

**Data Science Project 2018**

**TEAM 27**

**APS Failure at Scania Trucks**

&

**Quality Assessment of Digital Colposcopies**

Francisco Neves, ist424770

Francisco Catarrinho, ist424771

João Ramos, ist180915

**TABLE OF CONTENTS**

[1. Introduction 3](#_Toc529983058)

[2. Pre-processing 3](#_Toc529983059)

[**2.1 APS Failure at Scania Trucks 3**](#_APS_Failure_at)

[2.1.1 Unsupervised Learning 3](#_Toc529983061)

[2.1.2 Supervised Learning 3](#_Toc529983062)

[**2.2 Quality Assessment of Digital Colposcopies 3**](#_Toc529983063)

[2.2.1 Unsupervised Learning 3](#_Toc529983064)

[2.2.2 Supervised Learning 3](#_Toc529983065)

[3. Exploration 4](#_Toc529983066)

[**3.1 APS Failure at Scania Trucks 4**](#_Toc529983067)

[3.1.1 Unsupervised Learning 4](#_Toc529983068)

[3.1.1.1 Methods and Parametrization 4](#_Toc529983069)

[3.1.1.2 Results 4](#_Toc529983070)

[3.1.2 Supervised Learning 4](#_Toc529983071)

[3.1.2.1 Methods and Parametrization 4](#_Toc529983072)

[3.1.2.2 Results 4](#_Toc529983073)

[**3.2 Quality Assessment of Digital Colposcopies 4**](#_Toc529983074)

[3.2.1 Unsupervised Learning 4](#_Toc529983075)

[3.2.1.1 Methods and Parametrization 4](#_Toc529983076)

[3.2.1.2 Results 4](#_Toc529983077)

[3.2.2 Supervised Learning 5](#_Toc529983078)

[3.2.2.1 Methods and Parametrization 5](#_Toc529983079)

[3.2.2.2 Results 5](#_Toc529983080)

[4. Critical Analysis 5](#_Toc529983081)

[5. Conclusions 5](#_Toc529983082)

# Introduction

O presente relatório no âmbito da U.C. de Ciência de Dados tem como objetivo a análise de dois problemas distintos.

O primeiro problema, **APS Failure at Scania Trucks** fornece-nos um *dataset* com informação recolhida acerca do sistema APS de camiões da *Scania.* Este *dataset* apresenta 170 atributos contínuos e 1 atributo *target* binário, estando dividido num conjunto de treino com 60.000 observações e num conjunto de teste com 16.000 observações. O atributo *target* pode tomar os valores *pos* e *neg,* que indicam se ocorreu uma avaria ou não respetivamente. Para a avaliação dos resultados obtidos é-nos fornecida uma métrica dada por , onde é o número de falsos positivos e o número de falsos negativos na classificação, sendo o objetivo do problema minimizar a mesma.

No segundo problema, **Quality Assessment of Digital Colposcopies**, é-nos apresentada informação relativa a uma série de imagens capturadas em colposcopias usando 3 diferentes técnicas de visualização usadas neste tipo de exame. O *dataset* está dividido em 3 ficheiros, cada um correspondendo a uma das técnicas (*green, schiller* e *hinselmann*). Existem 62 atributos contínuos e 7 atributos *target*, sendo que 6 destes representam a avaliação dada por diferentes peritos do domínio acerca da qualidade das imagens. Já o último atributo representa o consenso entre os vários peritos. No total, o *dataset* tem 287 observações (98 no *green*, 92 no *schiller* e 97 no *hinselmann*).

# Pre-processing

## APS Failure at Scania Trucks

## Unsupervised Learning

* **Association Rules**

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

* **Clustering**

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

## Supervised Learning

Uma vez que o *dataset* apresenta cerca de 99% de linhas com *missing values* não foi considerada a remoção de todas estas linhas. Optou-se então por testar a sua substituição utilizando diferentes métodos (substituição por zero, média, moda, mediana) e para cada classificador foi escolhido o método que obteve melhores resultados. Decidiu-se, no entanto, remover as linhas e colunas que apresentassem mais de 90% de *missing values*.

Após a análise do desvio padrão de cada um dos atributos eliminou-se a coluna *cd\_000*, pois apresentava um desvio padrão de 0 (tomava sempre o mesmo valor), não acrescentado informação pertinente para o contexto do problema. Com o objetivo de reduzir a dimensionalidade dos atributos foi feita ainda uma análise utilizando o *Coeficiente de Correlação de Pearson* e descobriram-se 4 grupos de atributos que apresentavam uma correlação linear entre si. Foi então utilizado o *PCA* para reduzir cada um destes grupos em apenas um.

Para a execução do classificador *KNN* procedeu-se à normalização dos atributos com vista a reduzir o impacto das diferentes escalas de atributos sobre a distância calculada para os *vizinhos*.

Foi testada a remoção de *outliers* utilizando o método de *IQR* e verificou-se que os resultados da classificação pioravam. Após uma reflexão acerca da pioria dos resultados aquando da remoção dos *outliers,* chegou-se à conclusão que os mesmos são bastante importantes devido à natureza do problema. Uma vez que se pretende analisar a presença de avarias, os atributos que apresentem valores “fora do normal” (*outliers)* são essenciais para realizar tal previsão. Também não foram removidos valores potencialmente incorretos, pois todos os atributos estão anonimizados e é impossível saber quais os valores que cada atributo poderia tomar.

Por fim, devido ao elevado desequilíbrio entre as instâncias da classe positiva (1.000) e da classe positiva (59.000), procedeu-se ao balanceamento de ambas as classes utilizando diferentes métodos, de modo a evitar que os classificadores ficassem *biased* para a classe negativa, devido à elevada discrepância do número de instâncias de cada classe.

## Quality Assessment of Digital Colposcopies

## Unsupervised Learning

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

## Supervised Learning

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum



Figure 1. Figure caption.

# Exploration

## APS Failure at Scania Trucks

## Unsupervised Learning

## Methods and Parametrization

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

## Results

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

## Supervised Learning

## Methods and Parametrization

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

## Results

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

## Quality Assessment of Digital Colposcopies

## Unsupervised Learning

## Methods and Parametrization

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

## Results

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

## Supervised Learning

## Methods and Parametrization

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

## Results

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

# Critical Analysis

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.

# Conclusions

Vunc sed eede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh adsumes.  
Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum.