Projeto Aplicado em Ciência de Dados II

2023/2024 - 1º semestre

Enunciado do projeto

Versão 1.0 (2023-09-09)

- Este trabalho deverá ser realizado em grupos 4 ou 5 elementos (excecionalmente: 3 elementos)
- Data de entrega: **Sábado, 9 de dezembro até às 23:59**
- A apresentação do projeto será feita nos dias **12 e 13 de dezembro**, durante o horário da aula

Objetivo

Este trabalho tem como objetivo explorar um conjunto de dados reais e extrair daí conhecimento, usando um conjunto de ferramentas à sua escolha, tais como R ou Python (e o pandas). O grupo deve produzir um relatório onde, de uma forma transparente, descreve os seus objetivos, os dados utilizados, os procedimentos efetuados, os resultados obtidos e apresenta as contribuições do grupo. No final do semestre o grupo deverá também preparar uma apresentação para fazer durante a aula.

O relatório e a apresentação devem ter em conta uma audiência, que inclui os restantes colegas do curso e poderá incluír alguns membros da instituição que fornece os dados. Todos devem ser capazes de compreender a análise efetuada, os resultados apresentados e eventuais excertos de código, acompanhados das devidas explicações.

No decorrer do trabalho, deve ter em atenção que:

- o trabalho deve cobrir o maior número possível de elementos anteriormente lecionados na UC: tipos de medida, limpeza e pré-processamento de dados, análise exploratória e visualização de dados, aprendizagem supervisionada e/ou não supervisionada;
- será valorizada a combinação de diferentes tabelas de dados, incluindo tabelas adicionais obtidas ou criadas pelo grupo, que permitam ajudar a responder a determinada questão, de forma não óbvia.
- será desejável que o grupo aplique elementos de aprendizagem automática, por forma a procurar responder a questões relevantes sobre os dados.

Notas adicionais

- ao longo do trabalho poderá ser utilizado código ou ideias provenientes da web. No entanto, deve ser sempre incluída uma referência para os elementos originais
- durante o desenvolvimento do trabalho podem ser feitos os testes que forem necessários. No entanto, no relatório final devem apenas constar os passos e resultados mais relevantes.
- sejam razoáveis: nem sempre é necessário ter uma análise *super-extraordinária*, mas espera-se que pensem! Para cada etapa, perguntem a vocês próprios se um determinado passo faz sentido e depois expliquem-no no vosso relatório.

Conjuntos de dados

O grupo poderá optar por um dos dois conjuntos de dados reais fornecidos pela equipa docente ou, em alternativa, por outro conjunto de dados que se encontre disponível online. A Secção 3 descreve o procedimento a efetuar, caso o grupo pretenda optar por outro conjunto de dados.

Nota importante: *Os dados disponibilizados no âmbito desta Unidade Curricular devem ser usados exclusivamente para esse fim e devem ser totalmente eliminados no final do semestre*

Conjuntos de dados disponíveis:

1. Ocorrência de acidentes rodoviários nas várias regiões do país (recomendado)

Dados fornecidos pela Autoridade Nacional de Segurança Rodoviária (ANSR) sobre a ocorrência de acidentes rodoviários nas várias regiões do país, que cobrem o período de 2010 a 2019.

- Análise de pontos críticos em termos de segurança rodoviária
- Permite fazer tarefas de *Previsão*, *Classificação* e *Deteção de outliers*

No caso de optar por este conjunto de dados, poderá restringir os dados a uma determinada região do país, analisar o comportamento de acidentes rodoviários ao longo de vários anos, comparar diferentes zonas do país, etc. Poderá ser interessante, por exemplo, explorar as principais causas de acidentes em determinado local ou verificar se as condições metereológicas, pavimento, etc. podem estar relacionadas com algum tipo de acidentes. Poderá também ser relevante juntar outros dados disponíveis (IPMA, Proteção Civil, etc.), que nalguns casos se encontram disponíveis para algumas regiões do país e podem ser fornecidos a pedido. Como ponto de partida, poderá considerar algumas das tarefas já realizadas no âmbito de trabalhos anteriores, tais como:

- Como se caracterizam os acidentes no distrito do Porto?
- Quais os motivos que levam o distrito de Portalegre a apresentar a maior percentagem de mortes por acidente?
- Qual a relação entre acidentes com velocípedes e a existência de ciclovias?
- As condições condições climatéricas influenciam a ocorrência de acidentes?
- Indentificar as características de acidentes graves que envolvem morte de alguém
- Compreender a distribuição temporal e geográfica dos acidentes graves
- Fatores que influenciam a gravidade dos acidentes
- Identificar as vias com maior ocorrência de acidentes e os troços de estrada onde estes acontecem
- Estudar os acidentes graves e mortais ocorridos nos troços mais perigosos das vias com maior número de acidentes
- Previsão do tipo de acidente, com base nas características do condutor, veículo, estado da via e condições climatéricas

Caso pretenda obter mais informação sobre estes dados, poderá consultar o material que se encontra na pasta `documentos`, juntamente com os dados. Poderá também consultar o boletim estatístico de acidentes de viação (BEAV) e o respetivo manual de preenchimento (ANSR).

