PROGRAMA DE RESIDÊNCIA EM TECNOLOGIA DA INFORMAÇÃO – TRE-RN
EDITAL 001/2023 – PROVA DE CONHECIMENTOS ESPECÍFICOS
ÁREA DE CONCENTRAÇÃO 1 – BUSINESS INTELLIGENCE E ANALYTICS

Candidato: _____
CPF: _____ Telefone: _____

QUESTÕES OBJETIVAS

01. A seguir é apresentado um trecho de código escrito na linguagem Python.

```
x = [2,3,1,5,4]
for i in range(len(x)-1):
      x[i],x[i+1] = x[i+1],x[i]
print(x)
```

O valor da variável **x** após a execução desse trecho de código será:

A) `[2,3,1,5,4]`
B) `[3,1,5,4,2]`
C) `[3,4,1,2,5]`
D) `[4,2,5,3,1]`

02. Suponha que para resolver um determinado problema, você aplicou quatro técnicas de classificação, sendo: AD, k-NN, Naive Bayes e MLP. Ao comparar o resultado dos classificadores, você percebeu que o Naive Bayes teve uma taxa de acerto bem menor que as taxas dos outros classificadores. O que é possível afirmar sobre o conjunto de treinamento:

A) Possui os atributos dependentes e/ou os atributos numéricos obedecem a uma distribuição normal.
B) Possui os atributos independentes e/ou os atributos numéricos obedecem a uma distribuição normal.
C) Possui os atributos dependentes e/ou os atributos numéricos não obedecem a uma distribuição normal.
D) Possui os atributos independentes e/ou os atributos numéricos não obedecem a uma distribuição normal.

03. Seja a seguinte sequência de operações da Álgebra Relacional:

$$\pi_{A1,A2} (\sigma_{A1=5} (A \bowtie_{A1=B3} B))$$

Considerando essa sequência da esquerda para a direita, que operações foram empregadas?

A) Junção, seleção e projeção
B) Seleção, projeção e junção
C) Seleção, junção e projeção
D) Projeção, seleção e junção

04. Considere o esquema relacional abaixo, no qual **placa** é a chave primária.

VEICULO(Placa, Cor, Modelo, Marca, Ano, Valor)

Qual é a expressão em álgebra relacional a ser aplicada sobre esse esquema, de forma a obter as Placas dos VEICULOS com Ano igual a 2011 e Valor menor que 9000?

A) πPlaca (πAno = 2011; πValor < 9000 (VEICULO))
B) σPlaca (σAno = 2011; Valor < 9000)
C) πPlaca (σAno = 2011 AND Valor < 9000 (VEICULO))
D) σPlaca (πValor < 9000 AND Ano = 2011 (VEICULO))

05. Uma *black box* é como uma caixa opaca onde você não consegue observar o seu interior. No contexto de Aprendizado de Máquina, na *black box* você pode visualizar entradas e saídas, mas não o seu funcionamento. Qual dos algoritmos abaixo é considerado uma *black box*?

A) k-means
B) k-NN
C) Naive Bayes
D) MLP

06. Em um SGBDR, se ocupa respectivamente de alteração de tabela (ex. *Alter Tabl*e), atualização de uma linha de tabela (ex. *Update*) e exclusão de visão (ex. *Drop View*), a:

A) DDL, DML e DDL
B) DDL, DDL e DML
C) DML, DDL e DML
D) DML, DML e DDL

07. Duas tabelas de página são mantidas durante a vida de uma transação: a tabela de página atual e a tabela de página cópia. Quando a transação inicia, as duas tabelas são idênticas. A tabela de página cópia nunca é alterada durante a execução da transação. A tabela de página atual é alterada quando a transação processa uma operação de escrita. Quando a transação é parcialmente efetivada, a tabela de página cópia é descartada e a tabela de página atual torna-se a nova tabela de página. Se a transação for abortada, a tabela de página atual é descartada. Qual é a técnica de recuperação do banco de dados em caso da falha descrita acima?

A) Recuperação adiada
B) Paginação de sombra (*shadow*)
C) Recuperação baseada em Log
D) Recuperação imediata

08. O Power BI é uma coleção de serviços de *software*, aplicativos e conectores que trabalham juntos para transformar suas fontes de dados não relacionadas em informações coerentes. Sobre o Power BI, assinale a afirmativa correta.

