**Parte 2**

**(a) Orange environment and analysis of existing examples**

**(b)** Identify the list of supported file formats/types for datasets. Show a screen-shot of the

Uma imagem com texto, captura de ecrã, software, Ícone de computador

Descrição gerada automaticamentestatistical analysis.

Uma imagem com texto, software, diagrama, Gráfico

Descrição gerada automaticamente

**(c)** Explain the purpose of this example. Identify the information provided by the Data Info

widget. What are the most relevant features with the IG and FCBF criteria?

O widget do Data Info dá informação acerca do dataset, nomeadamente a sua dimensão, com 150 linhas e 6 colunas. Das 6 features, 4 são numéricas, 1 categórica, e outra de metadados.

Uma imagem com texto, captura de ecrã, Tipo de letra, número

Descrição gerada automaticamenteAcrescenta também informação adicional, como o nome completo, a descrição do dataset, os autores, o ano de publicação, e as referências.

Segundo o filtro FCBF para Feature Selection, as features mais relevantes são ‘petal\_length’ e ‘petal\_width’

Uma imagem com texto, captura de ecrã, número, Tipo de letra

Descrição gerada automaticamente

**(b) Feature ranking and selection**

**(a)** What seems to be the most relevant feature?

À exceção do filtro ReliefF, a feature ‘petal\_length’ é a mais relevante, sendo que a segunda mais relevante é a ‘petal\_width’, que supera a anterior apenas no filtro mencionado.

Uma imagem com texto, captura de ecrã, Tipo de letra, número

Descrição gerada automaticamente

Figura 1 - Petal Length bacano ya

Uma imagem com texto, captura de ecrã, Tipo de letra, número

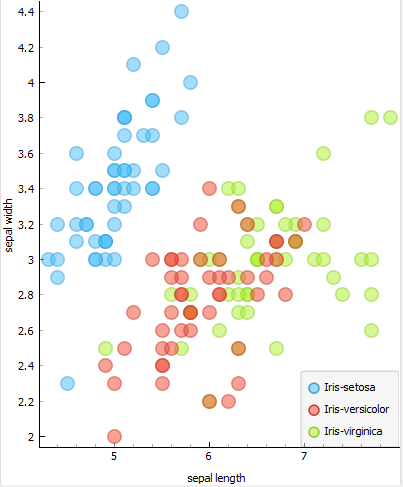
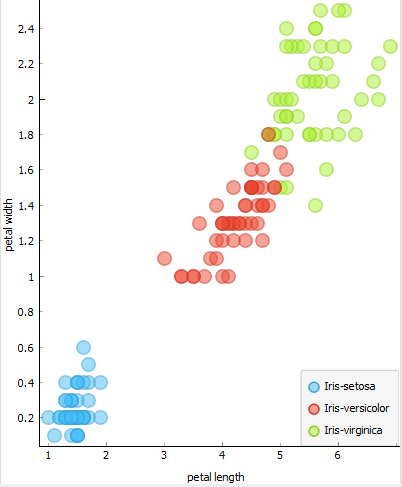
Descrição gerada automaticamente

Figura 2 - Aqui ja e a outra ya

(b) Show some screen-shots that of your analysis to find the most relevant feature and justify your answer.

Uma imagem com texto, captura de ecrã, Tipo de letra, número

Descrição gerada automaticamenteNo Scatter Plot, utiliza-se a função Find Informative Projections, que indica as combinações de features que possam produzir resultados mais relevantes. Comparando a primeira e a última classificada obtemos:



Pela lista obtida, concluímos que as features que aparecem nas melhores opções são ‘petal\_length’ e ‘petal\_width’.

Observando também os gráficos produzidos, podemos concluir que os dados mais relevantes são os que apresentam maior distância entre classes, representadas por cores distintas, e que os valores de cada classe se sobrepõem menos.

No exemplo do scatter plot da relação entre ‘sepal\_length’ e ‘sepal\_width’ podemos observar que as classes se sobrepõem bastante, com alguma sobreposição de valores dentro da mesma classe.

**(c) Feature reduction with principal component analysis and discretization**

1. Explain the key actions of this demo and find an adequate number of reduced dimensions.

O PCA é um método de Feature Reduction não supervisionado, que tem por objetivo reduzir o tamanho ‘d’ inicial do dataset num conjunto com ‘m’ componentes. Para tal, devem ser selecionadas componentes que permitam garantir cerca de 90%-99% da variância dos dados originais. Selecionando um número inferior de componentes, a variância dos dados originais diminui e estaríamos a perder informação acerca dos dados originais.

