 <b>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</b>	Tipo de Prova Projeto Prático	Ano letivo 2022/2023	Data
	Curso Mestrado em Engenharia Informática	Hora	
	Unidade Curricular Tecnologias Escaláveis para Análise de Dados	Duração	

#### Observações

- Versão inicial (1.0)

## A. Destinatários

Este trabalho prático destina-se a todos os estudantes inscritos na Unidade Curricular de Tecnologias Escaláveis para Análise de dados, do Mestrado em Engenharia Informática, que pretendam realizar a avaliação durante o período letivo.

O trabalho prático tem um peso de 100% na classificação final da UC e será desenvolvido de forma faseada ao longo do semestre.

## B. Objetivos

Este projeto funcionará como um elemento integrador dos conhecimentos adquiridos na UC de Tecnologias Escaláveis para Análise de Dados e, de uma forma mais geral, no Mestrado em Engenharia Informática, tentando simular de uma forma realista o ciclo de vida de um projeto de análise de dados. Para o efeito, será implementado um projeto proposto por uma empresa nacional: a e-GDS. Nomeadamente, serão trabalhadas competências fundamentais, incluindo:


- Desenvolver e defender abordagens de análise de dados para um domínio concreto, após levantamento dos seus desafios e das suas potencialidades
- Identificar as fontes de dados potencialmente relevantes para o problema
- Adquirir, processar e integrar diferentes fontes de dados públicas, analisando de forma crítica a sua relevância para o problema a resolver
- Criar e avaliar novas variáveis (*feature engineering*) que possam enriquecer um dataset já existente
- Treinar e avaliar modelos de Machine Learning com vista à resolução de problemas concretos
- Capacidades de comunicação e de apresentação

## C. Enunciado

Criada há mais de 25 anos, a e-GDS tem sido o parceiro tecnológico escolhido por milhares de hotéis, cadeias e outros estabelecimentos de hotelaria em todo o mundo, com uma plataforma online que permite gerir o comércio hoteleiro, facilitando a automatização de operações.

O vasto conjunto de integrações que a e-GDS implementa, com plataformas como a Booking.com, permitiu à empresa recolher um vasto conjunto de dados ao longo dos anos, sobre as reservas feitas em cada hotel.

Com este trabalho pretende-se abordar o problema da previsão de ocupação/afluência ao longo do tempo. Para este efeito, a e-GDS disponibiliza um dataset com dados de reservas efetuadas em clientes seus. No entanto, o objetivo deste trabalho vai muito mais além da utilização dos dados

 <b>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</b>	Tipo de Prova Projeto Prático	Ano letivo 2022/2023	Data
	Curso Mestrado em Engenharia Informática	Hora	
	Unidade Curricular Tecnologias Escaláveis para Análise de Dados	Duração	

fornecidos. Pretende-se que os Estudantes sejam capazes de identificar fontes de dados externas e públicas, que possam eventualmente ser indicadores importantes para a previsão de afluência.

Torna-se assim necessário, para além da identificação das fontes de dados potencialmente relevantes, a sua aquisição, o seu tratamento e a sua integração com os dados fornecidos pela e-GDS, para análise conjunta.


Em termos de metodologia, serão seguidas as seguintes principais fases ao longo do trabalho, cuja calendarização se apresenta na tabela 1:

- Business Understanding – Levantamento e análise de literatura sobre o tema, que permita identificar potenciais fontes de dados relevantes, bem como as principais características do domínio
- Data Understanding – Familiarização com os dados fornecidos pela e-GDS, bem como com as fontes de dados externas, e com os conceitos aí descritos
- Data Preparation (collect) – Aquisição de dados de fontes externas, que possam ser relevantes para o problema em análise
- Data Preparation (Integrate) – Integração dos dados externos adquiridos com os dados fornecidos pela e-GDS, de forma a obter um único dataset consolidado
- Data Processing and Analysis – Inclui todas as tarefas necessárias para melhorar a qualidade dos dados (e.g. limpeza, conversões de formatos), ou expor o seu conhecimento intrínseco (e.g. feature engineering). Inclui ainda a criação de visualizações que permitam obter conhecimento sobre problema
- Modeling + Evaluation – Treino e avaliação de modelos de Machine Learning que permitam, de forma satisfatória, dar resposta ao problema da previsão de reservas em unidades hoteleiras

Um problema secundário que poderá ser abordado pelos grupos é o da previsão do cancelamento de reservas. Isto é, num determinado momento, prever que reservas poderão vir a ser canceladas, de forma a que o hotel possa atuar sobre essa informação.

