Boa tarde, eu sou o João do grupo 05 e vou apresentar o nosso trabalho prático

INTRODUCAO

No âmbito da unidade curricular de ADI, devolvemos um projeto que envolvia a conceção dos diferentes modelos de aprendizagem estudados durante o semestre para dois conjuntos de dados. Para esse efeito, foi preciso aplicar técnicas de exploração e preparação de dados, assim como aplicar os modelos corretamente, dado o conhecimento extraído dos conjuntos de dados.  
  
METODOLOGIA  
Antes de avançar com o problema, é uma excelente prática definir uma metodologia.

No nosso caso, optamos por seguir a CRISP-DM, pois é uma configuração versátil, sendo apropriada para casos onde os analistas não são muito conhecedores do domínio.

A versão que utilizamos foi adaptada, pois fases como o deployment não eram uteis para os objetivos do projeto.

ESTUDO DE NEGÓCIO  
O estudo de negócio é precisamente a compreensão dos objetivos do projeto a realizar. Pelo repositório de onde os dados provêm, sabemos que o objetivo é classificar um paciente como doador de sangue ou doente contaminado com Hepatite C. No caso de estar infetado, são adicionadas novas categorias, dependendo do progresso da doença.

ESTUDO DE DADOS  
(Ler pp)

Por questões de tempo, não vamos conseguir expor o estudo feito para cada atributo, por isso fizemos uma seleção que consideramos relevante para apresentação.  
  
AGE

(ler pp)  
Pois é relevante perceber como é que as idades estão distribuídas na nossa amostra. De forma análoga, observamos também o histograma dos anos de nascimento

(mostrar histogramas na apresentação)

Aqui vemos que os diagramas são semelhantes, mas não equivalentes, então decidimos ver se a idade está concordante com a data de nascimento.

Facilmente vimos que não havia concordância, mas também é importante destacar que existe uma correlação inversa perfeita entre estes atributos, o que nos leva acreditar que as idades que estão no dataset são as idades da altura da recolha.  
  
SEX  
Para esta feature, (ler pp). Esta atribuição tanto pode ser uma falha do sistema, como pode ser o resultado de pacientes que não queiram partilhar os seus dados. Do nosso entendimento, o mais provável é ter sido um erro de escrita.

ALT  
Ainda dentro dos valores numéricos, temos a Alanina transaminase. Neste valor, para além das estatísticas, decidimos ver a relação em gráfico de barras com o nosso atributo alvo.

Da figura retiramos que os valores variam entre 0.9 e 325.3, com média de 28.451 e desvio padrão de 25.470. Os valores para um individuo saudável são geralmente inferiores a 35 g/L, logo há casos demasiados extremos fora do comum.  
  
PREPARACAO DE DADOS  
Nesta etapa, vamos referir algumas transformações que realizamos sobre os dados. Para dar continuidade à apresentação, optamos por focar-nos mais nos exemplos que têm sido mostrados.

(idade e data) (ler pp)

(Sex)(ler pp)

(elementos numéricos) (ler pp)

É de realçar que também existem outliers que, no contexto da medicina, podem ser bastante importantes, pois é esperado que alguém que esteja doente não tenha os valores dentro da referência por isso (ler pp)

MODELACAO

A título da apresentação, vamos focar mais nos modelos de classificação, regressão logística e clustering, deixando os restantes para o próximo dataset.

MODELOS DE CLASSIFICACAO

O objetivo desta experiencia era ver como o tratamento de valores em falta afetava o desempenho.

(ler pp)

REGRESSÃO LOGISTICA COM ALVO BINARIO

(ler pp)

CLUSTERING  
Precisamos de balancear o dataset, a performance ficou muito atrás dos algoritmos anteriores.  
  
(ler pp) o que pode levar a dados viciados.

ESTUDO DE NEGOCIOS

Para o nosso Dataset escolhemos um que (ler pp)

ESTUDO DE DADOS

(ler pp)

de forma semelhante ao dataset anterior, vamos focar-nos em atributos que consideramos mais intressantes.

YEAR

Começando pelo ano de montagem (ler pp)

Pelo seguinte gráfico, podemos perceber que (ler pp)

CILINDROS

Decidimos também analisar os cilindros porque (ler pp)

Do gráfico conseguimos perceber que o número de cilindros varia entre 3 e 12, com dois casos onde não é especificado. Interpretamos que este caso “other” seria principalmente para casos ondes os veículos eram elétricos ou outro tipo de motor não usual.

ODOMETER

Analisando a distância feita por um veículo, a primeira observação que deparamos ao analisar os dados é que existem (ler pp), que ultrapassam muitas vezes o valor médio esperado por um veículo.

Do gráfico retiramos que (ler pp)

PRECO

Finalmente, é muito importante (ler pp)

Da tabela conseguimos perceber que existem valores que não são realistas por isso removemos estes valores

Um aspeto importante de referir é o valor “0”. O que o grupo entendeu,que há dois casos possíveis para o colocar: (ler pp)

PREPARAÇÃO DE DADOS

Year: (ler pp) porque nao nos pareciam corretas

Cylinders: (ler pp)

Odometer: (ler pp)

Curiosamente, o próprio preço precisa de um pequeno ajuste. (ler pp)

MODELACAO

De forma semelhante ao dataset anterior para este conjunto de dados, vamos salientar (ler os que estão a negrito)

MODELOS DE REGRESSÃO

Os resultados a nosso ver são positivos porque houve (ler pp).

Dado que os anúncios são colocados por particulares os seus anúncios são especulativos e (ler pp)

RNA  
Os resultados das redes neurais não atingiram as expectativas estabelecidas Este desempenho insatisfatório foi observado em vários aspetos-chave do modelo, revelando não serem as melhores técnicas de previsão para este problema. (ler pp)

Para nossa surpresa a (ler pp)

MODELACAO SEM OUTLIERS

Relativamente à modelação sem outliers à primeira vista, podemos logo afirmar que (ler pp)

No caso dos modelos de regressão, esta técnica teve (ler pp). No caso da (ler pp)

CONCLUSÃO

Para terminar, deste trabalho, podemos afirmar que o grupo ficou consciencializado (ler pp).

Vimos também a (ler pp).

Apercebemos nos que a própria modelação não é uma ciência exata e requer bastante paciência e diversas tentativas para alcançar os melhores resultados possíveis.

(ler pp)

Refletindo à cerca do que foi feito para cada conjunto de dados, sentimos nos realizados com ambos, mas, (ler pp), achamos que certas situações poderiam ter sido mais bem exploradas.