Na ferramenta do PCA, se escolhermos uma variância de 90%, necessitamos de 25 componentes, ou seja, seria feita a redução de 76 componentes do dataset originais para 25. Para uma variância acumulada de 99% precisam-se de 59 componentes.

Para o cenário de variância 99%, estamos a preservar mais os valores originais, no entanto tornamos a introduzir um elevado número de componentes e voltamos ao problema da maldição da dimensionalidade. É necessário portanto encontrar um compromisso entre o numero de componentes e a variância que se pretende. Neste caso, o grupo optou por variância acumulada de 90%, com 25 componentes.

Uma imagem com texto, Gráfico, file, diagrama

Descrição gerada automaticamente

Pelo Scatter Plot, e com recurso às “Informative Projections”, obtemos que a melhor classificação é dada pela relação da primeira e terceira componente.

Observando a primeira e terceira componente no Scatter Plot, obtemos:

Uma imagem com texto, captura de ecrã, Saturação de cores

Descrição gerada automaticamente

1. Show the Orange widget that performs these actions as well as the resulting file.

Uma imagem com logótipo, Tipo de letra, captura de ecrã, Gráficos

Descrição gerada automaticamenteO Widget no Orange para fazer estas ações encontra-se em “Transform -> Discretize”:

Fazendo a discretização das 25 componentes determinadas pelo PCA, obtemos o seguinte Scatter Plot (Representação de PC1 e PC3)

Uma imagem com texto, captura de ecrã, Saturação de cores

Descrição gerada automaticamente

O esquema utilizado no Orange foi o seguinte:

**Uma imagem com texto, captura de ecrã, diagrama, file

Descrição gerada automaticamente**

**PART III. R STUDIO**

1. **Feature Selection**
2. Compute the (unsupervised) relevance of each feature, using variance and mean-median, as the relevance measures

Plot the sorted relevance values in decreasing order. Comment on the resulting plot.

Compare on the smallest and the largest relevance value.

Uma imagem com texto, número, file, Tipo de letra

Descrição gerada automaticamente

1. For the relevance values found in (a), compute an adequate number of features, m, by the cumulative sorted relevance criterion, with three different thresholds.

State the value of the considered thresholds as well as the corresponding values of m.

Uma imagem com texto, captura de ecrã, file, Tipo de letra

Descrição gerada automaticamenteUma imagem com texto, captura de ecrã, número, file

Descrição gerada automaticamente

Uma imagem com texto, captura de ecrã, Tipo de letra, número

Descrição gerada automaticamenteUma imagem com texto, captura de ecrã, número, file

Descrição gerada automaticamente

1. Repeat (a) and (b) using the Fisher ratio as the relevance measure instead of the variance/mean-median relevance. Comment on the results.
2. **Feature Reduction**
3. Compute the PCA decomposition.

Plot the corresponding eigenvalues sorted in decreasing order. What would be an adequate number of reduced dimensions, m, for this dataset?

