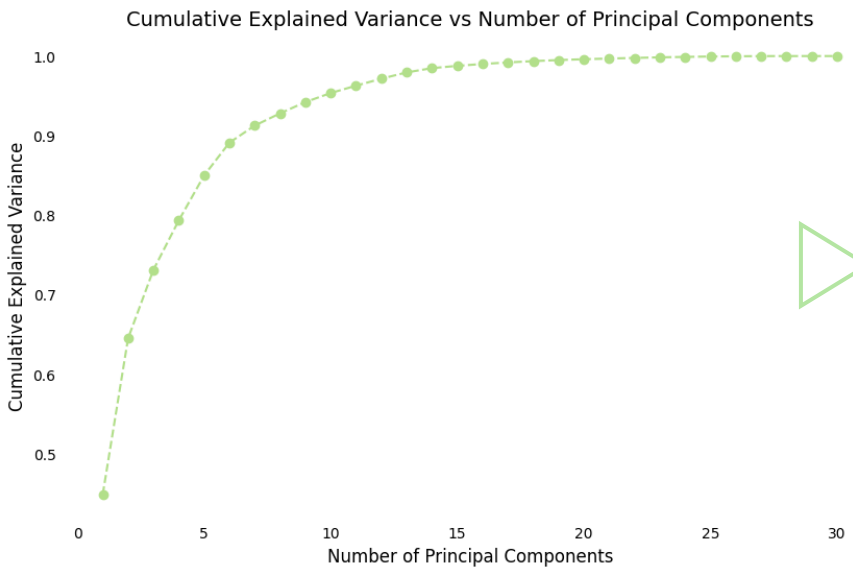
01	Objetivos do projeto
02	Dataset
03	Regressão Logística
04	GMMs, covariâncias e valor de silhueta
05	GMMs - Resultados
06	RBFN - Conceito
07	RBFN - Resultado
08	Conclusões
09	Agradecimento & Questões

NTT DATA TÉCNICO LISBOA

O dataset possui 30 atributos como o raio, a textura, o perímetro e a área e tem como objetivo classificar um tumor como maligno ou benigno

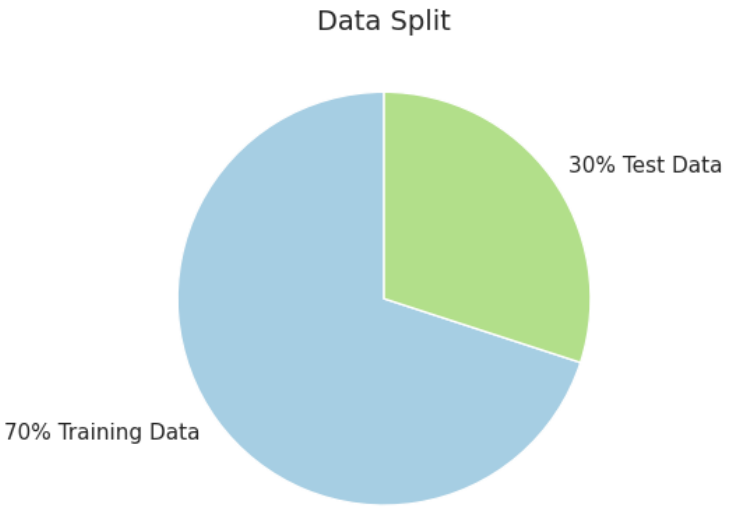


Distribuição das classificações dos tumores.



Número de componentes e respetiva variação. Conclui-se que serão necessários 10 componentes.

O conjunto de dados será dividido em três versões:
1.Base: sem quaisquer alterações ao conjunto original;
2.Normalizado: os atributos passam por normalização para uniformizar a escala;
3.PCA: redução dimensional para manter 95% da variância original dos dados.



Distribuição dos dados: 70% para treino e 30% para testes.