A) *Deplaymente* pipeline é um recurso disponível no Power BI Desktop que permite testar relatórios antes do lançamento para os usuários.
B) *Paginated Reports* são relatórios criados no Power Bi Service para serem exibidos em dispositivos com limitação de memória, como celulares e tablets.

C) *Power Query Editor* permite conectar a uma ampla variedade de tipos de fontes de dados, porém, é necessário usar a linguagem OQL para acesso aos dados
D) *Direct Query* permite criar visualizações de conjuntos de dados muito grandes, nos casos em que seria impraticável importar todos os dados com pré-agregação.

09. Com base nos sistemas de banco de dados NoSQL, assinale a alternativa que correlaciona corretamente os SGBD's no NoSQL e seus modelos estruturais.

A) Cassandra: Modelo Orientado a Colunas – Neo4J: Modelo Baseado em Grafos – MongoDB: Modelo Orientado a Documentos – Redis: Modelo Chave-Valor.
B) MongoDB: Modelo Orientado a Colunas – Cassandra: Modelo Baseado em Grafos – Neo4J: Modelo Orientado a Documentos – Redis: Modelo Chave-Valor.
C) Cassandra: Modelo Orientado a Colunas – Neo4J: Modelo Baseado em Grafos – Redis: Modelo Orientado a Documentos – MongoDB: Modelo Chave-Valor.
D) Redis: Modelo Orientado a Colunas – Cassandra: Modelo Baseado em Grafos – MongoDB: Modelo Orientado a Documentos – Neo4J: Modelo Chave-Valor.

10. Em Ciência de Dados, Python é uma das linguagens de programação mais utilizadas. A esse respeito, é correto afirmar que a linguagem de programação Python:

A) Mostra-se ideal para desenvolvimento rápido e criação de *scripts* em razão de sua natureza compilada.
B) Foi desenvolvida com o intuito de substituir a linguagem de programação C por causa de sua altíssima performance.
C) Pode ser utilizada como uma linguagem de programação funcional.
D) Possui recursos para controle de fluxo, como *if-else*, *switch-case*, *while* e *for* em todas as suas versões.

11. Considere o seguinte código Python:

```
df = pd.DataFrame({'A':[1,2,np.nan],
                   'B':[5,np.nan,np.nan],
                   'C':[1,2,3]})
```

Out[3]:

	A	B	C
0	1.0	5.0	1
1	2.0	NaN	2
2	NaN	NaN	3

Qual das linhas de código abaixo é responsável por excluir todos os dados ausentes presentes no DataFrame?

A) `df.dropna(thresh=2)`
B) `df.dropna()`
C) `df.dropna(axis=1)`
D) `df.dropna(axis=2)`

12. No contexto *software versus* ser humano, existem duas áreas em evolução: *User Experience Design* (UX), que se preocupa com o ponto de contato de um produto/serviço com as pessoas, e *User Interface* (UI). Apesar de estarem relacionadas, essas áreas são muito diferentes. No caso da UI, o primeiro ponto de avaliação do usuário e a maior influência para isso são:

A) Robustez e Eficácia.
B) Eficácia e Familiaridade.
C) Apresentação e Robustez.
D) Clareza e Eficácia.

13. Analise as seguintes afirmativas:

I. *K-means* é um algoritmo de aprendizado não supervisionado, em que se calcula a distância entre os objetos da base e cada um dos centroides; em que se atribui cada objeto ao centroide mais próximo;
II. K-NN é um método de *Clustering*. Uma vez que os agrupamentos e os seus centroides são identificados, é fácil atribuir novos objetivos para um cluster baseado na distância do objeto do centroide mais próximo.
III. O CRISP-DM é uma metodologia abrangente de mineração de dados e um modelo de processo que fornece, para os usuários de data mining (DM), um modelo completo para a realização de um projeto de DM.

Podemos afirmar corretamente que:

A) Todas as afirmativas estão corretas.
B) Apenas as afirmativas I e II estão corretas.
C) Apenas as afirmativas I e III estão corretas.
D) Apenas as afirmativas II e III estão corretas.

14. _____ é uma classe de métodos *ensemble* que utiliza classificadores de árvore de decisão. É uma combinação de preditores de árvores tal que cada árvore depende dos valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores.