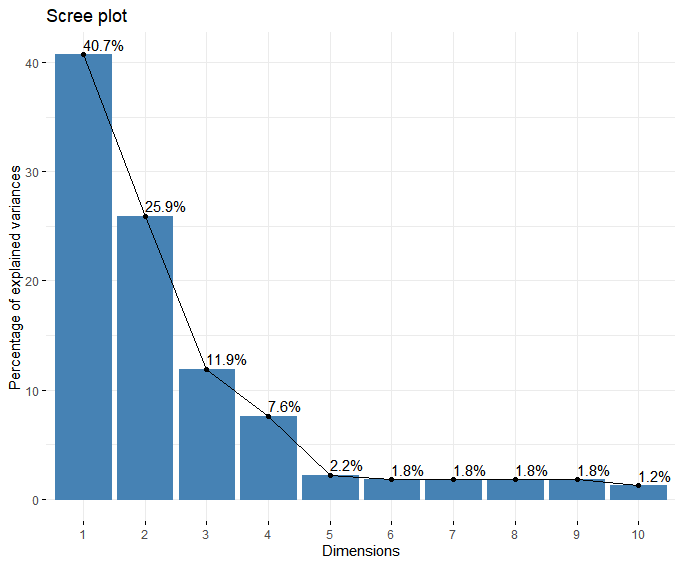


Figura 3 - Percentagens de variância de cada dimensão do dataset Lisbon

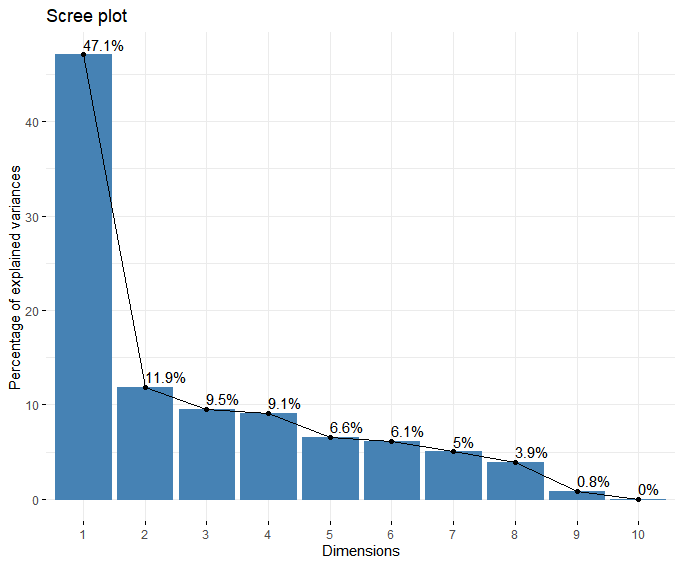


Figura 4 - Percentagem de variância de cada dimensão do dataset pima

O número reduzido de dimensões é obtido através da percentagem observada na Figura 3 e Figura 4. As dimensões escolhidas são as que, somando as percentagens, perfazem no mínimo 80%.

No caso do dataset Lisbon há uma redução para 4 dimensões. No caso do dataset pima há uma redução para 5 dimensões.

* + - * 1. Compute the SVD decomposition.

Plot the corresponding singular values sorted in decreasing order. What would be an adequate number of reduced dimensions, m, for this dataset?