Ao longo do semestre está prevista a entrega de 4 deliverables, que serão parte integrante do relatório final, a saber:

- Deliverable 1 – Problem description + Business Glossary + Data sources description. Este deliverable deve fazer uma caracterização geral do problema bem como dos fatores mais relevantes indicados na literatura. Deve ainda apresentar um glossário com os principais termos do domínio, bem como uma descrição das fontes de dados identificadas, que deve incluir, no mínimo, uma descrição da forma de aceder aos dados, e um dicionário de dados (por fonte).
- Deliverable 2 – Data description report. Este deliverable, que ocorre após a integração dos dados, deve descrever o dataset resultante e que será usado como base para as tarefas de processamento e análise de dados. Entre outros aspetos, o deliverable deve pelo menos analisar a quantidade e qualidade (e.g. distribuição das variáveis, dados em falta, problemas nos dados e tarefas de limpeza necessárias) dos dados disponíveis. Isto será usado como base para o trabalho a desenvolver na fase seguinte.
- Deliverable 3 – Data processing report. Este deliverable deve descrever todas as tarefas de limpeza, transformação, criação, etc. levadas a cabo de forma a melhorar a qualidade dos

 <b>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</b>	Tipo de Prova Projeto Prático	Ano letivo 2022/2023	Data
	Curso Mestrado em Engenharia Informática	Hora	
	Unidade Curricular Tecnologias Escaláveis para Análise de Dados	Duração	

dados, e a expor da melhor forma o conhecimento existente para a fase de Machine Learning que se segue.

- Deliverable 4 – Final Report. Este relatório final deve incluir, para além de todos os elementos já abordados, uma descrição das tarefas de modelação levadas a cabo, bem como uma análise crítica dos seus resultados. Deve incluir ainda, entre outros elementos que o grupo ache pertinente, uma secção de conclusões que auto-avale a qualidade do trabalho levado a cabo, as limitações ou dificuldades encontradas, e o potencial trabalho futuro.

O trabalho guiar-se-á pela calendarização detalhada na Tabela 1.


*Tabela 1: Calendarização do projeto*

Data	Deliverable	Tarefas
28/02/22		Análise da literatura
07/03/22		Constituição de grupos + Análise da literatura/Aquisição de dados
14/03/22		Análise da literatura/Aquisição de dados
21/03/22		Aquisição de dados
28/03/22		Aquisição de dados
04/04/22	Deliverable 1	Integração de dados
11/04/22		Integração de dados
18/04/22	Deliverable 2	Processamento + AD
25/04/22		Férias de Páscoa
02/05/22		Processamento + AD
09/05/22		Processamento + AD
16/05/22		Processamento + AD
23/05/22	Deliverable 3	Modelação
30/05/22		Modelação
06/06/22	Deliverable 4	Apresentação

## D. Realização

Este trabalho é realizado em grupos de 3 elementos, salvo situações excecionais a validar previamente com o docente da UC. A data limite para a comunicação da constituição dos grupos ao docente da Unidade Curricular é o dia 7 de março.

O trabalho consiste numa única entrega final. As deadlines detalhadas na secção anterior são as datas destinadas à discussão de cada entregável, em que o docente dará o seu contributo para a melhoria destes elementos. O grupo pode fazer melhorias aos elementos após esta data, e até à data de submissão, mas os mesmos não serão novamente discutidos após a deadline.

 <small>ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO</small>	Tipo de Prova Projeto Prático	Ano letivo 2022/2023	Data
	Curso Mestrado em Engenharia Informática	Hora	
	Unidade Curricular Tecnologias Escaláveis para Análise de Dados	Duração	

O trabalho desenvolvido, contendo para além do relatório e da apresentação, todos os elementos considerados relevantes pelo grupo, deve ser entregue na página da UC no Moodle, até às 23:55 do dia 1 de junho de 2023.

Apenas um elemento de cada grupo deverá submeter o trabalho em nome do grupo.

A apresentação e defesa do trabalho, de carácter obrigatório e em que todo o grupo deverá estar presente simultaneamente, decorrerá na aula de 6 de junho de 2023.

## E. Critérios de Avaliação

A nota final da UC será dada pelos seguintes elementos principais, cujas componentes são descritas de seguida:

$$NTP = 0.2D_1 + 0.2D_2 + 0.2 D_3 + 0.2D_4 + 0.2Ap$$

Em que:

- NTP – nota do trabalho prático
- $D_n$  – Nota de cada deliverable
- Ap – Nota da apresentação

Os deliverables 1 a 3 serão apenas avaliados no final, não obstante a data para a sua submissão ser anterior. Isto destina-se a permitir ao docente dar feedback sobre o processo do trabalho, e aos Estudantes a possibilidade de o melhorar.

A avaliação do deliverable 4 incidirá sobretudo nos elementos finais (i.e. ML + Conclusões) e não sobre todo o conteúdo do relatório, uma vez que cada deliverable anterior é avaliado separadamente. No entanto, será avaliada a qualidade do relatório enquanto tal (e.g. como as secções foram integradas, qualidade/clareza da escrita, etc.).