Agenda

01	Objetivos do projeto
02	Dataset
03	Regressão Logística
04	GMMs, covariâncias e valor de silhueta
05	GMMs - Resultados
06	RBFN - Conceito
07	RBFN - Resultado
08	Conclusões
09	Agradecimento & Questões

NTT DATA TÉCNICO LISBOA

A regressão logística utiliza a função sigmoid para modelar probabilidades em problemas de classificação binária

Resultados da regressão logística	
Dataset	**Accuracy**
Base	96.5%
Normalizado	97.7%
PCA	97.7%

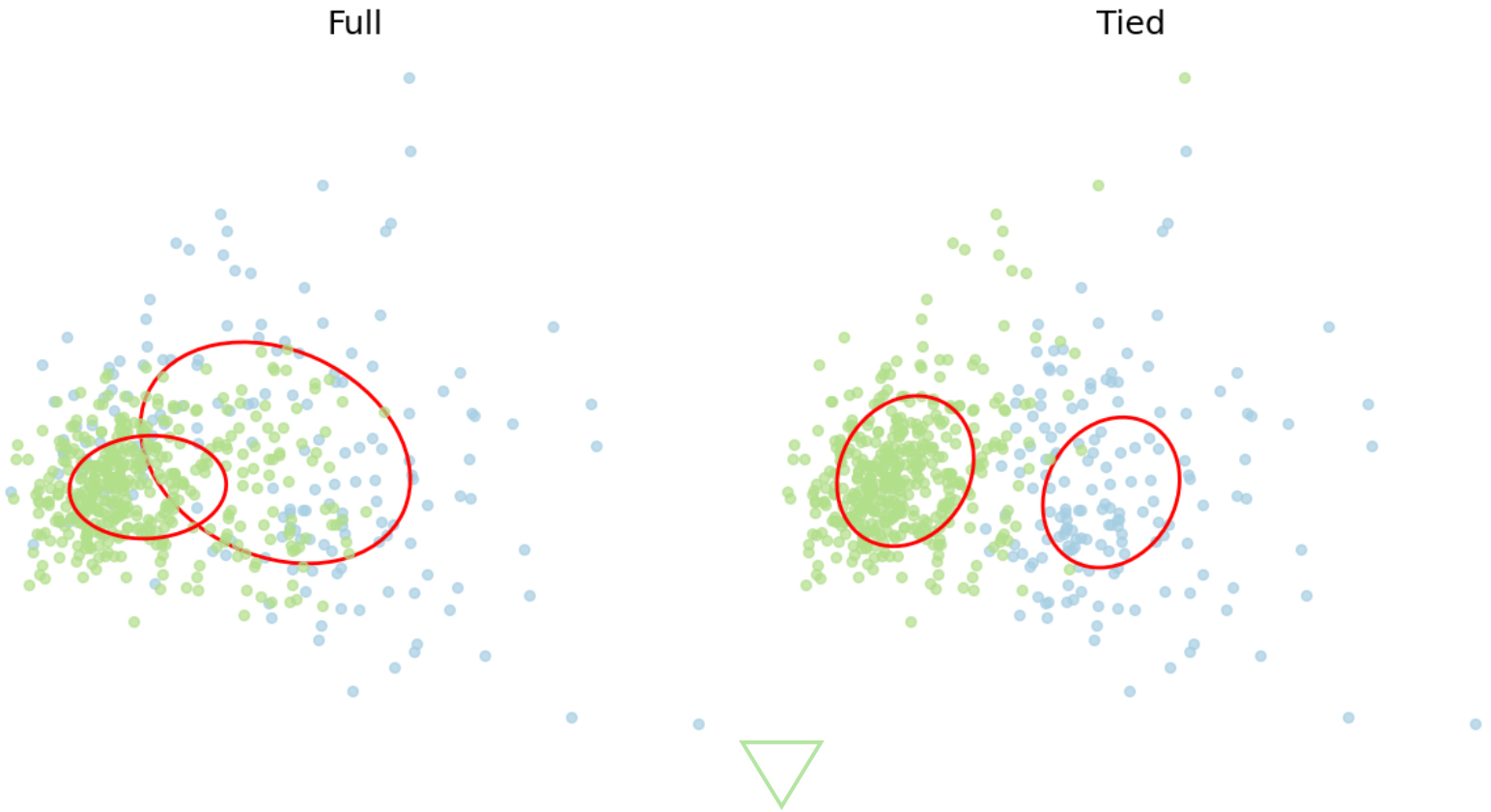
Tanto o dataset normalizado como o dataset PCA apresentaram resultados superiores ao dataset base.