Assinale a alternativa que preenche corretamente a lacuna do trecho acima:

A) Support Vector Machine (SVM)
B) k-Nearest Neighbors (k-NN)
C) Regressão logística
D) Random Forest

15. Considere o seguinte código Python:

```
import pandas as pd

data = {'Empresa':['GOOG','GOOG','MSFT','MSFT','FB','FB'],
        'Nome':['Sam','Charlie','Amy','Vanessa','Carl','Sarah'],
        'Venda':[200,120,340,124,243,350]}

df = pd.DataFrame(data)

por_companhia = df.groupby("Empresa")
```

Qual das linhas de código abaixo é responsável por permitir a seguinte saída (*output*)?

Out[20]:

Empresa	Nome	Venda
FB	Sarah	350
GOOG	Sam	200
MSFT	Vanessa	340

A) `por_companhia.max()`
B) `por_companhia.min()`
C) `por_companhia.mean()`
D) `por_companhia.std()`

16. A respeito das bibliotecas **NumPy** e **Pandas**, analise as seguintes afirmativas:

I. O método describe() da biblioteca Pandas retorna as linhas superiores e inferiores do DataFrame;
II. A biblioteca Pandas apresenta os dados em uma estrutura de DataFrame, composta por linhas e colunas;
III. A classe `numpy.poly1d()` permite a criação de arrays multidimensionais.

Podemos afirmar corretamente que:

A) Todas as afirmativas estão corretas.
B) Apenas a afirmativas I e II estão corretas.
C) Apenas a afirmativa II está correta.
D) Apenas as afirmativas II e III estão corretas.

17. São exemplos de banco de dados relacional e banco de dados não relacional, respectivamente:

A) Cassandra e PostgreSQL.
B) MySQL e MongoDB.
C) Redis e MongoDB.
D) MS Access e MySQL.

18. Por meio da ferramenta *Pentaho Data Integration* (PDI), é possível realizar processos de ETL (*Extraction, Transformation and Load*). A ferramenta gráfica que faz parte do PDI, usada para modelar as transformations e os jobs, é denominada de:

A) spoon.
B) kettle.
C) weka.
D) mondrian.

19. Sobre o Esquema Estrela (*Star Schema*) de um *Data Warehouse*, é **INCORRETO** afirmar que:

A) a dimensão descreve os dados que estão nos fatos, ou seja, as chaves.
B) o Esquema Estrela é composto por tabelas do tipo fato e dimensão.
C) a tabela fato é ligada sempre a duas ou mais dimensões.
D) a granularidade, também chamada de detalhe, é o menor nível da hierarquia da dimensão.

20. A modelagem de *Data Warehouses* pode ser feita seguindo diferentes esquemas. Sobre esse tópico, analise as afirmativas:

I. No esquema estrela, os dados são organizados em uma tabela dimensão e muitas tabelas fatos;
II. O esquema floco de neve é uma variação do esquema estrela, onde algumas tabelas fatos são normalizadas, dividindo, assim, os dados em tabelas adicionais;
III. Quando várias tabelas fatos compartilham as mesmas tabelas dimensões, temos nesse caso, uma constelação de fatos.

Podemos afirmar corretamente que:

A) somente as afirmativas I e II estão corretas.
B) somente as afirmativas II e III estão corretas.
C) somente a afirmativa II está correta.
D) somente a afirmativa III está correta.

21. *Big Data* surgiu a partir da necessidade de manipular um grande volume de dados e, com isso, novos conceitos foram introduzidos, como o D*ata Lake*, que:

A) pode ser considerado um conjunto de bancos de dados relacionais e com relacionamentos entre tabelas de diferentes esquemas de bancos de dados.
B) é projetado para armazenar dados de diversas fontes e formatos, não havendo a necessidade da definição de um esquema de dados para inserir novos itens.
C) pode ser considerado um repositório de dados relacionados, sendo, portanto, um armazém de dados orientado por assunto.
D) é o resultado de sucessivas operações de mineração de dados, sendo um ambiente no qual é possível ter relatórios e dashboards de maneira amigável para os analistas de negócio.