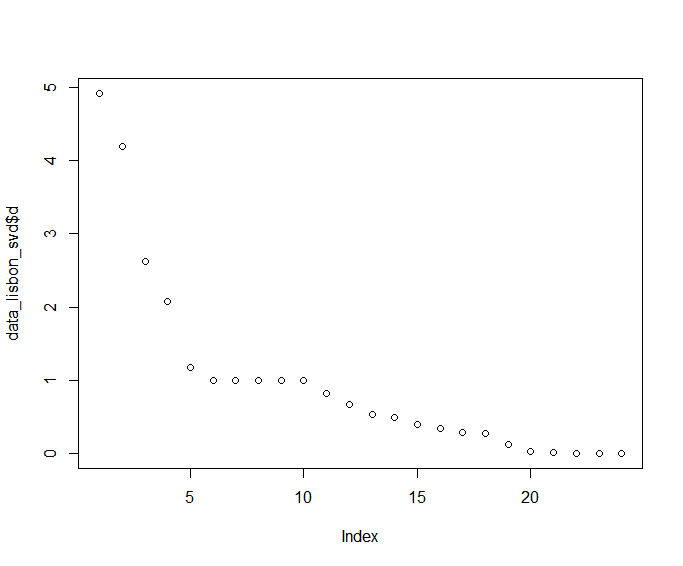


Figure 3 - SVD do dataset Lisbon

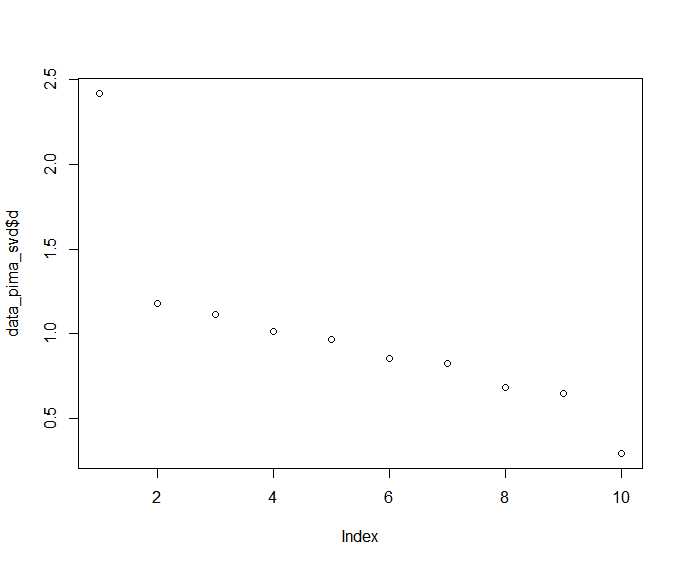


Figure 4 - SVD do dataset pima

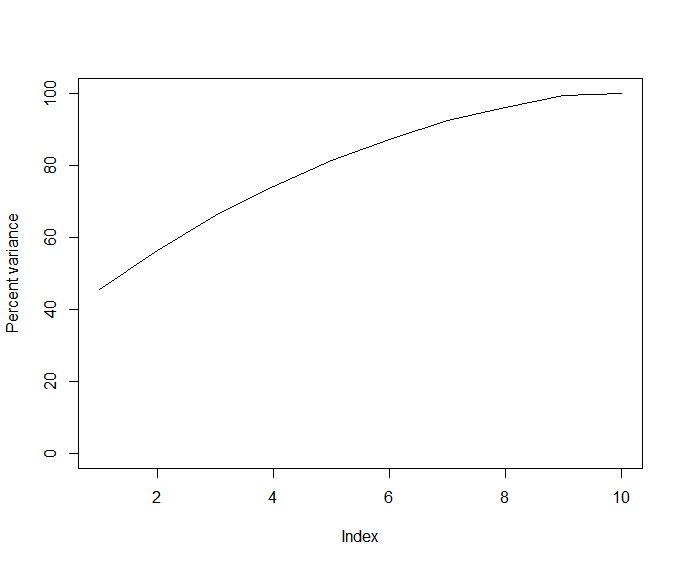


Figure 6 - Soma cumulativa da variância do dataset pima

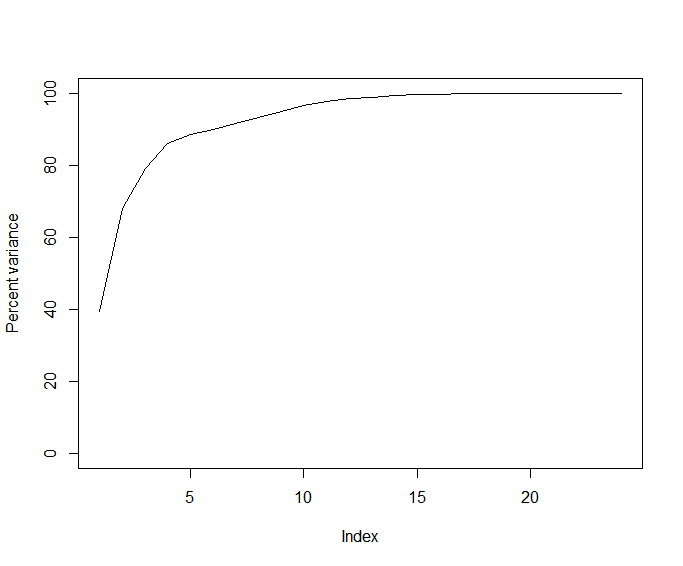


Figure 5 - Soma cumulativa da variância do dataset Lisbon

Observando os gráficos da soma cumulativa e tendo em conta que se deve escolher o número de dimensões com uma soma cumulativa de pelo menos 80% é possível concluir que para o dataset Lisbon a dimensão adequada é de 4 e para o dataset pima a dimensão adequada é 5.

* + - * 1. Using the decomposition results of (a) and (b), compute the dimensionality-reduced versions of both datasets.

Explain how you perform the dimensionality reduction. State the number of features of the reduced datasets.

Sendo que as dimensões adequadas para a redução são iguais tanto no PCA como no SVD as dimensões do dataset Lisbon passam para 4 e do dataset pima para 5.

Para efetuar essa redução basta utilizar apenas as primeiras 4 e 5 colunas dos datasets Lisbon e pima respetivamente.

1. **Feature Discretization**
   1. Discretização não supervisionada

O método de discretização não supervionada escolhido foi o Equal Frequency Binning. Os dados numéricos são divididos em grupos (bins) o que reduz a complexidade e ajuda quando os dados são usados em algoritmos de classificação por exemplo.

* 1. Discretização supervisionada