Agenda

01	Objetivos do projeto
02	Dataset
03	Regressão Logística
04	GMMs, covariâncias e valor de silhueta
05	GMMs - Resultados
06	RBFN - Conceito
07	RBFN - Resultado
08	Conclusões
09	Agradecimento & Questões

NTT DATA TÉCNICO LISBOA

Os GMMs serão utilizados para identificar o número ideal de clusters com base nos valores de silhueta e accuracy, variando entre 1 a 150 clusters. Comparemos os resultados utilizando as covariâncias tied e full



Valor de silhueta: Métrica utilizada para avaliar a qualidade de clusters. Mede o quão bem cada ponto está associado ao seu próprio cluster em comparação com os clusters vizinhos. Os valores variam de -1 a 1, onde valores próximos a 1 indicam que os pontos estão bem ajustados ao seu cluster, valores próximos de 0 sugerem sobreposição entre clusters e valores negativos indicam que os pontos podem estar no cluster errado.

Diferenças entre Full e Tied:

- **Full**: Cada cluster tem a sua própria matriz de covariância, permitindo que cada um assuma uma forma específica e mais flexível.
- **Tied**: Todos os clusters partilham a mesma matriz de covariância, resultando em clusters com a mesma forma.

Agenda

01	Objetivos do projeto
02	Dataset
03	Regressão Logística
04	GMMs, covariâncias e valor de silhueta
05	GMMs - Resultados
06	RBFN - Conceito
07	RBFN - Resultado
08	Conclusões
09	Agradecimento & Questões

NTT DATA TÉCNICO LISBOA

Clusters ideais identificados com base na avaliação da silhueta e da accuracy

Full		
Dataset	**Clusters que maximizam a silhueta**	**Clusters que maximizam a accuracy**
Base	2	2
Normalizado	2	7
PCA	2	5

Tied		
Dataset	**Clusters que maximizam a silhueta**	**Clusters que maximizam a accuracy**
Base	2	10
Normalizado	2	58
PCA	2	29

Agenda

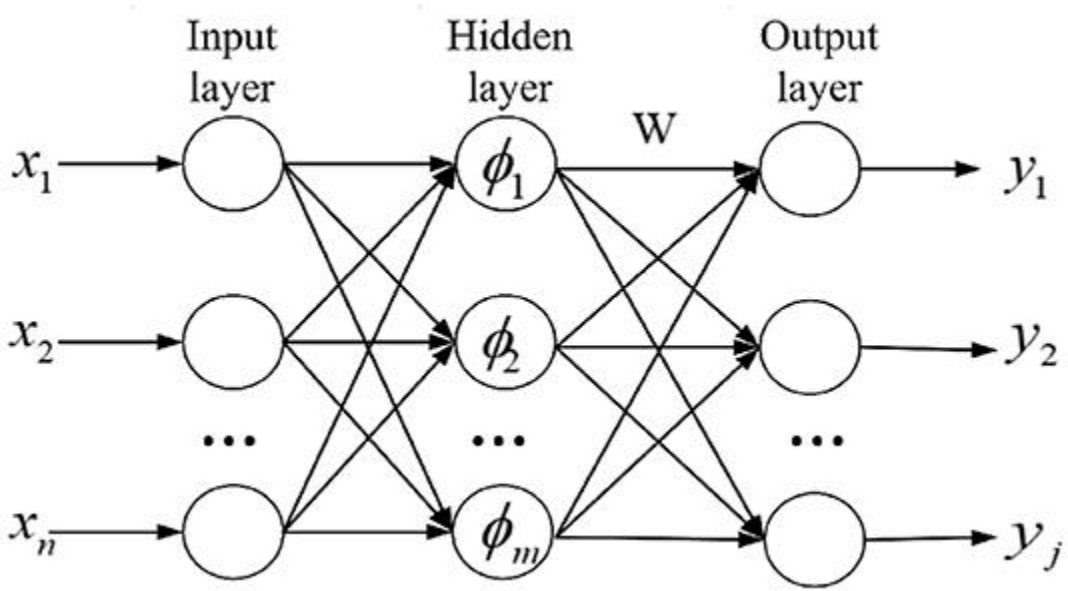
01	Objetivos do projeto
02	Dataset
03	Regressão Logística
04	GMMs, covariâncias e valor de silhueta
05	GMMs - Resultados
06	RBFN - Conceito
07	RBFN - Resultado
08	Conclusões
09	Agradecimento & Questões