Nota: Os 3 melhores trabalhos serão selecionados para uma apresentação em data posterior, onde estarão presentes elementos da ANSR.

2. Tráfego na cidade de Lisboa

Dados fornecidos pela Câmara Municipal de Lisboa (CML), contendo registos *waze* das vias de entrada em Lisboa e das principais vias da cidade, durante o período de 2019 e 1º semestre de 2020.

- Permite fazer tarefas de *Previsão*, *Classificação* e *Deteção de outliers*

- Permite obter a situação das principais vias de entrada em Lisboa, bem como as principais vias da cidade ao longo do tempo
- Permite estabelecer relações entre o tráfego, a hora do dia, o dia da semana, o dia do mês e o período temporal

No caso de optar por este conjunto de dados, poderá optar por restringir os dados a determinadas vias. Poderá analisar filas de trânsito em função do horário, dia da semana ou dia do mês e criar modelos para previsão do tempo de espera. Caso pretendam utilizar estes dados, pelo menos um dos elementos do grupo (representante) deve registar-se no site Lisboa LxDATALAB, através do qual poderá vir a ter acesso a informações adicionais e a dados mais recentes. O conjunto de dados atualmente disponível ocupa cerca de 6 GBytes (descompactado).

3. Outro conjunto de dados

O grupo poderá também optar por escolher outro conjunto de dados que se encontre disponível online, desde que seja abrangente e contenha uma quantidade considerável de informação. Nesse caso, o grupo deverá obter o conjunto de dados, analisar os dados e confirmar que contém informação relevante para extraír conhecimento acerca de pelo menos 3 questões relevantes acerca desses dados. No caso de optar por esta opção, antes de usar qualquer fonte de dados, o grupo deverá fazer um pedido de autorização aos docentes. Só depois de um feedback positivo será possível utilizar os dados.

Exemplos de organizações que fornecem dados:

- União Europeia: data.europa.eu
- Portal de dados abertos da Administração Pública: dados.gov
- DataHub.io - várias coleções de dados, organizadas por tópicos
- UCI Machine Learning Repository
- OCDE (dados económicos)
- Banco central europeu (dados económicos)
- Banco mundial (dados económicos). É usada uma API para acesso aos dados
- Dados públicos da França
- Kaggle

Etapas da elaboração do projeto

1. Análise exploratória e visualização

Deverá começar por fazer uma análise aos dados, identificando os diversos tipos e escalas de medida para cada uma das variáveis. Poderá ser necessário algum tipo de limpeza e pré-processamento de dados, envolvendo o tratamento de ruído, tratamento de valores omissos ou a identificação e tratamento de *outliers*. Nesse caso, deve explicar o que foi feito e as razões para esse procedimento. Pode também ser necessário fazer a transformação de algumas variáveis.

Para algumas das variáveis, poderá também ser relevante fazer uma análise exploratória e visualização, como forma de melhor compreender a informação presente nos dados.

2. Desenvolvimento do projeto

O trabalho deverá ser sempre desenvolvido a partir de questões claras, para as quais se pretendem obter respostas. Assim, o grupo deverá começar por preparar (pelo menos) 2 questões relevantes sobre os dados. De seguida, para cada uma das questões deverá:

- justificar a relevância da questão, de forma sucinta
- identificar e descrever a informação necessária à obtenção de uma resposta à questão
- explicar as etapas utilizadas no processo de obtenção de resultados
- reportar as conclusões obtidas
- incluir alguma discussão sobre os resultados obtidos e descrever outra eventual informação adicional que poderia contribuir para melhores resultados ou resultados mais relevantes

3. Relatório

O grupo deverá produzir um relatório, bem estruturado e informativo, que descreve o trabalho realizado. O relatório deverá ser entregue em **formato PDF**, não deverá ter mais do que 25 páginas em formato A4 (sem contar com bibliografia e eventuais anexos) e deverá incluir pelo menos os seguintes items:

- *título*;
- *resumo* - breve explicação do trabalho desenvolvido, devendo mencionar os dados utilizados, as tarefas realizadas, principais resultados obtidos e principais conclusões;
- *introdução* - contextualizar o trabalho, identificar os dados utilizados, apresentar os objetivos, e explicar a metodologia;
- *caracterização dos dados utilizados* - Pode incluir uma análise exploratória aos dados.
- *tarefas desenvolvidas e procedimentos efetuados*, para cada uma das questões (ou tarefas), deve enumerar os recursos utilizados, as opções tomadas e os resultados obtidos;
- *conclusões* - Enumerar as principais conclusões e mencionar **trabalho futuro** que poderia ser considerado para complementar este trabalho;
- *referências bibliográficas* - incluír pelo menos 5 referências com relevância para o trabalho desenvolvido.