22. Qual é a principal finalidade do *Data Lake* no contexto de análise de dados?

A) Armazenar grandes volumes de dados não estruturados.
B) Realizar análises estatísticas avançadas.
C) Facilitar a geração de relatórios em tempo real.
D) Servir como um banco de dados transacional.

23. Qual técnica é usada para segmentar clientes em grupos com base em características semelhantes?

A) Regressão linear.
B) Classificação.
C) Associação.
D) *Clustering*.

24. Dentro do contexto de Mineração de Dados, podemos dizer afirmar:

I. A limpeza dos dados é uma das etapas do processo.
II. O treinamento é uma parte fundamental para efetuar a limpeza dos dados.
III. A busca por padrões consiste em um dos possíveis objetivos da Mineração de Dados.
IV. O escalonamento dos valores é uma etapa possível na preparação dos dados.
Sobre as afirmativas anteriores, é correto dizer que:

A) Somente a alternativa I está correta.
B) Estão corretas somente as afirmativas III e IV
C) Estão incorretas as afirmativas II e IV.
D) Somente a alternativa II está incorreta.

25. Qual é a diferença entre aprendizado supervisionado e não supervisionado?

A) No aprendizado supervisionado, os resultados são conhecidos e usados para treinamento, enquanto no não supervisionado, os resultados não são conhecidos antecipadamente.
B) No aprendizado supervisionado, não há dados de entrada, enquanto no não supervisionado os resultados são conhecidos.
C) No aprendizado supervisionado, não há resultados conhecidos, enquanto no não supervisionada, os resultados são conhecidos.
D) Não há diferença entre elas.

26. Qual cláusula SQL é usada para combinar linhas de duas ou mais tabelas?

A) LINK
B) COMBINE
C) JOIN
D) MERGE

27. Qual cláusula SQL é usada para agrupar dados em uma consulta?

A) SORT BY
B) CATEGORIZE
C) ASSEMBLE
D) GROUP BY

28. Qual cláusula SQL é usada para limitar o número de registros retornados por uma consulta?

A) TOP
B) LIMIT
C) RESTRICT
D) COUNT

29. Qual é o resultado da seguinte consulta SQL, que envolve junção de tabelas?

```
SELECT clientes.nome, COUNT(pedidos.id)
FROM clientes
LEFT JOIN pedidos ON clientes.id = pedidos.cliente_id
GROUP BY clientes.nome;
```

A) Retorna o nome de cada cliente e o número de pedidos feitos por aquele cliente.
B) Retorna o nome de cada cliente e o número total de pedidos em toda a tabela de pedidos.
C) Retorna um erro, pois não é possível fazer uma junção entre essas tabelas.
D) Retorna o nome de cada cliente e a média de pedidos feitos por cliente.

30. Qual é o resultado da seguinte consulta SQL, que usa a função RANK()?

```
SELECT nome, salario,
RANK() OVER (ORDER BY salario DESC) AS ranking
FROM funcionarios;
```

A) Retorna o nome e o salário de cada funcionário, mas não inclui informações de classificação.
B) Retorna um erro, pois as funções de janela não são suportadas em banco de dados relacionais.
C) Retorna o nome, o salário e a classificação de cada funcionário com base no salário, em ordem decrescente.
D) Retorna o nome e o salário de cada funcionário, em ordem alfabética decrescente.

31. Qual é o resultado da seguinte consulta SQL, que usa uma subconsulta correlacionada?

```
SELECT nome, salario FROM funcionarios AS f
WHERE salario > (SELECT AVG(salario) FROM funcionarios WHERE departamento = f.departamento);
```

A) Retorna os nomes e salários de todos os funcionários.
B) Retorna os nomes e salários dos funcionários cujos salários são maiores do que a média de seus colegas no mesmo departamento.
C) Retorna um erro, pois subconsultas correlacionadas não são suportadas.
D) Retorna os nomes e salários dos funcionários cujos salários são menores do que a média de seus colegas no mesmo departamento.

32. O objetivo é obter a lista dos 3 funcionários com os maiores salários na tabela "funcionarios". Qual código SQL atinge esse objetivo?

