**Neste trabalho, o aluno terá de demonstrar de forma clara que compreendeu o problema de mineração de dados a resolver, e que consegue escolher e aplicar as soluções mais adequadas em cada momento do desenvolvimento do modelo de aprendizagem. O conjunto de dados está disponível aqui.**

**Objectivos de Aprendizagem - Fase 1**

**Nesta fase, o aluno deve ser capaz de:**

**Descrever o problema concreto de mineração de dados a resolver e enquadrá-lo no contexto geral das tarefas de mineração de dados;**

O problema em concreto de mineração de dados presente neste problema é de tentar encontrar uma corelação entre todos os dados que são fornecidos, especificamente, encontrar uma correlação entre o gasto de energia, o estado do tempo e algumas estatísticas demográficas.

Este problema encaixa no contexto geral das tarefas de mineração de dados na parte de modelação para previsões futuras. Se conseguirmos perceber e modular as relações mencionadas através de padrões presentes nos dados seremos capazes de criar previsões que são estatisticamente suportadas.

**Caracterizar, em detalhe, o conjunto de dados usado e os critérios seguidos para a sua construção;**

O conjunto de dados fornecidos estão divididos em 3 partes, o consumo horário de energia, histórico de informações sobre o tempo em Lisboa e os censos feitos a Lisboa em 2021.

Sobre o consumo horário de energia, temos a data e hora em que foram recolhidos os dados, o código zip relativo a zona onde foram recolhidos e por último o valor da energia que estava a ser consumida em si em KWh. Temos um pouco mais de 3.7 milhões de instâncias.  
Sobre o histórico de informações sobre o tempo em Lisboa, temos a data e a hora, tal como tínhamos no consumo de energia, em seguida temos vários dados sobre o tempo, incluindo a temperatura, humidade, velocidade do vento e cobertura das nuvens. Estes dados estão separados em 13 ficheiros diferentes e contêm a volta de 740 instâncias cada, para um total de mais de 9600 instâncias.

Sobre os censos feitos a Lisboa em 2021, temos muitos atributos incluindo a secção a que pertencem, o número de edifícios, o número de residentes e o número de agregados familiares. Existem um pouco mais de 2800 instâncias.

**Indicar os problemas e soluções adotadas nas fases de pré-processamento;**

Como é possível perceber estes conjuntos de dados contêm problemas no âmbito do exercício de mineração de dados, sendo estes:

* Apesar dos censos e do tempo serem apenas da região de Lisboa, o consumo de energia não é exclusivo a Lisboa.
* Os censos contêm quase 50 atributos, ou seja, a sua dimensionalidade é elevada.
* O tempo também contém muitos atributos incluindo alguns que serão claramente inúteis para o trabalho como por exemplo o “icon”.

Para resolver estes problemas foram feitas reduções de dimensionalidade. Começando pelos consumos de energia, foram identificados os seguintes atributos:

1. Date/Time;
2. Date;
3. Hour;
4. Zip Code;
5. Active Energy (kWh)

Removemos logo o atributo Date e Hour devido a estes valores já estarem presentes no atributo Date/Time que é do tipo TimeStamp tal como o atributo datetime das tabelas do tempo.

Outra alteração foi filtrar os Zip Code, para conter apenas valores entre 1000 e 1990 que é o intervalo dos códigos postais de Lisboa ([pesquisa](https://www.google.com/search?q=What%20is%20the%20range%20of%20postal%20codes%20in%20Lisbon%3F)), fizemos isto porque os restantes dados são apenas referentes a Lisboa.

Para os ficheiros que contem a informação do tempo, foram identificados todos os atributos. Destes foram removidos os que não tinham mais de 20 valores únicos ou seja:

1. "name",
2. "precipprob",
3. "preciptype",
4. "snow",
5. "snowdepth",
6. "uvindex",
7. "severerisk",
8. "icon",

A única exceção foi “conditions” porque acreditamos que pode ser relevante apesar de ser mais geral.

Também foi removido o “feelslike” por ser sempre muito similar ao “temp” e o “precip” por ter 98% dos valores como 0 e “visibility” por 86.6% dos seus valores serem 10, 10.8 ou 11.6.

Para os censos…

**Aplicar corretamente as técnicas de redução de dimensionalidade e de representação adequada dos dados;**

**Criar e disponibilizar o código R que implemente todas as análises e soluções encontradas;**

**Elaborar um relatório que explique, de forma clara, o processo adoptado e os resultados experimentais obtidos.**