NTT DATA

TÉCNICO LISBOA

RBFN (Radial Basis Function Network) é um tipo de rede neuronal que utiliza funções de base radial (como a Gaussiana) na camada oculta

RBFN



Representação de uma RBFN. A hidden layer terá um número de neurónios igual ao número de clusters previamente calculados. A função gaussiana utilizada será a seguinte:

$$\vartheta(x) = e^{-\beta_i \|x - c_i\|^2}$$

Onde c_i é o centro do cluster i e β_i é igual a $\frac{1}{2d_{mean}^2}$, sendo d_{mean} a distância média dos pontos do cluster i relativamente ao seu centro.

Agenda

01	Objetivos do projeto
02	Dataset
03	Regressão Logística
04	GMMs, covariâncias e valor de silhueta
05	GMMs - Resultados
06	RBFN - Conceito
07	RBFN - Resultado
08	Conclusões
09	Agradecimento & Questões

NTT DATA

TÉCNICO LISBOA

Accuracy das RBFNs calculados utilizando o número de clusters ideias

Full		
Dataset	**Accuracy com clusters que maximizam a silhueta**	**Accuracy com clusters que maximizam a accuracy**
Base	88.9%	88.9%
Normalizado	93.0%	96.5%
PCA	88.9%	94.7%

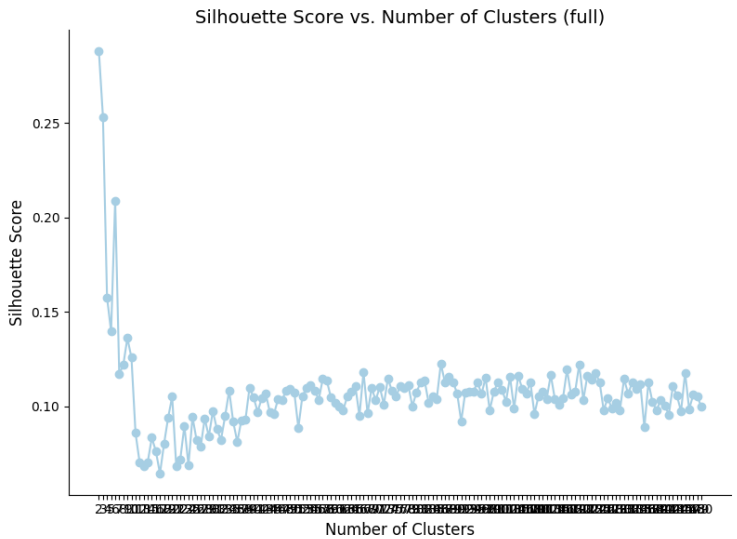
Tied		
Dataset	**Accuracy com clusters que maximizam a silhueta**	**Accuracy com clusters que maximizam a accuracy**
Base	89.5%	90.6%
Normalizado	93.6%	96.5%
PCA	94.7%	98.8%