A) SELECT nome, salario FROM funcionarios ORDER BY salario DESC OFFSET 3 FETCH NEXT 3 ROWS ONLY;
B) SELECT nome, salario FROM funcionarios ORDER BY salario ASC FETCH FIRST 3 ROWS ONLY;
C) SELECT nome, salario FROM funcionarios ORDER BY salario DESC LIMIT 3;
D) SELECT nome, salario FROM funcionarios ORDER BY salario ASC LIMIT 3;

33. Você deseja encontrar a quantidade de produtos em cada categoria na tabela "produtos". Qual código SQL atinge esse objetivo?

A) SELECT categoria, SUM(quantidade) FROM produtos GROUP BY categoria;
B) SELECT categoria, AVG(preco) FROM produtos GROUP BY categoria;
C) SELECT categoria, MAX(quantidade) FROM produtos GROUP BY categoria;
D) SELECT categoria, COUNT(*) FROM produtos GROUP BY categoria;

34. Você deseja criar uma nova tabela chamada "estoque_backup" que seja uma cópia exata da tabela "estoque". Qual código SQL realiza essa tarefa?

A) COPY estoque TO estoque_backup;
B) CREATE TABLE estoque_backup AS SELECT * FROM estoque;
C) INSERT INTO estoque_backup SELECT * FROM estoque;
D) BACKUP TABLE estoque TO estoque_backup;

35. Qual das seguintes afirmações é verdadeira sobre bancos de dados relacionais?

A) Eles são especialmente adequados para armazenar dados semiestruturados.
B) Usam tabelas para armazenar dados e relacionamentos entre eles.
C) Não requerem linguagem de consulta.
D) São altamente escaláveis horizontalmente.

36. O que é uma chave primária em um banco de dados relacional?

A) Uma chave usada para estabelecer relacionamentos entre tabelas.
B) Uma chave usada para acessar o sistema de gerenciamento de banco de dados.
C) Uma chave usada para identificar unicamente cada registro em uma tabela.
D) Uma chave usada para criptografar dados sensíveis.

37. O que é normalização em bancos de dados relacionais?

A) Um processo de adição de redundância aos dados para melhorar o desempenho.
B) Um processo de remoção de redundância dos dados para melhorar a integridade.
C) Um processo de criptografia de dados sensíveis.
D) Um processo de conversão de dados em formato JSON.

38. O que é consistência eventual em bancos de dados NoSQL?

A) Todos os dados são consistentes em tempo real.
B) A consistência é alcançada eventualmente, mas não é garantida em tempo real.
C) Os dados são consistentes apenas quando não há falhas no sistema.
D) A consistência é garantida por meio de transações.

39. O que é o teorema CAP (Consistência, Disponibilidade e Tolerância a Partições) em relação a bancos de dados NoSQL?

A) É um teorema que descreve a estrutura de tabelas em bancos de dados NoSQL.
B) É um teorema que define os tipos de dados que podem ser armazenados em bancos de dados NoSQL.
C) É um teorema que descreve as limitações na capacidade de um sistema distribuído em garantir simultaneamente consistência, disponibilidade e tolerância a partições.
D) É um teorema que descreve a escalabilidade de bancos de dados NoSQL.

40. Qual é a diferença entre as listas e as tuplas em Python?

A) Listas são mutáveis, enquanto tuplas são imutáveis.
B) Listas podem conter elementos de tipos diferentes, mas as tuplas não.
C) Tuplas são usadas para iterar sobre sequências, enquanto listas são usadas para armazenar elementos.
D) Não há diferença, os termos são usados indistintamente.

41. Considere o seguinte código Python que utiliza compreensão de lista:

```
numeros = [1, 2, 3, 4, 5]
resultado = [x for x in numeros if x % 2 != 0]
```

Qual é o valor da variável resultado após a execução deste código?

A) `[2, 4]`
B) `[1, 3, 5]`
C) `[2, 4, 6]`
D) Erro de sintaxe.

42. Qual dos códigos Pyton executa o seguinte procedimento: Dada uma *string*, verificar se ela contém a palavra "Python":

A)
```
texto = "Linguagem de programação"
contem_python = texto.contains("Python")
```
B)
```
texto = "Python é uma linguagem de programação"
contem_python = "Python" in texto
```

C)
```
texto = "Python é uma linguagem de programação"
contem_python = texto.indexOf("Python") != -1
```
D)
```
texto = "Linguagem de programação Python"
contem_python = texto.index("Python")
```

43. Qual o código Python executa o seguinte procedimento: Dada uma lista de palavras, criar uma nova lista que contenha apenas as palavras com mais de 5 caracteres.

