

# Processamento de Streams - Módulo II

## Fase I - Extended Abstract

João Funenga nº 61635, Catarina Afonso nº 62206

13 de maio de 2022

### Abstract

Neste *paper*, será descrito o dataset com que se vai trabalhar neste *assignment*, bem como enunciar algumas questões que se quer analisar numa fase posterior sobre os dados que se tem.

### Introdução

Neste projeto trabalhar-se-á sobre um *dataset* que engloba os dados dos tempos de espera nos hospitais na região de Lisboa. Com este conjunto de dados pretende-se resolver algumas perguntas que o grupo irá enunciar na secção seguinte recorrendo a técnicas de Machine Learning (ML) em *streaming*. Relativamente aos dados com que se vai trabalhar, o dataset será limitado para as observações feitas apenas no Santa Maria por ser o que contém mais observações (395858) e por ter menos tipos de urgência (apenas urgência geral e pediátrica).

### Dataset

A tabela 1 representa o *schema* dos dados. O *dataset* é definido por 1603384 instâncias com 8 *features*, sendo estas representadas abaixo:

Table 1: Descrição do dataset

AcquisitionTime	Hospital	UrgencyType	Service
TimeStamp	int	string	string
EmergencyState	WaitingTime	PeopleWaiting	H_name
int	int	int	string

### Desafios

O desafio focar-se-á essencialmente no desenvolvimento do modelo preditivo do tempo de espera sabendo a cor da pulseira e a altura do dia. O tempo de espera dos hospitais representa tudo o que advém de algo anormal. Isto é, no Inverno em que é recorrente haver crises de gripes, os hospitais já têm mais gente para trabalhar logo os tempos de espera já terão isso em conta. Tudo o que foge destes padrões esperados é que entram para o tempo de espera.

### Situação Atual em Portugal

Atualmente, de acordo com o ministério de saúde (0), a estimação de tempo de espera com uma determinada pulseira é recorrendo à mediana de todos os tempos de espera para cada um dos diferentes níveis de manchester, mas tendo em conta a dimensão da instituição. Para cada um dos diferentes níveis, os intervalos de tempo considerados para o cálculo da mediana dos dados variam. Embora, estes cálculos sejam feitos com um intervalo de 5 minutos, ou seja, em contexto de *streams* isto corresponderia a uma janela deslizante de tamanho correspondente ao definido para cada nível de urgência e cujo *stride* seria de 5 minutos. Esta forma de estimação, parece ser algo rudimentar por apenas ser calculada uma mediana e não com um algoritmo de ML mais avançado. No entanto, perceber-se-á se é de facto a melhor escolha ou não no que toca a este problema.

### State of the art

Neste momento, no estrangeiro os modelos para fazerem este tipo de previsões baseiam-se em vários atributos e no fundo são um modelo de regressão cujo valor a estimar é o tempo de espera. No paper é dito que esta metodologia obteve resultados bastante melhores que apenas uma média deslizante sobre uma janela (análogo ao que Portugal faz com a mediana). Com estas melhorias, é possível dar melhores estimativas para o bem estar dos pacientes. Este modelo não procura reduzir os tempos de espera porque esses são sempre devido a acontecimentos aleatórios e esporádicos mas sim de os prever.

### References

Serviço Nacional de Saúde (2019, 15 de Janeiro). *Informação do cálculo de tempos de espera nos hospitais*  
Acedido a 06/05/2022, em [http://tempos.min-saude.pt/attachments/Informacao\\_calculo\\_tempos\\_2019.pdf](http://tempos.min-saude.pt/attachments/Informacao_calculo_tempos_2019.pdf)  
Pak A., Gannon B. Staib A *Predicting waiting time to treatment for emergency department patients*  
Acedido a 06/05/2022, em <https://www.sciencedirect.com/science/article/pii/S1386505620305219>