PUCRS
Pontifícia Universidade Católica do Rio Grande do Sul

Coleta, Preparação e Análise de Dados

Prof. Lucas R. C. Pessutto

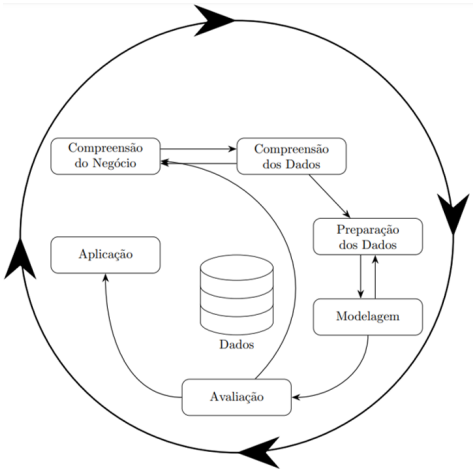
Turma 30

Prova

Aluno(a): _____ Matrícula: _____

1 ponto

1. O CRISP-DM é um processo padronizado de descoberta de conhecimento em grandes conjuntos de dados que consiste de seis fases, conforme mostrado na figura abaixo:



No que consiste a etapa de **compreensão dos dados**? Quais tarefas ela engloba?

1 ponto

2. Qual a finalidade de se colocar o arquivo `robots.txt` no diretório raiz de um site?

1 ponto

3. Existem quatro problemas que afetam a qualidade dos dados: ruídos, outliers, valores faltantes e dados duplicados. Forneça uma explicação de cada um desses problemas.

1 ponto

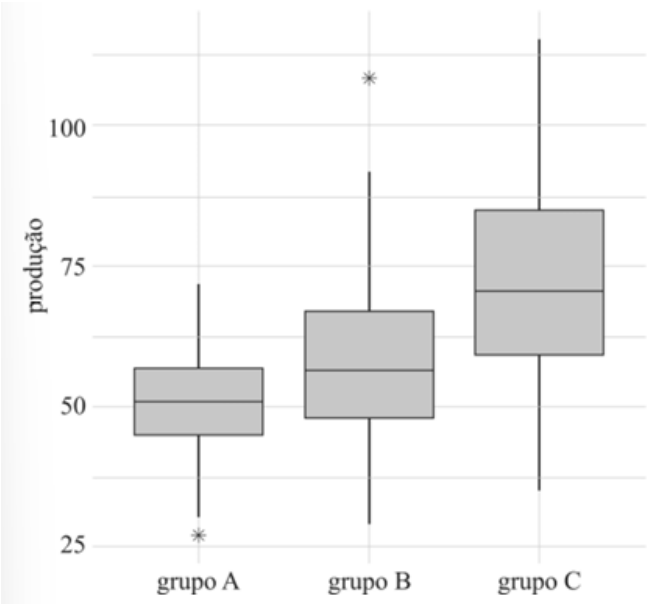
4. Descreva o que está envolvido no processo de **ETL** (Extraction, Transformation and Loading) no início do processo de um Data Warehousing.

1 ponto

5. Relacione as técnicas de amostragem com a sua respectiva descrição:

(1)	Amostragem aleatória com resposição	( )	Os dados são agrupados e após isso um dos grupos é selecionado para ser a amostra
(2)	Amostragem aleatória sem resposição	( )	Amostragem por sorteio, sem repetição
(3)	Amostragem estratificada	( )	os elementos são escolhidos segundo um sistema de escolhas preestabelecido
(4)	Amostragem sistemática	( )	Os valores da amostra mantém a proporção de uma das colunas
(5)	Amostragem por cluster	( )	Feita por sorteio com repetição nos valores

1 ponto 6. Considerando o gráfico boxplot abaixo, assinale a alternativa correta sobre ele:



A. somente o grupo B possui outliers

B. existem pessoas no grupo A com produção inferior a 25

C. o terceiro quartil do grupo A tem valor 80

D. Pelo menos 75% dos valores que estão no grupo A são inferiores do que a mediana do grupo C

E. a mediana do grupo C possui valor superior do que o maior valor presente no grupo B

1 ponto 7. Considerando as seguintes afirmações

I. A Análise Exploratória de Dados utiliza ferramentas como gráficos de dispersão, correlações e histogramas para descobrir padrões, relações ocultas e tendências importantes nos dados, como sazonalidade e correlação entre variáveis.

II. Em um conjunto de dados sobre renda familiar, um valor extremamente alto pode representar uma pessoa com renda muito superior à média, e, portanto, deve sempre ser removido, independentemente de sua relevância para a análise.

III. A imputação de valores ausentes com a média ou mediana é uma técnica comum para lidar com dados ausentes em análise de dados. Entretanto, ela pode distorcer a distribuição dos dados se a quantidade de dados ausentes for grande.

Quais delas estão corretas?

A. Apenas I

B. Apenas II

C. I e III

D. II e III

E. I, II e III

1 ponto 8. Por que o Selenium é uma alternativa para webscraping considerando páginas web dinâmicas?

A. Pois o Selenium simula o uso de um navegador web por um usuário normal

B. O Selenium tem uma interface de programação mais moderna, portanto facilita o desenvolvimento

C. O Selenium possui suporte a mais versões do python do que outras ferramentas

D. O selenium permite uso de regex

E. Pois ele consegue desativar recursos de Javascript das páginas web

1 ponto 9. Considerando as seguintes afirmações

I. BeautifulSoup é uma biblioteca Python que permite a realização de requisições HTTP na web, mas requer o uso de outros frameworks, como por exemplo Selenium, para poder tratar os resultados destas requisições.

II. DOM (Document Object Model) é uma interface independente de linguagem que permite tratar documentos HTML ou XML em estrutura do tipo árvore, onde cada nodo da árvore representa um elemento do documento.

III. Web scraping é uma técnica para coletar na web todos os links que são alcançáveis a partir de um site, sem se importar com o conteúdo dos recursos acessíveis por estes links.

Quais delas estão corretas?

A. Apenas I

B. Apenas II

C. I e III

D. II e III

E. I, II e III

1 ponto 10. Com relação ao Data Warehousing avalie as afirmativas a seguir e assinale V para a afirmativa verdadeira e F para a falsa.

( ) Em comparação com os bancos de dados transacionais, os Data Warehouses são tidos como voláteis. Isso significa que as informações no Data Warehouse mudam com muito mais frequência e podem ser considerados de tempo real.

( ) Para um data warehouse com grande volume de dados históricos e consultas frequentes, o esquema estrela geralmente apresenta melhor desempenho que o esquema snowflake, pois minimiza a complexidade das junções entre tabelas.

( ) Um Data Wharehouse visa apoiar e permitir análises de negócios e tomadas de decisões mais eficazes por parte dos trabalhadores do conhecimento.

As afirmativas são, respectivamente,

A. F – F – F

B. F – F – V

C. F – V – V

D. V – F – V

E. V – V – V