A)
```
palavras = ["python", "programação", "dados", "ciência"]
palavras_longas = [p for p in palavras if p.count() > 5]
```
B)
```
palavras = ["python", "programação", "dados", "ciência"]
palavras_longas = [p for p in palavras if p.size() > 5]
```
C)
```
palavras = ["python", "programação", "dados", "ciência"]
palavras_longas = [p for p in palavras if p.length() > 5]
```
D)
```
palavras = ["python", "programação", "dados", "ciência"]
palavras_longas = [p for p in palavras if len(p) > 5]
```

44. Qual é o principal objetivo do Apache Airflow?

A) Armazenar grandes volumes de dados.
B) Automatizar a execução de fluxos de trabalho.
C) Servir como um servidor *web*.
D) Gerenciar contêineres Docker.

45. O que é um DAG *(Directed Acyclic Graph*) no contexto do Apache Airflow?

A) Uma estrutura de dados para armazenar registros de log.
B) Um formato de arquivo usado para armazenar configurações.
C) Um tipo de banco de dados NoSQL.
D) Uma representação visual de um fluxo de trabalho com tarefas e suas dependências.

46. Qual é o objetivo principal do Metabase?

A) Armazenar grandes volumes de dados.
B) Criar *websites* dinâmicos.
C) Visualizar e analisar dados de forma simples e intuitiva.
D) Gerenciar servidores de banco de dados.

47. Qual é a finalidade dos "*cards*" no Metabase?

A) São cartões de crédito virtuais usados para pagamentos *online*.
B) São peças de um quebra-cabeça virtual.
C) São mini-relatórios ou visualizações individuais de dados.
D) São mensagens de erro do sistema.

48. Qual é o objetivo do "Power Query" no Power BI?

A) Criar cálculos avançados.
B) Criar gráficos e visualizações.
C) Importar, transformar e combinar dados de várias fontes.
D) Criar dashboards interativos.

49. Qual é a linguagem de fórmula usada para criar medidas personalizadas no Power BI?

A) JavaScript.
B) DAX (Data Analysis Expressions).
C) SQL.
D) VBA (Visual Basic for Applications)

50. Qual é o principal objetivo da IHC?

A) Desenvolver software sem interfaces de usuário.
B) Melhorar a eficiência das máquinas.
C) Projetar sistemas computacionais que sejam eficazes, eficientes e agradáveis de usar pelos usuários.
D) Substituir interações humanas por automação.

QUESTÕES DISCURSIVAS

51. Suponha que você esteja desenvolvendo um dispositivo para um teste de gravidez. Esse dispositivo vai analisar os componentes da urina de uma mulher e vai informar se ela está grávida (Positivo) ou não (Negativo). Para o dispositivo, você decidiu avaliar 3 técnicas supervisionados, sendo: Árvore de Decisão (AD), Redes Neurais (RNA) e k-NN (do Inglês *Nearest Neighbour*). Após várias execuções, você obteve o seguinte resultado