Tenha também em atenção que:

- deve identificar claramente quais os dados que estão a ser utilizados no trabalho. No caso de não estar a ser utilizado um dos conjuntos de dados propostos, deve incluír um link para esses dados e fazer uma breve descrição dos mesmos
- deve usar gráficos adequados para introduzir os dados e representar os resultados.
- deve explicar as várias etapas do trabalho, de forma a que o leitor/cliente consiga compreender facilmente o que foi feito.

Os seguintes aspetos de apresentação serão também avaliados

- as secções e subsecções devem ser numeradas
- as tabelas e os gráficos devem ser numerados, incluir uma descrição (caption) e devem ser sempre referidos no texto.
- deve evitar incluir tabelas e gráficos com base em *Print Screens*, pois em geral ficam com má qualidade.
- o **tamanho da letra das tabelas e dos gráficos deve permitir uma leitura confortável**. Isto significa que não deve adotar um tamanho de letra muito menor do que o tamanho da letra do texto do documento.
- os gráficos e tabelas não devem ocupar as margens do documento
- no caso do relatório incluir tabelas exaustivas, pouco relevantes para a compreensão do trabalho, estas devem ser colocadas numa secção final do documento, designada por Anexos

4. Apresentação

O grupo deverá preparar uma apresentação para fazer durante a aula. A apresentação deverá ser projetada para um máximo de 10 minutos, devendo identificar claramente os dados utilizados e os objetivos do trabalho (questões colocadas). Para cada uma das tarefas, deve indicar: *os dados utilizados*, *as abordagens adotadas* e *opções mais relevantes*, os *resultados obtidos* e propostas para *trabalho futuro*.

Instruções de submissão

O trabalho deverá ser entregue através do moodle, usando o respetivo grupo. Deverá ser enviado um ficheiro ZIP contendo o relatório, o código desenvolvido e a apresentação. Poderão ser realizadas várias submissões, no entanto tenham em conta que a submissão mais recente substitui a anterior.

Notem que os **dados não devem ser submetidos juntamente com o relatório**. No caso de não estar a usar um dos conjuntos propostos pela equipa docente, o relatório deverá incluir os respetivos links para acesso aos dados.

Agradecimentos

Agradecemos à Autoridade Nacional de Segurança Rodoviária a disponibilização dos dados ao abrigo do protocolo integrado na missão Inteligência Artificial para a Administração Pública (IA>AP).

Política em caso de fraude

Os alunos podem partilhar e/ou trocar ideias entre si sobre os trabalhos e/ou resolução dos mesmos. No entanto, o trabalho entregue deve corresponder ao esforço individual de cada grupo. São consideradas fraudes as seguintes situações:

- Trabalho parcialmente copiado
- Facilitar a copia através da partilha de ficheiros
- Utilizar material alheio sem referir a sua fonte.

Em caso de deteção de algum tipo de fraude, os trabalhos em questão não serão avaliados, sendo enviados à Comissão Pedagógica ou ao Conselho Pedagógico, consoante a gravidade da situação, que decidirão a sanção a aplicar aos alunos envolvidos. Serão utilizadas as ferramentas Moss e SafeAssign para detecção automática de cópias.

Recorda-se ainda que o Anexo I do Código de Conduta Académica, publicado a a 25 de Janeiro de 2016 em Diário da Republica, 2ª Série, nº 16, indica no seu ponto 2 que:

> Quando um trabalho ou outro elemento de avaliação apresentar um nível de coincidência elevado com outros trabalhos (percentagem de coincidência com outras fontes reportada no relatório que o referido software produz), cabe ao docente da UC, orientador ou a qualquer elemento do júri, após a análise qualitativa desse relatório, e em caso de se confirmar a suspeita de plágio, desencadear o respetivo procedimento disciplinar, de acordo com o Regulamento Disciplinar de Discentes do ISCTE - Instituto Universitário de Lisboa, aprovado pela deliberação n.o 2246/2010, de 6 de dezembro.

O ponto 2.1 desse mesmo anexo indica ainda que:

> No âmbito do Regulamento Disciplinar de Discentes do ISCTE-IUL, são definidas as sanções disciplinares aplicáveis e os seus efeitos, podendo estas variar entre a advertência e a interdição da frequência de atividades escolares no ISCTE-IUL até cinco anos.