Slide 1

|  |  |
| --- | --- |
| Destacar que o título já dá um ligeiro overview às questões colocadas e eventuais respostas |  |

Slide 2

|  |  |
| --- | --- |
| A apresentação seguirá o seguinte plano:  1 - Inicialmente começamos por descrever as características dos conjuntos de dados, nomeadamente o número de *features* e de observações, indicando que decisões tomámos em lidar com valores em falta, ou na união dos dois conjuntos de dados.  2 - Com os dados preparados, segue-se a exposição das questões formuladas, já indiciadas pelo título da apresentação e explicam-se que técnicas de redução de dimensionalidade foram aplicadas.  3 - Depois de preparados os conjuntos de dados para responder às questões colocadas, mostram-me os modelos explorados pelo grupo, comparando as métricas de performance em cenários com e sem amostragem.  4 - Por fim mostram-se conclusões, indicando alguns dos problemas que tivemos e que pretendemos corrigir, bem como possíveis sugestões a serem abordadas num cenário futuro. |  |

Slide 3

|  |  |
| --- | --- |
| Dos dois conjuntos de dados fornecidos, numa fase inicial, interpretaram-se as features de cada um, e o seu respetivo significado, e começou-se por explorar alguns valores que pudessem estar em falta.  Neste caso, o dataset das condições meteorológicas em Lisboa apresenta uma dimensionalidade de 24 *features* com 9504 observações. Da analise estatística, observação dos valores em falta, notou-se que a feature que indica o risco severo está ausente em grande parte das observações (8760 valores, representativo de aproximadamente 92% das observações). Em simultâneo, outras *features* apresentam valores em falta.  Desta forma, é afastada a ideia de remover estas observações, uma vez que a perda de informação seria consideravelmente grande, reduzindo o *dataset* a apenas 8% da dimensão original, pelo que se optou por fazer imputação.  Tendo por base o desvio padrão das features, num cenário onde este é elevado imputou-se com a mediana, utilizando a média para cenários onde o desvio padrão é menor.  Para o severerisk, de acordo com a documentação dos dados, pode ser feita a imputação com valor 0 |  |

Slide 4

|  |  |
| --- | --- |
| Seguindo para o conjunto de dados da energia, composto por 5 *features,* com um registo de um total de aproximadamente 3.8 milhões medições de energia ativa consumida para os vários códigos postais compreendidos ao longo do país. Verificou-se que todos as entradas possuem um valor não nulo, ou NA, para todas as *features,* pelo que não foi necessário fazer qualquer tipo de imputação.  Num cenário inicial, começando pelos códigos postais, com auxílio a dados externos indicadores de consumo industrial e doméstico, provenientes da PorData [4], o grupo considerou interessante a preparação de um *subset* dos dados originais, onde se categoriza cada código postal escolhido como Residencial ou Industrial, na perspetiva de treinar um classificador supervisionado  Após a normalização do novo *dataset*, verificou-se que em zonas industriais o consumo energético muito raramente desce abaixo de 70% do seu consumo máximo diário, ao contrário das zonas residenciais onde a tendência das horas “mortas” e a madrugada demonstra uma forte descida do consumo energético. |  |