AD
Acurácia:90,6%
Matriz de confusão

	Pos	Neg
Pos	553	47
Neg	47	353

k-NN
Acurácia:90,4%
Matriz de confusão

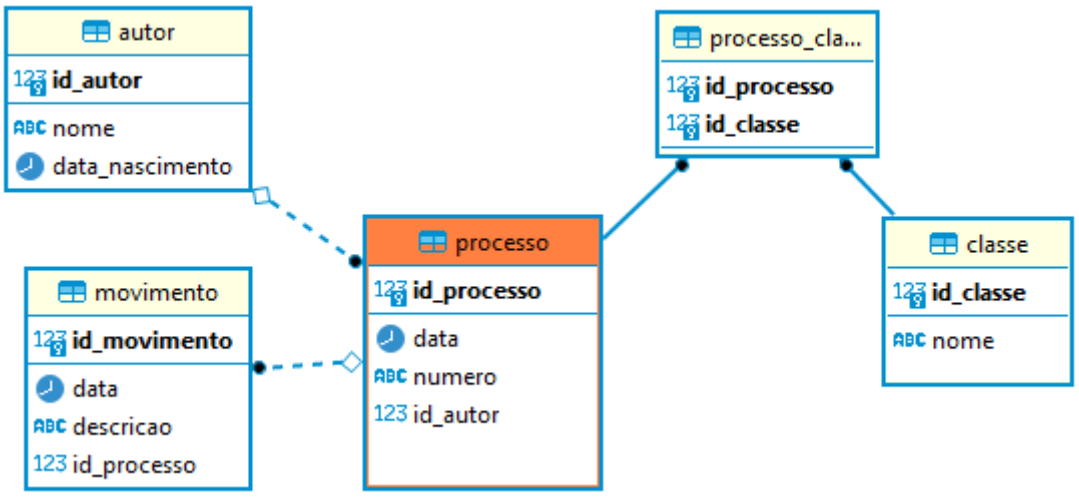
	Pos	Neg
Pos	600	0
Neg	96	304

RNA
Acurácia:90,5%
Matriz de confusão

	Pos	Neg
Pos	505	95
Neg	0	400

Baseado nos resultados acima, defina qual o melhor modelo supervisionado para ser utilizado no dispositivo que você está criando. Explique a sua resposta.

52. Suponha que você é o analista de negócios responsável pela análise de requisitos e análise exploratória de dados de uma empresa. Ciente disso, foi dado o acesso a um banco de dados relacional a você com a seguinte estrutura:



Diante desse modelo relacional, foi solicitada a confecção de um código SQL que forneça todos os dados descritivos de processo, incluindo todos os dados relacionados, exceto as chaves primária e estrangeira.

53. Imagine que você é o responsável pela parte de visualização de dados de uma empresa qualquer. Sabendo disso, o engenheiro de dados disponibilizou um arquivo CSV com os dados que serão visualizados através do painel (*dashboard*) que será desenvolvido por você. Veja abaixo uma pequena amostra dos dados (VendasProdutos.csv).

idProduto	Produto	Categoria	Segmento	Fabricante	idLoja	Cidade	Estado	Vendedor	idVendedor	DataVenda	ValorVenda
SKU-0000001	LG K10 TV Power	Celulares	Corporativo	LG	SP8821	São Paulo	São Paulo	Ana Teixeira	1009	04/10/2012	R$ 679,00
SKU-0000002	Geladeira Duplex	Eletrodomésticos	Doméstico	Brastemp	SP8821	São Paulo	São Paulo	Josias Silva	1006	01/01/2012	R$ 832,00
SKU-0000003	Lavadora 11 Kg	Eletrodomésticos	Doméstico	Brastemp	SP8821	São Paulo	São Paulo	Josias Silva	1006	02/02/2012	R$ 790,00
SKU-0000004	Lavadora 11 Kg	Eletrodomésticos	Doméstico	Brastemp	SP8821	São Paulo	São Paulo	Mateus Gonçalves	1003	03/03/2012	R$ 765,32
SKU-0000005	Lavadora 11 Kg	Eletrodomésticos	Doméstico	Electrolux	SP8821	São Paulo	São Paulo	Artur Moreira	1004	04/04/2012	R$ 459,89
SKU-0000006	Lavadora 11 Kg	Eletrodomésticos	Doméstico	Brastemp	SP8821	São Paulo	São Paulo	Rodrigo Fagundes	1005	04/05/2012	R$ 590,98
SKU-0000007	Geladeira Duplex	Eletrodomésticos	Doméstico	Brastemp	SP8821	São Paulo	São Paulo	Josias Silva	1006	04/06/2012	R$ 1.000,91

Porém, antes de implementar o painel você precisará aplicar os conceitos de modelagem dimensional, e para tal, precisará construir um modelo estrela. Dessa forma, mostre graficamente (diagrama) como ficaria o seu modelo.

54. Imagine que você passou a ser o engenheiro de dados responsável pela construção dos modelos dimensionais de uma empresa, onde todos os modelos são do tipo estrela. Com base no arquivo CSV, da questão anterior, faça uma função em Python, preferencialmente utilizando PySpark, para criação das dimensões denominada “cria_dimensao_csv”, onde o parâmetro dessa função é o arquivo CSV e o retorno dessa função é uma estrutura de dicionário onde a chave é o nome da dimensão e o valor é a própria dimensão. (Considere que cada coluna é uma dimensão e todas são do tipo string).