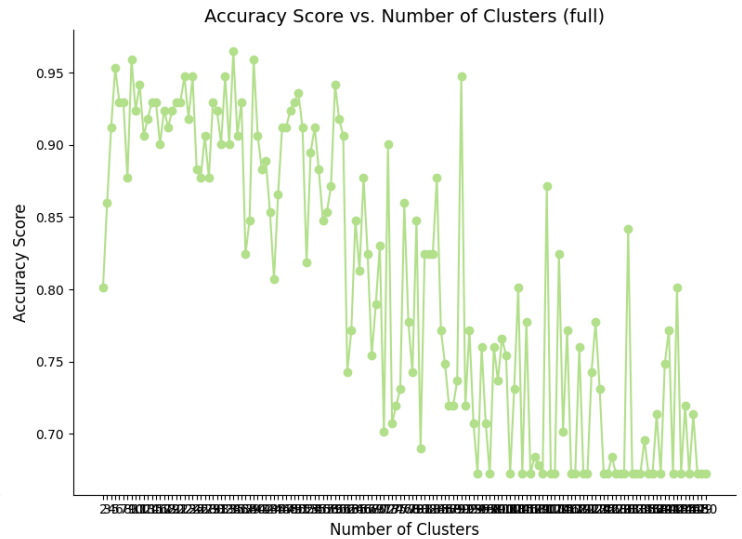
Agenda

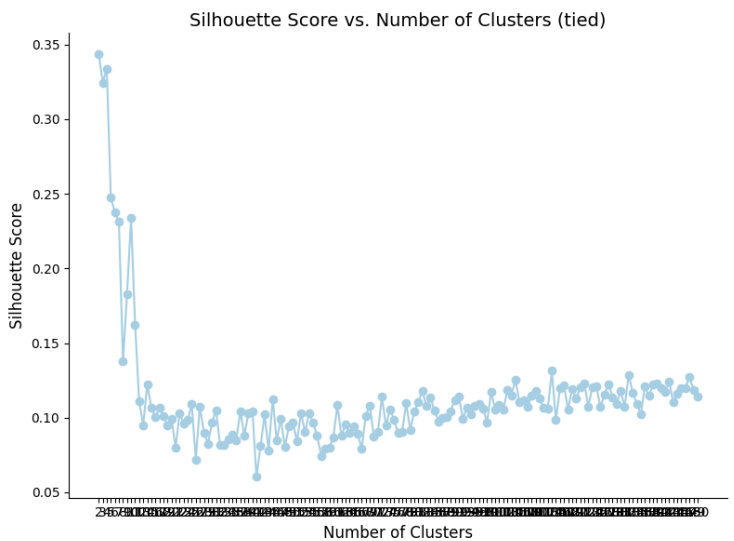
01	Objetivos do projeto
02	Dataset
03	Regressão Logística
04	GMMs, covariâncias e valor de silhueta
05	GMMs - Resultados
06	RBFN - Conceito
07	RBFN - Resultado
08	Conclusões
09	Agradecimento & Questões

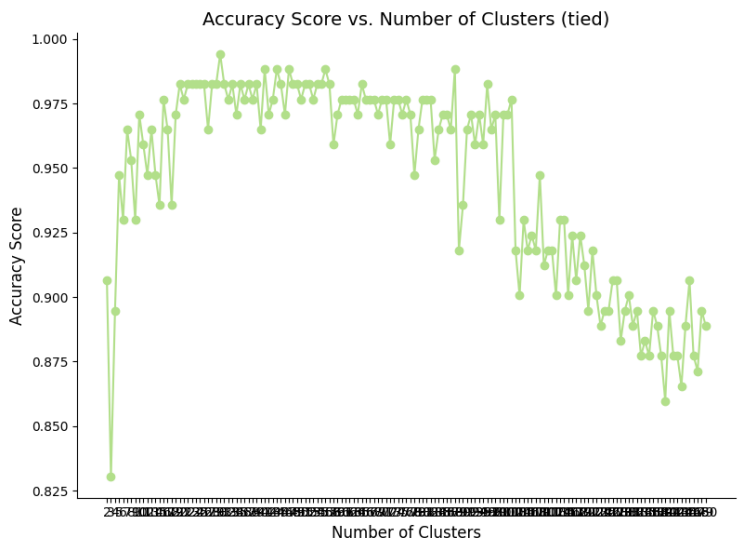
NTT DATA TÉCNICO LISBOA

Chegou-se a conclusão que o melhor modelo é uma RBFN com o dataset PCA, covariância tied e 29 neurónios na hidden layer









Gráficos a comparar a variação do valor de silhueta e da accuracy em ambas as covariâncias com o dataset PCA.

Agenda

01	Objetivos do projeto
02	Dataset
03	Regressão Logística
04	GMMs, covariâncias e valor de silhueta
05	GMMs - Resultados
06	RBFN - Conceito
07	RBFN - Resultado
08	Conclusões
09	Agradecimento & Questões

NTT DATA TÉCNICO LISBOA