Curso: Ciência de Dados

Disciplina: Ciência de Dados

Relatório de Projeto – Descrição do Conteúdo

A primeira etapa consiste na escolha da base de dados a ser utilizada. Existem dois repositórios clássicos que disponibilizam dados reais para usos em experimentos:

a) UCI Machine Learning Repository (“https://archive.ics.uci.edu/ml/datasets.php”): contém centenas de bases de dados reais usados em muitos experimentos. Algumas destas bases de dados são de menor tamanho (menos de 100 exemplos ou instânicas) e outras são maiores (mais de 1.000 exemplos ou instâncias); são bases adequadas para um primeiro contato e experiências na área;

b) Kaggle Repository (“https://www.kaggle.com/datasets”): contém uma grande quantidade de bases de dados reais, em sua maioria apresentadas por empresas em busca de soluções e/ou competições. Em geral as bases de dados disponíveis são grandes e apresentam problemas específicos para os quais se busca uma solução. Em função disto, é necessário um maior conhecimento prévio da área para trabalho com estas bases de dados.

Tendo a base escolhida, os seguintes aspectos devem ser apresentados no Relatório:

1. Descrição da Base de Dados:

* Indicar o nome da base;
* Que tipo de dado (ou problema) caracteriza esta base de dados, ou seja, o que se pretende extrair de conhecimento a partir dos dados lá disponíveis (classificação/regressão, agrupamento, associação, por exemplo);
* Indicar quem forneceu esta base, além da data da coleta dos dados (início e final), além de particularidades (por exemplo, a coleta foi interrompida durante um período de tempo e depois retomada);
* Quantos atributos (variáveis) possui e para cada um o tipo de atributo: nominal, real, discreto, além dos valores (ou faixa de valores);
* Indicar se existem informações já publicadas sobre seu uso e, em caso positivo, onde foi publicado, por quem, o quê;
* Quantos exemplos existem para cada classe, no caso de classificação, ou a faixa de valores da variável a ser prevista no caso de regressão.

2. Análise das principais estatísticas da base de dados:

* Para cada um dos atributos (variáveis) do banco de dados (ou para os cinco primeiros atributos) fornecer:
  1. O valor máximo e o valor mínimo do atributo, caso seja um atributo numérico, ou o conjunto de valores possíveis em caso contrário;
  2. O valor médio do atributo, caso seja um atributo numérico;
  3. A mediana do atributo, caso seja um atributo numérico;
  4. O desvio padrão do atributo, caso sejam um atributo numérico;
  5. Os percentis 0%, 25%, 50%, 75% e 100% no caso de atributos numéricos;
  6. A moda do atributo (se existir);
  7. A curtose e a obliquidade do atributo, caso seja numérico;
  8. Quantas instâncias (exemplos) existem para cada classe, no caso de classificação;
  9. Se o problema for de regressão e for transformado em classificação (de acordo com os percentis 0%, 25%, 50%, 75% e 100%) quantos exemplos tem-se para cada classe.
* Para cada par de atributos (ou um subconjunto deles, no caso de serem muitos), indique:
  1. A covariância entre os atributos;
  2. A correlação entre os atributos;
  3. O coeficiente de correlação de Spearman;
  4. O coeficiente de correlação de Pearson.

3. Transformação da base para adequação aos algoritmos posteriores:

* Crie uma nova base de dados a partir da original, em que os valores das variáveis contínuas estejam normalizados pelo método da:
  1. Normalização mín-máx;
  2. Normalização pelo desvio padrão;
* Existem dados faltantes na base?
  1. Em caso positivo, em quais atributos e qual o percentual?
  2. Em caso negativo, simule alguns valores faltantes (retire da base original) e encontre-os;
* Substitua os valores faltantes (ou que foram retirados), pelo valor médio do atributo, considerando os valores restantes. Feito isto, determine qual a nova média e novo desvio padrão de cada um destes atributos. Determine também o erro que está sendo cometido nos valores substituídos;
* Refaça o item anterior, substituindo os valores faltantes pelo valor que mantém o desvio padrão constante;
* Considere que neste projeto, um outlier é aquele valor que está a pelo menos uma distância de dois desvios-padrão da média.
  1. Com esta definição, existem outliers na base?
  2. Em caso positivo, com a sua retirada, há mudança significativa no valor médio e desvio padrão dos atributos?
* Realize uma discretização não supervisionada dos valores contínuos usando intervalos de mesmo tamanho – justifique o valor escolhido.

4. Visualização de dados:

* Baseado na sua base de dados, aplique as técnicas de visualização que achar apropriadas e indique que informações podem ser extraídas a partir das mesmas.
* Pesquise que bibliotecas permitem usar coordenadas paralelas em Python e aplique aos seus dados. Que conclusões ela permite obter?